SAND2019-13835R

# Status Report on Uncertainty Quantification and Sensitivity Analysis Tools in the Geologic Disposal Safety Assessment (GDSA) Framework

## Fuel Cycle Research and Development

***Prepared for***
***U.S. Department of Energy***
***Spent Fuel and Waste Science and Technology Campaign***

***L.P. Swiler, J.C. Helton, E. Basurto, D.M. Brooks, P.E. Mariner, L.M. Moore, S. Mohanty, S.D. Sevougian, and E.R. Stein***

***Sandia National Laboratories***

***September 30, 2019***

**M2SF-19SN010304031**
**SAND2019-xxxx R**


Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Revision 5
1/15/2018

# APPENDIX E
# NTRD DOCUMENT COVER SHEET[1]

| | |
|---|---|
| Name/Title of Deliverable/ Milestone/Revision No. | *Status Report on UQ/SA Tools in GDSA Framework*, M2SF-19SN010304031 |
| Work Package Title & Number | GDSA – Uncertainty and Sensitivity Analysis Methods – SNL, SF-19SN01030403 |
| Work Package WBS Number | 1.08.01.03.04 |
| Responsible Work Package Manager | S. David Sevougian S. David Sevougian |
| | (Name/Signature) |
| Date Submitted | September 30, 2019 |

| Quality Rigor Level for Deliverable/Milestone[2] | ☐ QRL-1 ☐ Nuclear Data | ☐ QRL-2 | ☒ QRL-3 | ☐ QRL-4 Lab QA Program[3] |
|---|---|---|---|---|

This deliverable was prepared in accordance with Sandia National Laboratories
*(Participant/National Laboratory Name)*

QA program which meets the requirements of

☒ DOE Order 414.1 ☐ NQA-1 ☐ Other

**This Deliverable was subjected to:**

☒ Technical Review ☐ Peer Review

**Technical Review (TR)**
**Review Documentation Provided**
☐ Signed TR Report or,
☐ Signed TR Concurrence Sheet or,
☒ Signature of TR Reviewer(s) below
**Name and Signature of Reviewers**
Robert Rechard
Aubrey Eckert

**Peer Review (PR)**
**Review Documentation Provided**
☐ Signed PR Report or,
☐ Signed PR Concurrence Sheet or,
☐ Signature of PR Reviewer(s) below
Gabriel Huerta
David Sassani

**NOTE 1:** *Appendix E should be filled out and submitted with the deliverable. Or, if the PICS:NE system permits, completely enter all applicable information in the PICS:NE Deliverable Form. The requirement is to ensure that all applicable information is entered either in the PICS:NE system or by using the NTRD Document Cover Sheet.*

- *In some cases there may be a milestone where an item is being fabricated, maintenance is being performed on a facility, or a document is being issued through a formal document control process where it specifically calls out a formal review of the document. In these cases, documentation (e.g., inspection report, maintenance request, work planning package documentation or the documented review of the issued document through the document control process) of the completion of the activity, along with the Document Cover Sheet, is sufficient to demonstrate achieving the milestone.*

**NOTE 2**: *If QRL 1, 2, or 3 is not assigned, then the QRL 4 box must be checked, and the work is understood to be performed using laboratory QA requirements. This includes any deliverable developed in conformance with the respective National Laboratory / Participant, DOE or NNSA-approved QA Program.*

**NOTE 3:** *If the lab has an NQA-1 program and the work to be conducted requires an NQA-1 program, then the QRL-1 box must be checked in the work Package and on the Appendix E cover sheet and the work must be performed in accordance with the Lab's NQA-1 program. The QRL-4 box should not be checked*

# ACKNOWLEDGEMENTS

This report has benefited from the insights and assistance of many. The authors thankfully acknowledge Jennifer Frederick and Michael Nole for their contributions to the crystalline reference case, and Adam Stephens for his help trouble-shooting Dakota input decks. Participants in the international uncertainty and sensitivity analysis collaboration provided valuable discussion and perspective. The authors are especially grateful to Klaus Röhlig (TU Clausthal) and Dirk Becker (GRS) for their leadership, Eef Weetjens and Joan Govaerts (SCK-CEN) for hosting the most recent workshop, and Elmar Plischke (TU Clausthal) for sharing the CUSUNORO method and software with us. The authors acknowledge the leadership and career contributions of Robert McKinnon (retired) and Peter Swift in repository science. The authors appreciate and thank the technical reviewers Aubrey Eckert, Gabriel Huerta, Rob Rechard, and Dave Sassani for comments and suggestions that improved the quality of the manuscript. Finally, the authors give a special thanks to Prasad Nair, Mark Tynan (retired), Bill Spezialetti, and Tim Gunter of the Department of Energy's Spent Fuel and Waste Science and Technology campaign for their on-going support and guidance.

# EXECUTIVE SUMMARY

The Spent Fuel and Waste Science and Technology (SFWST) Campaign of the U.S. Department of Energy (DOE) Office of Nuclear Energy (NE), Office of Fuel Cycle Technology (FCT) is conducting research and development (R&D) on geologic disposal of spent nuclear fuel (SNF) and high-level nuclear waste (HLW). Two high priorities for SFWST disposal R&D are design concept development and disposal system modeling. These priorities are directly addressed in the SFWST *Geologic Disposal Safety Assessment* (GDSA) control account, which is charged with developing a geologic repository system modeling and analysis capability, and the associated software, *GDSA Framework*, for evaluating disposal system performance for nuclear waste in geologic media. *GDSA Framework* is supported by SFWST Campaign and its predecessor the Used Fuel Disposition (UFD) campaign.

This report fulfills the GDSA Uncertainty and Sensitivity Analysis Methods work package (SF-19SN01030403) level 2 milestone – *Status Report on UQ/SA Tools in GDSA Framework* (M2SF-19SN010304031). It presents high level objectives and strategy for development of uncertainty and sensitivity analysis tools, documents and demonstrates uncertainty quantification (UQ) and sensitivity analysis (SA) tools in *GDSA Framework* in FY19 and describes additional UQ/SA tools whose future implementation would enhance the UQ/SA capability of *GDSA Framework*. This work was closely coordinated with the other Sandia National Laboratory GDSA work packages: the GDSA Framework Development work package (SF-19SN01030402), the GDSA Repository Systems Analysis work package (SF-19SN01030405), and the GDSA Modeling work package (SF-19SN01030406). This report builds on developments reported in previous *GDSA Framework* milestones, particularly M2SF-18SN01030406.

Tools for uncertainty quantification (UQ) and sensitivity analysis (SA) are essential components of performance/safety assessment. Probabilistic performance assessment (PA) relies on propagating uncertainties in model inputs through a predictive model (or models) to quantify expected outcomes (e.g., mean dose to an individual, groundwater concentration, or cumulative radionuclide release) and associated probability distributions for comparison to regulatory limits. Probabilistic, sampling-based methods of sensitivity analysis are used to understand both qualitatively and quantitatively the contributions of uncertain inputs to uncertainty in model outputs and to enhance confidence in predictive models by providing a check that model behavior is reasonable and expected. UQ/SA is an iterative process. Results may guide further collection of data to reduce uncertainty in model predictions, identify processes for inclusion or exclusion in PA models, help debug numerical models, and enable design of a parsimonious final (regulatory) uncertainty analysis (Figure 1).

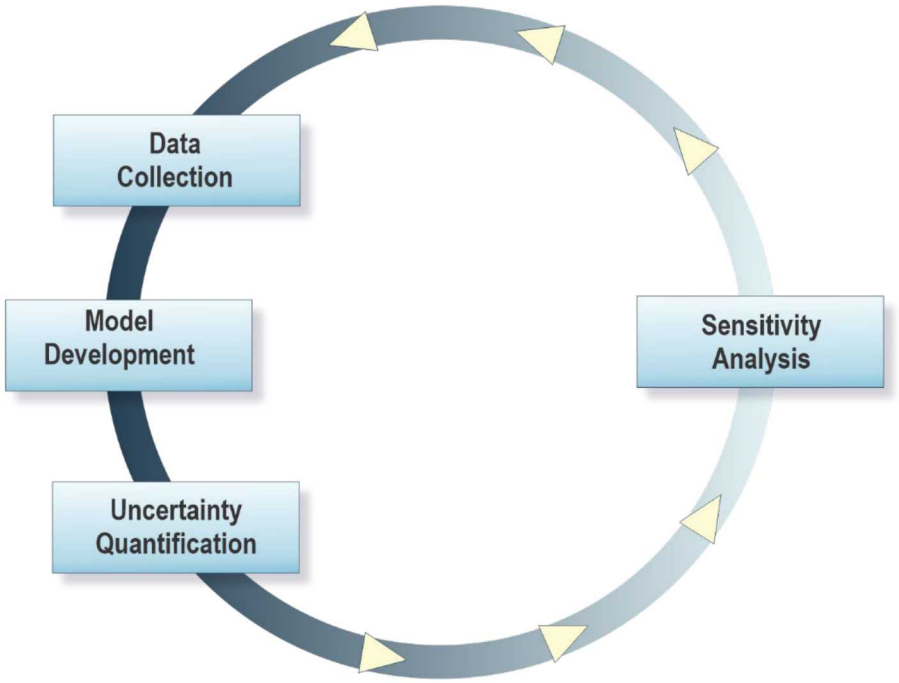



Figure 1. Uncertainty Quantification and Sensitivity Analysis as an iterative process

Development of UQ/SA tools within *GDSA Framework* is driven by the overarching objectives of *GDSA Framework* development including: 1) enabling increasingly coupled, mechanistic multi-physics modeling; 2) leveraging existing high-performance computing capabilities; 3) remaining flexible enough to take advantage of future advances in hardware, software, and simulation and analysis methods; and 4) developing in an open-source environment so that implementations are transparent and software is freely available to stakeholders. Objectives specific to UQ/SA development include: 1) to make available standard sampling-based methods of uncertainty propagation, sensitivity analysis, and uncertainty quantification typically used within U.S. nuclear waste disposal programs; 2) to enable adoption of new methods consistent with the current standard of practice in the UQ/SA community and appropriate for high-dimensional, highly coupled, nonlinear problems resulting from the implementation of mechanistic multi-physics simulations; and 3) to create a consistent, common framework that enables a user to perform a range of uncertainty and sensitivity analyses for a particular problem or set of simulations. These are important goals for performance assessments now and in the future.

There is a rich legacy of UQ/SA in repository safety assessments, predicated on a number of important factors, including regulatory requirements, confidence building for models and data, and optimization of R&D directed at quantifying and reducing uncertainties. In this document, regulatory considerations are documented in Chapter 2, the treatment of epistemic vs. aleatory uncertainty in Chapter 3, and the sampling strategies used in prior safety assessments in Chapter 4. Since the early 2000s, there have been advances in the development of surrogates or emulators for expensive computational simulations as well as advances in the UQ/SA methods, such as the increasing use of variance-based sensitivity indices. The variance-based sensitivity indices are documented in Chapter 5 and Appendix B, surrogates are discussed in Chapters 6 and Appendix

C, and Chapters 7-8 present some of the new UQ/SA methods as applied to two generic reference case models, that of a shale and crystalline repository, respectively. Chapter 9 outlines some recent UQ/SA algorithms, including methods to handle time-varying responses, new graphical sensitivity approaches, and adaptive sampling methods such as importance sampling which can focus additional samples on regions of interest without attempting to explore the entire parameter space with limited samples. Additional methods presented include Bayesian calibration methods which are used to estimate posterior parameter distributions for model parameters given experimental data (and such parameter distributions can be then carried forward in a UQ study), and multi-fidelity UQ methods which combine a small number of high-fidelity simulations with a large number of low-fidelity runs to obtain statistics on the high-fidelity model. Chapter 10 describes collaborations with the international community. Appendix A provides a list of software tools implementing some of the existing and new UQ/SA methods, and Chapter 11 provides a summary of the report.

Development, implementation, and demonstration of new tools and methods for uncertainty and sensitivity analysis in *GDSA Framework* will maintain leadership of the repository science community in UQ/SA methods, while also maintaining an infrastructure of proven tools. Geologic repository performance assessment in the U.S. involves coupled, multi-physics modeling at high resolution with large parameter spaces. It requires high-performance computing and costly sample evaluations. UQ/SA methods discussed in this report, including surrogate modeling to reduce computational expense, variance-based sensitivity analysis to quantify importance of parameter interactions in a multi-physics system, multi-fidelity methods, adaptive importance sampling, etc., will enable analysis methods to keep pace with the increasing sophistication of the physics models. *GDSA Framework* development seeks to keep abreast of improvements to existing UQ/SA methods, employ new methods, and maintain an infrastructure of proven tools that can be extended to support computationally expensive analyses.

# Contents

# Table of Figures

# Table of Tables

# 1. INTRODUCTION

This report presents high level objectives and strategy for development of uncertainty and sensitivity analysis tools in *Geologic Disposal Safety Assessment (GDSA) Framework*, a software toolkit for probabilistic post-closure performance assessment (PA) of systems for deep geologic disposal of nuclear waste. *GDSA Framework* is supported by the Spent Fuel and Waste Science and Technology (SFWST) Campaign of the U.S. Department of Energy (DOE) Office of Nuclear Energy (NE) and its predecessor the Used Fuel Disposition (UFD) campaign.

This report fulfills the GDSA Uncertainty and Sensitivity Analysis Methods work package (SF-19SN01030403) level 2 milestone – *Status Report on UQ/SA Tools in GDSA Framework* (M2SF-19SN010304031). It documents and demonstrates uncertainty quantification (UQ) and sensitivity analysis (SA) tools newly implemented in *GDSA Framework* in FY19, it summarizes the progress of UQ/SA tools during the UFD and SFWST campaigns, and it describes additional UQ/SA tools whose future implementation would enhance the UQ/SA capability of *GDSA Framework*.

This work was closely coordinated with the other SNL GDSA work packages: the GDSA Framework Development work package (SF-19SN01030402), the GDSA Repository Systems Analysis work package (SF-19SN01030405), and the GDSA Modeling work package (SF-19SN01030406). This report builds on developments reported in previous *GDSA Framework* milestones, particularly M2SF-18SN01030406 [1].

The repository community must maintain leadership in UQ/SA methods. Computational codes are becoming increasingly complex and require high performance computers to run, resulting in costly sample evaluations. Geologic repository performance assessment in the U.S. involves a code base that includes coupled, multiphysics modeling at high resolution. Having relatively few simulation samples highlights the need to consider surrogate models to sample and explore the input parameter space more extensively. However, this must be done in a careful way so that surrogate accuracy can be tracked and understood in the context of UQ/SA results. Another approach is to employ "multifidelity" UQ in which many low fidelity simulation runs (e.g., coarser mesh, simpler physics) augment a small number of high fidelity runs. Variance-based sensitivity indices are now a standard practice in the sensitivity analysis community but require a large number of function evaluations of the predictive model. Much research has focused on calculating variance-based sensitivity indices while keeping the computational cost reasonable. Keeping abreast of improvements to existing UQ/SA methods as well as employing new methods is critical to performing sensitivity and uncertainty analysis of new repository systems which will involve large parameter spaces and computationally expensive simulations.

## 1.1 Overview of this Report

This report provides an extensive explanation of UQ/SA methods used in prior geologic disposal performance assessment as well as a look at current and future methods. It is a comprehensive report meant to provide historical context, archive important methodological considerations, document and explain UQ/SA methods, and describe a vision for UQ/SA in Geologic Disposal Safety Assessment going forward. This report will serve as an important reference on these subjects.

The outline of this report is as follows: Chapter 2 provides the regulatory context in geologic disposal and motivation for why and how UQ/SA support regulatory requirements. Chapter 3 presents the conceptual structure of a performance assessment including the treatment of epistemic and aleatory uncertainty. Chapter 4 describes sampling methods for UQ/SA and Chapter 5 describes variance-based approaches for sensitivity analysis. Chapter 6 discusses surrogates or "metamodels" that may be used when the computational cost of performing the UQ, SA, or calibration is too expensive. Chapters 7 and 8 present recent examples applying various UQ/SA methods on two generic reference cases being developed in GDSA: a generic shale host-rock repository and a generic crystalline host-rock repository, respectively. Chapter 9 presents an overview of some new UQ/SA methods and possible future research directions of interest for more comprehensive SA, adaptive sampling, and a capability to inherently address time-varying quantities. Chapter 10 discusses international collaborations around UQ/SA. Chapter 11 presents conclusions. The Appendices include a description of the UQ/SA software tools available as well as more mathematical details on variance-based calculations and polynomial chaos expansions.

## 1.2 GDSA Framework

*GDSA Framework* (Figure 1.1) capabilities include multi-physics simulation of coupled processes affecting deep geologic repository performance, uncertainty and sensitivity analysis, pre- and post-processing, and visualization. For a given performance assessment, these tools will be linked to a version-controlled parameter database and an automated run-control system. The overall objectives of *GDSA Framework* development are to:

- create a framework that is flexible enough to take advantage of future advances in hardware, software, and simulation and analysis methods;
- leverage existing high-performance computing capabilities (e.g., meshing, simulation, analysis, and visualization);
- enable increasingly coupled, mechanistic multi-physics modeling;
- provide analysis methods for prioritization of SFWST Disposal Research R&D activities;
- provide transparent implementation of simulation and analysis methods;
- develop and distribute in an open-source environment so that software is freely available to stakeholders ([1;2;3;4]).

Objectives specific to the uncertainty and sensitivity analysis capability in *GDSA Framework* are to make available standard sampling-based methods of uncertainty propagation, sensitivity analysis, and uncertainty quantification typically used within U.S. nuclear waste disposal programs (e.g., DOE 2008 [5], DOE 2014 [6]); and to enable future adoption of new methods consistent with the current standard of practice in the UQ/SA community and appropriate for high-dimensional, highly coupled, nonlinear problems resulting from the implementation of mechanistic multi-physics simulations. Having a consistent, common framework which enables a user to perform a range of sensitivity analysis and UQ approaches for a particular problem or set of simulations allows for reproducibility, comparative analyses, use of verified algorithms, and

documentation of best practices. These are important goals for performance assessments now and in the future.

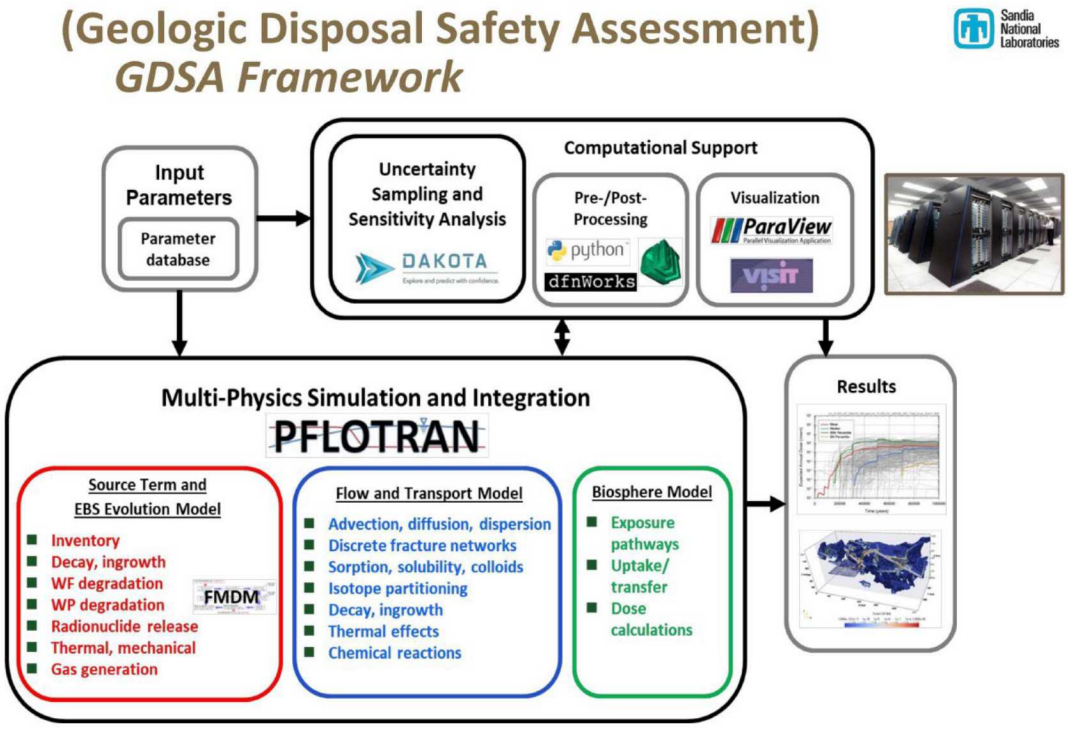



Figure 1.1 GDSA Framework

The following sections highlight the key components of *GDSA Framework*. More information about each can be found by following the links at https://pa.sandia.gov.

### 1.2.1 PFLOTRAN

PFLOTRAN is an open source, state-of-the-art, massively parallel subsurface flow and reactive transport simulator ([7;8;9]). Written in object-oriented Fortran 2003, PFLOTRAN is a porous medium continuum code for modeling multicomponent, multiphase flow and transport, heat conduction and convection, biogeochemical reactions, geomechanics, and isotope decay and ingrowth. The code is developed under a GNU LGPL license allowing for third parties to interface proprietary software with the code. The availability and continuing development of PFLOTRAN are due to an ongoing collaborative effort of several DOE laboratories led by Sandia. PFLOTRAN development for *GDSA Framework* is described by Mariner et al. ([1;2;3;4]) and Sevougian et al. 2018 [10]. PFLOTRAN installation instructions are available at: https://www.pflotran.org.

### 1.2.2 Dakota

Dakota is an open-source toolkit of algorithms that contains both state-of-the-art research and robust, usable software for optimization and uncertainty quantification (UQ). It is available at: https://dakota.sandia.gov [11]. The algorithms allow a user to explore a computational simulation to answer questions such as:

- what is the best design?

- how safe is this design?
- what are the most important parameters?
- what effects do uncertainties have on my system?

The Dakota software has a rich set of parametric analysis methods that enable design exploration, model calibration, risk analysis, and quantification of margins and uncertainty with computational models. Dakota provides a flexible, extensible interface between simulation codes and these iterative analysis methods, which include:

- optimization with gradient and nongradient-based methods;
- uncertainty quantification with sampling, reliability, stochastic expansion, and epistemic methods;
- parameter estimation using nonlinear least squares (deterministic) or Bayesian inference (stochastic); and
- sensitivity/variance analysis with design of experiments and parameter study methods.

These capabilities may be used on their own or as components within advanced strategies such as hybrid optimization, surrogate-based optimization, mixed-integer nonlinear programming, or optimization under uncertainty. Dakota is a C++ code and has been under development at Sandia since 1994 and has been primarily sponsored by DOE's Advanced Simulation and Computing (ASC) program. As such, it has a focus on interfacing to and running simulations which are computationally expensive, require high performance computing and parallel execution, and exhibit nonlinearities, non-monotonic and/or discontinuous responses, and often involve noisy responses and high-dimensional inputs. Thus, a focus of the algorithm development in Dakota has been on methods that are as efficient as possible and minimize the number of runs required of a high-fidelity simulation model. Such algorithms include surrogate or emulator models, adaptive sampling approaches, and multifidelity UQ methods which augment a small number of high fidelity runs with many low fidelity runs to obtain comparable accuracy in statistical estimators.

Dakota contains the uncertainty quantification and sensitivity analysis methods typically used in the U.S. repository program. Dakota implements Latin Hypercube Sampling (LHS) with correlation control on input parameters. It calculates moments on responses of interest as well as correlation matrices (simple, partial, and rank correlations) between inputs and outputs. Dakota also contains an algorithm for performing incremental LHS which allows one to double an initial LHS study such that the second LHS study is a Latin design and the combined initial and second LHS studies together form a Latin hypercube design. Dakota allows nested studies to perform an "outer loop" epistemic sampling and an "inner loop" aleatory sampling to generate ensembles of distributions. Dakota returns tables of input and output amenable to further processing and visualization with additional tools developed within *GDSA Framework* or by an individual user. Additional methods that have been implemented in Python for use in *GDSA Framework* include calculation of standardized regression coefficients via stepwise linear regression and calculation of partial correlation coefficients via iterative loop [2].

A graphical depiction of Dakota interfacing with a computational model such as a repository simulation in PFLOTRAN is shown in Figure 1.2. Based on the type of study being performed (optimization, uncertainty quantification, etc.), Dakota chooses the next set of parameters at which

to evaluate the simulator and runs the simulator, which returns the performance metrics of interest back to Dakota. Dakota then generates the next set of parameters according to the algorithm being used for the study and keeps iterating until the specified number of samples is reached.

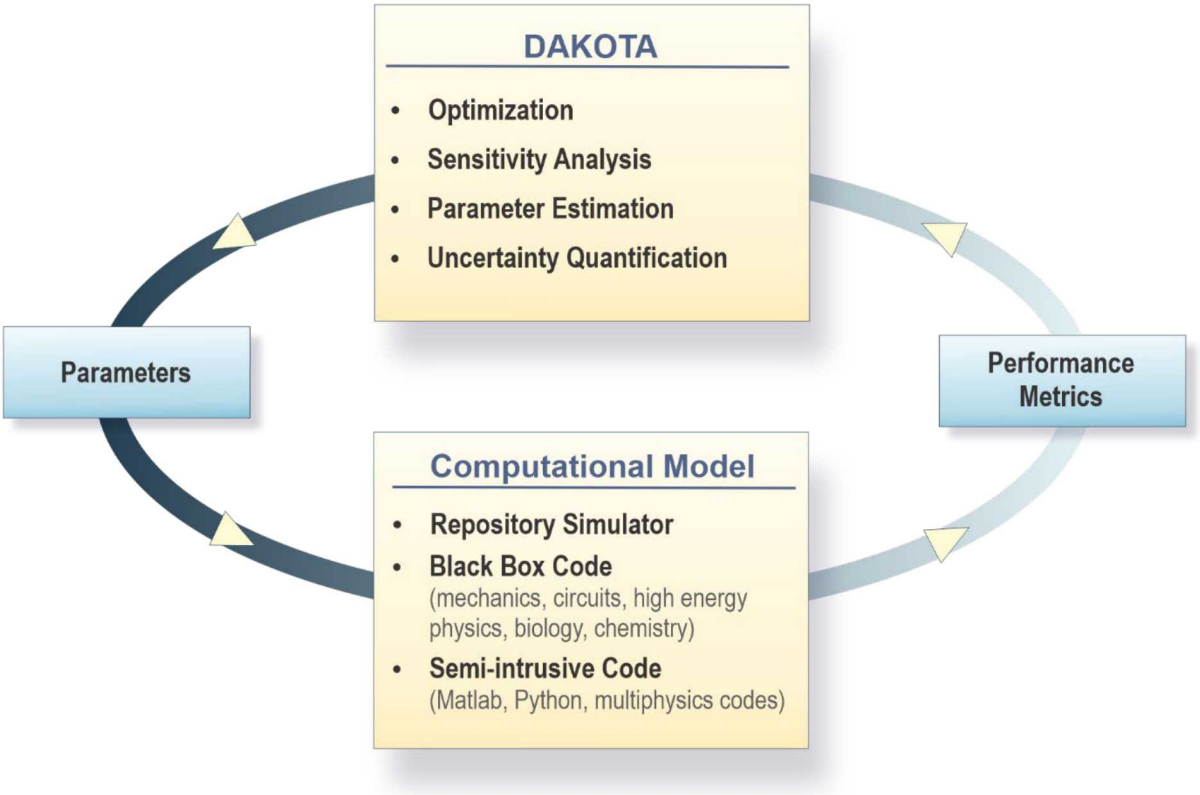



Figure 1.2 Dakota interfacing to a computational model such as a repository simulator

The UQ/SA methods in Dakota have evolved as the standard of practice evolves. Over the past ten years, the Dakota team has invested in methods which calculate the Sobol' variance-based sensitivity indices in an efficient manner. Currently, a Dakota user can calculate these by extensive sampling of the simulation code, by using surrogate methods such as regression or Gaussian process models, and by the use of polynomial chaos expansions. These advanced methods are presented in more detail later in this report for possible inclusion in *GDSA Framework*. Dakota is an actively maintained and developed code with formal releases issued twice per year. Dakota uses formal software quality development processes including advanced version control, unit and regression testing, agile programming practices, and software quality assessment.

### 1.2.3 Visualization Tools

*GDSA Framework* employs ParaView and/or VisIT for visualization of results. ParaView is an open-source, multi-platform data analysis and visualization application developed by Sandia National Laboratories. As stated on the ParaView website (https://www.paraview.org/): "ParaView users can quickly build visualizations to analyze their data using qualitative and quantitative techniques. The data exploration can be done interactively in 3D or or

programmatically using ParaView's batch processing capabilities. ParaView was developed to analyze extremely large datasets using distributed memory computing resources. It can be run on supercomputers to analyze datasets of petascale size as well as on laptops for smaller data, has become an integral tool in many national laboratories, universities and industry, and has won several awards related to high performance computation."

VisIT has been developed at Livermore National Laboratory. As stated on the VisIT website (https://wci.llnl.gov/simulation/computer-codes/visit): "VisIT is an open source, interactive, scalable, visualization, animation, and analysis tool. From Unix, Windows, or Mac workstations, users can interactively visualize and analyze data ranging in scale from small ($<10^1$ cores) desktop-sized projects to large ($>10^5$ cores) leadership-class computing facility simulation campaigns. Users can quickly generate visualizations, animate them through time, manipulate them with a variety of operators and mathematical expressions, and save the resulting images and animations for presentations. VisIt contains a rich set of visualization features to enable users to view a wide variety of data including scalar and vector fields defined on two- and three-dimensional (2D and 3D) structured, adaptive and unstructured meshes. Owing to its customizeable plugin design, VisIt is capabable of visualizing data from over 120 different scientific data formats."

### 1.2.4 GDSA Software Strategy

The software strategy for GDSA is to leverage and use open source software that is actively maintained and developed, whenever possible. That is why the *GDSA Framework* utilizes PFLOTRAN, Dakota, Paraview, and VisIT. Another goal is to support HPC computing, which is a primary focus for all of the software tools listed above. In addition, the *GDSA Framework* should have the flexibility to develop and adopt new capabilities as state-of-the-art hardware, software, and methodology evolves. Again, the codes chosen for GDSA exhibit this flexibility and are constantly evolving and adopting to utilize new software and hardware capabilities. Much of the interfacing between the codes is currently performed with Python scripts and other scripting tools, and there will be tighter integration as *GDSA Framework* progresses.

## 1.3 References: Chapter 1

1. Mariner, P. E., E. R. Stein, S. D. Sevougian, L. J. Cunningham, J. M. Frederick, G. E. Hammond, T. S. Lowry, S. Jordan, and E. Basurto (2018). *Advances in Geologic Disposal Safety Assessment and an Unsaturated Alluvium Reference Case*. SFWD-SFWST-2018-000509; SAND2018-11858R. Sandia National Laboratories, Albuquerque, NM.
2. Mariner, P. E., E. R. Stein, J. M. Frederick, S. D. Sevougian, and G. E. Hammond (2017). *Advances in Geologic Disposal System Modeling and Shale Reference Cases*. SFWD-SFWST-2017-000044 / SAND2017-10304R. Sandia National Laboratories, Albuquerque, NM.
3. Mariner, P. E., E. R. Stein, J. M. Frederick, S. D. Sevougian, G. E. Hammond, and D. G. Fascitelli (2016). *Advances in Geologic Disposal System Modeling and Application to*

*Crystalline Rock*. FCRD-UFD-2016-000440 / SAND2016-9610R. Sandia National Laboratories, Albuquerque, NM.

4. Mariner, P. E., W. P. Gardner, G. E. Hammond, S. D. Sevougian, and E. R. Stein (2015). *Application of Generic Disposal System Models*. FCRD-UFD-2015-000126 / SAND2015-10037R. Sandia National Laboratories, Albuquerque, NM.
5. DOE (2008). Yucca Mountain Repository License Application Safety Analysis Report. DOE/RW-0573, Revision 1, http://www.nrc.gov/waste/hlw-disposal/yucca-lic-app/yucca-lic-app-safety-report.html#1. US Department of Energy, Washington, DC.
6. DOE (2014). Title 40 CFR Part 191 Subparts B and C Compliance Recertification Application 2014 for the Waste Isolation Pilot Plant. DOE/WIPP 15-3503. U.S. Department of Energy, Carlsbad, New Mexico.
7. Lichtner, P. C. and G. E. Hammond (2012). Quick Reference Guide: PFLOTRAN 2.0 (LA-CC-09-047) Multiphase-Multicomponent-Multiscale Massively Parallel Reactive Transport Code. LA-UR-06-7048. December 8, 2012. Los Alamos National Laboratory, Los Alamos, New Mexico.
8. Hammond, G. E., P. C. Lichtner, R. T. Mills, and C. Lu. (2008). Toward petascale computing in geosciences: application to the Hanford 300 Area. In R. L. Stevens (Ed.), *Scidac 2008: Scientific Discovery through Advanced Computing* (Vol. 125).
9. Hammond, G. E., P. C. Lichtner, and R. T. Mills (2014). "Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN". Water Resources Research, 50(1), 208-228. doi: 10.1002/2012wr013483.
10. Sevougian, S. D., G. E. Hammond, P. E. Mariner, E. R. Stein, J. M. Frederick, and R. J. MacKinnon (2018). "*GDSA Framework*: High-Performance Safety Assessment Software to Support the Safety Case." in *Proceedings of the IGSC Safety Case Symposium 2018*, Rotterdam, The Netherlands, October 10-11, 2018, SAND2018-9975C, Sandia National Laboratories. Albuquerque, NM.
11. Adams, B. M., M. S. Ebeida, M. S. Eldred, G. Geraci, J. D. Jakeman, K. A. Maupin, J. A. Monschke, J. A. Stephens, L. P. Swiler, D. M. Vigil, T. M. Wildey and et. al. (2018). Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.8 User's Manual. Sandia National Laboratories. Albuquerque, NM: 384.

# 2. QUANTIFYING UNCERTAINTY IN A REGULATORY CONTEXT

The ultimate structure of a given post-closure performance assessment depends on the regulatory context in which it is performed. In the U.S., safety standards and implementing regulations are promulgated by the U.S. Environmental Protection Agency (EPA) and the U.S. Nuclear Regulatory Commission (NRC). Safety standards and guidance are also available from national programs in other countries and from international organizations (e.g., IAEA 2011[1]; 2012[2]). Regulations govern the criteria against which the performance of a geologic disposal system is to be judged (e.g., radionuclide releases to the accessible environment, groundwater concentrations, doses to members of the public, acceptable levels of risk), the time frame for which compliance must be demonstrated, how to quantify and present PA model outputs in terms of statistical estimators (e.g., mean) for comparison to regulatory limits, selection of future scenarios, and treatment of uncertainties [3].

This section describes the regulatory context in which it is assumed *GDSA Framework* might be used in the future. The content draws primarily from the safety case for (generic) deep borehole disposal [4] and articles describing the regulatory frameworks of the Waste Isolation Pilot Plant [3] and Yucca Mountain [5].

## 2.1 The Safety Case

A widely accepted approach for documenting the basis for the understanding of a geologic disposal system, describing the key justifications for its safety, and acknowledging the unresolved uncertainties and their safety significance is a structured document, or set of documents, known as a safety case. The formal concept of a safety case for the long-term disposal of SNF and HLW in an engineered facility located in a deep geologic formation was first introduced by the Nuclear Energy Agency (NEA) (NEA 1999a [6]). Initial discussion and documentation on the topic continued in NEA (2002 [7]), NEA (2004 [8]), and IAEA (2006 [9]). More recently, there have been a number of international symposia, conferences, working groups, and summary papers devoted to understanding, developing, and/or summarizing the nature, purpose, context, and elements of safety cases (e.g., NEA 2008 [10]; NEA 2009 [11]; IAEA 2011 [1]; IAEA 2012 [2]; NEA 2012 [12]; NEA 2013 [13]; and NEA 2018 [15]). In these recent summary and overview reports, it is observed that there is notable convergence in the understanding and development of safety case documents published by national and international organizations. In parallel, UFD/SFWST has published safety case overviews relevant to geologic disposal in the U.S. [14][4]. The following excerpt from NEA (2012, Section 3.1 [12]) provides a definition of a safety case that is current and consistent with the aforementioned documents:

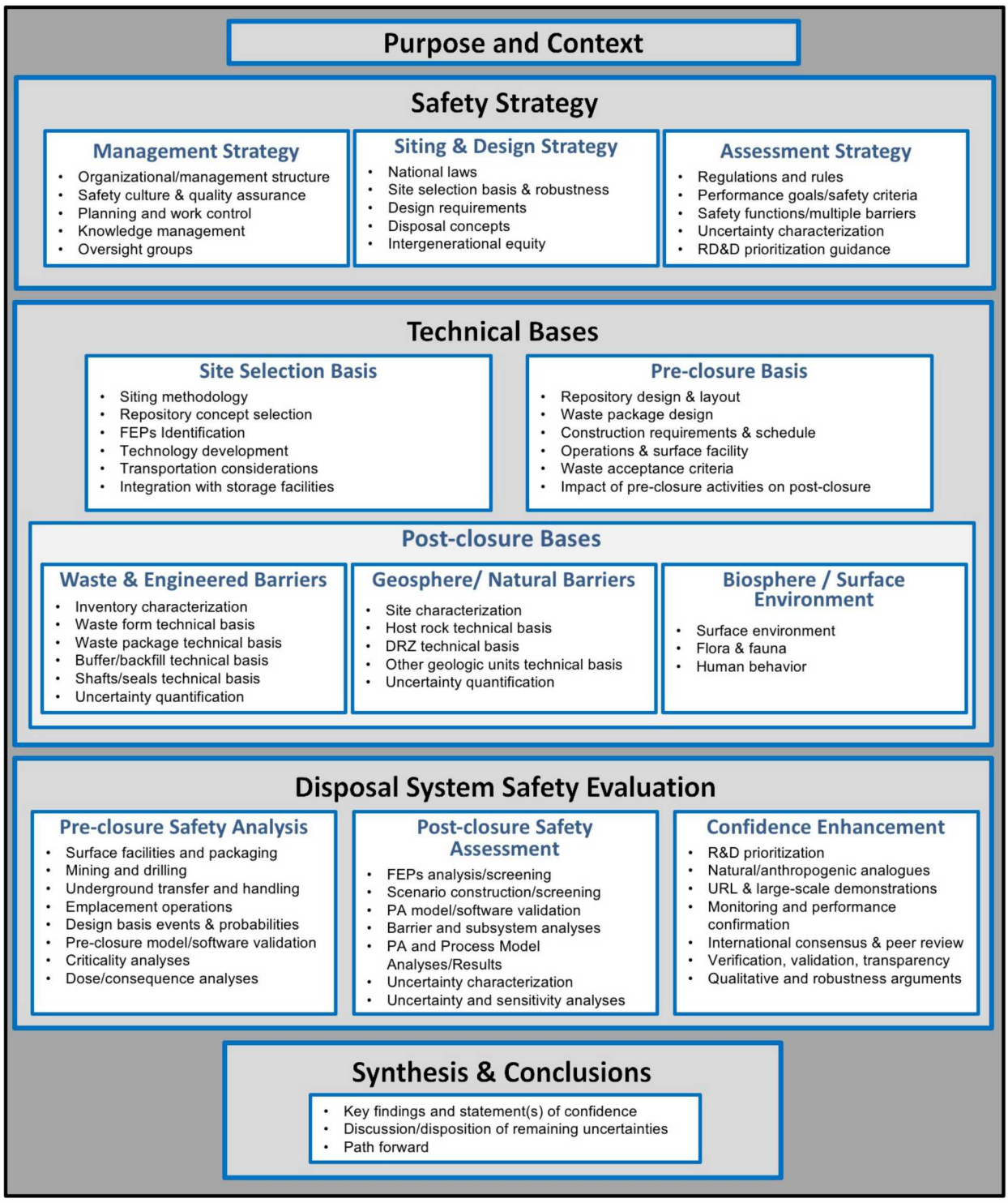
> *The safety case is an integration of arguments and evidence that describe, quantify and substantiate the safety of the geological disposal facility and the associated level of confidence.*

A number of elements contribute to and must be described in the safety case (Figure 2.1), central among them is the safety assessment. There are some differences in the use of the term safety assessment across national programs and over time; the definition used by Freeze et al. (2016) [4] is:

*Safety Assessment* – An iterative set of assessments for evaluating the performance of a repository system and its potential impact that aims to provide reasonable assurance that the

repository system will achieve sufficient safety and meet the relevant requirements for the protection of humans and the environment over a prolonged period. The role of a safety assessment, in a safety case, is (i) to quantify the repository system performance for all selected situations and (ii) to evaluate the level of confidence (taking into account the identified uncertainties) in the estimated performance of the system (NEA 2013, Section 5.1 [13]). This encompasses all aspects that are relevant for the safety of the development, operation and closure of the disposal facility, including qualitative aspects, non-radiological issues, and organizational and managerial aspects (IAEA 2012, Section 4.41 [2]).

The scope of this definition has broadened recently to include not just quantitative analyses, but also a broad range of complementary qualitative evidence and arguments that support the reliability of the quantitative analyses (NEA 2013, Section 1 [13]). The quantitative components of a safety assessment ("Disposal System Safety Evaluation" in Figure 2.1) include both "pre-closure safety analysis" and "post-closure performance assessment". The *GDSA Framework* development is for the purpose of post-closure performance assessment.




Modified from NEA (2013, Figure 2.1) [13]

Figure 2.1 Key Elements of a Safety Case [21; 22].

## 2.2 Post-Closure Performance Assessment

Post-closure performance assessment and its key requirements are defined by regulation, which may be written for the general case, or for a specific deep geologic repository. For instance, the general regulations at 40 CFR § 191.12 [16] define post-closure performance assessment as an analysis that:

1. Identifies the processes and events that might affect the disposal system;

2. Examines the effects of these processes and events on the performance of the disposal system; and

3. Estimates the cumulative releases of radionuclides, considering associated uncertainties, caused by all significant processes and events. These estimates shall be incorporated into an overall probability distribution of cumulative release to the extent practicable.

Regulations applicable to a repository at Yucca Mountain contain a similar definition of post-closure PA, but require that dose to an individual be estimated rather than cumulative release of radionuclides. At 40 CFR § 197.12 [16] performance assessment is defined as an analysis that:

1. Identifies the features, events, and processes, (except human intrusion), and sequences of events and processes (except human intrusion) that might affect the Yucca Mountain disposal system and their probabilities of occurring;

2. Examines the effects of those features, events, processes, and sequences of events and processes upon the performance of the Yucca Mountain disposal system; and

3. Estimates the annual committed effective dose equivalent incurred by the reasonably maximally exposed individual, including the associated uncertainties, as a result of releases caused by all significant features, events, processes, and sequences of events and processes, weighted by their probability of occurrence.

Both of these definitions require accounting for uncertainties in a probabilistic fashion. Development of UQ/SA capability in *GDSA Framework* is intended to create a post-closure PA platform for use in a regulatory environment in which probabilistic PA is required and post-closure safety criteria (such as standards for containment or dose) may be influenced as described in the following section.

## 2.3 Regulatory Criteria and Standards for Post-Closure PA

In a future regulatory environment, post-closure safety standards may be influenced by (following Freeze et al. 2016 [4]):

- existing general criteria and standards (10 CFR § 60 [17] and 40 CFR § 191 [18]),

- the risk-informed approach in 10 CFR § 63 [16] and 40 CFR § 197 [19],

- and dose or risk metrics recognized internationally to be important to establishing repository safety (e.g., IAEA 2011 [1]; 2012 [2]).

Relevant portions of the aforementioned regulations that might provide insights to future regulations are excerpted in Freeze et al. 2016 [4], Appendix A.

Existing general regulations in the U.S. for disposal of high-activity wastes in geologic repositories are 10 CFR 60 (NRC) [17] and 40 CFR 191 (EPA) [18], first promulgated in 1983 and 1985, respectively. Regarding post-closure PA, 10 CFR 60.112 together with 40 CFR 191.13, 191.15, and 191.24 establish a containment standard, an individual protection standard, a groundwater protection standard, and a time frame for compliance of 10,000 years after disposal. The following descriptions of the standards are excerpted from [4]:

- The containment requirements at 40 CFR 191.13(a) state that the cumulative releases of radionuclides to the accessible environment for 10,000 years after disposal shall: (1) have a likelihood of less than one chance in 10 of exceeding the release limits; and (2) have a likelihood of less than one chance in 1,000 of exceeding ten times the release limits. The release limits, listed in Appendix A of 10 CFR 191, are radionuclide-specific limits expressed in units of cumulative radionuclide release (curies) per 1,000 metric tons of heavy metal (MTHM) of spent nuclear fuel or equivalent.

- 40 CFR § 191.15(a) limits the annual dose to any member of the public in the accessible environment from undisturbed performance of the disposal system (i.e., "not disrupted by human intrusion or the occurrence of unlikely natural events" (40 CFR 191.12)) to ≤ 15 mrem/yr for 10,000 years after disposal.

- 40 CFR § 191.24(a)(1) limits the levels of radioactivity in any underground source of drinking water, in the accessible environment, from 10,000 years of undisturbed disposal system performance after disposal, to the limits specified in 40 CFR 141 (the National Primary Drinking Water Regulations). Maximum contaminant levels for radionuclides are specified in 40 CFR 141.66 as:

    - ≤ 5 pCi/L for combined radium-226 and radium-228,

    - ≤ 15 pCi/L for gross alpha activity (including radium-226 but excluding radon and uranium),

    - ≤ 4 mrem/yr (to the total body or any internal organ, based on drinking 2 liters of water per day) for combined beta particle and photon radioactivity, and

    - ≤ 30 μg/L for uranium.

The containment criteria are the focus of the performance assessment for the Waste Isolation Pilot Plant (WIPP), a deep geologic repository for transuranic waste in bedded salt [3]. Many of the features of the statistical and probabilistic treatment of uncertainty in the performance assessment are specified or constrained by 40 CFR § 194, including for instance [3]:

- 40 CFR § 194.34(a) requires that PA results be assembled into complementary, cumulative distribution functions (CCDFs) that represent the probability of exceeding various levels of cumulative releases caused by all significant processes and events.

- 40 CFR § 194.34 also discusses development of probability distributions for uncertain values, random sampling, and criteria for determining the number of CCDFs needed.

- 40 CFR § 194.25(a) specifies that PA "shall assume that characteristics of the future remain what they are at the time the compliance application is prepared, provided that such characteristics are not related to hydrogeologic, geologic or climatic conditions." This directive prescribes the human activities considered in the PA, and the conceptual models describing the effects of those human activities, i.e., potash mining and drilling.

- 40 CFR § 194.32(b) limits the occurrence of potash mining to once per 10,000 years. Therefore, multiple mining events are not modeled.

- 40 CFR § 194.33(b) requires that future drilling events be represented as random in time based on the historical drilling rate.

These and other regulations constrain the structure of the WIPP PA (described in Chapter 3), by requiring specific approaches to addressing uncertainty, including the use of random sampling, the construction of CCDFs, and comparison of results to cumulative release limits.

In contrast, more recent regulatory criteria in the U.S. and internationally ([1],[5]) emphasize risk to the individual. Risk-informed post-closure safety criteria are found in 10 CFR § 63 [16] and 40 CFR § 197 [19], both of which were first promulgated in 2001 [5]. These regulations apply specifically to geologic disposal of spent nuclear fuel and high-level waste at Yucca Mountain. They are the result of an evolution of safety criteria over the course of the project, driven in part by recommendations made by the National Academy of Science (NAS) in 1995 [20]. Rechard et al. [5] discuss the most influential recommendations, including: (1) use a maximum individual risk evaluated from an effective dose; (2) evaluate compliance at the time of peak risk; (3) evaluate only the potential consequences (not probability) of selected inadvertent human intrusion scenarios." The standards established in response to the recommendations include:

- 10 CFR § 63.311 and 40 CFR § 197.20 establish individual protection standards for two time periods after permanent closure. Dose limits to the reasonably maximally exposed individual (RMEI) are:
  - 0.15 mSv (15 mrem) for 10,000 years following disposal; and
  - 1.0 mSv (100 mrem) after 10,000 years, but within the period of geologic stability ($10^6$ y).
- 10 CFR § 63.321 establishes the same individual protection standards for a prescribed human intrusion scenario.
- 10 CFR § 63.331 and 40 CFR § 197.30 establish standards for protection of groundwater that are identical to those at 40 CFR § 191.24 except for the absence of a limit on U concentration.
  - ≤ 5 pCi/L for combined radium-226 and radium-228,
  - ≤ 15 pCi/L for gross alpha activity (including radium-226 but excluding radon and uranium), and

- ≤ 4 mrem/yr (to the total body or any internal organ, based on drinking 2 liters of water per day) for combined beta particle and photon radioactivity.

10 CFR § 63 requires that performance assessment be used to demonstrate a reasonable expectation that the standards will be met and requires PA to consider "the full range of defensible and reasonable parameter distributions" (10 CFR § 63.304). As in the WIPP PA, many of the features of the statistical and probabilistic treatment of uncertainty are specified or constrained by regulation, including:

- Assumptions that reduce uncertainty about the biosphere (10 CFR § 63.305), the RMEI (10 CFR § 63.312), and the human intrusion scenario (10 CFR § 63.322), for instance,
  - 10 CFR § 63.305(a) "Features, events, and processes that describe the reference biosphere must be consistent with present knowledge of the conditions in the region surrounding the Yucca Mountain site."
  - 10 CFR § 63.312(b) The RMEI "has a diet and living style representative of the people who now reside in the Town of Amargosa Valley, Nevada. DOE will use surveys of the people residing in the Town of Amargosa Valley, Nevada, to determine their current diets and living styles and use the mean of these factors in the assessments conducted for §§ 63.311 and 63.321."
  - 10 CFR § 63.322(a) "There is a single human intrusion as a result of exploratory drilling for groundwater;" and (b) "The intruders drill directly through a degraded waste package into the uppermost aquifer underlying the Yucca Mountain repository."
  - 10 CFR § 63.342(2) "DOE must assess the effects of climate change. The climate change analysis may be limited to the effects of increased water flow through the repository as a result of climate change, and the resulting transport and release of radionuclides to the accessible environment. The nature and degree of climate change may be represented by constant-in-time climate conditions. The analysis may commence at 10,000 years after disposal and shall extend through the period of geologic stability. The constant-in-time values to be used to represent climate change are to be the spatial average of the deep percolation rate within the area bounded by the repository footprint. The constant-in-time deep percolation rates to be used to represent climate change shall be based on a lognormal distribution with an arithmetic mean of 41 mm/year (1.6 in./year) and a standard deviation of 33 mm/year (1.3 in./year). The lognormal distribution is to be truncated so that the deep percolation rates vary between 10 and 100 mm/year (0.39 and 3.9 in./year)."
- Use of the mean for comparison to compliance criteria:
  - 10 CFR § 63.303 "Compliance is based upon the arithmetic mean of the projected doses from DOE's performance assessments for the period within 1 million years after disposal."

Later guidance from the NRC requires a display of the 5th and 95th percentiles of dose about the expected (mean) value, thereby retaining the intent of the earlier regulations to not focus on one measure of the distribution but to examine the entire distribution [23].

The shift toward assessment of peak dose results in conceptual changes to a PA. Some of the differences identified by Rechard et al. [5] include:

- Assessment of peak dose removes the arbitrary limit on the length of the regulatory period. Whereas the general requirements at 40 CFR 191.13(a) impose a 10,000 year regulatory period, the concept of peak dose effectively extends the regulatory period indefinitely, or for the period of "geologic stability."

- Calculation of a peak dose takes credit for (and requires characterization of) dilution and dispersion in the natural system.

- Peak dose is more sensitive to the release rate of radionuclides from the repository than cumulative release is, motivating a higher-fidelity source-term model.

- The dose calculation requires a biosphere model which (though having assumptions constrained by regulation) is defended by the licensee, removing the burden on the regulator of defending the logic behind the containment criteria.

- Peak dose is more susceptible to "risk dilution" than the containment requirement. Risk dilution occurs when overly broad uncertainty distributions are used for key parameters, thus shifting statistical estimators (mean, median, etc.) to lower values.

Assessing risk to an individual is consistent with the risk-informed approach recommended by the International Atomic Energy Agency (IAEA), which recommends a maximum health (fatality) risk of $10^{-5}$/yr or maximum dose limit from a disposal facility of 0.3 mSv/yr ([1][5]). Nations that have adopted an annual dose limit include Germany, France, the Czech Republic, Finland, Switzerland, Hungary, the Republic of Korea, the Netherlands, Spain, and the Slovak Republic. Sweden, Canada, and the United Kingdom set the cancer fatality frequency a factor of 10 below the recommended maximum health risk for $10^4$ years [5]. As suggested by Freeze et al. [4], it seems likely that future post-closure regulations will be dose and/or risk-based standards to at least 1,000,000 years after closure, consistent with IAEA guidelines, international standards, and Yucca Mountain post-closure PA requirements.

The objective of *GDSA Framework* development is to provide a post-closure PA toolkit that 1) has state-of-the-art subsurface flow and reactive transport modeling and simulation capabilities and 2) addresses the computational challenges of "demonstrating a reasonable expectation that standards will be met" in a regulatory framework that focuses on risk to the individual by developing mechanistic models of processes controlling radionuclide release and transport; leveraging high-performance computing; and implementing computationally efficient methods for quantifying and reducing uncertainty; and evaluating the effect of uncertainty on compliance criteria, such as dose, through sensitivity analysis methods. The rest of this report discusses these methods.

## 2.4 References: Chapter 2

1. IAEA (International Atomic Energy Agency) 2011. Disposal of Radioactive Waste, Specific Safety Requirements. IAEA Safety Standards Series No. SSR-5, IAEA, Vienna, Austria.
2. IAEA (International Atomic Energy Agency) 2012. The Safety Case and Safety Assessment for the Disposal of Radioactive Waste, Specific Safety Guide. IAEA Safety Standards Series No. SG-23, IAEA, Vienna, Austria.
3. Howard, B. A., M. B. Crawford, D. A. Galson, and M. G. Marietta 2000. "Regulatory basis for the Waste Isolation Pilot Plant performance assessment". *Reliability Engineering & System Safety, 69*(1-3), 109-127. doi: 10.1016/s0951-8320(00)00028-4
4. Freeze, G. A., E. R. Stein, L. Price, R. MacKinnon, and J. Tillman 2016. Deep Borehole Disposal Safety Analysis. FCRD-UFD-2016-000075 / SAND2016-10949R. Sandia National Laboratories, Albuquerque, NM.
5. Rechard, R. P., T. A. Cotton, and M. D. Voegele 2014. "Site selection and regulatory basis for the Yucca Mountain disposal system for spent nuclear fuel and high-level radioactive waste". *Reliability Engineering & System Safety, 122*, 7-31. doi: 10.1016/j.ress.2013.06.021
6. NEA (Nuclear Energy Agency) 1999a. *Confidence in the Long-term Safety of Deep Geological Repositories: Its Development and Communication*. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
7. NEA (Nuclear Energy Agency) 2002. *Establishing and Communicating Confidence in the Safety of Deep Geologic Disposal: Approaches and Arguments*. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
8. NEA (Nuclear Energy Agency) 2004. *Post-Closure Safety Case for Geological Repositories, Nature and Purpose.* NEA Report No. 3679. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
9. IAEA (International Atomic Energy Agency) 2006. Geologic Disposal of Radioactive Waste, Safety Requirements. IAEA Safety Standards Series No. WS-R-4, IAEA, Vienna, Austria.
10. NEA (Nuclear Energy Agency) 2008. *Safety Cases for Deep Geological Disposal of Radioactive Waste: Where Do We Stand? Symposium Proceedings Paris, France, 23-25 January 2007*, NEA Report No. 6319. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
11. NEA (Nuclear Energy Agency) 2009. *International Experiences in Safety Case for Geological Repositories (INTESC)*, NEA Report No. 6251. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
12. NEA (Nuclear Energy Agency) 2012. *Methods for Safety Assessment of Geological Disposal Facilities for Radioactive Waste: Outcomes of the NEA MeSA Initiative.* NEA No. 6923. Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.
13. NEA (Nuclear Energy Agency) 2013. *The Nature and Purpose of the Post-Closure Safety Cases for Geological Repositories*, NEA/RWM/R(2013)1, Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France.

14. Freeze, G., M. Voegele, P. Vaughn, J. Prouty, W.M. Nutt, E. Hardin, and S.D. Sevougian 2013a. *Generic Deep Geologic Disposal Safety Case*. FCRD-UFD-2012-000146 Rev. 1, SAND2013-0974P. Sandia National Laboratories, Albuquerque, NM.

15. NEA (Nuclear Energy Agency) 2018. *IGSC Safety Case Symposium 2018 – Current Understanding and Future Direction for the Geological Disposal of Radioactive Waste*, October 10-11, 2018, Rotterdam, Netherlands, Organisation for Economic Co-operation and Development, Nuclear Energy Agency, Paris, France (https://www.oecd-nea.org/rwm/workshops/igsc2018/ )

16. 10 CFR Part 63. Disposal of High-Level Radioactive Wastes in a Geologic Repository at Yucca Mountain, Nevada. Readily available. https://www.govinfo.gov/app/details/CFR-2005-title10-vol2/CFR-2005-title10-vol2-part63

17. 10 CFR Part 60. Disposal of High-Level Radioactive Wastes in Geologic Repositories. Readily available. https://www.govinfo.gov/app/details/CFR-2015-title10-vol2/CFR-2015-title10-vol2-part60

18. 40 CFR Part 191. Environmental Radiation Protection Standards for Management and Disposal of Spent Nuclear Fuel, High-Level and Transuranic Radioactive Wastes. Readily available. https://www.govinfo.gov/app/details/CFR-2011-title40-vol25/CFR-2011-title40-vol25-part191

19. 40 CFR Part 197. Public Health and Environmental Radiation Protection Standards for Yucca Mountain, Nevada. Readily available. https://www.govinfo.gov/app/details/CFR-2012-title40-vol26/CFR-2012-title40-vol26-part197

20. NAS (National Academy of Sciences) 1995. Technical Bases for Yucca Mountain Standards. National Research Council, Board on Radioactive Waste Management. National Academy Press. Washington, DC.

21. Sevougian, S. D., Mariner, P. E., Connolly, L. A., MacKinnon, R. J., Rogers, R. D., Dobson, D. C. and J. L. Prouty 2019a. *DOE SFWST Campaign R&D Roadmap Update*, *Rev. 1*, SAND2019-9033R, Sandia National Laboratories, Albuquerque, New Mexico, July 22, 2019.

22. Sevougian, S. D., G. E. Hammond, P. E. Mariner, R. J. MacKinnon, P. N. Swift, R. D. Rogers, D. C. Dobson, and M. C. Tynan 2019b. "Re-Evaluation of U.S. DOE R&D Efforts for Generic Deep Geologic Repositories – Roadmap Update," in *Proceedings of the International High-Level Radioactive Waste Management Conference (IHLRWM 2019)*, April 14 – 18, 2019, Knoxville, TN, USA, SAND2019-0340C.

23. NRC (US Nuclear Regulatory Commission) 2003. *Yucca Mountain Review Plan, Final Report*, NUREG-1804, Rev. 2, Office of Nuclear Material Safety and Safeguards, US Nuclear Regulatory Commission, Washington, DC.

# 3. CONCEPTUAL STRUCTURE OF PERFORMANCE ASSESSMENTS FOR THE GEOLOGIC DISPOSAL OF RADIOACTIVE WASTE

## 3.1 Introduction

A conceptual structure for performance assessments (PAs) for the geologic disposal of radioactive waste is described in this chapter. Illustrations of this structure are provided by past PAs for (i) a repository for transuranic radioactive waste near Carlsbad, New Mexico, known as the Waste Isolation Pilot Plant (WIPP) [1], and (ii) a proposed repository for spent nuclear fuel and high-level radioactive waste at Yucca Mountain (YM), Nevada [2]. Probabilistic risk assessments (PRAs) for nuclear power plants and other complex engineered facilities also have a similar conceptual structure [3-5].

The following topics are addressed: (i) Characterization of uncertainty (Sect. 3.2), (ii) Representation of aleatory uncertainty (Sect. 3.3), (iii) Estimation of consequences (Sect. 3.4), (iv) Representation of epistemic uncertainty (Sect. 3.5), (v) Propagation and display of uncertainty (Sect. 3.6), and (vi) Sensitivity analysis (Sect. 3.7). The chapter then ends with a brief summary (Sect. 3.8). This chapter is an edited and updated adaption of two prior conference presentations [6; 7] and several book chapters [8-10] and also draws on material in two special journal issues devoted to the indicated WIPP and YM PAs [1; 2].

It is recognized that not all analyses for the geologic disposal of radioactive waste will have the same structure as the two PAs [1; 2] used for illustration in this chapter (Chapter 3) describing the conceptual structure of PAs for the geologic disposal of radioactive waste and in the next chapter (Chapter 4) describing sampling-based procedures for uncertainty and sensitivity analysis. Individual PAs for proposed facilities for the geologic disposal of radioactive waste will most likely have custom designed structures that must incorporate multiple aspects of the proposed facility and its analysis, including (i) the design of the facility, (ii) the properties of the waste to be disposed of, (iii) the geologic properties of the site, (iv) the nature and availability of the models needed to perform an analysis of the facility and its geologic environment, (v) the need to appropriately treat the inevitable uncertainties that will be present in analysis results, (vi) the regulatory requirements that must be satisfied, and (vii) the needed, and hopefully available, resources required to perform a credible analysis. Although the designs for individual facilities for the geologic disposal radioactive are unlikely to be identical, they will probably have many aspects and components that are similar to the design described in Chapters 3 and 4.

## 3.2 Characterization of Uncertainty

A PA for a geologic repository for radioactive waste or any other complex facility is an analysis intended to answer three initial questions about the facility and one additional question about the analysis itself. The initial three questions are: Q1, "What could happen?"; Q2, "How likely is it to happen?"; and Q3, "What are the consequences if it does happen?". Formally, the answers to the preceding three questions can be represented by a set of ordered triples of the form

$$\left(\mathcal{S}_i, p\mathcal{S}_i, \mathbf{cS}_i\right), i = 1,2,\ldots,nS, \tag{3.2.1}$$

where (i) $\mathcal{S}_i$ is a set of similar occurrences, (ii) the sets $\mathcal{S}_i$ are disjoint (i.e., $\mathcal{S}_i \cap \mathcal{S}_j = \varnothing$ for $i \neq j$) and $\bigcup_i \mathcal{S}_i$ contains everything that could potentially occur at the particular facility under consideration, (iii) $p\mathcal{S}_i$ is the probability for $\mathcal{S}_i$, and (iv) $\mathbf{cS}_i$ is a vector of consequences associated with $\mathcal{S}_i$ [11]. The fourth question is: Q4, "What is the uncertainty in the answers to the first three questions?" or, equivalently, "What is the level of confidence in the answers to the first three questions?".

The set of ordered triples in Eq. (3.2.1) is often referred to as the Kaplan-Garrick ordered triple representation for risk [11] and, in essence, is an intuitive representation for a function $f$ (i.e., a random variable) defined in association with a probability space $(\mathcal{A}, \mathbb{A}, p_A)$, where (i) $\mathcal{A}$ is the set of everything that could occur in the particular universe under consideration, (ii) $\mathbb{A}$ is the collection of subsets of $\mathcal{A}$ for which probability is defined, and (iii) $p_A$ is the function that defines probability for the elements of $\mathbb{A}$ ([12], Sect. IV.3). Specifically, the sets $\mathcal{S}_i$ are elements of $\mathbb{A}$ with $\bigcup_i \mathcal{S}_i = \mathcal{A}$; $p_A(\mathcal{S}_i)$ is the probability $p\mathcal{S}_i$ of $\mathcal{S}_i$; and $f(\mathbf{a}_i)$ for a representative element $\mathbf{a}_i$ of $\mathcal{S}_i$ defines $\mathbf{cS}_i$. Another less commonly used possibility is that $\mathbf{cS}_i$ is the expected value of $f$ conditional on $\mathcal{S}_i$.

The uncertainty characterized by the sets $\mathcal{S}_i$ and associated probabilities $p\mathcal{S}_i$ in Eq. (3.2.1) introduced as answers to questions Q1 and Q2 is often referred to as aleatory uncertainty and results from a perceived randomness in future occurrences that could take place at the facility under consideration. The descriptor aleatory derives from the Latin āleae, which refers to games of dice, and is the reason for the notational use of "A" in the definition of the probability space $(\mathcal{A}, \mathbb{A}, p_A)$. Alternative designators for aleatory uncertainty include stochastic, type A, and irreducible. Similarly, the function (i.e., model) $f$ provides an answer to question Q3.

The fourth question (Q4) relates to uncertainty that results from a lack of knowledge with respect to the appropriateness of assumptions and/or parameter values used in an analysis. The basic idea is that an analysis has been developed to the point that it has a well-defined overall structure with established definitions for models and parameters but uncertainty remains with respect to appropriate parameter values and possibly submodel structure or choice for use in this overall structure. This form of uncertainty is usually indicated by the descriptor epistemic, which derives from the Greek epistēmē for knowledge. Alternative descriptors for epistemic uncertainty

include subjective, state-of-knowledge, type B, and reducible. Many analyses use probability to characterize epistemic uncertainty, which in turn means that in some way a probability space $(\mathcal{E}, \mathbb{E}, p_E)$ characterizing epistemic uncertainty must be defined and incorporated into the analysis. The representation of aleatory and epistemic uncertainty is an important, and sometimes contentious, part of the analysis of a complex system. As a result, an extensive literature exits on the incorporation of aleatory and epistemic uncertainty into analyses of complex systems (e.g., [13-26]).

The concepts of aleatory uncertainty and epistemic uncertainty are not new. The formal development of the concept of probability began around 1660 [29]. Even at its early nascent state, probability was used to characterize both (i) frequency concepts related to the outcomes of games of chance and the valuation of annuities (i.e., aleatory uncertainty) and (ii) degrees of belief with respect to the truth of propositions that were devoid of a frequency-based content (i.e., epistemic uncertainty).

Hacking [29] provides an excellent discussion of the early development of probability and the dual presence of aleatory uncertainty and epistemic uncertainty in this development. Additional discussions of interest appear in Refs. [30; 31]. An example of the entry of the concepts of aleatory uncertainty and epistemic uncertainty into economic analysis is provided by Knight [32], who uses the term "risk" in reference to aleatory uncertainty and the term "uncertainty" in reference to epistemic uncertainty.

Significant impetus to the separation and assessment of the effects and implications of aleatory uncertainty and epistemic uncertainty in analyses of complex systems was given by the Lewis Committee's 1978 review [33] of the Reactor Safety Study (commonly known as WASH-1400 after its report number [34]), which was completed by the U.S. Nuclear Regulatory Commission (NRC) in 1975. In its review, the Lewis committee praised the WASH-1400 analysis for its assessment of potential accidents at commercial nuclear power plants (essentially an analysis of the effects and implications of aleatory uncertainty) but noted that the analysis had not adequately characterized the (epistemic) uncertainty in its results. The Lewis Committee did not use the descriptor epistemic in their references to a need for additional uncertainty characterization in the WASH-1400 analysis but context clearly indicates that epistemic uncertainty was what was being referred to.

The Lewis Committee's review of the WASH-1400 analysis initiated an intense interest at the NRC with respect to the appropriate treatment of aleatory uncertainty and epistemic uncertainty in programs and analyses carried out in support of the NRC's regulatory responsibilities for assuring the safety of the nuclear industry in the United States, including (i) development of a methodology to assess the risk associated with the geologic disposal of radioactive waste [35-41], (ii) development of the MELCOR code system for the assessment of reactor accident consequences [42-44], (iii) an updated assessment of the risk from commercial nuclear power plants (commonly known as NUREG-1150 after its report number [45]) [3-5; 45-48], and (iv) various studies to assess the uncertainty in estimates of the off-site consequences associated with reactor accidents [49-61]. Of particular note is the extensive review and characterization of epistemic uncertainty performed in support of the NUREG-1150 analyses [62-68]. In addition, the separation of aleatory uncertainty and epistemic uncertainty played an major role in the design and implementation of the analysis that supported the U.S. Department of Energy's (DOE's) successful Compliance Certification

Application to the U.S Environmental Protection Agency (EPA) for the Waste Isolation Pilot Plant (WIPP) [27; 69-73]. Coinciding with the increasing recognition in the 1980s and 1990s of the importance of an adequate treatment of uncertainty in an analysis of a complex system, a number of discussions of this topic appeared [11; 13-15; 17; 19; 21; 22; 74-76].

A PA for a geologic repository for radioactive waste or any other complex facility is a highly complex analysis. The understanding of such an analysis is greatly facilitated if it can be seen "in the large" before having to work through all the details of the implementation of the analysis. Fortunately, such an analysis is typically underlain by three basic entities (EN) or components: EN1, a probability space $(\mathcal{A}, \mathbb{A}, p_A)$ that characterizes aleatory uncertainty; EN2, a function $f$ that estimates consequences for individual elements $\mathbf{a}$ of the sample space $\mathcal{A}$ for aleatory uncertainty; and EN3, a probability space $(\mathcal{E}, \mathbb{E}, p_E)$ that characterizes epistemic uncertainty [7; 8; 27; 28]. A recognition and understanding of these three basic entities make it possible to understand the conceptual and computational structure of a large PA without having basic concepts obscured by fine details of the analysis and leads to an analysis that results in informative uncertainty and sensitivity analyses. The nature and role of these three basic entities in PA and associated uncertainty and sensitivity analyses are elaborated on and illustrated in the following sections.

## 3.3 EN1, Representation of Aleatory Uncertainty

As indicated in Sect. 3.2, aleatory uncertainty is formally characterized by a probability space $(\mathcal{A}, \mathbb{A}, p_A)$. The elements $\mathbf{a}$ of $\mathcal{A}$ are vectors

$$\mathbf{a} = [a_1, a_2, \ldots, a_{nA}] \tag{3.3.1}$$

characterizing individual futures that potentially could occur at the facility under consideration. The set $\mathcal{A}$ and the individual futures $\mathbf{a}$ contained in $\mathcal{A}$ are typically defined for some specified time interval. For example, the time interval [0, $10^4$ yr] is specified in the regulations for WIPP [27; 72], and time intervals of [0, $10^4$ yr] and [0, $10^6$ yr] are specified in different parts of the regulations for the proposed YM repository ([28], Sect. 6). In contrast, futures are, in effect, defined for one year time intervals in PAs for nuclear power plants [3-5]. In practice, the probability space $(\mathcal{A}, \mathbb{A}, p_A)$ is usually defined by specifying distributions for the individual elements of $\mathbf{a}$. For notational purposes, it is often convenient to represent the distribution associated with the elements $\mathbf{a}$ of $\mathcal{A}$ with a density function $d_A(\mathbf{a})$.

For the 1996 PA supporting the Compliance Certification Application (CCA) for the WIPP, the only disruptions with sufficient potential for occurrence over the $10^4$ yr regulatory period to merit inclusion in the analysis were drilling for petroleum resources and mining of potash above the excavated waste disposal drifts [69; 72]. As a consequence, the individual futures underlying the 1996 WIPP PA have the form

$$\mathbf{a} = [\underbrace{t_1, l_1, e_1, b_1, p_1, \mathbf{w}_1}_{\text{1st intrusion}}, \quad \underbrace{t_2, l_2, e_2, b_2, p_2, \mathbf{w}_2}_{\text{2nd intrusion}}, \ldots, \quad \underbrace{t_n, l_n, e_n, b_n, p_n, \mathbf{w}_n}_{n\text{th intrusion}}, \quad t_{min}], \tag{3.3.2}$$

where (i) $n$ is the number of drilling intrusions over the $10^4$ yr regulatory period, (ii) $t_i$ is the time (yr) of the $i^{\text{th}}$ intrusion, (iii) $l_i$ designates the location of the $i^{\text{th}}$ intrusion, (iv) $e_i$ designates the penetration of an excavated or nonexcavated area by the $i^{\text{th}}$ intrusion, (v) $b_i$ designates whether or not the $i^{\text{th}}$ intrusion penetrates pressurized brine in the Castile Formation, (vi) $p_i$ designates the borehole plugging procedure used with the $i^{\text{th}}$ intrusion, (vii) $\mathbf{w}_i$ designates the type of waste penetrated by the $i^{\text{th}}$ intrusion, and (viii) $t_{min}$ is the time (yr) at which potash mining occurs within the land withdrawal boundary. The distributions for the number $n$ and times $t_i$ for drilling intrusions and the time $t_{min}$ for potash mining are defined by Poisson processes; the distributions for the remaining elements of $\mathbf{a}$ are defined on the basis of the properties of the waste disposal drifts associated with the WIPP and the geologic environment that surrounds these drifts. Additional information on the elements of $\mathbf{a}$ and their probabilistic characterization is available in Table III of Ref. [70] and in Ref. [72].

For the 2008 PA supporting the license application for the proposed YM repository,

$$\mathbf{a} = \{nEW, nED, nII, nIE, nSG, nSF, \mathbf{a}_{EW}, \mathbf{a}_{ED}, \mathbf{a}_{II}, \mathbf{a}_{IE}, \mathbf{a}_{SG}, \mathbf{a}_{SF}\}, \tag{3.3.3}$$

where (i) $nEW$ = number of early waste package (WP) failures, (ii) $nED$ = number of early drip shield (DS) failures, (iii) $nII$ = number of igneous intrusive events, (iv) $nIE$ = number of igneous eruptive events, (v) $nSG$ = number of seismic ground motion events, (vi) $nSF$ = number of seismic fault displacement events, (vii) $\mathbf{a}_{EW}$ = vector defining the $nEW$ early WP failures, (viii) $\mathbf{a}_{ED}$ = vector defining the $nED$ early DS failures, (ix) $\mathbf{a}_{II}$ = vector defining the $nII$ igneous intrusive events, (x) $\mathbf{a}_{IE}$ = vector defining the $nIE$ igneous eruptive events, (xi) $\mathbf{a}_{SG}$ = vector defining the $nSG$ seismic ground motion events, and (xii) $\mathbf{a}_{SF}$ = vector defining the $nSF$ fault displacement events. In turn, the vectors $\mathbf{a}_{EW}$, $\mathbf{a}_{ED}$, $\mathbf{a}_{II}$, $\mathbf{a}_{IE}$, $\mathbf{a}_{SG}$ and $\mathbf{a}_{SF}$ are of the form

$$\mathbf{a}_{EW} = \left[\mathbf{a}_{EW,1}, \mathbf{a}_{EW,2}, \ldots, \mathbf{a}_{EW,nEW}\right], \mathbf{a}_{ED} = \left[\mathbf{a}_{ED,1}, \mathbf{a}_{ED,2}, \ldots, \mathbf{a}_{ED,nED}\right], \tag{3.3.4}$$

$$\mathbf{a}_{II} = \left[\mathbf{a}_{II,1}, \mathbf{a}_{II,2}, \ldots, \mathbf{a}_{II,nII}\right], \mathbf{a}_{IE} = \left[\mathbf{a}_{IE,1}, \mathbf{a}_{IE,2}, \ldots, \mathbf{a}_{IE,nIE}\right], \tag{3.3.5}$$

and

$$\mathbf{a}_{SG} = \left[\mathbf{a}_{SG,1}, \mathbf{a}_{SG,2}, \ldots, \mathbf{a}_{SG,nSG}\right], \mathbf{a}_{SF} = \left[\mathbf{a}_{SF,1}, \mathbf{a}_{SF,2}, \ldots, \mathbf{a}_{SF,nSF}\right], \tag{3.3.6}$$

where (i) $\mathbf{a}_{EW,j}$ = vector defining early WP failure $j$ for $j$ = 1, 2, …, $nEW$, (ii) $\mathbf{a}_{ED,j}$ = vector defining early DS failure $j$ for $j$ = 1, 2, …, $nED$, (iii) $\mathbf{a}_{II,j}$ = vector defining igneous intrusive event $j$ for $j$ = 1, 2, …, $nII$, (iv) $\mathbf{a}_{IE,j}$ = vector defining igneous eruptive event $j$ for $j$ = 1, 2, …, $nIE$, (v) $\mathbf{a}_{SG,j}$ = vector defining seismic ground motion event $j$ for $j$ = 1, 2, …, $nSG$, and (vi) $\mathbf{a}_{SF,j}$ = vector defining seismic fault displacement event $j$ for $j$ = 1, 2, …, $nSF$. Definitions of the vectors $\mathbf{a}_{EW,j}$, $\mathbf{a}_{ED,j}$, $\mathbf{a}_{II,j}$, $\mathbf{a}_{IE,j}$, $\mathbf{a}_{SG,j}$ and $\mathbf{a}_{SF,j}$ and their associated probabilistic characterizations are given in Refs. [77-79].

As an example, the individual vectors $\mathbf{a}_{EW,j}$, $j$ = 1, 2, …, $nEW$, appearing in the definition of $\mathbf{a}_{EW}$ in Eq. (3.3.4) are defined by

$$\mathbf{a}_{EW,j} = \left[t_j, b_j, d_j\right] \tag{3.3.7}$$

and characterize the properties of failed WP $j$, where (i) $t_j$ designates WP type (i.e., $t_j$ = 1 ~ commercial spent nuclear fuel (CSNF) WP, $t_j$ = 2 ~ codisposed (CDSP) WP), (ii) $b_j$ designates percolation bin in which failed WP is located (i.e., $b_j$ = $k$ ~ location of failed WP in percolation bin $k$ for $k$ = 1, 2, 3, 4, 5; see Figure 2, Ref. [80]), and (iii) $d_j$ designates whether the failed WP experiences nondripping or dripping conditions (i.e., $d_j$ = 0 ~ nondripping conditions and $d_j$ = 1 ~ dripping conditions). In turn, the associated probilities are based on the assumption that the number early failed WPs follows a binomial distribution with the failed WPs distributed randomly over WP types, percolation bins, and nondrippin/dripping conditions.

Although potentially complex, representations of aleatory uncertainty of the form illustrated for the WIPP CCA PA and the YM license application PA permit a complete and unambiguous description of what constitutes aleatory uncertainty and the probabilistic characterization of this uncertainty.

## 3.4 EN2, Model that Estimates Consequences

As indicated in Sect. 3.2, the second entity that underlines a PA for a complex system is a function $f$ that estimates a vector $f(\mathbf{a})$ of consequences for individual elements $\mathbf{a}$ of the sample space $\mathcal{A}$ for aleatory uncertainty. In most analyses, the function $f$ corresponds to a sequence of models for multiple physical processes that must be implemented and numerically evaluated in one or more computer programs. However, for notational and conceptual purposes, it is useful to represent this sequence of models as a function of elements of the sample space $\mathcal{A}$ for aleatory uncertainty. In most analyses associated with radioactive waste disposal, the analysis results of interest are functions of time. In this situation, the model used to predict system behavior can be represented by $f(\tau|\mathbf{a})$, where $\tau$ corresponds to time and the indication of conditionality (i.e., "$|\mathbf{a}$") emphasizes that the results at time $\tau$ depend on the particular element of $\mathcal{A}$ under consideration. In most analyses, $f(\tau|\mathbf{a})$ corresponds to a large number of results. In general, $\tau$ could also correspond to designators other than time such as spatial coordinates; however, it is possible that spatial coordinates and associated location dependent results would simply be viewed as part of a very large number of results corresponding to $f(\tau|\mathbf{a})$.

As an example, the sequence of linked models that corresponds to $f$ in the 1996 WIPP PA is indicated in Figure 3.1. As summarized in Table 3.1, the models indicated in Figure 3.1 represent a variety of physical processes and also involve a variety of mathematical structures (see Refs. [1; 69] for additional information on the models used in the 1996 WIPP PA).

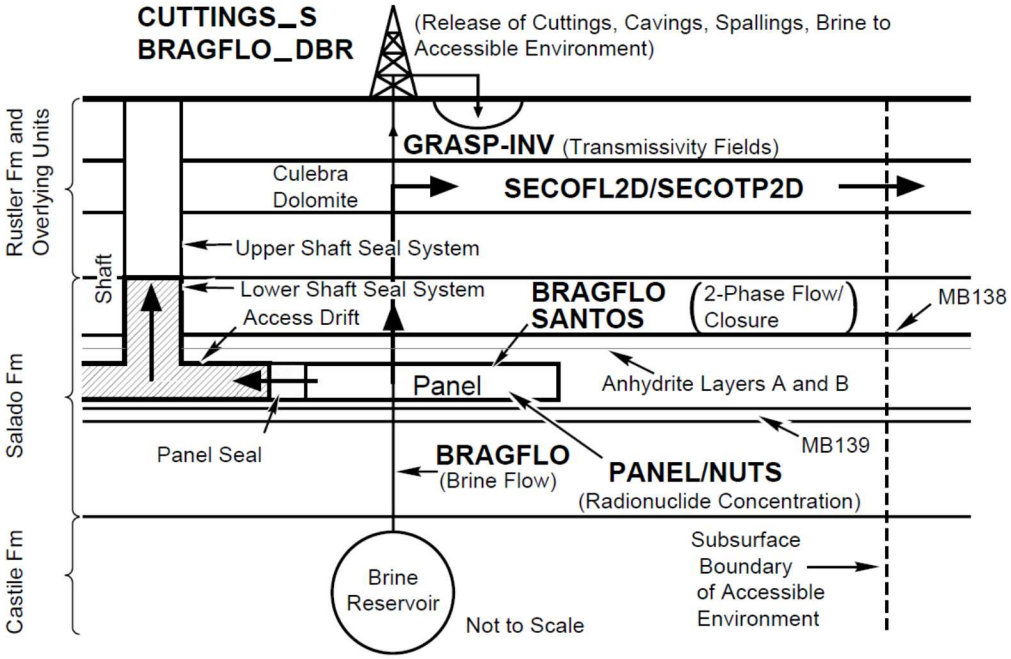



Figure 3.1 Models used in the 1996 WIPP PA ([81], Figure 4); see Table 3.1 for a description of the individual models.

Table 3.1 Summary of individual models used in the 1996 WIPP PA ([81], Table 1).

| |
|---|
| BRAGFLO: Calculates multiphase flow of gas and brine through a porous heterogeneous reservoir. Uses finite difference procedures to solve system of nonlinear partial differential equations (PDEs) that describes the conservation of gas and brine along with appropriate constraint equations, initial conditions, and boundary conditions. |
| BRAGFLO-DBR: Special configuration of BRAGFLO model used to calculate dissolved radionuclide releases to the surface at the time of a drilling intrusion. Uses initial value conditions obtained from calculations performed with BRAGFLO and CUTTINGS_S. |
| CUTTINGS_S: Calculates quantity of radioactive material brought to the surface in cuttings, cavings and spallings generated by a drilling intrusion. Spallings calculation uses initial value conditions obtained from calculations performed with BRAGFLO. |
| GRASP-INV: Generates transmissivity fields conditioned on measured transmissivity values and calibrated to steady-state and transient pressure data at well locations using an adjoint sensitivity and pilot-point technique |
| NUTS: Solves system of PDEs for dissolved radionuclide transport in the vicinity of the repository. Uses brine volumes and flows calculated by BRAGFLO as input. |
| PANEL: Calculates rate of discharge and cumulative discharge of radionuclides from a waste panel through an intruding borehole. Uses brine volumes and flows calculated by BRAGFLO as input. |
| SANTOS: Determines quasistatic, large deformation, inelastic response of two-dimensional solids with finite element techniques. Used to determine porosity of waste as a function of time and cumulative gas generation, which is an input to calculations performed with BRAGFLO. |
| SECOFL2D: Calculates single-phase Darcy flow for groundwater flow in two dimensions based on a PDE for hydraulic head. Uses transmissivity fields generated by GRASP-INV. |
| SECOTP2D: Simulates transport of radionuclides in a fractured porous medium. Solves two PDEs: one provides a two-dimensional representation for convective and diffusive radionuclide transport in fractures and the other provides a one-dimensional representation for diffusion of radionuclides into the rock matrix surrounding the fractures. Uses flow fields calculated by SECOFL2D. |

As another example, a subset of the models that correspond to $f$ in the 2008 PA for the proposed YM repository is indicated in Figure 3.2. The particular configuration of models shown in Figure 3.2 was used in the YM PA to calculate consequences associated with elements of the sample space $\mathcal{A}$ for aleatory uncertainty that involved seismic disruptions. Similar model configurations were used to calculate consequences for early WP failures, early DS failures, and igneous intrusive events; a very different suite of models was used to calculate consequences for igneous eruptive events. The number of individual models used in the 2008 YM PA is too great to permit their individual description here. However, descriptions of these models are available in Ref. [82; 83] and in the more detailed model-specific technical reports cited in Refs. [82; 83].

As is evident from Figure 3.1, Figure 3.2 and Table 3.2, the function $f$ that constitutes the second of the three entities that underlie a PA for radioactive waste repository can be quite complex.

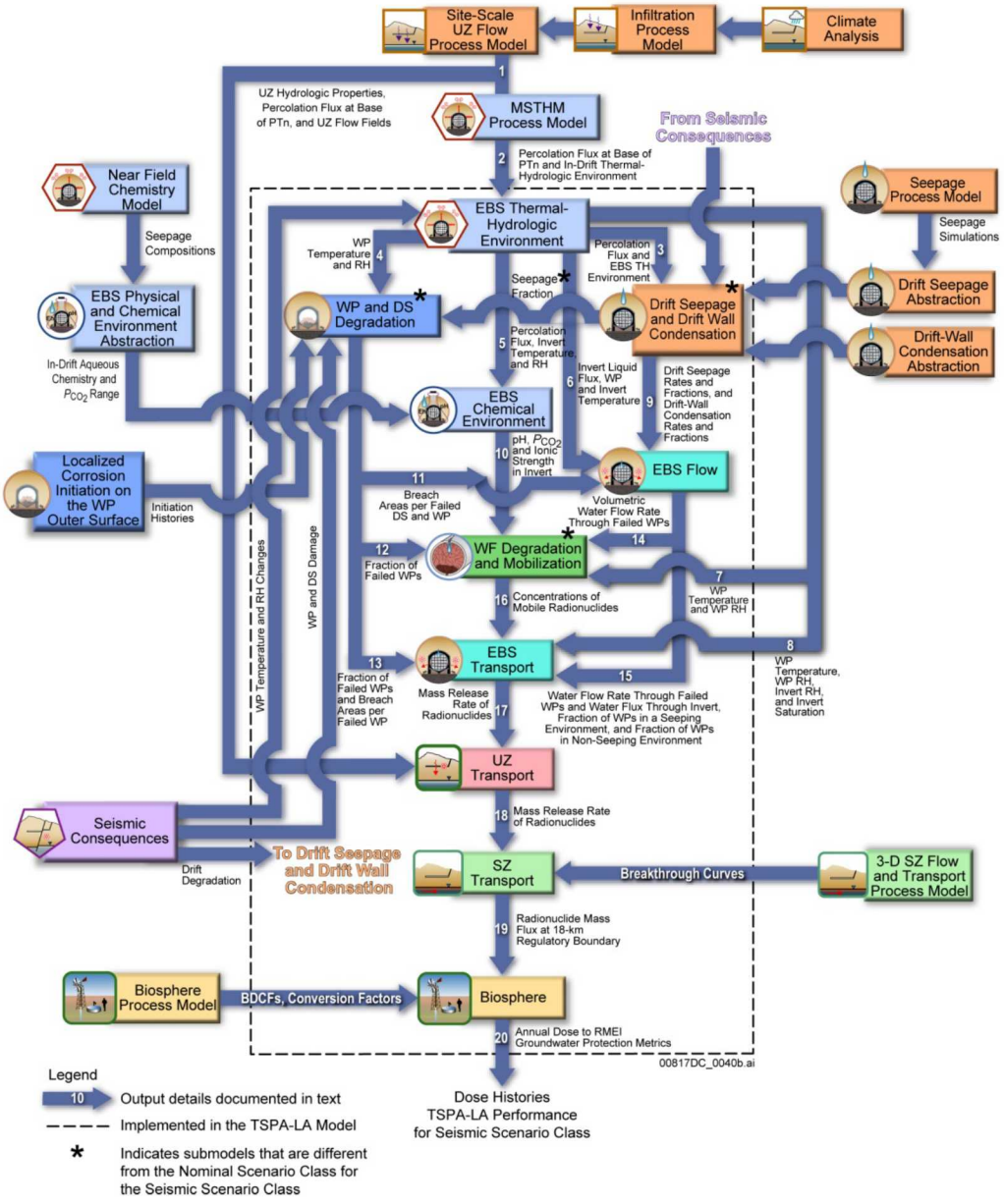



Figure 3.2 Models used in the 2008 YM PA for seismic disruptions ([82], Figure 6.1.4-6).

## 3.5 EN3, Representation of Epistemic Uncertainty

As indicated in Sect. 3.2, epistemic uncertainty is formally characterized by a probability space $(\mathcal{E}, \mathbb{E}, p_E)$. The elements **e** of $\mathcal{E}$ are vectors of the form

$$\mathbf{e} = [\mathbf{e}_A, \mathbf{e}_M] = [e_1, e_2, ..., e_{nE}], \tag{3.5.1}$$

where $\mathbf{e}_A$ is a vector of epistemically uncertain quantities involved in the definition of the probability space $(\mathcal{A}, \mathbb{A}, p_A)$ for aleatory uncertainty that constitutes the first of the three entities that underlies a PA and $\mathbf{e}_M$ is a vector of epistemically uncertain quantities involved in the definition of the function *f* that constitutes the second of the three entities that underlies a PA. The probability space $(\mathcal{E}, \mathbb{E}, p_E)$ for epistemic uncertainty and its associated density function $d_E(\mathbf{e})$ are usually developed through an expert review process that involves assigning a distribution to each element $e_i$ of **e** [62; 84-90]. With the introduction of the epistemically uncertain quantities that constitute the elements of **e** = $[\mathbf{e}_A, \mathbf{e}_M]$, the notation for the density function $d_A(\mathbf{a})$ associated with the probability space $(\mathcal{A}, \mathbb{A}, p_A)$ for aleatory uncertainty becomes $d_A(\mathbf{a} \mid \mathbf{e}_A)$ to indicate the dependence on $\mathbf{e}_A$; similarly, the notation for the time-dependent values for the function *f* becomes $f(\tau|\mathbf{a}, \mathbf{e}_M)$.

In effect, the probability space $(\mathcal{E}, \mathbb{E}, p_E)$ is the result of combining a probability space $(\mathcal{E}_A, \mathbb{E}_A, p_{EA})$ for epistemically uncertain quantities that affect the definition of the probability space $(\mathcal{A}, \mathbb{A}, p_A)$ for aleatory uncertainty and a probability space $(\mathcal{E}_M, \mathbb{E}_M, p_{EM})$ for epistemically uncertain quantities required in the modeling of physical processes. In turn, the sample space for epistemic uncertainty is defined by $\mathcal{E} = \mathcal{E}_A \times \mathcal{E}_M$, with $\mathcal{E}$ containing vectors of the form defined in Eq. (3.5.1).

As examples, the probability spaces for epistemic uncertainty in the 1996 WIPP PA and the 2008 YM PA involved *nE* = 57 and *nE* = 392 epistemically uncertain quantities, respectively. Example elements of the vector **e** for the WIPP and YM PAs are described in Table 3.2 and Table 3.3. The example variables in Table 3.2 and Table 3.3 were selected to include all variables that appear in the example sensitivity analyses discussed in Sect. 3.7.

Table 3.2 Examples of the *nE* = 57 elements of the vector **e** of epistemically uncertain variables in the 1996 WIPP PA (see (i) [73], Table 1, for a complete listing of the indicated 57 epistemically uncertain variables and sources of detailed information on the individual variables and (ii)[150;151] for a description of the procedures used to develop uncertainty distributions for these variables).

| |
|---|
| *ANHPRM* – Logarithm of anhydrite permeability ($m^2$). *Distribution:* Student's with 5 degrees of freedom. *Range:* –21.0 to –17.1. *Correlation:* –0.99 rank correlation with *ANHCOMP* (anhydrite compressibility). |
| *BHPRM* – Logarithm of borehole permeability ($m^2$). *Distribution:* Uniform. *Range:* –14.0 to –11.0. |
| *HALPOR* – Halite porosity (dimensionless). *Distribution:* Piecewise uniform. *Range:* 0.001 to 0.03. |
| *SHBCEXP* – Brooks-Corey pore distribution parameter for shaft (dimensionless). *Distribution:* Piecewise uniform. *Range:* 0.11 to 8.10. |
| *WRGSSAT* – Residual gas saturation in waste (dimensionless). *Distribution:* Uniform. *Range:* 0 to 0.15. |

Table 3.3 Examples of the nE = 392 elements of the vector e of epistemically uncertain variables in the 2008 YM PA (see [28], App. B, for a complete listing of the indicated 392 epistemically uncertain variables and sources of detailed information on the individual variables).

| |
|---|
| *CORRATSS*. Stainless steel corrosion rate (μm/yr). *Distribution:* Truncated log normal. *Range:* 0.01 to 0.51. *Mean/Median/Mode:* 0.267. *Standard Deviation:* 0.209. |
| *CSNFMASS*. Scale factor used to characterize uncertainty in radionuclide content of commercial spent nuclear fuel (CSNF) (dimensionless). Distribution: Uniform. Range: 0.85 to 1.4. |
| *DSNFMASS*. Scale factor used to characterize uncertainty in radionuclide content of defense spent nuclear fuel (DSNF) (dimensionless). *Distribution:* Triangular. *Range:* 0.45 to 2.9. *Mode:* 0.62. |
| *DTDRHUNC*. Selector variable used to determine the collapsed drift rubble thermal conductivity (dimensionless). Distribution: Discrete. Range: 1 to 2. |
| *HLWDRACD*. Effective rate coefficient (affinity term) for the dissolution of high level waste (HLW) glass in codisposed spent nuclear fuel(CDSP) waste packages (WPs) under low pH conditions ($g/(m^2 d)$). Distribution: Triangular. Range: 8.41E+03 to 1.15E+07. Mode: 8.41E+03. |
| *IGRATE*. Frequency of intersection of the repository footprint by a volcanic event ($yr^{-1}$). *Distribution:* Piecewise uniform. *Range:* 0 to 7.76E-07. |
| *INFIL*. Pointer variable for determining infiltration conditions: $10^{th}$, $30^{th}$, $50^{th}$ or $90^{th}$ percentile infiltration scenario (dimensionless). *Distribution:* Discrete. *Range:* 1 to 4. |
| *INRFRCTC*. The initial release fraction of $^{99}Tc$ in a CSNF waste package (dimensionless). Distribution: Triangular. Range: 0.0001 to 0.0026. Mode: 0.001. |
| *MICC14*. Groundwater Biosphere Dose Conversion Factor (BDCF) for $^{14}C$ in modern interglacial climate ((Sv/year)/($Bq/m^3$)). *Distribution:* Discrete. *Range:* 7.18E–10 to 2.56E–08. *Mean:* 1.93E–09. *Standard Deviation:* 1.85E–09. |
| *MICTC99*. Groundwater BDCF for $^{99}Tc$ in modern interglacial climate ((Sv/year)/($Bq/m^3$)). *Distribution:* Discrete. *Range:* 5.28E–10 to 2.85E–08. *Mean:* 1.12E–09. *Standard Deviation:* 1.26E–09. |
| S*CCTHR*. Residual stress threshold for stress corrosion cracking (MPa). *Distribution:* Uniform. *Range:* 315.9 to 368.55. |
| *SCCTHRP*. Residual stress threshold for stress corrosion crack (SCC) nucleation of Alloy 22 (as a percentage of yield strength in MPa) (dimensionless). *Distribution:* Uniform. *Range:* 90 to 105. |
| *SZFIPOVO*. Logarithm of flowing interval porosity in volcanic units (dimensionless). *Distribution:* Piecewise uniform. *Range:* –5 to –1. *Mean/Median/Mode:* –3. |
| *Table 3.3 continued on page 3-13* |

| |
|---|
| *SZGWSPDM*. Logarithm of the scale factor used to characterize uncertainty in groundwater specific discharge (dimensionless). *Distribution:* Piecewise uniform. *Range:* –0.951 to 0.951. |
| *THERMCON*. Selector variable for one of three host-rock thermal conductivity scenarios (low, mean, and high) (dimensionless). *Distribution:* Discrete. *Range:* 1 to 3. |
| *WDGCA22*. Temperature dependent slope term of Alloy 22 general corrosion rate (K). *Distribution:* Truncated normal. *Range:* 666 to 7731. *Mean:* 4905. *Standard Deviation:* 1413. |
| *WDGCUA22*. Variable for selecting distribution for general corrosion rate (low, medium, or high) (dimensionless). *Distribution:* Discrete. *Range:* 1 to 3. |
| *WDNSCC*. Stress corrosion cracking growth rate exponent (repassivation slope) (dimensionless). *Distribution:* Truncated normal. *Range:* 0.935 to 1.395. *Mean:* 1.165. *Standard Deviation:* 0.115. |
| *WDZOLID*. Deviation from median yield strength range for outer lid (dimensionless). *Distribution:* Truncated normal. *Range:* –3 to 3. *Mean:* 0. *Standard Deviation:* 1. |

As done in the WIPP and YM PAs, probability is the most widely used mathematical structure for the representation of epistemic uncertainty. However, a number of additional structures for the representation of epistemic uncertainty in the presence of limited information have also been developed, including interval analysis, evidence theory and possibility theory (e.g., [91-100]). As yet, these uncertainty structures have not been used in large-scale PAs.

## 3.6 Propagation and Display of Uncertainty

Two possibilities exist when the propagation and display of uncertainty are considered in a PA that involves a separation of aleatory uncertainty and epistemic uncertainty: (i) Presentation of the effects of epistemic uncertainty conditional on a specific realization of aleatory uncertainty, and (ii) Presentation of the effects of aleatory uncertainty conditional on a specific realization of epistemic uncertainty. The presentation of the effects of epistemic uncertainty conditional on a specific realization of aleatory uncertainty is considered first.

### 3.6.1 Effects of epistemic uncertainty conditional on a specific realization of aleatory uncertainty

Epistemic uncertainty in an outcome of a PA conditional on a specific realization **a** of aleatory uncertainty can be formally summarized with a cumulative distribution function (CDF) or a complementary cumulative distribution function (CCDF). For a real-valued analysis outcome $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$, the CDF and CCDF for $y$ resulting from epistemic uncertainty in $\mathbf{e}_M$ are defined by

$$p_E\left(\tilde{y} \le y \mid \mathbf{a}\right) = \int_{\mathcal{E}} \underline{\delta}_y\left[f\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right)\right] d_E\left(\mathbf{e}_M\right) \mathrm{d}\mathcal{E} \tag{3.6.1}$$

and

$$p_E\left(y < \tilde{y} \mid \mathbf{a}\right) = 1 - p_E\left(\tilde{y} \le y \mid \mathbf{e}\right) = \int_{\mathcal{E}} \overline{\delta}_y\left[f\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right)\right] d_E\left(\mathbf{e}_M\right) \mathrm{d}\mathcal{E}, \tag{3.6.2}$$

respectively, where (i) $\mathbf{e}_M$ is a vector of epistemically uncertain quantities affecting the function $f$ as indicated in conjunction with Eq. (3.5.1), (ii) $d_E(\mathbf{e}_M)$ is the density function associated with the probability space that characterizes the epistemic uncertainty associated with $\mathbf{e}_M$, (iii) $\mathcal{E}$ and $d_E(\mathbf{e}_M)$ are restricted to possible values for $\mathbf{e}_M$, and (iv)

$$\underline{\delta}_y\left(\tilde{y}\right) = \begin{cases} 1 \text{ for } \tilde{y} \le y \\ 0 \text{ otherwise} \end{cases} \text{ and } \overline{\delta}_y\left(\tilde{y}\right) = 1 - \underline{\delta}_y\left(\tilde{y}\right) = \begin{cases} 1 \text{ for } y \le \tilde{y} \\ 0 \text{ otherwise.} \end{cases} \tag{3.6.3}$$

Specifically, plots of the points $[y, p_E(\tilde{y} \le y \mid \mathbf{a})]$ and $[y, p_E(y \le \tilde{y} \mid \mathbf{a})]$ define the CDF and CCDF for $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$, with these plots being conditional on the vector **a** from the sample space for aleatory uncertainty. In addition,

$$E_E\left(y \mid \mathbf{a}\right) = \int_{\mathcal{E}} f\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right) d_E\left(\mathbf{e}_M\right) \mathrm{d}\mathcal{E} \tag{3.6.4}$$

defines the expected value of $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$ over epistemic uncertainty.

In most PAs, the integrals in Eqs. (3.6.1), (3.6.2) and (3.6.4) are too complex to estimate with formal quadrature procedures, with the result that these integrals are typically estimated with

procedures based on simple random sampling or Latin hypercube sampling. In turn, the indicated sampling procedures result in estimates of the form

$$p_E\left(\tilde{y} \le y \mid \mathbf{a}\right) \cong \sum_{k=1}^{n} \underline{\delta}_y \left[ f\left(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}\right)\right] / n, \tag{3.6.5}$$

$$p_E\left(y < \tilde{y} \mid \mathbf{a}\right) \cong \sum_{k=1}^{n} \overline{\delta}_y \left[ f\left(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}\right)\right] / n, \tag{3.6.6}$$

and

$$E_E\left(y \mid \mathbf{a}\right) \cong \sum_{k=1}^{n} f\left(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}\right) / n, \tag{3.6.7}$$

where $\mathbf{e}_{Mk}$, $k = 1, 2, \ldots, n$, is a sample from $\mathcal{E}$ obtained in consistency with the density function $d_E(\mathbf{e}_M)$ for epistemic uncertainty.

Dose (mrem/yr) to the reasonably maximally exposed individual (RMEI) determined in the 2008 YM PA under the assumption of nominal (i.e., undisturbed) conditions provides an illustration of the results formally defined in Eqs. (3.6.1)-(3.6.7). In the 2008 YM PA, nominal conditions correspond to the occurrence of the aleatory future $\mathbf{a}_N$ defined in Eq. (3.3.3) for which no disruptions of any type occur (i.e., the future in which $nEW = nED = nII = nIE = nSG = nSF = 0$). The 2008 YM PA used a Latin hypercube sample (LHS)

$$\mathbf{e}_k = \left[\mathbf{e}_{Ak}, \mathbf{e}_{Mk}\right], k = 1, 2, \ldots, n = 300, \tag{3.6.8}$$

of size 300 in the propagation of epistemic uncertainty [28; 101; 102]. Latin hypercube sampling operates in the following manner to generate a sample of size $n$ from the distributions $D_1$, $D_2$, …, $D_{nE}$ associated with the elements of $\mathbf{e} = [e_1, e_2, \ldots, e_{nE}]$. The range of each variable $e_l$ is divided into $n$ disjoint intervals of equal probability and one value $e_{kl}$ is randomly selected from each interval. The $n$ values for $e_1$ are randomly paired without replacement with the $n$ values for $e_2$ to produce $n$ pairs. These pairs are then randomly combined without replacement with the $n$ values for $e_3$ to produce $n$ triples. This process is continued until a set of $n$ $nE$-tuples $\mathbf{e}_k = [e_{k1}, e_{k2}, \ldots, e_{k,nE}]$, $k = 1, 2, \ldots, n$, is obtained, with this set constituting the LHS. Owing to its efficient stratification properties, Latin hypercube sampling has also been used for the propagation of epistemic uncertainty in the 1996 WIPP PA [73], the NUREG-1150 probabilistic risk assessments for five nuclear power stations [3; 4], and many other analyses.

As indicated in Sect. 3.5, only elements of the vector $\mathbf{e}_M$ in Eq. (3.6.8) are involved in this example for dose to the RMEI as the elements of the vector $\mathbf{e}_A$ are involved in the definition of probability distributions related to the characterization of aleatory uncertainty. In consistency with

the notation used in the 2008 YM PA, $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ is used to represent dose to the RMEI at time $\tau$ conditional on (i) the future corresponding to $\mathbf{a}_N$ and (ii) the values for epistemically uncertain quantities contained in the vector $\mathbf{e}_M$. A complex sequence of calculations similar to those indicated in Figure 3.2 is used to determine

$$D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right) \text{ for } 0 \le \tau \le 10^6 \text{ yr and } k = 1,2,\ldots,n = 300. \tag{3.6.9}$$

In turn, approximations to the CDF, CCDF and expected value for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ as defined in Eqs. (3.6.5)-(3.6.7) are given by

$$p_E\left[D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_M\right) \le D\right] \cong \sum_{k=1}^{n=300} \underline{\delta}_D\left[D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right)\right] / n, \tag{3.6.10}$$

$$p_E\left[D < D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_M\right)\right] \cong \sum_{k=1}^{n=300} \overline{\delta}_D\left[D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right)\right] / n, \tag{3.6.11}$$

$$E_E\left[D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_M\right)\right] \cong \sum_{k=1}^{n=300} D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right) / n, \tag{3.6.12}$$

and illustrated in Figure 3.3a for $\tau = 600{,}000$ yr.

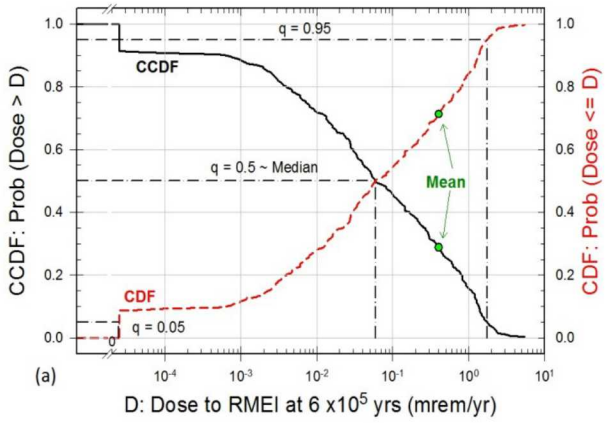
Also identified in Figure 3.3a are selected quantile values for $q$ = 0.05, 0.5 and 0.95. Once a CDF is available, quantile values are easily determined as indicated in Figure 3.3a by (i) starting at a desired quantile value $q$ on the ordinate of the CDF plot, (ii) drawing a horizontal line to the CDF, and (iii) then dropping a vertical line to the abscissa to obtain the $q$-quantile of the variable under consideration. More formally and with the same notation as used in Eqs. (3.6.1)-(3.6.7), the $q$-quantile value $Q_{Eq}[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ for $y = f(\tau \mid \mathbf{a}, \mathbf{e}_M)$ conditional on $\mathbf{a}$ and arising from the epistemic uncertainty associated with $\mathbf{e}_M$ is the value of $y$ such that

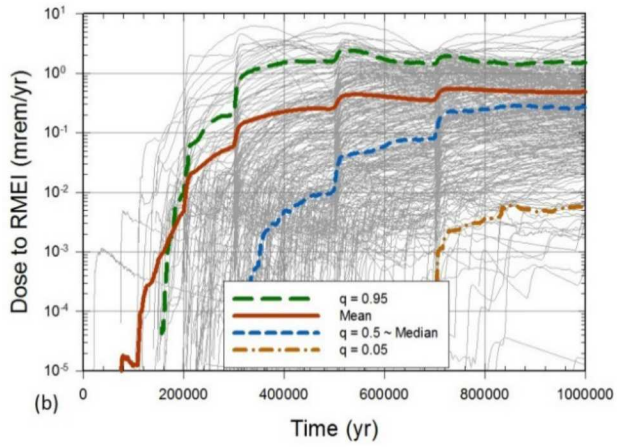
$$\begin{aligned} q &= \int_{\mathcal{E}} \underline{\delta}_y\left[f\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right)\right] d_E\left(\mathbf{e}_M\right) \mathrm{d}\mathcal{E} \\ &\cong \sum_{k=1}^{n} \underline{\delta}_y\left[f\left(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}\right)\right] / n. \end{aligned} \tag{3.6.13}$$

Specifically, the quantiles $Q_{Eq}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ at $\tau = 600{,}000$ yr indicated in Figure 3.3a are defined by the values of $D$ such that

$$q \cong \sum_{k=1}^{n=300} \underline{\delta}_D\left[D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right)\right] / n \tag{3.6.14}$$

for $q$ = 0.05, 0.5 and 0.95.




| (c) *DOSTOT*: 600,000 year | | | |
|---|---|---|---|
| **Step**[a] | **Variable**[b] | **$R^2$**[c] | **SRRC**[d] |
| 1 | *WDGCA22* | 0.85 | -0.94 |
| 2 | *WDZOLID* | 0.87 | 0.14 |
| 3 | *THERMCON* | 0.88 | -0.10 |
| 4 | *INFIL* | 0.89 | -0.10 |
| 5 | *SCCTHR* | 0.90 | -0.09 |
| 6 | *WDNSCC* | 0.91 | -0.08 |
| 7 | *CORRATSS* | 0.91 | -0.07 |
| 8 | *WDGCUA22* | 0.92 | 0.08 |
| 9 | *MICTC99* | 0.92 | 0.07 |
| 10 | *DTDRHUNC* | 0.92 | 0.04 |
| 11 | *CSNFMASS* | 0.92 | 0.04 |

a: Steps in stepwise rank regression analysis
b: Variables listed in order of selection in stepwise regression
c: Cumulative $R^2$ value with entry of each variable into regression model
d: Standardized rank regression coefficients (SRRCs) in final regression model

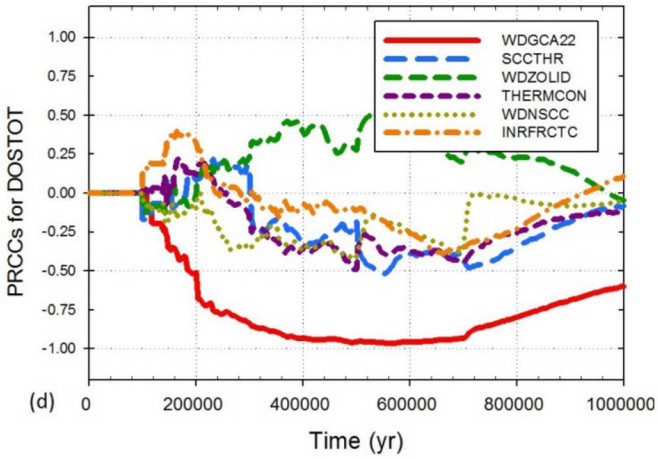



Figure 3.3 Estimates obtained with an LHS of size $n$ = 300 of the epistemic uncertainty in dose $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ to RMEI conditional on the realization $\mathbf{a}_N$ of aleatory uncertainty corresponding to nominal conditions: (a) CDF and CCDF for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ with $\tau$ = 600,000 yr ([103], Figure 1b), (b) Individual results $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk})$, $k$ = 1, 2,…, $n$, and associated expected (mean) values $E_E[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ and quantile values $Q_{Eq}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ $q$ = 0.05, 0.5, 0.95, for $0 \le \tau \le 10^6$ yr ([103], Figure 1a), (c) Stepwise rank regressions for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ at $\tau$ = 600,000 yr ([80], Table 10), and (d) Partial rank correlation coefficients (PRCCs) for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ for $0 \le \tau \le 10^6$ yr ([80], Figure 23b).

Also identified in Figure 3.3a are selected quantile values for $q$ = 0.05, 0.5 and 0.95. Once a CDF is available, quantile values are easily determined as indicated in Figure 3.3a by (i) starting at a desired quantile value $q$ on the ordinate of the CDF plot, (ii) drawing a horizontal line to the CDF, and (iii) then dropping a vertical line to the abscissa to obtain the $q$-quantile of the variable under consideration. More formally and with the same notation as used in Eqs. (3.6.1)-(3.6.7), the $q$-quantile value $Q_{Eq}[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ for $y = f(\tau \mid \mathbf{a}, \mathbf{e}_M)$ conditional on $\mathbf{a}$ and arising from the epistemic uncertainty associated with $\mathbf{e}_M$ is the value of $y$ such that

$$\begin{aligned} q &= \int_{\mathcal{E}} \underline{\delta}_y \left[ f\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right) \right] d_E\left(\mathbf{e}_M\right) \mathrm{d}\mathcal{E} \\ &\cong \sum_{k=1}^{n} \underline{\delta}_y \left[ f\left(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}\right) \right] \Big/ n. \end{aligned} \tag{3.6.15}$$

Specifically, the quantiles $Q_{Eq}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ at $\tau$ = 600,000 yr indicated in Figure 3.3a are defined by the values of $D$ such that

$$q \cong \sum_{k=1}^{n=300} \underline{\delta}_D \left[ D_N\left(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}\right) \right] \Big/ n \tag{3.6.16}$$

for $q$ = 0.05, 0.5 and 0.95.

As is the case for many PA results, $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ is a function of time. Specifically, $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ is determined for $0 \le \tau \le 10^6$ yr in the 2008 YM PA. Although plots of the form shown in Figure 3.3a provide a complete display of the effects of epistemic uncertainty at the time under consideration (e.g., $\tau$ = 600,000 yr in Figure 3.3a), they are not practical for displaying the effects of epistemic uncertainty at a large number of times. In such situations, an effective presentation format is to use a single plot to present both (i) the individual results $f(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$, $k$ = 1, 2, …, $n$, calculated for the time interval $t_{mn} \le \tau \le t_{mx}$ under consideration and (ii) the expected value $E_E[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ and selected quantile values $Q_{Eq}[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ for $f(\tau \mid \mathbf{a}, \mathbf{e}_M)$ also calculated for the time interval $t_{mn} \le \tau \le t_{mx}$. As an example, this presentation format is illustrated in Figure 3.3b for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ and $0 \le \tau \le 10^6$ yr, where (i) the lighter lines correspond to the results $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk})$, $k$ =1, 2, …, $n$ = 300, calculated as indicated in conjunction with Eq. (3.6.9) and (ii) the heavier lines correspond to the expected value $E_E[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ calculated as indicated in Eq. (3.6.12) and the quantiles $Q_{Eq}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$, $q$ = 0.05, 0.5, 0.95, calculated as indicated in Eq. (3.6.16). The presentation format illustrated in Figure 3.3b provides an effective way to present a large amount of information in a small amount of space. The sensitivity analysis results in Figure 3.3c and d are discussed in Sect. 3.7.

### 3.6.2 Latin hypercube sampling and the propagation of epistemic uncertainty

Latin hypercube sampling was developed for use in the analysis of computationally demanding analysis structures for which computational demands severely limit the number of system simulations that can be performed ([101]; [104], App. A). As a result, the LHS size in use in such analyses is often perceived as being small relative to the number of epistemically uncertain variables being sampled (e.g., an LHS of size 100 from 57 variables in the 1996 WIPP PA and an LHS of size 300 from 392 variables in the 2008 YM PA). In response to this concern, a replicated sampling procedure has been developed to provide a way to assess the adequacy of the LHS size used in the analysis of a complex and computationally demanding analysis structure ([105]; [73], Sect. 7). Specifically, the goal of the procedure is to determine how much variability exists in estimated analysis outcomes when different LHSs of size $n$ are used.

The procedure is based on independently generating LHSs of size $n$ with different random seeds $nR$ times. At an intuitive level, the adequacy of the sample size can be assessed by simply comparing the results obtained with the individual samples. At a more formal level, the $t$-distribution can be used to place confidence intervals around summary results (e.g., expected values over epistemic uncertainty) obtained from individual samples. Specifically, assume that the LHS has been replicated $nR$ times to produce $r = 1, 2, \ldots, nR$ independently generated LHSs

$$\mathbf{e}_{rk} = \left[\mathbf{e}_{Ark}, \mathbf{e}_{Mrk}\right], k = 1, 2, \ldots, n, \tag{3.6.17}$$

of size $n$ and that $C_r$ is an analysis result of interest obtained for replicated sample $r$. For example, $C$ might be the expected value $E_E[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ at time $\tau$ that is estimated with use of Latin hypercube sampling and $C_r = E_{Er}[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ would be the estimate for $C = E_E[f(\tau \mid \mathbf{a}, \mathbf{e}_M)]$ obtained with replicated sample $r$. Then,

$$\overline{C} = \sum_{r=1}^{nR} C_r / nR \quad \text{and} \quad SE\left(\overline{C}\right) = \left\{\sum_{r=1}^{nR}\left[C_r - \overline{C}\right]^2 \Big/ nR\left(nR-1\right)\right\} \tag{3.6.18}$$

provide an additional estimate of $C$ and an estimate of the standard error associated with this additional estimate. The $t$-distribution with $nR - 1$ degrees of freedom can now be used to place confidence intervals around the estimate $\overline{C}$ for $C$ in Eq. (3.6.18). Specifically, the $1 - \alpha$ confidence interval for $C$ is given by $\overline{C} \pm t_{1-\alpha/2} SE(\overline{C})$, where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t$-distribution with $nR - 1$ degrees of freedom. For example, $t_{1-\alpha/2} = 4.303$ for $\alpha = 0.05$ and $nR = 3$.

The 2008 YM PA used $nR = 3$ replicated LHSs of size 300 to assess the adequacy of an LHS of size 300 from the 392 epistemically uncertain variables under consideration. As an example, the results obtained for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ are shown in Figure 3.4. Specifically, the results for the individual replicates (i.e., $E_{Er}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ and $Q_{Eqr}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ for $q = 0.05, 0.5, 0.95$) are shown in Figure 3.4a, and the time-dependent 95% confidence intervals for $E_E[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ are shown in Figure 3.4b. As indicated by the results in Figure 3.4, an LHS

of size 300 is adequate for assessing the epistemic uncertainty present in estimates for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$. There is a substantial amount of variability in the LHS results prior to about 300,000 yr; however, this is due to the presence of results that are either 0 or so small as to be below the resolution of the calculation (see Figure 3.3b).

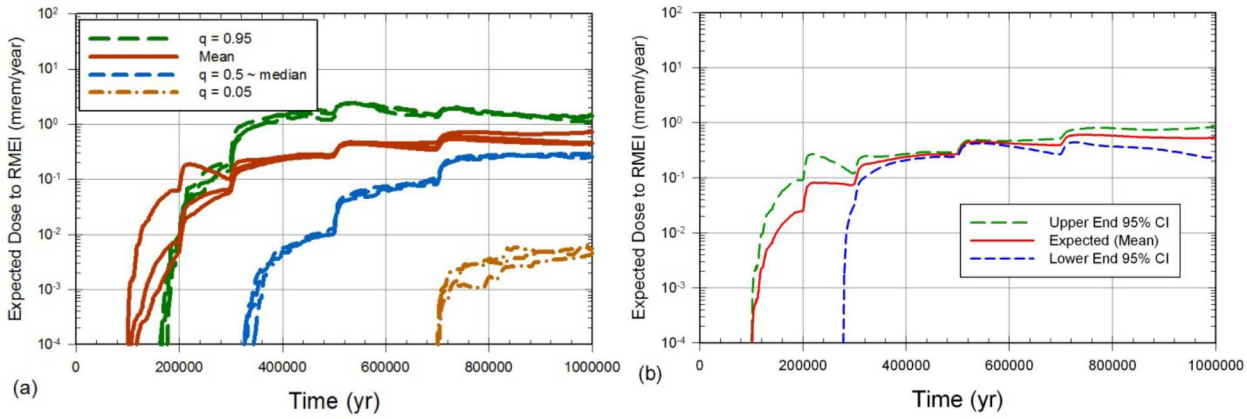



Figure 3.4 Illustration of replicated sampling to determine the adequacy of an LHS of size 300 for the assessment of the epistemic uncertainty present in estimates for the dose $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ to the RMEI conditional on undisturbed (i.e., nominal) repository conditions as indicated by the vector $\mathbf{a}_N$ ([103], Figure 4): (a) Expected values $E_{Er}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ and quantiles $Q_{Eqr}[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$, $q$ = 0.05, 0.5, 0.95, for $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)$ obtained for each of the 3 replicated samples indicated in Eq. (3.6.17), and (b) Time-dependent 95% confidence intervals (i.e., for $\alpha = 0.05$) for $E_E[D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_M)]$ obtained as indicated in conjunction with Eq. (3.6.18)

As part of the 2008 YM PA, the effects of sample size for a large number of different analysis outcomes were assessed with the three indicated replicated LHSs (e.g., see: [77], Figure 10; [78], Figs. 10 and 19; [79], Figs. 11 and 23; [106], Figs. 5 and 10; [107]). In each case, an LHS of size 300 is adequate for assessing the epistemic uncertainty present in the analysis outcome under consideration.

Another approach to assess sample size adequacy is to observe the results obtained with a sequence of LHSs of increasing size. An adequate sample size then corresponds to the sample size at which estimates of the analysis results of primary interest stabilize with respect to increasing sample size (e.g., [108], Figs. 13 and 14). Unlike simple random sampling, the size of an LHS cannot be extended by sampling additional sample elements without consideration of the sample elements already present in the sample. However, a technique does exist to extend an LHS of a given size to an LHS of a larger size [109].

Another concern that arises from the small sample sizes used with Latin hypercube sampling is that these small sample sizes could result in spurious correlations between variables within a sample. However, this problem can be controlled with a restricted pairing technique developed by Iman and Conover to control rank correlations between variables within random samples and LHSs [110; 111]. Specifically, this technique can be used to ensure that a sample has (i) rank correlations

close to zero for uncorrelated variables and (ii) desired rank correlations between correlated variables.

### 3.6.3 Effects of aleatory uncertainty conditional on a specific realization of epistemic uncertainty

The formal representation of the effects of aleatory uncertainty conditional on a specific realization of epistemic uncertainty is similar to the formal representation of the effects of epistemic uncertainty conditional on a specific realization of aleatory uncertainty but with the roles of the probability spaces for epistemic uncertainty and aleatory uncertainty reversed. Specifically, aleatory uncertainty in an outcome of a PA conditional on a specific realization $\mathbf{e} = [\mathbf{e}_A, \mathbf{e}_M]$ of epistemic uncertainty can be formally summarized with a CDF or a CCDF. For a real-valued analysis outcome $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$, the CDF and CCDF for $y$ resulting from aleatory uncertainty are defined by

$$p_A(\tilde{y} \le y \mid \mathbf{e}) = \int_{\mathcal{A}} \underline{\delta}_y \left[ f(\tau \mid \mathbf{a}, \mathbf{e}_M) \right] d_A(\mathbf{a} \mid \mathbf{e}_A) \, \mathrm{d}\mathcal{A} \tag{3.6.19}$$

and

$$p_A(y < \tilde{y} \mid \mathbf{e}) = 1 - p_A(\tilde{y} \le y \mid \mathbf{e}) = \int_{\mathcal{A}} \overline{\delta}_y \left[ f(\tau \mid \mathbf{a}, \mathbf{e}_M) \right] d_A(\mathbf{a} \mid \mathbf{e}_A) \, \mathrm{d}\mathcal{A}, \tag{3.6.20}$$

respectively, where (i) $\mathbf{a}$ is a vector defining one realization of aleatory uncertainty, (ii) $d_A(\mathbf{a} \mid \mathbf{e}_A)$ is the density function conditional on $\mathbf{e}_A$ associated with the probability space that characterizes the aleatory uncertainty associated with $\mathbf{a}$, and (iii) $\underline{\delta}_y(\tilde{y})$ and $\overline{\delta}_y(\tilde{y})$ are defined the same as in Eq. (3.6.3). In turn, plots of the points $[y, p_A(\tilde{y} \le y \mid \mathbf{e})]$ and $[y, p_A(y \le \tilde{y} \mid \mathbf{e})]$ define the CDF and CCDF for $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$, with these plots being conditional on the vector $\mathbf{e} = [\mathbf{e}_A, \mathbf{e}_M]$ from the sample space for epistemic uncertainty. Similarly,

$$E_A(y \mid \mathbf{e}) = \int_{\mathcal{A}} f(\tau \mid \mathbf{a}, \mathbf{e}_M) d_A(\mathbf{a} \mid \mathbf{e}_A) \, \mathrm{d}\mathcal{A} \tag{3.6.21}$$

defines the expected value of $y = f(\tau|\mathbf{a}, \mathbf{e}_M)$ over aleatory uncertainty.

In most PAs, the integrals in Eqs. (3.6.19)-(3.6.21) are too complex to estimate with formal quadrature procedures, with the result that these integrals are typically estimated with procedures based on some variant of random sampling or stratified sampling. In turn, the indicated sampling procedures result in estimates of the form

$$p_A(\tilde{y} \le y \mid \mathbf{e}) \cong \begin{cases} \sum_{i=1}^{m} \underline{\delta}_y \left[ f(\tau \mid \mathbf{a}_i, \mathbf{e}_M) \right] / m & \text{or} \\ \sum_{i=1}^{\tilde{m}} \underline{\delta}_y \left[ f(\tau \mid \tilde{\mathbf{a}}_i, \mathbf{e}_M) \right] p_A(\mathcal{A}_i \mid \mathbf{e}_A), \end{cases} \tag{3.6.22}$$

$$p_A(y < \tilde{y} \mid \mathbf{e}) \cong \begin{cases} \sum_{i=1}^{m} \overline{\delta}_y \left[ f(\tau \mid \mathbf{a}_i, \mathbf{e}_M) \right] / m & \text{or} \\ \sum_{i=1}^{\tilde{m}} \overline{\delta}_y \left[ f(\tau \mid \tilde{\mathbf{a}}_i, \mathbf{e}_M) \right] p_A(\mathcal{A}_i \mid \mathbf{e}_A) \end{cases} \tag{3.6.23}$$

and

$$E_A(y \mid \mathbf{e}) \cong \begin{cases} \sum_{i=1}^{m} f(\tau \mid \mathbf{a}_i, \mathbf{e}_M) / m & \text{or} \\ \sum_{i=1}^{\tilde{m}} f(\tau \mid \tilde{\mathbf{a}}_i, \mathbf{e}_M) p_A(\mathcal{A}_i \mid \mathbf{e}_A), \end{cases} \tag{3.6.24}$$

where (i) $\mathbf{a}_i$, $i = 1, 2, \ldots, m$, is a sample from $\mathcal{A}$ generated in consistency with the density function $d_A(\mathbf{a} \mid \mathbf{e}_A)$ for aleatory uncertainty and (ii) the sets (i.e., strata) $\mathcal{A}_i$, $i = 1, 2, \ldots, \tilde{m}$, satisfy $\cup_i \mathcal{A}_i = \mathcal{A}$ with $\mathcal{A}_i \cap \mathcal{A}_j = \varnothing$ for $i \neq j$, $\tilde{\mathbf{a}}_i$ is a representative element of $\mathcal{A}_i$, and $p_A(\mathcal{A}_i \mid \mathbf{e}_A)$ is the probability of $\mathcal{A}_i$. The summations with $\mathbf{a}_i$ and $\tilde{\mathbf{a}}_i$ correspond to approximations with simple random or quasi-random sampling and stratified sampling, respectively.

Use of the alternative approximations with the sets $\mathcal{A}_i$ and associated values $\tilde{\mathbf{a}}_i$ in Eqs. (3.6.22) -(3.6.24) corresponds to use of the ordered triple representation for risk in Eq. (3.2.1) with $\mathcal{S}_i = \mathcal{A}_i$, $p\mathcal{S}_i = p_A(\mathcal{A}_i \mid \mathbf{e}_A)$, and $c\mathcal{S}_i = f(\tau \mid \tilde{\mathbf{a}}_i, \mathbf{e}_M)$. In effect, representations of this form are being employed when an event tree is used to represent the effects of aleatory uncertainty, with each path through the tree being used to both define one set $\mathcal{A}_i$ and determine the probability of this set.

As an example, the CCDFs for normalized radionuclide release to the accessible environment obtained in the 1996 WIPP PA are shown in Figure 3.5a. In this example, the individual CCDFs are calculated as indicated in Eq. (3.6.23) with a random sample of size $10^4$ from the sample space for aleatory uncertainty, and the distribution of CCDFs results from repeating this calculation for each element $\mathbf{e}_k = [\mathbf{e}_{Ak}, \mathbf{e}_{Mk}]$ of an LHS of size $n$ = 100 from the sample space for epistemic uncertainty (See Ref. [112] for computational details). As indicated immediately after Eq. (3.6.20), each CCDF in Figure 3.5a is a plot of points of the form $[y, p_A(y \leq \tilde{y} \mid \mathbf{e}_k)]$ with $y$ corresponding to normalized release $R$ at 10,000 yr. Specifically, the indicated release is the integrated (i.e., total) release that takes place from repository closure (i.e., time 0 yr) out to 10,000 yr; thus, in the formal model representation $f(\tau \mid \mathbf{a}, \mathbf{e}_M)$ used in Eqs. (3.6.1)-(3.6.24), the time variable $\tau$ corresponds to 10,000 yr in Figure 3.5a.

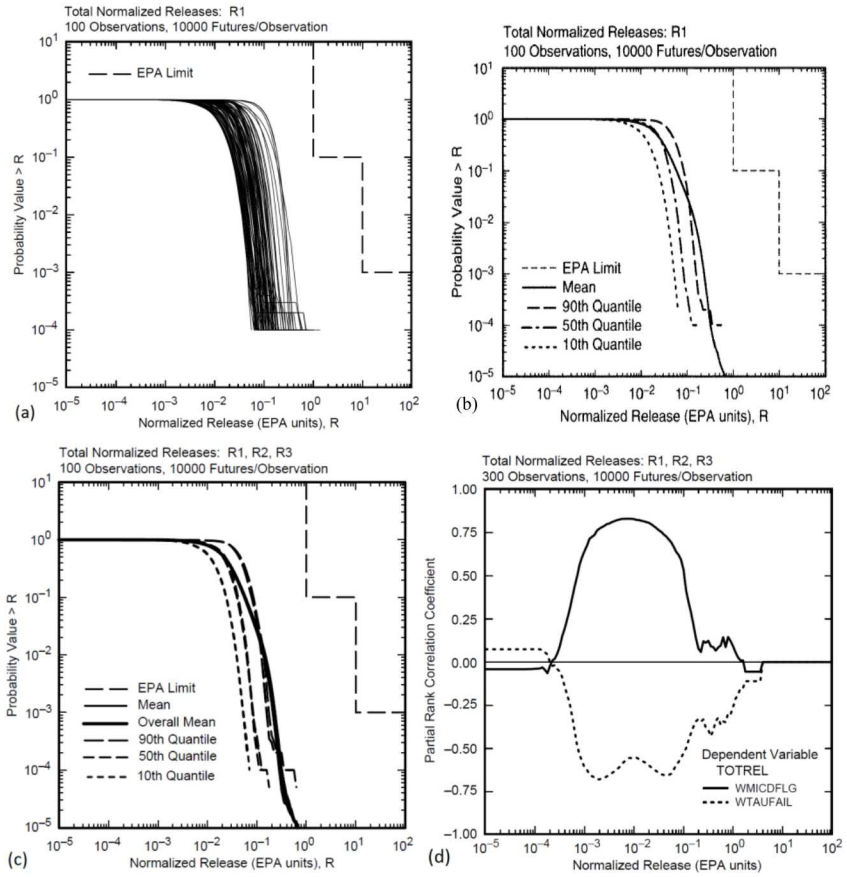



Figure 3.5 Distribution of CCDFs for normalized release to the accessible environment over $10^4$ yr obtained in the 1996 WIPP PA: (a) Individual CCDFs for LHS of size 100 ([112], Figure 1a), (b) Mean and quantile curves for CCDFs in Frame a ([112], Figure 1b), (c) Replicated mean and quantile curves for CCDFs in Frame a for 3 independent LHSs (i.e., R1, R2, R3) of size 100 ([112], Figure 2a), and (d) Partial rank correlation coefficients (PRCCs) for all 300 CCDFs used to generate results in Frame c ([112], Figure 5).

Displays of the form shown in Figure 3.5a provide a visual impression of the effects of epistemic uncertainty in the determination of the presented results. A more quantitative presentation of the effects of epistemic uncertainty is provided by a plot of the expected value and selected quantile values (e.g., $q = 0.05, 0.5, 0.95$) for the exceedance probabilities associated with individual values on the abscissa (e.g., normalized release $R$ as in Figure 3.5a). In general and with the same notation as used in Eqs. (3.6.22)-(3.6.24), the indicated expected value $E_E[p_A(y < \tilde{y} \mid \mathbf{e})]$ is given by

$$E_E\left[p_A\left(y<\tilde{y}\,|\,\mathbf{e}\right)\right]=\int_{\mathcal{E}}\left\{\int_{\mathcal{A}}\overline{\delta}_y\left[f\left(\tau\,|\,\mathbf{a},\mathbf{e}_M\right)\right]d_A\left(\mathbf{a}\,|\,\mathbf{e}_A\right)\mathrm{d}\mathcal{A}\right\}d_E(\mathbf{e})\mathrm{d}\mathcal{E}$$

$$\cong\begin{cases}\displaystyle\sum_{k=1}^{n}\left\{\sum_{i=1}^{m}\overline{\delta}_y\left[f\left(\tau\,|\,\mathbf{a}_i,\mathbf{e}_{Mk}\right)\right]/m\right\}/n & \text{or}\\ \displaystyle\sum_{k=1}^{n}\left\{\sum_{i=1}^{\tilde{m}}\overline{\delta}_y\left[f\left(\tau\,|\,\tilde{\mathbf{a}}_i,\mathbf{e}_{Mk}\right)\right]p_A\left(\mathcal{A}_i\,|\,\mathbf{e}_{Ak}\right)\right\}/n\end{cases}\qquad(3.6.25)$$

and the $q$ quantile value $Q_{Eq}[p_A(y<\tilde{y}\,|\,\mathbf{e})]$ is the value of $p=p_A(y<\tilde{y}\,|\,\mathbf{e})$ such that

$$\begin{aligned}q&=\int_{\mathcal{E}}\underline{\delta}_p\left[p_A\left(y<\tilde{y}\,|\,\mathbf{e}\right)\right]d_E(\mathbf{e})\mathrm{d}\mathcal{E}\\&=\int_{\mathcal{E}}\underline{\delta}_p\left\{\int_{\mathcal{A}}\overline{\delta}_y\left[f\left(\tau\,|\,\mathbf{a},\mathbf{e}_M\right)\right]d_A\left(\mathbf{a}\,|\,\mathbf{e}_A\right)\mathrm{d}\mathcal{A}\right\}d_E(\mathbf{e})\mathrm{d}\mathcal{E}\\&\cong\begin{cases}\displaystyle\sum_{k=1}^{n}\underline{\delta}_p\left\{\sum_{i=1}^{m}\overline{\delta}_y\left[f\left(\tau\,|\,\mathbf{a}_i,\mathbf{e}_{Mk}\right)\right]/m\right\}/n & \text{or}\\ \displaystyle\sum_{k=1}^{n}\underline{\delta}_p\left\{\sum_{i=1}^{\tilde{m}}\overline{\delta}_y\left[f\left(\tau\,|\,\tilde{\mathbf{a}}_i,\mathbf{e}_{Mk}\right)\right]p_A\left(\mathcal{A}_i\,|\,\mathbf{e}_{Ak}\right)\right\}/n.\end{cases}\end{aligned}\qquad(3.6.26)$$

As an example, mean and quantile (i.e., percentile) results for the CCDFs in Figure 3.5a are presented in Figure 3.5b.

Results analogous to those illustrated in Figure 3.5a and b can also be obtained and presented for CDFs. However, CCDFs rather than CDFs usually provide the preferred presentation format for results obtained in PAs for two reasons. First, CCDFs provide an answer to questions of the form "How likely is it to be this bad or worse?", which is typically the type of question of most interest in a PA. Second, CCDFs facilitate the display of the probabilities associated with low probability but high consequence aleatory occurrences (especially when a $\log_{10}$ scale is used on the ordinate); in contrast, such probabilities are difficult to obtain from a CDF owing to the lack of resolution in displayed probabilities as cumulative probability approaches 1.0.

As previously indicated, the 1996 WIPP PA used an LHS of size 100 in the propagation of epistemic uncertainty. To assess the adequacy of this sample size, the analysis was repeated with 3 replicated LHSs of size 100 denoted R1, R2 and R3. As shown in Figure 3.5c, there is little variability in the results obtained with the individual replicated samples. The sensitivity analysis results with PRCCs in Figure 3.5d are discussed in Sect. 7.

An interesting potential regulatory requirement is also illustrated in Figure 3.5a and b. The two step boundary labeled "EPA Limit" in the upper right corner of Figure 3.5a derives from a licensing regulation for the WIPP established by the U.S. Environmental Agency (EPA) which requires for the time interval [0, $10^4$ yr] after repository closure that (i) the probability of exceeding a normalized release of size $R = 1.0$ shall be less than 0.1 and (ii) the probability of exceeding a normalized release of size $R = 10.0$ shall be less than 0.001. As specified by the EPA, the indicated

requirement is violated if the mean CCDF in Figure 3.5b crosses the indicated boundary line (see Ref. [27] for additional details). A requirement of this type places stronger restrictions on system outcomes as the severity of these outcomes increases and is known as the Farmer limit line approach to the definition of acceptable risk [113-115].

An additional example of the representation of the effects of aleatory uncertainty conditional on a specific realization of epistemic uncertainty is provided by the treatment in the 2008 YM PA of expected dose $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ over aleatory uncertainty to the RMEI conditional on a realization $\mathbf{e} = [\mathbf{e}_A, \mathbf{e}_M]$ of epistemic uncertainty. In the preceding, $D(\tau|\mathbf{a}, \mathbf{e}_M)$ represents dose (mrem/yr) to the RMEI at time $\tau$ (yr) conditional on the "future" defined by the vector $\mathbf{a}$ of aleatory quantities (see Eq. (3.3.3)) and the vector $\mathbf{e}_M$ of epistemically uncertainty quantities (see Table 3.3) used in the calculation of dose to the RMEI; and $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ is defined as indicated in Eq. (3.6.21). Specifically,

$$E_A\left[D\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right) \mid \mathbf{e}_A\right] = \int_{\mathcal{A}} D\left(\tau \mid \mathbf{a}, \mathbf{e}_M\right) d_A\left(\mathbf{a} \mid \mathbf{e}_A\right) \mathrm{d}\mathcal{A}. \tag{3.6.27}$$

The time-dependent expected values $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ for $D(\tau|\mathbf{a}, \mathbf{e}_M)$ were estimated for the time intervals [0, 20,000 yr] and [0, $10^6$ yr] for each of the LHS sample elements $\mathbf{e}_k = [\mathbf{e}_{Ak}, \mathbf{e}_{Mk}]$ indicated in Eq. (3.6.8).

The resultant expected dose curves $[\tau, E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]]$, $k = 1, 2, \ldots, n = 300$, the associated curves for the quantiles $Q_{Eq}\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$, $q = 0.05, 0.5, 0.95$, and the expected value $E_E\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ for dose over aleatory and epistemic uncertainty are shown in Fig. 3.6a. The quantities $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ and $E_E\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ are both expected doses but with expectations calculated with respect to different probability spaces. In consistency with regulatory requirements that place bounds on $E_E\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ for the YM repository, the 2008 YM PA refers to $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ and $E_E\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ as "expected dose" and "expected (mean) dose", respectively [28]. A more detailed summary of the distribution results for $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ at $\tau = 10^4$ yr is presented in Fig. 3.6b, and the stability of the results obtained with the three replicated LHSs of size 300 is illustrated in Fig. 3.6c by the similarity of the resultant plots for estimates of $Q_{Eq}\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$, $q = 0.05, 0.5, 0.95$, and $E_E\{E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$. The sensitivity analysis results in Fig. 3.6d are discussed in Sect. 3.7.

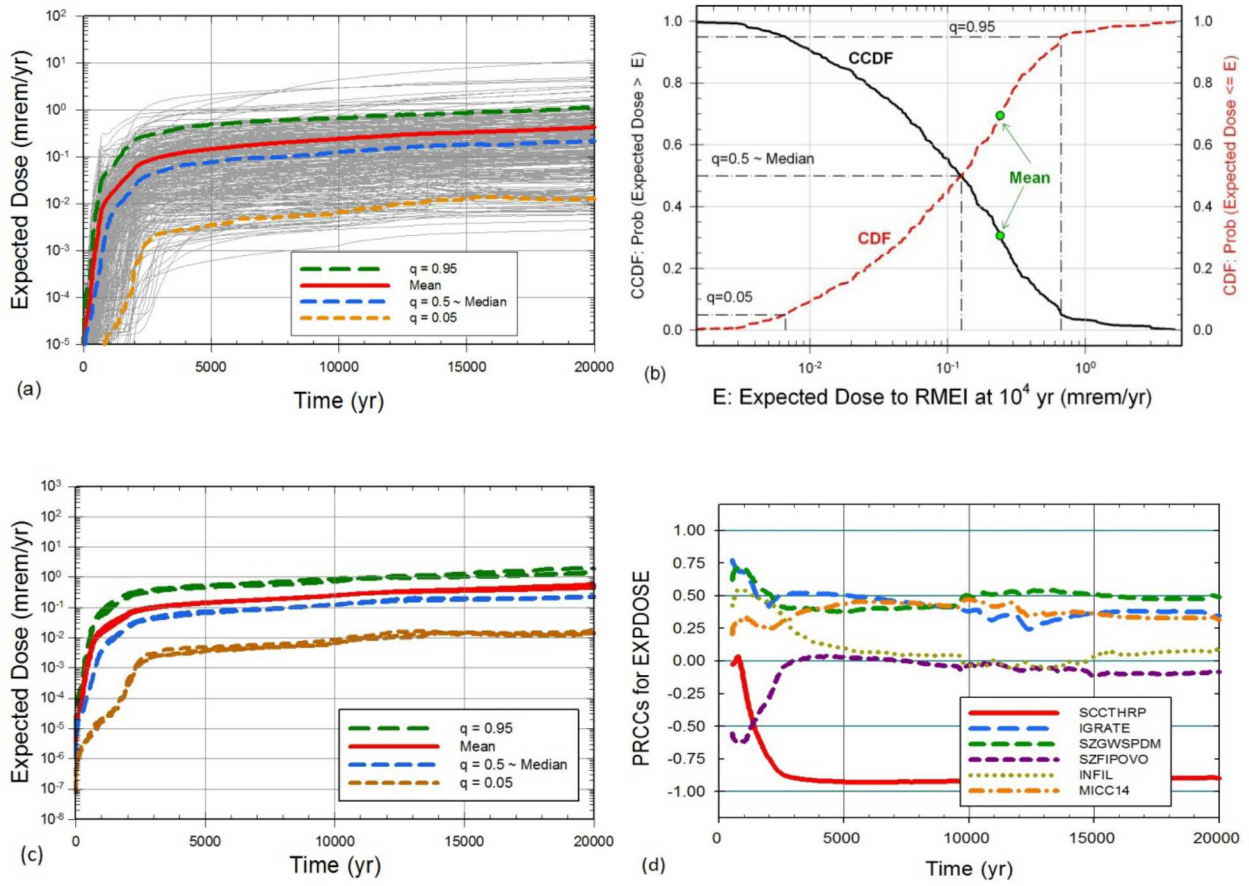



Fig. 3.6 Estimates obtained with an LHS of size $n$ = 300 of expected dose $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]$ to the RMEI in the 2008 YM PA for the time interval [0, 20,000 yr]: (a) Expected dose $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$, $k$ = 1, 2, …, $n$ = 300, expected (mean) dose $E_E\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, and quantiles $Q_{Eq}\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, $q$ = 0.05, 0.5, 0.95 ([106], Fig. 1a), (b) Cumulative probabilities $p_E\{E_A[D(10^4 \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A] \le D\}$, exceedance probabilities $p_E\{D < E_A[D(10^4 \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, expected (mean) dose $E_E\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, and quantiles $Q_{Eq}\{E_A[D(10^4 \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, $q$ = 0.05, 0.5, 0.95, for $\tau = 10^4$ yr ([106], Fig. 1b), (c) Estimates of $E_E\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$ and $Q_{Eq}\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$, $q$ = 0.05, 0.5, 0.95, obtained with three replicated LHSs of size $n$ = 300 as indicated in Eq. (3.6.17) ([106], Fig. 5), and (d) Partial rank correlation coefficients (PRCCs) for $E_E\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$ ([106], Fig. 6).

The numerics of determining the expected doses $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]]$ in Fig. 3.6a are quite complicated and will be discussed briefly after this paragraph. Once the expected doses $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ are available, the estimation of $Q_{Eq}\{E_A[D(\tau \mid \mathbf{a},\mathbf{e}_M) \mid \mathbf{e}_A]\}$ and

$E_E\{E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_M)\,|\,\mathbf{e}_A]\}$ is straight forward. Specifically, $Q_{Eq}\{E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_M)\,|\,\mathbf{e}_A]\}$ corresponds to the value $y$ defined and approximated by

$$\begin{aligned} q &= \int_{\mathcal{E}} \underline{\delta}_y \left\{E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_M\right)|\,\mathbf{e}_A\right]\right\} d_E\left(\mathbf{e}\right) \mathrm{d}\mathcal{E} \\ &\cong \sum_{k=1}^{n} \underline{\delta}_y \left\{E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk}\right)|\,\mathbf{e}_{Ak}\right]\right\} / n, \end{aligned} \tag{3.6.28}$$

and $E_E\{E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_M)\,|\,\mathbf{e}_A]\}$ is defined and approximated by

$$\begin{aligned} E_E\left\{E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_M\right)|\,\mathbf{e}_A\right]\right\} &= \int_{\mathcal{E}} E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_M\right)|\,\mathbf{e}_A\right] d_E\left(\mathbf{e}\right) \mathrm{d}\mathcal{E} \\ &\cong \sum_{k=1}^{n} E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk}\right)|\,\mathbf{e}_{Ak}\right] / n. \end{aligned} \tag{3.6.29}$$

Similarly, $Q_{Eq}\{E_A[y(\tau\,|\,\mathbf{a},\mathbf{e}_M)\,|\,\mathbf{e}_A]\}$ and $E_E\{E_A[y(\tau\,|\,\mathbf{a},\mathbf{e}_M)\,|\,\mathbf{e}_A]\}$ for an arbitrary analysis result $y(\tau|\mathbf{a},\ \mathbf{e}_M)$ are defined and approximated as indicated in Eqs. (3.6.28) and (3.6.29) with $y(\tau|\mathbf{a},\ \mathbf{e}_M)$ and $y(\tau|\mathbf{a},\ \mathbf{e}_{Mk})$ replacing $D(\tau|\mathbf{a},\ \mathbf{e}_M)$ and $D(\tau|\mathbf{a},\ \mathbf{e}_{Mk})$.

The numerics of determining the expected doses $E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk})\,|\,\mathbf{e}_{Ak}]$ are now considered. In concept, $E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk})\,|\,\mathbf{e}_{Ak}]$ can be approximated by

$$E_A\left[D\left(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk}\right)|\,\mathbf{e}_A\right] \cong \sum_{i=1}^{m} D\left(\tau\,|\,\mathbf{a}_i,\mathbf{e}_{Mk}\right) / m, \tag{3.6.30}$$

where $\mathbf{a}_i$, $i$ = 1, 2, …, $m$, is a sample from the probability space $\mathcal{A}$ for aleatory uncertainty (see Eq. (3.3.3)) generated in consistency with the density function $d_A(\mathbf{a}\,|\,\mathbf{e}_A)$. As is the case in many complex analyses, the overall structure of the 2008 YM PA was too complex (see Eq. (3.3.3) and Fig. 3.2) for this sampling-based approach for the estimation of $E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk})\,|\,\mathbf{e}_{Ak}]$ to be practicable. In such situations, it is necessary to (i) decompose the analysis into a number of intermediate calculations smaller than a complete calculation of the result of interest (e.g., $E_A[D(\tau\,|\,\mathbf{a},\mathbf{e}_{Mk})\,|\,\mathbf{e}_{Ak}]$ in Eq. (3.6.30) and Fig. 3.6) and then (ii) combine these intermediate results in a computationally efficient manner to obtain the final result of interest. The indicated combination of intermediate results involves taking advantage of computational efficiencies that derive from specific properties of the analysis and can include (i) division of the original integral over the sample space $\mathcal{A}$ for aleatory uncertainty into a sum of integrals over subsets of $\mathcal{A}$, (ii) use of additivity, linearity and interpolations in conjunction with intermediate results to obtain

additional results, and (iii) development of one or more metamodels that are approximations to the original model.

In the 2008 YM PA, the first step in the evaluation of $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ was to divide the defining integral for $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ in Eq. (3.6.27) into a sum of integrals over selected subsets of the sample space $\mathcal{A}$ for aleatory uncertainty. These subsets are referred to as scenario classes (and sometimes as modeling cases) in the 2008 YM PA and are denoted and defined with the notation used in Eq. (3.3.3)-(3.3.6) by (i) $\mathcal{A}_{EW} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nEW \geq 1\}$ for the early WP failure class, (ii) $\mathcal{A}_{ED} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nED \geq 1\}$ for the early DS failure class, (iii) $\mathcal{A}_{II} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nII \geq 1\}$ for the igneous intrusion failure class, (iv) $\mathcal{A}_{IE} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nIE \geq 1\}$ for the igneous eruption failure class, (v) $\mathcal{A}_{SG} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nSG \geq 1\}$ for the seismic ground motion failure class, (vi) $\mathcal{A}_{SF} = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nSF \geq 1\}$ for the seismic fault displacement failure class, and (vii) $\mathcal{A}_N = \{\mathbf{a} \mid \mathbf{a} \in \mathcal{A} \text{ and } nEW = nED = nII = nIE = nSG = nSF = 0\} = \{\mathbf{a}_N\}$ for the nominal failure class. In turn, the indicated sum of integrals is given by

$$
\begin{aligned}
E_A\left[D\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) \mid \mathbf{e}_{Ak}\right] &= \int_{\mathcal{A}} D\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) d_A\left(\mathbf{a} \mid \mathbf{e}_{Ak}\right) \mathrm{d}\mathcal{A} \\
&\cong D_N\left(\tau \mid \mathbf{a}_N,\mathbf{e}_{Mk}\right) + \int_{\mathcal{A}} \sum_{C \in \mathcal{MC}} D_C\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) d_A\left(\mathbf{a} \mid \mathbf{e}_{Ak}\right) \mathrm{d}\mathcal{A} \\
&= D_N\left(\tau \mid \mathbf{a}_N,\mathbf{e}_{Mk}\right) + \sum_{C \in \mathcal{MC}} \int_{\mathcal{A}} D_C\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) d_A\left(\mathbf{a} \mid \mathbf{e}_{Ak}\right) \mathrm{d}\mathcal{A} \qquad (3.6.31) \\
&= D_N\left(\tau \mid \mathbf{a}_N,\mathbf{e}_{Mk}\right) + \sum_{C \in \mathcal{MC}} \int_{\mathcal{A}_C} D_C\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) d_A\left(\mathbf{a} \mid \mathbf{e}_{Ak}\right) \mathrm{d}\mathcal{A} \\
&= D_N\left(\tau \mid \mathbf{a}_N,\mathbf{e}_{Mk}\right) + \sum_{C \in \mathcal{MC}} E_A\left[D_C\left(\tau \mid \mathbf{a},\mathbf{e}_{Mk}\right) \mid \mathbf{e}_{Ak}\right],
\end{aligned}
$$

with (i) $\mathcal{MC} = \{EW, ED, II, IE, SG, SF\}$ and (ii) the doses $D_C(\tau \mid \mathbf{a},\mathbf{e}_{Mk})$ ) for scenario class $\mathcal{A}_C$ calculated with use of only the elements of $\mathbf{a}$ and $\mathbf{e} = [\mathbf{e}_A,\mathbf{e}_M]$ that constitute part of the defining conditions for scenario class $\mathcal{A}_C$ (i.e., early WP failure, early DS failure, igneous intrusion, igneous eruption, seismic ground motion, or seismic fault displacement). The preceding approximation to $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ was justified for use in the 2008 YM PA on the basis of tradeoffs between the effects of high probability-low consequence scenario classes and low probability-high consequence scenario classes ([106], Sect. 5).

Use of the decomposition in Eq. (3.6.31) reduces the determination of $E_A[D(\tau \mid \mathbf{a},\mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ from the evaluation of a single very complex integral to the evaluations of (i) $D_N(\tau \mid \mathbf{a}_N,\mathbf{e}_{Mk})$ and (ii) six individual integrals (i.e., one integral for each element of the set $\mathcal{MC}$). The six integrals

for $D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ over $\mathcal{A}_C$ for $C \in \mathcal{MC}$ that need to be evaluated are still complex, but less complex than a single integral for $D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ over $\mathcal{A}$. The advantage of the decomposition in Eq. (3.6.31) is that the integrals for $D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ over $\mathcal{A}_C$ for $C \in \mathcal{MC}$ can be implemented in ways that obtain computational efficiencies that take advantage of specific properties of the dose function $D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ and the restriction of the density function to the set $\mathcal{A}_C$ for scenario class $C$.

As a single example, a summary description of the evaluation of $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ over the time interval [0, 20,000 yr] for the seismic ground motion scenario class will be used for illustration; a full description of the analysis for $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ is given in Refs. [79; 116; 117]. In effect, the consideration of only seismic ground motion events reduces the sample space $\mathcal{A}$ for aleatory uncertainty to the previously indicated set $\mathcal{A}_{SG}$ with elements $\mathbf{a}$ defined by

$$\mathbf{a} = [nSG, t_1, v_1, A_1, t_2, v_2, A_2, ..., t_{nSG}, v_{nSG}, A_{nSG}], \tag{3.6.32}$$

where (i) $nSG$ = number of seismic ground motion events in 20,000 yr, (ii) $t_i$ = time (yr) of event $i$, (iii) $v_i$ = peak ground motion velocity (m/s) for event $i$, (iv) $A_i$ = damaged area (m$^2$) on individual WPs for peak ground motion velocity $v_i$, (v) the occurrence of seismic ground motion events is characterized by a hazard curve for peak ground motion velocity, and (vi) damaged area is characterized by distributions conditional on peak ground motion velocity. In effect, $\mathcal{A}_{SG}$ is the sample space for aleatory uncertainty when only seismic ground motion events are considered over the time interval [0, 20,000 yr].

To evaluate $E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$, it is necessary to integrate the function $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ over the set $\mathcal{A}_{SG}$ with $\mathbf{a}$ defined as indicated in Eq. (3.6.32). In full detail, $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ is defined by the model system indicated in Fig. 3.2. Evaluation of this system is too computationally demanding to permit its evaluation 1000's of times for each element $\mathbf{e}_k = [\mathbf{e}_{Ak}, \mathbf{e}_{Mk}]$ of the LHS in Eq. (3.6.8). This is a common situation in analyses of complex systems, where very detailed physical models are developed which then turn out to be too computationally demanding to be naively used in the propagation of aleatory uncertainty. In such situations, it is necessary to find ways to efficiently use the results of a limited number of model evaluations to predict outcomes for a large number of different possible realizations of aleatory uncertainty.

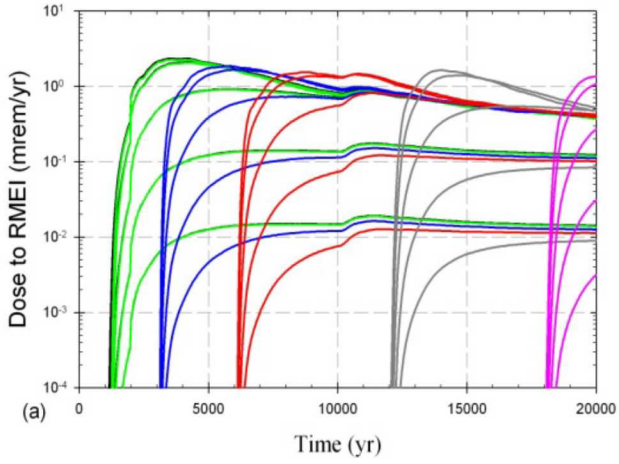
For the seismic ground motion scenario class and the time interval [0, $2\times10^4$ yr], the needed computational efficiency was achieved by evaluating $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ at a sequence of seismic event times (i.e., 100, 1000, 3000, 6000, 12000, 18000 yrs) and for a sequence of damaged areas (i.e., $10^{-8+s}$(32.6 m$^2$) for $s$ = 1, 2, …, 5 with 32.6 m$^2$ corresponding to the surface area of a WP) at each of the indicated times (Fig. 3.7a). This required 6×5 = 30 evaluations of the system indicated in Fig. 3.2 for each LHS element in Eq. (3.6.8). Once obtained, these evaluations can be used with

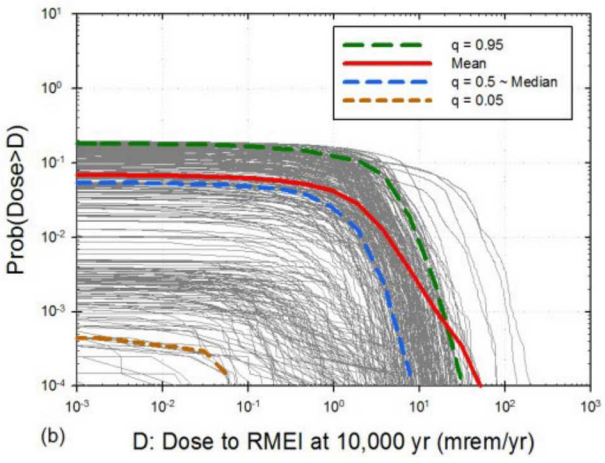
appropriate interpolation and additive procedures to evaluate $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ for different values of $\mathbf{a}$ for each LHS element $\mathbf{e}_k = [\mathbf{e}_{Ak}, \mathbf{e}_{Mk}]$ (see Ref. [79], Sect. 4, for computational details).

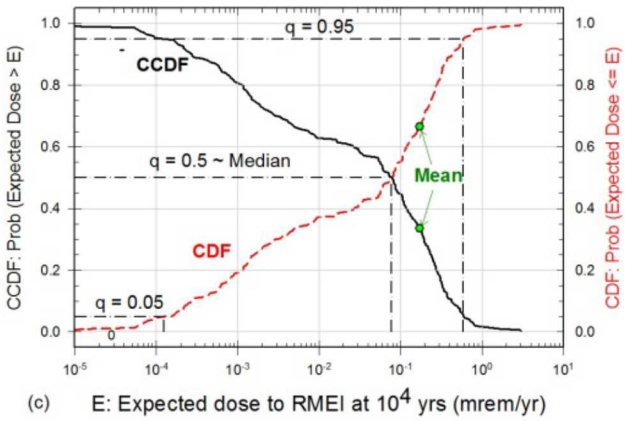
The individual CCDFs in Fig. 3.7b are defined by probabilities of the form shown in Eq. (3.6.20) with (i) $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$ and $\mathbf{e}_{Ak}$ replacing $y(\tau|\mathbf{a}, \mathbf{e}_M)$ and $\mathbf{e}_A$ and (ii) $\tau = 10^4$ yr. Numerically, the integrals that define exceedance probabilities for the individual CCDFs are approximated with (i) random sampling from the possible values for $\mathbf{a}$ as indicated in Eq. (3.6.23) and (ii) estimated values $\hat{D}_{SG}(10^4 \mid \mathbf{a}_i, \mathbf{e}_{Mk})$ for $D_{SG}(10^4 \mid \mathbf{a}_i, \mathbf{e}_{Mk})$ constructed from results of the form shown in Fig. 3.7a (see Ref. [79], Sect. 4, for computational details). Specifically,

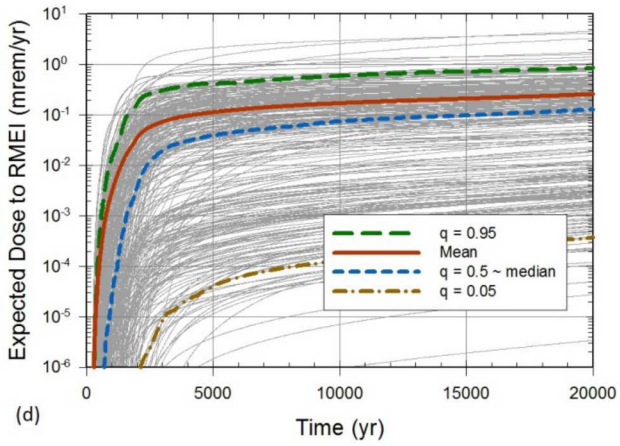
$$\hat{p}_A\left[y < D_{SG}\left(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}\right) \mid \mathbf{e}_{Ak}\right] = \sum_{i=1}^{m} \overline{\delta}_y\left[\hat{D}_{SG}\left(10^4 \mid \mathbf{a}_i, \mathbf{e}_{Mk}\right)\right] / m, \tag{3.6.33}$$

with the $\mathbf{a}_i$, $i = 1, 2, \ldots, m$, sampled in consistency with the density function $d_A(\mathbf{a} \mid \mathbf{e}_{Ak})$ for vectors of the form shown in Eq. (3.6.32). The mean and quantile curves in Fig. 3.7b are (i) defined and approximated as indicated in Eqs. (3.6.25) and (3.6.26) and (ii) provide a summary of the epistemic uncertainty present in the estimation of exceedance probabilities (i.e., $p_A[y < D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for $D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk})$.




| *EXPDOSE* : 10,000 yr | | |
|---|---|---|
| **Variable[a]** | **$R^{2\,b}$** | **SRRC[c]** |
| *SCCTHRP* | 0.88 | -0.93 |
| *MICTC99* | 0.90 | 0.12 |
| *DSNFMASS* | 0.91 | 0.14 |
| *HLWDRACD* | 0.91 | 0.08 |
| *MICC14* | 0.92 | 0.08 |

a: Variables listed in order of selection
b: Cumulative $R^2$ value
c: SRRCs in final regression model

(e)

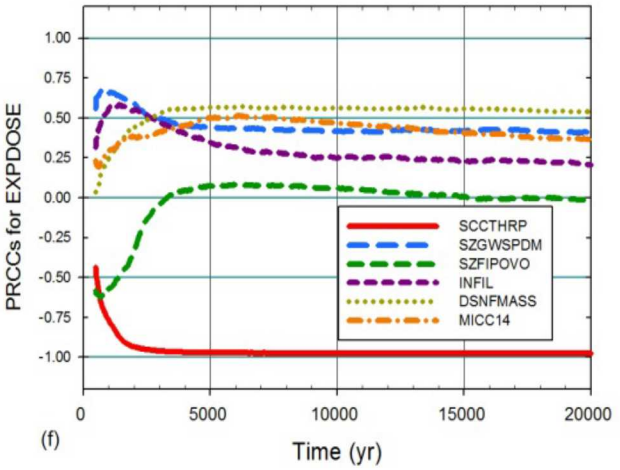



Figure 3.7 Example results for dose (mrem/yr) to RMEI in the YM PA for seismic ground motion scenario class: (a) Dose for seismic events occurring at different times and causing different damaged areas on WPs ([79], Figure 4a), (b) CCDFs for dose at 10,000 yr ([79], Figure 10), (c) CCDF for expected dose at 10,000 yr ([79], Figure 6b), (d) Time-dependent expected dose ([79], Figure 6a), (e) Stepwise rank regression for expected dose at 10,000 yr ([117], Table 4), and (f) Time-dependent PRCCs for expected dose ([117], Figure 4b).

As indicated in Eqs. (3.6.21) and (3.6.24), the expected value $E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ of $D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk})$ over aleatory uncertainty can also be defined and estimated, with the estimates $\hat{E}_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ obtained with the random samples used to construct the CCDFs in Figure 3.7b. The expected values $E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ and their corresponding estimates are the result of reducing each CCDF in Figure 3.7b to a single number. As illustrated in Figure 3.7c, the epistemic uncertainty associated with $E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ can be summarized by (i) an expected (mean) value $E_E\{E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ over epistemic uncertainty, (ii) a CDF defined by the cumulative probabilities for $E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$ or (iii) a CCDF defined by the complementary cumulative probabilities for $E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]$. The indicated expected value, CDF and CCDF are defined and approximated in a manner analogous to that used to obtain the results in Figure 3.3a with $\hat{E}_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ replacing $D_N(600{,}000 \mid \mathbf{a}_N, \mathbf{e}_M)$. The expected (mean) value $E_E\{E_A[D_{SG}(10^4 \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ illustrated in Figure 3.7c is the outcome of reducing all the information in Figure 3.7b to a single number.

Expected doses $E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for individual LHS elements correspond to the lighter lines in Figure 3.7d, and quantile values $Q_{Eq}\{E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]\}$ and expected (mean) values $E_E\{E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_M) \mid \mathbf{e}_A]\}$ for $E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ that summarize the effects of epistemic uncertainty correspond to the darker dashed and solid lines. The results on Figure 3.7d at 10,000 years correspond to the results shown in more detail in Figure 3.7c. For reasons of computational efficiency, the individual expected dose curves $E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for $0 \le \tau \le 20{,}000$ yr in Figure 3.7d were estimated with a quadrature procedure that used the results in Figure 3.7a rather than with a sampling-based procedure as illustrated in Figure 3.7b; details of this procedure are given in Ref. [79]. The sensitivity analysis results in Figure 3.7e and f are discussed in Sect. 3.7.

The expected doses $E_A[D_{SG}(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ in Fig. 3.7d constitute one of the six expected doses appearing in the approximation to $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ in Eq. (3.6.31) for $0 \le \tau \le 20{,}000$ yr. The other five expected doses $E_A[D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$, $C = EW, ED, II, IE, SF$, were determined in a generally similar manner in which a representative set of calculations analogous to those indicated in Fig. 3.7a were performed for each LHS element and then used to estimate $E_A[D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$. Details of these calculations are given in Refs. [77-79]. The resultant values for all six expected doses $E_A[D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for the LHS in Eq. (3.6.8) are shown in Fig. 3.8. As indicated in Eq. (3.6.31), the doses $E_A[D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ in Fig. 3.8 are added to obtain the approximations to $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for $0 \le \tau \le 20{,}000$ yr in Fig. 3.6a. Associated analyses showed that $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk}) = 0$ mrem/yr for $0 \le \tau \le 20{,}000$ yr (see Fig.

3.3b); thus, $D_N(\tau \mid \mathbf{a}_N, \mathbf{e}_{Mk})$ made no contribution to $E_A[D(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ for $0 \le \tau \le 20{,}000$ yr.

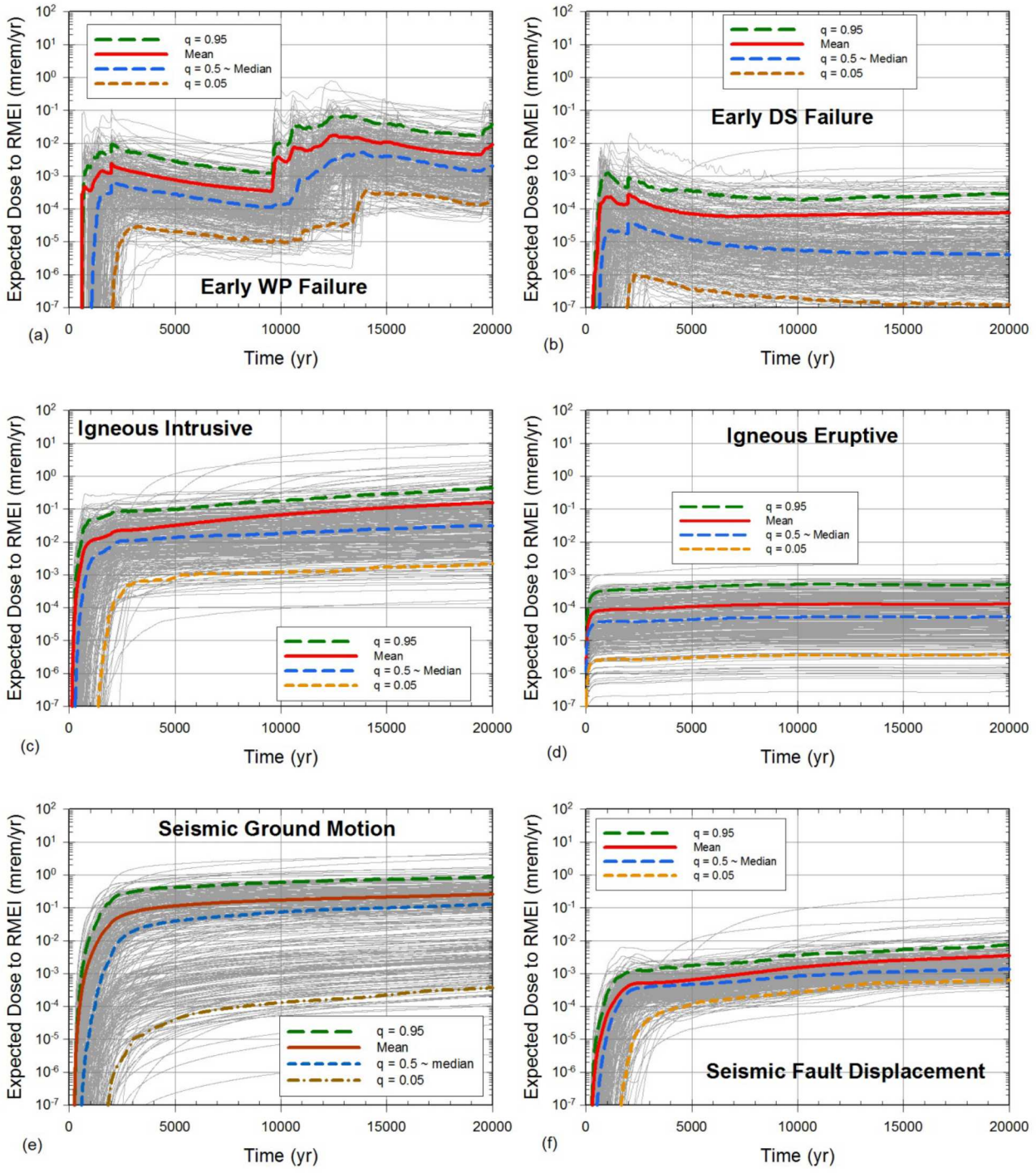



Figure 3.8 Expected dose $E_A[D_C(\tau \mid \mathbf{a}, \mathbf{e}_{Mk}) \mid \mathbf{e}_{Ak}]$ to RMEI for $0 \le \tau \le 20{,}000$ yr for $C \in \{EW, ED, II, IE, SG, SF\}$: (a) Early WP failure, (b) Early DS failure, (c) Igneous intrusion, (d) Igneous eruption, (e) Seismic ground motion, and (f) Seismic fault displacement ([106], Figure 2).

## 3.7 Sensitivity Analysis

Sensitivity analysis is an important part of a PA and, indeed, any analysis. Specifically, while uncertainty analysis is intended to determine the total uncertainty in analysis outcomes, sensitivity analysis is intended to determine the contributions of individual uncertain analysis inputs to the total uncertainty in analysis results of interest. The descriptor "sensitivity analysis" is usually used in reference to the determination of the effects of epistemic uncertainty on analysis outcomes of interest. As a result, sensitivity analysis is typically closely tied to the previously indicated propagation of epistemic uncertainty. In particular, a sampling-based propagation of epistemic uncertainty generates a mapping between uncertain analysis inputs and uncertain analysis results that can then be explored with a variety of sensitivity analysis procedures, including examination of scatterplots, correlation analysis, regression analysis, partial correlation analysis, rank transformations, statistical tests for patterns based on gridding, entropy tests for patterns based on gridding, nonparametric regression analysis, squared rank differences/rank correlation test, two dimensional Kolmogorov-Smirnov test, tests for patterns based on distance measures, top down coefficient of concordance, and variance decomposition [118; 119].

Example sensitivity analyses based on partial rank correlation coefficients (PRCCs) are presented in Figure 3.3d, Figure 3.5d, Fig. 3.6d and Figure 3.7f. In these examples, the PRCCs are determined by analyzing the uncertainty associated with the analysis results above individual values on the abscissas in the indicated figures and then connecting these results to form curves of sensitivity analysis results for individual analysis inputs ([118], Sects. 6.4 and 6.5; [120]) . As a result of the use of a rank transform ([118], Sect. 6.5; [121]), a PRCC provides a measure of the strength of the monotonic effect of an uncertain analysis input on an analysis result after the removal of the monotonic effects of all other uncertain analysis inputs. In the presence of nonlinear but monotonic relationships between a dependent variable and multiple independent (i.e., sampled) variables, use of the rank transform can substantially improve the resolution of sensitivity analysis results. Specifically, a sensitivity analysis based on rank transformed data operates in the following manner: (i) the smallest value for each variable is assigned a rank of 1, (ii) next largest value is assigned a rank of 2, (iii) tied values are assigned their average rank, and (iv) so on up to the largest value, which is assigned a rank equal to the sample size in use; then, the analysis is performed with these rank-transformed values. The rank transformation is a very useful procedure in many sensitivity analyses as it converts nonlinear but monotonic relationships into linear relationships.

The results in Figures 3.3d, Figure 3.5d, Fig. 3.6d and Figure 3.7f illustrate a spectrum of the possibilities in which sensitivity analyses can be performed in a PA, including on: (i) results conditional on a specific realization of aleatory uncertainty (Figure 3.3d), (ii) exceedance probabilities that define CCDFs that summarize the effects of aleatory uncertainty (Figure 3.5d), (iii) expected values over aleatory uncertainty (Fig. 3.6d and 3.7f), and (iv) intermediate results that are part of a very complex calculation (Figure 3.9).

As another example, stepwise rank regressions are presented in Figure 3.3c and Figure 3.7e for results at specific points in time. In stepwise regression, a regression model is first constructed with the most influential variable (i.e., $x_1$ as determined based on $R^2$ values for regression models containing only single variables). Then, a regression model is constructed with $x_1$ and the next

most influential variable (i.e., $x_2$ as determined based on $R^2$ values for regression models containing $x_1$ and each of the remaining variables). The process then repeats to determine $x_3$ in a similar manner and continues until no more variables with an identifiable effect on dependent variable can be found. Variable importance (i.e., sensitivity) is then indicated by (i) the order in which variables are selected in the stepwise process, (ii) the changes in cumulative $R^2$ values as additional variables are added to the regression model, and (iii) the standardized regression coefficients for the variables in the final regression model ([118], Sect. 6.3). In a rank regression, the variables are rank-transformed before the regression analysis is performed. Plots of standardized regression coefficients analogous to the plots of PRCCs in Figure 3.3d, Figure 3.5d, Fig. 3.6d and Figure 3.7f are also possible.

When relationships between individual independent variables and a dependent variable are both nonlinear and nonmonotonic, sensitivity analyses with partial correlation coefficients and stepwise regression procedures will perform poorly with both untransformed data and rank transformed data. In such situations, examination of scatterplots may be sufficient to reveal the effects of nonlinear relationships between individual independent variables and a dependent variable ([118], Sect. 6.6; [122]). Specifically, for a sample $\mathbf{x}_k = [x_{k1}, x_{k1}, ..., x_{k,nV}]$, $k$ = 1, 2, …, $n$, from $nV$ uncertain analysis inputs (e.g., as in Eq. (3.6.8)) and a corresponding analysis result $y_k = y(\mathbf{x}_k)$, the scatterplot for variable $x_l$ is a plot of the points $(x_{kl}, y_k)$ for $k$ = 1, 2, …, $n$. As an example, the well-defined relationship between *BHPRM* and pressure is clearly shown by the scatterplot in Figure 3.9 but missed by partial correlation and regression analyses with untransformed and rank transformed data. More formal procedures to identify nonlinear relationships in sampling-based analyses are provided by grid-based procedures ([118], Sect. 6.6; [123]), nonparametric regression procedures ([118], Sect. 6.8; [124-127]), and the squared rank differences/rank correlation test ([118], Sect. 6.9; [128]).

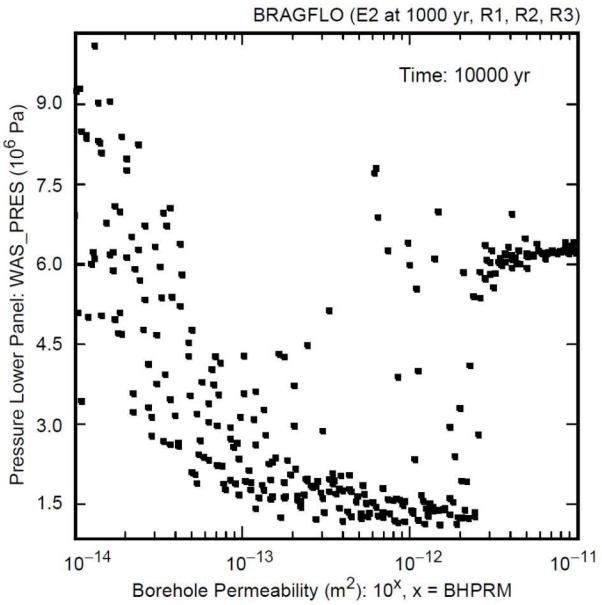



Figure 3.9 Example of a scatterplot obtained in the sampling-based uncertainty/sensitivity analysis for the WIPP ([129], Figure 28; see [129], Sect. 4, for additional discussion of this example).

Many additional examples of sampling-based sensitivity analysis results are available in Ref. [1] for the 1996 WIPP PA and in Ref. [2] for the 2006 YM PA. Further, Chapter 4 provides additional information on and illustrations of sampling-based procedures for uncertainty and sensitivity analysis. Additional information on both sampling-based approaches to uncertainty and sensitivity analysis and other approaches to uncertainty and sensitivity analysis (e.g., differential analysis and variance decomposition) is available in a number of reviews [118; 119; 123; 130-142]. In addition, Chapter 5 and Appendices B and C describe approaches to sensitivity analysis based on variance decomposition.

## 3.8 Summary

As described and illustrated in this presentation, a conceptual structure and an associated computational structure for PAs for radioactive waste disposal facilities and other complex engineered facilities can be based on the following three conceptual entities : EN1, a probability space $(\mathcal{A}, \mathbb{A}, p_A)$ that characterizes aleatory uncertainty; EN2, a function $f$ that estimates consequences for individual elements **a** of the sample space $\mathcal{A}$ for aleatory uncertainty; and EN3, a probability space $(\mathcal{E}, \mathbb{E}, p_E)$ that characterizes epistemic uncertainty [7; 8; 27; 28]. A recognition and understanding of these three basic entities makes it possible to understand the conceptual and computational structure of a large PA without having basic concepts obscured by fine details of the analysis and leads to an analysis that results in insightful uncertainty and sensitivity analyses. For example, calculations that are basically integrations of very complex functions can be recognized and described as such before the introduction of the often highly-complex numerical procedures needed to perform such integrations.

Use of the indicated uncertainty and sensitivity analysis procedures provide important insights into a PA and the factors that affect its outcomes. Uncertainty analysis results enhance the quality of decisions that can be made on the basis of a PA and also enhance the credibility of a PA by showing the resolution (i.e., uncertainty) in its results. Sensitivity analysis provides guidance on both (i) what are the dominant contributors to the uncertainty in results obtained in a PA and (ii) if necessary, where resources can be best invested to reduce the uncertainty in the results of a PA.

In addition, sampling-based uncertainty and sensitivity analysis procedures are important analysis verification tools. The extensive exercising of the models used in a PA in a sampling-based uncertainty and sensitivity analysis provides a very effective test of the operation of these models. Further, the associated sensitivity results are effective in revealing inappropriate relationships between variables that are indicative of errors in the formulation or implementation of the PA and its underlying models.

Owing to both the insights that can be gained and the potential for the identification of analysis errors, the examination of uncertainty and sensitivity results should never be limited to only the final outcomes of greatest interest in a PA. Rather, uncertainty and sensitivity results should be examined at multiple points in the chain of computations that lead to the final result or results of greatest interest (e.g., regulatory requirements). Examples of the analysis of intermediate results in a PA are provided by Refs. [129; 143-147] for the WIPP PA and Refs. [80; 117; 148; 149] for the YM PA.

## 3.10 References: Chapter 3

1. Helton J.C. and M.G. Marietta (Editors). 2000. Special Issue: The 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:1-451.
2. Helton J.C., C.W. Hansen and P.N. Swift (Editors). 2014. Special Issue: Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:1-456.
3. Helton J.C. and R.J. Breeding. 1993. Calculation of Reactor Accident Safety Goals. *Reliability Engineering and System Safety* 39:129-158.
4. Breeding R.J., J.C. Helton, E.D. Gorham and F.T. Harper. 1992. Summary Description of the Methods Used in the Probabilistic Risk Assessments for NUREG-1150. *Nuclear Engineering and Design* 135:1-27.
5. Breeding R.J., J.C. Helton, W.B. Murfin, L.N. Smith, J.D. Johnson, H.-N. Jow and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Surry Nuclear Power Station. *Nuclear Engineering and Design* 135:29-59.
6. Helton J.C., C.W. Hansen and C.J. Sallaberry. Conceptual Structure of Performance Assessments for the Geologic Disposal of Radioactive Waste *Proc. 10th International Probabilistic Safety Assessment & Management Conference Seattle, WA*, *2010*:
7. Helton J.C. and C.J. Sallaberry. 2012. Uncertainty and Sensitivity Analysis: From Regulatory Requirements to Conceptual Structure and Computational Implementation. *IFIP Advances in Information and Communication Technology* 377 AICT:60-76.
8. Helton J.C. 2003. Mathematical and Numerical Approaches in Performance Assessment for Radioactive Waste Disposal: Dealing with Uncertainty. In *Modelling Radioactivity in the Environment*, ed. EM Scott:353-390. New York, NY: Elsevier Science.
9. Helton J.C. and C.J. Sallaberry. 2010. Treatment of Uncertainty in Performance Assessments for the Geological Disposal of Radioactive Waste. In *Geological Reopositories for Safe Disposal of Spent Fuels and Radioactive Materials: Advanced Technologies*, ed. J Ahn, MJ Apted:547-579. Cambridge, UK: Woodhead Publishing.
10. Helton J.C., Hanson, C.W., Sallaberry, C.J. 2017. Conceptual Structure of Performance Assessments for Geologic Disposal of Radioactive Waste In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H., pp. 1-37. Switzerland: Springer International Publishing.
11. Kaplan S. and B.J. Garrick. 1981. On the Quantitative Definition of Risk. *Risk Analysis* 1:11-27.
12. Feller W. 1971. *An Introduction to Probability Theory and Its Applications*. New York, NY: John Wiley & Sons.
13. Parry G.W. and P.W. Winter. 1981. Characterization and Evaluation of Uncertainty in Probabilistic Risk Analysis. *Nuclear Safety* 22:28-42.
14. Parry G.W. 1996. The Characterization of Uncertainty in Probabilistic Risk Assessments of Complex Systems. *Reliability Engineering and System Safety* 54:119-126.

15. Apostolakis G. 1990. The Concept of Probability in Safety Assessments of Technological Systems. *Science* 250:1359-1364.

16. Apostolakis G. 1999. The Distinction between Aleatory and Epistemic Uncertainties is Important: An Example from the Inclusion of Aging Effects into PSA In *Proceedings of the International Topical Meeting on Probabilistic Safety Assessment, PSA '99: Risk Informed and Performance-Based Regulation in the New Millennium* 1:135-142. La Grange Park, IL: American Nuclear Society.

17. Helton J.C. 1994. Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis* 14:483-511.

18. Helton J.C. and D.E. Burmaster. 1996. Guest Editorial: Treatment of Aleatory and Epistemic Uncertainty in Performance Assessments for Complex Systems. *Reliability Engineering and System Safety* 54:91-94.

19. Helton J.C. 1997. Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty. *Journal of Statistical Computation and Simulation* 57:3-76.

20. Helton J.C., J.D. Johnson and C.J. Sallaberry. 2011. Quantification of Margins and Uncertainties: Example Analyses from Reactor Safety and Radioactive Waste Disposal Involving the Separation of Aleatory and Epistemic Uncertainty. *Reliability Engineering and System Safety* 96:1014-1033.

21. Hoffman F.O. and J.S. Hammonds. 1994. Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability. *Risk Analysis* 14:707-712.

22. Paté-Cornell M.E. 1996. Uncertainties in Risk Analysis: Six Levels of Treatment. *Reliability Engineering and System Safety* 54:95-111.

23. Winkler R.L. 1996. Uncertainty in Probabilistic Risk Assessment. *Reliability Engineering and System Safety* 54:127-132.

24. Ferson S. and L. Ginzburg. 1996. Different Methods are Needed to Propagate Ignorance and Variability. *Reliability Engineering and System Safety* 54:133-144.

25. Aven T. 2011. On Different Types of Uncertainties in the Context of the Precautionary Principle. *Risk Analysis* 31:1515-1525.

26. Aven T. and G. Reniers. 2013. How to Define and Interpret a Probability in a Risk and Safety Setting. *Safety Science* 51:223-231.

27. Helton J.C., D.R. Anderson, H.-N. Jow, M.G. Marietta and G. Basabilvazo. 2000. Conceptual Structure of the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:151-165.

28. Helton J.C., C.W. Hansen and C.J. Sallaberry. 2014. Conceptual Structure and Computational Organization of the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:223-248.

29. Hacking I. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. London; New York: Cambridge University Press.

30. Bernstein P.L. 1996. *Against the Gods: The Remarkable Story of Risk*. New York: John Wiley & Sons.

31. Rechard R.P. 1999. Historical Relationship Between Performance Assessment for Radioactive Waste Disposal and Other Types of Risk Assessment. *Risk Analysis* 19:763-807.

32. Knight F.H. 1921. *Risk, Uncertainty, and Profit*. New York, NY: Hart, Schaffner and Marx.

33. Lewis H.W., R.J. Budnitz, H.J.C. Kouts, W.B. Loewenstein, W.D. Rowe, F. von Hippel and F. Zachariasen. 1978. Risk Assessment Review Group Report to the U.S. Nuclear Regulatory Commission. *Rpt. NUREG/CR-0400*, U.S. Nuclear Regulatory Commission, Washington, D.C.

34. U.S. NRC (U.S. Nuclear Regulatory Commission). 1975. Reactor Safety Study-An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants. *Rpt. WASH-1400 (NUREG-75/014)*, U.S. Nuclear Regulatory Commission, Washington, DC.

35. Campbell J.E., R.T. Dillon, M.S. Tierney, H.T. Davis, P.E. McGrath, F.J. Pearson, H.R. Shaw, J.C. Helton and F.A. Donath. 1978. *Risk Methodology for Geologic Disposal of Radioactive Waste: Interim Report. Rpt. SAND78-0029, NUREG/CR-0458*, Sandia National Laboratories, Albuquerque, NM.

36. Iman R.L., J.C. Helton and J.E. Campbell. 1978. *Risk Methodology for Geologic Disposal of Radioactive Waste: Sensitivity Analysis Techniques. Rpt. SAND78-0912, NUREG/CR-0390*, Sandia National Laboratories, Albuquerque, NM.

37. Iman R.L., J.C. Helton and J.E. Campbell. 1981. An Approach to Sensitivity Analysis of Computer Models, Part 1. Introduction, Input Variable Selection and Preliminary Variable Assessment. *Journal of Quality Technology* 13:174-183.

38. Iman R.L., J.C. Helton and J.E. Campbell. 1981. An Approach to Sensitivity Analysis of Computer Models, Part 2. Ranking of Input Variables, Response Surface Validation, Distribution Effect and Technique Synopsis. *Journal of Quality Technology* 13:232-240.

39. Iman R.L., W.J. Conover and J.E. Campbell. 1980. Risk Methodology for Geologic Disposal of Radioactive Waste: Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment. *Rpt. SAND80-0020, NUREG-CR-1397*, Sandia National Laboratories, Albuquerque, NM.

40. Iman R.L. and W.J. Conover. 1980. Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment. *Communications in Statistics: Theory and Methods* A9:1749-1842.

41. Cranwell R.M., J.E. Campbell, J.C. Helton, R.L. Iman, D.E. Longsine, N.R. Ortiz, G.E. Runkle and M.J. Shortencarier. 1987. *Risk Methodology for Geologic Disposal of Radioactive Waste: Final Report. Rpt. SAND81-2573, NUREG/CR-2452*, Sandia National Laboratories, Albuquerque, NM.

42. Sprung J.L., D.C. Aldrich, D.J. Alpert, M.A. Cunningham and G.G. Weigand. Overview of the MELCOR Risk Code Development Program. *Proc. Proceedings of the International*

*Meeting on Light Water Reactor Severe Accident Evaluation*, *Cambridge, MA, August 28-September 1, 1983. Boston, MA: Stone and Webster Engineering Corporation*, *1983*:TS-10.11-11 to TS-10.11-18:

43. Iman R.L. and J.C. Helton. 1985. A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Rpt. NUREG/CR-3904, SAND84-1461*, Sandia National Laboratories, Albuquerque, NM.

44. Iman R.L. and J.C. Helton. 1988. An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Risk Analysis* 8:71-90.

45. U.S. NRC (U.S. Nuclear Regulatory Commission). 1990-1991. Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants. *Rpt. NUREG-1150, Vols. 1-3*, U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Division of Systems Research, Washington, DC.

46. Payne A.C., Jr., R.J. Breeding, J.C. Helton, L.N. Smith, J.D. Johnson, H.-N. Jow and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Peach Bottom Atomic Power Station. *Nuclear Engineering and Design* 135:61-94.

47. Gregory J.J., R.J. Breeding, J.C. Helton, W.B. Murfin, S.J. Higgins and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Sequoyah Nuclear Plant. *Nuclear Engineering and Design* 135:95-115.

48. Brown T.D., R.J. Breeding, J.C. Helton, H.-N. Jow, S.J. Higgins and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Grand Gulf Nuclear Station. *Nuclear Engineering and Design* 135:117-137.

49. Helton J.C., J.A. Rollstin, J.L. Sprung and J.D. Johnson. 1992. An Exploratory Sensitivity Study with the MACCS Reactor Accident Consequence Model. *Reliability Engineering and System Safety* 36:137-164.

50. Helton J.C., J.D. Johnson, A.W. Shiver and J.L. Sprung. 1995. Uncertainty and Sensitivity Analysis of Early Exposure Results with the MACCS Reactor Accident Consequence Model. *Reliability Engineering and System Safety* 48:91-127.

51. Helton J.C., J.D. Johnson, J.A. Rollstin, A.W. Shiver and J.L. Sprung. 1995. Uncertainty and Sensitivity Analysis of Food Pathway Results with the MACCS Reactor Accident Consequence Model. *Reliability Engineering & System Safety* 49:109-144.

52. Helton J.C., J.D. Johnson, J.A. Rollstin, A.W. Shiver and J.L. Sprung. 1995. Uncertainty and Sensitivity Analysis of Chronic Exposure Results with the MACCS Reactor Accident Consequence Model. *Reliability Engineering and System Safety* 50:137-177.

53. Helton J.C., J.D. Johnson, M.D. McKay, A.W. Shiver and J.L. Sprung. 1995. Robustness of an Uncertainty and Sensitivity Analysis of Early Exposure Results with the MACCS Reactor Accident Consequence Model. *Reliability Engineering and System Safety* 48:129-148.

54. Goossens L.H.J., F.T. Harper, B.C.P. Kraan and H. Metivier. 2000. Expert Judgement for a Probabilistic Accident Consequence Uncertainty Analysis. *Radiation Protection Dosimetry* 90:295-301.

55. Goossens L.H.J. and F.T. Harper. 1998. Joint EC/USNRC Expert Judgement Driven Radiological Protection Uncertainty Analysis. *Journal of Radiological Protection* 18:249-264.

56. Goossens L.H.J., J.D. Harrison, B.C.P. Kraan, R.M. Cooke, F.T. Harper and S.C. Hora. 1998. Probabilistic Accident Consequence Uncertainty Analysis: Uncertainty Assessment for Internal Dosimetry. *Rpt. NUREG/CR-6571, EUR 16773, SAND98-0119, Vols. 1 & 2*, U.S. Nuclear Regulatory Commission, Washington, D.C.

57. Goossens L.H.J., J. Boardman, B.C.P. Kraan, R.M. Cooke, J.A. Jones, F.T. Harper, M.L. Young and S.C. Hora. 1997. Probabilistic Accident Consequence Uncertainty Analysis: Uncertainty Assessment for Deposited Material and External Doses. *Rpt. NUREG/CR-6526, EUR 16772, SAND97-2323, Vols. 1 & 2*, U.S. Nuclear Regulatory Commission, Washington, D.C.

58. Haskin F.E., F.T. Harper, L.H.J. Goossens, B.C.P. Kraan and J.B. Grupa. 1997. Probabilistic Accident Consequence Uncertainty Analysis: Early Health Effects Uncertainty Assessment. *Rpt. NUREG/CR-6545, EUR 16775, SAND97-2689, Vols. 1 & 2*, U.S. Nuclear Regulatory Commission, Washington, D.C.

59. Little M.P., C.R. Muirhead, L.H.J. Goossens, B.C.P. Kraan, R.M. Cooke, F.T. Harper and S.C. Hora. 1997. Probabilistic Accident Consequence Uncertainty Analysis: Late Health Effects Uncertainty Assessment. *Rpt. NUREG/CR-6555, EUR 16774, SAND97-2322, Vols. 1 & 2*, U.S. Nuclear Regulatory Commission, Washington, D.C.

60. Harper F.T., L.H.J. Goossens, R.M. Cooke, S.C. Hora, M.L. Young, J. Päsler-Sauer, L.A. Miller, B. Kraan, C.H. Lui, M.D. McKay, J.C. Helton and J.A. Jones. 1995. Probabilistic Accident Consequence Uncertainty Analysis: Dispersion and Deposition Uncertainty Assessment. *Rpt. NUREG/CR-6244, EUR 15855EN, SAND94-1453, Vols. 1-3*, U.S. Nuclear Regulatory Commission, Washington, D.C.

61. Harper F.T., L.H.J. Goosens, R.M. Cooke, J.C. Helton, S.C. Hora, J.A. Jones, B. Kraan, C. Lui, M.D. McKay, L.A. Miller, J. Pasler-Sauer and M.L. Young. 1994. Joint USNRC/CEC Consequence Uncertainty Study: Summary of Objectives, Approach, Application, and Results for the Dispersion and Deposition Uncertainty Assessment. *Rpt. NUREG/CR-6244, SAND94-1453*, Sandia National Laboratories, Albuquerque.

62. Hora S.C. and R.L. Iman. 1989. Expert Opinion in Risk Analysis: The NUREG-1150 Methodology. *Nuclear Science and Engineering* 102:323-331.

63. Ortiz N.R., T.A. Wheeler, R.J. Breeding, S. Hora, M.A. Meyer and R.L. Keeney. 1991. Use of Expert Judgment in NUREG-1150. *Nuclear Engineering and Design* 126:313-331.

64. Harper F.T., R.J. Breeding, T.D. Brown, J.J. Gregory, A.C. Payne, E.D. Gorham and C.N. Amos. 1990. Evaluation of Severe Accident Risks: Quantification of Major Input Parameters, Expert Opinion Elicitation on In-Vessel Issues. *Rpt. NUREG/CR-4551, SAND86-1309, Vol. 2, Part 1, Rev. 1*, Sandia National Laboratories, Albuquerque, NM.

65. Harper F.T., A.C. Payne, R.J. Breeding, E.D. Gorham, T.D. Brown, G.S. Rightley, J.J. Gregory, W. Murfin and C.N. Amos. 1991. Evaluation of Severe Accident Risks: Quantification of Major Input Parameters, Experts' Determination of Containment Loads and

Molten Core Containment Interaction Issues. *Rpt. NUREG/CR-4551, SAND86-1309, Vol. 2, Part 2, Rev. 1*, Sandia National Laboratories, Albuquerque, NM.

66. Breeding R.J., C.N. Amos, T.D. Brown, E.D. Gorham, J.J. Gregory, F.T. Harper, W. Murfin and A.C. Payne. 1992. Evaluation of Severe Accident Risks: Quantification of Major Input Parameters, Experts' Determination of Structural Response Issues. *Rpt. NUREG/CR-4551, SAND86-1309, Vol. 2, Part 3, Rev. 1*, Sandia National Laboratories, Albuquerque, NM.

67. Harper F.T., R.J. Breeding, T.D. Brown, J.J. Gregory, H.-N. Jow, A.C. Payne, E.D. Gorham, C.N. Amos, J.C. Helton and G. Boyd. 1992. Evaluation of Severe Accident Risks: Quantification of Major Input Parameters, Experts' Determination of Source Term Issues. *Rpt. NUREG/CR-4551, SAND86-1309, Vol. 2, Part 4, Rev. 1*, Sandia National Laboratories, Albuquerque, NM.

68. Wheeler T.A., S.C. Hora, W.R. Cramond and S.D. Unwin. 1990. Analysis of Core Damage Frequency: Expert Judgment Elicitation. *Rpt. NUREG/CR-4550, SAND86-2084, Vol. 2, Rev. 1*, Sandia National Laboratories, Albuquerque, NM.

69. U.S. DOE (U.S. Department of Energy). 1996. *Title 40 CFR Part 191 Compliance Certification Application for the Waste Isolation Pilot Plant. Rpt. DOE/CAO-1996-2184, Vols. I-XXI*, U.S. Department of Energy, Carlsbad Area Office, Waste Isolation Pilot Plant, Carlsbad, NM.

70. Helton J.C., D.R. Anderson, H.-N. Jow, M.G. Marietta and G. Basabilvazo. 1999. Performance Assessment in Support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant. *Risk Analysis* 19:959 - 986.

71. Helton J.C. 1999. Uncertainty and Sensitivity Analysis in Performance Assessment for the Waste Isolation Pilot Plant. *Computer Physics Communications* 117:156-180.

72. Helton J.C., F.J. Davis and J.D. Johnson. 2000. Characterization of Stochastic Uncertainty in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:167-189.

73. Helton J.C., M.-A. Martell and M.S. Tierney. 2000. Characterization of Subjective Uncertainty in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:191-204.

74. Parry G.W. 1988. On the Meaning of Probability in Probabilistic Safety Assessment. *Reliability Engineering and System Safety* 23:309-314.

75. Paté-Cornell M.E. 1986. Probability and Uncertainty in Nuclear Safety Decisions. *Nuclear Engineering and Design* 93:319-327.

76. Apostolakis G.E. 1989. Uncertainty in Probabilistic Risk Assessment. *Nuclear Engineering and Design* 115:173-179.

77. Helton J.C., C.W. Hansen and C.J. Sallaberry. 2014. Expected Dose for the Early Failure Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:297-309.

78. Sallaberry C.J., C.W. Hansen and J.C. Helton. 2014. Expected Dose for the Igneous Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive

Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:339-353.

79. Helton J.C., M.B. Gross, C.W. Hansen, C.J. Sallaberry and S.D. Sevougian. 2014. Expected Dose for the Seismic Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:380-398.

80. Hansen C.W., G.A. Behie, A. Bier, K.M. Brooks, Y. Chen, J.C. Helton, S.P. Hommel, K.P. Lee, B. Lester, P.D. Mattie, S. Mehta, S.P. Miller, C.J. Sallaberry, S.D. Sevougian and P. Vo. 2014. Uncertainty and Sensitivity Analysis for the Nominal Scenario Class in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:272-296.

81. Helton J.C., J.D. Johnson, H.-N. Jow, R.D. McCurley and L.J. Rahal. 1998. Stochastic and Subjective Uncertainty in the Assessment of Radiation Exposure at the Waste Isolation Pilot Plant. *Human and Ecological Risk Assessment* 4:469-526.

82. SNL (Sandia National Laboratories). 2008. Total System Performance Assessment Model/Analysis for the License Application. *Rpt. MDL-WIS-PA-000005 Rev 00, AD 01*, U.S. Department of Energy Office of Civilian Radioactive Waste Management Las Vegas, NV.

83. Hansen C.W., J.T. Birkholzer, J. Blink, C.R. Bryan, Y. Chen, M.B. Gross, E. Hardin, J. Houseworth, R. Howard, R. Jarek, K.P. Lee, B. Lester, P. Mariner, P.D. Mattie, S. Mehta, F.V. Perry, B. Robinson, D. Sassani, S.D. Sevougian, J.S. Stein and M. Wasiolek. 2014. Overview of Total System Model Used for the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:249-266.

84. Thorne M.C. and M.M.R. Williams. 1992. A Review of Expert Judgement Techniques with Reference to Nuclear Safety. *Progress in Nuclear Safety* 27:83-254.

85. Budnitz R.J., G. Apostolakis, D.M. Boore, L.S. Cluff, K.J. Coppersmith, C.A. Cornell and P.A. Morris. 1998. Use of Technical Expert Panels: Applications to Probabilistic Seismic Hazard Analysis. *Risk Analysis* 18:463-469.

86. McKay M. and M. Meyer. 2000. Critique of and Limitations on the Use of Expert Judgements in Accident Consequence Uncertainty Analysis. *Radiation Protection Dosimetry* 90:325-330.

87. Meyer M.A. and J.M. Booker. 2001. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Philadelphia, PA: SIAM.

88. Ayyub B.M. 2001. *Elicitation of Expert Opinions for Uncertainty and Risks*, CRC Press, Boca Raton, FL.

89. Cooke R.M. and L.H.J. Goossens. 2004. Expert Judgement Elicitation for Risk Assessment of Critical Infrastructures. *Journal of Risk Research* 7:643-656.

90. Garthwaite P.H., J.B. Kadane and A. O'Hagan. 2005. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* 100:680-700.

91. Baudrit C. and D. Dubois. 2006. Practical Representations of Incomplete Probabilistic Knowledge. *Computational Statistics & Data Analysis* 51:86-108.

92. Dubois D. and H. Prade. 2006. Possibility Theory and Its Applications: A Retrospective and Prospective View. *CISM Courses and Lectures* 482:89-109.

93. Klir G.J. 2006. *Uncertainty and Information: Foundations of Generalized Information Theory*. New York, NY: Wiley-Interscience.

94. Ross T.J. 2004. *Fuzzy Logic with Engineering Applications*. New York, NY: Wiley.

95. Helton J.C., J.D. Johnson and W.L. Oberkampf. 2004. An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions. *Reliability Engineering and System Safety* 85:39-71.

96. Ross T.J., J.M. Booker and W.J. Parkinson (eds.). 2002. *Fuzzy Logic and Probability Applications: Bridging the Gap*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

97. Klir G.J. and M.J. Wierman. 1999. Uncertainty-Based Information, Physica-Verlag, New York, NY.

98. Bardossy G. and J. Fodor. 2004. Evaluation of Uncertainties and Risks in Geology, Springer-Verlag, New York, NY.

99. Helton J.C., J.D. Johnson, W.L. Oberkampf and C.J. Sallaberry. 2010. Representation of Analysis Results Involving Aleatory and Epistemic Uncertainty. *International Journal of General Systems* 39:605-646.

100. Helton J.C. and J.D. Johnson. 2011. Quantification of Margins and Uncertainties: Alternative Representations of Epistemic Uncertainty. *Reliability Engineering and System Safety* 96:1034-1052.

101. McKay M.D., R.J. Beckman and W.J. Conover. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21:239-245.

102. Helton J.C. and F.J. Davis. 2003. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Reliability Engineering and System Safety* 81:23-69.

103. Helton J.C., C.W. Hansen and C.J. Sallaberry. 2014. Expected Dose for the Nominal Scenario Class in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:267-271.

104. Helton J.C. and F.J. Davis. 2002. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Rpt. SAND2001-0417*, Sandia National Laboratories, Albuquerque, NM.

105. Iman R.L. Statistical Methods for Including Uncertainties Associated with the Geologic Isolation of Radioactive Waste Which Allow for a Comparison with Licensing Criteria. *Proc. Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste, Gatlinburg, TN, March 9-13, 1981*, *1982*:145-157: Washington, DC: U.S. Nuclear Regulatory Commission, Directorate of Technical Information and Document Control.

106. Helton J.C., C.W. Hansen and C.J. Sallaberry. 2014. Expected Dose and Associated Uncertainty and Sensitivity Analysis Results for All Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:421-435.

107. Hansen C.W., J.C. Helton and C.J. Sallaberry. 2012. Use of Replicated Latin Hypercube Sampling to Estimate Sampling Variance in Uncertainty and Sensitivity Analysis Results for the Geologic Disposal of Radioactive Waste. *Reliability Engineering and System Safety* 107:139-148.

108. Helton J.C. and C.J. Sallaberry. 2009. Computational Implementation of Sampling-Based Approaches to the Calculation of Expected Dose in Performance Assessments for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 94:699-721.

109. Sallaberry C.J., J.C. Helton and S.C. Hora. 2008. Extension of Latin Hypercube Samples with Correlated Variables. *Reliability Engineering and System Safety* 93:1047-1059.

110. Iman R.L. and W.J. Conover. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics: Simulation and Computation* B11:311-334.

111. Iman R.L. and J.M. Davenport. 1982. Rank Correlation Plots for Use with Correlated Input Variables. *Communications in Statistics: Simulation and Computation* B11:335-360.

112. Helton J.C., D.R. Anderson, G. Basabilvazo, H.-N. Jow and M.G. Marietta. 2000. Summary Discussion of the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:437-451.

113. Farmer F.R. 1967. Reactor Safety and Siting: A Proposed Risk Criterion. *Nuclear Safety* 8:539-548.

114. Cox D.C. and P. Baybutt. 1982. Limit Lines for Risk. *Nuclear Technology* 57:320-330.

115. Munera H.A. and G. Yadigaroglu. 1986. On Farmer's Line, Probability Density Functions, and Overall Risk. *Nuclear Technology* 74:229-232.

116. Helton J.C., M.B. Gross and C.J. Sallaberry. 2014. Representation of Aleatory Uncertainty Associated with the Seismic Ground Motion Scenario Class in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:399-405.

117. Hansen C.W., G.A. Behie, A. Bier, K.M. Brooks, Y. Chen, J.C. Helton, S.P. Hommel, K.P. Lee, B. Lester, P.D. Mattie, S. Mehta, S.P. Miller, C.J. Sallaberry, S.D. Sevougian and P. Vo. 2014. Uncertainty and Sensitivity Analysis for the Seismic Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:406-420.

118. Helton J.C., J.D. Johnson, C.J. Sallaberry and C.B. Storlie. 2006. Survey of Sampling-Based Methods for Uncertainty and Sensitivity Analysis. *Reliability Engineering and System Safety* 91:1175-1209.

119. Helton J.C. 1993. Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal. *Reliability Engineering and System Safety* 42:327-367.

120. Iman R.L., M.J. Shortencarier and J.D. Johnson. 1985. *A FORTRAN 77 Program and User's Guide for the Calculation of Partial Correlation and Standardized Regression Coefficients. Rpt. NUREG/CR-4122, SAND85-0044*, Sandia National Laboratories, Albuquerque, NM.

121. Iman R.L. and W.J. Conover. 1979. The Use of the Rank Transform in Regression. *Technometrics* 21:499-509.

122. Cooke R.M. and J.M. van Noortwijk. 2000. Graphical Methods. In *Sensitivity Analysis*, Eds. A. Saltelli, K. Chan, E.M. Scott. New York, NY: Wiley. pp. 245-264.

123. Kleijnen J.P.C. and J.C. Helton. 1999. Statistical Analyses of Scatterplots to Identify Important Factors in Large-Scale Simulations, 1: Review and Comparison of Techniques. *Reliability Engineering and System Safety* 65:147-185.

124. Storlie C.B. and J.C. Helton. 2008. Multiple Predictor Smoothing Methods for Sensitivity Analysis: Description of Techniques. *Reliability Engineering and System Safety* 93:28-54.

125. Storlie C.B. and J.C. Helton. 2008. Multiple Predictor Smoothing Methods for Sensitivity Analysis: Example Results. *Reliability Engineering and System Safety* 93:55-77.

126. Storlie C.B., L.P. Swiler, J.C. Helton and C.J. Sallaberry. 2009. Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models. *Reliability Engineering and System Safety* 94:1735-1763.

127. Storlie C.B., B.J. Reich, J.C. Helton, L.P. Swiler and C.J. Sallaberry. 2013. Analysis of Computationally Demanding Models with Continuous and Categorical Inputs. *Reliability Engineering and System Safety* 113:30-41.

128. Hora S.C. and J.C. Helton. 2003. A Distribution-Free Test for the Relationship Between Model Input and Output when Using Latin Hypercube Sampling. *Reliability Engineering and System Safety* 79:333-339.

129. Helton J.C., J.E. Bean, K. Economy, J.W. Garner, R.J. MacKinnon, J. Miller, J.D. Schreiber and P. Vaughn. 2000. Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Disturbed Conditions. *Reliability Engineering and System Safety* 69:263-304.

130. Hamby D.M. 1994. A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models. *Environmental Monitoring and Assessment* 32:135-154.

131. Frey H.C. and S.R. Patil. 2002. Identification and Review of Sensitivity Analysis Methods. *Risk Analysis* 22:553-578.

132. Ionescu-Bujor M. and D.G. Cacuci. 2004. A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems--I: Deterministic Methods. *Nuclear Science and Engineering* 147:189-2003.

133. Cacuci D.G. and M. Ionescu-Bujor. 2004. A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems--II: Statistical Methods. *Nuclear Science and Engineering* 147:204-217.

134. Cacuci D.G. 2003. *Sensitivity and Uncertainty Analysis, Vol. 1: Theory*. Boca Raton, FL: Chapman and Hall/CRC Press.

135. Cacuci D.G., M. Ionescu-Bujor and I.M. Navon. 2005. *Sensitivity and Uncertainty Analysis, Vol. II, Application to Large-Scale Systems*. Boca Raton, FL: Chapman & Hall/CRC.

136. Saltelli A., K. Chan and E.M. Scott (eds). 2000. *Sensitivity Analysis*. New York, NY: Wiley.

137. Saltelli A., M. Ratto, S. Tarantola and F. Campolongo. 2005. Sensitivity Analysis for Chemical Models. *Chemical Reviews* 105:2811-2828.

138. Matott L.S., J.E. Badendreier and S.T. Purucker. 2009. Evaluating Uncertainty in Integrated Environmental Models: A Review of Concepts and Tools. *Water Resources Research* 45:W06421.

139. Wei P., Z. Lu and S. Song. 2015. Variable Inportance Analysis: A Conprehensive Review. *Reliability Engineering and System Safety* 142:399-432.

140. Pianosi F., K. Beven, J. Freer, J.W. Hall, J. Rougier, D.B. Stephenson and T. Wagener. 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling and Software* 79:214-232.

141. Borgonovo E. and E. Plischke. 2016. Sensitivity Analysis: A Review of Recent Advances. *European Journal of Operational Research* 248:869-887.

142. Ghanem R., Higdon, D., Owhadi, H., ed. 2017. *Handbook of Uncertainty Quantification*: Springer International Publishing Switzerland.

143. Helton J.C., J.E. Bean, K. Economy, J.W. Garner, R.J. MacKinnon, J. Miller, J.D. Schreiber and P. Vaughn. 2000. Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Undisturbed Conditions. *Reliability Engineering and System Safety* 69:227-261.

144. Berglund J.W., J.W. Garner, J.C. Helton, J.D. Johnson and L.N. Smith. 2000. Direct Releases to the Surface and Associated Complementary Cumulative Distribution Functions in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Cuttings, Cavings and Spallings. *Reliability Engineering and System Safety* 69:305-330.

145. Stoelzel D.M., D.G. O'Brien, J.W. Garner, J.C. Helton, J.D. Johnson and L.N. Smith. 2000. Direct Releases to the Surface and Associated Complementary Cummulative Distribution Functions in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Direct Brine Release. *Reliability Engineering and System Safety* 69:343-367.

146. Stockman C.T., J.W. Garner, J.C. Helton, J.D. Johnson, A. Shinta and L.N. Smith. 2000. Radionuclide Transport in the Vicinity of the Repository and Associated Complementary Cumulative Distribution Functions in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:370-396.

147. Ramsey J.L., R. Blaine, J.W. Garner, J.C. Helton, J.D. Johnson, L.N. Smith and M. Wallace. 2000. Radionuclide and Colloid Transport in the Culebra Dolomite and Associated

Complementary Cumulative Distribution Functions in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:397-420.

148. Hansen C.W., G.A. Behie, A. Bier, K.M. Brooks, Y. Chen, J.C. Helton, S.P. Hommel, K.P. Lee, B. Lester, P.D. Mattie, S. Mehta, S.P. Miller, C.J. Sallaberry, S.D. Sevougian and P. Vo. 2014. Uncertainty and Sensitivity Analysis for the Early Failure Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:310-338.
149. Sallaberry C.J., G.A. Behie, A. Bier, K.M. Brooks, Y. Chen, C.W. Hansen, J.C. Helton, S.P. Hommel, K.P. Lee, B. Lester, P.D. Mattie, S. Mehta, S.P. Miller, S.D. Sevougian and P. Vo. 2014. Uncertainty and Sensitivity Analysis for the Igneous Scenario Classes in the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:354-379.
150. Rechard R.P. and M.S. Tierney. 2005. Assignment of Probability Distributions for Parameters in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. Part 1: Description of Process. *Reliability Engineering and System Safety* 88:1-32.
151. Rechard R.P. and M.S. Tierney. 2005. Assignment of Probability Distributions for Parameters in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. Part 2. Application of Process. *Reliability Engineering and System Safety* 88:33-80.

# 4. SAMPLING-BASED PROCEDURES FOR UNCERTAINTY AND SENSITIVITY ANALYSIS

## 4.1 Introduction

This chapter provides an introduction to sampling-based procedures for uncertainty analysis and sensitivity analysis in analyses of complex systems, where (i) uncertainty analysis involves the determination of the uncertainty in analysis results that derives from uncertainty in analysis inputs, and (ii) sensitivity analysis involves the determination of the contributions of individual uncertain analysis inputs to the uncertainty in analysis results. The indicated uncertainty derives from a lack of knowledge about the appropriate value to use for quantities that are assumed to have fixed values in the context of a particular analysis and is usually referred to as epistemic uncertainty. In the conceptual and computational organization of an analysis for a complex system, epistemic uncertainty is usually considered to be distinct from aleatory uncertainty, which arises from an assumed inherent randomness in the behavior of the system under study [1-12].

Sampling-based procedures for uncertainty and sensitivity analysis are both effective and widely used [13-20] and involve the generation and exploration of a mapping from uncertain analysis inputs to uncertain analysis results. The underlying idea is that analysis results $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$ are functions of uncertain analysis inputs $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$. In turn, uncertainty in $\mathbf{x}$ results in a corresponding uncertainty in $\mathbf{y}(\mathbf{x})$. This leads to two questions: (i) "What is the uncertainty in $\mathbf{y}(\mathbf{x})$ given the uncertainty in $\mathbf{x}$?", and (ii) "How important are the individual elements of $\mathbf{x}$ with respect to the uncertainty in $\mathbf{y}(\mathbf{x})$?". The goal of uncertainty analysis is to answer the first question, and the goal of sensitivity analysis is to answer the second question. In practice, the implementation of an uncertainty analysis and the implementation of a sensitivity analysis are closely connected on both a conceptual and a computational level.

The following sections summarize the five basic components that underlie the implementation of a sampling-based uncertainty and sensitivity analysis: (i) Definition of distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the epistemic uncertainty in the components $x_1, x_2, \ldots, x_{nX}$ of $\mathbf{x}$ (Sect. 4.2), (ii) Generation of a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{nS}$ from the $\mathbf{x}$'s consistent with the distributions $D_1, D_2, \ldots, D_{nX}$ (Sect. 4.3), (iii) Propagation of the sample through the analysis to produce a mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results (Sect. 4.4), (iv) Presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of $\mathbf{y}$ constructed from the corresponding elements of $\mathbf{y}(\mathbf{x}_i)$, $i = 1, 2, \ldots, nS$) (Sect. 4.5), and (v) Determination of sensitivity analysis results (i.e., exploration of the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$) (Sect. 4.6). The presentation then ends with several concluding observations (Sect. 4.7).

This chapter is based on material contained in a previously published technical report [21]. The content of this technical report has also formed the basis for three invited conference presentations and corresponding publications in the associated conference proceedings [22-24]. The content of Ref. [23] is effectively the same as the content of Ref. [21].

## 4.2 Characterization of Uncertainty

Definition of the distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the epistemic uncertainty in the components $x_1, x_2, \ldots, x_{nX}$ of **x** is the most important part of a sampling-based uncertainty and sensitivity analysis as these distributions determine both the uncertainty in **y** and the sensitivity of the elements of **y** to the elements of **x**. The distributions $D_1, D_2, \ldots, D_{nX}$ are typically defined through an expert review process [25-32], and their development can constitute a major analysis cost. A possible analysis strategy is to perform an initial exploratory analysis with rather crude definitions for $D_1, D_2, \ldots, D_{nX}$ and use sensitivity analysis to identify the most important analysis inputs; then, resources can be concentrated on characterizing the uncertainty in these inputs, and a second presentation or decision-aiding analysis can be carried out with these improved uncertainty characterizations.

The scope of an expert review process can vary widely depending on the purpose of the analysis, the size of the analysis, and the resources available to carry out the analysis. At one extreme is a relatively small study in which a single analyst both develops the uncertainty characterizations (e.g., on the basis of personal knowledge or a cursory literature review) and performs the associated uncertainty and sensitivity analyses. At the other extreme is a large analysis on which important societal decisions will be based and for which uncertainty characterizations are carried out for a large number of variables by teams of outside experts who support the analysts actually performing the analysis.

Given the breadth of analysis possibilities, it is beyond the scope of this presentation to provide an exhaustive review of how the distributions $D_1, D_2, \ldots, D_{nX}$ might be developed. However, as general guidance, it is best to avoid trying to obtain these distributions by specifying the defining parameters (e.g., mean and standard deviation) for a specific distribution type. Rather, distributions can be defined by specifying selected quantiles (e.g., 0.0, 0.1, 0.25, …, 0.9, 1.0) of the corresponding cumulative distribution function (CDF), which should keep the individual supplying the information in closer contact with the original sources of information or insight than is the case when a particular named distribution is specified (Figure 4.1a). Distributions from multiple experts can be aggregated by averaging (Figure 4.1b).

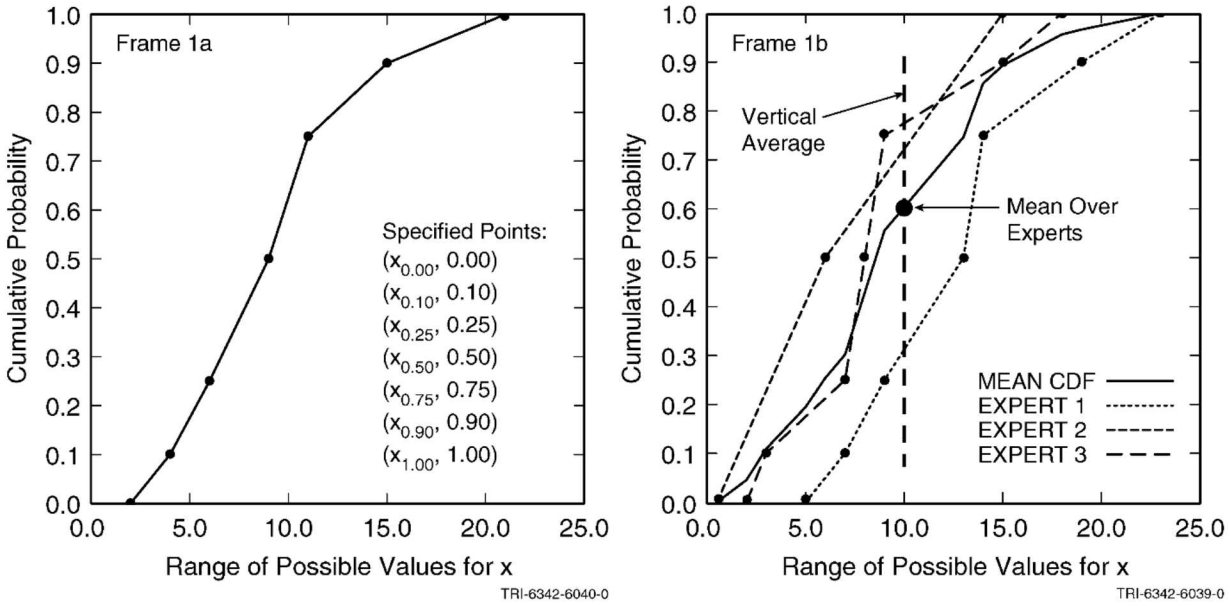



Figure 4.1 Characterization of epistemic uncertainty: (a) Construction of CDF from specified quantile values (Figure 4.1, Ref. [33]), and (b) Construction of mean CDF by vertical averaging of CDFs defined by individual experts with equal weight (i.e., $1/nE = 1/3$, where $nE = 3$ is the number of experts) given to each expert (Figure 4.2, Ref. [33]).

This presentation draws its examples from an uncertainty and sensitivity analysis carried out for a two-phase flow model implemented with the BRAGFLO program [34-36] in support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant [37; 38]. A subset of the 31 uncertain variables considered in the example results (i.e., $x_1, x_2, \ldots, x_{nX}$ with $nX = 31$) and their associated distributions (i.e., $D_1, D_2, \ldots, D_{31}$) is summarized in Table 4.1. A full listing of the indicated 31 variables is available in Table 1 of Ref. [39]. Additional information on the use of these variables in the two phase flow model and on the development of the associated uncertainty distributions is available in the original analysis documentation [34; 40].

Additional information: Sect. 6.2, Ref. [41]; Refs. [25-28; 30; 31; 42-54]. As an example, Ref. [25] describes the approach used in the extensive expert review process that supported the U.S. Nuclear Regulatory Commission's (NRC's) reassessment of the risk from commercial nuclear power plants (i.e., NUREG-1150; see Refs. [55-60]).

Table 4.1 Definition of the 15 Out of a Total of 31 Independent (i.e., Sampled) Variables that Are Identified in the Example Sensitivity Analyses for the Two-Phase Flow Model Presented in Sect. 4.6 (adapted from Table 1, Ref. [39]).

| |
|---|
| *ANHBCEXP* – Brooks-Corey pore distribution parameter for anhydrite (dimensionless). Distribution: Student's with 5 degree of freedom. Range: 0.491 – 0.842. Mean, median: 0.644, 0.644. |
| *ANHBCVGP* – Pointer variable for selection of relative permeability model for use in anhydrite. Distribution: Discrete with 60% 0, 40% 1. Value of 0 implies Brooks-Corey model; value of 1 implies van Genuchten-Parker model. |
| *ANHPRM*– Logarithm of anhydrite permeability ($m^2$). Distribution: Student's with 5 degrees of freedom. Range: –21.0 to –17.1 (i.e., permeability range is $1 \times 10^{-21}$ to $1 \times 10^{-17.1}$ $m^2$). Mean, median: –18.9, –18.9. Correlation: –0.99 rank correlation with *ANHCOMP*. |
| *ANRGSSAT* – Residual gas saturation in anhydrite (dimensionless). Distribution: Student's with 5 degrees of freedom. Range $1.39 \times 10^{-2}$ to $1.79 \times 10^{-1}$. Mean, median: $7.71 \times 10^{-2}$, $7.71 \times 10^{-2}$. |
| *BHPRM* – Logarithm of borehole permeability ($m^2$). Distribution: Uniform. Range: –14 to –11 (i.e., permeability range is $1 \times 10^{-14}$ to $1 \times 10^{-11}$ $m^2$). Mean, median: –12.5, –12.5. |
| *HALPOR* – Halite porosity (dimensionless). Distribution: Piecewise uniform. Range: $1.0 \times 10^{-3}$ to $3 \times 10^{-2}$. Mean, median: $1.28 \times 10^{-2}$, $1.00 \times 10^{-2}$. |
| *HALPRM* – Logarithm of halite permeability ($m^2$). Distribution: Uniform. Range: –24 to –21 (i.e., permeability range is $1 \times 10^{-24}$ to $1 \times 10^{-21}$ $m^2$). Mean, median: –22.5, –22.5. Correlation: –0.99 rank correlation with *HALCOMP*. |
| *SALPRES* – Initial brine pressure, without the repository being present, at a reference point located in the center of the combined shafts at the elevation of the midpoint of MB 139 (Pa). Distribution: Uniform. Range: $1.104 \times 10^7$ to $1.389 \times 10^7$ Pa., Mean, median: $1.247 \times 10^7$ Pa, $1.247 \times 10^7$ Pa. |
| *SHPRMDRZ* – Logarithm of permeability ($m^2$) of DRZ surrounding shaft. Distribution: Triangular. Range: –17.0 to –14.0 (i.e., permeability range is $1 \times 10^{-17}$ to $1 \times 10^{-14}$ $m^2$). Mean, mode: –15.3, –15.0. |
| *SHRBRSAT* – Residual brine saturation in shaft (dimensionless). Distribution: Uniform. Range: 0 to 0.4. Mean, median: 0.2, 0.2. |
| *SHRGSSAT* – Residual gas saturation in shaft (dimensionless). Distribution: Uniform. Range: 0 to 0.4. Mean, median: 0.2, 0.2. |
| *WASTWICK* – Increase in brine saturation of waste owing to capillary forces (dimensionless). Distribution: Uniform: Range: 0 – 1. Mean, median: 0.5, 0.5. |
| *WGRCOR* – Corrosion rate for steel under inundated conditions in the absence of $CO_2$ (m/s). Distribution: Uniform. Range: 0 to$1.58 \times 10^{-14}$ m/s. Mean, median: $7.94 \times 10^{-15}$ m/s, $7.94 \times 10^{-15}$ m/s. |

Table 4.1 Continued

| |
|---|
| *WGRMICI* – Microbial degradation rate for cellulose under inundated conditions (mol/kg s). Distribution: Uniform. Range: $3.17 \times 10^{-10}$ to $9.51 \times 10^{-9}$ mol/kg s. Mean, median: $4.92 \times 10^{-9}$ mol/kg s, $4.92 \times 10^{-9}$ mol/kg s. |
| *WMICDFLG* – Pointer variable for microbial degradation of cellulose. Distribution: Discrete, with 50% 0, 25% 1, 25% 2, *WMICDFLG* = 0, 1, 2, implies no microbial degradation of cellulose, microbial degradation of only cellulose, microbial degradation of cellulose, plastic and rubber. |

## 4.3 Generation of Sample

Several sampling strategies are available, including random sampling (Sect. 7.5.1, [61];[62]), quasi random sampling [63-66], importance sampling (Chap. 4, [67] ; Chap. 3, [68] ; Chap. 6, [69]), and Latin hypercube sampling [70; 71]. Latin hypercube sampling is very popular for use in analyses with computationally demanding models because its efficient stratification properties facilitate the extraction of a large amount of uncertainty and sensitivity information with a relatively small sample size [39; 72; 73].

Latin hypercube sampling operates in the following manner to generate a sample of size *nS* from the distributions $D_1, D_2, \ldots, D_{nX}$ associated with the elements of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$. The range of each $x_j$ is exhaustively divided into *nS* disjoint intervals of equal probability and one value $x_{ij}$ is randomly selected from each interval. The *nS* values for $x_1$ are randomly paired without replacement with the *nS* values for $x_2$ to produce *nS* pairs. These pairs are then randomly combined without replacement with the *nS* values for $x_3$ to produce *nS* triples. This process is continued until a set of *nS* *nX*-tuples $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,nX}]$, $i = 1, 2, \ldots, nS$, is obtained, with this set constituting the Latin hypercube sample (LHS) (Figure 4.2).

Latin hypercube sampling is a good choice for a sampling procedure when computationally demanding models are being studied. The popularity of Latin hypercube sampling has led to the original article being designated a *Technometrics* classic in experimental design [74]. When the model is not computationally demanding, many model evaluations can be performed and random sampling works as well as Latin hypercube sampling.

Control of correlations is an important aspect of sample generation. Specifically, correlated variables should have correlations close to their specified values, and uncorrelated variables should have correlations close to zero. In general, the imposition of complex correlation structures is not easy. However, Iman and Conover have developed a broadly applicable procedure to impose rank correlations on sampled values that (i) is distribution free (i.e., does not depend on the assumed marginal distributions for the sampled variables), (ii) can impose complex correlation structures involving multiple variables, (iii) works with both random and Latin hypercube sampling, and (iv) preserves the intervals used in Latin hypercube sampling [75; 76]. Details on the implementation of the procedure are available in the original reference [75]; illustrative results are provided in Figure 4.3.

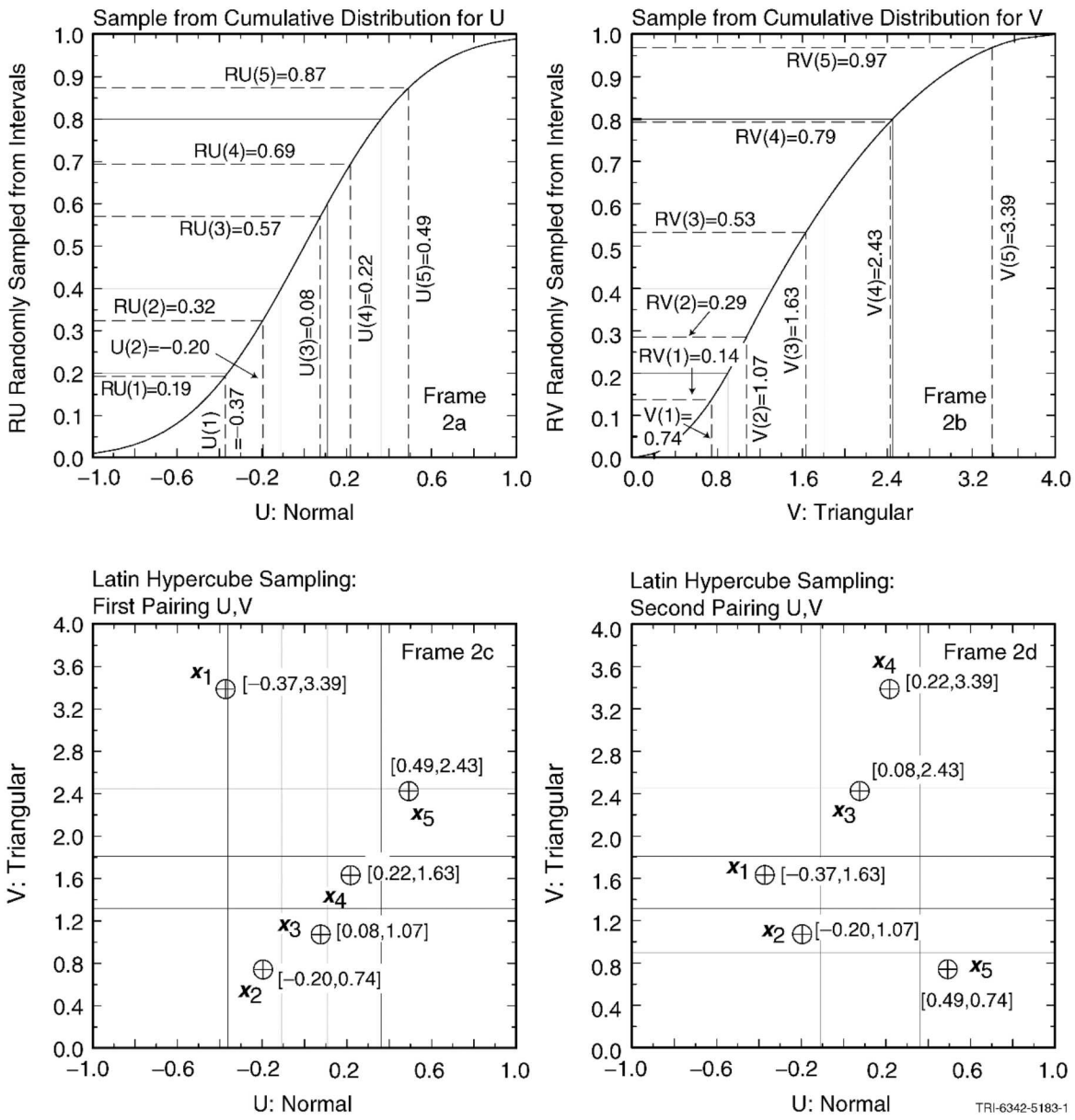



Figure 4.2 Example of Latin hypercube sampling to generate a sample of size $nS = 5$ from $\mathbf{x} = [U, V]$ with $U$ normal on [–1, 1] (mean = 0.0; 0.01 quantile = –1; 0.99 quantile = 1) and $V$ triangular on [0, 4] (mode = 1): (a) Upper frames illustrate sampling of values for $U$ and $V$, and (b) Lower frames illustrate two different pairings of the sampled values of $U$ and $V$ in the construction of a Latin hypercube sample (Figure 5.3, Ref. [33]).

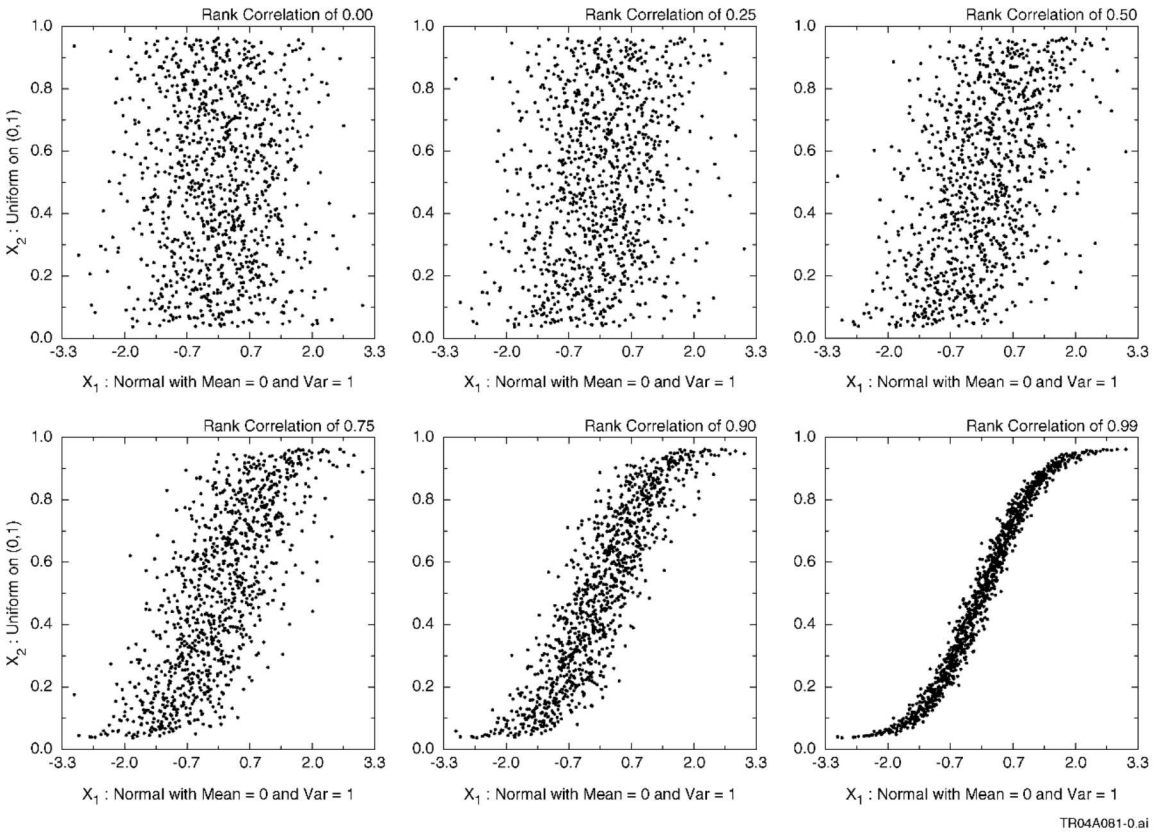



Figure 4.3 Examples of rank correlations of 0.00, 0.25, 0.50, 0.75, 0.90 and 0.99 imposed with the Iman/Conover restricted pairing technique for an LHS of size $nS = 1000$ (Figure 5.1, Ref. [77]).

Unlike simple random sampling, the size of an LHS cannot be increased by simply adding one sample element at a time. However, it is possible to retain the elements of an initial LHS in an expanded LHS [78-80]. This can be important in a computationally demanding analysis in which it is desired both to increase the size of an LHS and also to retain already performed calculations in the analysis. Further, the stability of results obtained with Latin hypercube sampling for a given sample size can be assessed with a replicated sampling technique developed by R.L. Iman [40; 81].

The example analysis involving the model for two-phase fluid flow used three independently generated (i.e., replicated) LHSs of size $nS = 100$ each from the 31 epistemically uncertain variables under consideration. The purpose of the replication was to provide a basis for testing the stability of uncertainty and sensitivity analysis results obtained with Latin hypercube sampling (Sects. 7, 8, Ref. [40]). The Iman/Conover restricted pairing technique indicated in the preceding paragraph was used to control correlations within the individual samples. The analyses with the three replicated samples were sufficiently similar that each analysis would have independently lead to the same insights with respect to model behavior [39]. However, to make full use of all model evaluations, final presentation results [35; 36] were calculated with the three replicated samples pooled together to produce a single sample of size $nS = 300$.

Additional information: Sect. 6.3, Ref. [41]; Refs. [70; 71; 82-84].

## 4.4 Propagation of Sample through the Analysis

Propagation of the sample through the analysis to produce the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results is often the most computationally demanding part of a sampling-based uncertainty and sensitivity analysis. The details of this propagation are analysis specific and can range from very simple for analyses that involve a single model to very complicated for large analyses that involve complex systems of linked models [38; 55].

When a single model is under consideration, this part of the analysis can involve little more than putting a DO loop around the model that (i) supplies the sampled input to the model, (ii) runs the model, and (iii) stores model results for later analysis. When more complex analyses with multiple models are involved, considerable sophistication may be required in this part of the analysis. Implementation of such analyses can involve (i) development of simplified models to approximate more complex models, (ii) clustering of results at model interfaces, (ii) reuse of model results through interpolation or linearity properties, and (iv) complex procedures for the storage and retrieval of analysis results.

As is typical of many large analyses, propagation of the three replicated LHSs through the two-phase flow model produced a large number of time-dependent results. The subset of these results used for illustration of uncertainty and sensitivity analysis procedures in this presentation is summarized in Table 4.2.

Additional information: Examples of complex analyses that have used Latin hypercube sampling in the propagation of epistemic uncertainty are provided by (i) the NUREG-1150 analyses for five nuclear power plants [25; 55-60], (ii) the analyses carried out in support of the Compliance Certification Application for the Waste Isolation Pilot Plant [38], and (iii) the analyses carried out in support of the licensing application for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada [85].

Table 4.2 Definition of Dependent Variables Calculated by BRAGFLO Program for Two Phase Flow and Used in the Illustration of Uncertainty and Sensitivity Analysis Procedures.

| |
|---|
| *BNBHDNUZ* – Cumulative brine flow ($m^3$) down borehole at Market Bed (MB) 138 (i.e., from cell 223 to cell 575 in Figure 3, Ref. [34]). |
| *BRAABNIC* – Cumulative brine flow ($m^3$) out of north anhydrites A and B into disturbed rock zone (DRZ) (i.e., from cell 556 to cell 527 in Figure 3, Ref. [34]). |
| *BRAABSIC* – Cumulative brine flow ($m^3$) out of south anhydrites A and B into DRZ (i.e., from cell 555 to cell 482 in Figure 3, Ref. [34]). |
| *BRAALIC* – Cumulative brine flow ($m^3$) out of all MBs into DRZ (i.e., *BRAALIC* = *BRM38NIC* + *BRAABNIC* + *BRM39NIC* + *BRM38SIC* + *BRAABSIC* + *BRM39SIC*). |
| *BRM38NIC* – Cumulative brine flow ($m^3$) out of north MB138 into DRZ (i.e., from cell 588 to cell 587 in Figure 3, Ref. [34]). |
| *BRM38SIC* – Cumulative brine flow ($m^3$) out of south MB138 into DRZ (i.e., from cell 571 to cell 572 in Figure 3, Ref. [34]). |
| *BRM39NIC* – Cumulative brine flow ($m^3$) out of north MB139 to DRZ (i.e., from cell 540 to cell 465 in Figure 3, Ref. [34]). |
| *BRM39SIC* – Cumulative brine flow ($m^3$) out of south MB139 into DRZ (i.e., from cell 539 to cell 436 in Figure 3, Ref. [34]). |
| *BRNREPTC* – Cumulative brine flow ($m^3$) into repository (i.e., into regions corresponding to cells 596 – 625, 638 – 640 in Figure 3, Ref. [34]). |
| *REP_SATB* – Brine saturation in upper waste panels (i.e., average brine saturation calculated over cells 617 – 625 in Figure 3, Ref. [34]). |
| *WAS_PRES* – Pressure (Pa) in lower waste panel (i.e., average pressure calculated over cells 596 – 616 in Figure 3, Ref. [34]). |
| *WAS_SATB* – Brine saturation in lower waste panel (i.e., average brine saturation calculated over cells 596 – 616 in Figure 3, Ref. [34]) |
| Notation: The designator E0 is used to indicate results calculated for undisturbed conditions, and the designator E2 is used to indicate results calculated for disturbed conditions due to a drilling intrusion that penetrates the lower waste panel of the repository 1000 yr after repository closure. Further, the designator R1 indicates results calculated for the first of the three replicated Latin hypercube samples indicated in Sect. 4.3, and the designators R1, R2, R3 collectively are used to indicate results calculated with the three replicates pooled together. |

## 4.5 Presentation of Uncertainty Analysis Results

Presentation of uncertainty analysis results is generally straight forward and involves little more than displaying the results associated with the already calculated mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$. Presentation possibilities include means and standard deviations, density functions, cumulative distribution function (CDFs), complementary cumulative distribution functions (CCDFs), and box plots [71]. Presentation formats such as CDFs (Figure 4.1a), CCDFs (Figure 4.4a) and box plots (Figure 4.4b) are usually preferable to means and standard deviations because of the large amount of uncertainty information that is lost in the calculation of means and standard deviations. Owing to their flattened shape, box plots are particularly useful when it is desired to the display and compare the uncertainty in a number of related variables.

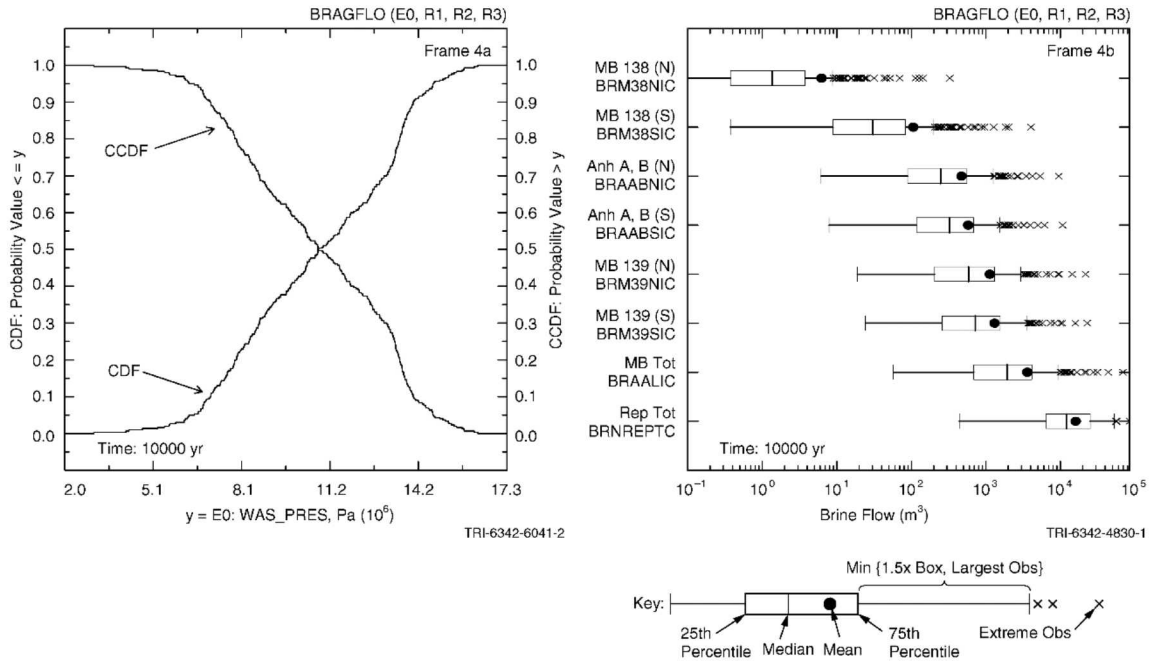



Figure 4.4 Representation of uncertainty in scalar-valued analysis results: (a) CDFs and CCDFs (Figure 7.2, Ref. [33]) and (b) box plots (Figure 7.4, Ref. [33]).

The representational challenge is more complex when the analysis outcome of interest is a function rather than a scalar. For example, a system property that is a function of time is a common analysis outcome. As another example, a CCDF that summarizes the effects of aleatory uncertainty is a standard analysis outcome in risk assessments. An effective display format for such analysis outcomes is to use two plot frames, with first frame displaying the analysis results for the individual sample elements and the second frame displaying summary results for the outcomes in the first frame (e.g., quantiles and means) (Figure 4.5).

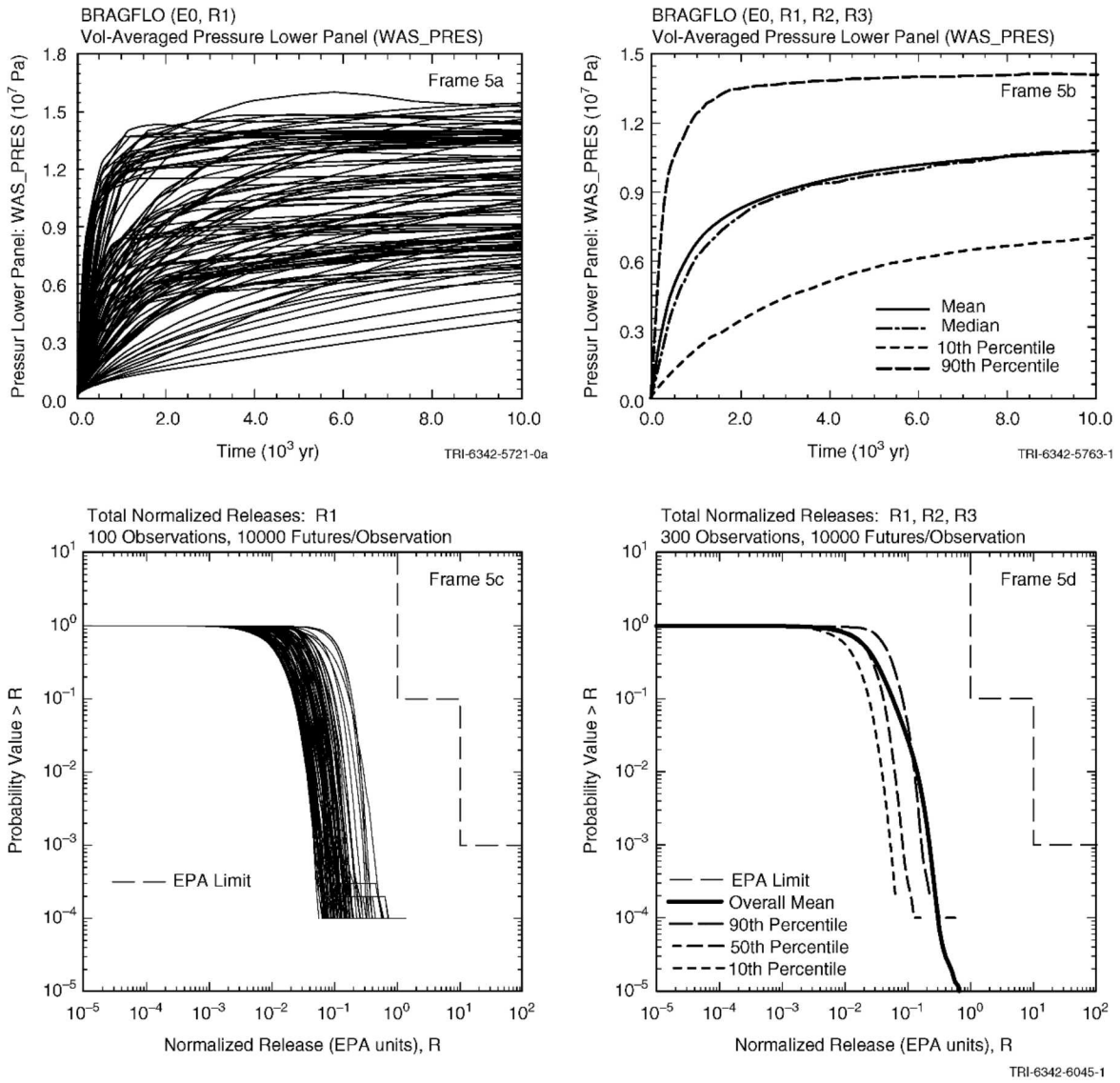



Figure 4.5 Representation of uncertainty in analysis results that are functions: (a, b) Pressure as a function of time (Figs. 7.5, 7.9, Ref. [33]), and (c, d) Effects of aleatory uncertainty summarized as a CCDF (Figure 10.5, Ref. [33]).

## 4.6 Determination of Sensitivity Analysis Results

Determination of sensitivity analysis results is usually more demanding than the presentation of uncertainty analysis results due to the need to actually explore the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, to assess the effects of individual components of $\mathbf{x}$ on the components of $\mathbf{y}$. A number of approaches to sensitivity analysis that can be used in conjunction with a sampling-based uncertainty analysis are listed and briefly summarized below. In this summary, (i) $x_j$ is an element of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$, (ii) $y_k$ is an element of $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$, (iii) $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,nX}]$, $i = 1, 2, \ldots, nS$, is a random or Latin hypercube sample from the possible values for $\mathbf{x}$ generated in consistency with the joint distribution assigned to the $x_j$, (iv) $\mathbf{y}_i = \mathbf{y}(\mathbf{x}_i)$ for $i = 1, 2, \ldots, nS$, and (v) $x_{ij}$ and $y_{ik}$ are elements of $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively.

### 4.6.1 Scatterplots

Scatterplots are plots of the points $[x_{ij}, y_{ik}]$ for $i = 1, 2, \ldots, nS$ and can reveal nonlinear or other unexpected relationships (Figure 4.6). In many analyses, scatterplots provide all the information that is needed to understand the sensitivity of analysis results to the uncertainty in analysis inputs. Further, scatterplots constitute a natural starting point in a complex analysis that can help in the development of a sensitivity analysis strategy using one or more additional techniques.

Additional information: Sect. 6.6.1, Ref. [41]; Sect. 6.1, Ref. [23].

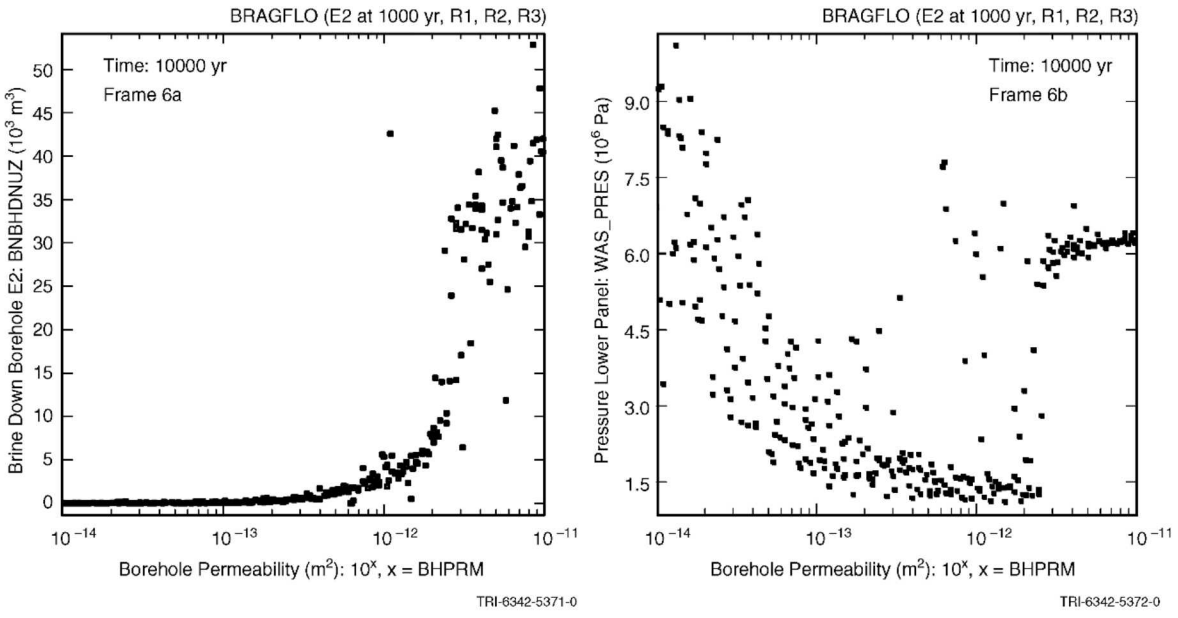



Figure 4.6 Examples of scatterplots obtained in a sampling-based uncertainty/sensitivity analysis (Figs. 8.1, 8.2, Ref. [33]).

### 4.6.2 Cobweb Plots

Cobweb plots are plots of the points $[\mathbf{x}_i, y_{ik}]=[x_{i1}, x_{i2}, \ldots, x_{i,nX}, y_{ik}]$ for $i$ = 1, 2, …, $nS$ and provide a two-dimensional representation for $[\mathbf{x}_i, y_{ik}]$, which is a $nX + 1$ dimensional quantity. Specifically, values for the $y_{ik}$ and also for the elements $x_{ij}$ of $\mathbf{x}_i$ appear on the ordinate of a cobweb plot and the variables themselves are designated by fixed locations on the abscissa. Then, the values $y_{ik}$, $i$ = 1, 2, .., $nS$, for $y_k$ and the values $x_{ij}$, $i$ = 1, 2, …, $nS$, for each $x_j$ are plotted above the locations for $y_k$ and $x_j$ on the abscissa and each $nX + 1$ dimensional point $[\mathbf{x}_i, y_{ik}]$ is represented by a line connecting the values for the individual components of $[\mathbf{x}_i, y_{ik}]$. When $y_k$ and the $x_j$'s have significantly different ranges, their values will have to rank transformed in order to produce a practical and readable plot. Cobweb plots provide more information in a single plot frame than a scatterplot but are harder to read.

Additional information: Sect. 11.7, Ref. [86].

### 4.6.3 Regression Analysis

Regression analysis provides an algebraic representation of the relationships between $y_k$ and one or more $x_j$'s. Unless stated otherwise, regression analysis is usually assumed to involve the construction of linear models of the form

$$\hat{y} = b_0 + \sum_{j=1}^{nX} b_j x_j \tag{4.6.1}$$

for multiple independent variables (i.e., $x_{i1}, x_{i2}, \ldots, x_{i,nX}$). The regression coefficients in Eq. (4.6.1) are determined such that the sum

$$\sum_{i=1}^{nS} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{nS} \left[ y_i - \left( b_0 + \sum_{j=1}^{nS} b_j x_{ij} \right) \right]^2 \tag{4.6.2}$$

is minimized. As a result, such regression models in are often referred to as least squares models due to the minimization of the sums of squares in Eq. (4.6.2).

An important property of least squares regression models is the equality

$$\sum_{i=1}^{nS} (y_i - \overline{y})^2 = \sum_{i=1}^{nS} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{nS} (\hat{y}_i - y_i)^2 \text{ with } \overline{y} = \sum_{i=1}^{nS} y_i / nS. \tag{4.6.3}$$

For notational convenience, the preceding equality is often written

$$SS_{\text{tot}} = SS_{\text{reg}} + SS_{\text{res}} \tag{4.6.4}$$

where

$$SS_{\text{tot}} = \sum_{i=1}^{nS} (y_i - \bar{y})^2, SS_{\text{reg}} = \sum_{i=1}^{nS} (\hat{y}_i - \bar{y})^2, SS_{\text{res}} = \sum_{i=1}^{nS} (\hat{y}_i - y_i)^2 \quad (4.6.5)$$

and the three preceding summations are called the total sum of squares ($SS_{\text{tot}}$), regression sum of squares ($SS_{\text{reg}}$) and residual sum of squares ($SS_{\text{res}}$), respectively.

The ratio

$$R^2 = SS_{reg} / SS_{tot} = \sum_{i=1}^{nS} (\hat{y}_i - \bar{y})^2 \Big/ \sum_{i=1}^{nS} (y_i - \bar{y})^2 \quad (4.6.6)$$

provides a measure of the extent to which the regression model can match the observed data. Specifically, when the variation about the regression model is small (i.e., $SS_{\text{res}}$ is small relative to $SS_{\text{reg}}$), then the corresponding $R^2$ value is close to 1, which indicates that the regression model is accounting for most of the uncertainty in *y*. Conversely, an $R^2$ value close to 0 indicates that the regression model is not very successful in accounting for the uncertainty in *y*. When the individual $x_j$ in the regression model in Eq. (4.6.1) are linearly independent, the $R^2$ value for the regression model can be expressed as

$$R^2 = SS_{reg} / SS_{tot} = R_1^2 + R_2^2 + \ldots + R_{nX}^2, \quad (4.6.7)$$

where $R_j^2$ is the $R^2$ value that results from regressing *y* on only $x_j$. Thus, $R_j^2$ is equal to the contribution of $x_j$ to the $R^2$ value for the regression model in Eq. (4.6.1) when the $x_j$'s are linearly independent.

The regression coefficients $b_j$, $j$ = 1, 2, …, *nX*, are not very useful in sensitivity analysis because each $b_j$ is influenced by the units in which $x_j$ is expressed and also does not incorporate any information on the distribution assigned to $x_j$. Because of this, the regression model in Eq. (4.6.1) is usually reformulated as

$$(\hat{y} - \bar{y})/\hat{s} = \sum_{j=1}^{nX} (b_j \hat{s}_j / \hat{s})(x_j - \bar{x}_j)/\hat{s}_j, \quad (4.6.8)$$

where

$$\hat{s} = \left[ \sum_{i=1}^{nS} (y_i - \bar{y})^2 / (nS - 1) \right]^{1/2}, \hat{s}_j = \left[ \sum_{i=1}^{nS} (x_{ij} - \bar{x}_j)^2 / (nS - 1) \right]^{1/2}, \bar{x}_j = \sum_{i=1}^{nS} x_{ij} / nS. \quad (4.6.9)$$

The coefficients $b_j\hat{s}_j / \hat{s}$ in Eq. (4.6.8) are referred to as standardized regression coefficients (SRCs).

The SRC $b_j\hat{s}_j / \hat{s}$ provides a measure of variable importance based on the effect on $y$ relative to the standard deviation $\hat{s}$ of $y$ of moving $x_j$ away from its expected value $\overline{x}_j$ by a fixed fraction of its standard deviation $\hat{s}_j$. Further, when the $x_j$'s are linearly independent, the inclusion or exclusion of an individual $x_j$ from the regression models in Eqs. (4.6.1) and (4.6.8) has no effect on the SRCs for the remaining variables in the model. Thus, as long as the $x_j$'s are linearly independent, the SRCs $b_j\hat{s}_j / \hat{s}$ in Eq. (4.6.8) provide a useful measure of variable importance, with (i) the absolute values of the coefficients $b_j\hat{s}_j / \hat{s}$ providing a comparative measure of variable importance (i.e., variable $x_u$ is more important than variable $x_v$ if $|b_u\hat{s}_u / \hat{s}| > |b_v\hat{s}_v / \hat{s}|$) and (ii) the sign of $b_j\hat{s}_j / \hat{s}$ indicating whether $x_j$ and $y$ tend to move in the same direction or in opposite directions. However, when $x_j$'s are not linearly independent, SRCs do not provide reliable indications of variable importance (Sect. 6.6.7, Ref. [41]).

For purposes of sensitivity analysis, there is usually no reason to construct a regression model containing all the uncertain variables (i.e., $x_1, x_2, \ldots, x_{nX}$) as indicated in Eqs. (4.6.1) and (4.6.8). Rather, a more appropriate procedure is to construct regression models in a stepwise manner. With this procedure, a regression model is first constructed with the most influential variable (e.g., $\tilde{x}_1$ as determined based on $R^2$ values for regression models containing only single variables). Then, a regression model is constructed with $\tilde{x}_1$ and the next most influential variable (e.g., $\tilde{x}_2$ as determined based on $R^2$ values for regression models containing $\tilde{x}_1$ and each of the remaining variables). The process then repeats to determine $\tilde{x}_3$ in a similar manner and continues until no more variables with an identifiable effect on $y$ can be found. Variable importance (i.e., sensitivity) is then indicated by the order in which variables are selected in the stepwise process, the changes in cumulative $R^2$ values as additional variables are added to the regression model, and the SRCs for the variables in the final regression model. An example of a sensitivity analysis of this form is presented in Table 4.3.

Table 4.3 Example of Stepwise Regression Analysis to Identify Uncertain Variables Affecting the Uncertainty in Repository Pressure at 10,000 yr in Figure 4.5a (Table 8.6, Ref. [33]).

| Step[a] | Variable[b] | SRC[c] | $R^{2}$[d] |
|---|---|---|---|
| 1 | *WMICDFLG* | 0.718 | 0.508 |
| 2 | *HALPOR* | 0.466 | 0.732 |
| 3 | *WGRCOR* | 0.246 | 0.792 |
| 4 | *ANHPRM* | 0.129 | 0.809 |
| 5 | *SHRGSSAT* | 0.070 | 0.814 |
| 6 | *SALPRES* | 0.063 | 0.818 |

[a] Steps in stepwise regression analysis.
[b] Variables listed in the order of selection in regression analysis.
[c] SRCs for variables in final regression model.
[d] Cumulative $R^2$ value with entry of each variable into regression model.

A display of regression results of the form shown in Table 4.3 is very unwieldy when results at a sequence of times are under consideration. In this situation, a more compact display of regression results is provided by plotting SRCs as functions of time for all $x_j$ that appear to have a significant effect on $y$ at some point in the time interval under consideration (Figure 4.7a).

Additional information: Sects. 6.6.2, 6.6.3, 6.6.5, Ref. [41]; Sect. 6.3, Ref. [23]. Further, general information on regression analysis is available in a number of texts (e.g., Refs. [87-91]).

### 4.6.4 Correlation

The (Pearson or sample) correlation coefficient (CC) $c(x_j, y)$ provides a measure of the strength of the linear relationship between $x_j$ and $y$ and is defined by

$$c\left(x_j, y\right) = \frac{\sum_{i=1}^{nS}\left(x_{ij} - \overline{x}_j\right)\left(y_i - \overline{y}\right)}{\left[\sum_{i=1}^{nS}\left(x_{ij} - \overline{x}_j\right)^2\right]^{1/2}\left[\sum_{i=1}^{nS}\left(y_i - \overline{y}\right)^2\right]^{1/2}}. \tag{4.6.10}$$

The CC $c(x_j, y)$ has a value between −1 and 1, with a positive value indicating that $x_j$ and $y$ tend to increase and decrease together and a negative value indicating that $x_j$ and $y$ tend to move in opposite directions. Further, gradations in the absolute value of $c(x_j, y)$ between 0 and 1 correspond to a trend from no linear relationship between $x_j$ and $y$ to an exact linear relationship between $x_j$ and $y$. As an example, the CCs associated with the scatterplots in Figure 4.8 are $c$(*HALPOR*, *REP_SATB*) = 0.75 (Figure 4.8a) and $c$(*WGRCOR*, *REP_SATB*) = –0.41 (Figure 4.8b).

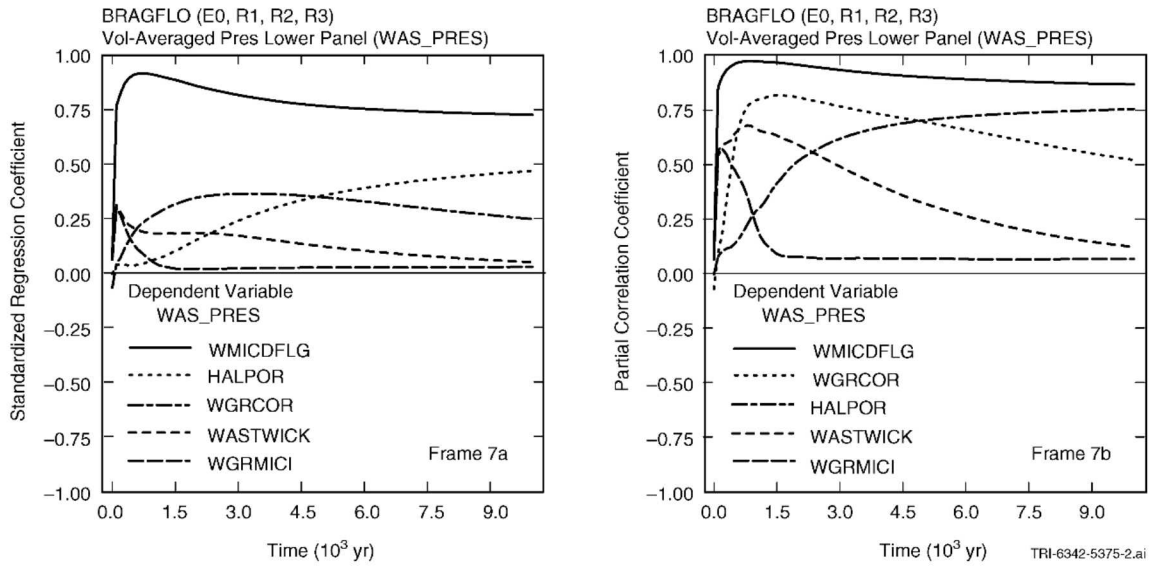



Figure 4.7 Time-dependent sensitivity analysis results for uncertain pressure curves in Figure 4.5a: (a) SRCs as a function of time, and (b) PCCs as a function of time (Figure 8.3, Ref. [33]).

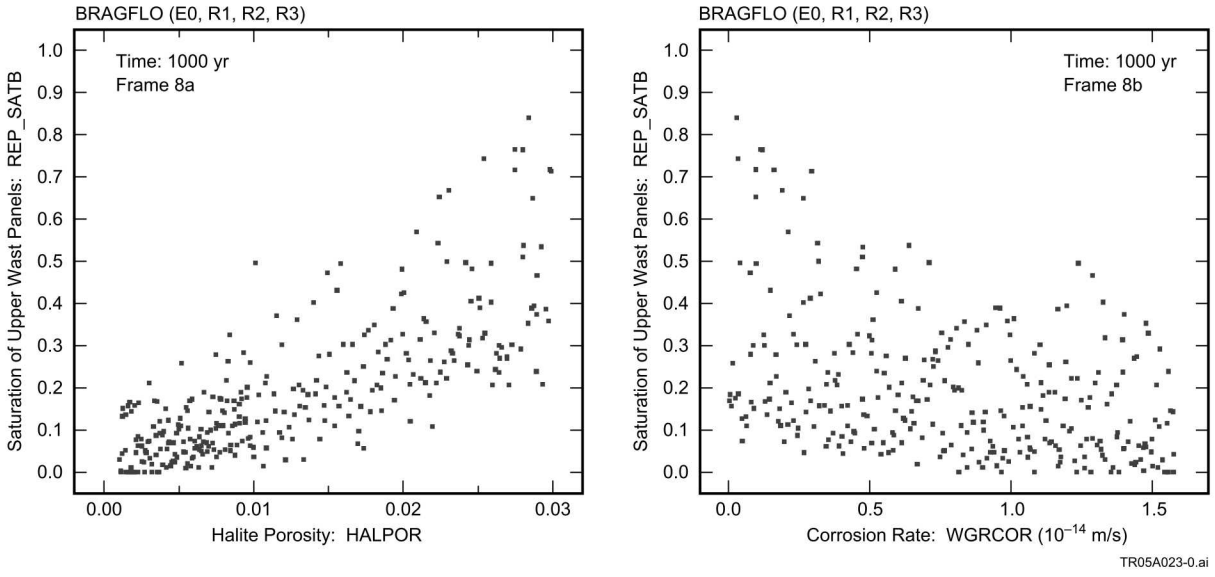



Figure 4.8 Illustration of correlation coefficients: (a) $c(x_j, y) = 0.75$ with $x_j$ = *HALPOR* and $y$ = *REP_SATB* and (b) $c(x_j, y) = -0.41$ with $x_j$ = *WGRCOR* and $y$ = *REP_SATB* (Figure 8, Ref. [21]).

The CC $c(x_j, y)$ is closely related to results obtained in a linear regression relating $y$ to $x_j$. Specifically, $c(x_j, y)$ is equal to the standardized regression coefficient in the indicated regression, and the absolute value of $c(x_j, y)$ is equal to the square root of the corresponding $R^2$ value. As a

correlation of 0 only indicates the absence of a linear association between $x_j$ and $y$, it does not preclude the existence of a well-defined nonlinear relationship between $x_j$ and $y$ (e.g., $y$ =sin $x_j$ ).

Additional information: Sect. 6.6.4, Ref. [41]; Sect. 6.2, Ref. [23].

### 4.6.5 Partial Correlation

The partial correlation coefficient (PCC) between $x_j$ and $y$ can be defined in the following manner. First, the two regression models indicated below are constructed:

$$\hat{x}_j = c_0 + \sum_{\substack{p=1 \\ p \neq j}}^{nX} c_p x_p \text{ and } \hat{y} = b_0 + \sum_{\substack{p=1 \\ p \neq j}}^{nX} b_p x_p. \tag{4.6.11}$$

Then, the results of the two preceding regressions are used to define the new variables $x_j - \hat{x}_j$ and $y - \hat{y}$. The PCC between $x_j$ and $y$ is the CC $c(x_j - \hat{x}_j, y - \hat{y})$ (see Eq. (4.6.10)) between $x_j - \hat{x}_j$ and $y - \hat{y}$. As for SRCs, PCCs are often defined for variables that are functions of time and presented as time-dependent plots (Figure 4.7b).

The PCC characterizes the linear relationship between $x_j$ and $y$ after a correction has been made for the linear effects on $y$ of the remaining elements of **x**, and the SRC characterize the effect on $y$ that results from perturbing $x_j$ by a fixed fraction of its standard deviation. Thus, PCCs and SRCs provide related, but not identical, measures of variable importance. In particular, the PCC between $x_j$ and $y$ provides a measure of variable importance that tends to exclude the effects of the other elements of **x**, the assumed distribution for $x_j$, and the magnitude of the impact of the uncertainty in $x_j$ on the uncertainty in $y$. In contrast, the SRC relating $x_j$ to $y$ is more influenced by the distribution assigned to $x_j$ and the magnitude of the impact of the uncertainty in $x_j$ on the uncertainty in $y$. However, when the elements of **x** are linearly independent, PCCs and SRCs give the same rankings of variable importance. Specifically, an ordering of variable importance based on the absolute value of PCCs is the same as an ordering based on either the absolute value of CCs or the absolute value of SRCs (Sect. 6.6.4, Ref. [41]). A cosmetic benefit of using PCCs is that PCCs tend to be spread out in value more than SRCs and thus produce results that are easier to read (e.g., compare Figure 4.7a and Figure 4.7b); however, the downside to this is that a variable can appear to have a larger effect on the uncertainty in $y$ than is actually the case.

As for analyses based on SRCs, analyses based on PCCs can give very misleading results when correlations exist between the elements of **x**. Specifically, if **x** contains two highly correlated variables, then each variable will cancel the other's effect when PCCs with $y$ are calculated.

Additional information: Sect. 6.6.4, Ref. [41]; Sect. 6.4, Ref. [23]; Ref. [92].

### 4.6.6 Rank Transformations

A rank transformation replaces values for $y_k$ and $x_j$ with their corresponding ranks. Specifically, the smallest value for a variable is assigned a rank of 1; next largest value is assigned a rank of 2; tied values are assigned their average rank; and so on up to the largest value, which is assigned a rank of *nS*. Use of the rank transformation converts a nonlinear but monotonic relationship between $y_k$ and $x_j$ to a linear relationship and produces rank (i.e., Spearman) correlations, rank regressions, standardized rank regression coefficients (SRRCs) and partial rank correlation coefficients (PRCCs). In the presence of a nonlinear but monotonic relationships between the $x_j$ and $y_k$, the use of the rank transform can substantially improve the resolution of sensitivity analysis results (Table 4.4).

Table 4.4 Comparison of Stepwise Regression Analyses with Raw and Rank-Transformed Data for Variable *BRAALIC* in Figure 4b (Table 8.8, Ref. [33]).

| | Raw Data | | | Rank-Transformed Data | | |
|---|---|---|---|---|---|---|
| **Step**[a] | **Variable**[b] | **SRC**[c] | $R^{2}$[d] | **Variable**[b] | **SRRC**[e] | $R^{2}$[d] |
| 1 | *ANHPRM* | 0.562 | 0.320 | *WMICDFLG* | –0.656 | 0.425 |
| 2 | *WMICDFLG* | –0.309 | 0.423 | *ANHPRM* | 0.593 | 0.766 |
| 3 | *WGRCOR* | –0.164 | 0.449 | *HALPOR* | –0.155 | 0.802 |
| 4 | *WASTWICK* | –0.145 | 0.471 | *WGRCOR* | –0.152 | 0.824 |
| 5 | *ANHBCEXP* | –0.120 | 0.486 | *HALPRM* | 0.143 | 0.845 |
| 6 | *HALPOR* | –0.101 | 0.496 | *SALPRES* | 0.120 | 0.860 |
| 7 | | | | *WASTWICK* | –0.010 | 0.869 |

[a] Steps in stepwise regression analysis.
[b] Variables listed in order of selection in regression analysis.
[c] SRCs for variables in final regression model.
[d] Cumulative $R^2$ value with entry of each variable into regression model.
[e] SRRCs for variables in final regression model.

An effective analysis strategy with stepwise regression analysis with either untransformed or rank-transformed variables is to have a scatterplot printed for each variable selected in the stepwise process. Even when the regression performs poorly (i.e., has a low $R^2$ value), it often selects the dominant variables even though they have nonlinear effects on the output variable of interest. Examination of the indicated scatterplots will often identify the dominant variables and provide a clear indication of why a model with a low $R^2$ value was obtained.

Additional information: Sect. 6.6.6, Ref. [41]; Sect. 6.6, Ref. [23]; Ref. [93].

### 4.6.7 Tests for Patterns Based on Gridding.

Analyses on raw and rank-transformed data can fail when the underlying relationships between the $x_j$ and $y_k$ are nonlinear and nonmonotonic (Figure 4.9). The scatterplot in Figure 4.6b is for the pressure at 10,000 yr in Figure 4.9a versus the uncertain variable *BHPRM*. The analyses with

PRCCs summarized in Figure 4.9b fail at later times because the pattern appearing in Figure 4.6b is too complex to be captured with a regression analysis based on raw or rank-transformed data.

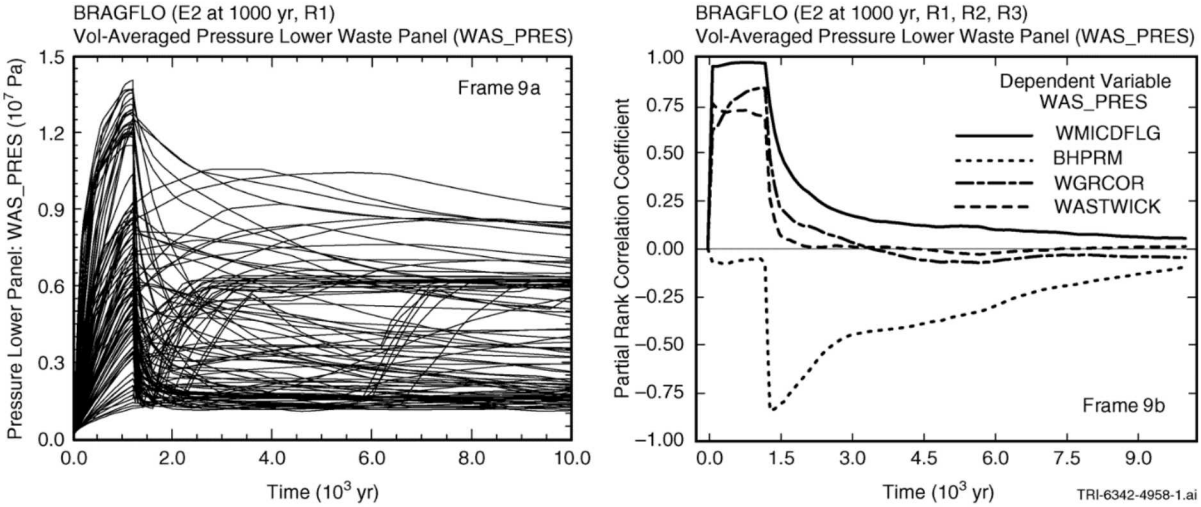



Figure 4.9 Illustration of failure of a sensitivity analysis based on rank-transformed data: (a) Pressures as a function of time and (b) PRCCs as a function of time (Figure 8.7, Ref. [33]).

An alternative analysis strategy for situations of this type is to place grids on the scatterplot for $y_i$ and $x_{ij}$ and then perform various statistical tests to determine if the distribution of points across the grid cells appears to be nonrandom. Appearance of a nonrandom pattern indicates that $x_j$ has an effect on *y*. Possibilities include (i) tests for common means (CMNs) and common distributions (denoted by CL for common locations) for values of $y_i$ based on partitioning the range of $x_{ij}$ (Figure 4.10a) and (ii) tests for common medians (CMDs) and statistical independence (SI) based on partitioning the ranges of $x_j$ and *y* (Figure 4.10a for CMDs and 4.10b for SI).

Descriptions of the preceding tests follow. The CMNs test is based on dividing the values of $x_j$ (i.e., $x_{ij}$, $i$ = 1, 2, …, $nS$) into $nI$ classes and then using the *F*-test to determine if *y* has a common mean across these classes (Sect. 3.1, Ref. [94]). The required classes are obtained by dividing the range of $x_j$ into a sequence of mutually exclusive and exhaustive subintervals containing equal numbers of sampled values (Figure 4.10a).

The CLs test employs the Kruskal-Wallis test statistic *T*, which is based on rank-transformed data and uses the same classes of $x_j$ values as the CMNs test (pp. 229-230, Ref. [95]). Specifically, if the *y* values conditional on each class of $x_j$ values have the same distribution, then the statistic *T* approximately follows a $\chi^2$ distribution with $nI - 1$ degrees of freedom (pp. 230 - 231, Ref. [95]). Thus, the probability $prob(\tilde{T} > T \mid nI - 1)$ of obtaining a value $\tilde{T}$ that exceeds $\hat{T}$

in the presence of identical $y$ distributions for the individual classes can be obtained from a $\chi^2$ distribution with $nI - 1$ degrees of freedom.

The CMDs test is based on the $\chi^2$-test for contingency tables, which can be used to test for the equality of the median values of $y$ for the classes into which the values of $x_j$ have been divided (pp. 143-144, Ref. [95]).

The SI test also uses the $\chi^2$-test to indicate if the pattern appearing in a scatterplot appears to be nonrandom. The SI test uses the same partitioning of $x_j$ values as used for the CMNs, CLs and CMDs tests. In addition, the $y$ values are also partitioned in a manner analogous to that used for the $x_j$ values (Figure 4.10b). The four tests described in this section are illustrated in Table 4.5 for $y$ = *WAS_PRES* at 10,000 yr under disturbed conditions (Figure 4.9a).

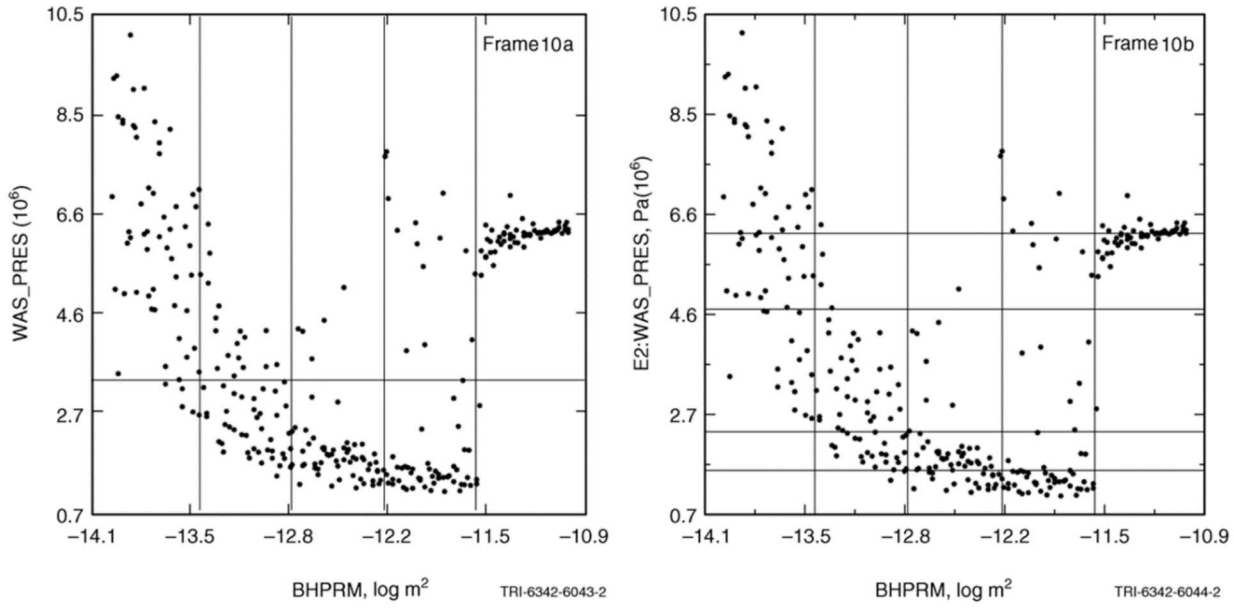



Figure 4.10 Grids used to test for nonrandom patterns: (a) Partitioning of range of $x_j$ for tests based on common means and common distributions and ranges of $x_j$ and $y_k$ for test based on common medians (Figure 8.8, Ref. [33]), and (b) Partitioning of ranges of $x_j$ and $y_k$ for test of statistical independence (Figure 8.9, Ref. [33]).

Table 4.5 Statistical Tests for Patterns Based on Gridding for Pressure (*WAS_PRES*) at 10,000 yr under Disturbed (i.e., E2) Conditions (Figure 4.9a) (adapted from Tables 4 and 21 of Ref. [96]).

| Variable[a] | CMNs: 1×5[b] | | CLs: 1×5[c] | | CMDs: 2×5[d] | | SI: 5×5[e] | |
|---|---|---|---|---|---|---|---|---|
| | Rank | *p*-val | Rank | *p*-val | Rank | *p*-val | Rank | *p*-val |
| *BHPRM* | 1 | 0.0000 | 1 | 0.0000 | 1 | 0.0000 | 1 | 0.0000 |
| *HALPRM* | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.0000 | 2 | 0.0002 |
| *ANHPRM* | 3 | 0.0002 | 3 | 0.0000 | 3 | 0.0007 | 4 | 0.0049 |
| *ANHBCEXP* | 4 | 0.0405 | 4 | 0.0602 | 4 | 0.0595 | 14 | 0.4414 |
| *HALPOR* | 5 | 0.0415 | 5 | 0.0940 | 5 | 0.0700 | 11 | 0.3142 |
| *WGRCOR* | 17 | 0.5428 | 9 | 0.2242 | 14. | 50.5249 | 3 | 0.0002 |

[a] Table includes only variables that had a *p*-value less than 0.05 for at least one of the procedures although the variable rankings for a specific procedure are based on the *p*-values obtained for that procedure for all variables considered in the analysis (see Table 1).

[b] Variable ranks and *p*-values for CMNs test with 1 × 5 grid; see Eq. (15), Ref. [23]. Exceptions for CMNs, CLs, CMDs and SI tests: because variables *ANHBCVGP* and *WMICDFLG* are discrete with 2 and 3 values, respectively (see Table 1), *nD* = 2 and 3 rather than 5 for these two variables.

[c] Variable ranks and *p*-values for CLs test with 1 × 5 grid; see Eq. (16), Ref. [23].

[d] Variable ranks and *p*-values for CMDs test with 2 × 5 grid; see Eq.(18), Ref. [23].

[e] Variable ranks and *p*-values for SI test with 5 × 5 grid; see Eq. (20), Ref. [23].

Additional information: Ref. [96]; Sects. 6.6.8 and 6.6.9, Ref. [41]; Sect. 6.6, Ref. [23]. As discussed and illustrated in Sect. 6.7 of Ref. [23], several grid-based tests for patterns in scatterplots based on measures of entropy give results similar to those obtained with the SI test.

### 4.6.8 Nonparametric Regression

Nonparametric regression seeks more general models than those obtained by least squares regression and can succeed in situations such as the one illustrated in Figure 4.9 where regression and correlation analysis based on raw and rank-transformed data fail. Nonparametric regression attempts to find models that are local in the approximation to the relationship between $y$ and multiple $x_j$'s, and, as a result, are better at capturing complex nonlinear relationships than models obtained with traditional linear regression or rank regression. Nonparametric regression models can be constructed in a stepwise manner with (i) incremental changes in $R^2$ values with the addition of successive variables to the model providing an indication of variable importance and (ii) a *p*-value (e.g., 0.02) used to determine a stopping point in the stepwise procedure (Sect. 4, Ref. [97]). As an example, the 10, 000 yr pressure results in Figure 4.9a are analyzed with a stepwise implementation of the following nonparametric regression procedures (Table 4.6): (i) locally weighted regression (commonly known as LOESS; see Sect. 3.3.1 and Table 2, Ref. [97]), (ii) generalized additive models (GAMs; see Sect. 3.3.2 and Table 3, Ref. [97]), (iii) projection pursuit regression (PP_REG; see Sect. 3.3.3 and Table 4, Ref. [97], and (iv) recursive partitioning regression (RP_REG; see Sect. 3.3.4 and Table 5, Ref. [97]). For comparison, stepwise results are also presented in Table 4.6 for stepwise implementations of (i) linear regression (LIN_REG), (ii) rank regression (RANK_REG), (iii) quadratic regression (QUAD_REG; i.e., linear regression with

the original variables, their squares, and binary products being the candidates for inclusion in the regression model; see Sect. 4.2 and Table 1, Ref. [97]), and (iv) a technique to be discussed later designated the SRD/RCC test. Brief descriptions of QUAD_REG, LOESS, GAMs, PP_REG and RP_REG are presented after the next paragraph.

Table 4.6 Nonparametric Sensitivity Analyses for Pressure (*WAS_PRES*) at 10,000 yr under Disturbed (i.e., E2) Conditions (Figure 4.9a) (adapted from Table 9, Ref. [21]).

| Variable[a] | $R^{2b}$ | *df*[c] | *p*-Val[d] | Variable | $R^2$ | *df* | *p*-Val | Variable | $R^2$ | *df* | *p*-Val |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pressure, Disturbed (i.e., E2) Conditions at 10,000 yr (Figure 9a) | | | | | | | | | | | |
| **LIN_REG** | | | | **RANK_REG** | | | | **QUAD_REG** | | | |
| *HALPRM* | 0.1410 | 1.0 | 0.0000 | *HALPRM* | 0.1289 | 1.0 | 0.0000 | *BHPRM* | 0.6098 | 2.0 | 0.0000 |
| *ANHPRM* | 0.1999 | 1.0 | 0.0000 | *ANHPRM* | 0.1866 | 1.0 | 0.0000 | *HALPRM* | 0.7006 | 3.0 | 0.0000 |
| *HALPOR* | 0.2203 | 1.0 | 0.0057 | *HALPOR* | 0.2049 | 1.0 | 0.0094 | *ANHPRM* | 0.7902 | 4.0 | 0.0000 |
| **LOESS** | | | | **PP_REG** | | | | *HALPOR* | 0.8291 | 5.0 | 0.0000 |
| *BHPRM* | 0.6625 | 8.8 | 0.0000 | *BHPRM* | 0.6646 | 9.0 | 0.0000 | *ANHBCVGP* | 0.8400 | 6.0 | 0.0023 |
| *ANHPRM* | 0.7321 | 12.8 | 0.0000 | *ANHPRM* | 0.7603 | 10.7 | 0.0000 | *WGRCOR* | 0.8532 | 7.0 | 0.0013 |
| *HALPRM* | 0.7894 | 10.5 | 0.0000 | *HALPRM* | 0.8440 | 9.8 | 0.0000 | *SHRBRSAT* | 0.8654 | 8.0 | 0.0030 |
| *ANHBCVGP* | 0.8286 | 28.9 | 0.0058 | *HALPOR* | 0.8965 | 10.4 | 0.0000 | **RP_REG** | | | |
| **GAM** | | | | **SRD/RCC TEST** | | | | *BHPRM* | 0.7163 | 17.0 | 0.0000 |
| *BHPRM* | 0.6654 | 10.0 | 0.0000 | *BHPRM* | NA | 4.0 | 0.0000 | *HALPRM* | 0.8474 | 15.0 | 0.0000 |
| *ANHPRM* | 0.7555 | 4.0 | 0.0000 | *HALPRM* | NA | 4.0 | 0.0000 | *ANHPRM* | 0.8894 | -9.0 | 0.0000 |
| *HALPRM* | 0.8242 | 2.0 | 0.0000 | *ANHPRM* | NA | 4.0 | 0.0001 | *ANRGSSAT* | 0.9726 | 81.0 | 0.0000 |
| *HALPOR* | 0.8590 | 2.0 | 0.0000 | *SHPRMDRZ* | NA | 4.0 | 0.0150 | | | | |

[a] Variables listed in order of selection.
[b] Cumulative $R^2$ value with entry of each variable into model.
[c] Incremental degrees of freedom with entry of each variable into model for all cases except SRD/RCC test; *df* fixed at 4.0 for all variables for SRD/RCC test.
[d] *p*-value for model with addition of each new variable. Stepwise procedure terminates at a *p*-value of 0.02.
[e] NA indicates that result is not applicable.

As examination of Table 4.6 shows, the LIN_REG and RANK_REG techniques perform very poorly as indicated by their $R^2$ values. In contrast, QUAD_REG and the four nonparametric regression procedures are able to capture most of the uncertainty in the variable under consideration and provide consistent rankings of the variables that contribute to this uncertainty.

Each of the traditional regression approaches (i.e., LIN_REG, RANK_REG and QUAD_REG) can be implemented the same way with a forward stepwise selection procedure and a *p*-value criterion of $\alpha = 0.02$ (Table 4.3). The forward selection procedure with QUAD_REG requires some additional explanation as there are many ways to structure this procedure to incorporate variable interactions and squares. The approach taken to obtain the QUAD_REG results in Table 4.6 is to consider a variable, its square, and all two-way interactions at each step of the selection procedure. Thus, if $\tilde{x}_1$ is the first variable selected, then the corresponding model is of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1 + \hat{\beta}_{11} \tilde{x}_1^2 . \tag{4.6.12}$$

Then, if $\tilde{x}_2$ is the second variable selected, the corresponding model is of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\tilde{x}_1 + \hat{\beta}_2\tilde{x}_2 + \hat{\beta}_{12}\tilde{x}_1\tilde{x}_2 + \hat{\beta}_{11}\tilde{x}_1^2 + \hat{\beta}_{22}\tilde{x}_2^2, \tag{4.6.13}$$

and so on (see Sect. 4.2 and Table 1, Ref. [97]).

The LOESS technique is based on the assumption that the relationship between $y$ and $\mathbf{x}$ is of the form

$$y = f(\mathbf{x}) = \alpha(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x})\mathbf{x}, \tag{4.6.14}$$

where $\boldsymbol{\beta}(\mathbf{x}) = [\beta_1(\mathbf{x}), \beta_2(\mathbf{x}), \ldots, \beta_{nX}(\mathbf{x})]$ and $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]^T$. In turn, an approximate relationship of the form

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\alpha}(\mathbf{x}) + \hat{\boldsymbol{\beta}}(\mathbf{x})\mathbf{x} \tag{4.6.15}$$

is sought with LOESS. The quantities $\hat{\alpha}(\mathbf{x})$ and $\hat{\boldsymbol{\beta}}(\mathbf{x})$ for a given value of $\mathbf{x}$ are defined to be the values for $\alpha$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{nX}]$ that minimize the sum

$$\sum_{i=1}^{nS} (\alpha + \boldsymbol{\beta}\mathbf{x}_i - y_i)^2 \left[1 - \left(\frac{\| \mathbf{x} - \mathbf{x}_i \|}{d_r(\mathbf{x})}\right)^3\right]^3 I_{[0,d_r(\mathbf{x}))}(\| \mathbf{x} - \mathbf{x}_i \|), \tag{4.6.16}$$

where (i) $d_r(\mathbf{x})$ is the distance to the $r^{\text{th}}$ nearest neighbor of $\mathbf{x}$ in $nX$-dimensional Euclidean space, (ii) $I_{[0,d_r(\mathbf{x}))}(\| \mathbf{x} - \mathbf{x}_i \|)$ equals 1 if $\|\mathbf{x} - \mathbf{x}_i\| < d_r(\mathbf{x})$ and equals 0 otherwise, and (iii) the individual independent variables (i.e., $x_1, x_2, \ldots, x_{nX}$) are normalized to mean zero and standard deviation one so that the value of the norm $\|\sim\|$ is not dominated by the units used for these variables. The determination of $\alpha$ and $\boldsymbol{\beta}$ is straightforward with the use of appropriate matrix techniques (p. 139, Ref. [98]). As a reminder, (i) in the first step of a stepwise analysis with LOAS, the model in Eq. (4.6.15) is constructed for each of the uncertain variables under consideration and the variable $\tilde{x}_j$ that contributes the most to the uncertainty in $y$ is selected, (ii) in the second step, the model in Eq. (4.6.15) is constructed for $\tilde{x}_j$ and each of the remaining uncertain variables and the variable $\tilde{x}_k$ is selected for which $\tilde{x}_j$ and $\tilde{x}_k$ contribute the most to the uncertainty in $y$, and (iii) the process continues in this manner until no additional variables are identified that contribute significantly to the uncertainty in $y$. An algorithm for stepwise variable selection with LOESS is described in Sect. 4.3 and Table 2 of Ref. [97].

For GAMs, the function $f(\mathbf{x})$ is assumed to have the form

$$f(\mathbf{x}) = \sum_{j=1}^{nX} f_j(x_j), \tag{4.6.17}$$

where the $f_j$ are arbitrary functions that will be determined as part of the analysis process. In turn, the observed values for $y$ are assumed to be of the form

$$y_i = f(\mathbf{x}_i) = \sum_{j=1}^{nX} f_j(x_{ij}). \tag{4.6.18}$$

Given initial estimates $\hat{f}_2$, $\hat{f}_3$, …, $\hat{f}_{nX}$ for $f_2$, $f_3$, …, $f_{nX}$, an estimate $\hat{f}_1$ for $f_1$ can be obtained through use of the relationship

$$y_i - \sum_{j=2}^{nX} \hat{f}_j(x_{ij}) \cong f_1(x_{i1}) \tag{4.6.19}$$

for $i$ = 1, 2, …, $nS$. In particular, a scatterplot smoother (e.g., LOESS with only one inependent variable) can be used to smooth the partial residuals on the left hand side of Eq. (4.6.19) across $x_1$. This produces an estimate $\hat{f}_1$ for $f_1$ defined across the range of values for $x_1$. Given this estimate for $f_1$, the estimate $\hat{f}_2$ for $f_2$ can be refined in the same manner across the range of values for $x_2$ with $\hat{f}_1$, $\hat{f}_3$, $\hat{f}_4$, …, $\hat{f}_{nX}$. This procedure then continues and repetitively cycles through the variables. The cycling continues until convergence is achieved. The result is $\hat{f}_j$ defined at $x_{1j}, x_{2j}, ..., x_{nS,j}$ for $j$ = 1, 2, …, $nX$. An algorithm for stepwise variable selection with GAMs is described in Sect. 4.4 and Table 3 of Ref. [97]. Additional detail on GAMs is available elsewhere (pp. 90 – 91, Ref. [99]; pp. 300 – 302, Ref. [100]).

The PP_REG procedure involves both dimension reduction and additive modeling and is based on the assumption that $f(\mathbf{x})$ has the form

$$f(\mathbf{x}) = \sum_{s=1}^{nD} g_s(\alpha_s \mathbf{x}), \tag{4.6.20}$$

where $\boldsymbol{\alpha}_s = [\alpha_{1s}, \alpha_{2s}, ..., \alpha_{nX,s}]$, $\boldsymbol{\alpha}_s$ and $\boldsymbol{\alpha}_t$ are orthogonal for $s \neq t$, $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]^{\mathrm{T}}$, $\boldsymbol{\alpha}_s \mathbf{x}$ corresponds to a linear combination of the elements of $\mathbf{x}$, and $g_s$ is an arbitrary function. Values for $g_s$, $\boldsymbol{\alpha}_s$ and $nD$ are determined as part of the analysis procedure. The expression in Eq. (4.6.20) is an additive model with the quantities $\boldsymbol{\alpha}_s \mathbf{x}$ replacing the elements $x_j$ of $\mathbf{x}$ as the independent variables. Further, this expression involves a reduction in dimension as $nD$ is usually smaller than $nX$. The entities $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$, …, $\hat{\boldsymbol{\alpha}}_{nD}$ and $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{nD}$ are estimated as part of the construction process. This is accomplished by first estimating $\boldsymbol{\alpha}_1$ and $g_1$. Specifically, $\hat{\boldsymbol{\alpha}}_1$ and $\hat{g}_1$ are defined to be the values for $\boldsymbol{\alpha}$ and $g_{\boldsymbol{\alpha}}$ that minimize the sum

$$\sum_{i=1}^{nS} \left[ y_i - g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} \mathbf{x}_i) \right]^2, \tag{4.6.21}$$

where $\boldsymbol{\alpha} \in R^{nX}$, $||\boldsymbol{\alpha}|| = 1$, and $g_{\boldsymbol{\alpha}}$ is the outcome of using a scatterplot smoother (e.g., LOESS) on the points $[y_i, \boldsymbol{\alpha}\mathbf{x}_i]$, $i = 1, 2, \ldots, nS$. Once $\hat{\boldsymbol{\alpha}}_1$ and $\hat{g}_1$ are estimated, the partial residuals $y_i - \hat{g}_1(\hat{\boldsymbol{\alpha}}_1\mathbf{x}_i)$, $i = 1, 2, \ldots, nS$, are used to obtain $\hat{\boldsymbol{\alpha}}_2$ and $\hat{g}_2$. Specifically, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{g}_2$ are defined to be the values for $\boldsymbol{\alpha}$ and $g_{\boldsymbol{\alpha}}$ that minimize the sum

$$\sum_{i=1}^{nS}\left\{\left[y_i - \hat{g}_1\left(\hat{\boldsymbol{\alpha}}_1\mathbf{x}_i\right) - g_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\mathbf{x}_i\right)\right]\right\}^2, \tag{4.6.22}$$

where $\boldsymbol{\alpha} \in R^{nX}$, $||\boldsymbol{\alpha}|| = 1$, and $g_{\boldsymbol{\alpha}}$ is the outcome of using a scatterplot smoother on the points $[y_i - \hat{g}_1(\hat{\boldsymbol{\alpha}}_1\mathbf{x}_i), \boldsymbol{\alpha}\mathbf{x}_i]$, $i = 1, 2, \ldots, nS$. This process continues until no appreciable improvement based on a relative error criterion is observed. An algorithm for stepwise variable selection with PP_REG is described in Sect. 4.5 and Table 4 of Ref. [97]. Additional details on PP_REG is available elsewhere ([101]; pp. 389-392, Ref. [102]).

The RP_REG procedure is based on splitting the data into subgroups where observations within each subgroup are more homogeneous than they are over the set of all observations. Then, $f(\mathbf{x})$ is estimated with linear regression models defined for each subgroup. Specifically, $f(\mathbf{x})$ is estimated by

$$\hat{f}\left(\mathbf{x}\right) = \sum_{s=1}^{nP}\left(\hat{\alpha}_s + \hat{\boldsymbol{\beta}}_s\mathbf{x}\right) I_s\left(\mathbf{x}\right), \tag{4.6.23}$$

where (i) $\mathcal{A}_s$, $s = 1, 2, \ldots, nP$, designate the subgroups into which the data are partitioned, (ii) $\hat{y} = \hat{\alpha}_s + \boldsymbol{\beta}_s\mathbf{x}$ is the least squares approximation to $y$ associated with $\mathcal{A}_S$, and (iii) $I_s$ is the indicator function such $I_s(\mathbf{x}) = 1$ if $\mathbf{x}$ is associated with $\mathcal{A}_s$ and $I_s(\mathbf{x}) = 0$ otherwise. The subgroups $\mathcal{A}_s$, $s = 1, 2, \ldots, nP$, are developed algorithmically from the observations $[\mathbf{x}_i, y_i]$, $i = 1, 2, \ldots, nS$. Additional details of RP_REG are discussed in Sect. 3.3.4 of Ref. [97] and an algorithm for stepwise variable selection with RP_REG is described in Sect. 4.6 and Table 5 of Ref. [97]. Closely related to RP_REG are tree-based sensitivity analyses that search for relationships between $y$ and multiple $x_j$'s by successively subdividing the sample elements $\mathbf{x}_i$ on the basis of observed effects of individual $x_j$'s on $y$ [103; 104].

An additional use of nonparametric regression models in sampling-based sensitivity analysis is to first fit a nonparametric regression model to the sampled results and then use this model in a Sobol' variance decomposition ([105]; see Sect. 6.13, Ref. [23] and Refs. [106-111] for details on Sobol' variance decomposition). Nonparametric regression models can also be effective in sensitivity analyses that involve categorical variables (i.e., epistemically uncertain analysis inputs with discrete distributions) [112].

Additional information, background references and examples with respect to the use of nonparametric regression procedures in sensitivity analysis is available in Refs. [97; 105; 113].

### 4.6.9 Squared Differences of Ranks

The SRD/RCC test is the result of combining a test for nonrandomness in the relationship between an independent and a dependent variable called the squared rank differences (SRD) test with the Spearman rank correlation coefficient (RCC) [114]. The SRD component of the test is based on the statistic

$$Q_j = \sum_{i=1}^{nS-1} \left( r_{i+1,j} - r_{ij} \right)^2, \tag{4.6.24}$$

where $r_{ij}$, $i = 1, 2, \ldots, nS$, is the rank of $y$ obtained with the sample element in which $x_j$ has rank $i$. Under the null hypothesis of no relationship between $x_j$ and $y$, the quantity

$$S_j = \left\{ Q_j - \left[ nS\left(nS^2 - 1\right)/6 \right] \right\} \Big/ \left\{ \sqrt{nS^5}/6 \right\} \tag{4.6.25}$$

approximately follows a standard normal distribution for $nS > 40$. Thus, a $p$-value $p_{ri}$ indicative of the strength of the nonlinear relationship between $x_j$ and $y$ can be obtained from $Q_j$. Specifically, $p_{ri}$ is the probability that a value $\tilde{Q}_j > Q_j$ would occur due to chance if there was no relationship between $x_j$ and $y$.

The RCC component of the test is based on the rank (i.e., Spearman) correlation coefficient

$$rc\left(x_j, y\right) = \frac{\sum_{i=1}^{nS} \left[ r\left(x_{ij}\right) - \left(nS+1\right)/2 \right] \left[ r\left(y_i\right) - \left(nS+1\right)/2 \right]}{\left\{ \sum_{i=1}^{nS} \left[ r\left(x_{ij}\right) - \left(nS+1\right)/2 \right]^2 \right\}^{1/2} \left\{ \sum_{i=1}^{nS} \left[ r\left(y_i\right) - \left(nS+1\right)/2 \right]^2 \right\}^{1/2}}, \tag{4.6.26}$$

where $r(x_{ij})$ and $r(y_i)$ are the ranks associated $x_j$ and $y$ for sample element $i$. Under the null hypothesis of no rank correlation between $x_j$ and $y$, the quantity $rc(x_j, y)$ has a known distribution (Table 10, Ref. [95]). Thus, a $p$-value $p_{cj}$ indicative of the strength of the monotonic relationship between $x_j$ and $y$ can be obtained from $rc(x_j, y)$.

The SRD/RCC test is obtained from combining the $p$-values $p_{ri}$ and $p_{cj}$ to obtain the statistic

$$\chi_4^2 = -2\left[ \ln\left(p_{rj}\right) + \ln\left(p_{cj}\right) \right], \tag{4.6.27}$$

which has a chi-square distribution with four degrees of freedom. The $p$-value associated with $\chi_4^2$ constitutes the SRD/RCC test for the strength of the relationship between $x_j$ and $y$.

Results obtained with SRD/RCC test are illustrated in Table 4.6. Like the nonparametric regression procedures, the SRD/RCC test is able to identify the nonlinear effect associated with *BHPRM* for the result in Figure 4.9a, which is completely missed with the linear regression procedures with raw and rank-transformed data.

Additional information: A detailed description of the SRD/RCC test and the determination of the associated *p*-value is available in the original article [114].

### 4.6.10 Tests for Patterns Based on Distance Measures

Tests based on distance measures consider relationships within the scatterplot for $y_i$ and $x_{ij}$ such as the distribution of distances between nearest neighbors and provide a way to identify nonrandom relationships between $y_i$ and $x_{ij}$. A positive feature of these tests is the avoidance of the problem of defining an appropriate grid as is the case with grid-based methods. However, some initial tests indicate that these tests may not be as effective as the grid-base techniques illustrated in Table 4.5 (see Sect. 6.11, Ref. [23]).

Additional information: Sect. 6.11, Ref. [23]; Refs. [115-118].

### 4.6.11 Two-Dimensional Kolmogorov-Smirnov Test

The two-dimensional Kolmogorov-Smirnov test [119-121] provides another way to test for nonrandom patterns in the scatterplot for $y_i$ and $x_{ij}$ that does not require the imposition of a grid (Sect. 6.10, Ref. [23]). However, some initial tests indicate that this test may not be as effective as the grid-base techniques illustrated in Table 4.5 (see Sect. 6.10, Ref. [23]).

Additional information: Sect. 6.10, Ref. [23]; Refs. [119-121].

### 4.6.12 Top-Down Concordance with Replicated Samples

This procedure uses the top-down coefficient of concordance to assess the consistency of results obtained in independent analyses. Possible applications include assessing the agreement of sensitivity analysis results obtained in (i) the analysis of the same sampled input with different sensitivity analysis procedures and (ii) the analysis of replicated sets of sampled input with the same sensitivity analysis procedure. In such analyses, the top-down coefficient is used to identify important variables by seeking variables with similar rankings across all replicates.

Additional information: Sect. 6.12, Ref. [23]; Refs. [39; 122].

## 4.7 Concluding Observations

Latin hypercube sampling in conjunction with examination of scatterplots, stepwise regression analysis with raw and rank-transformed data, and PCCs determined with raw and rank-transformed data is a successful and widely used approach to sensitivity analysis (e.g., [13-20]). However, this approach performs poorly when nonmonotonic relationships between sampled and calculated variables are present. As indicated in this presentation, there are a number of additional sensitivity analysis procedures that can be effective in the presence of nonmonotonic relationships between sampled and calculated variables. Of these, stepwise nonparametric regression analysis can be very effective, and it is anticipated that this approach to sensitivity analysis will experience greater use in the future when nonmonotonic relationships between sampled and calculated variables are present.

Verification of implementation correctness is an essential part of any analysis (e.g., [123-130]). Sampling-based uncertainty and sensitivity analysis can play an important role in analysis verification in two ways. First, the extensive running of the computational implementation of an analysis required to propagate the sample needed for uncertainty and sensitivity analysis entails a systematic exercising of analysis programming and numerics. This can reveal errors by a run failure or the production of results that are obviously incorrect. Second, sensitivity analysis can reveal errors by the demonstration that some input variables have effects that they should not have. Unless a well-tested analysis structure is under consideration, the first run of a sampling-based sensitivity analysis may reveal one or more errors and thus planning should allow for rerunning the analysis one or more times as the identified problems are fixed.

Definition of uncertainty distributions is the crucial first step in a sampling-based uncertainty and sensitivity analysis. As often emphasized, for analyses that will support import important decisions it is important to avoid uncertainty characterizations that are conservative (i.e., that will overemphasize negative analysis outcomes) or in some other way lead to inappropriate analysis outcomes (e.g., [131-135]). The appropriate goal in characterizing the uncertainty in analysis inputs should be to be neither deliberatively optimistic nor deliberatively pessimistic in this characterization while being honest about the uncertainty that is present.

As done in this chapter, probability is the most widely used mathematical structure for the representation of epistemic uncertainty in uncertainty and sensitivity analyses. However, a number of additional structures for the representation of epistemic uncertainty in the presence of limited information have been developed, including interval analysis, evidence theory and possibility theory (e.g., [136-145]). At present, procedures for sensitivity analysis in the context of these newer uncertainty representations are not extensively developed. As an example, an approach to sensitivity analysis in the context of evidence theory is illustrated in Refs. [146; 147].

This chapter has concentrated on a single model with the presence of only epistemic uncertainty. Many real-world analyses are more challenging as multiple linked models are involved and the effects and implications of both aleatory uncertainty and epistemic uncertainty must be assessed. Such analyses require an appropriate conceptual structure and a computational implementation that is consistent with this structure if informative uncertainty and sensitivity analysis results are to be obtained (e.g., [4; 148-150]). Examples of analyses of this type include the NUREG-1150 analyses performed to assess the safety of representative nuclear power plants

in the United States [55-60], the compliance certification analysis for the Waste Isolation Pilot Plant [38], and a performance assessment for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada [85].

Additional information on both sampling-based approaches to uncertainty and sensitivity analysis and other approaches to uncertainty and sensitivity analysis (e.g., differential analysis and variance decomposition) is available in a number of publications [151-166]. In addition, sensitivity analysis procedures based on variance decomposition are discussed in Chapter 5 and Appendices B and C.

## 4.8 References: Chapter 4

1. Apostolakis G. 1999. The Distinction between Aleatory and Epistemic Uncertainties is Important: An Example from the Inclusion of Aging Effects into PSA In *Proceedings of the International Topical Meeting on Probabilistic Safety Assessment, PSA '99: Risk Informed and Performance-Based Regulation in the New Millennium* 1:135-142. La Grange Park, IL: American Nuclear Society.
2. Apostolakis G. 1990. The Concept of Probability in Safety Assessments of Technological Systems. *Science* 250:1359-1364.
3. Aven T. 2011. On Different Types of Uncertainties in the Context of the Precautionary Principle. *Risk Analysis* 31:1515-1525.
4. Helton J.C. and C.J. Sallaberry. 2012. Uncertainty and Sensitivity Analysis: From Regulatory Requirements to Conceptual Structure and Computational Implementation. *IFIP Advances in Information and Communication Technology* 377 AICT:60-76.
5. Helton J.C., J.D. Johnson and C.J. Sallaberry. 2011. Quantification of Margins and Uncertainties: Example Analyses from Reactor Safety and Radioactive Waste Disposal Involving the Separation of Aleatory and Epistemic Uncertainty. *Reliability Engineering and System Safety* 96:1014-1033.
6. Helton J.C. 1997. Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty. *Journal of Statistical Computation and Simulation* 57:3-76.
7. Helton J.C. and D.E. Burmaster. 1996. Guest Editorial: Treatment of Aleatory and Epistemic Uncertainty in Performance Assessments for Complex Systems. *Reliability Engineering and System Safety* 54:91-94.
8. Helton J.C. 1994. Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis* 14:483-511.
9. Hoffman F.O. and J.S. Hammonds. 1994. Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability. *Risk Analysis* 14:707-712.
10. Parry G.W. 1996. The Characterization of Uncertainty in Probabilistic Risk Assessments of Complex Systems. *Reliability Engineering and System Safety* 54:119-126.
11. Parry G.W. 1988. On the Meaning of Probability in Probabilistic Safety Assessment. *Reliability Engineering and System Safety* 23:309-314.
12. Paté-Cornell M.E. 1996. Uncertainties in Risk Analysis: Six Levels of Treatment. *Reliability Engineering and System Safety* 54:95-111.
13. Das P., C. Shrubsole, B. Jones, I. Hamilton, Z. Chalabi, M. Davies, A. Mavrogianni and J. Taylor. 2014. Using Probabilistic Sampling-Based Sensitivity Analyses for Indoor Air Quality Modelling. *Building and Environment* 78:171-182.
14. Feyissa A.H., K.V. Gernaey and J. Adler-Nissen. 2014. Uncertainty and Sensitivity Analysis: Mathematical Model of Coupled Heat and Mass Transfer for a Contact Baking Process. *Journal of Food Engineering* 109:281-290.

15. Esmaeili S., N.R. Thomson, B.A. Tolson, B.J. Zebarth, S.H. Kuchta and D. Neilsen. 2014. Quantitative Global Sensitivity Analysis of the RZWQM to Warrant a Robust and Effective Calibration. *Journal of Hydrology* 511:567-579.

16. Ramin E., G. Sin, P.S. Mikkelsen and B.G. Plosz. 2014. Significance of Settling Model Structures and Parameter Subsets in Modelling WWTPs under Wet-Weather Flow and Filamentous Bulking Conditions. *Water Research* 63:209-221.

17. Calleja Rodriguez G., A. Carrillo Andres, F. Dominguez Munoz, J.M. Cejudo Lopez and Y. Zhang. 2013. Uncertainties and Sensitivity Analysis in Building Energy Simulation Using Macroparameters. *Energy and Buildings* 67:79-87.

18. Lu Z., Y.T. Grohn, R.L. Smith, D.R. Wolfgang, J.A.S. Van Kessel and Y.H. Schukken. 2009. Assessing the Potential Impact of Salmonella Vaccines in an Endemically Infected Dairy Herd. *Journal of Theoretical Biology* 259:770-784.

19. Larocque G.R., J.S. Bhatti, R. Boutin and O. Chertov. 2008. Uncertainty Analysis in Carbon Cycle Models of Forest Ecosystems: Research Needs and Development of a Theoretical Framework to Estimate Error Propagation. *Ecological Modelling* 219:400-412.

20. Mackay F.J., G.E. Apostolakis and P. Hejzlar. 2008. Incorporating Reliability Analysis into the Design of Passive Cooling Systems with an Application to a Gas-Cooled Reactor. *Nuclear Engineering and Design* 238:217-228.

21. Helton J.C., J.D. Johnson, C.J. Sallaberry and C.B. Storlie. 2006. Survey of Sampling-Based Methods for Uncertainty and Sensitivity Analysis. *Rpt. SAND2006-2901*, Sandia National Laboratories, Albuquerque, NM.

22. Helton J.C. 2004. Uncertainty and Sensitivity Analysis for Models of Complex Systems. In *Verification & Validation of Complex Models for Design and Performance Evaluation of High-Consequence Engineering Systems*, ed. J.T. Fong and R. deWit:237-263. Gaithersburg, MD: U.S. National Institute of Standards and Technology.

23. Helton J.C., J.D. Johnson, C.J. Sallaberry and C.B. Storlie. 2006. Survey of Sampling-Based Methods for Uncertainty and Sensitivity Analysis. *Reliability Engineering and System Safety* 91:1175-1209.

24. Helton J.C. 2008. Uncertainty and Sensitivity Analysis for Models of Complex Systems. In *Computational Methods in Transport: Verification and Validation.* ed. F Graziani:207-228. New York, NY: Springer-Verlag.

25. Hora S.C. and R.L. Iman. 1989. Expert Opinion in Risk Analysis: The NUREG-1150 Methodology. *Nuclear Science and Engineering* 102:323-331.

26. Thorne M.C. and M.M.R. Williams. 1992. A Review of Expert Judgement Techniques with Reference to Nuclear Safety. *Progress in Nuclear Safety* 27:83-254.

27. Budnitz R.J., G. Apostolakis, D.M. Boore, L.S. Cluff, K.J. Coppersmith, C.A. Cornell and P.A. Morris. 1998. Use of Technical Expert Panels: Applications to Probabilistic Seismic Hazard Analysis. *Risk Analysis* 18:463-469.

28. McKay M. and M. Meyer. 2000. Critique of and Limitations on the Use of Expert Judgements in Accident Consequence Uncertainty Analysis. *Radiation Protection Dosimetry* 90:325-330.

29. Meyer M.A. and J.M. Booker. 2001. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Philadelphia, PA: SIAM.

30. Ayyub B.M. 2001. *Elicitation of Expert Opinions for Uncertainty and Risks*, CRC Press, Boca Raton, FL.

31. Cooke R.M. and L.H.J. Goossens. 2004. Expert Judgement Elicitation for Risk Assessment of Critical Infrastructures. *Journal of Risk Research* 7:643-656.

32. Garthwaite P.H., J.B. Kadane and A. O'Hagan. 2005. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* 100:680-700.

33. Helton J.C. and F.J. Davis. 2000. Sampling-Based Methods for Uncertainty and Sensitivity Analysis. *Rpt. SAND99-2240*, Sandia National Laboratories, Albuquerque, NM.

34. Vaughn P., J.E. Bean, J.C. Helton, M.E. Lord, R.J. MacKinnon and J.D. Schreiber. 2000. Representation of Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:205-226.

35. Helton J.C., J.E. Bean, K. Economy, J.W. Garner, R.J. MacKinnon, J. Miller, J.D. Schreiber and P. Vaughn. 2000. Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Undisturbed Conditions. *Reliability Engineering and System Safety* 69:227-261.

36. Helton J.C., J.E. Bean, K. Economy, J.W. Garner, R.J. MacKinnon, J. Miller, J.D. Schreiber and P. Vaughn. 2000. Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Disturbed Conditions. *Reliability Engineering and System Safety* 69:263-304.

37. U.S. DOE (U.S. Department of Energy). 1996. *Title 40 CFR Part 191 Compliance Certification Application for the Waste Isolation Pilot Plant. Rpt. DOE/CAO-1996-2184, Vols. I-XXI*, U.S. Department of Energy, Carlsbad Area Office, Waste Isolation Pilot Plant, Carlsbad, NM.

38. Helton J.C. and M.G. Marietta (Editors). 2000. Special Issue: The 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:1-451.

39. Helton J.C., F.J. Davis and J.D. Johnson. 2005. A Comparison of Uncertainty and Sensitivity Analysis Results Obtained with Random and Latin Hypercube Sampling. *Reliability Engineering and System Safety* 89:305-330.

40. Helton J.C., M.-A. Martell and M.S. Tierney. 2000. Characterization of Subjective Uncertainty in the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:191-204.

41. Helton J.C. and F.J. Davis. 2000. Sampling-Based Methods. In *Sensitivity Analysis*, Eds. A. Saltelli, K. Chan, E.M. Scott. New York, NY: Wiley. pp. 101-153.

42. Cooke R.M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford; New York: Oxford University Press.

43. Meyer M.A. and J.M. Booker. 1991. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. New York, NY: Academic Press.

44. Goossens L.H.J., F.T. Harper, B.C.P. Kraan and H. Metivier. 2000. Expert Judgement for a Probabilistic Accident Consequence Uncertainty Analysis. *Radiation Protection Dosimetry* 90:295-301.

45. Goossens L.H.J. and F.T. Harper. 1998. Joint EC/USNRC Expert Judgement Driven Radiological Protection Uncertainty Analysis. *Journal of Radiological Protection* 18:249-264.

46. Siu N.O. and D.L. Kelly. 1998. Bayesian Parameter Estimation in Probabilistic Risk Assessment. *Reliability Engineering and System Safety* 62:89-116.

47. Evans J.S., G.M. Gray, R.L. Sielken Jr., A.E. Smith, C. Valdez-Flores and J.D. Graham. 1994. Use of Probabilistic Expert Judgement in Uncertainty Analysis of Carcinogenic Potency. *Regulatory Toxicology and Pharmacology* 20:15-36.

48. Chhibber S., G. Apostolakis and D. Okrent. 1992. A Taxonomy of Issues Related to the Use of Expert Judgments in Probabilistic Safety Studies. *Reliability Engineering and System Safety* 38:27-45.

49. Kaplan S. 1992. Expert Information Versus Expert Opinions: Another Approach to the Problem of Eliciting Combining Using Expert Knowledge in PRA. *Reliability Engineering and System Safety* 35:61-72.

50. Otway H. and D.V. Winterfeldt. 1992. Expert Judgement in Risk Analysis and Management: Process, Context, and Pitfalls. *Risk Analysis* 12:83-93.

51. Bonano E.J. and G.E. Apostolakis. 1991. Theoretical Foundations and Practical Issues for Using Expert Judgments in Uncertainty Analysis of High-Level Radioactive Waste Disposal. *Radioactive Waste Management and the Nuclear Fuel Cycle* 16:137-159.

52. Keeney R.L. and D.V. Winterfeldt. 1991. Eliciting Probabilities from Experts in Complex Technical Problems. *IEEE Transactions on Engineering Management* 38:191-201.

53. Svenson O. 1989. On Expert Judgments in Safety Analyses in the Process Industries. *Reliability Engineering and System Safety* 25:219-256.

54. Mosleh A., V.M. Bier and G. Apostolakis. 1988. A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment. *Reliability Engineering and System Safety* 20:63-85.

55. Breeding R.J., J.C. Helton, E.D. Gorham and F.T. Harper. 1992. Summary Description of the Methods Used in the Probabilistic Risk Assessments for NUREG-1150. *Nuclear Engineering and Design* 135:1-27.

56. Breeding R.J., J.C. Helton, W.B. Murfin, L.N. Smith, J.D. Johnson, H.-N. Jow and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Surry Nuclear Power Station. *Nuclear Engineering and Design* 135:29-59.

57. Payne A.C., Jr., R.J. Breeding, J.C. Helton, L.N. Smith, J.D. Johnson, H.-N. Jow and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Peach Bottom Atomic Power Station. *Nuclear Engineering and Design* 135:61-94.

58. Gregory J.J., R.J. Breeding, J.C. Helton, W.B. Murfin, S.J. Higgins and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Sequoyah Nuclear Plant. *Nuclear Engineering and Design* 135:95-115.

59. Brown T.D., R.J. Breeding, J.C. Helton, H.-N. Jow, S.J. Higgins and A.W. Shiver. 1992. The NUREG-1150 Probabilistic Risk Assessment for the Grand Gulf Nuclear Station. *Nuclear Engineering and Design* 135:117-137.

60. Helton J.C. and R.J. Breeding. 1993. Calculation of Reactor Accident Safety Goals. *Reliability Engineering and System Safety* 39:129-158.

61. Cullen A.C. and H.C. Frey. 1999. *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. London; New York: Plenum Press.

62. Marseguerra M. and E. Zio. 2002. *Basics of the Monte Carlo Method with Application to System Reliability*. Hagen, Germany: LiLoLe Verlag.

63. Dick J., F.Y. Kuo and I.H. Sloan. 2013. High-Dimensional Integration: The Quasi-Monte Carlo Way. *Acta Numerica* 22:133-288.

64. Sobol' I.M. 1967. On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Computational Mathematics and Mathematical Physics* 7:86-112.

65. Bratley P. and B.L. Fox. 1988. ALGORIHTM 659 Implementing Sobol's Quasirandom Sequence Generator. *ACM Transactions on Mathematical Software* 14:88-100.

66. Niederreiter H. 1992. *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

67. Rubinstein R.Y. 1981. *Simulation and the Monte Carlo Method*. New York: John Wiley & Sons.

68. Robert C.A. and G. Casella. 1999. *Monte Carlo Statistical Methods*. New York, NY: Springer.

69. Evans M. and T. Swartz. 2000. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

70. McKay M.D., R.J. Beckman and W.J. Conover. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21:239-245.

71. Helton J.C. and F.J. Davis. 2003. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Reliability Engineering and System Safety* 81:23-69.

72. Helton J.C., J.D. Johnson, M.D. McKay, A.W. Shiver and J.L. Sprung. 1995. Robustness of an Uncertainty and Sensitivity Analysis of Early Exposure Results with the MACCS Reactor Accident Consequence Model. *Reliability Engineering and System Safety* 48:129-148.

73. Iman R.L. and J.C. Helton. 1991. The Repeatability of Uncertainty and Sensitivity Analyses for Complex Probabilistic Risk Assessments. *Risk Analysis* 11:591-606.

74. Morris M.D. 2000. Three Technometrics Experimental Design Classics. *Technometrics* 42:26-27.

75. Iman R.L. and W.J. Conover. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics: Simulation and Computation* B11:311-334.

76. Iman R.L. and J.M. Davenport. 1982. Rank Correlation Plots for Use with Correlated Input Variables. *Communications in Statistics: Simulation and Computation* B11:335-360.

77. Helton J.C. and F.J. Davis. 2002. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Rpt. SAND2001-0417*, Sandia National Laboratories, Albuquerque, NM.

78. Tong C. 2006. Refinement Strategies for Stratified Sampling Methods. *Reliability Engineering and System Safety* 91:1257-1265.

79. Sallaberry C.J., J.C. Helton and S.C. Hora. 2006. Extension of Latin Hypercube Samples with Correlated Variables. *Rpt. SAND2006-6135*, Sandia National Laboratories, Albuquerque, NM.

80. Sallaberry C.J., J.C. Helton and S.C. Hora. 2008. Extension of Latin Hypercube Samples with Correlated Variables. *Reliability Engineering and System Safety* 93:1047-1059.

81. Iman R.L. Statistical Methods for Including Uncertainties Associated With the Geologic Isolation of Radioactive Waste Which Allow for a Comparison With Licensing Criteria. *Proc. Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste*, *Gatlinburg, TN*, *1981*:145-157: Washington, DC: US Nuclear Regulatory Commission, Directorate of Technical Information and Document Control.

82. Iman R.L. 1992. Uncertainty and Sensitivity Analysis for Computer Modeling Applications. In *Reliability Technology - 1992, The Winter Annual Meeting of the American Society of Mechanical Engineers, Anaheim, California, November 8-13, 1992*, ed. TA Cruse, pp. 153-168. New York, NY: American Society of Mechanical Engineers, Aerospace Division.

83. Iman R.L. and W.J. Conover. 1980. Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment. *Communications in Statistics: Theory and Methods* A9:1749-1842.

84. Kleinjen J.P.C. 2005. An Overview of the Design and Analysis of Simulation Experiments for Sensitivity Analysis. *European Journal of Operational Research* 164:287-300.

85. Helton J.C., C.W. Hansen and P.N. Swift (Editors). 2014. Special Issue: Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:1-456.

86. Cooke R.M. and J.M. van Noortwijk. 2000. Graphical Methods. In *Sensitivity Analysis*, Eds. A. Saltelli, K. Chan, E.M. Scott. New York, NY: Wiley, pp. 245-264.

87. Myers R.H. 1989. *Classical and Modern Regression with Applications*. Boston, MA: Duxbury Press.

88. Draper N.R. and H. Smith. 1981. *Applied Regression Analysis*. New York, NY: John Wiley & Sons.

89. Daniel C., F.S. Wood and J.W. Gorman. 1980. *Fitting Equations to Data: Computer Analysis of Multifactor Data*. New York, NY: John Wiley & Sons.

90. Seber G.A. 1977. *Linear Regression Analysis*. New York, NY: John Wiley & Sons.

91. Neter J. and W. Wasserman. 1974. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL: Richard D. Irwin.

92. Iman R.L., M.J. Shortencarier and J.D. Johnson. 1985. *A FORTRAN 77 Program and User's Guide for the Calculation of Partial Correlation and Standardized Regression Coefficients. Rpt. NUREG/CR-4122, SAND85-0044*, Sandia National Laboratories, Albuquerque, NM.

93. Iman R.L. and W.J. Conover. 1979. The Use of the Rank Transform in Regression. *Technometrics* 21:499-509.

94. Scheffé H. 1959. *The Analysis of Variance*. New York, NY: John Wiley & Sons, Inc.

95. Conover W.J. 1980. *Practical Nonparametric Statistics*. New York, NY: John Wiley & Sons.

96. Kleijnen J.P.C. and J.C. Helton. 1999. Statistical Analyses of Scatterplots to Identify Important Factors in Large-Scale Simulations, 1: Review and Comparison of Techniques. *Reliability Engineering and System Safety* 65:147-185.

97. Storlie C.B. and J.C. Helton. 2008. Multiple Predictor Smoothing Methods for Sensitivity Analysis: Description of Techniques. *Reliability Engineering and System Safety* 93:28-54.

98. Simonoff J.S. 1996. *Smoothing Methods in Statistics*. New York, NY: Springer-Verlag.

99. Hastie T.J. and R.J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall.

100. Chambers J.M. and T.J. Hastie. 1992. *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks.

101. Friedman J.H. and W. Stuetzle. 1981. Projection Pursuit Regression. *Journal of the American Statistical Association* 76:817-823.

102. Hastie T., R. Tibshirani and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

103. Mishra S., N.E. Deeds and B.S. RamaRao. 2003. Application of Classification Trees in the Sensitivity Analysis of Probabilistic Model Results. *Reliability Engineering and System Safety* 79:123-129.

104. Breiman L., J.H. Friedman, R.A. Olshen and C.J. Stone. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth Intl.

105. Storlie C.B., L.P. Swiler, J.C. Helton and C.J. Sallaberry. 2009. Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models. *Reliability Engineering and System Safety* 94:1735-1763.

106. Li G., C. Rosenthal and H. Rabitz. 2001. High-Dimensional Model Representations. *The Journal of Physical Chemistry* 105:7765-7777.

107. Rabitz H. and O.F. Alis. 1999. General Foundations of High-Dimensional Model Representations. *Journal of Mathematical Chemistry* 25:197-233.

108. Saltelli A., S. Tarantola and K.P.-S. Chan. 1999. A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output. *Technometrics* 41:39-56.

109. Sobol' I.M. 1993. Sensitivity Estimates for Nonlinear Mathematical Models. *Mathematical Modeling & Computational Experiment* 1:407-414.

110. Saltelli A., S. Tarantola, F. Campolongo and M. Ratto. 2004. *Sensitivity Analysis in Practice*. New York, NY: Wiley.

111. Cukier R.I., H.B. Levine and K.E. Shuler. 1978. Nonlinear Sensitivity Analysis of Multiparameter Model Systems. *Journal of Computational Physics* 26:1-42.

112. Storlie C.B., B.J. Reich, J.C. Helton, L.P. Swiler and C.J. Sallaberry. 2013. Analysis of Computationally Demanding Models with Continuous and Categorical Inputs. *Reliability Engineering and System Safety* 113:30-41.

113. Storlie C.B. and J.C. Helton. 2008. Multiple Predictor Smoothing Methods for Sensitivity Analysis: Example Results. *Reliability Engineering and System Safety* 93:55-77.

114. Hora S.C. and J.C. Helton. 2003. A Distribution-Free Test for the Relationship Between Model Input and Output when Using Latin Hypercube Sampling. *Reliability Engineering and System Safety* 79:333-339.

115. Ripley B.D. 1979. Tests of Randomness for Spatial Point Patterns. *Journal of the Royal Statistical Society* 41:368-374.

116. Diggle P.J. and T.F. Cox. 1983. Some Distance-Based Tests of Independence for Sparsely-Sampled Multivariate Spatial Point Patterns. *International Statistical Review* 51:11-23.

117. Zeng G. and R.C. Dubes. 1985. A Comparison of Tests for Randomness. *Pattern Recognition* 18:191-198.

118. Assunçao R. 1994. Testing Spatial Randomness by Means of Angles. *Biometrics* 50:531-537.

119. Peacock J.A. 1983. Two-Dimensional Goodness-Of-Fit Testing in Astronomy. *Monthly Notices of the Royal Astronomical Society* 202:615-627.

120. Fasano G. and A. Franceschini. 1987. A Multidimensional Version of the Kolmogorov-Smirnov Test. *Monthly Notices of the Royal Astronomical Society* 225:155-170.

121. Garvey J.E., E.A. Marschall and R.A. Wright. 1998. From Star Charts to Stoneflies: Detecting Relationships in Continuous Bivariate Data. *Ecology* 79:442-447.

122. Iman R.L. and W.J. Conover. 1987. A Measure of Top-Down Correlation. *Technometrics* 29:351-357.

123. American Institute of Aeronautics and Astronautics. 1998. Guide for the Verification and Validation of Computational Fluid Dynamics Simulations. *Rpt. AIAA G-077-1998*, American Institute of Aeronautics and Astronautics, Reston, VA.

124. Roache P.J. 1998. *Verification and Validation in Computational Science and Engineering*. Albuquerque, NM: Hermosa Publishers.

125. Roache P.J. 2004. Building PDE Codes to be Verifiable and Validatable. *Computing in Science & Engineering* 6:30-38.

126. Babuska I. and J.T. Oden. 2004. Verification and Validation in Computational Engineering and Science: Basic Concepts. *Computer Methods in Applied Mechanics and Engineering* 193:4057-4066.

127. Oberkampf W.L. and T.G. Trucano. 2002. Verification and Validation in Computational Fluid Dynamics. *Progress in Aerospace Sciences* 38:209-272.

128. Oberkampf W.L., T.G. Trucano and C. Hirsch. 2004. Verification, Validation, and Predictive Capability in Computational Engineering and Physics. *Applied Mechanics Review* 57:345-384.

129. Oberkampf W.L. and C.J. Roy. 2010. *Verification and Validation in Scientific Computing* New York, NY: Cambridge University Press.

130. Trucano T.G., L.P. Swiler, T. Igusa, W.L. Oberkampf and M. Pilch. 2006. Calibration, Validation, and Sensitivity Analysis: What's What. *Reliability Engineering and System Safety* 91:1331-1357.

131. Diaz N.J. Realism and Conservatism, Remarks by Chairman Diaz at the 2003 Nuclear Safety Research Conference, October 20, 2003. *Proc. NRC News*, *No. S-03-023. Washington, D.C.*, *2003*: U.S. Nuclear Regulatory Commission.

132. Paté-Cornell E. 2002. Risk and Uncertainty Analysis in Government Safety Decisions. *Risk Analysis* 22:633-646.

133. Caruso M.A., M.C. Cheok, M.A. Cunningham, G.M. Holahan, T.L. King, G.W. Parry, A.M. Ramey-Smith, M.P. Rubin and A.C. Thadani. 1999. An Approach for Using Risk Assessment in Risk-Informed Decisions on Plant-Specific Changes to the Licensing Basis. *Reliability Engineering and System Safety* 63:231-242.

134. Sielken R.L., Jr., R.S. Bretzlaff and D.E. Stevenson. 1995. Challenges to Default Assumptions Stimulate Comprehensive Realism as a New Tier in Quantitative Cancer Risk Assessment. *Regulatory Toxicology and Pharmacology* 21:270-280.

135. Nichols A.L. and R.J. Zeckhauser. 1988. The Perils of Prudence: How Conservative Risk Assessments Distort Regulation. *Regulatory Toxicology and Pharmacology* 8:61-75.

136. Baudrit C. and D. Dubois. 2006. Practical Representations of Incomplete Probabilistic Knowledge. *Computational Statistics & Data Analysis* 51:86-108.

137. Dubois D. and H. Prade. 2006. Possibility Theory and Its Applications: A Retrospective and Prospective View. *CISM Courses and Lectures* 482:89-109.

138. Klir G.J. 2006. *Uncertainty and Information: Foundations of Generalized Information Theory*. New York, NY: Wiley-Interscience.

139. Ross T.J. 2004. *Fuzzy Logic with Engineering Applications*. New York, NY: Wiley.

140. Helton J.C., J.D. Johnson and W.L. Oberkampf. 2004. An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions. *Reliability Engineering and System Safety* 85:39-71.

141. Ross T.J., J.M. Booker and W.J. Parkinson (eds.). 2002. *Fuzzy Logic and Probability Applications: Bridging the Gap*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

142. Klir G.J. and M.J. Wierman. 1999. Uncertainty-Based Information, Physica-Verlag, New York, NY.

143. Bardossy G. and J. Fodor. 2004. Evaluation of Uncertainties and Risks in Geology, Springer-Verlag, New York, NY.

144. Helton J.C., J.D. Johnson, W.L. Oberkampf and C.J. Sallaberry. 2010. Representation of Analysis Results Involving Aleatory and Epistemic Uncertainty. *International Journal of General Systems* 39:605-646.

145. Helton J.C. and J.D. Johnson. 2011. Quantification of Margins and Uncertainties: Alternative Representations of Epistemic Uncertainty. *Reliability Engineering and System Safety* 96:1034-1052.

146. Helton J.C., J.D. Johnson, W.L. Oberkampf and C.J. Sallaberry. 2006. Sensitivity Analysis in Conjunction with Evidence Theory Representations of Epistemic Uncertainty. *Reliability Engineering and System Safety* 91:1414-1434.

147. Helton J.C., J.D. Johnson, W.L. Oberkampf and C.B. Storlie. 2007. A Sampling-Based Computational Strategy for the Representation of Epistemic Uncertainty in Model Predictions with Evidence Theory. *Computational Methods in Applied Mechanics and Engineering* 196:3980-3998.

148. Helton J.C., C.W. Hansen and C.J. Sallaberry. 2014. Conceptual Structure and Computational Organization of the 2008 Performance Assessment for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety* 122:223-248.

149. Helton J.C. 2011. Quantification of Margins and Uncertainties: Conceptual and Computational Basis. *Reliability Engineering and System Safety* 96:976-1013.

150. Helton J.C., D.R. Anderson, H.-N. Jow, M.G. Marietta and G. Basabilvazo. 2000. Conceptual Structure of the 1996 Performance Assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69:151-165.

151. Helton J.C. 1993. Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal. *Reliability Engineering and System Safety* 42:327-367.

152. Blower S.M. and H. Dowlatabadi. 1994. Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: an HIV Model, as an Example. *International Statistical Review* 62:229-243.

153. Hamby D.M. 1994. A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models. *Environmental Monitoring and Assessment* 32:135-154.

154. Frey H.C. and S.R. Patil. 2002. Identification and Review of Sensitivity Analysis Methods. *Risk Analysis* 22:553-578.

155. Ionescu-Bujor M. and D.G. Cacuci. 2004. A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems--I: Deterministic Methods. *Nuclear Science and Engineering* 147:189-2003.

156. Cacuci D.G. and M. Ionescu-Bujor. 2004. A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems--II: Statistical Methods. *Nuclear Science and Engineering* 147:204-217.

157. Cacuci D.G. 2003. *Sensitivity and Uncertainty Analysis, Vol. 1: Theory*. Boca Raton, FL: Chapman and Hall/CRC Press.

158. Cacuci D.G., M. Ionescu-Bujor and I.M. Navon. 2005. *Sensitivity and Uncertainty Analysis, Vol. II, Application to Large-Scale Systems*. Boca Raton, FL: Chapman & Hall/CRC.

159. Saltelli A., K. Chan and E.M. Scott (eds). 2000. *Sensitivity Analysis*. New York, NY: Wiley.

160. Saltelli A., M. Ratto, S. Tarantola and F. Campolongo. 2005. Sensitivity Analysis for Chemical Models. *Chemical Reviews* 105:2811-2828.

161. Wu J., R. Dhingra, M. Gambhir and J.V. Remais. 2013. Sensitivity Analysis of Infectious Disease Models: Methods, Advances and Their Application. *Journal of the Royal Society Interface* 10:Article number 1018.

162. Matott L.S., J.E. Badendreier and S.T. Purucker. 2009. Evaluating Uncertainty in Integrated Environmental Models: A Review of Concepts and Tools. *Water Resources Research* 45:W06421.

163. Wei P., Z. Lu and S. Song. 2015. Variable Inportance Analysis: A Conprehensive Review. *Reliability Engineering and System Safety* 142:399-432.

164. Pianosi F., K. Beven, J. Freer, J.W. Hall, J. Rougier, D.B. Stephenson and T. Wagener. 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling and Software* 79:214-232.

165. Borgonovo E. and E. Plischke. 2016. Sensitivity Analysis: A Review of Recent Advances. *European Journal of Operational Research* 248:869-887.

166. Ghanem R., Higdon, D., Owhadi, H., ed. 2017. *Handbook of Uncertainty Quantification*: Springer International Publishing Switzerland.

# 5. VARIANCE-BASED PROCEDURES FOR UNCERTAINTY AND SENSITIVITY ANALYSIS

## 5.1 Introduction

This chapter provides an introduction to variance-based procedures for uncertainty and sensitivity analysis. The goal of these methods is to provide a decomposition of the output or response variance in terms of the effect and variance of individual input parameters. For example, an analysis could show that the effect and variance in input $x_1$ contributed 15% of the output variance while the effect and variance in $x_2$ contributed 48% of the output variance. Variance-based analysis is a form of sensitivity analysis. It also highlights which inputs merit spending more effort on uncertainty characterization to improve the estimation of output variance. Variance-based sensitivity analysis can also be used to identify which inputs are insignificant in terms of contributing to the output variance, thus helping simplify an uncertainty analysis. Variance-based methods were not used in the WIPP and YM performance assessments. Their usage has become popular in the past two decades due to computational and algorithmic improvements. *GDSA Framework* will incorporate the use of variance-based procedures.

The following topics are considered in this chapter: differential analysis (Sect. 5.2), Sobol' variance decomposition (Sect. 5.3), variance decompositions based on polynomial chaos expansions (Sect. 5.4), and variance decomposition based on Fourier series (Sect. 5.5). In addition, Sobol' variance decomposition and polynomial chaos expansions are described in more detail in Appendices B and C.

## 5.2 Differential Analysis

Differential analysis is based on using a Taylor series to approximate the model under consideration. Once constructed, this series is used as a surrogate for the original model in uncertainty and sensitivity analyses. The following description of differential analysis is a lightly edited adaptation of Sect. 2.1 of Ref. [1].

For notational convenience, it is assumed that the model under consideration can be represented by a function of the form

$$y = f(x_1, x_2, ..., x_n) = f(\mathbf{x}), \tag{5.2.1}$$

where $y$ is the model prediction and $\mathbf{x} = [x_1, x_2, ..., x_n]$ is the input to the model. The output variable $y$ is assumed to be single valued to keep the notation from becoming unwieldy although, in practice, there is usually more than one predicted variable of interest.

A differential analysis involves four steps. In the first step, base values, ranges and distributions are selected for the input variables $x_i, i = 1, 2, ..., n$, in Eq. (5.2.1). The base values can be represented by the vector

$$\mathbf{x}_0 = [x_{10}, x_{20}, ..., x_{n0}] \tag{5.2.2}$$

and form the point about which the Taylor series will be developed. The ranges and distributions will be used later in the analysis when the effects on the output variable $y$ of perturbations away from $\mathbf{x}_0$ are investigated.

In the second step, a Taylor series approximation to $y$ is developed. If the series is restricted to first order terms, the approximation has the form

$$y(\mathbf{x}) \cong y(\mathbf{x}_0) + \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right] [x_i - x_{i0}]. \tag{5.2.3}$$

If first and second order terms are used, the approximation becomes

$$y(\mathbf{x}) \cong y(\mathbf{x}_0) + \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right] [x_i - x_{i0}] + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} \right] [x_i - x_{i0}][x_j - x_{j0}]. \tag{5.2.4}$$

Higher order expansions are also possible if $f$ has the necessary partial derivatives. In concept, the order of the approximation should be determined by the curvature of the surface $y = f(\mathbf{x})$. In practice, determination of an appropriate order can be difficult. Evaluation of the partial derivatives in Eqs. (5.2.3) and (5.2.4) is usually the most computationally demanding part of a differential analysis.

In the third step, variance propagation techniques are used to estimate the uncertainty in $y$. If the approximation in Eq. (5.2.3) is used, the expected value and variance of $y$ can be estimated by

$$E(y) \cong y(\mathbf{x}_0) + \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right] E\left[ x_i - x_{i0} \right] = y(\mathbf{x}_0) \tag{5.2.5}$$

and

$$V(y) \cong \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right]^2 V(x_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right] \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_j} \right] Cov(x_i, x_j), \tag{5.2.6}$$

where $E$, $V$ and $Cov$ denote expected value, variance and covariance, respectively, and $x_{i0}$ is assumed to equal $E(x_i)$. If the $x_i$ are uncorrelated, the variance decomposition in Eq. (5.2.6) reduces to

$$V(y) \cong \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right]^2 V(x_i). \tag{5.2.7}$$

If the approximation in Eq. (5.2.4) is used, the $x_i$ are uncorrelated, and fourth order and higher order terms are ignored in the representation for $V(y)$, the representations for expected value and variance are

$$E(y) \cong y(\mathbf{x}_0) + \frac{1}{2} \sum_{i=1}^{n} \left[ \frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 x_i} \right] V(x_i) \tag{5.2.8}$$

and

$$V(y) \cong \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right]^2 V(x_i) + \sum_{i=1}^{n} \left[ \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \right] \left[ \frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 x_i} \right] \mu_3(x_i), \tag{5.2.9}$$

where $\mu_3(x_i)$ denotes the third central moment for $x_i$. As correlations between the $x_i$ and higher order terms are added, approximations to the expected value and variance for $y$ rapidly become very complicated [2-4]. Together, $E(y)$ and $V(y)$ provide a representation for the uncertainty in $y$ that results from the uncertainty in the $x_i$. It is also possible to obtain a representation for the uncertainty in $y$ by using the approximation in Eq. (5.2.3) or (5.2.4) and letting the $x_i$ vary over their ranges.

In the fourth and final step, the Taylor series approximations are used to estimate the importance of the individual $x_i$ in contributing to the uncertainty in $y$. Fractional contributions to variance can be estimated from the expressions in Eqs. (5.2.7) and (5.2.9) and provide one means of measuring variable importance. If the representation for $y$ in Eq. (5.2.3) is used, then the contribution of $x_i$ to the variance of $y$ can be estimated with the ratio

$$\left[\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\right]^2 V(x_i)/V(y), \tag{5.2.10}$$

where $V(y)$ is defined in Eq. (5.2.7). Similarly, if the representation for $y$ in Eq. (5.2.4) is used, then the contribution of $x_i$ to the variance of $y$ can be estimated with the ratio

$$\left\{\left[\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\right]^2 V(x_i)+\left[\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\right]\left[\frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 x_i}\right]\mu_3(x_i)\right\}/V(y), \tag{5.2.11}$$

where $V(y)$ is defined in Eq. (5.2.9). The fractional contributions to variance shown in Eqs. (5.2.10) and (5.2.11) can be used to order the $x_i$ with respect to their contribution to the uncertainty in $y$. This ordering involves the actual effects of the $x_i$, as measured by their partial derivatives, and the effects of the distributions assigned to the $x_i$, as measured by $V(x_i)$ and $\mu_3(x_i)$.

Another way of measuring variable importance is by normalizing the partial derivatives in the first order Taylor series approximation to $y$ in Eq. (5.2.3). Specifically, this approximation can be restated in the following two forms:

$$\frac{y(\mathbf{x})-y(\mathbf{x}_0)}{y(\mathbf{x}_0)} \cong \sum_{i=1}^{n}\left[\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\frac{x_{i0}}{y(\mathbf{x}_0)}\right]\left[\frac{x_i-x_{i0}}{x_{i0}}\right] \tag{5.2.12}$$

and

$$\frac{y(\mathbf{x})-y(\mathbf{x}_0)}{SD(y)} \cong \sum_{i=1}^{n}\left[\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\frac{SD(x_i)}{SD(y)}\right]\left[\frac{x_i-x_{i0}}{SD(x_i)}\right], \tag{5.2.13}$$

where SD is used to denote standard deviation and it is assumed that there is no problem with division by zero. The relation in Eq. (5.2.7) is used to provide an estimate for $SD(y)$. The normalized coefficients (i.e., derivatives)

$$\frac{\partial f(\mathbf{x}_0)}{\partial x_i}\frac{x_{i0}}{y(\mathbf{x}_0)} \quad \text{and} \quad \frac{\partial f(\mathbf{x}_0)}{\partial x_i}\frac{SD(x_i)}{SD(y)} \tag{5.2.14}$$

provide related, but different, measures of the effect of individual $x_i$ on $y$. The coefficients from Eq. (5.2.12) measure the effect on $y$ that results from perturbing each $x_i$ by a fixed fraction of its base value $x_{i0}$. Thus, these coefficients can be used to compare the individual $x_i$ based on the effects of perturbing each $x_i$ by the same fraction of its base value. The coefficients from Eq. (5.2.13) measure the effect on $y$ that results from perturbing each $x_i$ by a fixed fraction of its standard deviation $SD(x_i)$. Thus, these coefficients can be used to compare the individual $x_i$ based on the effects of perturbing each $x_i$ by the same fraction of its standard deviation. Further, the sign of the coefficient indicates whether $x_i$ and $y$ tend to move in the same direction or in opposite directions. The absolute values of these coefficients can be used to rank the relative importance of the individual $x_i$. A ranking based on the coefficients from Eq. (5.2.13) includes the effects of distribution assumptions while a ranking based on the coefficients from Eq. (5.2.12) does not.

As a reminder, the individual derivatives $\partial f(\mathbf{X}_0)/\partial x_i$ by themselves are not very useful as a means to compare the effects on $y$ of the individual $x_i$ because value of $\partial f(\mathbf{X}_0)/\partial x_i$ is determined in part by the units in which $x_i$ is defined. Thus, the "normalized" derivatives in Eq. (5.2.14) provide more meaningful representations of the effects of individual variables than the partial derivatives from which they are derived. Specifically, the normalizations in Eq. (5.2.14) remove the effects of the units specified for the individual $x_i$.

As noted earlier, the largest effort in a differential analysis is usually the determination of the partial derivatives in Eqs. (5.2.3) and (5.2.4). Because of this, a number of specialized techniques have been developed to facilitate the calculation of these derivatives. Included in these are adjoint techniques [5-10], Green's function techniques [11-13], and the GRESS/ADGEN compiler [14-18]. Adjoint and Green's function techniques can provide significant computational savings in the calculation of partial derivatives but can be very tedious and time consuming to implement because they require detailed manipulations involving the equations that define the model under consideration. The GRESS/ADGEN compiler was developed to avoid the human cost associated with these manipulations by performing the necessary manipulations directly from the FORTRAN code that implements the model. However, the basic ideas in differential analysis remain the same regardless of the procedure used to calculate the necessary partial derivatives.

As indicated, the determination of the needed derivatives for a Taylor series expansion for a complex model can be a significant numerical challenge. In general, the needed derivatives are not obtained by simple differentiations but rather are obtained by appropriate numerical procedures as will be illustrated by the following example.

For this example, the quantity of interest $y(t, \mathbf{X}_0)$ is defined by the differential equation

$$\mathrm{d}y(t, \mathbf{X}_0)/\mathrm{d}t = f[t, \mathbf{X}_0, y(t, \mathbf{X}_0)], y(0, \mathbf{X}_0) = g(\mathbf{X}_0) \quad \text{for} \quad 0 \le t \tag{5.2.15}$$

with $\mathbf{x}_0 = [x_{10}, x_{20}, ..., x_{n0}]$ corresponding to the base value for a vector $\mathbf{x} = [x_1, x_2, ..., x_n]$ of epistemically uncertain quantities. In turn, the desire is to approximate $y(t, \mathbf{x}_0)$ with the Taylor series

$$y(t, \mathbf{x}) \cong y(t, \mathbf{x}_0) + \sum_{i=1}^{n} \left[ \frac{\partial y(t, \mathbf{x}_0)}{\partial x_i} \right] [x_i - x_{i0}], \tag{5.2.16}$$

which requires determination of the derivatives $\partial y(t, \mathbf{x}_0) / \partial x_i$.

For this example, the desired derivatives can be obtained as solutions to appropriately defined differential equations. The starting point in development of the indicated equations is to differentiate the expressions in Eq. (5.2.15) to obtain the relationships

$$\partial \left[ \mathrm{d}y(t, \mathbf{x}_0) / \mathrm{d}t \right] / \partial x_i = \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial x_i \quad \text{and} \quad \partial \left[ y(0, \mathbf{x}_0) \right] / \partial x_i = \partial \left[ g(\mathbf{x}_0) \right] / \partial x_i. \tag{5.2.17}$$

Next, for the first equality in the preceding equation, a change in the order of differentiation produces

$$\begin{aligned} \mathrm{d}\left[ \partial y(t, \mathbf{x}_0) / \partial x_i \right] / \mathrm{d}t &= \partial \left[ \mathrm{d}y(t, \mathbf{x}_0) / \mathrm{d}t \right] / \partial x_i \\ &= \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial x_i \\ &= \left[ \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial x_i \right] \left[ \partial x_i / \partial x_i \right] + \left[ \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial y \right] \left[ \partial y(t, \mathbf{x}_0) / \partial x_i \right] \\ &= \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial x_i + \left[ \partial \left\{ f[t, \mathbf{x}_0, y(t, \mathbf{x}_0)] \right\} / \partial y \right] \left[ \partial y(t, \mathbf{x}_0) / \partial x_i \right], \end{aligned} \tag{5.2.18}$$

with the third equality following from the relationship

$$\begin{aligned} \partial \left\{ f[\mathbf{c}, u(x), v(x)] \right\} / \partial x = &\left[ \partial \left\{ f[\mathbf{c}, u(x), v(x)] \right\} / \partial u \right] \left[ u(x) / \partial x \right] \\ &+ \left[ \partial \left\{ f[\mathbf{c}, u(x), v(x)] \right\} / \partial v \right] \left[ v(x) / \partial x \right] \end{aligned} \tag{5.2.19}$$

in which $\mathbf{c}$ is a vector of constants. Specifically,

$$\mathbf{c} = [x_{10}, ..., x_{i-1,0}, x_{i+1,0}, ..., x_n] \quad \text{and} \quad u(x) = u(x_i) = x_i \tag{5.2.20}$$

when the derivative in Eq. (5.2.19) is used to obtain the third equality in Eq. (5.2.18).

The following system of 2 differential equations

$$\begin{aligned} \mathrm{d}y(t,\mathbf{x}_0)/\mathrm{d}t &= f[t,\mathbf{x}_0,y(t,\mathbf{x}_0)] \text{ with } y(0,\mathbf{x}_0)=g(\mathbf{x}_0), \\ \mathrm{d}\left[\partial y(t,\mathbf{x}_0)/\partial x_i\right]/\mathrm{d}t &= \partial\left\{f[t,\mathbf{x}_0,y(t,\mathbf{x}_0)]\right\}/\partial x_i \\ &\quad +\left[\partial\left\{f[t,\mathbf{x}_0,y(t,\mathbf{x}_0)]\right\}/\partial y\right]\left[\partial y(t,\mathbf{x}_0)/\partial x_i\right] \\ &\quad \text{with } \partial y(0,\mathbf{x}_0)/\partial x_i = \partial\left[g(\mathbf{x}_0)\right]/\partial x_i \end{aligned} \tag{5.2.21}$$

for $0 \le t$ defining $y(t,\mathbf{x}_0)$ and $\partial y(t,\mathbf{x}_0)/\partial x_i$ results from combining the differential equations in Eqs. (5.2.15) and (5.2.18). Approximations to $y(t,\mathbf{x}_0)$ are obtained by numerically solving Eq. (5.2.15). Similarly, approximations to $\partial y(t,\mathbf{x}_0)/\partial x_i$ can be obtained by simultaneously solving the system of 2 equations in Eq. (5.2.21).

Solving the system of 2 equations in Eq. (5.2.21) yields approximations to $y(t,\mathbf{x}_0)$ and $\partial y(t,\mathbf{x}_0)/\partial x_i$. However, because the vector $\mathbf{x}=[x_1,x_2,...,x_n]$ of epistemically uncertain quantities has *n* elements, obtaining the Taylor series in Eq. (5.2.16) requires solving the system of 2 equations in Eq. (5.2.21) *n* times. This has the potential to be very computationally demanding. The computational demands become significantly greater when the single differential equation in Eq. (5.2.15) is replaced by a system of *m* differential equations defining $y_j(t,\mathbf{x}_0)$ for $j=1,2,...,m$. In this case, construction of Taylor series approximations of the form in Eq. (5.2.16) for the solutions of the indicated system of *m* differential equations would require the solution of *n* systems 2*m* differential equations.

To avoid the computational cost of solving multiple systems of differential equations, computational procedures have been developed to solve the original differential equation or system of differential equations only once and then use the resultant solutions in conjunction with appropriate integration procedures to obtain the desired derivatives with respect to the elements of $\mathbf{x}=[x_1,x_2,...,x_n]$. For notational simplicity, these procedures are illustrated below with the differential equation in Eq. (5.2.15) replaced by a system of 2 differential equations.

The system of 2 differential equations used for illustration can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y_1(t,\mathbf{x}_0) \\ y_2(t,\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right] \\ f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right] \end{bmatrix} \text{ with } \begin{bmatrix} y_1(0,\mathbf{x}_0) \\ y_2(0,\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} g_1(\mathbf{x}_0) \\ g_2(\mathbf{x}_0) \end{bmatrix} \tag{5.2.22}$$

Similarly, the system of 2 differential equations defining partials with respect to $x_i$ can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\partial y_1(t,\mathbf{x}_0)/\partial x_i\\ \partial y_2(t,\mathbf{x}_0)/\partial x_i\end{bmatrix}=\begin{bmatrix}\partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial x_i\\ \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial x_i\end{bmatrix}$$
$$+\begin{bmatrix}\partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_1 & \partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_2\\ \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_1 & \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_2\end{bmatrix}\begin{bmatrix}\partial y_1(t,\mathbf{x}_0)/\partial x_i\\ \partial y_2(t,\mathbf{x}_0)/\partial x_i\end{bmatrix}$$
$$\text{with}\quad\begin{bmatrix}\partial y_1(0,\mathbf{x}_0)/\partial x_i\\ \partial y_2(0,\mathbf{x}_0)/\partial x_i\end{bmatrix}=\begin{bmatrix}\partial g_1(\mathbf{x}_0)/\partial x_i\\ \partial g_2(\mathbf{x}_0)/\partial x_i\end{bmatrix}. \tag{5.2.23}$$

The needed partial derivatives to approximate $y_1(t,\mathbf{x}_0)$ and $y_2(t,\mathbf{x}_0)$ with Taylor series can be obtained by solving the system of 4 equations in Eqs. (5.2.22) and (5.2.23) *n* times.

An alternative and potentially less computationally intensive way to obtain the needed partial derivatives is to solve the system of 2 equations in Eq. (5.2.22) once and then use the resultant solutions for $y_1(t,\mathbf{x}_0)$ and $y_2(t,\mathbf{x}_0)$ in an integration procedure to obtain $\partial y_1(t,\mathbf{x}_0)/\partial x_i$ and $\partial y_2(t,\mathbf{x}_0)/\partial x_i$. This is accomplished by taking advantage of the fact that Eq. (5.2.23) is a linear differential equation of the form

$$\mathrm{d}\mathbf{w}_i(\mathrm{t})/\mathrm{d}t=\mathbf{u}_i(t)+\mathbf{M}(t)\mathbf{w}_i(\mathrm{t})\ \text{ with }\ \mathbf{v}_i=\mathbf{w}_i(0), \tag{5.2.24}$$

with

$$\mathbf{w}_i(t)=\begin{bmatrix}\partial y_1(t,\mathbf{x}_0)/\partial x_i\\ \partial y_2(t,\mathbf{x}_0)/\partial x_i\end{bmatrix},\ \mathbf{u}_i(t)=\begin{bmatrix}\partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial x_i\\ \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial x_i\end{bmatrix}, \tag{5.2.25}$$

$$\mathbf{M}(t)=\begin{bmatrix}\partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_1 & \partial f_1\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_2\\ \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_1 & \partial f_2\left[t,\mathbf{x}_0,y_1(t,\mathbf{x}_0),y_2(t,\mathbf{x}_0)\right]/\partial y_2\end{bmatrix}, \tag{5.2.26}$$

$$\mathbf{v}_i=\mathbf{w}_i(0)=\begin{bmatrix}\partial y_1(0,\mathbf{x}_0)/\partial x_i\\ \partial y_2(0,\mathbf{x}_0)/\partial x_i\end{bmatrix}=\begin{bmatrix}\partial g_1(\mathbf{x}_0)/\partial x_i\\ \partial g_2(\mathbf{x}_0)/\partial x_i\end{bmatrix}. \tag{5.2.27}$$

Given that the solution to the equation in Eq. (5.2.23) has been determined, the solution to the equation in Eq. (5.2.24) is given by

$$\mathbf{w}_i(\mathrm{t})=\prod_t^0\left[\mathbf{I}-\mathbf{M}(\tau)\mathrm{d}\tau\right]\mathbf{w}_i(0)+\int_0^t\prod_t^s\left[\mathbf{I}-\mathbf{M}(\tau)\mathrm{d}\tau\right]\mathbf{u}_i(s)\mathrm{d}s, \tag{5.2.28}$$

where

$$\prod_{t}^{0}\left[\mathbf{I}-\mathbf{M}(\tau)\mathrm{d}\tau\right]$$

$$= \lim_{\Delta t_j \to 0} \prod_{j=1}^{n}\left[\mathbf{I}-\mathbf{M}(t_j)\Delta t_j\right] \quad \text{for } t=t_0<t_1<\cdots<t_n=0 \text{ and } \Delta t_j = t_j - t_{j-1} \tag{5.2.29}$$

$$= \lim_{\Delta t_j \to 0} \prod_{j=n}^{1}\left[\mathbf{I}+\mathbf{M}(t_j)\Delta t_j\right] \quad \text{for } 0=t_0<t_1<\cdots<t_n=t \text{ and } \Delta t_j = t_j - t_{j-1}.$$

and $\mathbf{I}$ is the $2\times 2$ identity matrix. The quantity defined in the preceding equation (i.e., $\prod_{t}^{0}\left[\mathbf{I}-\mathbf{M}(\tau)\mathrm{d}\tau\right]$) is known as the product integral from *t* to 0 and is analogous to a sum integral except that, as indicated in Eq. (5.2.29), it is defined as a limit of products rather than as limit of sums [19].

Once the initial equation in Eq. (5.2.22) is solved, the matrix $\mathbf{M}(\tau)$ and associated product integral can be determined and the desired partial derivatives $\partial y_1(t,\mathbf{x}_0)/\partial x_i$ and $\partial y_2(t,\mathbf{x}_0)/\partial x_i$ obtained by numerically evaluating the sum integral in Eq. (5.2.28). In general, this is an efficient way to obtain $\partial y_1(t,\mathbf{x}_0)/\partial x_i$ and $\partial y_2(t,\mathbf{x}_0)/\partial x_i$ because (i) Eq. (5.2.22) has to be solved only once and (ii) numerical evaluation of the expression in Eq. (5.2.28) less computationally demanding than solving Eqs. (5.2.22) and (5.2.23) for each $x_i$.

At present, differential analysis is of more historical interest than current computational interest. This is the case because analyses based on Taylor series are inherently local. When the function $f(x_1,x_2,...,x_n)$ under consideration is nonlinear and the variables $x_1,x_2,...,x_n$ have wide ranges, analyses based on Taylor series expansions are unlikely to provide meaningful results. Taylor series analyses are most likely to be useful when only small perturbations of the variables $x_1,x_2,...,x_n$ from their base values are under consideration.

Additional information on differential analysis is available in a number of publications (e.g., [20-29]).

## 5.3 Sobol' Variance Decomposition

The Sobol' variance decomposition of a function $f(\mathbf{x})$ is based on a representation of $f(\mathbf{x})$ by a sum of orthogonal functions as described in Refs. [30; 31]. As the notation for this decomposition becomes rather cumbersome and tends to obscure the nature of the decomposition when $\mathbf{x} = [x_1, x_2, ..., x_n]$ is a vector of arbitrary length *n*, the decomposition will be initially described for $\mathbf{x} = [x_1, x_2, x_3]$.

As required for the Sobol' decomposition, the epistemically uncertain variables $x_1$, $x_2$ and $x_3$ are assumed to be independent; further, the uncertainty in $x_1$, $x_2$ and $x_3$ is characterized by the probability spaces $(\mathcal{X}_1, \mathbb{X}_1, m_1)$, $(\mathcal{X}_2, \mathbb{X}_2, m_2)$ and $(\mathcal{X}_3, \mathbb{X}_3, m_3)$ with associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$.

For $\mathbf{x} = [x_1, x_2, x_3]$, the Sobol' variance decomposition is based on the following representation for $f(\mathbf{x})$:

$$f(\mathbf{x}) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{13}(x_1, x_3) + f_{23}(x_2, x_3) + f_{123}(x_1, x_2, x_3), \tag{5.3.1}$$

where

$$\begin{aligned} f_0 &= \mathrm{E}[f(\mathbf{x})] \\ &= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\ &= \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \quad \text{with} \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \quad \text{and} \quad d(\mathbf{x}) = d_3(x_3) d_2(x_2) d_1(x_1), \end{aligned} \tag{5.3.2}$$

is the expected value for $f(\mathbf{x})$ and the functions $f_1(x_1), f_2(x_2), ..., f_{123}(x_1, x_2, x_3)$ are orthogonal. In the context of the present example, two functions *g* and *h* defined on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ are said to be orthogonal if

$$\begin{aligned} \langle g, h \rangle &= \int_{\mathcal{X}} g(\mathbf{x}) h(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} g(x_1, x_2, x_3) h(x_1, x_2, x_3) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\ &= 0, \end{aligned} \tag{5.3.3}$$

where $\langle g, h \rangle$ is the commonly used notation for the inner product of *g* and *h* defined by the preceding integrals.

As examples, $f_1(x_1), f_2(x_{12}),..., f_{123}(x_1, x_2, x_3)$ in Eq. (5.3.1) are defined by

$$\begin{aligned} f_1(x_1) &= \mathrm{E}[f(\mathbf{x} \mid x_1)] - f_0 \quad \text{for } x_1 \in \mathcal{X}_1 \\ &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f(\mathbf{x} \mid x_1) d_3(x_3) d_2(x_2) \mathrm{d}x_3 \mathrm{d}x_2 - f_0 \\ &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) d_2(x_2) \mathrm{d}x_3 \mathrm{d}x_2 - f_0, \end{aligned} \tag{5.3.4}$$

$$\begin{aligned} f_{12}(x_1, x_2) &= \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_1(x_1) - f_2(x_2) - f_0 \quad \text{for } (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \\ &= \int_{\mathcal{X}_3} f(\mathbf{x} \mid x_1, x_2) d_3(x_3) \mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0 \\ &= \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0, \end{aligned} \tag{5.3.5}$$

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2, x_3] - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) \\ &\quad - f_{12}(x_1, x_2) - f_{13}(x_1, x_3) - f_{23}(x_2, x_3) \\ &\quad \text{for } (x_1, x_2, x_3) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \\ &= f(x_1, x_2, x_3) - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) \\ &\quad - f_{12}(x_1, x_2) - f_{13}(x_1, x_3) - f_{23}(x_2, x_3). \end{aligned} \tag{5.3.6}$$

and the remaining functions in Eq. (5.3.1) are defined similarly. In addition to being orthogonal as can be shown by evaluating the integrals defined in Eq. (5.3.3) for pairs of the functions $f_1(x_1), f_2(x_2),..., f_{123}(x_1, x_2, x_3)$, these functions also have the two following important properties: (i) their use in Eq. (5.3.1) results in the indicated equality to $f(\mathbf{x})$ as can be shown by simply summing these functions, and (ii) expected values of

$$E[f_1(x_1)] = E[f_2(x_{12})] = \cdots = E[f_{123}(x_1, x_2, x_3)] = 0 \tag{5.3.7}$$

hold as can be shown by evaluation of the integrals that define the indicated expected values.

Given the preceding properties of the functions $f_1(x_1), f_2(x_2),..., f_{123}(x_1, x_2, x_3)$, the following fundamental property of Sobol' variance decomposition can be established with

$$[g_1, g_2,..., g_7] = [f_1(x_1), f_2(x_2),..., f_{123}(x_1, x_2, x_3)] \tag{5.3.8}$$

used for notational compactness:

$$
\begin{aligned}
\mathrm{V}(f) &=_1 \int_{\mathcal{X}} [f(\mathbf{x}) - f_0]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&=_2 \langle f - f_0, f - f_0 \rangle \\
&=_3 \left\langle \sum_{i=1}^{7} g_i, \sum_{i=1}^{7} g_i \right\rangle \\
&=_4 \sum_{i=1}^{7}\sum_{j=1}^{7} \langle g_i, g_j \rangle \\
&=_5 \sum_{i=1}^{7} \langle g_i, g_i \rangle \ \text{ from Eq. (3.24)} \\
&=_6 \sum_{i=1}^{7} \langle g_i - \mathrm{E}(g_i), g_i - \mathrm{E}(g_i) \rangle \\
&=_7 \mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23}) + \mathrm{V}(f_{123}),
\end{aligned}
\tag{5.3.9}
$$

with (i) Equality 1 following from the definition of variance, (ii) Equality 2 following from the definition of inner product in Eq. (5.3.3),(iii) Equality 3 following from Eqs. (5.3.1) and (5.3.8), (iv) Equality 4 following from the properties of an inner product, (v) Equality 5 following from the orthogonality of the functions $f_1(x_1), f_2(x_2), ..., f_{123}(x_1, x_2, x_3)$, (vi) Equality 6 following from Eqs.(5.3.7) and (5.3.8), and (iv) Equality 7 following from Eqs. (5.3.8) and the definitions of variance for the functions $f_1(x_1), f_2(x_2), ..., f_{123}(x_1, x_2, x_3)$.

Division of both sides of Eq. (5.3.9) by V(*f*) yields

$$1 = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123}, \tag{5.3.10}$$

with

$$S_i = \mathrm{V}(f_i) / \mathrm{V}(f) \ \text{ for } \ i = 1, 2, 3, \tag{5.3.11}$$

$$S_{ij} = \mathrm{V}(f_{ij}) / \mathrm{V}(f) \ \text{ for } \ 1 \le i < j \le 3, \tag{5.3.12}$$

$$S_{123} = \mathrm{V}(f_{123}) / \mathrm{V}(f). \tag{5.3.13}$$

The terms on the right-hand side of Eq. (5.3.10) equal the fractions of the variance V(*f*) of $f$ that can be accounted for by $f_1, f_2, ..., f_{123}$ and can be used as measures of sensitivity. An additional measure of sensitivity is given by

$$
\begin{aligned}
S_{1T} &= [\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{123})] / \mathrm{V}(f) \\
&= S_1 + S_{12} + S_{13} + S_{123},
\end{aligned}
\tag{5.3.14}
$$

which measures the effect on V($f$) of $x_1$ and all interactions of $x_1$ with $x_2$ and $x_3$. Corresponding measures $S_{2T}$ and $S_{3T}$ for $x_2$ and $x_3$ are defined analogously. However, there is overlap in the effects measured by $S_{1T}$, $S_{2T}$ and $S_{3T}$.

Completing an analysis for the sensitivity measures in Eqs. (5.3.11)-(5.3.14) requires representing the integrals that define the variances $\mathrm{V}(f_1), \mathrm{V}(f_2), ..., \mathrm{V}(f_{123})$ in configurations that are amenable to numerical approximation. Without going into details which are available in App. B, (i) the following representations

$$\begin{aligned} \mathrm{V}(f_1) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1]\} \\ &= \int_{\mathcal{X}_1} \left[ \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f(x_1, x_2, x_3 \mid x_1) d_3(x_3) d_2(x_2) \mathrm{d}x_3 \mathrm{d}x_2 \right]^2 d(x_1) \mathrm{d}x_1 - f_0^2, \end{aligned} \tag{5.3.15}$$

$$\begin{aligned} \mathrm{V}(f_{12}) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\ &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \left[ \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 \right]^2 d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 \\ &\quad - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2), \end{aligned} \tag{5.3.16}$$

$$\mathrm{V}(f_{123}) = \mathrm{V}(f) - \{\mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23})\}. \tag{5.3.17}$$

exist for $\mathrm{V}(f_1)$, $\mathrm{V}(f_{12})$ and $\mathrm{V}(f_{123})$, and (ii) similar representations exist for $\mathrm{V}(f_2)$, $\mathrm{V}(f_3)$, $\mathrm{V}(f_{13})$ and $\mathrm{V}(f_{23})$. Due to the complexity of the function (i.e., model) $f$ in most analyses, the integrals defining E($f$), V($f$) and $\mathrm{V}(f_1), \mathrm{V}(f_2), ..., \mathrm{V}(f_{123})$ are usually approximated with sampling-based procedures.

The material contained in this section on uncertainty and sensitivity procedures based on Sobol' variance decomposition is presented in more detail for functions of the form $f(x_1, x_2, x_3)$ and $f(x_1, x_2, ..., x_n)$ in App. B. Further, additional information on these procedures is available in a number of publications (e.g., [30-39]).

## 5.4 Polynomial Chaos Variance Decomposition

The polynomial chaos approach to uncertainty and sensitivity analysis is based on the use of sets of orthonormal polynomials to obtain functional representations that can be used to obtain variance decompositions. As a descriptive example, a function $f(\mathbf{x})$ is used, with (i) $\mathbf{x}=[x_1,x_2,x_3]$, (ii) the epistemically uncertain variables $x_1$, $x_2$ and $x_3$ assumed to be independent, and (iii) the uncertainty in $x_1$, $x_2$ and $x_3$ characterized by the probability spaces $(\mathcal{X}_1,\mathbb{X}_1,m_1)$, $(\mathcal{X}_2,\mathbb{X}_2,m_2)$ and $(\mathcal{X}_3,\mathbb{X}_3,m_3)$ with $\mathcal{X}_i=[\underline{x}_i,\overline{x}_i]$ and associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$. Further, each uncertain variable $x_i, i=1,2,3,$ is assumed to have an associated set

$$\mathcal{P}_i=\{p_{id}(x_i): p_{id}(x_i) \text{ is a polynomial of degree } d \text{ for } x_i\in\mathcal{X}_i \text{ and } d=0,1,2,...\} \tag{5.4.1}$$

of orthonormal polynomials with inner products defined by

$$\begin{aligned}\left\langle p_{ir}(x_i), p_{is}(x_i)\right\rangle &= \int_{\mathcal{X}_i} p_{ir}(x_i)p_{is}(x_i)d_i(x_i)\mathrm{d}x_i \quad \text{for } \mathcal{I}=\{0,1,2,3,....\} \text{ and } (r,s)\in\mathcal{I}\times\mathcal{I} \\ &= \begin{cases}1 & \text{if } r=s \\ 0 & \text{if } r\neq s.\end{cases}\end{aligned} \tag{5.4.2}$$

As a reminder, the second equality in Eq. (5.4.2) is simply a statement of what it means for the polynomials in the set $\mathcal{P}_i$ to be orthonormal with respect to the inner product defined in the first equality in Eq. (5.4.2). Examples of possible definitions for the set $\mathcal{P}_i$ are shown in conjunction with Tables C.1 and C.2 of App. C.

As a consequence of the orthonormality of the polynomials contained in $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$, the multivariable polynomials contained in the set

$$\begin{aligned}\mathcal{P}=\{p_{rst}(\mathbf{x}): p_{rst}(\mathbf{x})=p_{1r}(x_1)p_{2s}(x_2)p_{3t}(x_3) \text{ for } p_{1r}(x_1)\in\mathcal{P}_1, p_{2s}(x_2)\in\mathcal{P}_2,\\ p_{3t}(x_3)\in\mathcal{P}_3, \text{ and } (r,s,t)\in\mathcal{I}_3=\mathcal{I}\times\mathcal{I}\times\mathcal{I}\}\end{aligned} \tag{5.4.3}$$

are also orthonormal for the inner product

$$\begin{aligned}
\left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle &= \int_{\mathcal{X}} p_{rst}(\mathbf{x}) p_{uvw}(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \quad \text{for } (r,s,t), (u,v,w) \in \mathcal{I}_3 = \mathcal{I} \times \mathcal{I} \times \mathcal{I} \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} p_{1r}(x_1) p_{2s}(x_2) p_{3t}(x_3) \times p_{1u}(x_1) p_{2v}(x_2) p_{3w}(x_3) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} p_{1r}(x_1) p_{1u}(x_1) d_1(x_1) \mathrm{d}x_1 \times \int_{\mathcal{X}_2} p_{2s}(x_2) p_{2v}(x_2) d_2(x_2) \mathrm{d}x_2 \\
&\quad \times \int_{\mathcal{X}_3} p_{3t}(x_3) p_{3w}(x_3) d_3(x_3) \mathrm{d}x_3 \\
&= \begin{cases} 1 \times 1 \times 1 = 1 \text{ if } p_{rst}(\mathbf{x}) = p_{uvw}(\mathbf{x}), \text{ i.e.; all integrals = 1 from normality} \\ 0 \text{ if } p_{rst}(\mathbf{x}) \neq p_{uvw}(\mathbf{x}), \text{ i.e., at least one integral = 0 from orthogonality.} \end{cases}
\end{aligned} \tag{5.4.4}$$

The last equality in Eq. (5.4.4) establishes that the multivariable polynomials contained in the set $\mathcal{P}$ defined in Eq. (5.4.3) are orthonormal with respect to the indicated inner product.

The core idea in the polynomial chaos approach to variance decomposition is to represent $f(\mathbf{x})$ by

$$\begin{aligned}
f(\mathbf{x}) &= \sum_{(r,s,t) \in \mathcal{I}^3} c_{rst} p_{rst}(\mathbf{x}) \quad \text{for } \mathcal{I}^3 = \mathcal{I} \times \mathcal{I} \times \mathcal{I} \\
&= \sum_{(r,s,t) \in \mathcal{I}^3} c_{rst} p_{1r}(x_1) p_{2s}(x_2) p_{3t}(x_3) \\
&= f_0 + \sum_{(r,s,t) \neq (0,0,0) \in \mathcal{I}^3} c_{rst} p_{1r}(x_1) p_{2s}(x_2) p_{3t}(x_3),
\end{aligned} \tag{5.4.5}$$

where (i) $c_{rst}$ is an appropriately determined coefficient as indicated in Eqs. (5.4.8) - (5.4.10), and (ii) the summation notation $(r,s,t) \neq (0,0,0) \in \mathcal{I}^3$ in the last equality indicates that the summation is over all $(r,s,t) \in \mathcal{I}^3$ except $(0,0,0)$. Specifically, the variance V($f$) of $f$ is given by

$$
\begin{aligned}
\mathrm{V}(f) &= \mathrm{E}\left(\left[f(\mathbf{x}) - \mathrm{E}(f)\right]^2\right) \\
&= \int_{\mathcal{X}} \left[f(\mathbf{x}) - f_0\right]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x})\mathrm{d}\mathbf{x} - f_0^2 \\
&= \left\langle f(\mathbf{x}), f(\mathbf{x}) \right\rangle - f_0^2 \\
&= \left\langle \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{rst}(\mathbf{x}), \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{rst}(\mathbf{x}) \right\rangle - f_0^2 \\
&= \sum_{(r,s,t)\in\mathcal{I}^3} \sum_{(u,v,w)\in\mathcal{I}^3} c_{rst} c_{uvw} \left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle - f_0^2 \\
&= \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}^2 - f_0^2 \\
&= \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} c_{rst}^2 ,
\end{aligned}
\tag{5.4.6}
$$

with (i) the next to last equality following from the equalities

$$
\left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle = \begin{cases} 1 & \text{for } (r,s,t) = (u,v,w) \\ 0 & \text{for } (r,s,t) \neq (u,v,w) \end{cases}
\tag{5.4.7}
$$

and (ii) the final equality following from the equality $c_{000} = f_0$ as shown in Eq. (5.4.9).

The coefficients $c_{rst}$ in Eqs. (5.4.5) and (5.4.6) are defined by

$$
\begin{aligned}
\left\langle p_{rst}(\mathbf{x}), f(\mathbf{x}) \right\rangle &= \int_{\mathcal{X}} p_{rst}(\mathbf{x}) f(\mathbf{x}) d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} p_{rst}(\mathbf{x}) \left( \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw}\, p_{uvw}(\mathbf{x}) \right) d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw} \int_{\mathcal{X}} p_{rst}(\mathbf{x}) p_{uvw}(\mathbf{x}) d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw} \left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle \\
&= c_{rst} \left\langle p_{rst}(\mathbf{x}), p_{rst}(\mathbf{x}) \right\rangle \\
&= c_{rst} ,
\end{aligned}
\tag{5.4.8}
$$

with the final equality following from the equalities in Eq. (5.4.7). For $(r,s,t)=(0,0,0)$, the corresponding orthonormal polynomial of degree 0 is $p_{000}(\mathbf{x})=1$ with an associated coefficient $c_{000}$ defined by

$$c_{000}=\left\langle p_{000}(\mathbf{x}),f(\mathbf{x})\right\rangle=\int_{\mathcal{X}}p_{000}(\mathbf{x})f(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x}=\int_{\mathcal{X}}f(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x}=\mathrm{E}[f(\mathbf{x})]=f_0. \tag{5.4.9}$$

As shown in Eq. (5.4.8), the remaining coefficients are defined by

$$c_{rst}=\int_{\mathcal{X}}p_{rst}(\mathbf{x})f(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{5.4.10}$$

with the indicated integral requiring numerical evaluation with quadrature or sampling-based methods. An alternate evaluation procedure based on regression analysis techniques is described in Sect. C.7 of App. C.

The variances defined in Eqs. (5.3.15) -(5.3.17) for $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ can also be represented by appropriate sums of subsets of the squared coefficients $c_{rst}^2$ in the representation for V(*f*) in Eq. (5.4.6). The representation for $\mathrm{V}(f_1)$ is considered first. To start, $f_1(x_1)$ can be represented by

$$f_1(x_1)=\sum_{r=0}^{\infty}c_r p_{1r}(x_1)=\sum_{r=1}^{\infty}c_r p_{1r}(x_1)=\sum_{r=1}^{\infty}c_{r00}p_{1r}(x_1), \tag{5.4.11}$$

where (i) $p_{1r}(x_1)$, $r=0,1,2,3,...,$ correspond to the set of orthonormal polynomials defined in conjunction with Eq. (5.4.1) for the variable $x_1$ and (ii) *r* rather than *d* is used to represent the degree of the polynomial $p_{1r}(x_1)$ to match with notational use of *r*, *s* and *t* to represent the degrees of the polynomials $p_{1r}(x_1)$, $p_{2s}(x_2)$ and $p_{3t}(x_3)$ in Eq. (5.4.5) and elsewhere. Specifically,

$$
\begin{aligned}
c_r &= \left\langle p_{1r}(x_1), f_1(x_1) \right\rangle \\
&= \int_{\mathcal{X}_1} p_{1r}(x_1) f_1(x_1) d_1(x_1) \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} p_{1r}(x_1) \left[ \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f_1(x_1, x_2, x_3) d_3(x_3) d_2(x_2) \mathrm{d}x_3 \mathrm{d}x_2 - f_0 \right] d_1(x_1) \mathrm{d}x_1 \\
&= \int_{\mathcal{X}} p_{1r}(x_1) f_1(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0 \int_{\mathcal{X}} p_{1r}(x_1) d_1(x_1) \mathrm{d}x_1 \\
&= \begin{cases} 0 \text{ for } r = 0 \text{ from } f_0 - f_0 = 0 \\ \int_{\mathcal{X}} p_{1r}(x_1) f_1(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \text{ for } r > 0 \text{ from } \int_{\mathcal{X}} p_{1r}(x_1) d_1(x_1) \mathrm{d}x_1 = 0 \end{cases} \\
&= c_{r00} \text{ in the notation of Eq. (5.4.5).}
\end{aligned}
\tag{5.4.12}
$$

In turn,

$$
\begin{aligned}
\mathrm{V}(f_1) &= \left\langle f_1 - \mathrm{E}(f_1), f_1 - \mathrm{E}(f_1) \right\rangle \\
&= \left\langle \sum_{r=1}^{\infty} c_r p_{1r}(x_1), \sum_{r=1}^{\infty} c_r p_{1r}(x_1) \right\rangle \text{ from Eq. (5.4.11) and } \mathrm{E}(f_1) = 0 \\
&= \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_r c_s \left\langle p_{1r}(x_1), p_{1s}(x_1) \right\rangle \\
&= \sum_{r=1}^{\infty} c_r^2 \text{ from orthonormality} \\
&= \sum_{r=1}^{\infty} c_{r00}^2,
\end{aligned}
\tag{5.4.13}
$$

with the final equality following from Eq. (5.4.12). Thus, the variance $\mathrm{V}(f_1)$ of $f_1(x_1)$ can be obtained from a sum of squared coefficients obtained in a polynomial expansion of $f(x_1, x_2, x_3)$.

Similarly,

$$
f_2(x_2) = \sum_{s=1}^{\infty} c_{0s0} p_{2s}(x_2) \text{ and } \mathrm{V}(f_2) = \sum_{s=1}^{\infty} c_{0s0}^2, \tag{5.4.14}
$$

$$
f_3(x_3) = \sum_{t=1}^{\infty} c_{00t} p_{3t}(x_3) \text{ and } \mathrm{V}(f_3) = \sum_{t=1}^{\infty} c_{00t}^2, \tag{5.4.15}
$$

$$
f_{12}(x_1, x_2) = \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_{rs0} p_{1r}(x_1) p_{2s}(x_2) \text{ and } \mathrm{V}(f_{12}) = \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_{rs0}^2, \tag{5.4.16}
$$

$$f_{13}(x_1,x_3)=\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}p_{1r}(x_1)p_{3t}(x_3)\ \text{ and }\ \mathrm{V}(f_{13})=\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}^2, \tag{5.4.17}$$

$$f_{23}(x_2,x_3)=\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}p_{2s}(x_2)p_{3t}(x_3)\ \text{ and }\ \mathrm{V}(f_{23})=\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}^2, \tag{5.4.18}$$

and

$$f_{123}(x_1,x_2,x_3)=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\sum_{t=0}^{\infty}c_{rst}p_{1r}(x_1)p_{2s}(x_2)p_{3t}(x_3)\ \text{ and }\ \mathrm{V}(f_{123})=\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{rst}^2. \tag{5.4.19}$$

The polynomial chaos variance decomposition of $f(x_1,x_2,x_3)$ is based on the equality

$$\begin{aligned}\mathrm{V}(f)&=_1\sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3}c_{rst}^2\\&=_2\sum_{r=1}^{\infty}c_{r00}^2+\sum_{s=1}^{\infty}c_{0s0}^2+\sum_{t=1}^{\infty}c_{00t}^2+\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}c_{rs0}^2\\&\quad+\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}^2+\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}^2+\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{rst}^2\\&=_3\mathrm{V}(f_1)+\mathrm{V}(f_2)+\mathrm{V}(f_3)+\mathrm{V}(f_{12})+\mathrm{V}(f_{13})+\mathrm{V}(f_{23})+\mathrm{V}(f_{123}),\end{aligned} \tag{5.4.20}$$

with (i) Equality 1 obtained from Eq. (5.4.6), (ii) Equality 2 obtained by a rearrangement of the summation in Equality 1, and (iii) Equality 3 obtained from the variances defined in Eqs. (5.4.13) -(5.4.19). In turn, as previously discussed in connection with Eqs. (5.3.10)-(5.3.14), division of both sides of Eq. (5.4.20) by V(*f*) yields

$$1=S_1+S_2+S_3+S_{12}+S_{13}+S_{23}+S_{123}, \tag{5.4.21}$$

with

$$\begin{aligned}S_i&=\mathrm{V}(f_i)/\mathrm{V}(f)\ \text{ for }\ i=1,2,3\\&=\begin{cases}\sum_{r=1}^{\infty}c_{r00}^2/\mathrm{V}(f)\ \text{ for }\ i=1\\\sum_{s=1}^{\infty}c_{0s0}^2/\mathrm{V}(f)\ \text{ for }\ i=2\\\sum_{t=1}^{\infty}c_{00t}^2/\mathrm{V}(f)\ \text{ for }\ i=3,\end{cases}\end{aligned} \tag{5.4.22}$$

$$
\begin{aligned}
S_{ij} &= \mathrm{V}(f_{ij}) / \mathrm{V}(f) \quad \text{for} \quad 1 \le i < j \le 3 \\
&= \begin{cases}
\sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_{rs0}^2 / \mathrm{V}(f) & \text{for} \quad (i,j) = (1,2) \\
\sum_{r=1}^{\infty} \sum_{t=1}^{\infty} c_{r0t}^2 / \mathrm{V}(f) & \text{for} \quad (i,j) = (1,3) \\
\sum_{s=1}^{\infty} \sum_{t=1}^{\infty} c_{0st}^2 / \mathrm{V}(f) & \text{for} \quad (i,j) = (2,3),
\end{cases}
\end{aligned} \tag{5.4.23}
$$

$$
S_{123} = \mathrm{V}(f_{123}) / \mathrm{V}(f) = \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} c_{rst}^2 / \mathrm{V}(f) \tag{5.4.24}
$$

and

$$
\mathrm{V}(f) = \sum_{(r,s,t) \neq (0,0,0) \in \mathcal{I}^3} c_{rst}^2 \tag{5.4.25}
$$

as previously indicated in Eq. (5.4.6). The terms on the right-hand side of Eq. (5.3.10) equal the fractions of the variance V($f$) of $f$ that can be accounted for by $f_1, f_2, ..., f_{123}$ and can be used as measures of sensitivity. An additional measure of sensitivity is given by

$$
\begin{aligned}
S_{1T} &= [\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{123})] / \mathrm{V}(f) \\
&= S_1 + S_{12} + S_{13} + S_{123},
\end{aligned} \tag{5.4.26}
$$

which measures the effect on V($f$) of $x_1$ and all interactions of $x_1$ with $x_2$ and $x_3$. Corresponding measures $S_{2T}$ and $S_{3T}$ for $x_2$ and $x_3$ are defined analogously.

In computational practice, the infinite series of coefficients in Eqs. (5.4.22)-(5.4.25) used in the determination of the indicated sensitivity measures must be truncated as discussed in Sect. C.7 of App. C. Once series are truncated, the remaining coefficients must be approximated by either numerically evaluating their defining integrals (e.g., with sampling-based procedures) or with a regression-based procedure described in Sect. C.7 of App. C.

The material contained in this section on uncertainty and sensitivity procedures based on polynomial chaos expansions of functions is presented in more detail for functions of the form $f(x_1, x_2, x_3)$ and $f(x_1, x_2, ..., x_n)$ in App. C. Further, additional information on these procedures is available in a number of publications (e.g., [40-56]).

## 5.5 Fourier Amplitude Sensitivity Test (FAST)

The FAST approach to uncertainty and sensitivity analysis is based on performing a numerical calculation to obtain the expected value and variance of a model prediction. The basis of this calculation is a transformation that converts a multidimensional integral over all the uncertain model inputs to a one-dimensional integral. Further, a decomposition of the Fourier series representation of the model is used to obtain the fractional contributions of the individual input variables to the variance of the model prediction. As before, the model under consideration is assumed to be a function of the form shown in Eq. (5.2.1).

An analysis based on the FAST approach involves four steps. In the first step, ranges and distributions are developed for the elements $x_j$ of the vector

$$\mathbf{x} = [x_1, x_2, ..., x_n] \tag{5.5.1}$$

of uncertain model inputs. The outcome of this step is a set $\mathcal{X}$ of possible values for $\mathbf{x}$ and a density function

$$d(\mathbf{x}) = \prod_{i=1}^{n} d_i(x_i) \tag{5.5.2}$$

that characterizes the distribution of the possible values for $\mathbf{x}$ within $\mathcal{X}$, where $d_i(x_i)$ is the density function for the distribution characterizing the epistemic uncertainty associated with $x_i$.

Once $d(\mathbf{x})$ and $\mathcal{X}$ are known, the expected value and variance for $y = f(x_1, x_2, ..., x_n) = f(\mathbf{x})$ are defined by

$$E(y) = \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{5.5.3}$$

and

$$V(y) = \int_{\mathcal{X}} [f(\mathbf{x}) - E(y)]^2 d(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{5.5.4}$$

The preceding integrals are multidimensional and, in real analyses, can potentially be difficult to evaluate. This leads to the second step in the FAST approach.

In the second step, the multidimensional integral in Eq. (5.5.3) is converted to a one dimensional integral by the definition of a suitable space filling curve in $\mathcal{X}$ (i.e., a search curve that comes close to all variable combinations in $\mathcal{X}$). The indicated curve is formally defined by space

filling curves for the individual variables $x_i$ and their associated density functions $d_i(x_i)$ and sample spaces $\mathcal{X}_i$. Specifically, the space filling curve for $x_i$ is defined by

$$x_i(s) = g_i[u_i(s)] \text{ with } u_i(s) = G_i[\sin(\omega_i s)], \tag{5.5.5}$$

with (i) $u_i(s) = G_i[\sin(\omega_i s)]$ defined to provide a systematic pattern of movements that are consistent with the distribution characterized by the density function $d_i(x_i)$ and associated sample space $\mathcal{X}_i$ (i.e., a curve that comes close to each value for $x_i$ in $\mathcal{X}_i$), (ii) $\omega_i$ an appropriately chosen integer, and (iii) $g_i[u_i(s)]$ defined to assure that the movements defined by $u_i(s) = G_i[\sin(\omega_i s)]$ match with the values for $x_i$ in the sample space $\mathcal{X}_i$ for $0 \le s \le 2\pi$.

As developed in Ref. [57] and elaborated on in conjunction with Eq. (2.6) of Ref. [58], the function $G_i$ can be defined by the solution of the differential equation

$$\pi(1-x^2)^{1/2} d_i(G_i) \frac{\mathrm{d}G_i(x)}{\mathrm{d}x} = 1 \text{ with } G_i(0) = 0. \tag{5.5.6}$$

With a suitable definition of the integer $\omega_i$, $u_i(s) = G_i[\sin(\omega_i s)]$ can be used in the definition a search curve for $\mathcal{X}_i$ and the associated density function $d_i(x_i)$ for $0 \le s \le 2\pi$.

As a simple example, if $x_i$ has a uniform distribution on $\mathcal{X}_i = [a,b]$, then

$$d_i(x_i) = 1/(b-a) \text{ and } G_i(x_i) = (b-a)\sin^{-1}(x_i)/\pi \tag{5.5.7}$$

In turn, because the range of $\sin^{-1}[\sin(\omega_i s)]$ is $[-\pi/2, \pi/2]$, the range of

$$G_i[\sin(\omega_i s)] = (b-a)\sin^{-1}[\sin(\omega_i s)]/\pi \text{ is } [-(b-a)/2, (b-a)/2]. \tag{5.5.8}$$

Thus, to have $G_i[\sin(\omega_i s)]$ used to define a path that covers the range $\mathcal{X}_i = [a,b]$ of $x_i$, the function

$$g_i[u_i(s)] = g_i\{G_i[\sin(\omega_i s)]\} \tag{5.5.9}$$

introduced in Eq. (5.5.5) is defined by

$$g_i\{G_i[\sin(\omega_i s)]\} = \frac{a+b}{2} + G_i[\sin(\omega_i s)] = \frac{a+b}{2} + (b-a)\sin^{-1}[\sin(\omega_i s)]/\pi. \tag{5.5.10}$$

As a special case,

$$g_i\{G_i[\sin(\omega_i s)]\} = \frac{1}{2} + G_i[\sin(\omega_i s)] = \frac{1}{2} + \sin^{-1}[\sin(\omega_i s)]/\pi \tag{5.5.11}$$

if the density function $d_i(x_i)$ corresponds to a uniform distribution on [0, 1].

The following is an interesting possibility that merits investigation for the case when $d_i(x_i)$ corresponds to an arbitrary distribution (e.g., normal, triangular, log uniform, …) with an associated cumulative distribution function $CDF_i(x_i)$. Then, it may be possible to define $g_i$ and $G_i$ by

$$\begin{aligned} x_i(s) &= g_i\{G_i[\sin(\omega_i s)]\} \\ &= CDF_i^{-1}\{G_i[\sin(\omega_i s)]\} \\ &= CDF_i^{-1}\left\{\frac{1}{2} + \sin^{-1}[\sin(\omega_i s)]/\pi\right\} \end{aligned} \tag{5.5.12}$$

with $G_i$ defined for a uniform distribution on [0, 1] as indicated in Eq. (5.5.11). If this is a mathematically correct procedure, its use may be simpler than using Eq. (5.5.6) in conjunction with the density function $d_i(x_i)$ to define $g_i$ and $G_i$.

The integers $\omega_i, i = 1,2,...,n,$ need to be selected so that there is effectively little relationship (i.e., interference) between the individual curves generated with use of the integers $\omega_i$. To avoid the indicated interference, the $\omega_i$'s need to be defined so that an individual $\omega_i$ cannot be closely approximated by a linear combination of other $\omega_i$'s (e.g., $\omega_3 \cong \omega_1 + \omega_2$ or possibly $\omega_3 = \omega_1 + \omega_2$). An arbitrary sequence $\alpha_1, \alpha_2, ..., \alpha_m$ of integers is said to be of order *M*, an integer, if

$$\sum_{i=1}^{m} r_i \alpha_i \neq 0 \text{ for } r_1,\ r_2, ...,\ r_m \text{ integers and } \sum_{i=1}^{m} |\, r_i\,| \leq M + 1. \tag{5.5.13}$$

Further, a sequence of integers $\omega_i, i = 1,2,...,n,$ is said to be free of, or to avoid, interferences through order *M* provided every subset $\{\alpha_1, \alpha_2, ..., \alpha_m\}$ of $\{\omega_1, \omega_2, ..., \omega_n\}$ of size $m \leq M$ is of order *M*. Definition of $\omega_1, \omega_2, ..., \omega_n$ to be free of interferences through an appropriately chosen order *M* effectively eliminates interference between the individual curves generated with use of the integers $\omega_i$.

The definition of integer sequences $\omega_1, \omega_2, ..., \omega_n$ that are free of interferences through order *M* is discussed in Refs.[59; 60]. In addition, Table 1 of Ref. [59] provides examples of integer

sequences $\omega_1, \omega_2, ..., \omega_n$ for $n = 5, 6, ..., 19$ that are free of interferences through order $M = 4$; and Eq. (5.1) and Table VI of Ref. [60] define and illustrate a procedure to construct integer sequences $\omega_1, \omega_2, ..., \omega_n$ for $n = 3, 4, ..., 50$ that are free of interferences through order $M = 4$.

The ultimate outcome of Step 2 is a sequence $\omega_i, i = 1, 2, ..., n,$ integers and a corresponding sequence $g_i[G_i(\sin \omega_i s)], i = 1, 2, ..., n,$ of functions such that

$$\begin{aligned} E(y) &= \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &\cong \frac{1}{2\pi} \int_0^{2\pi} f\left[g_1[G_1(\sin \omega_1 s)], g_2[G_2(\sin \omega_2 s)], ..., g_n[G_n(\sin \omega_n s)]\right] \mathrm{d}s \\ &= \hat{E}(y). \end{aligned} \tag{5.5.14}$$

The second integral, which is one dimensional, is easier to evaluate than the first integral, which is multidimensional. In essence, construction of the $\omega_i$ and $G_i$ has converted $f$ from a function of $n$ variables (i.e., the elements $x_i$ of $\mathbf{x}$) to function of one variable (i.e., $s$).

In the third step, the one-dimensional representation for $f$ is used to estimate the expected value and variance of $y$. This provides a representation of the uncertainty in $y$ that results from the specified distributions for the $x_i$. As already indicated in Eq. (5.5.14), the expected value for $y$ can be approximated by

$$\hat{E}(y) = \frac{1}{2\pi} \int_0^{2\pi} f\left[g_1[G_1(\sin \omega_1 s)], g_2[G_2(\sin \omega_2 s)], ..., g_n[G_n(\sin \omega_n s)]\right] \mathrm{d}s \tag{5.5.15}$$

In turn, the variance of $y$ can be approximated by

$$V(y) \cong \frac{1}{2\pi} \int_0^{2\pi} \left\{ f\left[g_1[G_1(\sin \omega_1 s)], g_2[G_2(\sin \omega_2 s)], ..., g_n[G_n(\sin \omega_n s)]\right] \right\}^2 \mathrm{d}s - \hat{E}^2(y). \tag{5.5.16}$$

In practice, a numerical procedure must be written to evaluate the integrals in the approximations to $E(y)$ and $V(y)$. Further, this procedure will require values for $f$, which come from evaluating the model that predicts $y$.

In the fourth and final step, the sensitivity of $y$ to the individual $x_i$ is determined on the basis of fractional contribution to the variance of $y$. By using properties of the Fourier series representation for $f$, it can be shown that

$$V(y) \cong \sum_{k=1}^{\infty} (A_k^2 + B_k^2), \tag{5.5.17}$$

where

$$A_k = \frac{1}{\pi}\int_0^{2\pi} f\left[g_1[G_1(\sin \omega_1 s)], g_2[G_2(\sin \omega_2 s)], ..., g_n[G_n(\sin \omega_n s)]\right]\cos(ks)\mathrm{d}s \qquad (5.5.18)$$

and

$$B_k = \frac{1}{\pi}\int_0^{2\pi} f\left[g_1[G_1(\sin \omega_1 s)], g_2[G_2(\sin \omega_2 s)], ..., g_n[G_n(\sin \omega_n s)]\right]\sin(ks)\mathrm{d}s. \qquad (5.5.19)$$

Further, it can also be shown that the part of the variance of $y$ that derives from $x_i$ can be approximated by

$$V_i(y) \cong \sum_{k=1}^{\infty}(A_{k\omega_i}^2 + B_{k\omega_i}^2), \qquad (5.5.20)$$

where $\omega_i$ is the integer associated with $G_i$ in the conversion from a multidimensional integral to a one dimensional integral in Eq. (5.5.14) as described in Ref. [58]. Then, the ratio

$$V_i(y)/V(y) \qquad (5.5.21)$$

gives the fractional contribution of $x_i$ to the variance of $y$. This ratio can be used to rank the importance of the individual $x_i$ with respect to their impact on the uncertainty in $y$. In practice, the integrals that define $A_k$ and $B_k$ must be approximated numerically. Further, the series in Eqs. (5.5.17) and (5.5.20) must be truncated.

Additional information on the FAST approach to uncertainty and sensitivity analysis is available in Refs. [34; 57-63].

## 5.6 References: Chapter 5

1. Helton J.C. 1993. Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal. *Reliability Engineering and System Safety* 42:327-367.
2. Tukey J.W. 1957. *The Propagation of Errors, Fluctuations and Tolerances: An Exercise in Partial Differentiation. Rpt. Report No. 12*, Statistical Techniques Research Group, Princeton University Press, Princeton, NJ.
3. Tukey J.W. 1957. *The Propagation of Errors, Fluctuations and Tolerances: Basic Generalized Formulas. Rpt. Report No. 10*, Statistical Techniques Research Group, Princeton University Press, Princeton, NJ.
4. Tukey J.W. 1957. *The Propagation of Errors, Fluctuations and Tolerances: Supplementary Formulas. Rpt. Report No. 11*, Statistical Techniques Research Group, Princeton University Press, Princeton, NJ.
5. Cacuci D.G., C.F. Weber, E.M. Oblow and J.H. Marable. 1980. Sensitivity Theory for General Systems of Nonlinear Equations. *Nuclear Science and Engineering* 75:88-110.
6. Cacuci D.G. 1981. Sensitivity Theory for Nonlinear Systems. I. Nonlinear Functional Analysis Approach. *Journal of Mathematical Physics* 22:2794-2802.
7. Cacuci D.G. 1981. Sensitivity Theory for Nonlinear Systems. II. Extensions to Additional Classes of Responses. *Journal of Mathematical Physics* 22:2803-2812.
8. Cacuci D.G. 2003. *Sensitivity and Uncertainty Analysis, Vol. 1: Theory*. Boca Raton, FL: Chapman and Hall/CRC Press.
9. Cacuci D.G., M. Ionescu-Bujor and I.M. Navon. 2005. *Sensitivity and Uncertainty Analysis, Vol. II, Application to Large-Scale Systems*. Boca Raton, FL: Chapman & Hall/CRC.
10. Cacuci D.G. and M. Ionescu-Bujor. 2000. Adjoint sensitivity analysis of the RELAP5/MOD3.2 two-fluid thermal-hydraulic code system - I: Theory. *Nuclear Science And Engineering* 136:59-84.
11. Hwang J.-T., E.P. Dougherty, S. Rabitz and H. Rabitz. 1978. The Green's Function Method of Sensitivity Analysis in Chemical Kinetics. *Journal of Chemical Physics* 69:5180-5191.
12. Dougherty E.P., J.T. Hwang and H. Rabitz. 1979. Further Developments and Applications of the Green's Function Method of Sensitivity Analysis in Chemical Kinetics. *Journal of Chemical Physics* 71:1794-1808.
13. Dougherty E.P. and H. Rabitz. 1979. A Computational Algorithm for the Green's Function Method of Sensitivity Analysis in Chemical Kinetics. *International Journal of Chemical Kinetics* 11:1237-1248.
14. Oblow E.M. 1985. GRESS: Gradient Enhanced Software System, Version D, User's Guide. *Rpt. ORNL/TM-9658*, Oak Ridge National Laboratory, Oak Ridge.
15. Oblow E.M., F.G. Pin and R.Q. Wright. 1986. Sensitivity Analysis Using Computer Calculus: A Nuclear Waste Isolation Example. *Nuclear Science and Engineering* 94:46-65.

16. Pin F.G., B.A. Worley, E.M. Oblow, R.Q. Wright and W.V. Harper. 1986. An Automated Sensitivity Analysis Procedure for the Performance Assessment of Nuclear Waste Isolation Systems. *Nuclear and Chemical Waste Management* 6:255-263.
17. Worley B.A. and J.E. Horwedel. 1986. A Waste Package Performance Assessment Code with Automated Sensitivity-Calculation Capability. *Rpt. ORNL/TM-9976*, Oak Ridge National Laboratory, Oak Ridge.
18. Berz M., C. Bischof, G. Corliss and A. Griewank. 1996. *Computational Differentiation: Techniques, Applications, and Tools*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
19. Helton J. and S. Stuckwisch. 1976. Numerical Approximation of Product Integrals. *Journal of Mathematical Analysis and Applications* 56:410-437.
20. Tomovic R. and M. Vukobratovic. 1972. *General Sensitivity Theory*. New York, NY: Elsevier.
21. Frank P.M. 1978. *Introduction to System Sensitivity Theory*. New York, NY: Academic Press.
22. Varma A., M. Morbidelli and H. Wu. 1999. *Parametric Sensitivity in Chemical Systems*. Cambridge: Cambridge University Press.
23. Ronen Y. 1988. *Uncertainty Analysis.* Boca Raton, FL: CRC Press, Inc.
24. Turányi T. 1990. Sensitivity Analysis of Complex Kinetic Systems. Tools and Applications. *Journal of Mathematical Chemistry* 5:203-248.
25. Turanyi T., Rabitz, H. 2000. Local Methods. In *Sensitivity Analysis*, ed. A Satelli, Chan, K., Scott, E. M.:81-99. Chichester, England: Wiley.
26. Griewank A., Wather, A. 2008. *Principles and Techniques of Algorithmic Differentiation*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
27. Griewank A. 2014. On automatic differentiation and algorithmic linearization. *Pesquisa Operacional* 34:621-645.
28. Nodet M. and A. Vidard. 2017. Variational Methods. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1123-1142: Springer International Publishing Switzerland.
29. Kucherenko S. and B. looss. 2017. Derivative-based global sensitivity measures. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1241-1263: Spinger Internaltional Publishing Switzerland.
30. Sobol' I.M. 1993. Sensitivity Analysis for Nonlinear Mathematical Models. *Mathematical Modeling and Computational Experiment* 1:407-414.
31. Homma T. and A. Saltelli. 1996. Importance Measures in Global Sensitivity Analysis of Nonlinear Models. *Reliability Engineering and System Safety* 52:1-17.
32. Sobol' I.M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* 55:271-280.
33. Sobol' I.M. 2003. Theorems and examples on high dimensional model representation. *Reliability Engineering and System Safety* 79:187-193.

34. Chan K., S. Tarantola and A. Saltelli. 2000. Variance-Based Methods. In *Sensitivity Analysis*, ed. A Saltelli, K Chan, EM Scott:167-197. New York, NY: Wiley.

35. Saltelli A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto and S. Tarantola. 2010. Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index. *Computer Physics Communications* 181:259-270.

36. Kucherenko S., S. Tarantola and P. Annoni. 2012. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications* 183:937-946.

37. Mara T.A. and S. Tarantola. 2012. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering and System Safety* 107:115-121.

38. Prieur C. and S. Tarantola. 2017. Variance-based sensitivity analysis: Theory and estimation algorithms. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1217-1239: Springer International Publishing Switzerland.

39. Kucherenko S. and S. Song. 2017. Different numerical estimators for main effect global sensitivity indices. *Reliability Engineering and System Safety* 165:222-238.

40. Crestaux T., O. Le Maître and J.M. Martinez. 2009. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering and System Safety* 94:1161-1172.

41. Sudret B. 2008. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety* 93:964-979.

42. Blatman G. and B. Sudret. 2010. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering and System Safety* 95:1216-1229.

43. Le Gratiet L.L., S. Marelli and B. Sudret. 2017. Metamodel-based sensitivity analysis: Polynomial chaos expansions and gaussian processes. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1289-1325: Springer International Publishing Switzerland.

44. Ghanem R. and J. Red-Horse. 2017. Polynomial chaos: Modeling, estimation, and approximation. In *Handbook of Uncertainty Quantification*, ed. R ghanem, Higdon, D., Owhadi, H.:521-551: Springer International Publishing Switzerland.

45. Soize C. and R. Ghanem. 2004. Physical Systems with Random Characterristics: Chaos Representations with Arbitrary Probability Measure. *SIAM J. Scientific Computing* 26:395-410.

46. Hadigol M. and A. Doostan. 2018. Least squares polynomial chaos expansion: A review of sampling strategies. *Computer Methods in Applied Mechanics and Engineering* 332:382-407.

47. Sudret B. and C.V. Mai. 2015. Computing derivative-based global sensitivity measures using polynomial chaos expansions. *Reliability Engineering and System Safety* 134:241-250.

48. Konakli K. and B. Sudret. 2016. Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions. *Journal of Computational Physics* 321:1144-1169.

49. Debusschere B. 2017. Intrusive polynomial chaos methods for forward uncertainty propagation. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:617-636: Springer International Publishing Switzerland.

50. Eldred M.S. 2009. Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design C3 - Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.

51. Burnaev E., I. Panin and B. Sudret. 2017. Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions. *Annals of Mathematics and Artificial Intelligence* 81:187-207.

52. Garcia-Cabrejo O. and A. Valocchi. 2014. Global Sensitivity Analysis for multivariate output using Polynomial Chaos Expansion. *Reliability Engineering and System Safety* 126:25-36.

53. Eldred M.S. and J. Burkardt. 2009. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification C3 - 47th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition.

54. Blatman G. and B. Sudret. 2011. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics* 230:2345-2367.

55. Jakeman J.D., F. Franzelin, A. Narayan, M. Eldred and D. Plfüger. 2019. Polynomial chaos expansions for dependent random variables. *Computer Methods in Applied Mechanics and Engineering* 351:643-666.

56. Jakeman J.D., M.S. Eldred and K. Sargsyan. 2015. Enhancing ℓ1-minimization estimates of polynomial chaos expansions using basis selection. *Journal of Computational Physics* 289:18-34.

57. Cukier R.I., C.M. Fortuin, K.E. Shuler, A.G. Petschek and J.H. Schaibly. 1973. Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients, I. Theory. *The Journal of Chemical Physics* 59:3873-3878.

58. Cukier R.I., H.B. Levine and K.E. Shuler. 1978. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics* 26:1-42.

59. Schaibly J.H. and K.E. Shuler. 1973. Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients, II. Applications. *The Journal of Chemical Physics* 59:3879-3888.

60. Cukier R.I., J.H. Schaibly and K.E. Shuler. 1975. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations. *The Journal of Chemical Physics* 63:1140-1149.

61. McRae G.J., J.W. Tilden and J.H. Seinfeld. 1982. Global Sensitivity Analysis - A Computational Implementation of the Fourier Amplitude Sensitivity Test (FAST). *Computers and Chemical Engineering* 6:15-25.

62. Saltelli A. and R. Bolado. 1998. An Alternative Way to Compute Fourier Amplitude Sensitivity Test (FAST). *Computational Statistics & Data Analysis* 26:445-460.

63. Saltelli A., S. Tarantola and K.P.S. Chan. 1999. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* 41:39-56.

# 6. SURROGATE MODELS

A surrogate model (sometimes called meta-model, emulator, or response surface model) is an inexpensive input-to-output mapping that can emulate the results of computationally expensive multiphysics model and numerical simulator. Once constructed, this meta-model is relatively inexpensive to evaluate so it is often used as a surrogate for the computational simulation model in uncertainty propagation, sensitivity analysis, or optimization problems that may require thousands to millions of function evaluations [1-4]. Surrogate models may be used in various ways as part of geologic disposal repository assessment: they may be used as surrogates for complex multi-physics numerical simulators (such as PFLOTRAN) and probabilistic PA simulators (such as *GDSA Framework*) to perform sensitivity analysis, such as the variance-based decomposition indices presented in Chapter 5 that require tens of thousands of function evaluations. Surrogates may also be used as less expensive replacements for constitutive models or materials models within larger codes, such as the preliminary work demonstrating surrogates for the Fuel Matrix Degradation Model (FMDM) within PFLOTRAN [1]. Finally, surrogates may be used in Bayesian calibration (parameter estimation), which is also a computationally expensive task.

There are many different types of surrogate models, including polynomial regression, Gaussian processes, neural networks, radial basis functions, splines, etc. The subsections below cover popular types of surrogate models. There are many good overview articles that compare various meta-model strategies [1-4;2]. Simpson et al. provide an excellent overview not just of various statistical meta-model methods but also approaches that use low-fidelity models as surrogates for high-fidelity models [1]. Haftka and his students developed an approach that uses ensembles of emulators or hybrid emulators [6]. Finally, polynomial chaos expansions (PCE) have become popular surrogate models over the past fifteen years [3;4]. These stochastic expansion methods approximate the functional dependence of the simulation response on uncertain model parameters by expansion in an orthogonal polynomial basis. The polynomials used are tailored to the characterization of the uncertain input variables.

## 6.1 Polynomial Regression

A linear regression model $\hat{f}$ as a function of an m-dimensional input vector $\boldsymbol{x} \in \Re^m$ is defined as:

$$\hat{f}(\boldsymbol{x}) \approx c_0 + \sum_{i=1}^{m} c_i \, x_i \tag{6.1.1}$$

Similarly, a second order polynomial regression (also called a quadratic regression model) is defined as:

$$\hat{f}(\boldsymbol{x}) \approx c_0 + \sum_{i=1}^{m} c_i \, x_i + \sum_{i=1}^{m} \sum_{j \geq i}^{m} c_{ij} \, x_i x_j \tag{6.1.2}$$

To determine the coefficients of the polynomial regression model, a least-squares formulation that minimizes the sum-of-squared error (SSE) between the surrogate model and the training data from the simulation is typically used [5]. The SSE is the standard error metric for overdetermined polynomial regression. It is a quadratic loss function which tends to find solutions near zero SSE

well. Assume that a matrix of $n$ training samples is available, where each training sample has an input $\boldsymbol{x}_i$ and a corresponding output $y_i$. The coefficients minimize the SSE:

$$SSE = \sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i\right)^2 \tag{6.1.3}$$

For general nonlinear regression problems, one needs to use optimization methods to find the vector of coefficients $\boldsymbol{c}$ which minimize the SSE. However, for linear regression models, the least squares problem reduces to a linear solve. If the entire sample matrix of inputs is denoted as $\boldsymbol{X}$ (of dimension $n$ x $m$) and the sample matrix of outputs as $\mathbf{y}$ (of dimension $n$ x 1), the optimization problem becomes:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{argmin} \, \|\boldsymbol{X} \cdot \boldsymbol{c} - \boldsymbol{y}\|^2 = \boldsymbol{X}^{-1}\boldsymbol{y} \tag{6.1.4}$$

In practice, one does not take the explicit inverse of the input sample matrix $\boldsymbol{X}^{-1}$ to solve for the optimal c but instead use a matrix factorization such as a QR factorization. This makes the determination of $\hat{\boldsymbol{c}}$ computationally efficient. Note also that this system is typically overdetermined meaning that the number of training points $n$ is greater than the number of regression coefficients $m$.

## 6.2 Gaussian Process

One popular surrogate approach is to develop an emulator that is a stationary smooth Gaussian process [2;3;6]. The popularity of Gaussian processes is due to their ability to model complicated functional forms and to provide an uncertainty estimate of their predicted response value at a new input point. A Gaussian process (GP) is defined as follows: A stochastic process is a collection of random variables $\{y(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbf{X}\}$ indexed by a set $\mathbf{X}$ (in most cases, $\mathbf{X}$ is $\Re^d$, where $d$ is the number of inputs and $\Re$ is the space of real numbers). The stochastic process is defined by giving the joint probability distribution for every finite subset of variables $y(\boldsymbol{x}_1)$, ... $y(\boldsymbol{x}_k)$. A Gaussian process is a stochastic process for which any finite set of Y-variables has a joint multivariate Gaussian distribution. A Gaussian process is fully specified by its mean function $\mu(\boldsymbol{x}) = E[y(\boldsymbol{x})]$ and its covariance function $K(y(\boldsymbol{x}), y(\boldsymbol{x}'))$. For an extensive tutorial on Gaussian process models, see [7].

The notation for defining a GP is as follows: $\boldsymbol{x}$ is one set of inputs of dimension $d$. There are $n$ samples, $\boldsymbol{x}_i$, for i=1…n. Each $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$. $\mathbf{X}$ denotes the ($d$*$n$) set of all samples and Y=y($\mathbf{X}$) is an $n$*$1$ vector of the responses from the computational simulation.

The basic steps for using a GP are:

1. Define the mean function. Often the mean is taken to be zero or a constant, but this is not necessary. A common representation, for example, is a regression model $E[y(\boldsymbol{x})] = f(\boldsymbol{x})^T\boldsymbol{\beta}$. Note that $\boldsymbol{\beta}$ is the $d$*$1$ vector of regression coefficients.
2. Define the covariance. There are many different types of covariance functions that can be used, such as squared exponential, Matern, cubic, etc. A common choice is the squared

exponential function as defined in Equation 6.2.1. The full *n*n* covariance matrix between all points is **K**.

$$K(\boldsymbol{x},\boldsymbol{x'}) = \exp\left\{-\sum_{j=1}^{d}\theta_j (x_j - x_j')^2\right\} = \prod_{j=1}^{d}\exp\left(-\theta_j (x_j - x_j')^2\right) \quad (6.2.1)$$

This covariance function involves the product of $d$ squared-exponential covariance functions with different length scales $\theta_j$ on each dimension. The form of this covariance function captures the idea that nearby inputs have highly correlated outputs.

3. Perform the prediction calculations. Given a set of $n$ input data points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$ and a set of associated observed responses or targets $\{y_1, y_2, \ldots y_n\}$, the GP can predict the response value $y(\boldsymbol{x}^*)$ at a new input point $\boldsymbol{x}^*$. The prediction is represented as the sum of the "true" response, y, plus an error term representing noise in the observational data: $y_i + \varepsilon_i$, where $\varepsilon_i$ is a zero-mean Gaussian random variable with constant variance $\sigma^2$. If **K** is the *n*n* covariance matrix with entries $K(\boldsymbol{x}, \boldsymbol{x'})$, then the distribution of the predicted response, which is conditional on the data, is Gaussian with the following mean and variance:

$$\begin{aligned} \mathrm{E}[y(\boldsymbol{x}^*)|\boldsymbol{Y}] &= \mathrm{E}[y(\boldsymbol{x}^*)|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x},\boldsymbol{x}^*] = f(\boldsymbol{x}^*)\boldsymbol{\beta} + k(\boldsymbol{x}^*)^T[\mathbf{K}+\sigma^2\mathbf{I}]^{-1}[\mathbf{Y}-\mathbf{F}\boldsymbol{\beta}] \\ \mathrm{Var}[y(\boldsymbol{x}^*)|\boldsymbol{Y}] &= k(\boldsymbol{x}^*,\boldsymbol{x}^*) - k(\boldsymbol{x}^*)^T[\mathbf{K}+\boldsymbol{\sigma^2}\mathbf{I}]^{-1}k(\boldsymbol{x}^*) \end{aligned} \quad (6.2.2)$$

where $k(\boldsymbol{x}^*)$ is the *n*1* vector of covariances between the $n$ known points $\boldsymbol{X}$ and the new data point $\boldsymbol{x}^*$. **K** is the $n \times n$ covariance matrix of the original data, **I** is the $n \times n$ identity matrix, **F** is the set of basis functions for the full data set $\boldsymbol{X}$, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{Y}$ is the *n*1* vector of response values.

The equations for the mean and variance of the predictive distribution both require the inversion of K, an $n \times n$ matrix. In general, this is a $O(n^3)$ operation. Also, the covariance matrix may be near singular. Several approaches have been developed to deal both with the ill-conditioning and with large data sets (e.g. greater than 1000 data points) [8;10].

Steps 1 – 3 give the general framework for defining a Gaussian process and using it for prediction. However, the length scale parameters in the covariance matrix must be calculated to perform the prediction in Steps 2 and 3. There are two main approaches. One is to use maximum likelihood estimation, in which the likelihood function is maximized. This results in point estimates of the covariance parameters. The other approach is to use Markov chain Monte Carlo (MCMC) sampling to generate posterior distributions on the hyperparameters that govern the mean and covariance function. The assumption of zero-mean GPs is often made, so the Bayesian updating only involves hyperparameters governing the covariance function. Since these may be quite complicated, a MCMC sampling method is usually still needed to generate the posterior. Maximum likelihood methods are more commonly used to obtain the correlation lengths.

An example of a Gaussian process surrogate is shown in Figure 6.1. The left figure shows a GP based on the original training data, where the black line shows the mean prediction of the GP and the grey shaded region shows the prediction uncertainty (typically represented by $\boldsymbol{E}[y(\boldsymbol{x}^*)|\boldsymbol{X},\boldsymbol{Y}] \pm \mathbf{2}\sqrt{\mathrm{Var}[y(\boldsymbol{x}^*)|\boldsymbol{X},\boldsymbol{Y}]}$).

The right figure shows how the prediction uncertainty is reduced when additional data is added to GP build process.

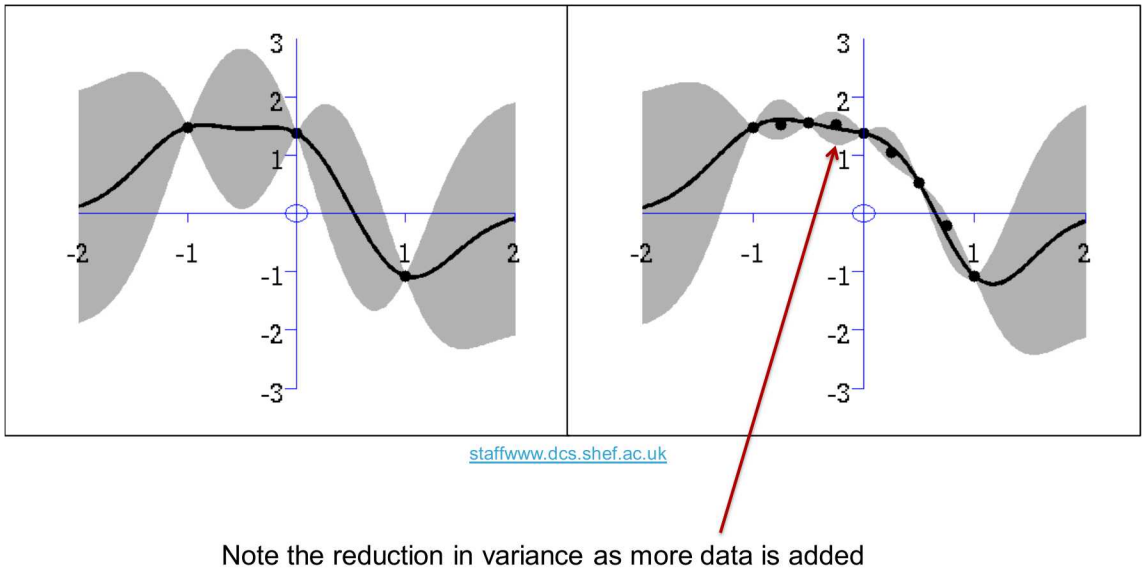



Figure 6.1 Example of a Gaussian process model mean prediction (shown in black) and uncertainty bounds on the mean prediction (shown in grey) with three data points (left) and nine data points (right).

## 6.3 Other advanced metamodels

There are other surrogates that are used in various settings. Surrogate models typically come with various mathematical assumptions, statistical models for the handling of error, etc. A short description of some advanced surrogates is given in the list below:

- Splines are one class of surrogate models which are frequently used for data interpolation. Splines are typically specified by piecewise polynomials that are constructed over disjoint intervals of the inputs. The build points for splines are called "spline knots." A commonly used multidimensional spline is called MARS (MARS- multivariate adaptive regression splines) [8].
- Radial basis functions (RBF) are functions whose value typically depends on the distance from a center point, called the centroid. The surrogate model approximation is constructed as the weighted sum of individual radial basis functions [9].
- Moving Least Squares (MLS) is a weighted least squares approach where the weighting is "moved" or recalculated for every new point where a prediction is desired [10].
- Artificial Neural Networks (ANN) are a machine learning approach that constructs a network meant to mimic neurons, with input nodes, output nodes, and hidden layers of nodes. The connections between nodes have weights associated with them, and there are activation functions which calculate the amplitude of the response between nodes. The

calculations of a neural network involve summing the weighted, activated nodes [11]. Neural networks have become wildly popular in the last decade for a variety of artificial intelligence usage [12]. The usage of ANNs as surrogate models is simply to construct an input-output map of simulation input parameters to simulation responses.

- Random Forests are another machine learning technique that were originally used as classifiers [13]. They are an example of a regression tree, where the input space is split into disjoint regions by dividing the observations at a cutpoint on one variable. That is, each branch represents the division of an input into two sets, and the leaves of the tree represent a particular partition or region of input space. A constant (e.g., the mean of the data in that region) is used as the surrogate approximation within that region. A random forest is obtained from an ensemble of trees, each of which is constructed on a random subset of inputs. They are many modifications of this basic concept.
- Non-parametric regression models involve adding a regularization term to a typical least-squares model, typically to penalize terms in the regression that may be unimportant and thus help perform variable selection. These include ridge regression, projection pursuit regression, LARS (least angle regression) [14], LASSO (Least absolution shrinkage and selection) [15], and ACOSSO (Adaptive COmponent Selection and Shrinkage Operator) [16]. ACOSSO is based on a smoothing spline which has penalty terms both for the the magnitude of the overall trend of each term over the domain as well as the magnitude of the curvature or roughness over the domain. There is a tuning parameter which controls the trade-off between the smoothness of the ACOSSO surrogate and the fidelity to the data. Storlie et al. [17] performed a rigorous comparison of a number of these non-parametric regression methods as applied to several challenge problems.

Analyses frequently do not have a large amount of training data on which to construct a surrogate model. Furthermore, depending on the optimization approach used to determine the governing parameters of the surrogate (e.g., regression coefficients, neural network weights) and the amount of data available, there may be significant error in the surrogate model. It is very important for users to examine the surrogate error and estimate it for prediction purposes. There are a variety of surrogate diagnostics or metrics used to calculated "goodness of fit." These include root-mean-squared-error (RMSE), mean absolute error (MAE), and $R$-squared (the proportion of the variance in the dependent variable (output or response) attributed to the independent variables (inputs). Note that these metrics can only be calculated with respect to the available build data (e.g., the training data) because they involve taking differences of the surrogate model and the "true" data. It is good practice to hold back some of the training data and use it as testing data to validate the surrogate model. This is often done in a framework called "cross-validation." In cross-validation, the training data is randomly divided into $k$ partitions or folds (sometimes called $k$-folds). Then $k$ models are computed, each excluding the corresponding $k$th partition of the data. Each model is evaluated at the points that were excluded in its generation and any metrics specified above are computed with respect to the held-out data. The overall metric is then calculated by averaging the individual metric (e.g., RMSE, MAE, or $R$-squared) value from each partition.

Note that it is also important to understand the how the actual surrogate model for a particular set of training data is obtained. Typically, this is done using some type of optimization. Many surrogate models may exhibit poor-identifiability of the hyperparameter estimation (e.g., many combinations of parameters governing the surrogate give about the same fit and same prediction error). There are many ways to test this, such as repeating the optimization many times at different starting points in parameter space to see how many "local minima" there are. Another approach is to perform sensitivity analysis before surrogate construction so that unimportant parameters are left out of the surrogate, making the surrogate estimation problem better posed. Finally, remember that in surrogate modeling of computational simulations, typically a user can only afford a relatively small number of training data points (e.g., the "truth data" may be 100 or 200 simulation runs in a 5-10 dimensional parameter space). This means that the surrogate could be quite different if another 200 data points were taken. Thus, the effects of limited data must be determined through diagnostic metrics and cross-validation.

## 6.4 Polynomial Chaos Expansion

A final class of surrogate models presented are polynomial chaos expansions (PCEs). These are presented in more mathematical detail in Appendix B, but a high-level overview is given here. Polynomial chaos expansion methods approximate the functional dependence of the simulation response on uncertain model parameters by expansion in an orthogonal polynomial basis. The polynomials used are tailored to the characterization of the uncertain input variables [7;8].

Polynomial chaos expansion is a stochastic expansion method whereby the output response is modeled as a function of the input random variables using a carefully chosen set of polynomials—typically, Hermite polynomials. The series expansion is typically based on unit-variance independent Gaussian random variables. Non-normal and correlated uncertain variables $\mathbf{X}$ in the problem must be transformed to a space of independent zero-mean and unit-variance Gaussian random variables, $\mathbf{u}$ (standard normal variables). The Nataf and Rosenblatt transformations [18;19] are used to transform the uncertain variables to standard normal variables. Then the expansion is formulated in terms of Hermite polynomials as functions of the Gaussian random variables. The goal of a PCE analysis is to determine the unknown coefficients of the Hermite polynomials in the expansion. For other distributions, the polynomials are usually chosen according to the Weiner-Askey scheme that provides an orthogonal basis with respect to the probability density function for the input random variables [20].

There are a variety of approaches for generating the sample points, but typically tensor product grids or sparse grids are used. Structured sampling schemes typically result in more accurate estimates of the coefficients of the polynomials than random sampling does, but this depends on the order of the polynomials needed to resolve the uncertainty in the response functions [21].

One advantage of stochastic expansion methods is that the moments of the expansion (e.g., the mean or variance of the response) can be written analytically, along with analytic formulations of the derivatives of these moments with respect to the uncertain variables. This feature can be exploited in design optimization under uncertainty problems [22] or epistemic uncertainty problems [23]. A disadvantage of polynomial chaos is that the tensor product grids upon which they are often constructed do not scale well, especially as the number of variables and the

polynomial order increases. Recent research in adaptive methods and sparse methods is addressing this limitation [24].

The propagation of input uncertainty through a model using PCE consists of the following steps: (1) input uncertainties are transformed in terms of a set of uncorrelated random variables; (2) a functional form such as Hermite polynomials is assumed; and (3) the parameters of the functional approximation are determined.

The general polynomial chaos expansion for a response R has the form:

$$R = a_0 B_0 + \sum_{i_1=1}^{\infty} a_{i_1} B_1(\xi_{i_1}) + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{\infty} a_{i_1,i_2} B_2(\xi_{i_1}, \xi_{i_2}) + \dots \tag{6.4.1}$$

where each B term represents an orthogonal polynomial that is a function of a number of the uncertain inputs $\xi_i$. The number of random variables and the order of the expansion are unbounded prior to truncation. This expression is usually written in terms of the order-based indexing:

$$R = \sum_{j=0}^{\infty} \alpha_j \Psi_j(\xi). \tag{6.4.2}$$

where there is a one-to-one correspondence between $a_{i1,i2}$ in Eq. 6.4.1 and $\alpha_j$ in Eq. 6.4.2, and between $B_n(\xi_{i1}, \xi_{i2}, ..., \xi_{in})$ in Eq. 6.4.1 and $\psi_j(\xi)$ in Eq. 6.4.2. Each $\psi_j(\xi)$ is a multivariate polynomial that involves products of the one-dimensional polynomials. In practice, both the number of random variables and the order of the expansion are truncated yielding an expansion of the form:

$$R \approx \sum_{j=0}^{P} \alpha_j \Psi_j(\xi). \tag{6.4.3}$$

If a total-order polynomial is constructed (e.g., a total order of 2 would involve terms whose exponents are less than or equal to 2, such as $x_1^2$, $x_2^2$ , and $x_1x_2$ but not $x_1^2x_2^2$), the total number of terms $N$ in a polynomial chaos expansion of arbitrary order $p$ for a response function involving $n$ uncertain input variables is given by: $(n + p)!/n!\,p!$. If an isotropic tensor product expansion is used with order $p$ in each dimension, $(p+1)^n$ sample points must be used. If the order $p$ of the expansion captures the behavior of the true function and if the number of quadrature points taken is sufficient to estimate the coefficient terms, polynomial chaos methods will give very accurate results for the output statistics of the response.

## 6.5 References: Chapter 6

1. Mariner, P.E., L. P. Swiler, D. T. Seidl, B. J. Debusschere, J. Vo, J. L. Jerden. "High fidelity surrogate modeling of fuel dissolution for probabilistic assessment of repository performance." *International High-Level Radioactive Waste Management conference,* April 2019. SAND2019-1917C.
2. Viana, F.A.C., R.T. Haftka, And V. Steffen, "Multiple surrogates: how cross-validation errors can help us to obtain the best predictor," *Structural and Multidisciplinary Optimization*, 39(4), 439-457 (2009).
3. Xiu, D., *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press (2010).
4. Ghanem, R. And P. Spanos, *Stochastic Finite Elements: A Spectral Approach*. New York, New York: Springer Verlag (2002).
5. Seber, G.A.F. And C.J. Wild, *Nonlinear Regression*. New York, New York: Wiley & Sons (2003).
6. Rasmussen, C.E. And C.K.I. Williams, *Gaussian Processes for Machine Learning.* MIT Press (2006).
7. Swiler, L.P. "Gaussian Process Models." SAND2019-9565C. Presented at the DOE FASTMATH/USC Summer School on Uncertainty Quantification, Aug. 2019.
8. Friedman, J.H. "Multivariate adaptive regression splines." *Annals of Statistics*, 19(1):1–141, March 1991.
9. Orr, M.J.L. "Introduction to radial basis function networks." Technical report, University of Edinburgh, Scotland, 1996.
10. Nealen.A. "A short-as-possible introduction to the least squares, weighted least squares, and moving least squares methods for scattered data approximation and interpolation." Technical report, Discrete Geometric Modeling Group, Technishe Universitaet, Berlin, Germany, 2004.
11 Bishop, C.M. *Neural Networks for Pattern Recognition.* Oxford University Press, 1996.
12 Geron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media. 2017.
13 Breiman L. "Random forests." *Machine Learning* 2001;45(1):5–32.
14 Efron, B., T. Hastie, I. Johnstone, and R.Tibshirani. 2004. "Least Angle Regression". The Annals of Statistics 32 (2). *Institute of Mathematical Statistics*: 407–51.
15 Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society*. Series B (methodological). Wiley. 58 (1): 267–88.
16 Storlie, C.B., H.D. Bondell, B.J. Reich, and H.H. Zhang (2011). "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property." *Statistica Sinica* 21.
17 Storlie, C. B., L.P. Swiler, J.C. Helton, and C.J. Sallaberry. "Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models." *Reliability Engineering and System Safety* Vol. 94, pp. 1735–1763, 2009.

18. Rosenblatt, M. "Remarks on a multivariate transformation." *Annals of Mathematical Statistics*, 23(3):470–472, 1952.
19. Der Kiureghian, A. and P.L. Liu. "Multivariate distribution models with prescribed marginals and covariances." *Probabilistic Engineering Mechanics,* 1 (2) (1986), pp. 105–112.
20. Askey, R. and J. Wilson. "Some basic hypergeometric polynomials that generalize Jacobi polynomials." In *Mem. Amer. Math. Soc*. 319, Providence, RI, 1985. AMS.
21. Constantine, P.G., Eldred, M.S., and Phipps, E.T., "Sparse Pseudospectral Approximation Method." *Computer Methods in Applied Mechanics and Engineering*, Volumes 229-232, pp. 1-12, July 2012.
22. Eldred, M.S., "Design Under Uncertainty Employing Stochastic Expansion Methods," *International Journal for Uncertainty Quantification*, Vol. 1, No. 2, Feb. 2011, pp. 119-146.
23. Eldred, M.S., Swiler, L.P., and Tang, G., "Mixed Aleatory-Epistemic Uncertainty Quantification with Stochastic Expansions and Optimization-Based Interval Estimation," *Reliability Engineering and System Safety*, Vol. 96, No. 9, Sept. 2011, pp. 1092-1113.
24. Najm, H. N, B. Debusschere, K. Sargsyan, J.D. Jakeman, C. Safta, M.S. Eldred, "Sparse Polynomial Representations of High Dimensional Models," Conference Paper, 10th *World Congress on Computational Mechanics*, June 2012.

# 7. SHALE REFERENCE CASE

In this chapter, the UQ/SA methods outlined in the previous chapters are demonstrated on a particular generic repository model referred to as the shale (or argillite) reference case. The reference case is notionally depicted in Figure 7.1.

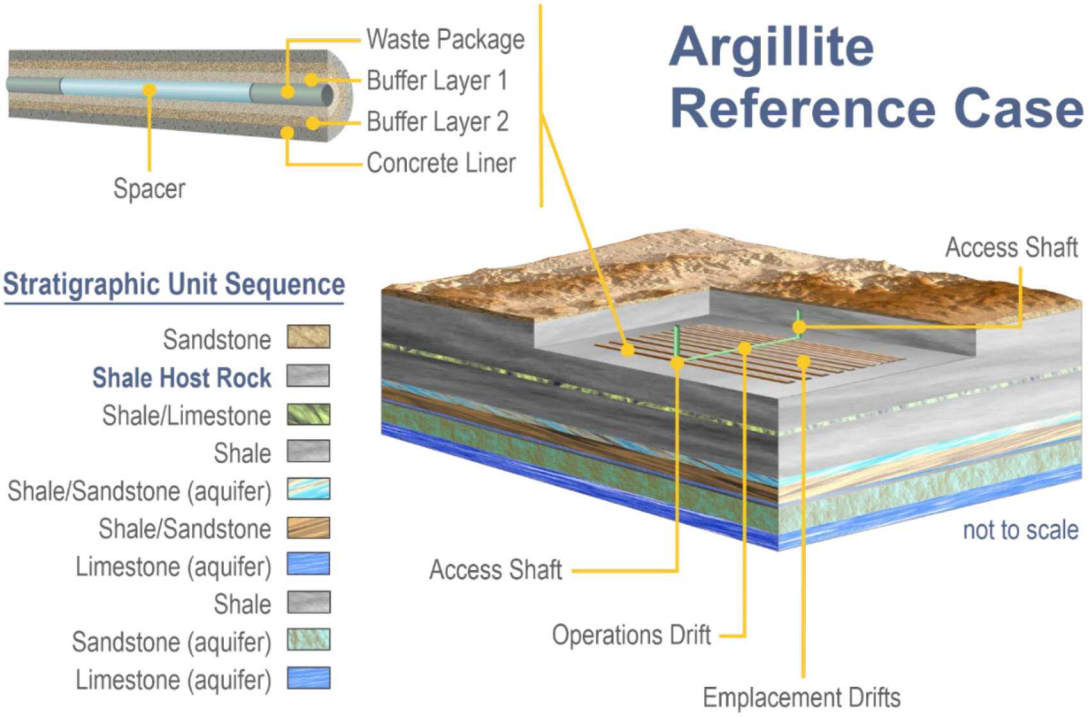



Figure 7.1 Shale Reference Case

*GDSA Framework* is used for these simulations. The sampling methods used in Dakota are Latin Hypercube sampling (presented in Chapter 4) and incremental LHS (to augment LHS samples with additional samples that also form a Latin design) as well as variance-based indices (presented in Chapter 5). The variance-based indices are calculated with two different surrogates: Gaussian processes and polynomial chaos expansions. Further, the results of the analysis with and without a log-transformation of the response quantity (maximum concentration of the radioactive isotope iodine-129 ($^{129}$I) at various observation points in the model domain) are compared. It is found that log-transforming the data is very important in obtaining reliable sensitivity estimates for this case. This chapter is meant to demonstrate a practical workflow using UQ/SA methods within the *GDSA Framework*.

## 7.1 Shale Reference Case Description

*GDSA Framework* simulations of the generic shale reference case [1] assume a mined repository for commercial spent nuclear fuel in a thick unit of a low-permeability shale or argillaceous formation. A thin limestone aquifer lies below the shale and a sandstone aquifer lies above the shale. A second sandstone aquifer lies at depth. A pressure gradient drives regional flow from west to east (left to right). Figure 7.2 shows a 2-dimensional vertical cross section of the

three-dimensional model domain, in which the repository is visible as a series of drift cross-sections toward the left side of the model domain. The model domain contains 6,925,936 cells.

Simulations assume a thick low-permeability shale with aquifers (potential paths to the biosphere) above and below the host rock. Epistemic uncertain inputs include properties of the engineered and natural systems. The output variables of interest, maximum $^{129}$I concentrations (independent of time) at observation points in the aquifers, vary over several orders of magnitude.

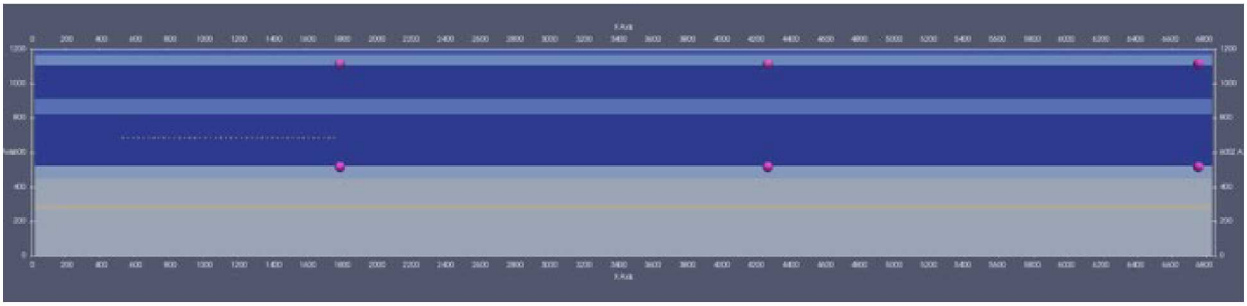


Figure 7.2 2-D cross section of 3-D model domain showing aquifer observation points (pink dots). From top to bottom and left to right: sand_obs1, sand_obs2, sand_obs3, lime_obs1, lime_obs2, lime_obs3.

Multi-physics simulations on the 7-million-cell grid are run in an HPC environment with the PFLOTRAN component of *GDSA Framework*. Each one-million-year simulation took approximately three hours to complete using 512 processor cores per simulation. Physical-chemical processes accounted for in the simulations include waste package degradation, waste form dissolution, radionuclide sorption, solubility-controlled precipitation and dissolution, radioactive decay and ingrowth in all phases (aqueous, adsorbed, precipitated), coupled heat and fluid flow, and radionuclide transport via advection and diffusion. Simulations included 18 radionuclides.

Two hundred realizations were generated by sampling ten input parameters (Table 7.1) using Dakota's incremental Latin hypercube sampling (LHS) capability and an initial sample size of 50. The concentration of $^{129}$I as a function of time was recorded at three locations in the upper sandstone and three locations in the limestone (Figure 7.3). The maximum $^{129}$I concentrations (regardless of time) at each of these six observation points are the output variables for the sensitivity analysis.

Two of the sampled parameters are linear distribution coefficients ($K_d$s) for neptunium sorption in various media. However, no further discussion of $^{237}$Np transport is included because Np concentration at the observation points remain very close to the numerically-imposed background concentration of $10^{-20}$ mol/L in all simulations.

Table 7.1 Sampled inputs for the shale reference case.

| Input | Description | Range | Units | Distribution |
|---|---|---|---|---|
| rateSNF | SNF Dissolution Rate | $10^{-8}$ – $10^{-6}$ | $yr^{-1}$ | log uniform |
| rateWP | Mean Waste Package Degradation Rate | $10^{-5.5}$ – $10^{-4.5}$ | $yr^{-1}$ | log uniform |
| kSand | Upper Sandstone Permeability | $10^{-15}$ – $10^{-13}$ | $m^2$ | log uniform |
| kLime | Limestone Permeability | $10^{-17}$ – $10^{-14}$ | $m^2$ | log uniform |
| kLSand | Lower Sandstone Permeability | $10^{-14}$ – $10^{-12}$ | $m^2$ | log uniform |
| kBuffer | Buffer Permeability | $10^{-20}$ – $10^{-16}$ | $m^2$ | log uniform |
| kDRZ | DRZ Permeability | $10^{-18}$ – $10^{-16}$ | $m^2$ | log uniform |
| pShale | Host Rock (Shale) Porosity | 0.1 – 0.25 | - | uniform |
| bNpKd | Np $K_d$ Buffer | 0.1 – 702 | $m^3kg^{-1}$ | log uniform |
| sNpKd | Np $K_d$ Shale | 0.047 – 20 | $m^3kg^{-1}$ | log uniform |

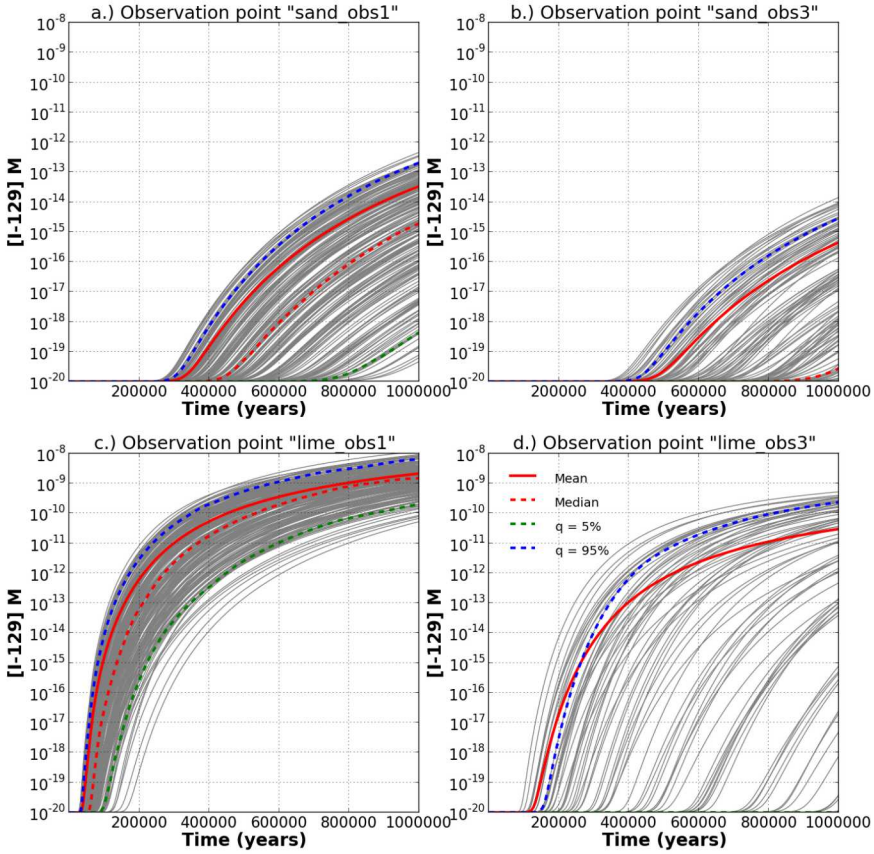



Figure 7.3 $^{129}I$ concentration versus time for 200 realizations at four aquifer observation points.

### 7.1.1 Incremental LHS

As described in Chapter 4.3, LHS is a sampling method in which the range of each input variable is divided into segments of equal probability, and a value is randomly selected from each segment [2]. Values thus chosen for each input variable are randomly combined to create vectors of inputs. LHS generally produces more stable estimates and results in faster convergence of statistics than Monte Carlo sampling. Dakota's incremental LHS method incorporates the existing LHS sample into a subsequent LHS sample of twice the size [3]. Thus, with an initial sample size of 50, the 200 realizations provide three LHS samples of size 50, 100, and 200 for sensitivity analysis.

### 7.1.2 Variance Decomposition/Sensitivity Indices

Variance-based global sensitivity analyses (i.e., calculations of sensitivity indices) conducted with Dakota on this case use polynomial chaos expansion and Gaussian process surrogate models [4]. Results of analyses conducted with raw output concentrations and with log-transformed output concentrations were compared. Using log-transformed concentrations results in larger sensitivity indices for more influential input variables, smaller sensitivity indices for less influential input variables, and more consistent values for sensitivity indices between methods (PCE and GP) and between analyses repeated with samples of different sizes.

The details of the variance-based based sensitivity indices are described in Chapter 5. Sensitivity indices are a measure of the fraction of the variance in the output due to the variance in uncertain inputs. Due to the large number of simulations required to calculate sensitivity indices, surrogate models are often employed—see Chapter 6. Two such approaches were employed in this analysis. In both cases, some number of realizations must be run with the original simulation code in order to derive the surrogate model describing the response of the output variable of interest to the sampled input variables. In the first approach, the surrogate model is then evaluated $m(d+2)$ times, where $m$ is the number of samples and $d$ is the number of sampled inputs, to find the main and total sensitivity indices [5]. In the second approach, a surrogate model is chosen from which sensitivity indices can be analytically calculated, eliminating the need for multiple evaluations of the surrogate. For the shale reference case, the first approach is demonstrated using Gaussian process surrogate models [6], and the second approach using polynomial chaos expansion [7]. Gaussian processes and polynomial chaos expansion are described in more detail in Chapter 6.

#### *7.1.2.1 Variance-based decomposition with Gaussian processes*

Gaussian process surrogate models were fit to the data using global optimization (least squares regression), a quadratic polynomial as the trend function, and a stationary isotropic squared exponential covariance function. The resulting surrogate model (the mean of the Gaussian process) and a sample size ($m$) of 10,000 was used to estimate main and total sensitivity indices.

#### *7.1.2.2 Variance-based decomposition with Polynomial Chaos Expansion*

PCEs were generated using orthogonal matching pursuit and 10-fold cross validation to select the order of the expansion (up to a user-specified maximum of 4$^{th}$ order) that resulted in the smallest cross validation error.

Because maximum $^{129}I$ concentrations are consistently very small (< $10^{-13}$ mol/L) at some observation points, the PCE method sometimes failed to return sensitivity indices due to "negligible variance." To overcome this problem, concentrations were multiplied by a factor of $10^{13}$, which enabled the PCE method to return sensitivity indices. Increasing the scaling factor to $10^{20}$ (1/[background $^{129}I$ concentration]), resulted in sensitivity indices within approximately 1% of those calculated using the smaller scaling factor. Scaling the concentrations was not necessary when working with log-transformed values.

### 7.1.3 Results

Sensitivity indices calculated for maximum $^{129}I$ concentration (scaled by $10^{13}$ for the PCE method) and log-transformed maximum $^{129}I$ concentration at two points in the upper sandstone aquifer and two points in the limestone aquifer are plotted in Figure 7.4 and Figure 7.5, respectively. Main sensitivity indices are indicated with hatch marks and plotted on top of total sensitivity indices. Regardless of method (PCE or GP) and sample size (50, 100, or 200), calculated sensitivity indices indicate that maximum $^{129}I$ concentration at all observation points is sensitive to the porosity of the shale host rock (pShale). At the observation points furthest from the repository (sand_obs3 and lime_obs3), maximum $^{129}I$ concentration is also strongly influenced by the permeability of the aquifer (kSand and kLime, respectively). The same conclusions can be drawn on the basis of $\log_{10}$[maximum $^{129}I$ concentration].

However, the sensitivity indices calculated on the basis of maximum $^{129}I$ concentration and log maximum $^{129}I$ concentration differ in the additional conclusions that might be drawn from them and in the stability of their values. In particular, sensitivity indices calculated using log-transformed concentrations indicate greater sensitivity to fewer input variables than sensitivity indices calculated using untransformed (raw if using the GP method, or linearly scaled if using the PCE method) concentrations. Sensitivity indices of more influential input variables are generally larger (and those of less influential input variables are generally smaller) when calculated using log-transformed concentrations than when calculated using raw/scaled concentrations. This pattern holds true at sand_obs1, where use of the log transform increases the total sensitivity index for pShale from approximately 0.7 to almost 1 and decreases the total sensitivity index for rateWP from approximately 0.2 to 0.02 (Figure 7.4); and at lime_obs1, where use of the log transform increases the total sensitivity index for pShale by approximately 0.1 and decreases the total sensitivity index for rateSNF by approximately 0.1 (Figure 7.5). Additionally, sensitivity indices calculated using log-transformed concentrations indicate less sensitivity to interactions between input variables than sensitivity indices calculated using raw/scaled concentrations, i.e., main sensitivity indices calculated using log-transformed concentrations sum to a larger number (and account for a larger fraction of the variance in the output) than main sensitivity indices calculated using raw/scaled concentrations (Table 7.2).

Use of the log transform results in greater stability among sensitivity indices calculated with different sample sizes and different methods. Three sample sizes (50, 100, 200) and two methods (PCE and GP) result in 6 estimates for each sensitivity index – the standard deviation of these 6 values can be taken as a measure of stability. At sand_obs3 (Figure 7.6), sensitivity indices calculated using raw/scaled concentrations have standard deviations up to 0.14, while those calculated using log-transformed concentrations have standard deviations approximately one order

of magnitude smaller (< 0.014). The comparison is similar at the other observation points (not shown).

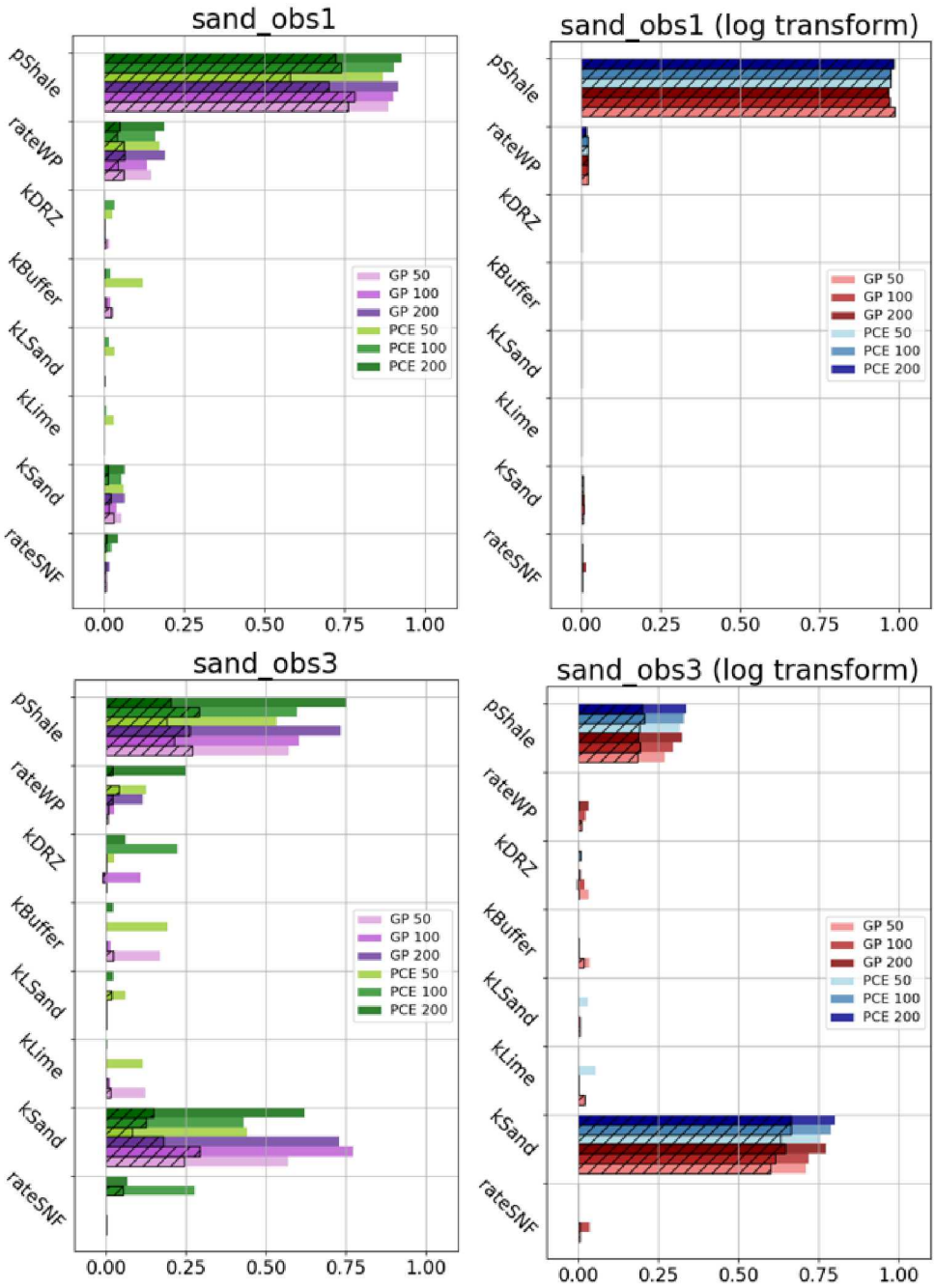



Figure 7.4 Sensitivity indices in the sandstone aquifer. Main sensitivity indices are indicated with hatch marks and plotted on top of total sensitivity indices.

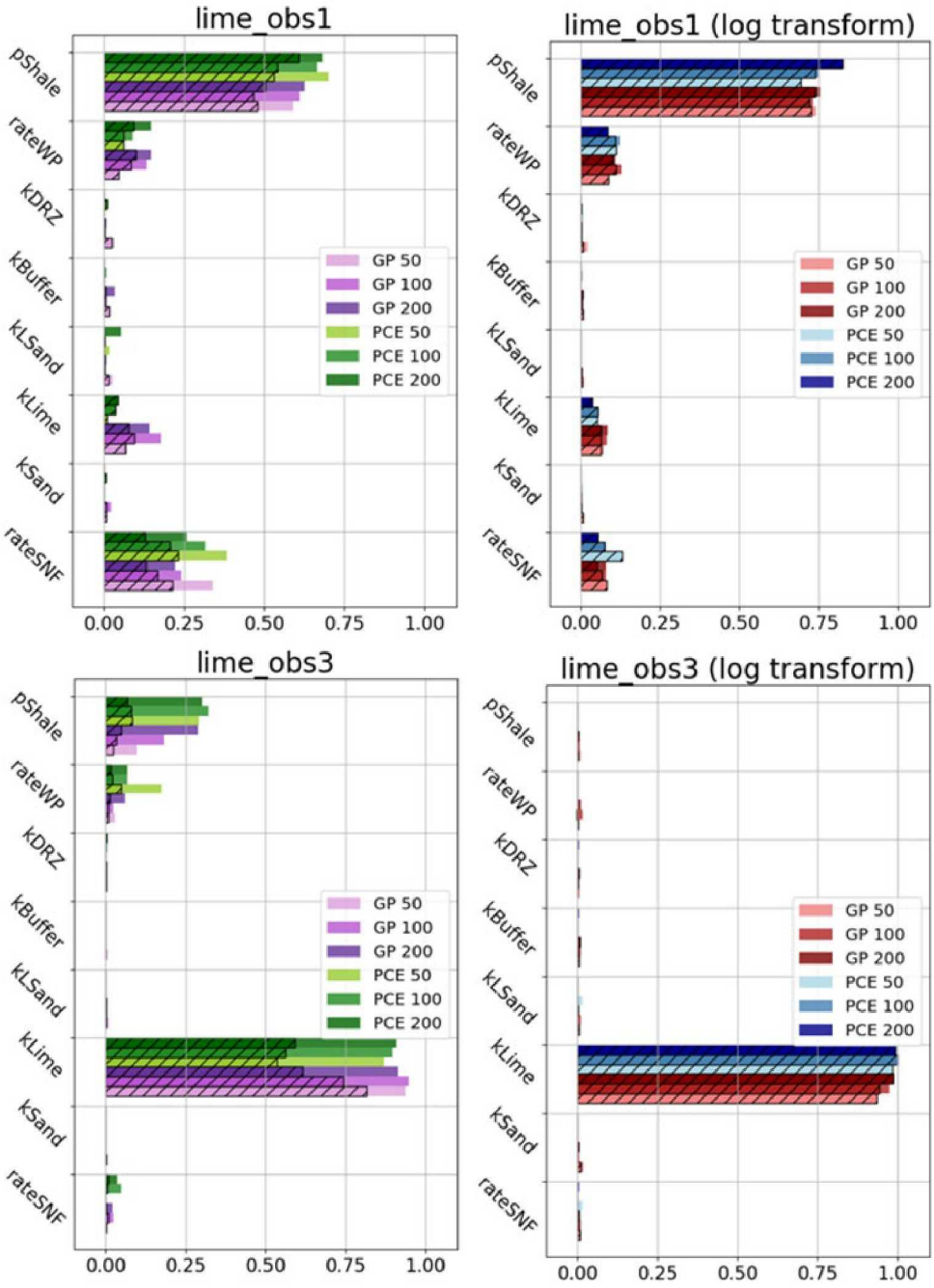



Figure 7.5 Sensitivity indices in the limestone aquifer. Main sensitivity indices are indicated with hatch marks and plotted on top of total sensitivity indices.

Table 7.2 Fraction of variance due to sum of main effects.

| | sand_obs1 | sand_obs3 | lime_obs1 | lime_obs3 |
|---|---|---|---|---|
| *without log transform* | | | | |
| **GP 200** | 0.79 | 0.46 | 0.82 | 0.68 |
| **PCE 200** | 0.79 | 0.37 | 0.87 | 0.68 |
| *with log transform* | | | | |
| **GP 200** | 1.00 | 0.84 | 0.96 | 0.99 |
| **PCE 200** | 1.00 | 0.86 | 1.00 | 0.99 |

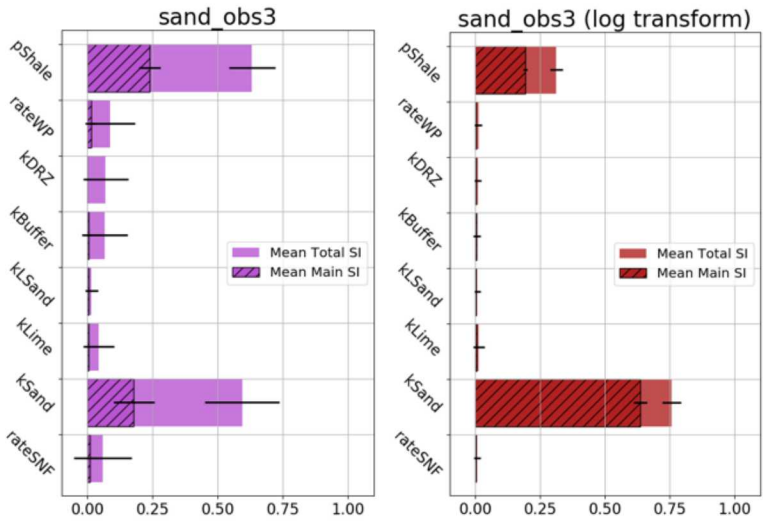



Figure 7.6 Mean sensitivity indices and standard deviation (error bars) at sand_obs3.

At sand_obs3, approximately 80% of the realizations result in maximum $^{129}I$ concentrations two orders of magnitude (or more) less than the largest value (Figure 7.7). Surrogate models built across an output range of several orders of magnitude are strongly influenced by the largest values, and accuracy of such surrogates at the low end of the output range tends to be poor. Use of log-transformed concentrations increases the contribution of smaller concentrations to both the variance of the output and the fit of the surrogate models used to calculate variance. The small standard deviations achieved for sensitivity indices calculated using log-transformed concentrations suggests that for the shale reference case, a sample size of 50 is sufficiently large to reliably estimate sensitivity indices if log-transformed concentrations are used. The large standard deviations (and lack of a trend with increasing sample size) obtained using raw or linearly scaled concentrations indicate that a sample size of 200 may not be large enough to reliably estimate sensitivity indices if raw/scaled concentrations are used.

Sensitivity indices calculated using raw or linearly scaled concentrations identify sensitivity to input variables and variable interactions that the log-transformed results do not. This result suggests that the combination of uncertain inputs contributing to variance at the upper end of the concentration range is somewhat different than the combination that controls the order-of-

magnitude differences over the entire range of output concentrations. To better quantify the contribution of uncertain inputs to output variance at the upper end of the concentration range, more realizations that result in high maximum $^{129}$I concentration are necessary. Importance sampling, a sampling method designed to preferentially sample regions of an input sample space that result in an outcome of interest, may be one means of achieving increased sample density in the region that results in high $^{129}$I concentration without significantly increasing the overall computational cost of the analysis, and will be explored in subsequent work.

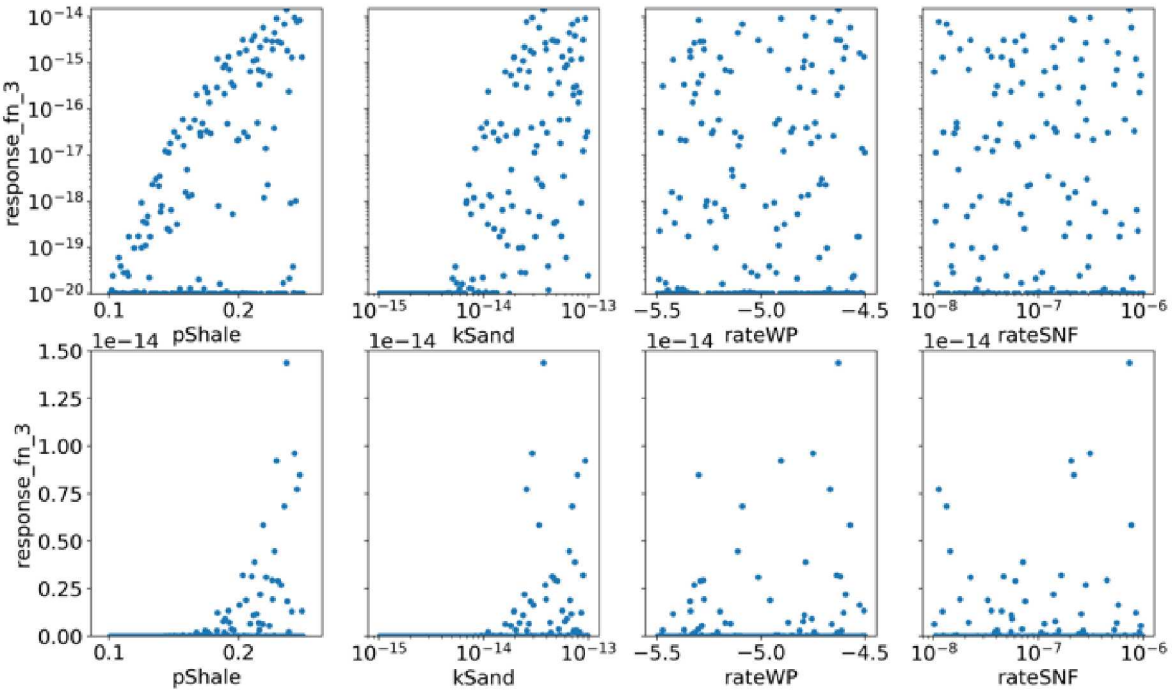



Figure 7.7 Maximum $^{129}$I concentration at sand_obs3 plotted against pShale, kSand, rateWP, and rateSNF on log scale (top) and linear scale (bottom).

Probabilistic simulations of a generic commercial SNF repository in shale provide a test case for applying and comparing sensitivity analysis methods available in *GDSA Framework*. Sensitivity indices measuring the contribution of uncertain inputs to variance in the output (either maximum $^{129}$I concentration or log maximum $^{129}$I concentration) were calculated with Dakota using polynomial chaos expansion and Gaussian process surrogate models. Using log-transformed concentration results in larger sensitivity indices for more influential input variables, smaller sensitivity indices for less influential input variables, and more consistent estimates of sensitivity indices between methods (PCE and GP) and between analyses repeated with samples of different sizes. However, sensitivity indices calculated using raw (or linearly scaled) concentrations identify sensitivity to input variables and variable interactions that the log-transformed results do not, perhaps because sensitivity indices calculated using raw concentrations discriminate amongst influences contributing to variance at the upper end of the concentration range. Importance sampling may provide a means of more accurately quantifying the sensitivity of the highest $^{129}$I concentrations to uncertain inputs without unduly increasing the computational expense of the analysis.

## 7.2 References: Chapter 7

1. Mariner, P.E., Stein, E., Frederick, J.M, Sevougian, S.D., and Hammond, G.E. *Advances in Geologic Disposal System Modeling and Shale Reference Cases*. SFWD-SFWST-2017-000044 / SAND2017-10304R. Sandia National Laboratories, Albuquerque, NM (2017).
2. Helton, J.C. and F.J. Davis. "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems". *Reliability Engineering & System Safety,* 81(1), 23-69 (2003).
3. Sallaberry, C.J., Helton, J.C., and S.A. Hora. "Extension of Latin Hypercube Samples with correlated variables." *Reliability Engineering & System Safety,* 93(7), 1047-1059 (2008).
4. Adams, B.M. et al., "*Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.8 User's Manual*. SAND2014-4253. Sandia National Laboratories, Albuquerque, NM (2018).
5. Saltelli A., Annoni P., Azzini I., Campolongo F., Ratto M., Tarantola S., "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". *Computer Physics Communications*, 181(2), 259-270 (2010).
6. Le Gratiet, L., Marelli, L.S., and B. Sudret, "Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes," in R. Ghanem et al. (Eds.), *Handbook of Uncertainty Quantification* (pp. 1289-1325), Springer, 2017.
7. Sudret, B., "Global sensitivity analysis using polynomial chaos expansions". *Reliability Engineering & System Safety*, 93(7), 964-979 (2008).

# 8. CRYSTALLINE REFERENCE CASE

For a nuclear waste repository located in crystalline rock, a major source of uncertainty in performance assessment is the spatial heterogeneity of potential fracture flow paths through the host rock. Conceptually, a long-lived radionuclide released from a waste package will initially migrate through the buffer material and into the surrounding damaged rock zone (DRZ). From there it will migrate along the DRZ until it enters a fracture that takes it farther into the host rock where connected fractures can provide a path to a nearby fracture zone. It might then migrate along this fracture zone and through connected fractures zones to the biosphere. Along the flow path, the radionuclide will undergo radioactive decay and ingrowth and diffuse into and out of dead-end pores and fractures. Additionally, depending on its properties and the environmental conditions along the flow path, it will adsorb and desorb from colloids and immobile mineral surfaces, chemically react with aqueous species, possibly change oxidation state, and if solubility-limited, precipitate and dissolve.

This chapter presents three new uncertainty analyses of a reference case repository for commercial spent nuclear fuel in fractured crystalline rock. The reference case is identical to that in Stein et al. [1], Mariner et al. [4], and Sevougian et al. [9] except for improvements to the fracture network implementation (described in Section 8.1.2) and the design of the uncertainty analyses.

In this report, spatial heterogeneity in fracture distribution and in waste package corrosion rate are treated as aleatory uncertainty (random, stochastic, or irreducible uncertainty; see Chapter 3) and uncertainties in input parameter values are treated as epistemic uncertainty (state of knowledge uncertainty; see Chapter 3). Epistemic uncertainties include porosity and permeability of the bentonite buffer (pBuffer and kBuffer, respectively), permeability of the DRZ (kDRZ) and of the overlying sedimentary unit (kGlacial), and rate of spent (used) nuclear fuel dissolution (rateUNF).

Stochastic (spatial) heterogeneity in waste package corrosion rate is introduced by sampling from a truncated log normal distribution for the base corrosion rate in a temperature-dependent rate expression for general corrosion [4]; this log normal distribution for the (spatially distributed) waste package corrosion rate is sampled within PFLOTRAN [6, 7]. Both the mean (meanWPrate) and standard deviation (stdWPrate) of this log normal distribution are treated as epistemically uncertain inputs. Stochastic (spatial) heterogeneity in fracture distribution in the host rock is generated using DFNWorks [3], which creates random realizations of the discrete fracture network (DFN).

Latin hypercube sampling (see Chapter 4) of epistemically uncertain inputs is performed using Dakota [8]. All three uncertainty analyses include pBuffer, kBuffer, kDRZ, kGlacial, and rateUNF among the epistemic uncertain inputs, and use peak $^{129}$I concentration and $^{129}$I breakthrough times as the output variables of interest. The three uncertainty analyses (UA) differ in the treatment of uncertainty in timing and in spatial heterogeneity of waste package corrosion rate, as well as the sample size for the DFN:

1. UA 1 (Section 8.3) comprises an aleatory loop of 30 DFN realizations, and an epistemic loop of sample size 50 for a total of 1500 simulations. The same epistemic sample is run on all 30 DFN realizations, and the mean of the waste package corrosion rate distribution (meanWPrate) is included as an epistemically uncertain input, but the standard deviation

(stdWPrate) is not. Within PFLOTRAN, each simulation uses the same random seed when sampling the waste package corrosion rate distribution, so that regardless of the mean of the distribution (which controls overall timing of waste package breach), the order in which waste packages breach is the same in all simulations – i.e., the location of the waste package with the fastest corrosion rate does not vary, etc.

2. UA 2 (Section 8.4) consists of a single DFN realization (DFN 8 from UA 1). Epistemic uncertain inputs include meanWPrate and stdWPrate (both of which affect the overall timing of waste package breach); an epistemic sample size of 50 is used. Results of UA 2 are compared to results of DFN 8 in UA 1 to demonstrate the effect of uncertainty in stdWPrate on uncertainty in model outputs.
3. UA 3 (Section 8.5) comprises an aleatory loop of sample size 20, and an epistemic loop of sample size 40 for a total of 800 simulations. For each DFN realization in the aleatory loop, a different random seed is used when sampling the waste package corrosion rate distribution, so that the order of waste package breach associated with each DFN is different. The same epistemic sample is run on all 20 aleatory realizations, and both meanWPrate and stdWPrate are included as epistemically uncertain inputs. Although UA 3 does not comprise the largest number of simulations, it does encompass the largest number of uncertainties.

## 8.1 Model Set-up

### 8.1.1 Model Domain

The model domain [1] is 3015-m in length, 2025-m in width, and 1260-m in height. Overlying the host rock is a 15-m thick overburden of glacial sediments (not shown). The repository is located at a depth of 585 m. Forty-two disposal drifts contain 80 12-PWR waste packages each (3360 12-PWR waste packages in total). Drifts are backfilled with bentonite buffer and are surrounded by a 1.67-m thick DRZ. Within the repository grid cells are as small as 1.67-m on a side; elsewhere grid cells are 15-m on a side. The model domain contains 4,848,260 cells; of these, approximately 2.5 million are the smaller cells in and around the repository that allow representation of individual waste packages with surrounding buffer materials. Additional information on the grid and dimensions may be found [1] available for download at https://pa.sandia.gov.

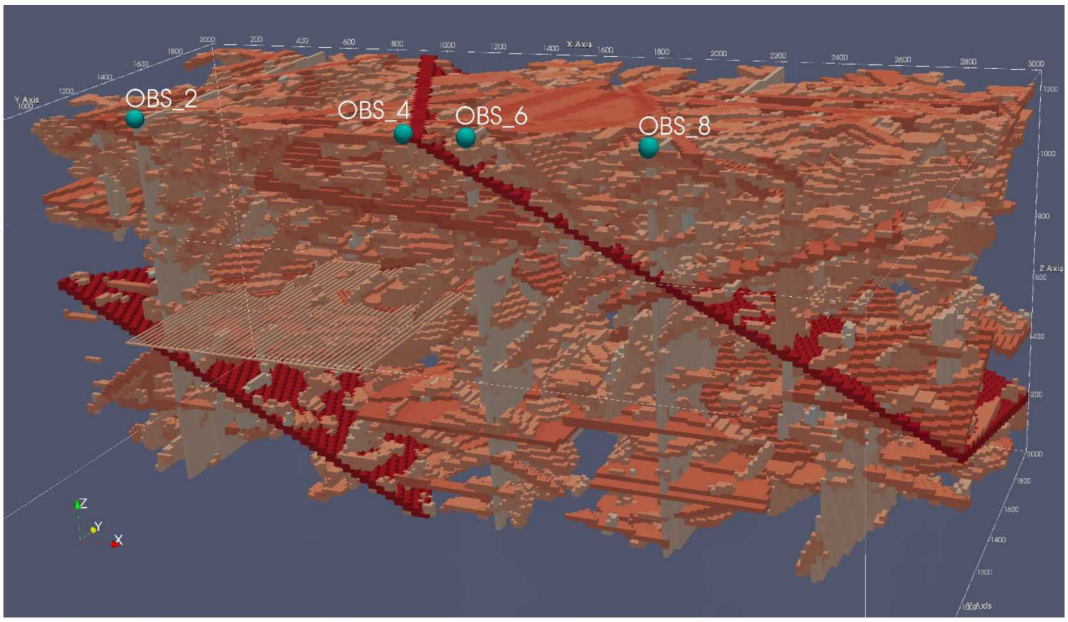



Figure 8.1 Cut-away of DFN 1 realization mapped to porous medium grid, showing the full repository and the far half of the model domain.

### 8.1.2 Discrete Fracture Networks

As described by Mariner et al. [4], the representation of fractured crystalline rock in the GDSA reference case is based on the well-characterized, sparsely fractured metagranite at Forsmark, Sweden [2]. At Forsmark, large-scale mappable features of concentrated brittle and/or ductile deformation (termed "deformation zones") bound volumes of relatively undeformed rock. Each volume of relatively undeformed rock (termed a "fracture domain") is sparsely fractured, and the fractures within each can be described in terms of a number of "fracture sets," distinguished from each other on the basis of fracture orientation. At Forsmark six fracture domains are defined, each containing five fracture sets. As appropriate, three depth zones are defined (<200 m below sea level (mbsl), 200-400 mbsl, and >400 mbsl) in order to account for the decrease in fracture density and fracture transmissivity with depth. Each fracture set within a particular fracture domain and depth zone is described using a 3-dimensional Fisher distribution to describe the orientation of fracture poles in space, a truncated power-law distribution for fracture radii, and a fracture density, $P_{32}$, which is defined as the surface area of fractures per volume of rock ($m^2/m^3$). For each depth zone within a fracture domain, a relationship is given between fracture radius and fracture transmissivity.

The crystalline host-rock reference case analyzed here [4], based on the Forsmark data set, contains 5 deterministic fracture zones and three depth intervals in which fracture density of the stochastic network decreases with increasing depth (Table 8.1). The deterministic, user-defined fracture zones represent large, mappable features, such as faults, and are common to all realizations. There are three subvertical fracture zones (Figure 8.1, in gray) and two fracture zones with a dip of approximately 30 degrees (in red). The stochastic discrete fracture networks, two-dimensional planes distributed in the three-dimensional model domain, are generated using DFNWorks [3], and mapped to the equivalent continuous porous medium domain using

mapDFN.py, a code that approximates hydraulic fracture properties by calculating and assigning permeability and porosity on a cell-by-cell basis [1].

The fracture set properties and deterministic fracture zones employed in this study provided sufficient fracture connectedness such that each DFN realization resulted in direct fracture pathways from the repository to the top boundary. The existence of connected fracture pathways was determined by computer code.

Table 8.1 Parameters used to generate discrete fracture networks in which fracture density decreases with depth.

| | **Orientation: Fisher Distribution for Poles** | | | **Size: Truncated Power Law for Radii** | | | **Fracture Density (Requested)** |
|---|---|---|---|---|---|---|---|
| **Depth (m) / Fracture Set** | Mean Trend | Mean Plunge | $\kappa$ | $\alpha$ | Min Radius $r_0$ (m) | Max Radius $r_x$ (m) | Number of fractures in 1 km$^3$ |
| 0-200 / NS | 90° | 0.0° | 22 | 2.5 | 30 | 500 | 184 |
| 0-200 / EW | 180° | 0.0° | 22 | 2.7 | 30 | 500 | 274 |
| 0-200 / HZ | 360° | 90.0° | 10 | 2.4 | 30 | 500 | 2217 |
| 200-400 / NS | 90° | 0.0° | 22 | 2.5 | 30 | 500 | 357 |
| 200-400 / EW | 180° | 0.0° | 22 | 2.7 | 30 | 500 | 296 |
| 200-400 / HZ | 360° | 90.0° | 10 | 2.4 | 30 | 500 | 1290 |
| >400 / NS | 90° | 0.0° | 22 | 2.5 | 30 | 500 | 236 |
| >400 / EW | 180° | 0.0° | 22 | 2.7 | 30 | 500 | 140 |
| >400 / HZ | 360° | 90.0° | 10 | 2.4 | 30 | 500 | 576 |

### 8.1.3 Waste Package Corrosion Model

The waste package corrosion model implemented in PFLOTRAN (Mariner et al. [4], Section 3.2.1) calculates normalized thickness of the waste package wall at each time step as a function of a base waste package corrosion rate, a canister material constant, and temperature. Waste package breach occurs when the normalized thickness reaches zero. The normalized thickness is initialized to 1, and is reduced at each time step as a function of the effective waste package corrosion rate $R_{eff}$,

$$R_{eff} = R \cdot e^{C\left(\frac{1}{333.15} - \frac{1}{T(t,\vec{x})}\right)} \tag{8-1}$$

where $R$ is the base corrosion rate at 60ºC, $T$ is the local temperature (in Kelvin), and $C$ is the canister material constant. This equation assumes that reaction rates are a function of temperature as described by the Arrhenius equation. Assuming waste package corrosion occurs via general corrosion, $R$ represents the normalized general corrosion rate at 60ºC in units of 1/T (i.e., units of L/T normalized by the thickness of the waste package wall).

PFLOTRAN assigns a base normalized general corrosion rate ($R$) for each waste package by sampling on a truncated log normal distribution whose mean, standard deviation, and upper truncation limit may be treated as epistemic uncertain inputs sampled by Dakota [8]. The mean of the distribution is sampled in all three uncertainty analyses (UA 1, UA 2, and UA 3) over the range −5.5 to −4.5 log units. In UA 1, a constant standard deviation of 0.5 log units is used; in UA 2 the standard deviation is sampled over the range 0 to 0.5 log units; and in UA 3 it is sampled over the range 0.15 to 0.4 log units. All three uncertainty analyses use a constant upper truncation limit of −3.5 log units. A mean of −5 log units results in a mean waste package breach time of 100,000 years if waste packages are held at a constant temperature of 60 °C.

### 8.1.4 Initial and Boundary Conditions

Initial conditions specified are pressure, temperature, and radionuclide concentrations. Initial pressures and temperatures throughout the model domain are calculated by applying a liquid flux of 0 m/s and an energy flux of 60 mW/m$^2$ to the base of the domain and holding temperature (10°C) and pressure (approximately atmospheric) constant at the top of the domain and allowing the simulation to run to $10^6$ years. Pressure at the top of the domain decreases from west (left) to east (right) with a head gradient of –0.0013 (m/m). This technique results in initial conditions that represent a geothermal temperature gradient and hydrostatic pressure gradient in the vertical direction, and a horizontal pressure gradient that drives flow from west to east.

The initial concentration of $^{129}$I in all cells is $10^{-22}$ mol/L. A non-zero value is necessary, because PFLOTRAN transport equations are formulated in terms of the log of concentration. A concentration of $10^{-22}$ mol/L is approximately 60 atoms of $^{129}$I per liter of water.

At all six faces of the model domain, pressures and temperatures are held constant at initial values. Concentration of $^{129}$I is held at the initial concentration at inflow boundaries. At outflow boundaries, the concentration gradient is set to zero.

### 8.1.5 Observation Points

Concentrations of $^{129}$I are monitored at observation points located on the vertical plane that intersects the midline of the repository [1]. Each observation point is located at the top of the crystalline host rock (just beneath the sedimentary overburden) within a deterministic fracture zone.

### 8.1.6 Timestep Size

During the development of the study, oscillations in $^{129}$I concentrations were found at observation points which could impact estimates of $^{129}$I breakthrough times. An analysis was conducted to identify the cause of the oscillations and determine how the effect could be minimized in future studies.

The strategy for this analysis was to focus on a single observation point from one epistemic realization that showed many oscillations. The only change made to the subsequent realizations was in respect to the maximum timestep size, which started at 5,000, decreased down to 500, and ended at 50 years. Observation Point 8 (Obs_8) was chosen for the analysis because results showed a high number of oscillations throughout many of the realizations. Oscillations were thought to be

caused by the maximum timestep value. To test this, the maximum timestep size was iteratively reduced and the sequence of results for Observation Point 8 were compared.

The $^{129}$I concentration oscillations become less frequent with smaller timesteps but this timestep reduction resulted in an increase of computational time. The maximum timestep size for the initial sensitivity study that resulted in oscillations was set to 5,000 years. By decreasing the maximum timestep size down, first to 500 and then to 50 years, oscillations became less frequent and more concentrated towards the beginning of the simulation. With a maximum timestep size of 5000 years, oscillations are present between 120,000 to 220,000 years and occur in the range of $10^{-22}$ to $10^{-14}$ mol/L. With a maximum timestep size of 500 years, oscillations are present between 118,000 to 148,500 years and occurred in the range of $10^{-22}$ to $10^{-18}$ mol/L. With a maximum timestep size of 50 years, no oscillations occurred. The computational time needed for the 5,000-year maximum timestep was 1 hour 27 minutes; this increased to 2 hours 13 minutes for the 500-year maximum timestep, and 7 hours 20 minutes for the 50-year maximum timestep. These results are shown in Figure 8.2 for one simulation, where peak $^{129}$I concentration at Observation Point 8 is plotted against time. Regions of the figure highlighted in red exhibit the discontinuous behavior. Results in this region would be misleading if a concentration threshold less than $10^{-14}$ mol/L were used as a measure of breakthrough. However, as a point of reference, assuming the RB1 dose ingestion model (2 L/day of drinking water) discussed by Mariner et al. [10, Sec. 4.3], this concentration level corresponds to a dose of only about $7 \times 10^{-7}$ mSv/yr, which is very low.

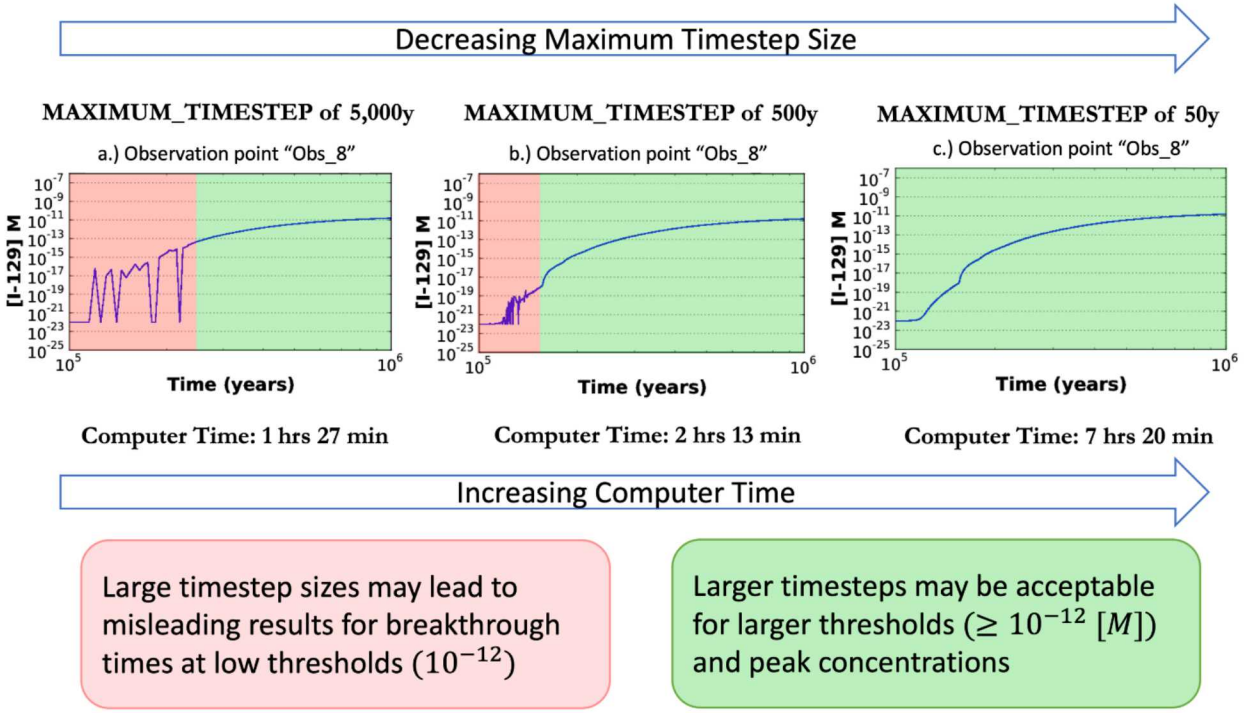



Figure 8.2 Peak $^{129}$I concentration at Observation Point 8 plotted against time with decreasing maximum timestep size and increasing computer time (left to right).

A maximum timestep size of 5,000 years can be used for uncertainty and sensitivity analyses on peak concentration, since peak concentrations occur later in the simulation period than

oscillations occur (i.e., the green region in Figure 8.2). Similarly, breakthrough time for concentrations greater than $10^{-12}$ mol/L tend to occur where there are no oscillations, depending on the observation point. Using the 5,000-year maximum timestep size and limiting the quantities of interest to those that are not sensitive to discontinuities saves about 6 hours of computer time per realization when compared to runs having a maximum timestep size of 50 years. Thus, a maximum timestep of 5000 years is used in all analyses.

## 8.2 Sampling Strategy

### 8.2.1 Uncertainty in Shape of the Distribution

PFLOTRAN assigns a base normalized general corrosion rate ($R$) for each waste package by sampling on a truncated log normal distribution. In this study, the mean and standard deviation of the distribution are treated as epistemically uncertain inputs. Figure 8.3 schematically represents the effect of varying the mean of the distribution and varying the mean and standard deviation of the distribution. When only the mean is uncertain (left), the distributions on corrosion rate are the same shape and width (neglecting the upper truncation limit) and are just shifted left or right. When both the mean and standard deviation are epistemically uncertainty, the result is both a shifting of the distribution left and right and a variation in the height and width of the distribution. UA 1 varies only the mean of the distribution. UA 2 and UA 3 vary both the mean and the standard deviation.

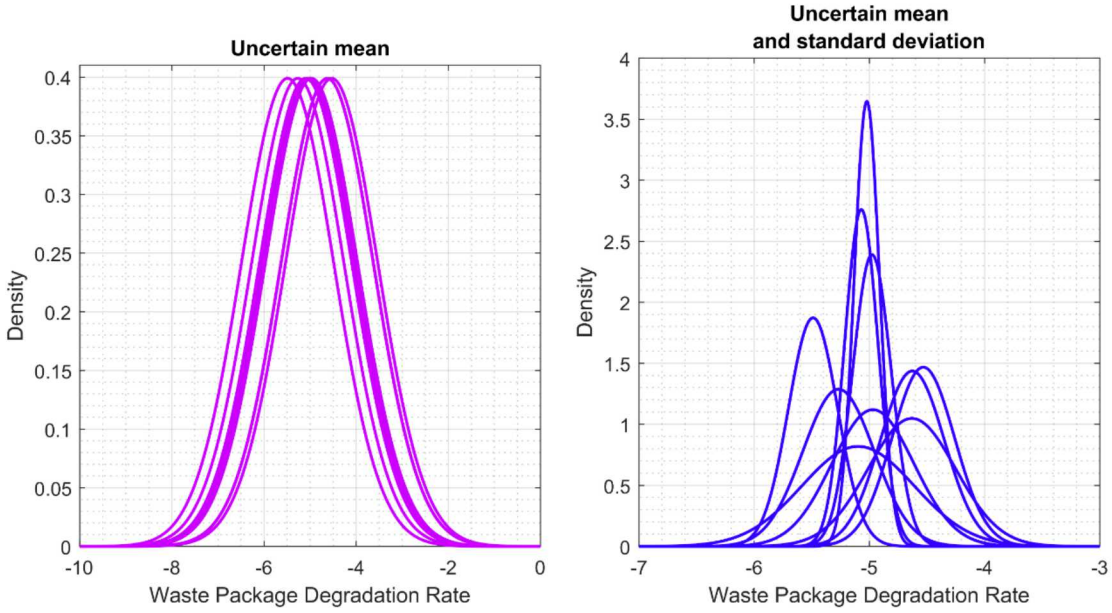



Figure 8.3 Comparison of waste package corrosion rate distribution shapes.

### 8.2.2 Uncertainty in Spatial Location of Waste Package Breach

The value of the random seed used to initiate sampling on the distribution of normalized general corrosion rate affects the spatial location of the waste package with the largest corrosion rate, the second largest, etc.

If the same random seed is used for multiple simulations, then the order in which waste packages breach will be the same even when the timing (controlled by the mean and standard deviation of the distribution) is different. A notional example of this situation is shown in Figure 8.4. In the example, each realization of the epistemic uncertain inputs has its own mean waste package corrosion rate, and the corrosion rates for five waste packages are sampled from the distribution with that mean. The corrosion rates are denoted $WP_t$ where the subscript $t$ denotes the location (a, b, c, d, or e) of the waste package in the repository. Notice that for both realizations, the waste package in location $a$ has the highest corrosion rate, the waste package in location $c$ has the next highest waste package corrosion rate, etc. Therefore, the waste package breach times are different between simulations, but the breach orders are the same with respect to location within the repository for every simulation. In UA 1, this holds true across DFNs and for each epistemic realization within a DFN and removes any spatial variation in waste package breach order that might be expected.

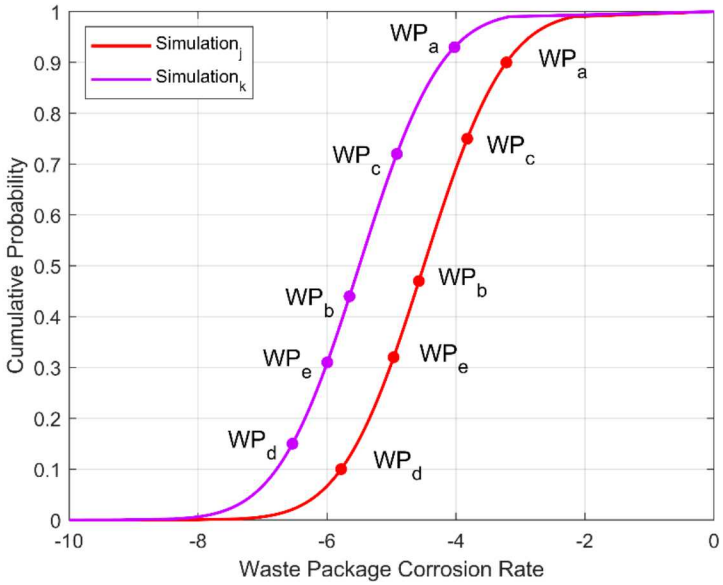



Figure 8.4 Notional example of waste package corrosion rate sampling

If a different random seed is used for different simulations, then the order in which waste packages breach will be different. This strategy was used in UA 3, in which a different random seed was associated with each DFN. A notional example of this is shown in Figure 8.5. The waste package order between the two simulations is different; in simulation j, the order is c, b, a and in simulation k, the order is c, a, b. This strategy effectively incorporates spatial uncertainty at the aleatory level in UA 3.

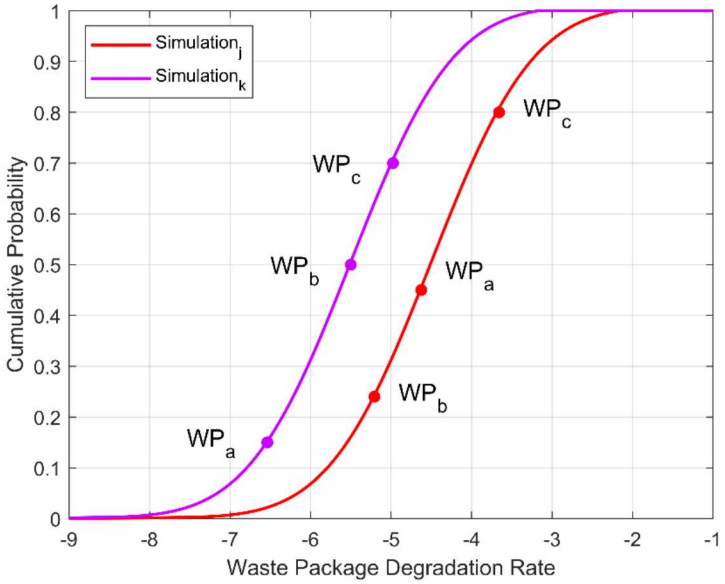



Figure 8.5 Notional example of waste package corrosion rate location assignment randomized at the aleatory level.

## 8.3 Uncertainty Analysis 1 (UA 1)

UA 1 comprises an aleatory loop of 30 DFN realizations, and an epistemic loop of sample size 50 for a total of 1500 simulations. The same epistemic sample is run on all 30 DFN realizations, and the mean of the truncated log normal distribution for waste package corrosion rate is included as an epistemically uncertain input. Epistemically uncertain inputs and their distributions are listed in Table 8.2. Within PFLOTRAN, each simulation uses the same random seed when sampling the waste package corrosion rate distribution, so that regardless of the mean of the distribution (which controls overall timing of waste package breach), the order in which waste packages breach is the same in all simulations – i.e., the location of the waste package with the fastest corrosion rate does not vary, etc. Output variables of interest are maximum $^{129}$I concentration and breakthrough time of $^{129}$I at each of the four observation points in Figure 8.1. Breakthrough time is defined as the time at which $^{129}$I concentration surpasses $10^{-10}$ mol/L.

Repeating the epistemic samples allows explicit separation of effects from the epistemic parameters and effects from the DFN. This comes at the cost of covering less of the epistemic sample space, so overall uncertainties may not accurately reflect the total uncertainty induced by epistemic parameters. For full quantification of uncertainty, larger sample sizes may be desired.

### 8.3.1 Results

Figure 8.6 shows the $^{129}$I concentration contours at 300 years for one of the epistemic realizations of DFN 1. Several waste packages have breached by 300 years, and radionuclide transport through the host rock is dominated by fracture flow.

Table 8.2 Epistemic uncertainty distributions propagated in crystalline reference case UA 1.

| Input | Description | Range | Units | Distribution |
|---|---|---|---|---|
| rateUNF | Fractional dissolution rate of spent (used) nuclear fuel | $10^{-8}$ – $10^{-6}$ | $yr^{-1}$ | log uniform |
| kGlacial | Glacial till permeability | $10^{-16}$ – $10^{-13}$ | $m^2$ | log uniform |
| pBuffer | Buffer porosity | 0.3 – 0.5 | - | uniform |
| kDRZ | DRZ permeability | $10^{-19}$ – $10^{-16}$ | $m^2$ | log uniform |
| kBuffer | Buffer permeability | $10^{-20}$ – $10^{-17}$ | $m^2$ | log uniform |
| meanWPrate | Mean of the truncated log normal distribution on base normalized general corrosion rate ($R$) | -5.5 – (-4.5) | $yr^{-1}$ | uniform |

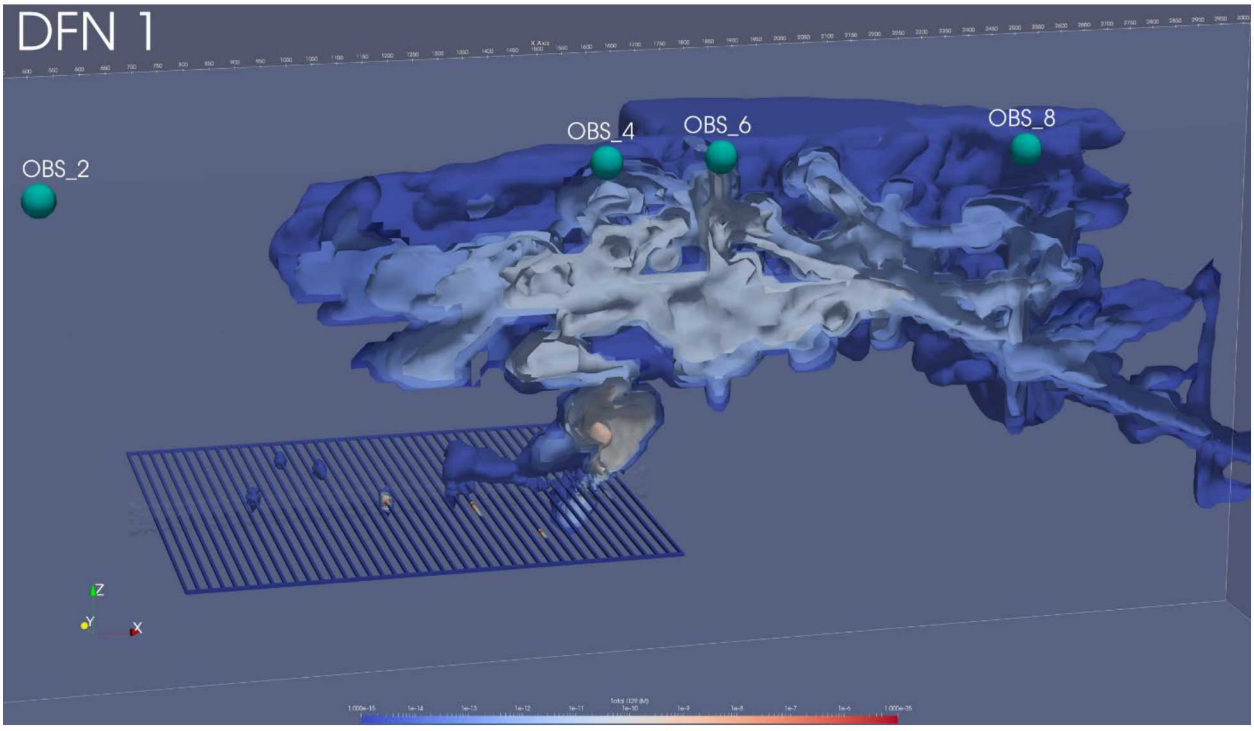



Figure 8.6 $^{129}$I concentration contours for DFN 1 at 300 years for one epistemic realization in UA 1, showing the full repository and the far half of the model domain.

The four horsetail plots in Figure 8.7 show $^{129}$I concentration as a function of time at Observation Points 4 and 8 for all epistemic realizations on two DFNs (DFN 1 and DFN 8). Propagation of the epistemic uncertainty distributions listed in Table 8.2 is responsible for the variation observed for a given DFN. Because the same values of rateUNF, kGlacial, pBuffer, kDRZ, kBuffer, and meanWPrate were used for each DFN, differences between the DFN 1 and DFN 8 $^{129}$I concentrations reflect the effects of the DFN structure, i.e., differences in the distribution of fractures among realizations of the DFN. At Observation Point 4, $^{129}$I concentrations decrease after the initial increase. This occurs for both DFNs but is more pronounced and occurs between a larger range of times than in DFN 8. At Observation Point 8, the breakthrough threshold

concentration of $10^{-10}$ mol/L is surpassed around 1000 yr in many simulations on DFN 1; whereas this concentration is not surpassed until several tens of thousands of years on DFN 8. At Observation Point 8, peak concentrations occur around 40,000 yr on both DFNs, but are almost an order of magnitude higher on DFN 1 than DFN 8. These differences show that the aleatory uncertainty (DFN realization) has an appreciable effect on key quantities of interest like breakthrough time and peak concentration.

Median $^{129}$I concentration versus time at Observation Point 4 for all 30 of the DFNs is shown in Figure 8.8. Differences between the DFNs show that aleatory uncertainty (differences in the location, size, and orientation of fractures) will have a non-negligible effect on quantities of interest like breakthrough time or peak concentration.

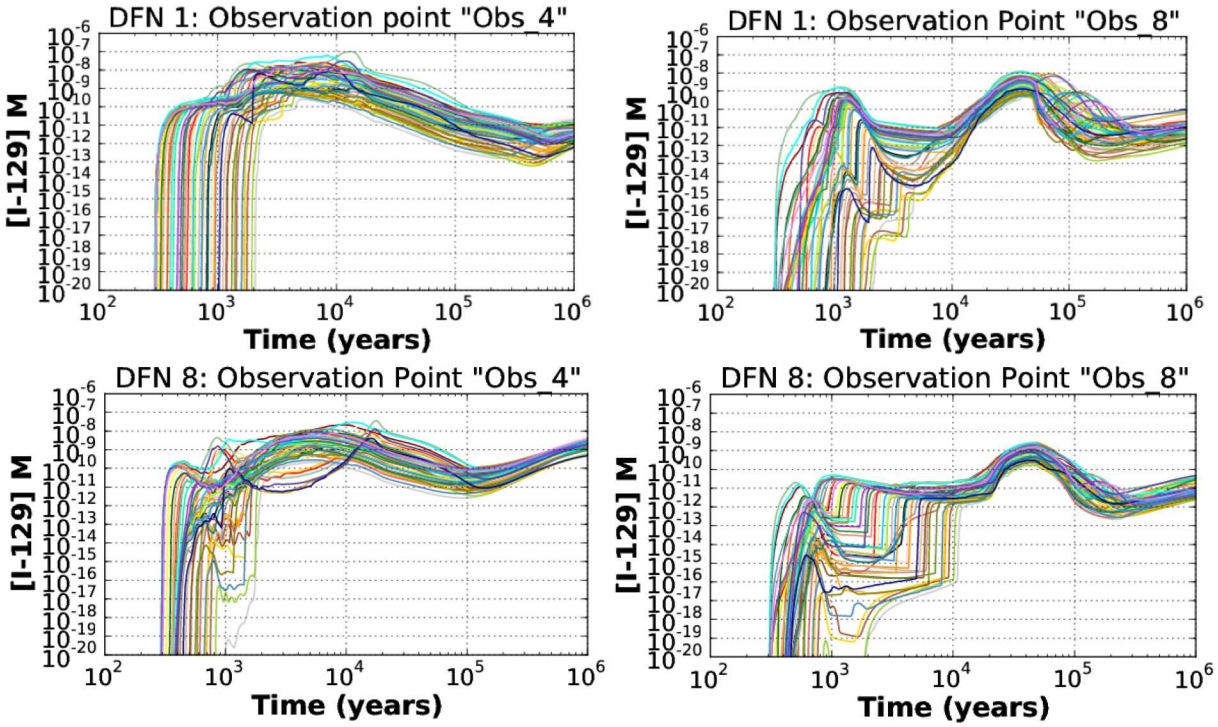



Figure 8.7 $^{129}$I concentration versus time at Observation Points 4 and 8 in DFN 1 and DFN 8 in UA 1.

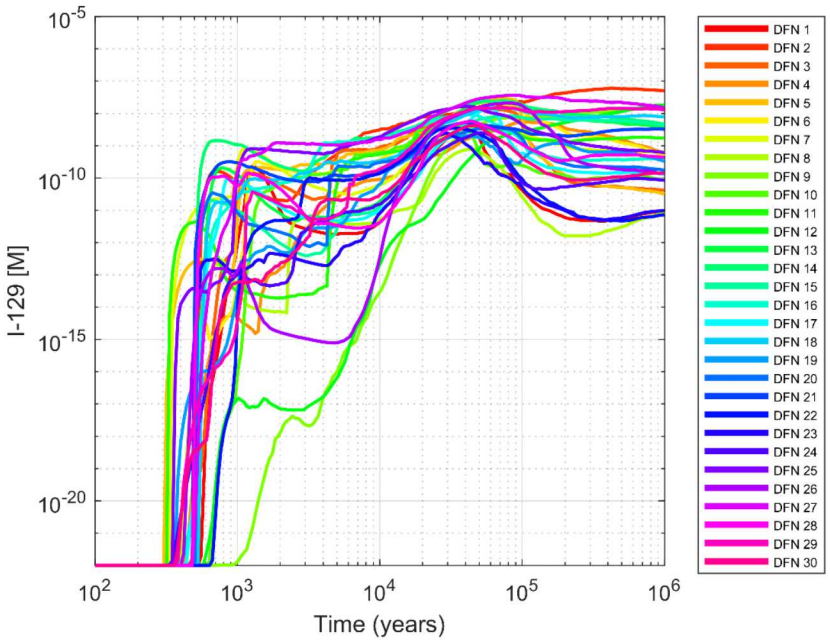



Figure 8.8 Median $^{129}$I concentration over time taken over the epistemic realizations for each DFN at Observation Point 4 in UA 1.

The effect of aleatory uncertainty (DFN realization) is further demonstrated by comparing the full set of epistemic realizations between DFNs. A comparison for DFNs 9 and 13 is shown in Figure 8.9. These specific DFNs are chosen for comparison to highlight difference; some DFNs have more similar $^{129}$I concentrations to each other. The $^{129}$I concentration curves for the epistemic realizations (i.e. individual simulations) within DFN 9 and DFN 13 have similar shapes, but the behavior differs between DFNs. The earliest increase in concentration for DFN 9 is on the order of thousands of years, whereas the earliest increase for DFN 13 is on the order of hundreds of years. Similarly, the peak concentrations differ by orders of magnitude. The difference in $^{129}$I concentrations between the DFNs is representative of the effects of DFN uncertainty (i.e., differences in the distribution of fractures among realizations of the DFN). The difference in $^{129}$I concentrations for a single DFN are representative of the effects of epistemic uncertainty. However, there is clearly less variability between the $^{129}$I concentrations for DFN 9 than for DFN 13. This is because the epistemic and aleatory uncertainties interact.

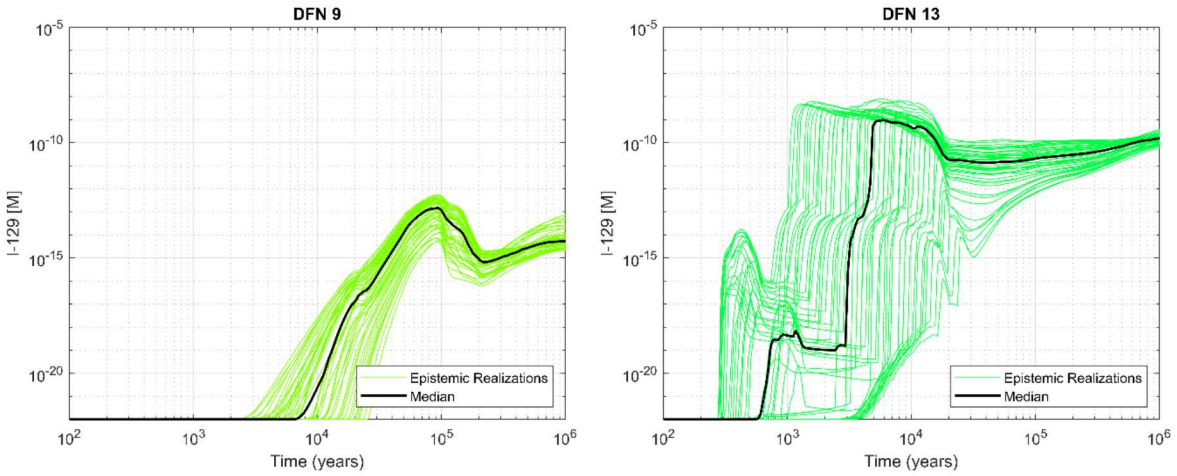



Figure 8.9 Comparison of $^{129}I$ concentrations for DFN 9 and DFN 13 at Observation Point 2 in UA 1.

The effects of both aleatory uncertainty and epistemic uncertainty on quantities of interest can be visualized by calculating the estimated cumulative distribution functions (ECDF) for quantitities of interest over the epistemic realizations, resulting in an ECDF for each DFN. This was done for the breakthrough time for Observation Points 4 and 8 in Figure 8.10. The two types of uncertainty are represented by shape and spread of the curves (epistemic) and in the variation between curves (aleatory). The variation between the CDFs shows the effect of the DFN on breakthrough time. The earliest median breakthrough time across all DFNs is at less than 1000 years, and the latest median breakthrough time is at approximately 10,000 years at Observation Point 4. At Observation Point 8, median breakthrough times across all DFNs are shifted to values between a few thousand years and a few tens of thousands of years. Variation in breakthrough time due to epistemic uncertainties in slightly smaller. At Observation Point 4, differences in breakthrough time among epistemic realizations on the same DFN are generally less than one order of magnitude; at Observation Point 8, differences among epistemic realizations are about one order of magnitude.

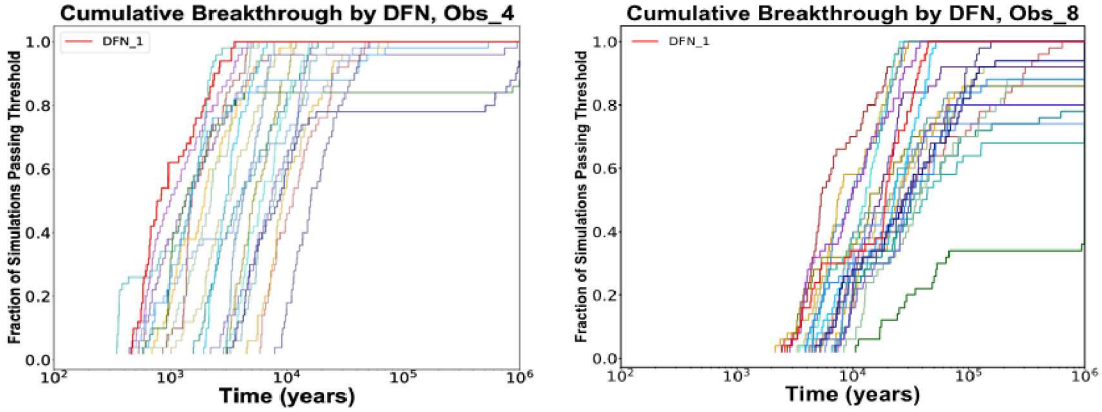



Figure 8.10 Cumulative occurrence of $^{129}I$ breakthrough (>$10^{-10}$ mol/L) at Observation Points 4 (left) and 8 (right) for all DFNs in UA 1.

### 8.3.2 Sensitivity Analysis for UA 1

The results of UA 1 suggest that there may be interaction effects between epistemically uncertain parameters and the DFN. There appears to be a significant interaction effect between the DFN and the mean waste package corrosion rate (meanWPrate). This interaction can be seen in Figure 8.11. The color scale indicates a relationship between waste package corrosion rate and peak $^{129}$I concentrations that is generally evident for most, but not all, DFNs. The DFNs also differ in the magnitude and spread of peak $^{129}$I concentrations.

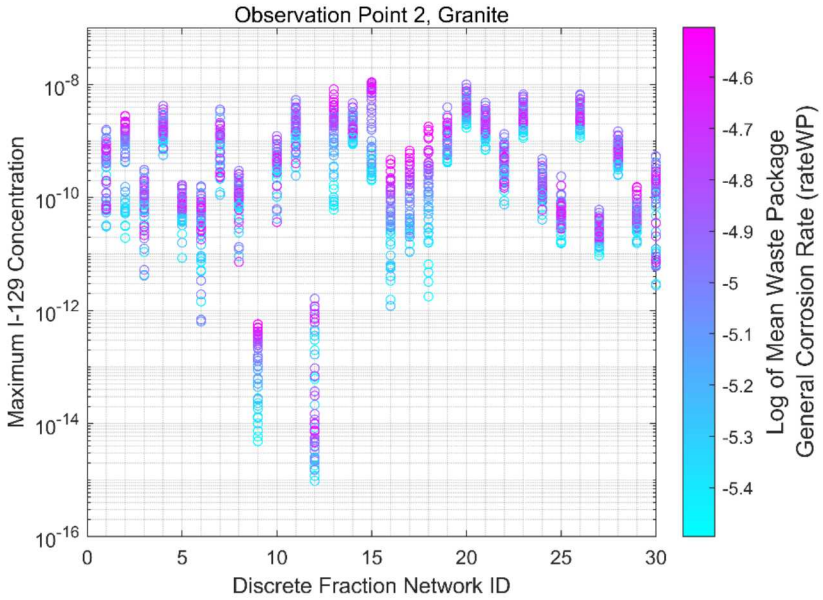



Figure 8.11 Interaction between mean waste package corrosion rate and DFN with respect to peak $^{129}$I concentrations in UA 1.

The conjoint influence of DFN and waste package corrosion rate on $^{129}$I concentration may be further complicated by spatial uncertainty in waste package corrosion rate. In UA 1, waste package corrosion rates were assigned to repository locations in the same order within each realization. There should be spatial uncertainty in the waste package corrosion rates, however, and this uncertainty is incorporated in UA 3 (Section 8.5).

## 8.4 Uncertainty Analysis 2 (UA 2)

UA 2 consists of a single DFN realization (DFN 8 from UA 1). Epistemically uncertain inputs include meanWPrate and stdWPrate (both of which affect the overall timing of waste package breach); a sample size of 50 is used. Epistemically uncertain inputs and their distributions are listed in Table 8.3. Results of UA 2 are compared to results of DFN 8 in UA 1 in order to demonstrate the effect of uncertainty in stdWPrate on uncertainty in maximum concentration and breakthrough time of $^{129}$I.

### 8.4.1 Results

Concentrations of $^{129}$I versus time at Observation Points 4 and 8 are shown in Figure 8.12 (top). The comparable results from UA 1 (Section 8.3) are duplicated on the same scale in the figure (bottom). Notably, the peak concentration over all realizations did not change as a result of including epistemic uncertainty in the standard deviation of the waste package corrosion rate. The

impact of including epistemic uncertainty in the standard deviation is seen in the increased variation in timing. For both DFNs, the range of times at which the initial increase in $^{129}I$ occurs is at least an order of magnitude wider than it was when the standard deviation was fixed.

Table 8.3 Epistemic uncertainty distributions propagated in crystalline reference case UA 2.

| Input | Description | Range | Units | Distribution |
|---|---|---|---|---|
| rateUNF | Fractional dissolution rate of spent (used) nuclear fuel | $10^{-8}$ – $10^{-6}$ | $yr^{-1}$ | log uniform |
| kGlacial | Glacial till permeability | $10^{-16}$ – $10^{-13}$ | $m^2$ | log uniform |
| pBuffer | Buffer porosity | 0.3 – 0.5 | - | uniform |
| kDRZ | DRZ permeability | $10^{-19}$ – $10^{-16}$ | $m^2$ | log uniform |
| kBuffer | Buffer permeability | $10^{-20}$ – $10^{-17}$ | $m^2$ | log uniform |
| meanWPrate | Mean of the truncated log normal distribution on base normalized general corrosion rate (*R*) | -5.5 – (-4.5) | $yr^{-1}$ | uniform |
| stdWPrate | Standard deviation of the truncated log normal distribution | 0 – 0.5 | - | uniform |

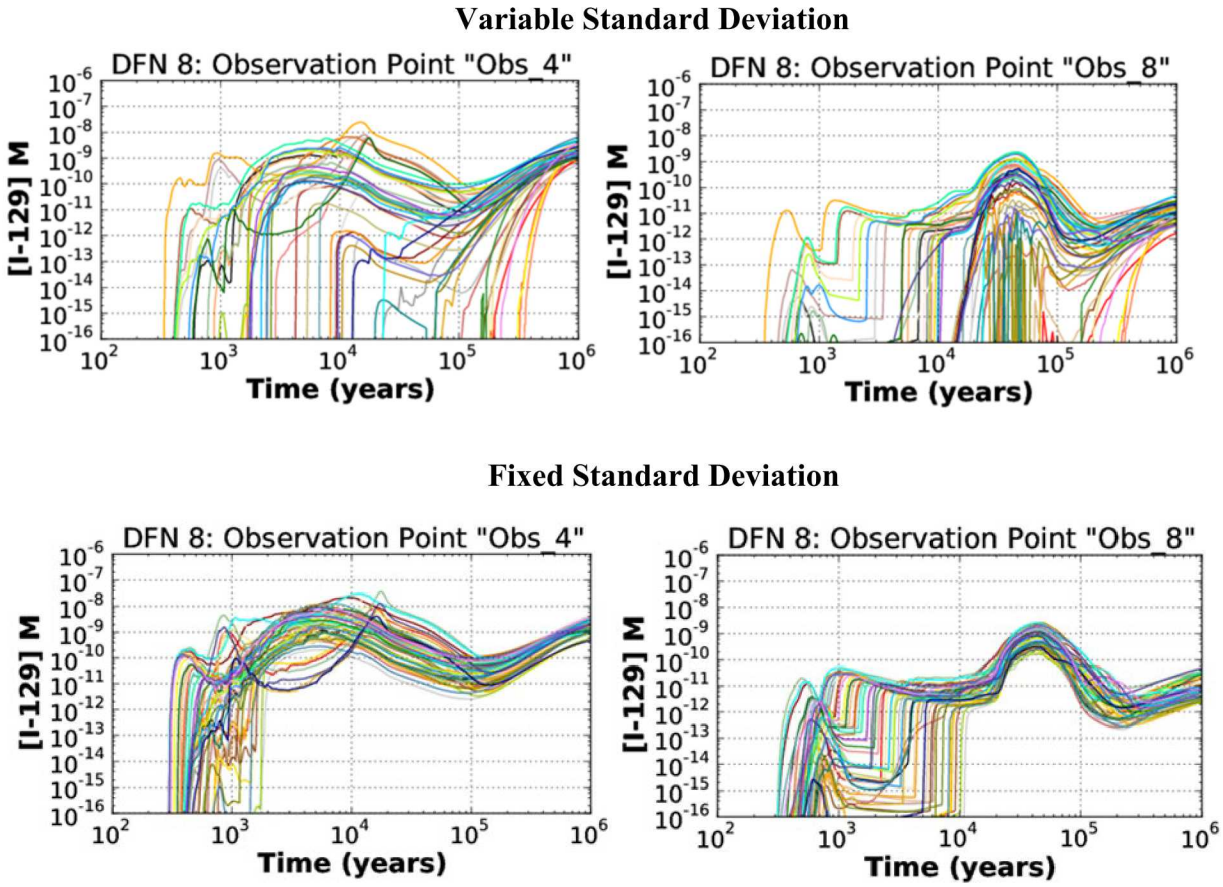



Figure 8.12 $^{129}I$ concentration versus time showing the effect of changing the standard deviation from fixed (top) to variable (bottom) at Observations Points 4 (left) and 8 (right).

The cumulative occurrence of $^{129}$I breakthrough greater than $10^{-10}$ mol/L for Observation Points 4 and 8 are shown in Figure 8.13. The breakthrough at Observation Point 4 occurs more often (and earlier) than that at Observation Point 8. This is due to the location of these observation points relative to the repository. Observation Point 4 is closest to the repository, while Observation Point 8 is farther downgradient from the repository (see Figure 8.1).

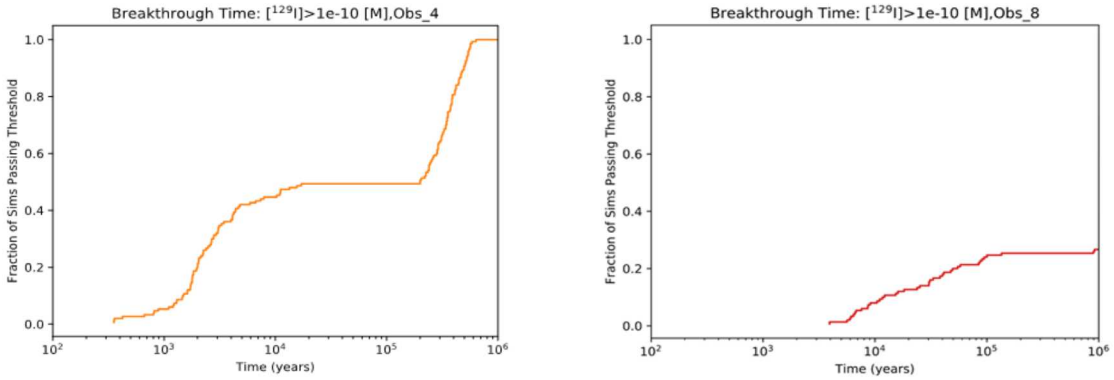



Figure 8.13 Cumulative occurrence of $^{129}$I breakthrough (>$10^{-10}$ mol/L) at Observation Points 4 (left) and 8 (right) for DFN 8 in UA 2.

### 8.4.2 Sensitivity Analysis for UA 2

A sensitivity analysis for UA 2 was performed using polynomial chaos expansion (see Sections 0 and 6.4) in Dakota to estimate Sobol' sensitivity indices. The results for main effects and total effects for peak $^{129}$I concentration at the five observation points are shown in Figure 8.14. Recall that the main effect to a parameter is the proportion of the variance in the quantity of interest that it accounts for alone. The total effect is the sum of the main effect and any proportion of the variance in the output accounted for by higher order relationships between parameters. In general, the waste package corrosion rate distribution parameters (meanWPrate, stdWPrate) have the most effect on uncertainty in peak $^{129}$I concentrations. The kGlacial and rateUNF parameters also appear to have a significant main effect though this does not hold for all observation points. The kDRZ parameter has a total sensitivity index of about 0.2 at Observation Point 4 but has a small main effect (<0.05), which suggests that its effect may be due to interactions. The rateUNF, stdWPrate, and meanWPrate parameters also have noticeably larger total effects than main effect, indicating interaction effects.

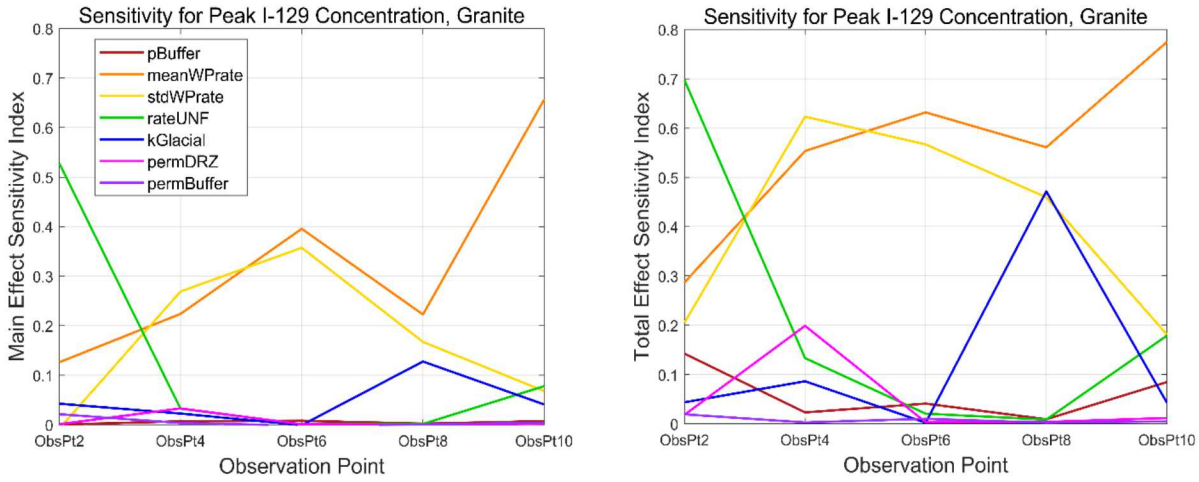



Figure 8.14 Main and total effect sensitivity indices for peak concentration of $^{129}$I in UA 2.

As an example, consider Observation Point 4. The only parameters that appear to have an important main effect on uncertainty in peak concentration of $^{129}$I are meanWPrate and stdWPrate. These parameters have higher total sensitivity indices and all other parameters have total indices less than 0.2. This suggests there is an interaction effect between meanWPrate and stdWPrate, and some interactions with kDRZ and rateUNF. Figure 8.15 shows the mean of the waste package corrosion rate distribution plotted against the standard deviation of that distribution at Observation Point 4. The points are colored by the magnitude of the peak $^{129}$I concentration. This plot shows that higher peak concentrations tend to occur when the mean and standard deviation are both high. This makes sense because the highest corrosion rates, and thus the highest number of breached containers, will be sampled when the corrosion rate distribution is centered at a high value and the standard deviation is large enough to allow samples significantly higher than the mean.

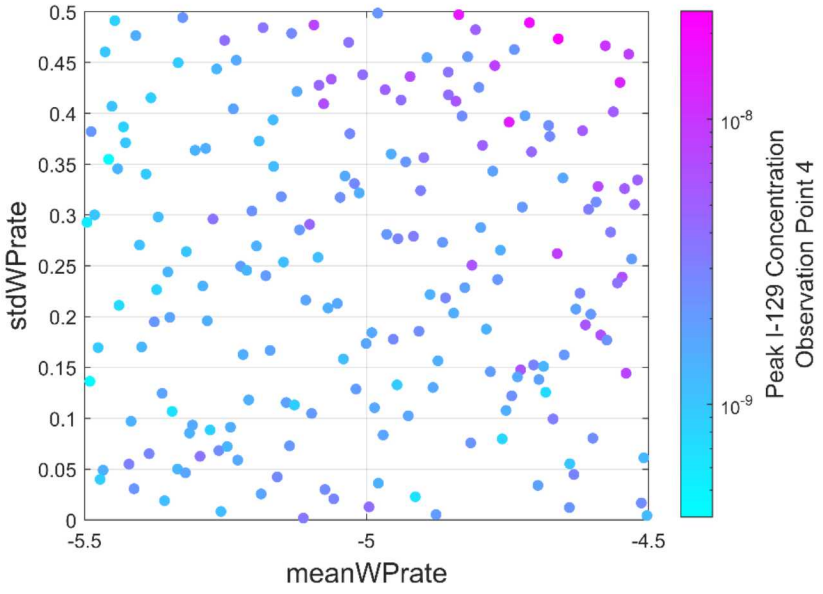



Figure 8.15 Interaction plot for meanWPrate, stdWPrate, and peak $^{129}$I concentration at Observation Point 4 in UA 2.

A similar type of interaction effect on peak $^{129}$I concentration can be seen in Figure 8.16 and Figure 8.17. Higher peak concentrations appear to occur mostly when corrosion rate parameters are high and kDRZ is high, or when the corrosion rate parameters are high and kGlacial is low. However, the pattern is weak and may require more investigation.

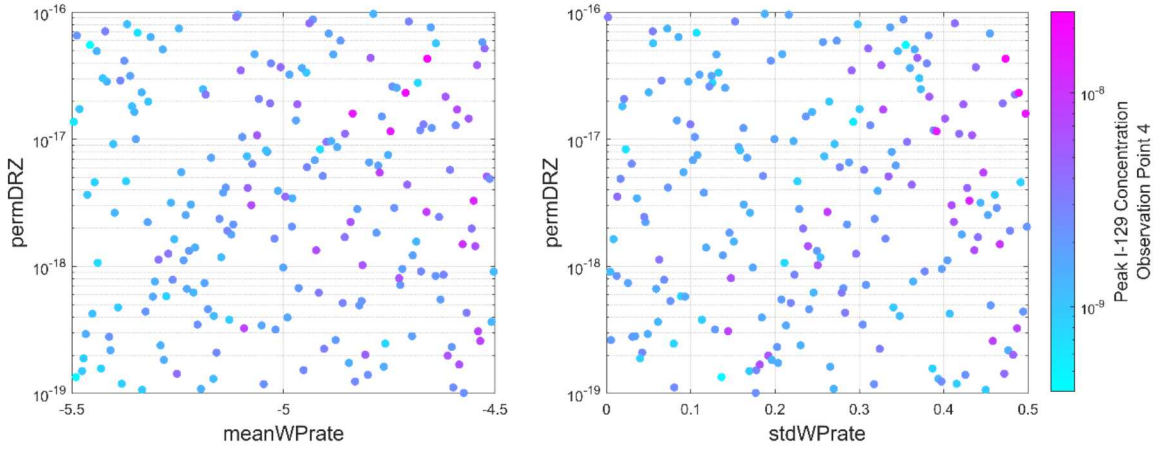



Figure 8.16 Interaction plots for meanWPrate, stdWPrate, kDRZ, and peak $^{129}$I concentration at Observation Point 4 in UA 2.

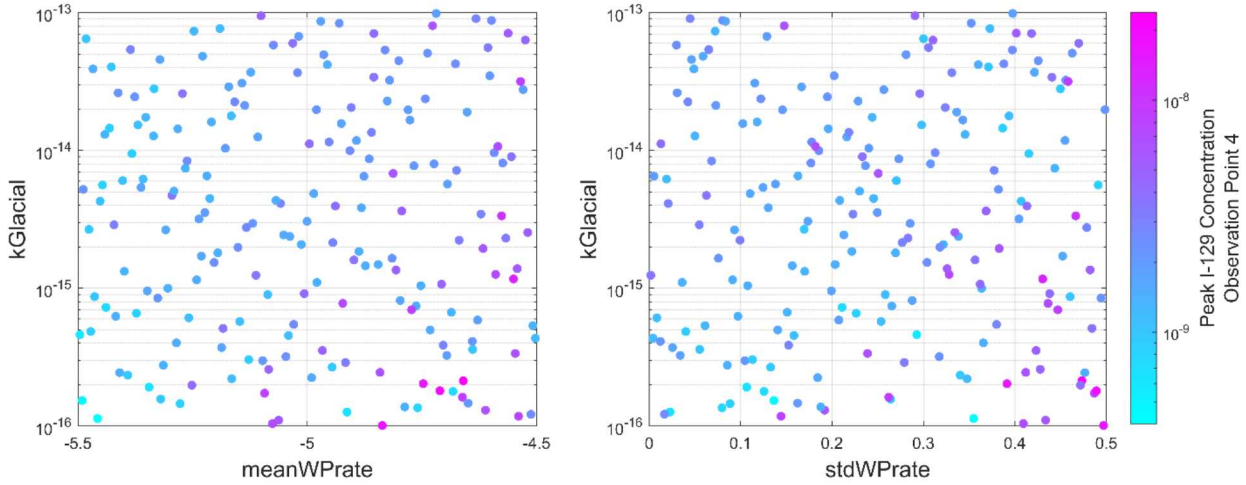



Figure 8.17 Interaction plots for meanWPrate, stdWPrate, kGlacial, and peak $^{129}$I concentration at Observation Point 4 in UA 2.

The effect of kGlacial also appears nonlinear, depending on the observation point, as shown in Figure 8.18. For kGlacial values between $10^{-16}$ and $10^{-15}$, the peak $^{129}$I concentration appears to decrease as kGlacial increases at Observation Point 8 (right), whereas for other values of kGlacial, the peak $^{129}$I concentration appears to increase with kGlacial. However, a similar trend is not apparent at Observation Point 4 (left).

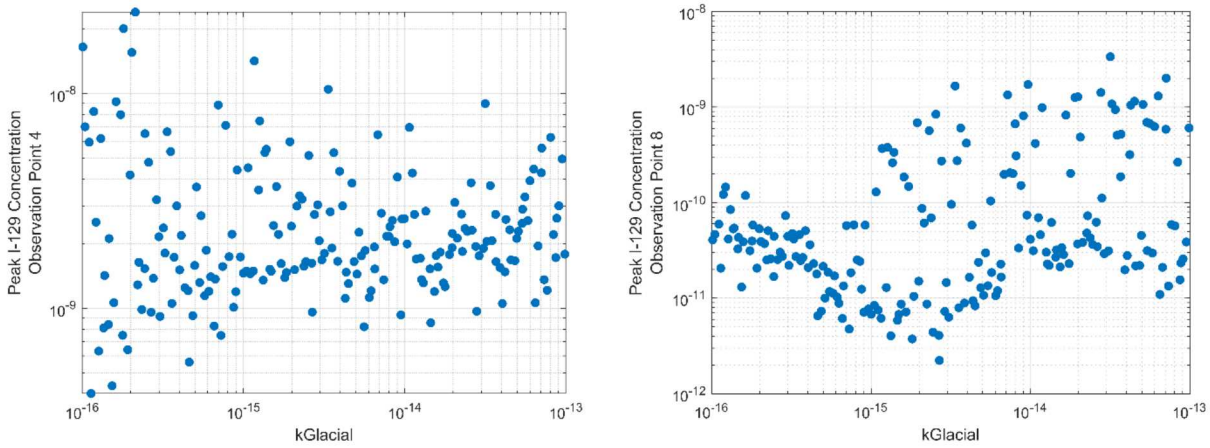



Figure 8.18 Peak $^{129}$I concentration versus kGlacial at Observation Points 4 (left) and 8 (right) in UA 2.

Sensitivity analysis for UA 2 was also performed for breakthrough time, with breakthrough defined as the time at which the concentration exceeds a threshold. Results were generated for multiple thresholds, but only results from $10^{-10}$ M are presented in Figure 8.19. As in the sensitivity results for peak concentration, the waste package corrosion rate parameters appear to have a significant interaction effect. This effect on breakthrough time is shown in Figure 8.20 for both $10^{-12}$ mol/L and $10^{-10}$ mol/L thresholds and appears stronger than the interaction effect on peak concentration (Figure 8.17). When both the mean and standard deviation of the corrosion rate distribution are low, the corrosion rates sampled within the simulation are also low, and hence breakthrough generally occurs later in time.

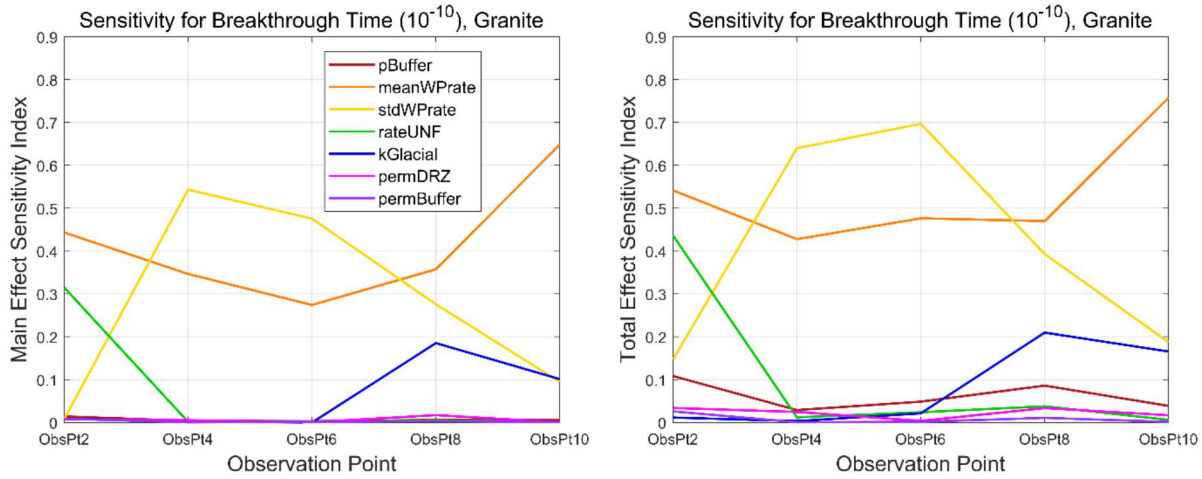



Figure 8.19 Main and total effect sensitivity indices for breakthrough time with a $10^{-10}$mol/L threshold in UA 2.

The results from the sensitivity analysis on breakthrough time are generally consistent across observation points, with main and total effects indices for the corrosion rate distribution parameters dominating all of the other uncertain parameters.

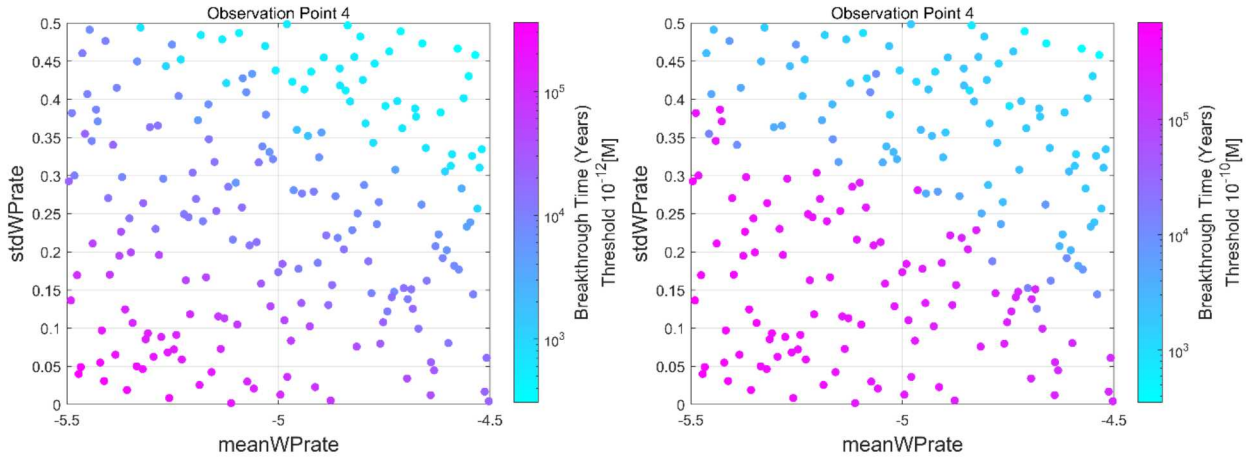



Figure 8.20 Interaction plot for meanWPrate, stdWPrate, and breakthrough time at a threshold of $10^{-12}$ mol/L (left) and $10^{-10}$mol/L (right) at Observation Point 4 in UA 2.

The UA 2 study was designed to test the effects of changing the treatment of uncertainty for the waste package corrosion rate distribution. The sensitivity analysis results show that breakthrough time and peak $^{129}$I concentrations are sensitive to the corrosion rate distribution. This sensitivity is due to main effects from the mean and the standard deviation and due to the conjoint influence of these parameters together defining the corrosion rate distribution. Peak $^{129}$I concentrations are high when the corrosion rate distribution is centered at a high mean value and has a large standard deviation. The breakthrough time is earlier when the corrosion rate distribution parameters support sampling of high corrosion rates, and breakthrough occurs later when the corrosion rate distribution is centered at a low value with a small standard deviation.

## 8.5 Uncertainty Analysis 3 (UA 3)

UA 3 comprises an aleatory loop of sample size 20, and an epistemic loop of sample size 40 for a total of 800 simulations. For each DFN realization in the aleatory loop, a different random seed is used when sampling the normalized general corrosion rate distribution, so that the order of waste package breach associated with each DFN is different. The same epistemic sample is run on all 20 aleatory realizations, and both meanWPrate and stdWPrate are included as epistemically uncertain inputs. Although UA 3 does not comprise the largest number of simulations, it does encompass the largest number of uncertainties. Output variables of interest are maximum $^{129}$I concentration and breakthrough time of $^{129}$I at each of the four observation points in Figure 8.1. Breakthrough time is defined as the time at which $^{129}$I concentration surpasses $10^{-10}$ mol/L.

A subset of the DFNs used in the UA 1 study were used here in the UA 3 study. The epistemically uncertain parameters and their distributions are listed in Table 8.4. The range of the distribution on kGlacial and on stdWPrate are smaller than the ranges of these distributions in UA 2. This change will decrease the contribution of these epistemically uncertain inputs to the uncertainty in the output variables.

As in UA 1, repeating the epistemic samples allows explicit separation of effects from the epistemic uncertainty and effects from the aleatory uncertainty (the DFN and the spatial

distribution of base normalized general corrosion rate) but comes at the cost of covering less of the epistemic sample space.

Table 8.4 Epistemic uncertainty distributions propagated in crystalline reference case UA 3.

| Input | Description | Range | Units | Distribution |
|---|---|---|---|---|
| rateUNF | Fractional dissolution rate of spent (used) nuclear fuel | $10^{-8}$ – $10^{-6}$ | $yr^{-1}$ | log uniform |
| kGlacial | Glacial till permeability | $10^{-15}$ – $10^{-13}$ | $m^2$ | log uniform |
| pBuffer | Buffer porosity | 0.3 – 0.5 | - | uniform |
| kDRZ | DRZ permeability | $10^{-19}$ – $10^{-16}$ | $m^2$ | log uniform |
| kBuffer | Buffer permeability | $10^{-20}$ – $10^{-17}$ | $m^2$ | log uniform |
| meanWPrate | Mean of the truncated log normal distribution on base normalized general corrosion rate ($R$) | -5.5 – (-4.5) | $yr^{-1}$ | uniform |
| stdWPrate | Standard deviation of the truncated log normal distribution | 0.15 – 0.4 | - | uniform |

### 8.5.1 Results

The four horsetail plots in Figure 8.21 show $^{129}I$ concentration as a function of time at Observation Points 4 and 8 for all epistemic realizations on two DFNs (DFN 1 and DFN 8). Propagation of the epistemic uncertainty distributions listed in Table 8.4 is responsible for the variation observed for a given DFN. As in UA 1, because the same epistemic samples were used for each DFN, differences between the DFN 1 and DFN 8 $^{129}I$ concentrations reflect the effects of aleatory uncertainty, i.e., the DFN structure plus spatial uncertainty in the waste package corrosion rates. At Observation Point 4, $^{129}I$ concentrations decrease after the initial increase. This occurs for both DFNs but is more pronounced and occurs between a larger range of times than in DFN 8. Similarly, all of the simulations see an increased $^{129}I$ concentration at Observation Point 4 by 10,000 years in DFN 1, but there are simulations that do not see an increased $^{129}I$ concentration until closer to 100,000 years in DFN 8. At Observation Point 8, the range of times over which $^{129}I$ concentrations increase is wider for DFN 1 than for DFN 8, and the peak concentrations are about an order of magnitude higher. These differences suggest that the aleatory uncertainty will have an appreciable effect on key quantities of interest like breakthrough time and peak concentration.

The effects of epistemic and aleatory uncertainty are also visualized by calculating the empirical cumulative distribution functions for quantities of interest over the epistemic realizations, resulting in an ECDF for each DFN. This was done for the breakthrough time at Observation Points 4 and 8 in Figure 8.22. The variation between the CDFs shows the effect of the DFN and spatial uncertainty on breakthrough timing. The earliest median breakthrough time is around 10,000 years, and the latest median breakthrough time is at over 1,000,000 years for both observation points. The epistemic uncertainties account for the spread in breakthrough times for a single DFN. Heuristically, these are comparable; the variation between ECDFs is around two orders of magnitude and the domain of any single ECDF covers about two orders of magnitude, though there are differences between observation points. In UA 1, differences in breakthrough time due to aleatory uncertainty were generally larger than differences due to epistemic

uncertainty. In UA 3, neither type of uncertainty dominates the other, and the relative contributions from each type of uncertainty can be different depending on the observation point.

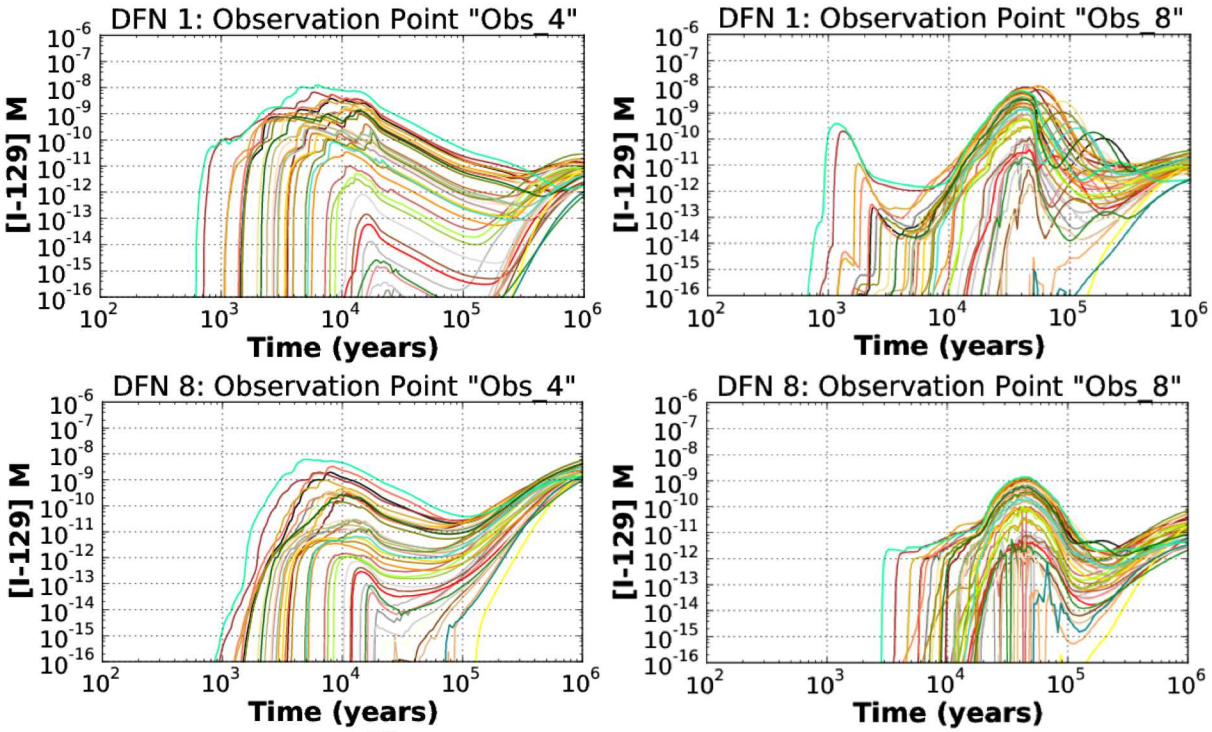



Figure 8.21 Comparison of $^{129}$I concentration as a function of time at Observation Points 4 (left) and 8 (right) for DFN 1 and DFN 8 in UA 3.

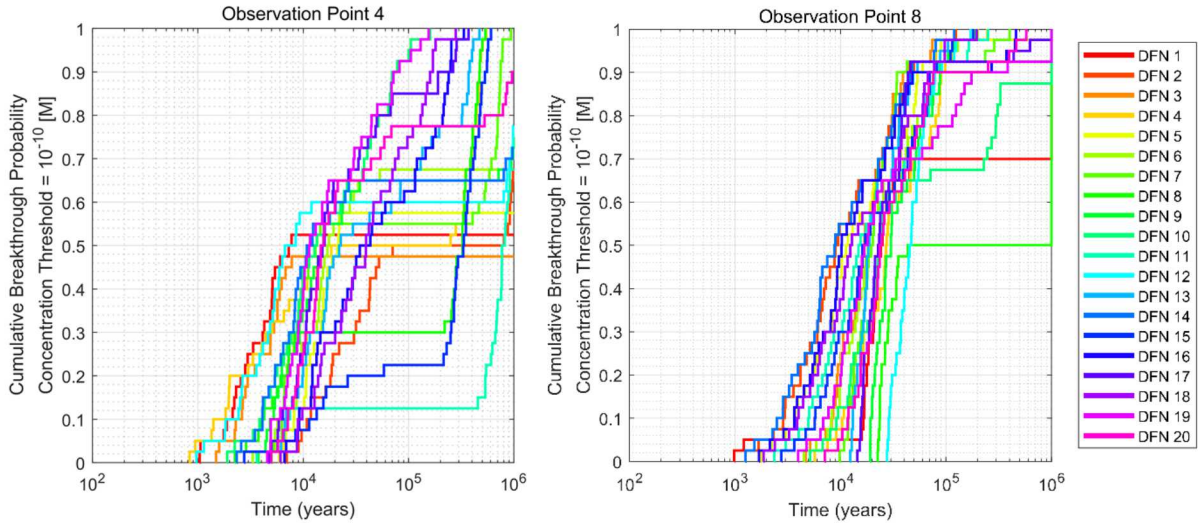



Figure 8.22 Cumulative occurrence of $^{129}$I breakthrough (>$10^{-10}$ mol/L) at Observation Points 4 (left) and 8 (right) for all DFNs in UA 3.

A similar analysis can be made on the basis of Figure 8.23 for peak $^{129}I$ concentrations. The median peak concentration varies by as much as three orders of magnitude between DFNs; and peak concentration varies by as much as three orders of magnitude within a single DFN. In UA 3, aleatory uncertainty (DFN realization and spatial uncertainty in waste package corrosion rate) seems roughly equally as important as epistemic uncertainty in driving uncertainty in both the breakthrough and peak concentration quantities of interest.

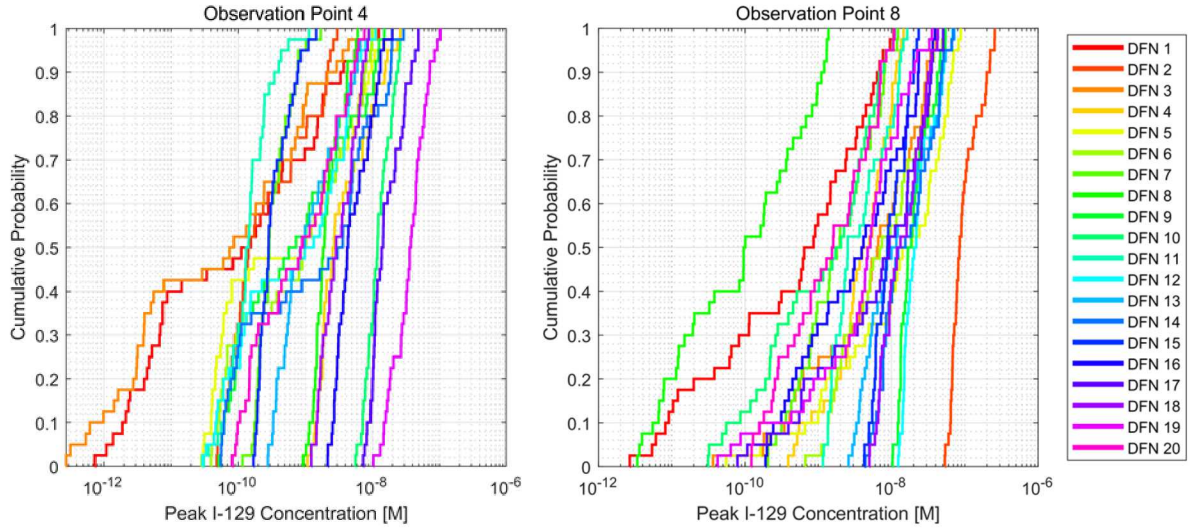



Figure 8.23 ECDFs for peak concentration at Observation Points 4 (left) and 8 (right) for all DFNs in UA 3.

### 8.5.2 Sensitivity Analysis for UA 3

A full sensitivity analysis has not been performed on the simulations for UA 3. However, some insights can be gained from simple scatter plots. The plot in Figure 8.24 shows the combined effect of the waste package corrosion rate parameters on breakthrough time for DFN 8 at Observation Point 8. The plot shows the same trend as Figure 8.20. The same trend occurs for the other DFNs at each observation point, indicating that the introduction of the aleatory uncertainties did not remove the sensitivity of breakthrough to uncertainty in the waste package corrosion rate distribution.

Figure 8.25 shows the effects of each of the waste package corrosion rate parameters separately on the breakthrough time. Across most DFNs, the low values of either parameter are generally associated with later breakthrough times and higher values of either parameter are generally associated with earlier breakthrough times.

This effect can be contrasted with the other epistemic parameters, plotted in Figure 8.26. The relationship between each of these parameters and breakthrough time is less distinct than that for the corrosion rate parameters.

The interaction effect of the waste package corrosion rate parameters on the peak concentration is shown in Figure 8.27 for DFN 8 at Observation Point 4. This is consistent with the result from the UA 2 (Figure 8.15), showing that high values of the corrosion rate parameters are associated with larger peak concentrations.

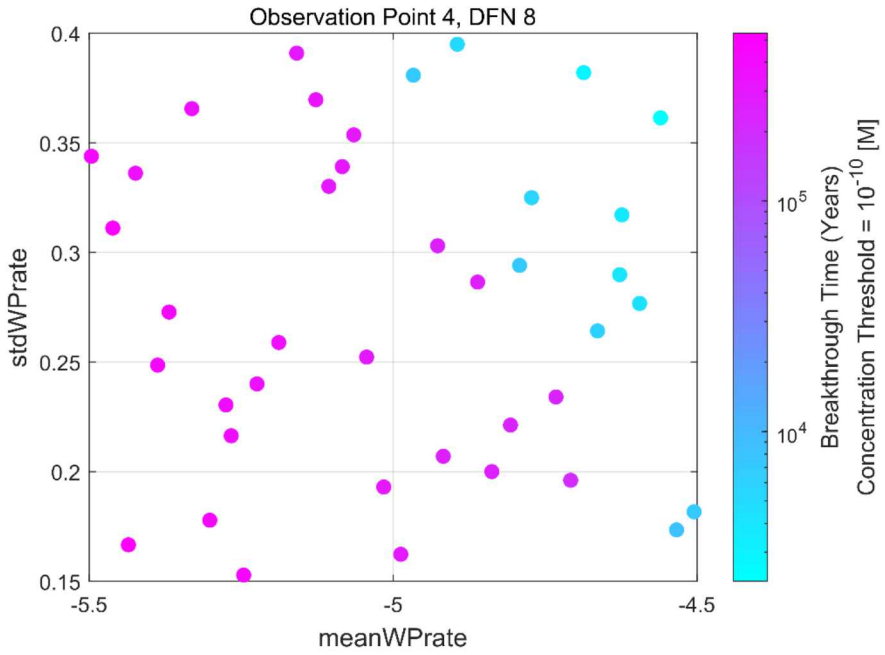



Figure 8.24 Interaction plot for meanWPrate, stdWPrate, and breakthrough time at Observation Point 4 for DFN 8 in UA 3.

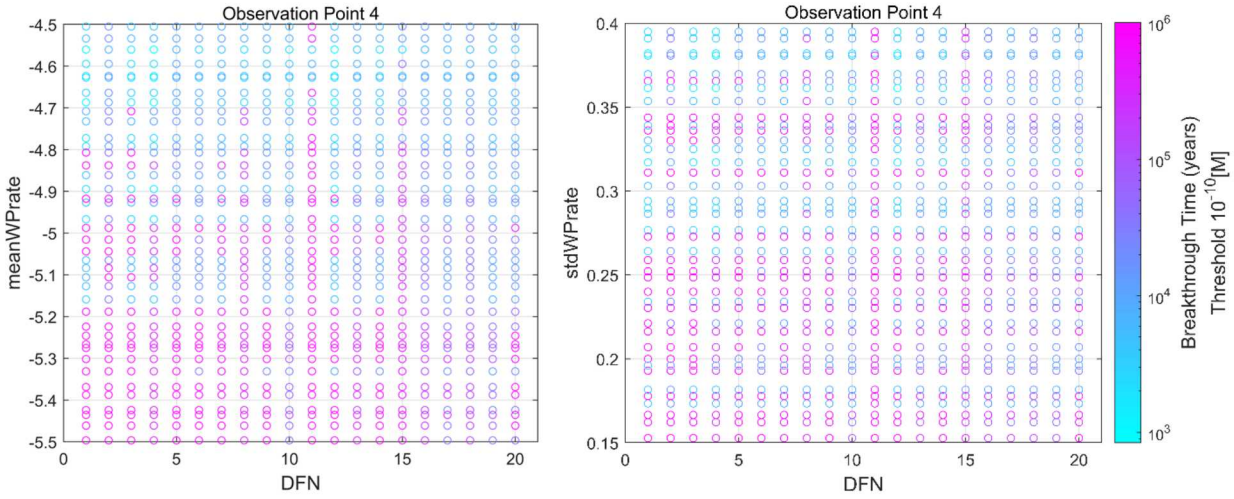



Figure 8.25 Waste package corrosion rate parameters colored by breakthrough time at Observation Point 4 for all DFNs in UA 3.

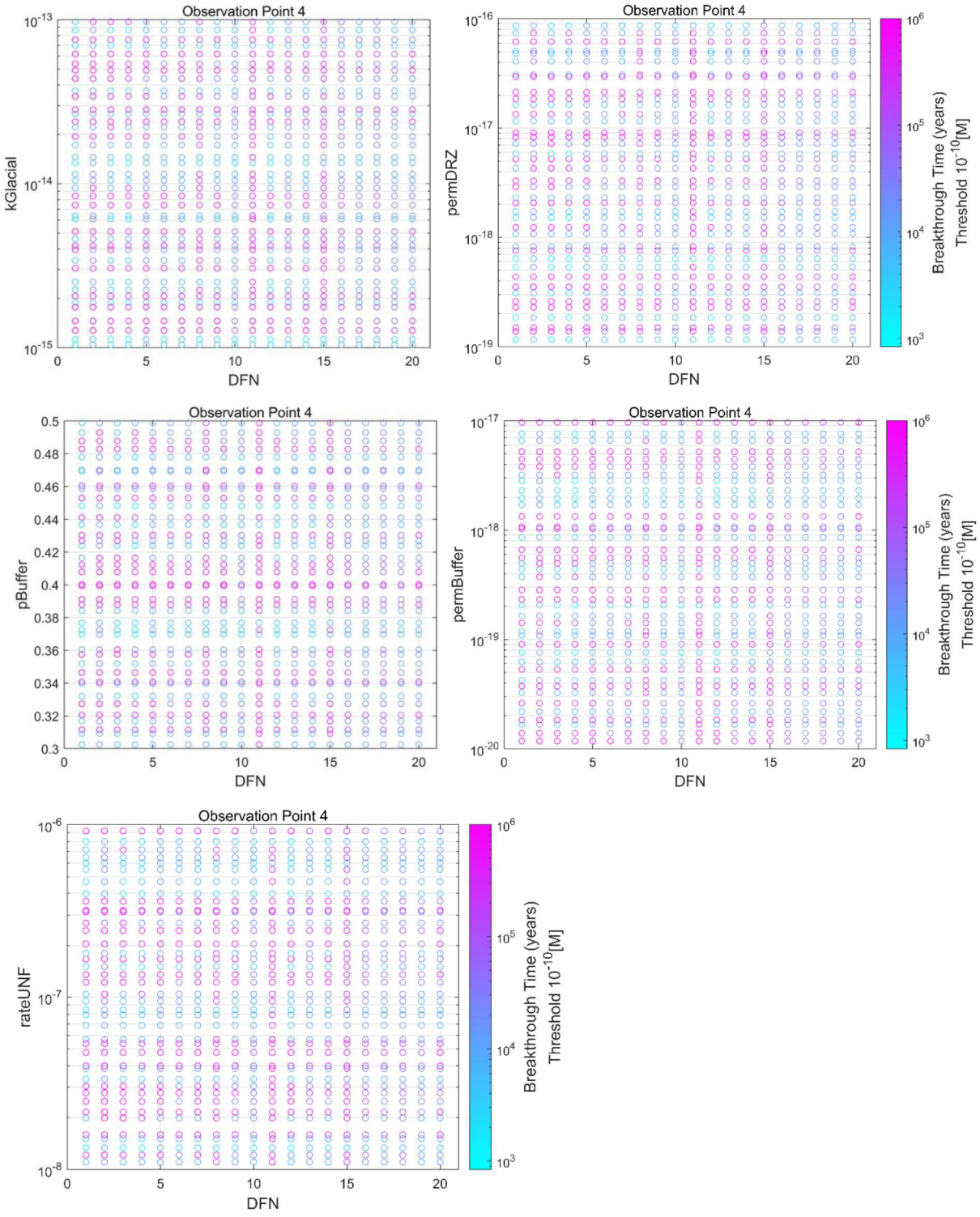



Figure 8.26 Selected epistemically uncertain inputs colored by breakthrough time at Observation Point 4 for all DFNs in UA 3.

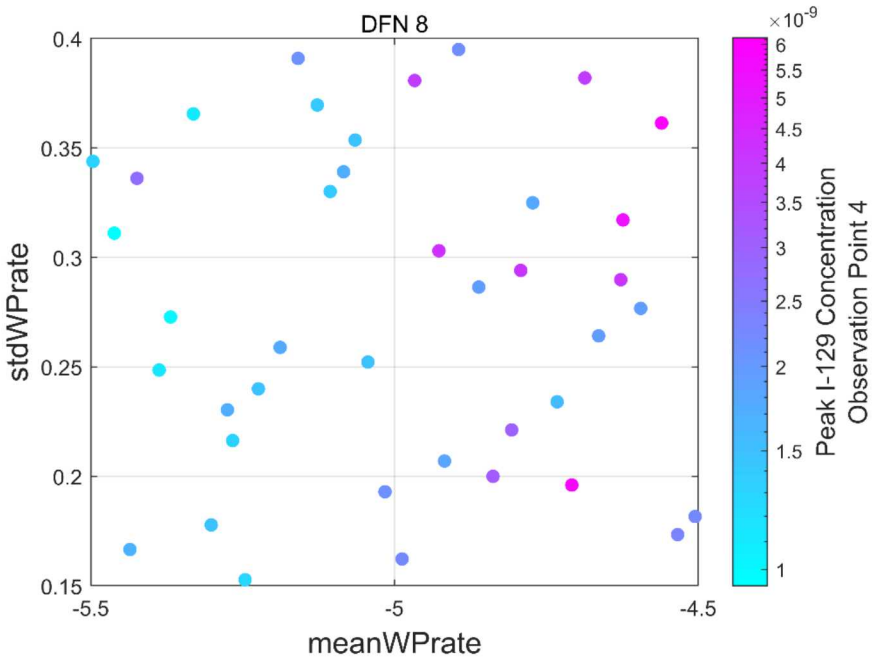



Figure 8.27 Interaction plot for meanWPrate, stdWPrate, and peak $^{129}I$ concentration at Observation Point 4 for DFN 8 in UA 3.

In UA 3, both aleatory and epistemic uncertainties are significant and neither dominates the overall uncertainty in quantities of interest (breakthrough time; peak concentration). Initial sensitivity results show that of the epistemically uncertain inputs, the parameters defining the normalized general corrosion rate distribution have the most influence on breakthrough time and peak concentration. This conclusion is supported by the more extensive sensitivity analysis performed for UA 2.

## 8.6 Conclusions for Crystalline Reference Case UA

Three new uncertainty analyses performed for a reference case repository for commercial spent nuclear fuel in fractured crystalline host rock indicate that both aleatory uncertainty associated with spatial heterogeneity and epistemic uncertainty associated with waste package corrosion rates contribute significantly to uncertainties in output variables of interest including peak concentrations and breakthrough times of $^{129}I$ at several observation points in the model domain. With a constant and relatively large standard deviation for the distribution of normalized general corrosion rates (as in UA 1), aleatory uncertainty associated with stochastically-generated DFNs has a larger influence on uncertainty in quantities of interest than does epistemic uncertainty associated with waste package corrosion rate. However, when the standard deviation of the corrosion rate distribution is included as an epistemically uncertain parameter, as in UA 3, epistemic uncertainty has as large an influence on the output as aleatory uncertainty. A sensitivity analysis performed across epistemically uncertain inputs on a single DFN realization (UA 2) confirms the contribution of uncertainties in mean and standard deviation of the corrosion rate distribution to uncertainty in the output.

Fracture distribution was treated as an aleatory uncertainty because it is assumed that the uncertainty associated with heterogeneous fracture distribution in the subsurface cannot be

reduced. However, it may be possible to reduce uncertainty associated with spatial heterogeneity by avoiding emplacement of waste packages at locations where connected fractures intersect the repository.

When site-specific data is available and design choices are made, the state of knowledge regarding uncertainty in the waste package corrosion rate can be improved through well-designed experimental work aimed at measuring corrosion rates and understanding corrosion processes in the site-specific system.

## 8.7 References: Chapter 8

1. Stein, E. R., J. M. Frederick, G. E. Hammond, K. L. Kuhlman, P. E. Mariner and S. D. Sevougian (2017). "Modeling Coupled Reactive Flow Processes in Fractured Crystalline Rock," *Proceedings of the 16th International High-Level Radioactive Waste Management (IHLRWM 2017) Conference*, Charlotte, North Carolina, April 9-13, American Nuclear Society.
2. Joyce, S., L. Hartley, D. Applegate, J. Hoek and P. Jackson (2014). "Multi-scale groundwater flow modeling during temperate climate conditions for the safety assessment of the proposed high-level nuclear waste repository site at Forsmark, Sweden," *Hydrogeology Journal*, **22**(6):1233-1249
3. Hyman, J. D., S. Karra, N. Makedonska, C. W. Gable, S. L. Painter and H. S. Viswanathan (2015). "dfnWorks: A discrete fracture network framework for modeling subsurface flow and transport," *Computers & Geoscience*, **84**:10-19.
4. Mariner, P.E., E.R. Stein, J.M. Frederick, S.D. Sevougian, G.E. Hammond, and D.G. Fascitelli (2016), *Advances in Geologic Disposal System Modeling and Application to Crystalline Rock*, FCRD-UFD-2016-000440, SAND2016-9610 R, Sandia National Laboratories, Albuquerque, NM.
5. SKB 2006. *Long-term safety for KBS-3 repositories at Forsmark and Laxemar – a first evaluation*. SKB TR-06-09. Svensk Kärnbränslehantering AB, Stockholm, Sweden.
6. Hammond, G.E., P.C. Lichtner, C. Lu and R.T. Mills (2011). "PFLOTRAN: Reactive Flow and Transport Code for Use on Laptops to Leadership-Class Supercomputers." *Groundwater Reactive Transport Models*. F. Zhang, G. T. Yeh and J. Parker. Bentham Science Publishers.
7. Lichtner. P.C. and G.E. Hammond (2012). *Quick Reference Guide: PFLOTRAN 2.0 (LA-CC-09-047) Multiphase-Multicomponent-Multiscale Massively Parallel Reactive Transport Code*. LA-UR-06-7048. December 8, 2012. Los Alamos National Laboratory, Los Alamos, New Mexico.
8. Adams, B.M., Bauman, L.E., Bohnhoff, W.J., Dalbey, K.R., Ebeida, M.S., Eddy, J.P., Eldred, M.S., Hough, P.D., Hu, K.T., Jakeman, J.D., Stephens, J.A., Swiler, L.P., Vigil, D.M., and Wildey, T.M., *Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 User's Manual*, SAND2014-4633, July 2014. Other versions released December 2009 (Version 5.0), December 2010 (Version 5.1), November 2011 (Version 5.2), February 2013 (Version 5.3), May 2013 (Version 5.3.1), November 2013 (Version 5.4), November 2014 (Version 6.1), May 2015 (Version 6.2), November 2015 (Version 6.3), May 2016 (Version 6.4), November 2016 (Version 6.5), May 2017 (Version 6.6), November

2017(Version 6.7), May 2018 (Version 6.8), November 2018 (Version 6.9), May 2019 (Version 6.10).

9. Sevougian, S. D., E. R. Stein, G. E. Hammond, P. E. Mariner, J. M. Frederick, and E. Basurto 2018. "Simulating the Effect of Fracture Connectivity on Repository Performance with GDSA Framework – 18589," in *Proceedings of WM2018 Conference*, March 18-22, 2018, Phoenix, AZ, USA, SAND2017-12677C.
10. Mariner, P.E., Stein, E., Frederick, J.M, Sevougian, S.D., and Hammond, G.E. *Advances in Geologic Disposal System Modeling and Shale Reference Cases*. SFWD-SFWST-2017-000044 / SAND2017-10304R. Sandia National Laboratories, Albuquerque, NM (2017).

# 9. EXPLORATION OF NEW UQ/SA DEVELOPMENTS

There have been many exciting developments in the fields of uncertainty quantification and sensitivity analysis the past twenty years. One is the development of many types of surrogates or metamodels, which are presented in Chapter 6. These have allowed more advanced UQ/SA methods to be used which could not be used previously because of the thousands of function evaluations they require. The use of Sobol' variance-decomposition to perform sensitivity analysis (presented in Chapter 5) has been enabled by surrogate model development.

Additionally, the last two decades have seen a large increase in the use of Bayesian methods to perform parameter calibration. The posterior parameter distributions generated as part of Bayesian calibration can be used to perform forward uncertainty propagation. Finally, there are some new research areas outlined below. These include dimension reduction methods which address functional decomposition of time series data; the CUSUNORO sensitivity approach; importance sampling; Bayesian methods; and multi-fidelity approaches. This chapter provides a brief introduction to these topics. It is our goal to investigate and understand these methods with respect to their application to GDSA in the coming years.

## 9.1 Dimension reduction

There are a variety of methods which take "functional" or "field" data that involves a response as a function of time and/or spatial location and reduces it in some manner. The idea of creating surrogate representations as part of temporal or spatial field data is addressed in some detail in [1-5]. These methods are very applicable to GDSA, because many geologic disposal safety assessments involve prediction of various radionuclide quantiies as a function of time and/or location. These radionuclide quantities are often correlated with each other (e.g., the concentration of $^{129}$I at a particular location at 5000 years is correlated with the concentration at that location at 6000 years. See Figure 7.3 for example data).

One classical approach to reducing the dimension of functional data is principal components analysis (PCA). To understand the basics, first assume that an LHS sample has been run to compute the functional or field responses. For notation purposes, there are *d* input parameters, N samples, and the field length is L. This will result in a set of field responses which can be put into a matrix **X** of dimension N*L. For example, for 100 samples of a 1000 length field, **X** is 100*1000.

To perform a PCA on the field responses, one first centers the mean of each column of **X**. That is, take **X** and subtract the mean of that column from each value in that column. PCA is then performed on the mean-corrected **X**. This ensures the columns have a sample mean of zero which simplifies the variance calculations.

The idea in PCA is to construct a linear representation of variable values to highlight the variance present in the data. To find the first principal component, define $f_i$ as a linear combination of variable values: $f_i = \sum_{j=1}^{L} \beta_j x_{ij} = X_i \boldsymbol{\beta}$. Then, find the set of (L*1) coefficients $\boldsymbol{\beta_1}$ such that the linear combination $f_{i1} = X_i \boldsymbol{\beta_1}$ has the largest possible mean square (defined as $\frac{\sum_i f_{i1}^2}{N}$) subject to the constraint that the $\|\boldsymbol{\beta_1}\|^2 = 1$. The optimal $\boldsymbol{\beta_1}$ when applied to all observations will result in a vector $\mathbf{Y_1}$ of dimension N*1: $Y_1 = \boldsymbol{X\beta_1}$. This vector identifies the strongest and most important modes of variation. The process is repeated with subsequent steps to determine the sets

of coefficients $\boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \ldots$ that define the second, third, and subsequent coefficient vectors. The coefficient vectors are called "loadings or "factor loadings" and the Y vectors are called "factor scores" or "principal component scores."

In summary, PCA results in a system:

$$\boldsymbol{Y} = \boldsymbol{X\beta} \tag{9.1.1}$$

To calculate the coefficients $\boldsymbol{\beta}$, it is necessary to calculate the eigenvectors of the covariance matrix of **X.** That is, calculate the matrix:

$$\boldsymbol{V} = \frac{\boldsymbol{X^T X}}{N} \tag{9.1.2}$$

Then, perform a Singular Value Decomposition (SVD) on the covariance matrix **V** to obtain the eigenvectors. Typically, the process is only carried out until the M principal components are identified that account for a certain percentage of the variance. The first eigenvector (with the largest eigenvalue) of **V** corresponds to $\boldsymbol{\beta_1}$, the second eigenvector of **V** corresponds to $\boldsymbol{\beta_2}$, etc. Note that the full $\boldsymbol{\beta}$ can be of dimension L*L but typically is smaller, L*M. Also, note that to perform a prediction in the original space, the loadings and factor scores are used. Equation 9.1.3, derived from Equation 9.1.1, uses the fact that the inverse of an orthogonal matrix is its transpose:

$$\boldsymbol{X} = \boldsymbol{Y\beta^T} \tag{9.1.3}$$

"Functional Data Analysis" by J.O. Ramsay and B.W. Silverman, Springer, 2005 is an excellent reference for PCA on functional data (fPCA). This book goes into more treatment of methods with functional data (e.g., treating the covariance matrix of the data as a function of time), but essentially the data is being treated as discretized values and thus can be used in a standard PCA.

Finally, note that PCA is not just a way to obtain efficient representations of the original data as given by the product of the M most significant factor scores times the loadings as presented in Equation 9.1.3. One can also use the PCA as a way to generate new predictions. This is done by creating a predictive model which involves summing the eigenvectors but treating their coefficients (the factor scores) as uncertain variables which are functions of the original uncertain parameters used to generate the sample matrix **X**.

More specifically: one can construct Gaussian process surrogates for the factor scores of the M principal components. The GP surrogates will be function of the uncertain inputs. The idea is that PCA has just been performed on (for example) the covariance matrix of 100 samples. Typically, those 100 samples will be generated by sampling over some *d* uncertain input parameters $\boldsymbol{u}$, so there should be a mapping from $\boldsymbol{u}$ to the field data, specifically to the loading coefficients and the factor scores. The goal is to create a surrogate model so that one can predict a new time series or set of field data for parameter values $\boldsymbol{u}$ that are not in the particular sample of 100. A prediction is given as:

$$\boldsymbol{X_{pred} = GP_1(u)\beta_1 + GP_2(u)\beta_2 + \cdots GP_M(u)\beta_M + \mu_X} \quad (9.1.4)$$

where the estimate of each factor score is a GP that is a function of the *d* uncertain input parameters, denoted $\boldsymbol{u}$. To obtain the data to estimate the GPs, there are (for example) 100 samples of $\mathbf{Y_1}$ and a corresponding 100 samples of $\boldsymbol{u}$, so $\mathbf{GP_1}$ is a surrogate for $\mathbf{Y_1}(\boldsymbol{u})$. Note that $\boldsymbol{\mu_X}$ is the mean that was originally subtracted from the original data matrix $\mathbf{X}$. Equation 9.1.4 is a generalization of Equation 9.1.3 in the case of new inputs $\boldsymbol{u_{new}}$.

The process used to build such a surrogate model has been refined specifically for uncertainty quantification of functional data. The concept developed by Tucker et. al. is to use fPCA twice to build the surrogate: once to characterize the horizontal uncertainty and once to characterize the vertical uncertainty. The resulting probabilistic models on horizontal and vertical uncertainty are combined to build the surrogate.

As in the general PCA case, the functions being modeled are aligned to the mean before application of PCA. Alignment to the mean is performed using warping functions (e.g., time-warping for time series data). Each function has an associated warping function that stretches and shrinks portions of the function with respect to the abscissa to minimize a measure of distance to the mean. Because this warping function removes horizontal uncertainty, its form characterizes the uncertainty it has removed. fPCA is applied to induce a probabilistic model on the warping functions to account for horizontal uncertainty. fPCA is applied separately to the aligned functions to induce a probabilistic model that characterizes the vertical uncertainty.

The details involved in this construction ensure that each step is invertible. The probabilistic models and vertical and horizontal uncertainty can be sampled separately to predict a new set of warping functions and aligned functions, which together define a new prediction for the population.

One of the primary benefits of this construction is that separating uncertainty and aligning to the Karcher mean rather than a cross-sectional mean prevents oscillations in functions from cancelling each other out. A notional example is provided in Figure 9.1. The population of functions being modeled have horizontal uncertainty that results in a cross-sectional mean that loses the maxima and minima characteristics of the entire population. This mean is not representative. The Karcher mean, however, is less affected and maintains a shape more consistent with the population.

In general, these dimension reduction methods which construct functional approximations can be directly applied to GDSA applications which involve time-series output. For example, many geologic disposal safety assessments involve prediction of various radionuclide quantiies as a function of time and/or location. Although the output of one simulation may involve predictions of a radionuclide at 10,000 time points, there is often correlated information in this vector of timepoints (e.g., a function over time) which is well-suited to dimension reduction.

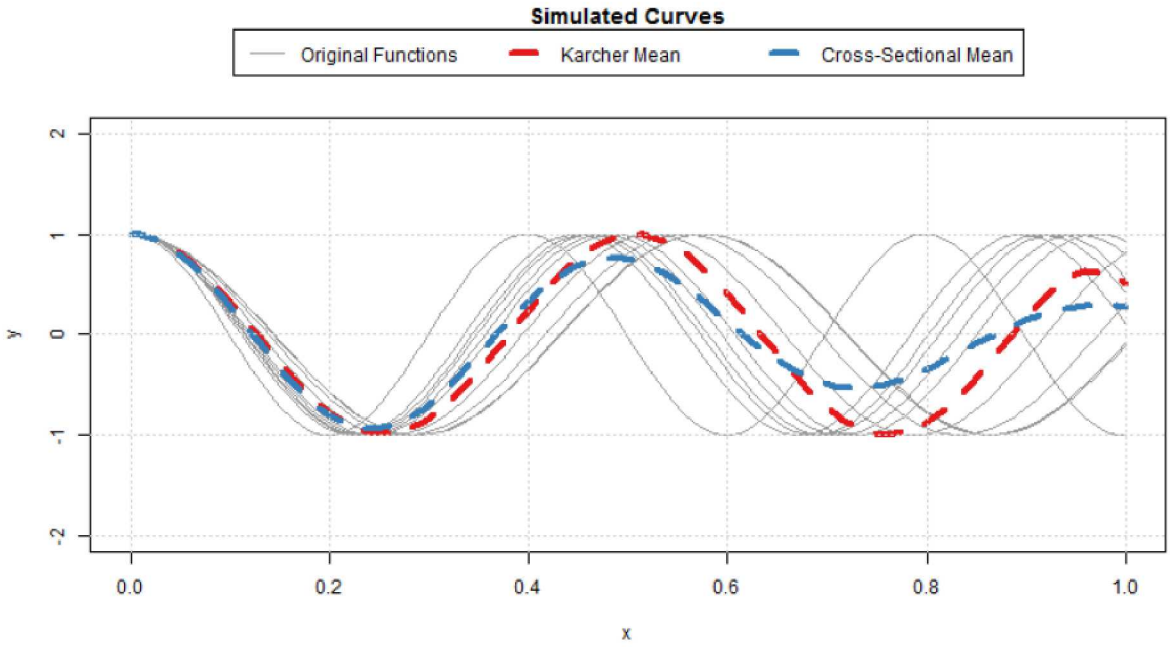



Figure 9.1 Notional example of the Karcher mean used in some functional data analysis methods (from [5]).

## 9.2 CUSUNORO

CUSUNORO, for "cumulative sums of normalized reordered output", is a basis for investigating sensitivity analysis of model output to inputs (Chapter 4, section 4.6) and presenting aggregate results graphically, proposed in Plischke [6]. It is an adaptation of Pearson's correlation ratio as a variance-based sensitivity indicator as well as the graphical presentation of sensitivity analysis results with "contribution to the sample mean", or CSM plots proposed in Bolado-Lavin, et al. [7]. The relevance of this strategy for GDSA is in determining whether a reduction in the dimension of the input space, via sensitivity analysis identification of important or unimportant inputs, is possible while still maintaining response distribution accuracy or adequacy of surrogate models based on a reduced input space.

For description of the CUSUNORO approach, a single response variable $\mathsf{y}(\mathbf{x})$ at input vector $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$, with $nX$ components is assumed and $nS$ responses $\{\mathsf{y}_i, i=1,\ldots nS\}$ are available for a set of inputs $\{\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{nX,i}], i=1,\ldots nS\}$. Although a judicially selected set of calculation points $\{\mathbf{x}_i\}$ as described in Chapter 4 is warranted, the description or implementation of the CUSONORO approach does not depend on this. Further, it is assumed the responses $\{\mathsf{y}_i\}$ are raw calculations, not result of any surrogate or meta-model, although further development might consider some alternative interim assumptions of model forms for exploring various mappings $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, including use of more than a single input at a time.

It is also useful to assume that the scale of individual inputs, say the $j$th input $\{x_{ji}, i=1,\ldots nS\}$, are the same and in the interval [0,1]. If this is not the case, then normalized inputs can be achieved

in various ways, for example by associating values of the input with ranks scaled by *nS*, creating equally spaced indices associated with the *j*th input. Plischke [6] considered additional options: scaled Chebychev points and an optimized partition layout. For purposes here, assume the inputs are in the unit intervals, so consequently all plots of response against input may overlay on [0,1] for visual comparison.

Ordering of the *j*th input imposes an associated ordering on the responses, $\{y_i, i=1,...nS\}$ and it is this ordering for a particular input, or input combination, that is relevant in the 'reordered output' concept of CUSUNORO and the basis for cumulative sums of the responses. Specifically, if $p(i;j)$ is the index of the *i*th ordered value of the *j*th input $x_{j(i)}$ of $\{x_{ji}, i=1,...nS\}$ then $y_{p(i;j)}$, $i=1,...nS$ is a reordering of the response that corresponds to the rank ordering $x_{j(i)}$, $i=1,...nS$. Normalization of responses is achieved by centering the responses and dividing by the square root of the sum of squared deviances of the responses from their mean: $ssy = \sum_{i=1}^{nS}(y_i - y.)^2$ . The scaled cumulative sum of responses associated with $x_{j(i)}$ is the *i*-left cumulative sum of normalized ordered responses, further divided by the square root of the sample size *nS* (as an adjustment for sample size):

$$z(i;j) = \frac{1}{sqrt(nS*ssy)} \sum_{k=1}^{i} \left(y_{p(k;j)} - y.\right) \tag{9.2.1}$$

Associating $z(0;j) = 0$ and noting $z(nS;j) = 0$, the plot $z(i;j)$ versus $x_{j(i)}$ is referred to as the CUSUNORO curve and provides a visual comparison of the relative influence of different inputs on an output (response), similar in spirit with the contribution to the mean, CSM concept. A similar development is possible to define contribution to the sample variance (CSV):

$$csv(i;j) = \frac{1}{ssy} \sum_{k=1}^{i} \left(y_{p(k;j)} - y.\right)^2 \tag{9.2.2}$$

Again, $csv(0;j) = 0$ and $csv(nS;j) = 1$ is noted. Plischke [6] develops these concepts further demonstrating a relation between Pearson's correlation ratio and gradients of the CUSUNORO curve and a strategy for adaptively partitioning the input range where piecewise constant values are fit to the response.

An example of a CUSUNORO plot for two sample sizes are shown in Figure 9.2. Both plots use a 5-parameter analytical test function where $x_1$, $x_2$, and $x_3$ are the active variables: $y=5x_1+5x_1x_2+3x_3x_3$. Figure 9.2a was generated based on 16 samples selected as an orthogonal array – based LHS, with a maximin distance criteria and minimum pairwise input correlation criteria implemented, and Figure 9.2b was generated based on 100 simple random samples. It is seen that the CUSUNORO curve for $x_1$ indicates its importance by a fairly steady decrease from zero to a minimum then steadily increasing to zero again. The first plotted value, after the degenerate zero, is the response value at the minimum input value differenced from the mean and is neccesarily negative since that response value is also minimum due to the increasing relationship of the response to input $x_1$. This behavior is as expected of the cumulative sum differenced from the mean of the response and correctly identifies that the response is an increasing function of the first input, $x_1$. In Figure 9.2a with 16 samples, that the response is also an increasing function of $x_2$ and $x_3$ may be harder to differentiate because the initial decline is somewhat less consistent, due

to the impact of the interaction relationship with $x_1$ diminishing the impact of $x_2$ and the smaller coefficient and compounding squared term in $x_3$ diminishing its impact (due also to that the inputs are on the interval between 0 and 1). However, CUSONORO curves for both $x_2$ and $x_3$ behave similiarly to that for $x_1$ with neither crossing the zero center-line. For Figure 9.2a with 16 samples, the remaining $x_4$ and $x_5$ CUSONORO curves, although they deviate only somewhat less from the zero center-line seemingly than the curves for $x_2$ and $x_3$, both cross the zero center-line. In Figure 9.2b with 100 samples, the ranking of the most important variables is clearer. Further investigation is underway on the behavior and interpretation of CUSUNORO curves as a function of number of samples and number of input parameters. It is worth noting that co-linearity in the sample of inputs can significantly confuse interpretation of these plots, particularly for small sample sizes, although with increasing sample size this problem is diminished.

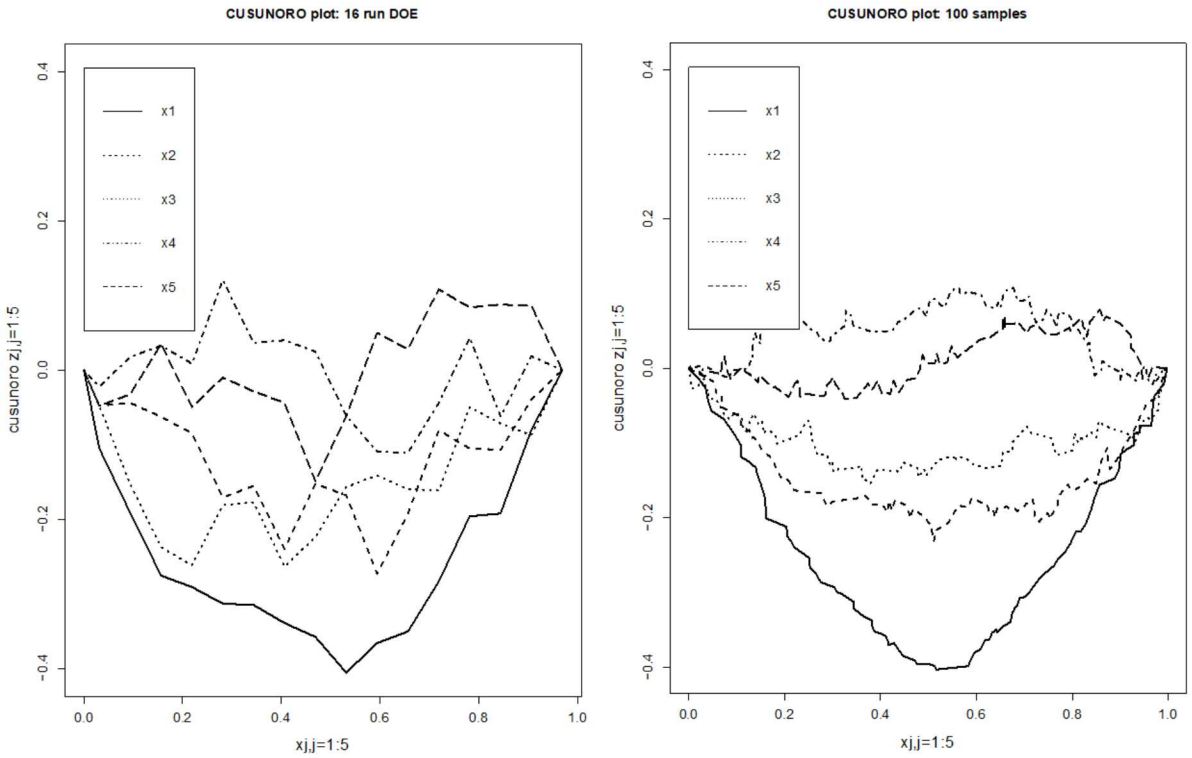



Figure 9.2 CUSUNORO plot examples for a test function with 5 parameters. 9.2a (left): results based on 16 samples. 9.2b (right): results based on 100 samples.

Figure 9.3 shows two CUSUNORO plots for the shale reference case described in Chapter 7. Figure 9.3a shows the cumulative sum of normalized output for the location "sand_obs1" depicted in Figure 7.2 and Figure 9.3b shows the cumulative sum of normalized output for the location "sand_obs3" in Figure 7.2. These plots show that the input parameters pShale and kSand are the most important parameters for these two locations as evidenced by their deviation from the nominal zero line and grey ellipse area. This is consistent with the earlier sensitivity analysis that was performed using the Sobol' variance-based indices presented in Figure 7.4.

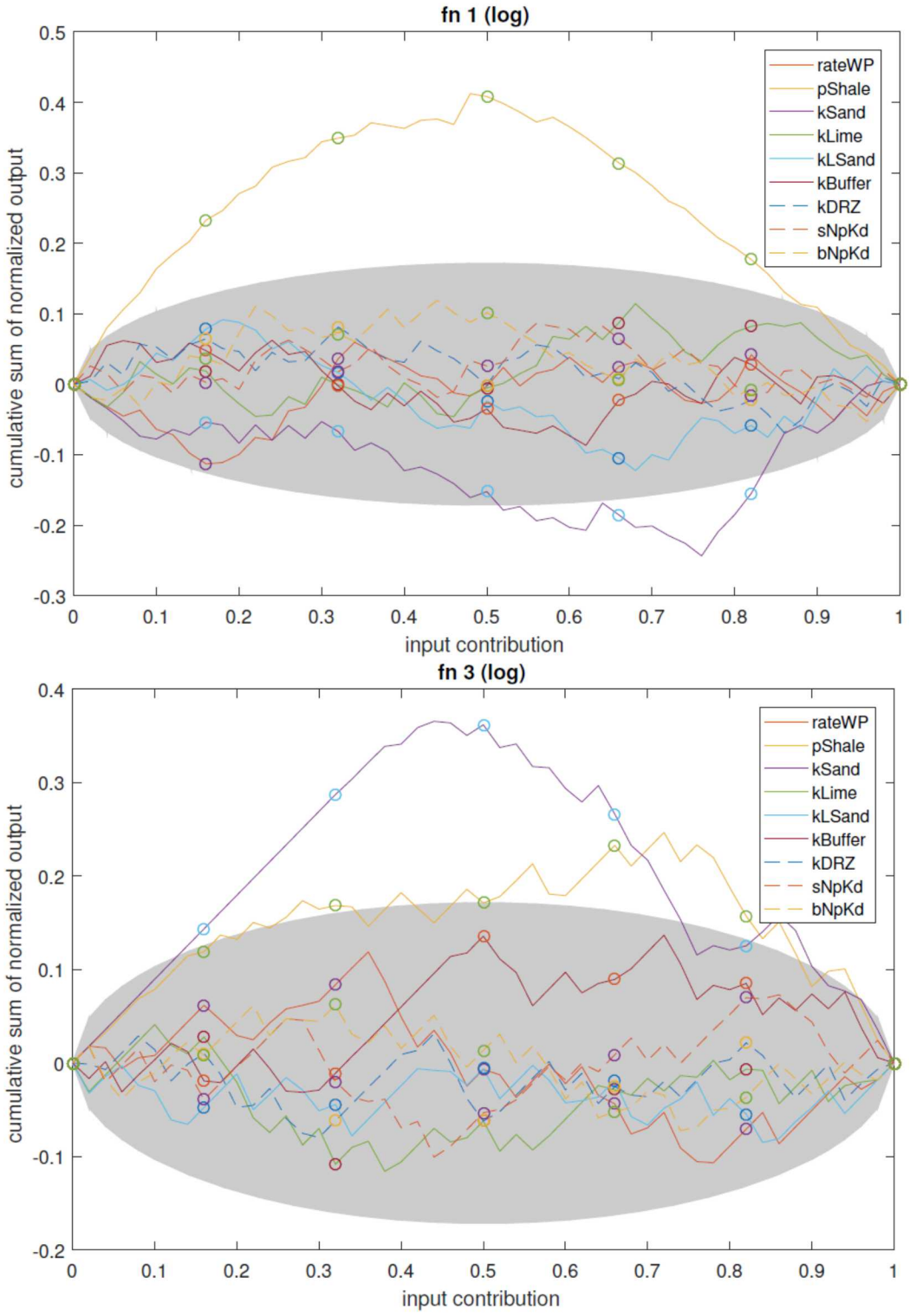



Figure 9.3 CUSUNORO curves for two locations in the shale reference case: (a) top plot shows results at sand_obs1 location and (b) bottom plot shows results at sand_obs3 location.

## 9.3 Importance Sampling

Importance sampling is a biased sampling method used to sample random variables from different densities than originally defined [8;9]. These importance sampling densities are constructed to pick "important" values of input random variables to improve the estimation of a statistical response of interest, such as a mean or probability of failure of a response from a simulation model. The use of biased input sampling densities will result in biased estimators if they are applied directly to the simulation results. However, the simulation results are weighted to correct for the use of the importance sampling densities, and this ensures that the importance sampling estimators are unbiased. Conceptually, importance sampling is very attractive: for example, when one wants to generate more samples in a failure region when estimating failure probabilities. In practice, however, importance sampling can be challenging to implement efficiently, especially in a general framework that will allow solutions for many classes of problems.

Importance sampling is often explained through integration. Accurate computation of high-dimensional integrals is common to many engineering and scientific applications. Monte Carlo methods have been commonly used for many years to approximation the expectation of functions of random variables by using the sample mean. That is, when calculating the expectation:

$$\mathrm{E}(r(X)) = \int r(x) f_X(x) dx \tag{9.3.1}$$

where $r$(X) is a response function: $r$: $\Re^{\mathrm{d}} \rightarrow \Re^{1}$ and $X$ is a multidimensional random variable having probability density function $f_X$, the Monte Carlo estimator of E($r$(X)) is:

$$\hat{\mathrm{E}}_{\mathrm{n}}(r(X)) = \frac{1}{n}\sum_{i=1}^{n} r(x_i) \tag{9.3.2}$$

Note that many quantities of interest can be cast as expectations, including probabilities of failure. For calculation of failure probabilities, the probability that $r$(X) takes on a value in set A of interest (defined as the failure region) can be written as:

$$\mathrm{Prob}(r(X) \in \mathrm{A}) = \mathrm{E}[\mathrm{I}_{\{\mathrm{A}\}}(r(X))] \tag{9.3.3}$$

where $\mathrm{I}_{\{\mathrm{A}\}} = 1$ when $r$(X)$\in$A and 0 when $r$(X)$\in$A.

The purpose of importance sampling is to sample the random variables from a different distribution than the original distribution of interest and use those samples to calculate an estimate $\hat{\mathrm{E}}_{\mathrm{n}}(r(X))$, with the goal of reducing the variance in the estimate. To do this, the Monte Carlo estimate must be weighted appropriately. If $h_X(x)$ is the new distribution from which the random variables will be sampled (note $h$ must have the same support as $f$), the new estimate is derived as follows:

$$\mathbf{E}(r(X)) = \int r(x)\frac{f(x)}{h(x)} h(x) dx = \hat{\mathbf{E}}_h\left[ r(X)\frac{f(X)}{h(X)} \right] = \frac{1}{n}\sum_{i=1}^{n} \frac{r(x_i) f(x_i)}{h(x_i)}, x_i \sim h(X) \tag{9.3.4}$$

where *h(x)* is called the importance sampling density, the product of the integrand and the original density is called the importance function, and *f(x)/h(x)* is called the weight function. There are certain choices that one should make when picking *h(x)*. The variance of the importance sampling estimator $\hat{E}_h$ is minimized when $h(x) \propto |r(x)f(x)|$.

Much of the work in importance sampling has been finding minimum variance estimators for specific cases. The fundamental issue in implementing importance sampling is the choice of the biased distribution which encourages the important regions of the input variables. If you have a good distribution, the payoff is a much lower simulation cost, but if you choose a poor importance sampling density, the run times could actually be longer than standard Monte Carlo (i.e., simple random sampling) simulation. There are various measures used to calculate the goodness of the importance sampling scheme. One is the ratio of the variance obtained by a pure Monte Carlo vs. the variance of the importance sampling result: $\sigma^2_{MC} / \sigma^2_{IS}$ ; another measure is the ratio of the number of samples required by each scheme, given the same output variance (this is called simulation gain): $N_{MC}/N_{IS}$. In general, a good importance sampling function should be as follows:

1. $h(x) > 0$ whenever $r(x)f(x) \neq 0$.
2. *h(x)* should be close to being proportional to $|r(x)f(x)|$.
3. It should be relatively easy to generate samples from *h(x)* and also to calculate the density *h(x)* for particular values of x.

Some standard approaches to determining *h(x)* including scaling, where the original random variable X is scaled by $\alpha$ or linearly shifted (e.g., $h(x) = f(x-c)$) to put more probability density into a particular region (for failure probability estimation, for example). Another classical approach is to assume that *h(x)* belongs to a parametric distribution family. Then, the problem is determining the values of the parameters governing that distribution (for example, determining the mean and variance for *h(x)* if *h* is assumed to be normal). Often these parameters are obtained by optimizing the variance of the importance sampling estimator $\hat{E}_h$, but this implies that one can calculate an analytic expression or approximation for this estimator. Oh and Berger (1993)[10] used an approach where they used a mixture distribution (a set of weighted individual distributions), where the individual distributions were multivariate-t distributions. They then had to determine the weights, location, and scale parameters of the *t*-distributions, which they did by numerical minimization of the estimate of the squared variation coefficient of the weight function.

For black-box simulations that have multiple uncertain inputs which may come from a wide variety of random input distribution types, one cannot generally assume that the importance sampling density will be normal or have a parametric form. The main problem is how to pick points and/or distributions around which to sample. Swiler and West investigated an adaptive process which generated LHS samples and then used Kernel Density Estimation (KDE) to generate an importance sampling kernel density around the points in a failure region [11]. This can be helpful but is very dependent on the number and location of points that are in the failure region from the initial sampling study.

A few importance sampling approaches are currently implemented in Dakota. They work in conjunction with reliability methods, which are methods tailored at answering the question: what

is the probability that my response will be greater than threshold T? Typically, a response greater than T indicates a failure and often these are extremely rare events, so the probability of failure could be only the order of 1.E-6 or less. In these situations, MC sampling to find the failure probability is prohibitive. Optimization is used in reliability methods to find the "most probable point" defining the boundary or interface between safe and failure regions. The theory of reliability methods [12] relies on transforming the problem to standard normal space, so the inputs are Gaussian. Because of this, after the most probable point is found, it is used as the mean of a Gaussian for the importance sampling process.

To demonstrate importance sampling on a problem of interest to GDSA, 200 samples were generated from a repository simulation where one of the radionuclides of interest is $^{129}$I concentration. One goal is to understand what will lead to high iodine concentrations. Figure 9.4 shows the $^{129}$I concentration at location sand_obs1 from the shale simulations as a function of the input parameters. The simulations with $^{129}$I concentrations greater than 1E-13 are highlighted. Note that of the 200 PFLOTRAN simulations, only 21 had high concentrations. The goal is not only in assessing the failure probability more accurately (if failure is defined as $^{129}$I concentration > 1E-13[M]) but also in understanding what input parameter combinations will lead to such cases. Importance sampling is ideal for this.

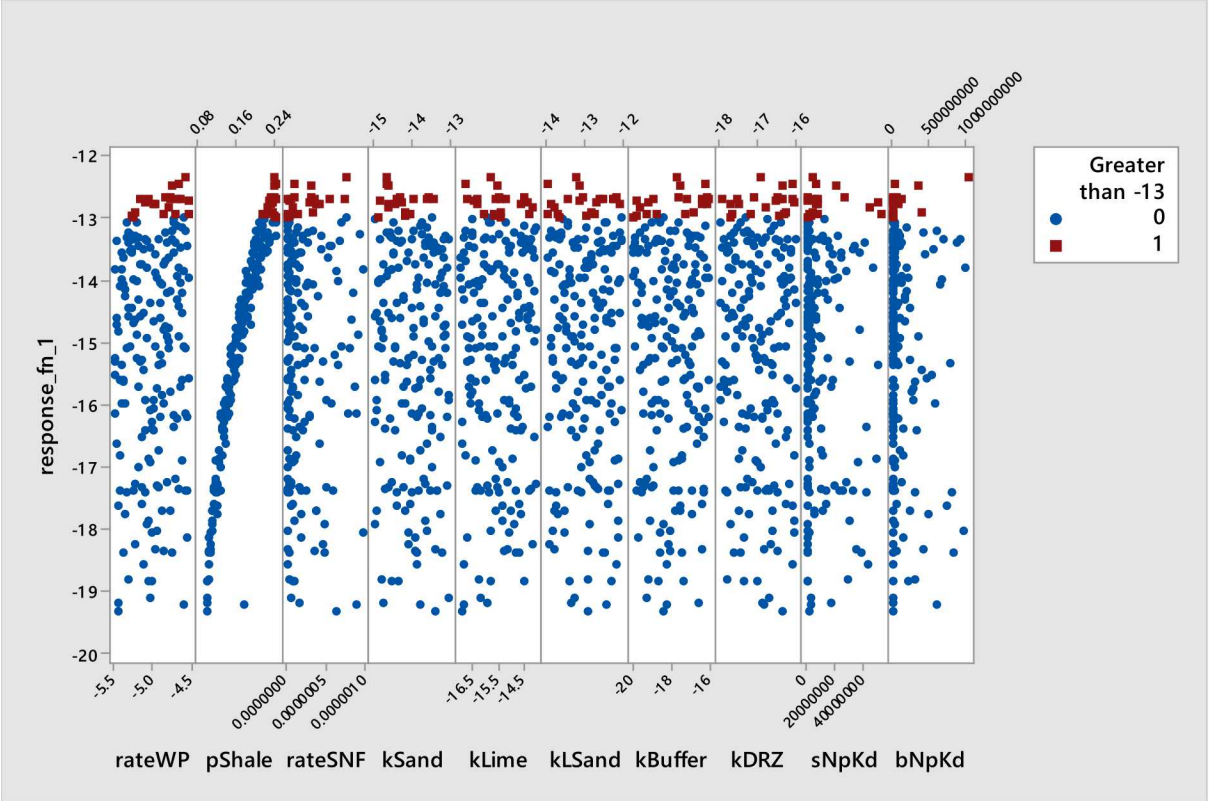



Figure 9.4 Monte Carlo sampling (200 samples) of PFLOTRAN results showing $^{129}$I concentration at location sand_obs1 from the shale simulations as a function of the input parameters. The samples with concentrations greater than 1.E-13[M] are highlighted in red.

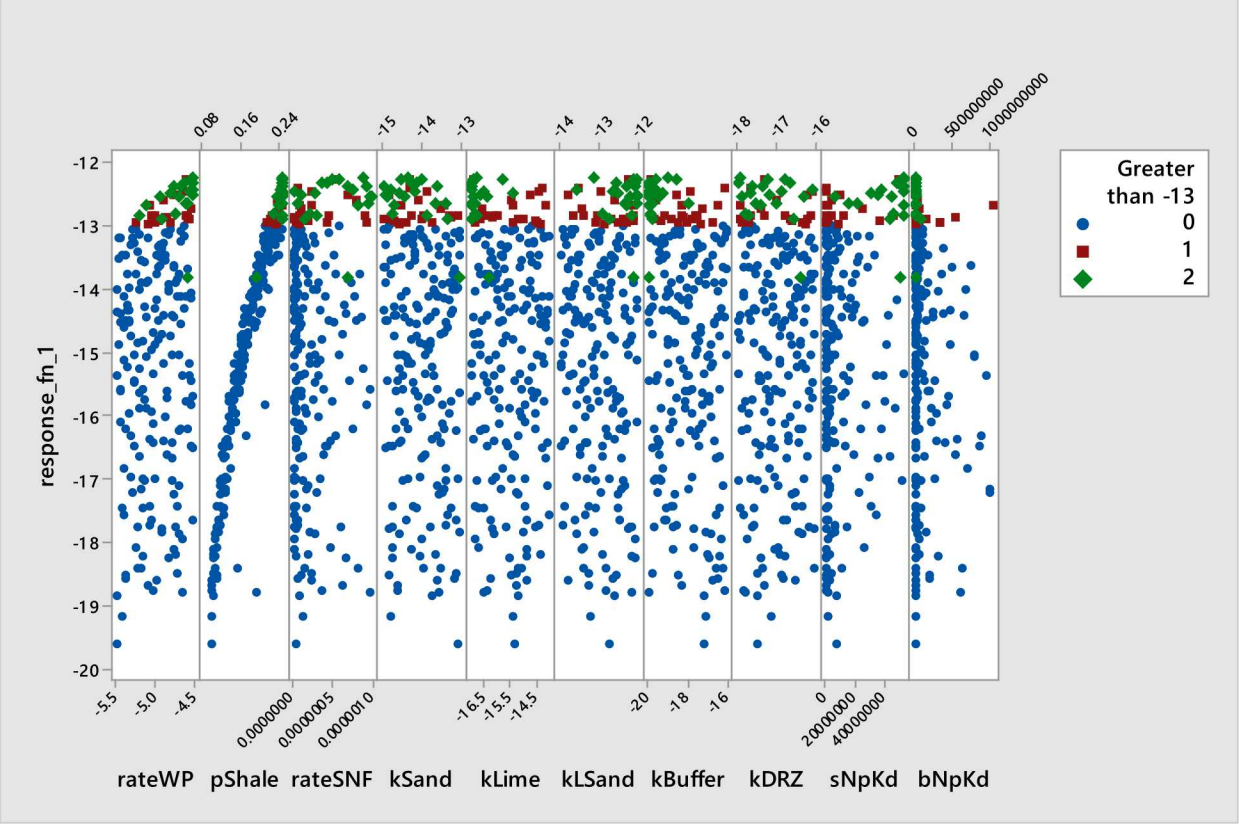



Figure 9.5 Importance sampling based on the original 200 LHS samples of Figure 9.4. The additional importance samples are shown in green.

Figure 9.5 shows the importance sampling points highlighted in green. There were 200 initial points and 20 points added with importance sampling. The blue and the red points are surrogates of the original points, where red points are in the high concentration region. Note that we did employ surrogates for the importance sampling calculations. This study demonstrated that importance sampling was able to determine additional, new sample points in a 10-dimensional input space which correspond to the region of interest (in this case, $^{129}$I concentrations greater than 1.E-13[M]). The green samples are much more strategically placed and focused on the region of interest than 20 new random samples taken over the input domain. Given the large number of sample inputs and the relatively small number of original samples, this study shows the potential to perform more targeted sampling using importance sampling for GDSA applications.

## 9.4 Bayesian Calibration Methods

Calibration is the process of determining parameters of a model given observational data. A common approach is to minimize the "error sum of squares" function. Assume that a model of the response $y$ as a function of the n-dimensional inputs $\mathbf{x}$ is given as:

$$y_i = f(\mathbf{x}_i;\boldsymbol{\theta}) + \varepsilon_i \tag{9.4.1}$$

where $f$ is the model (e.g., a computational simulation), $\boldsymbol{\theta}$ is a vector of parameters, $\varepsilon$ is the random error term. Assume that $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$ and $\varepsilon_i$ are identical and independently distributed (iid). Given observations of the response y, the goal of calibration is to find the optimal values $\boldsymbol{\theta}^*$ to minimize the error sum of squares function $S(\boldsymbol{\theta})$, also referred to as SSE:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})]^2 = \sum_{i=1}^{n} [R_i(\boldsymbol{\theta})]^2 \tag{9.4.2}$$

where $R_i(\boldsymbol{\theta})$ are the residual terms. Nonlinear least squares optimization algorithms have been designed to exploit the structure of such a sum of squares objective function. By assuming that the residuals $R_i(\boldsymbol{\theta})$ are close to zero near the solution, the Hessian matrix of second derivatives of $S(\boldsymbol{\theta})$ can be approximated using only first derivatives of $R_i(\boldsymbol{\theta})$. The Gauss Newton algorithm is often used [13].

The process above is often called a "deterministic" calibration: a point estimate or best estimate of the parameters is determined ($\boldsymbol{\theta}^*$). However, uncertainty should be incorporated in the parameters, including uncertainty in the test measurements, in the weighting of the objectives, and in the model form. One approach is to start the least squares optimizer at different locations in parameter space to see if the optimization always leads to the same answer or if there are many local optima (e.g., many combinations of parameters which indicate that the calibration problem is non-unique and/or poorly posed). The results from a multi-start optimization are helpful to identify if there are many local minima but they don't provide a distribution on the input parameters.

Bayesian calibration is often used to generate posterior parameter distributions. Such distributions are based on the prior parameter distributions (which must be specified based on expert judgement and/or historical data) and a likelihood function (which represents a goodness of fit and often involves a term such as the SSE). Bayesian calibration is expensive computationally because one generates the posterior distribution by a sampling process called Markov Chain Monte Carlo which generates possible samples in a sequential fashion, deciding at each point whether to reject or accept the sample and where to go next in parameter space. Some advanced MCMC algorithms are designed to accept every sample, which is more efficient. MCMC typically requires hundreds of thousands of function evaluations to obtain converged estimates of the posteriors and thus often involves the use of surrogate models. There have been many approaches to circumvent this, including the use of parallel chains, the use of adaptive methods, etc. This section does not go into detail, but the interested reader should review a few of the canonical references in the field [14-18] as well as some recent examples of the application of Bayesian methods to geophysical modeling [19-20].

In addition to generating posterior distributions of input parameters (and also posterior distributions of response quantities obtained by evaluating the model at the posterior parameter values), Bayesian calibration is attractive because it facilitates model selection (identification of

the best model if there are a few candidate models) and optimal experimental design. Model selection is a non-trivial computational task: it involves the calculation of the "model evidence" which involves averaging the model likelihood over all possible values of the parameters for each model. The likelihood is near zero in many portions of the parameter space which makes the accurate calculate of the integral challenging. There are various computational tricks to calculate model evidence [21;22].

In a Bayesian framework for optimal experimental design (OED), one identifies a set of candidate experiments that could be run, then selects the design that maximizes the expected value of the experiment [23;24]. To calculate this expected value, one needs to hypothesize the set of outcomes and their associated probabilities that one would get from performing each experiment. Then, these hypothesized outcomes are examined to see how they would change the posterior distributions of the parameter calibrations. Depending on the utility function, if an experiment would change or inform the posterior significantly, it is a likely candidate to be selected. Bayesian OED is extremely expensive computationally because it involves layers of model runs and integrals based on the results. To hypothesize model outcomes, one can use expert judgement but typically the model itself is used.

In the context of GDSA applications, one could use optimal experimental design to help identify which physical experiments to run to most inform one or more of the parameters. For example, if there are four possible locations at which to drill a well and only one well can be drilled, OED will identify which of the four is most useful to improve a parameter such as a permeability value. Bayesian calibration can help GDSA applications by incorporating experimental data such rock or material specimen testing into the uncertainty estimation of the model parameters.

## 9.5 Multifidelity UQ Methods

The scientific computing community has endeavored to develop methods which are as efficient as possible to perform UQ on computationally expensive simulation models. One set of approaches is called "multilevel" and "multifidelity" UQ methods. These multifidelity UQ techniques have gained popularity in the last decade or so when the need for UQ of high-fidelity numerical simulations led to the design of techniques capable of containing the overall computational burden. The main idea behind these techniques is to extract information from a limited number of high-fidelity model realizations and complement them with a much larger number of a set of lower fidelity evaluations. The final result is an estimator with a much lower variance, i.e. a more accurate and reliable estimator can be obtained. These approaches typically rely on a strong correlation between low and high-fidelity results and the use of control variates. The mathematics behind these approaches can be found in [25-30].

Multifidelity approaches can be used in many ways in the *GDSA Framework.* For example, one could have the same PLOTRAN model but have two different mesh resolutions for it: this would be referred to as a "multilevel" study. Or, one could have two different versions of a chemistry model such as the Fuel Matrix Degradation model, one of higher fidelity and one of lower fidelity. Another possibility is to have two codes such as PFLOTRAN and TOUGH2 serving as the high and low fidelity models, respectively.

## 9.6 References: Chapter 9

1. J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer, 2005.

2. D. Higdon, J. Gattiker, B.Williams, M.Rightley. "Computer model calibration using high-dimensional output." *Journal of the American Statistical Association* 103 (482), 570-583. 2007.

3. J.D. Tucker, J.R. Lewis, A. Srivastava. "Elastic functional principal component regression." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12 (2). Pages 101-115. 2019

4. J.D. Tucker. "Elastic functional data analysis." SAND2018-1615PE.

5. J.R. Lewis and D.M. Brooks. "Uncertainty Quantification and Comparison of Weld Residual Stress Measurements and Predictions." SAND2016-10932.

6. E. Plischke, "An adaptive correlation ratio method using the cumulative sum of the reordered output", *Reliability Engineering & System Safety*, Volume 107, 2012, p. 149-156,

7. R. Bolado-Lavin, W. Castaings, S. Tarantola, "Contribution to the sample mean plot for graphical and numerical sensitivity analysis" *Reliability Engineering & System Safety*, Volume 94, Issue 6 (2009), p. 1041-1049,

8. R. Srinivasan. Importance Sampling. Springer-Verlag, 2002.

9. S.T. Tokdar and R.E. Kass. "Importance sampling: a review." Wiley Interdisciplinary Reviews: *Computational Statistics.* 2010 Jan;2(1):54-60.

10. M.S. Oh and J.O. Berger. "Integration of multimodal functions by Monte Carlo importance sampling." *Journal of the American Statistical Association.* 1993 Jun 1;88(422):450-6.

11. L. P. Swiler and N.J. West. "Importance Sampling: Promises and Limitations." AIAA-2010-2850 in *Proceedings of the 51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference* (12th AIAA Non-Deterministic Approaches Conference), Orlando, FL, Apr 12-15, 2010.

12. A. Haldar and S. Mahadevan. *Probability, Reliability, and Statistical Methods in Engineering Design.* Wiley and Sons, 2000.

13. C.J. Wild, C. J. and G.A.F. Seber, (1989). *Nonlinear regression.* New York: Wiley.

14. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall/ CRC, 2004.

15. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

16. J.K. Kruschke. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Elsevier, 2015.

17. M. C. Kennedy and A. O'Hagan. "Bayesian calibration of computer models (with discussion)". *Journal of the Royal Statistical Society B*, 63:425–464, 2001.

18. H. Haario, M. Laine, A. Mira, and E. Saksman. "DRAM-Efficient adaptive MCMC". *Statistics and Computing*, 16(4):339–354, 2006.

19. J. Ray, Z. Hou, M. Huang, K. Sargsyan, and L. Swiler. "Bayesian calibration of the community landmodel using surrogates". *SIAM/ASA Journal on Uncertainty Quantification*, 3:199–233, 2015.

20. M. Huang, J.Ray, Z. Hou, H. Ren, Y. Liu, and L. Swiler. "On the applicability of surrogate-based Markov Chain Monte Carlo-bayesian inversion to the Community Land Model: Case studies at flux tower sites." *Journal of Geophysical Research: Atmospheres*, 121(13):7548–7563,2016.

21. L. Wasserman, "Bayesian Model Selection and Model Averaging," Journal of Mathematical Psychology, Volume 44, Issue 1, 2000, p. 92-107.

22. T. Portone, J. Niederhaus, J. Sanchez, and L. Swiler. "Bayesian model selection for metal yield models in high-velocity impact." *Proceedings of the 2019 Hypervelocity Impact Symposium,* HVIS2019-12, April 2019, Destin, FL, USA. SAND2019-4041C.

23. K. Chaloner and I. Verdinelli. "Bayesian Experimental Design: A Review." *Statistical Sciences* 10 (1995), no. 3, p. 273-304.

24. E.G. Ryan, C.C. Drovandi, J.M. McGree, and A.N. Pettitt. "A review of modern computational algorithms for Bayesian optimal design." *International Statistical Review*. 2016 Apr;84(1):128-54.

25. G. Geraci, M. S. Eldred, and G. Iaccarino. A multifidelity multilevel monte carlo method for uncertainty propagation in aerospace applications. In *19th AIAA Non-Deterministic Approaches Conference,* 2017.

26. M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

27. M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

28. A. A. Gorodetsky, G. Geraci, M. Eldred, and J. D. Jakeman. A generalized framework for approximate control variates. *arXiv preprint* arXiv:1811.04988v2, 2018.

29. R. Pasupathy, B. W. Schmeiser, M. R. Taaffe, and J. Wang. Control-variate estimation using estimated control means. *IIE Transactions*, 44(5):381–385, 2012.

30. B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016.

# 10. INTERNATIONAL ACTIVITIES AND COLLABORATION

## 10.1 International Working Group Activities 2018-2019

As early as 2012, an international collaboration has been ongoing in the areas of uncertainty quantification and sensitivity analysis, beginning with the formation of an informal Technical/Scientific Working Group in 2013 (see Becker and Ulrich 2014 in Hansen et al.[12], p. 195). Originally this began through multi-nation interactions under the auspices of the Implementing Geological Disposal of Radioactive Waste –Technology Platform (IGD-TP) of the European Union EU. This was followed by a series of annual meetings starting with a "Workshop on Handling Uncertainties" in September 2015 at Harwell, UK [13], followed by an "International Workshop on Sensitivity Analysis of Final Repository Systems" in Braunschweig, Germany in October 2016, and then by a "Quantification of Uncertainty Workshop" in August 2017 in Albuquerque, NM, USA. It was at this latter workshop in the U.S. where more collaborative work was initiated among the international participants and organizations, as discussed below.

A follow-on workshop in Brussels, Belgium in October 2018 [2;11] led to a more formal establishment of this collaboration during the subsequent Integration Group for the Safety Case (IGSC) Symposium on the Safety Case in Rotterdam, The Netherlands in October 2018. The current uncertainty task group, now under the umbrella of the IGSC of OECD/NEA, consists of waste management, consulting, and research organizations from several countries [4]. The task group involves at least 10 organizations from 8 countries. A subset of these organizations—GRS (Germany), Posiva (Finland), Radioactive Waste Management (RWM) (UK), and possibly SKB (Sweden)—have taken interest in uncertainty quantification whereas another group of organizations—GRS (Germany), Clausthal University of Technology (TUC) (Germany), Sandia National Laboratories (SNL) (USA), Posiva (Finland), and SCK-CEN (Belgium)—have been pursuing the subtopic of sensitivity analysis. Interest was also expressed at the IGSC Safety Case Symposium 2018 by organizations in France (IRSN), Switzerland (Nagra, ENSI), and Russia (IBRAE), who have been invited to the next annual meeting of the task group to be held at the GRS offices in Berlin in November 2019.

An Uncertainty Quantification workshop held August 2017 in Albuquerque, USA, led to a joint sensitivity analysis exercise effort initiated in October 2017, with participation from GRS, Posiva, SCK-CEN, SNL, and TUC. The Sensitivity Analysis subgroup identified six test cases ranked in order of complexity and since then has been comparing analysis methods on progressively more complex models. The group shared results at the next working group meeting—International SA-UQ Workshop—held in Brussels, Belgium, in 2018. The activities to date have resulted in three conference and symposium papers: Röhlig et al. [4], Becker [10] and Stein et al. [11] summarizing the results.

The six test cases that the Sensitivity Analysis subgroup has developed include five performance assessment (PA) models (Problems 1-5) and one hydrogeochemical process model describing spent nuclear fuel dissolution (Problem 6) to be used by all members of the subgroup. The test cases are a series of models with increasing level of complexity, but they are treated as black box models in the sense that the subgroup members were only provided with the input parameters in the form of sampled values and corresponding outputs from probabilistic model runs. Thus, for the purpose of these analyses, each model represents some unknown mapping from

the parameter space to the output variable space. The participants exchanged only rough descriptions of the systems and the parameters. In the absence of the detailed process- and implementation-level understanding of the systems by the sensitivity analysts, the goal was to determine how much information about the model can be derived unbiasedly from sensitivity analysis alone [10].

Table 10.1 and Table 10.2 show the six test problems and the global sensitivity analysis methods to be applied by various organizations, respectively.

Table 10.1 Performance assessment- and process model-based test cases.

| Problem No. | Case | Key Features |
|---|---|---|
| 1 | GRS (Germany):<br>High-level Waste (HLW)/ Spent Nuclear Fuel (SNF) repository in clay host rock | • Number of realizations (n) = 4,096 (Monte Carlo sampling)<br>• Time-dependent<br>• 6 independent inputs<br>• Smooth model behavior |
| 2 | Sandia National Laboratories (SNL) (USA):<br>Generic repository for SNF in shale (see Chapter 7 of this report) | • Case 1: n = 50 (LHS Sampling)<br>• Case 2: n = 200 (LHS Sampling) (to test convergence of sensitivity analysis results as the sample size increases)<br>• Time dependent, but sensitivity analysis performed on maximum radionuclide concentration (regardless of time) at several spatial locations<br>• 10 independent inputs<br>• Smooth model behavior |
| 3 | SCK-CEN (Belgium):<br>Dessel Surface Low and Intermediate Level Waste (LILW) repository [5] | • n = 1,024<br>• Quasi Monte Carlo sampling<br>• Time dependent<br>• 22 independent input parameters, some will change at a given point in time; material-dependent<br>• Non-monotonic behavior, regime change with time (i.e., standard sensitivity analysis methods do not sufficiently explain model behavior) |
| 4 | GRS (Germany):<br>Generic LILW repository in a salt mine | • Time-dependent<br>• 20 independent inputs<br>• highly non-linear behavior |
| 5 | SNL (USA):<br>Generic repository for SNF in fractured crystalline rock (see Chapter 8 of this report) | • Time-dependent<br>• 8 independent inputs plus the stochastically generated permeability field<br>• non-monotonic behavior |
| 6 | SNL (USA):<br>Reactive transport model describing spent nuclear fuel dissolution (see Mariner et al. 2019, Appendix A) [14] | •Time-dependent<br>•13 inputs, some dependent, two are a pre-determined function of time or some are outputs from the previous time step<br>•Non-linear behavior expected |

Table 10.2 Sensitivity analysis methods applied to test problems by various participants.

| Organization | Sensitivity Analysis Methods of Interest | Method Applied To |
|---|---|---|
| SNL (USA) | Correlation, rank correlation (simple, partial)<br>Meta-modeling via Gaussian process surrogate, polynomial chaos expansion (main, total effects, and higher order effects) | • GRS's HLW/SNF repository in Clay host rock problem [11]<br>• SCK-CEN's Dessel Surface LILW repository problem [11] |
| Posiva (Finland) | Correlation, rank correlation (simple, partial)<br>Regression (simple, partial)<br>First order sensitivity indices with Effective Algorithm for Computing Global Sensitivity (EASI) | • SCK-CEN's Dessel Surface LILW repository problem [8] |
| SCK.CEN (Belgium) | Regression based measures<br>First/Second order sensitivity indices based on its own LpTau sampler (n=24,000) | • SCK-CEN's Dessel Surface LILW repository problem [7] |
| GRS (Germany) | Linear, rank, and variance-based measures | • SCK-CEN's Dessel Surface LILW repository problem [1]<br>• GRS's generic LILW repository in a salt mine [6] |
| TUC (Germany) | Cumulative sums of normalized re-ordered output (CUSUNORO) curves | • SCK-CEN's Dessel Surface LILW repository problem [9] |

The Sensitivity Analysis subgroup members have been independently applying their own global sensitivity analysis methods (Table 10.2) to the Table 10.1 test cases. Two categories of sensitivity analysis methods were being applied: (i) linear and rank regression / correlation (Chapter 4 of this report) and (ii) variance-based methods (Chapter 5 of this report). In general, the applications of sensitivity analysis methods focused on number of parameters, sample size(s), model linearity, model monotonicity, continuity, or lack thereof, and parameter interactions.

Various subgroup members have applied their sensitivity analysis methods to Problem 1 and 2 and all subgroup members to Problem 3, providing a rich collection of sensitivity analysis outputs for cross comparison in the future [1-8]. Although it was considered too early to evaluate the exercise as a whole, several inferences were drawn from the analysis of some of the problem cases presented in Table 10.1. Problem 2 is described in detail in Chapter 7 of this report; problem 5, which has not yet been addressed by the group, is described in Chapter 8 of this report; and Problem 6 is described in [14].

Even with a relatively low number of runs (i.e., 50), the main sensitivities could be identified by standard methods when combined with appropriate parameter transformations. The small standard deviations achieved in Chapter 7 for sensitivity indices calculated using log-transformed model output values (concentrations) suggest that for the case of generic repository in shale (Problem 2 of Table 10.1) a sample size of 50 is sufficiently large to reliably estimate sensitivity indices. However, large standard deviations and lack of a trend with increasing sample size were

observed even with a sample size of 200 when raw model output values (i.e., concentrations) were used.

SNL's calculations using the GRS dataset (Problem 1 of Table 10.1) showed successful application of DAKOTA to compute higher-order sensitivity indices and interactions based on polynomial chaos expansion (PCE). Figure 10.1 shows sensitivity indices for first order (i.e., main) effect, second and third order interaction effects, and the total order effect using a PCE of order 5. The figure shows the sensitivities of all six sampled paramaters, and sensitivities change as a function of time. Successful computation of the higher order sensitivity indices is significant because challenges were encountered when computing higher order sensitivity indices for geologic disposal problems in the Performance Assessment Methodologies in Application (PAMINA) exercise (2006-2009).

All participants applied various sensitivity analysis methods to compute main (first order) sensitivity index for the Dessel Surface Low and Intermediate Level Waste (LILW) repository problem. Figure 10.2 compares the main (first order) effects calculated with different methods by different participants (SNL, GRS, and TUC). The results from these methods showed similar qualitative behavior, but considerable differences in details. The group determined that some of these differences could potentially be resolved by increasing the sample size [4].

Spiessl et al. [6] applied the Random-Sampling High Dimensional Model Representation (RS-HDMR) meta-modeling approach in the *SobolGSA* software to the Generic LILW Repository in a Salt Mine (Problem 4 of Table 10.1). They concluded that good estimates of first, second and total order sensitivity indices can be obtained with a reasonable choice of polynomial orders for the RS-HDMR method and number of simulations to build the meta-model. By using higher-order terms including more polynomial coefficients to build an accurate meta-model, they were able to obtain better estimates of the sensitivity indices. For robust estimation of the indices and also for better system understanding, they recommended checking the sum of all indices and obtaining regression coefficients ($R^2$). Inclusion of more insignificant parameters resulted in noisier results. Hence, they recommended excluding insignificant parameters to obtain reliable results, especially the second order and total order sensitivity indices. Comparison of the first order sensitivity from RS-HDMR with EASI and SDP methods showed that rough estimates can be obtained with all three methods at the sample sizes of more than 1,000. RS-HDMR and EASI required similar computation time whereas SDP required at least 94% more computation time.

The planned future activities of the Sensitivity Analysis subgroup include the following:

- Apply sensitivity analysis method to Problems 6 (Table 10.1)
- Investigate further higher-order effects in sensitivity analysis [11], effects of model output transformations, and density-based sensitivity analysis methods;
- Involve more test cases, especially to investigate the effects of sample-size effects, dependent input parameters, non-smooth model outputs, and discontinuous outputs;
- Develop formalism and/or performance criteria for uniform evaluation of the behavior of the algorithms used and results obtained by the participants;
- Develop sensitivity analysis guidelines.

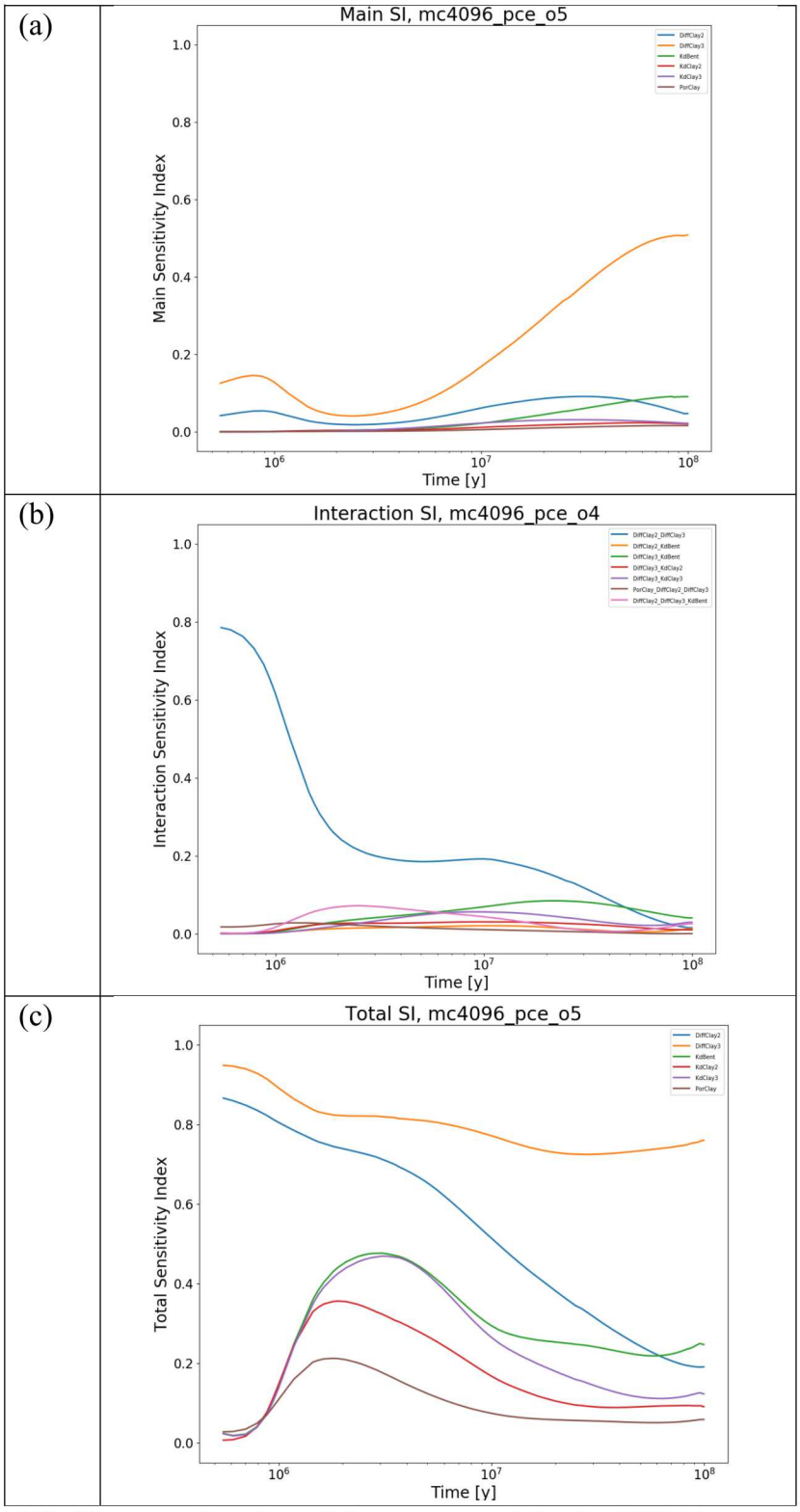



Figure 10.1. Sobol' indices as a function of time computed for the GRS Dataset (Problem 1) using the PCE method in DAKOTA: (a) main sensitivity index, (b) 2nd and 3rd order interactions, and (c) total sensitivity index.

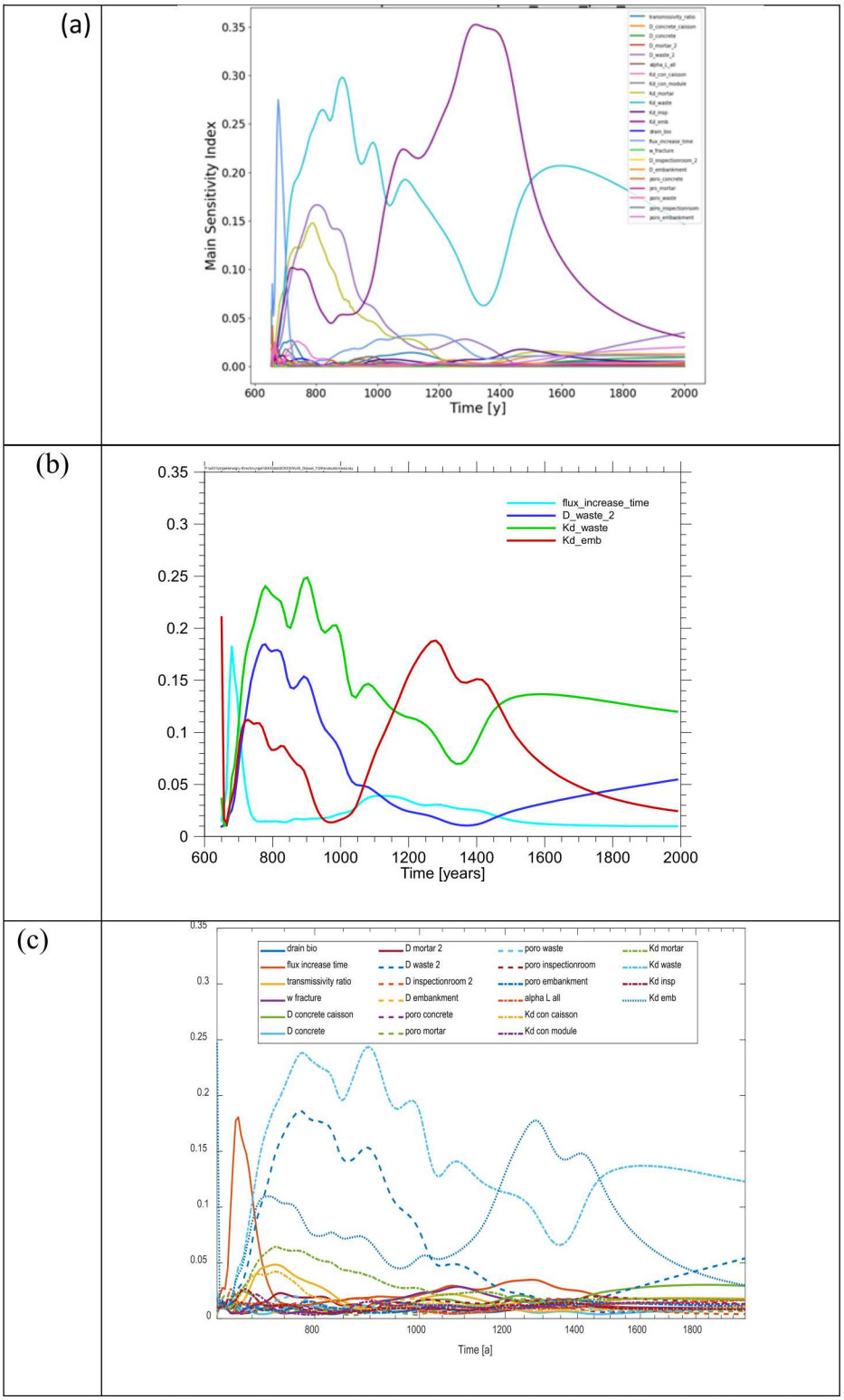



Figure 10.2 Comparision of first order sensitivity index calculated for Problem 3 with different methods by different participants: (a) PCE (SNL), (b) EASI (GRS), and (c) CUSUNORO (TUC)

## 10.2 References: Chapter 10

1. Becker, D.A. "Sensitivity Investigation of the Dessel Surface LILW Repository Mode l," Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
2. Stein, E., L. Swiler, and D. Sevougian. "Higher-order SIs and Interactions based on PCE (Using Dakota)," SAND2018-11081 PE. Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
3. Stein, E., L. Swiler, and D. Sevougian. "Sensitivity Analysis of Dessel Dataset with SNL's Dakota," SAND2018-11104PE. Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
4. Röhlig, K-J., E. Plischke, D-A. Becker, E.R. Stein, J. Govaerts, B.J. Debusschere, L. Koskinen, P. Kupiainen, C.D. Leigh, P. Mariner, O. Nummi, B. Pastina, D. Sevougian, S. Spießl, L.P. Swiler, E. Weetjens, and T. Zeitler. *"Sensitivity Analysis in Performance Assessment: Towards a Joint Approach,"* in the IGSC Symposium 2018, held in Rotterdam, The Netherlands, 10-11 October 2018. Nuclear Energy Agency (NEA), Issy-les-Moulineaux, France.
5. Weetjens, E. and J. Govaerts. "The cAt Project at Dessel: Overview of the Near Field Model Used in the SA Exercise," Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
6. Spiessl, S.M., S. Kucherenko, and D.A. Becker. "Application of the RS-HDMR Meta-Modeling Approach in the SobolGSA Software to LILW Models." Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
7. Govaerts, J. and E. Laloy. "Sensitivity Investigation of the Dessel Surface LILW Repository Model," Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
8. Kupiainen, P. "SCK Dataset Analysis Using Ecolego and Relevance of Analyses in Safety Assessment," Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
9. Plischke, E. "SA of the Dessel I-129 Example," Presented at the International SA-UQ Workshop 2018, Brussels, Belgium, October 4-5, 2018.
10. Becker, D-A. 2019. *Sensitivity Analysis in Repository Performance Assessment: Findings from an International Exercise,* in Proceedings of the International High-Level Radioactive Waste Management Conference (IHLRWMC) 2019, held in Knoxville, TN, April 14-18, 2019. American Nuclear Society, La Grange Park, IL.
11. Stein, E., L. Swiler, and D. Sevougian 2019. *Methods of Sensitivity Analysis in Geologic Disposal Safety Assessment (GDSA),* in Proceedings of the International High-Level Radioactive Waste Management Conference (IHLRWMC) 2019 held in Knoxville, TN, April 1, 2019. American Nuclear Society.
12. Hansen, F. D., C. D. Leigh, W. Steiniger, W. Bollingerfehr, and T. von Berlepsch 2015. *Proceedings of the 5th US/German Workshop on Salt Repository Research, Design, and*

*Operation*, FCRD-UFRD-2015-00514, SAND2015-0500R, Sandia National Laboratories, Albuquerque, NM, USA.

13. Sevougian, S. D. 2015. "Enhanced Performance Assessment Models for Generic Deep Geologic Repositories for HLW and SNF," *Workshop on Handling Uncertainties*, Harwell, Oxfordshire, UK, September 23-24, 2015, SAND2015-8006C, Sandia National Laboratories, Albuquerque, NM, USA.

14. Mariner, P. E., Connolly, L.A., Cunningham, L.J., Debusschere, B.J., Dobson, D.C., Frederick, J.M., Hammond, G.E., Jordan, S.H., LaForce, T.C., Nole, M.A., Park, H.D., Perry, F.V., Rogers, R.D., Seidl, D.T., Sevougian, S.D., Stein, E.R., Swift, P.N., Swiler, L.P., Vo. J., and M.G. Wallace 2019. "Progress in Deep Geologic Safety Assessment in the U.S. since 2010", M2SF-19SN010304041, U.S. Department of Energy, Spent Fuel and Waste Science and Technology Campaign, Office of Spent Fuel and Waste Disposition, Washington, DC, September 16, 2019.

# 11. SUMMARY AND FUTURE WORK

This report fulfills the GDSA Uncertainty and Sensitivity Analysis Methods work package (SF-19SN01030403) level 2 milestone – *Status Report on UQ/SA Tools in GDSA Framework* (M2SF-19SN010304031). It documents and demonstrates uncertainty quantification (UQ) and sensitivity analysis (SA) tools in the *GDSA Framework* in FY19, it summarizes the progress of UQ/SA tools during the UFD and SFWST Campaigns, and it describes additional UQ/SA tools whose future implementation would enhance the UQ/SA capability of *GDSA Framework*.

Development of UQ/SA tools within *GDSA Framework* is driven by the overarching objectives of *GDSA Framework* development including: 1) enabling increasingly coupled, mechanistic multi-physics modeling; 2) leveraging existing high-performance computing capabilities; 3) remaining flexible enough to take advantage of future advances in hardware, software, and simulation and analysis methods; and 4) developing in an open-source environment so that implementations are transparent and software is freely available to stakeholders.

Objectives specific to UQ/SA development include: 1) to make available standard sampling-based methods of uncertainty propagation, sensitivity analysis, and uncertainty quantification typically used within U.S. nuclear waste disposal programs; 2) to enable adoption of new methods consistent with the current standard of practice in the UQ/SA community and appropriate for high-dimensional, highly coupled, nonlinear problems resulting from the implementation of mechanistic multi-physics simulations; and 3) to create a consistent, common framework that enables a user to perform a range of uncertainty and sensitivity analyses for a particular problem or set of simulations. These are important goals for performance assessments now and in the future.

There is a rich legacy of UQ/SA being performed in repository assessment. In this report, the regulatory considerations have been documented along with the treatment of epistemic vs. aleatory uncertainty and the sampling strategies used in prior analyses. New approaches being used throughout the community have been presented, such as variance-based sensitivity indices and surrogate models. These new approaches have been demonstrated on two reference case studies, that of a shale and crystalline repository. Additional UQ/SA algorithms have been provided, including methods to handle time-varying responses, new graphical sensitivity approaches, and adaptive sampling methods such as importance sampling which can focus additional samples on regions of interest without attempting to explore the entire parameter space with limited samples. Advanced methods also include Bayesian calibration methods which are used to estimate posterior parameter distributions on model parameters given experimental data (and such parameter distributions can be then carried forward in a UQ study), and multifidelity UQ methods which augment a small number of high-fidelity code runs with a large number of low-fidelity runs to obtain statistics on the high-fidelity model. It is important to maintain awareness of and collaboration with the international community. Methods used by the international community have been described along with existing collaborations with that community. Finally, a list of software tools implementing some of the existing and new UQ/SA methods has been provided.

Development, implementation, and demonstration of new tools and methods for uncertainty and sensitivity analysis in *GDSA Framework* will maintain leadership of the repository science community in UQ/SA methods, while also maintaining an infrastructure of proven tools.

Geologic repository performance assessment in the U.S. involves coupled, multi-physics modeling at high resolution, large parameter spaces, and greater use of random (stochastic) field modeling. It requires high-performance computing and costly sample evaluations. UQ/SA methods discussed in this report, including surrogate modeling to reduce computational expense, variance-based sensitivity analysis to quantify importance of parameter interactions in a multi-physics system, multi-fidelity methods, and adaptive importance sampling will enable analysis methods to keep pace with the increasing sophistication of the physics models. *GDSA Framework* development seeks to keep abreast of improvements to existing UQ/SA methods, employ new methods, and maintain an infrastructure of proven tools that can be extended to support computationally expensive analyses.

There are several possible avenues of future UQ/SA development, including:

- Time-dependent analysis methods
- Dimension reduction
- Importance sampling
- Surrogate models
- Bayesian calibration
- Estimating/generating distributions/ranges for uncertain inputs
- Multifidelity UQ approaches
- Model uncertainty
- Density-based SA methods

Future work will involve development of the items above, which align with the *GDSA Framework* and the repository performance assessment needed.

# A: SOFTWARE TOOLS AND FRAMEWORKS

There are a variety of individual codes, code packages, and software frameworks or toolkits which implement various sensitivity analysis and uncertainty quantification methods. The listing we provide here is not a broad survey of open-source and commercial packages. Rather, it is a brief summary of tools that Sandia has developed and some tools which we are aware of through collaborations. References for each software tool are listed in the section describing that tool.

## A.1 LHS

Sandia developed one of the first implementation of Latin Hypercube sampling in the early 1980s, written by Ron Iman. The LHS code was written in F77. Greg Wyss and others converted it to a Fortran90 code in the late 1990s. Subsequently, it was adopted as a sampling tool within the Dakota software framework in 2003 by Laura Swiler. The Dakota team continues to maintain the Fortran90 version of LHS, including making small changes to maintain compatibility with various compilers and language updates. The following reports document the code evolution and an associated scatterplot tool.

- Iman, R.L., Davenport, J.M., and Ziegler, D.K. (1980). "*Latin Hypercube Sampling (Program User's Guide),*" Technical Report SAND79-1473, Sandia National Laboratories, Albuquerque, NM.
- Iman, R.L., and Shortencarier, M.J. (1984). "*A Fortran 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use with Computer Models*," NUREG/CR-3624, Technical Report SAND83-2365, Sandia National Laboratories, Albuquerque, NM.
- Swiler, L. P. and G. D. Wyss (2004). "*A User's Guide to Sandia's Latin Hypercube Sampling Software: LHS Unix Library/Standalone Version*." Technical Report SAND2004-2439. Sandia National Laboratories, Albuquerque, NM.
- Shortencarier M.J. and J.*C.* Helton (1999). "*A FORTRAN 77 Program and User's Guide for the Statistical Analyses of Scatterplots to Identify Important Factors in Large-Scale Simulations".* Technical Report SAND99-1058*,* Sandia National Laboratories, Albuquerque, NM.

## A.2 Correlation coefficients

The code to calculate simple and partial correlations for actual values (Pearson correlations) or for ranked values (Spearman correlation) exists in Dakota. The Dakota team re-wrote the calculations to make these as efficient as possible for large data sets with multiple inputs and outputs. The Dakota team also improved the calculations when the partial correlation calculation becomes ill-conditions. The original correlation code is listed in the SAND report below.

- Iman R.L., M.J. Shortencarier and J.D. Johnson (1985). *A FORTRAN 77 Program and User's Guide for the Calculation of Partial Correlation and Standardized Regression Coefficients.* Technical Report NUREG/CR-4122, SAND85-0044, Sandia National Laboratories, Albuquerque, NM.

## A.3 Stepwise

Stepwise regression was originally performed with the STEPWISE code, written in Fortran77 and meant to be used as a companion code using data from an LHS study. The original STEPWISE code is no longer maintained. We now use a set of Python scripts developed by Emily Stein to postprocess a Dakota sampling study and generate the stepwise information.

- Iman, R. (1976). "*Stepwise Regression."* Technical Report SAND1976-0364. Sandia National Laboratories, Albuquerque, NM.
- Iman R.L., J.M. Davenport, E.L. Frost and M.J. Shortencarier (1980). "*Stepwise Regression with PRESS and Rank Regression (Program User's Gui*de)*."* Technical Report SAND79-1472*,* Sandia National Laboratories, Albuquerque, NM.

## A.4 Dakota

Dakota is an open-source toolkit of algorithms that contains software for optimization and uncertainty quantification (UQ). It is available at: https://dakota.sandia.gov. The main classes of algorithms in Dakota include:

- uncertainty quantification with sampling, reliability, stochastic expansion, and epistemic methods
- parameter estimation using nonlinear least squares or Bayesian inference
- optimization with gradient and nongradient-based methods
- sensitivity/variance analysis with design of experiments and parameter study methods.

Dakota contains the uncertainty quantification and sensitivity analysis methods typically used in the U.S. repository program. Dakota implements Latin Hypercube Sampling (LHS) with correlation control on input parameters. It calculates moments on responses of interest as well as correlation matrices (simple, partial, and rank correlations) between inputs and outputs. Dakota also contains an algorithm for performing incremental LHS which allows one to double an initial LHS study such that the second LHS study is a Latin design and the combined initial and second LHS studies together form a Latin hypercube design. Dakota allows nested studies to perform an "outer loop" epistemic sampling and an "inner loop" aleatory sampling to generate ensembles of distributions. Dakota returns table of input and output amenable to further processing and visualization with additional tools developed within GDSA Framework or by individual user. Additional methods that have been implemented in Python for use in GDSA Framework include calculation of standardized regression coefficients via stepwise linear regression and calculation of partial correlation coefficients via iterative loop.

The UQ/SA methods in Dakota have evolved as the standard of practice evolves. Over the past ten years, the Dakota team has invested in methods which calculate the Sobol' variance-based sensitivity indices in an efficient manner. Currently, a Dakota user can calculate these by extensive sampling of the simulation code, by using surrogate methods such as regression or Gaussian process models, and by the use of polynomial chaos expansions. Dakota is an actively maintained and developed code with formal releases issued twice per year.

- Adams, B.M., Bauman, L.E., Bohnhoff, W.J., Dalbey, K.R., Ebeida, M.S., Eddy, J.P., Eldred, M.S., Hough, P.D., Hu, K.T., Jakeman, J.D., Stephens, J.A., Swiler, L.P., Vigil,

D.M., and Wildey, T.M. (2014, updated annually), "*Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 User's Manual*," Sandia Technical Report SAND2014-4633. Sandia National Laboratories, Albuquerque, NM Dakota version 6.10, 2019. https://dakota.sandia.gov

## A.5 UQTk

The UQ Toolkit (UQTk) is an open source collection of libraries and tools for the quantification of uncertainty in numerical model predictions. It offers intrusive and non-intrusive methods for propagating input uncertainties through computational models, tools for sensitivity analysis, methods for sparse surrogate construction, and Bayesian inference tools for inferring parameters from experimental data. The website is: http://www.sandia.gov/UQToolkit/
Note that Dakota is a large, integrated standalone executable which typically runs separately from the user's simulation code. In contrast, UQTk offers a number of "functions" or "libraries" which can be incorporated into simulation codes more directly. UQTk has several capabilities that are similar to Dakota but with a different implementation (Python and C++). Some focus areas include polynomial chaos expansion, compressive sensing and regularized regression, sparse quadrature, Karhunen-Loeve stochastic expansions, and the Sobol' variance-based indices.

- Sargsyan, K., C. Safta, K. Chowdhary, S. Castorena, S. de Bord, B. Debusschere (2017). "*UQTk Version 3.0.3 User Manual*." Technical Report SAND2017-5747. Sandia National Laboratories, Albuquerque, NM

## A.6 Variance-based decomposition

The Sobol' variance-based sensitivity indices can be calculated in Dakota or in UQTk.

## A.7 Other non-Sandia packages

There are a plethora of software tools and libraries available today. This is not a comprehensive list, but we wish to highlight a few:

- R: The R language is an open source statistical language which has thousands of contributed packages which perform various functions such as regression, surrogate construction, sensitivity analysis, etc. https://cran.r-project.org/
- UQLab: UQLab is a general purpose, MATLAB-based Uncertainty Quantification framework developed at ETH Zurich (Switzerland). It is made of open-source scientific modules which are smoothly connected through UQLab to carry out uncertainty propagation through Monte Carlo sampling, sensitivity analysis, reliability analysis (computation of rare event probabilities), build surrogate models (polynomial chaos expansions, Kriging, low-rank tensor approximations, etc.) and more. https://www.uqlab.com/
- Simlab: SimLab is a professional tool for model developers, scientists and professionals, to learn, use and exploit global uncertainty and sensitivity analysis techniques. SimLab provides a reference implementation of the most recent global sensitivity analysis techniques. SimLab is an ongoing project since 1985. The European Research Comission

group headed by Andrea Saltelli is a key contributor to Simlab. https://ec.europa.eu/jrc/en/samo/simlab

- MUQ: MUQ is a collection of tools for constructing models and a collection of uncertainty quantification (UQ)–focused algorithms for working on those models. Our goal is to provide an easy and clean way to set up and efficiently solve UQ problems. MUQ is developed under Prof. Youssef Marzouk's lab at MIT: http://muq.mit.edu/ (note: Prof. Marzouk was previously a von Neumann fellow at Sandia. MUQ is partly sponsored by the DOE QUEST SciDAC Institute). MUQ contains tools for:
    - Combining many simple model components into a single sophisticated model.
    - Propagating derivative information through sophisticated models.
    - Integrating ordinary differential equations and differential algebraic equations (via Sundials)
    - Performing Markov chain Monte Carlo (MCMC) sampling
    - Constructing polynomial chaos expansions (PCE)
    - Computing Karhunen-Loeve expansions
    - Building optimal transport maps
    - Solving nonlinear constrained optimization problems (both internally and through NLOPT)
    - Regression (including Gaussian process regression)
- Sobol GSA: SobolGSA is general purpose GUI driven global sensitivity analysis and metamodeling software developed by S. Kucherenko. It can be used to compute various sensitivity measures and/or to develop metamodels. All techniques implemented in SobolGSA make use of Quasi Monte Carlo sampling based on Sobol sequences. SobolGSA is primarily a MATLAB based toolkit. http://www.imperial.ac.uk/process-systems-engineering/research/free-software/sobolgsa-software/

# B. Sobol' Procedure for Variance Decomposition

## B.1 Introduction

The procedure that bears his name, Sobol' variance decomposition, was introduced by I.M. Sobol' [1]. This procedure for variance decomposition and the determination of associated variance-based sensitivity measures has become very popular. As a result, an extensive literature exists (e.g., [1-10]).

The notation needed to describe Sobol' variance decomposition is complicated and often described in a very compact manner in the journal literature in order to meet the length requirements imposed on journal publications. As a result, it can be difficult for the uninitiated to develop an understanding of exactly what is the mathematical structure that underlies uncertainty and sensitivity analyses based on Sobol' variance decomposition. The purpose of this appendix is to lead the interested reader through the details of this structure. To accomplish this, the structure of Sobol' variance decomposition is described separately for (i) the less complicated case involving a function $f(x_1, x_2, x_3)$ of only 3 uncertain variables and (ii) the notationally more complex case involving a function $f(x_1, x_2, ..., x_n)$ of $n$ uncertain variables.

The following topics are considered in this appendix: (i) Variance decomposition for $f(x_1, x_2, x_3)$ in Sect. B.2, (ii) Approximation of variance decomposition for $f(x_1, x_2, x_3)$ in Sect. B.3, (iii) Variance decomposition for $f(x_1, x_2, ..., x_n)$ in Sect. B.4, (iv) Approximation of variance decomposition for $f(x_1, x_2, ..., x_n)$ in Sect. B.5, (v) Additional considerations (i.e., surrogate models and correlated variables) in Sect. B.6.

## B.2 Variance Decomposition for $f(x_1, x_2, x_3)$

The Sobol' variance decomposition of a function $f(\mathbf{x})$ is based on a representation of $f(\mathbf{x})$ by a sum of orthogonal functions as described in Refs. [1; 2]. As the notation for this decomposition becomes rather cumbersome and tends to obscure the nature of the decomposition when $\mathbf{x} = [x_1, x_2, ..., x_n]$ is a vector of arbitrary length $n$, the decomposition will be initially described for $\mathbf{x} = [x_1, x_2, x_3]$. As required for the Sobol' decomposition, the epistemically uncertain variables $x_1$, $x_2$ and $x_3$ are assumed to be independent; further, the uncertainty in $x_1$, $x_2$ and $x_3$ is characterized by the probability spaces $(\mathcal{X}_1, \mathbb{X}_1, m_1)$, $(\mathcal{X}_2, \mathbb{X}_2, m_2)$ and $(\mathcal{X}_3, \mathbb{X}_3, m_3)$ with associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$.

For $\mathbf{x} = [x_1, x_2, x_3]$, the Sobol' variance decomposition is based on the following representation for $f(\mathbf{x})$:

$$f(\mathbf{x}) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{13}(x_1, x_3) + f_{23}(x_2, x_3) + f_{123}(x_1, x_2, x_3) \tag{B.2.1}$$

with

$$
\begin{aligned}
f_0 &= \mathrm{E}[f(\mathbf{x})] \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \qquad \text{(B.2.2)} \\
&= \int_{\mathcal{X}} f(\mathbf{x})d(\mathbf{x})\mathrm{d}V \quad \text{with} \quad \mathcal{X}=\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3 \quad \text{and} \quad d(\mathbf{x})=d_3(x_3)d_2(x_2)d_1(x_1),
\end{aligned}
$$

$$
\begin{aligned}
f_1(x_1) &= \mathrm{E}[f(\mathbf{x}\mid x_1)] - f_0 \quad \text{for } x_1\in\mathcal{X}_1 \\
&= \int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(\mathbf{x}\mid x_1)d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0 \qquad \text{(B.2.3)} \\
&= \int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0,
\end{aligned}
$$

$$
\begin{aligned}
f_2(x_2) &= \mathrm{E}[f(\mathbf{x}\mid x_2)] - f_0 \quad \text{for } x_2\in\mathcal{X}_2 \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_3} f(\mathbf{x}\mid x_2)d_3(x_3)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_1 - f_0 \qquad \text{(B.2.4)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2 - f_0,
\end{aligned}
$$

$$
\begin{aligned}
f_1(x_3) &= \mathrm{E}[f(\mathbf{x}\mid x_3)] - f_0 \quad \text{for } x_3\in\mathcal{X}_3 \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f(\mathbf{x}\mid x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_2\mathrm{d}x_1 - f_0 \qquad \text{(B.2.5)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f(x_1,x_2,x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_2\mathrm{d}x_1 - f_0,
\end{aligned}
$$

$$
\begin{aligned}
f_{12}(x_1,x_2) &= \mathrm{E}[f(\mathbf{x})\mid x_1,x_2] - f_1(x_1) - f_2(x_2) - f_0 \quad \text{for } (x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2 \\
&= \int_{\mathcal{X}_3} f(\mathbf{x}\mid x_1,x_2)d_3(x_3)\mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0 \qquad \text{(B.2.6)} \\
&= \int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)\mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0,
\end{aligned}
$$

$$
\begin{aligned}
f_{13}(x_1,x_3) &= \mathrm{E}[f(\mathbf{x})\mid x_1,x_3] - f_1(x_1) - f_3(x_3) - f_0 \quad \text{for } (x_1,x_3)\in\mathcal{X}_1\times\mathcal{X}_3 \\
&= \int_{\mathcal{X}_2} f(\mathbf{x}\mid x_1,x_3)d_2(x_2)\mathrm{d}x_2 - f_1(x_1) - f_3(x_3) - f_0 \qquad \text{(B.2.7)} \\
&= \int_{\mathcal{X}_2} f(x_1,x_2,x_3)d_2(x_2)\mathrm{d}x_2 - f_1(x_1) - f_3(x_3) - f_0,
\end{aligned}
$$

$$
\begin{aligned}
f_{23}(x_2,x_3) &= \mathrm{E}[f(\mathbf{x}) \mid x_2,x_3] - f_2(x_2) - f_3(x_3) - f_0 \quad \text{for } (x_2,x_3) \in \mathcal{X}_2 \times \mathcal{X}_3 \\
&= \int_{\mathcal{X}_1} f(\mathbf{x} \mid x_2,x_3) d_1(x_1) \mathrm{d}x_1 - f_2(x_2) - f_3(x_3) - f_0 \\
&= \int_{\mathcal{X}_1} f(x_1,x_2,x_3) d_1(x_1) \mathrm{d}x_1 - f_2(x_2) - f_3(x_3) - f_0,
\end{aligned} \tag{B.2.8}
$$

$$
\begin{aligned}
f_{123}(x_1,x_2,x_3) &= \mathrm{E}[f(\mathbf{x}) \mid x_1,x_2,x_3] - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) \\
&\quad - f_{12}(x_1,x_2) - f_{13}(x_1,x_3) - f_{23}(x_2,x_3) \\
&\quad \text{for } (x_1,x_2,x_3) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \\
&= f(x_1,x_2,x_3) - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) \\
&\quad - f_{12}(x_1,x_2) - f_{13}(x_1,x_3) - f_{23}(x_2,x_3).
\end{aligned} \tag{B.2.9}
$$

In the preceding equations, E denotes expected value; $\mathrm{E}[f(\mathbf{x} \mid x_i)]$ denotes the expected value of $f(\mathbf{x})$ evaluated with the $i^{\text{th}}$ component of $\mathbf{x}$ given the fixed value $x_i$; and $\mathrm{E}[f(\mathbf{x} \mid x_i,x_j)]$ denotes the expected value of $f(\mathbf{x})$ evaluated with the $i^{\text{th}}$ and $j^{\text{th}}$ components of $\mathbf{x}$ given the fixed values $x_i$ and $x_j$. The expected value $\mathrm{E}[f(\mathbf{x}) \mid x_1,x_2,x_3]$ is defined similarly, with the result that

$$
\mathrm{E}[f(\mathbf{x}) \mid x_1,x_2,x_3] = f(x_1,x_2,x_3) \tag{B.2.10}
$$

in Eq. (B.2.9).

The validity of the equality in Eq. (B.2.1) follows immediately by simply summing $f_0, f_1(x_1), ..., f_{123}(x_1,x_2,x_3)$, with all terms except $f(x_1,x_2,x_3)$ in the definition of $f_{123}(x_1,x_2,x_3)$ cancelling. Specifically,

$$
\begin{aligned}
&f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1,x_2) + f_{13}(x_1,x_3) + f_{23}(x_2,x_3) + f_{123}(x_1,x_2,x_3) \\
&= f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1,x_2) + f_{13}(x_1,x_3) + f_{23}(x_2,x_3) \\
&\quad + f(x_1,x_2,x_3) - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) - f_{12}(x_1,x_2) - f_{13}(x_1,x_3) - f_{23}(x_2,x_3) \\
&= f(x_1,x_2,x_3).
\end{aligned} \tag{B.2.11}
$$

The expected value of $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ with respect to any one of their arguments (i.e., $x_1$, $x_2$ or $x_3$ as appropriate) is zero. As discussed later, this property is central to establishing that the indicated functions are orthogonal.

For $f_1(x_1)$, the expected value with respect to $x_1$ is given by

$$
\begin{aligned}
\mathrm{E}[f_1(x_1)] &= \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0\right]d_1(x_1)\mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 - \int_{\mathcal{X}_1} f_0 d_1(x_1)\mathrm{d}x_1 \qquad \text{(B.2.12)} \\
&= f_0 - f_0 \\
&= 0.
\end{aligned}
$$

Similarly, $\mathrm{E}[f_2(x_2)] = \mathrm{E}[f_3(x_3)] = 0$.

For $f_{12}(x_1,x_2)$, the expected value with respect to $x_1$ with $x_2$ fixed is given by

$$
\begin{aligned}
\mathrm{E}[f_{12}(x_1,x_2)\,|\,x_2] &= \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)\mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0\right]d_1(x_1)\mathrm{d}x_1 \\
&= \left[\int_{\mathcal{X}_1}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_1 - \int_{\mathcal{X}_1} f_0 d_1(x_1)\mathrm{d}x\right] \\
&\quad - \int_{\mathcal{X}_1} f_1(x_1)d_1(x_1)\mathrm{d}x - \int_{\mathcal{X}_1} f_2(x_2)d_1(x_1)\mathrm{d}x \\
&= \left[\int_{\mathcal{X}_1}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_1 - f_0\right] - 0 - f_2(x_2) \\
&= f_2(x_2) - f_2(x_2) \\
&= 0.
\end{aligned}
\qquad \text{(B.2.13)}
$$

Similarly, $\mathrm{E}[f_{12}(x_1,x_2)\,|\,x_1]$, $\mathrm{E}[f_{13}(x_1,x_3)\,|\,x_3]$, $\mathrm{E}[f_{13}(x_1,x_3)\,|\,x_1]$, $\mathrm{E}[f_{23}(x_2,x_3)\,|\,x_3]$ and $\mathrm{E}[f_{23}(x_2,x_3)\,|\,x_2]$ are equal to zero.

In like manner, the expected value $\mathrm{E}[f_{123}(x_1,x_2,x_3)\,|\,x_2,x_3]$ of $f_{123}(x_1,x_2,x_3)$ with respect to $x_1$ with $x_2$ and $x_3$ fixed is given by

$$
\begin{aligned}
\mathrm{E}[f_{123}(x_1,x_2,x_3)\,|\,x_2,x_3)] &= \int_{\mathcal{X}_1} f_{123}(x_1,x_2,x_3)d_1(x_1)\mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} \big[\, f(x_1,x_2,x_3) - f_0 - f_1(x_1) - f_2(x_2) - f_3(x_3) \\
&\qquad - f_{12}(x_1,x_2) - f_{13}(x_1,x_3) - f_{23}(x_2,x_3)\big] d_1(x_1)\mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} f_{123}(x_1,x_2,x_3)d_1(x_1)\mathrm{d}x_1 - f_0 - 0 - f_2(x_2) - f_3(x_3) - 0 - 0 - f_{23}(x_2,x_3) \qquad \text{(B.2.14)} \\
&= \left[\int_{\mathcal{X}_1} f_{123}(x_1,x_2,x_3)d_1(x_1)\mathrm{d}x_1 - f_2(x_2) - f_3(x_3) - f_0\right] - f_{23}(x_2,x_3) \\
&= f_{23}(x_2,x_3) - f_{23}(x_2,x_3) \\
&= 0.
\end{aligned}
$$

Similarly, $\mathrm{E}[f_{123}(x_1,x_2,x_3)\,|\,x_1,x_3]$ and $\mathrm{E}[f_{123}(x_1,x_2,x_3)\,|\,x_1,x_2]$ are equal to zero.

Because the expected values of $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ with respect to any one of their arguments (i.e., $x_1$, $x_2$ or $x_3$ as appropriate) is zero, it follows immediately that the expected values of these functions with respect to all of their arguments are also zero. Specifically,

$$\mathrm{E}[f_i(x_i)] = 0 \quad \text{for } i = 1,2,3, \qquad \text{(B.2.15)}$$

$$\mathrm{E}[f_{ij}(x_i,x_j)] = 0 \quad \text{for } 1 \le i < j \le 3, \qquad \text{(B.2.16)}$$

$$\mathrm{E}[f_{ijk}(x_i,x_j,x_k)] = 0. \qquad \text{(B.2.17)}$$

The preceding equalities are immediate because the expected values will always involve one or more integrals with respect to an argument of the function under consideration.

The Sobol' variance decomposition is based on the orthogonality of the functions on the right-hand side of Eq. (B.2.1). In the context of the present example with probability spaces $(\mathcal{X}_1,\mathbb{X}_1,m_1)$, $(\mathcal{X}_2,\mathbb{X}_2,m_2)$, $(\mathcal{X}_3,\mathbb{X}_3,m_3)$ and associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$, two functions *g* and *h* defined on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ are said to be orthogonal if

$$
\begin{aligned}
\langle g,h\rangle &= \int_{\mathcal{X}} g(\mathbf{x})h(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} g(x_1,x_2,x_3)h(x_1,x_2,x_3)d_3(x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \qquad \text{(B.2.18)} \\
&= 0,
\end{aligned}
$$

where $\langle g,h\rangle$ is the commonly used notation for the inner product of $g$ and $h$ defined by the preceding integrals. The definition of orthogonality in Eq. (B.2.18) is a generalization of the concept that two vectors in $R^2$ or $R^3$ are orthogonal when their inner, or dot, product is equal to zero. Important properties of the inner product defined in Eq. (B.2.18) include

$$\langle g,h\rangle = \langle h,g\rangle, \quad \langle cg,h\rangle = c\langle g,h\rangle \quad \text{for } c \in R, \tag{B.2.19}$$

and

$$\langle g,h\rangle = \left\langle \sum_{i=1}^{m} g_i, \sum_{j=1}^{n} h_j \right\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} \langle g_i, h_j\rangle \quad \text{for } g(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{x}), \quad h(\mathbf{x}) = \sum_{j=1}^{n} h_j(\mathbf{x}). \tag{B.2.20}$$

A demonstration of the orthogonality of $f_0$, $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ follows. For $f_0$, and $f_1(x_1)$,

$$\begin{aligned} \langle f_0, f_1\rangle &= \int_{\mathcal{X}} f_0 f_1(x_1) d(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= f_0 \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f_1(x_1) d_3(x_3) d_2(x_2) d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \\ &= f_0 \ \mathrm{E}[f_1(x_1)] \\ &= 0, \end{aligned} \tag{B.2.21}$$

with the final equality following from the equality $\mathrm{E}[f_1(x_1)] = 0$ obtained in Eq. (B.2.12). Similarly, $\langle f_0, g\rangle = 0$ for $g$ corresponding to $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$.

For $f_1(x_1)$ and $f_2(x_2)$,

$$
\begin{aligned}
\langle f_1, f_2 \rangle &= \int_{\mathcal{X}} f_1(x_1) f_2(x_2) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f_1(x_1) f_2(x_2) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_3} \left[ \int_{\mathcal{X}_1} f_1(x_1) d_1(x_1) \mathrm{d}x_1 \right] \left[ \int_{\mathcal{X}_2} f_2(x_2) d_2(x_2) \mathrm{d}x_2 \right] d_3(x_3) \mathrm{d}x_3 \\
&= \int_{\mathcal{X}_3} \mathrm{E}[f_1(x_1)] \, \mathrm{E}[f_2(x_2)] d_3(x_3) \mathrm{d}x_3 \\
&= \int_{\mathcal{X}_3} 0 \times 0 \; d_3(x_3) \mathrm{d}x_3 \\
&= 0,
\end{aligned}
\tag{B.2.22}
$$

with the introduced zeros resulting from the equalities $\mathrm{E}[f_1(x_1)] = 0$ and $\mathrm{E}[f_2(x_2)] = 0$ obtained in conjunction with Eq. (B.2.12). Similarly, $\langle f_i, g \rangle = 0$ for $g$ corresponding to $f_{j \neq i}(x_j)$, $f_{12}(x_1, x_2)$, $f_{13}(x_1, x_3)$, $f_{23}(x_2, x_3)$ and $f_{123}(x_1, x_2, x_3)$.

For $f_{12}(x_1, x_2)$ and $f_{23}(x_2, x_3)$,

$$
\begin{aligned}
\langle f_{12}, f_{23} \rangle &= \int_{\mathcal{X}} f_{12}(x_1, x_2) f_{23}(x_2, x_3) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f_{12}(x_1, x_2) f_{23}(x_2, x_3) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f_{12}(x_1, x_2) \left[ \int_{\mathcal{X}_3} f_{23}(x_2, x_3) d_3(x_3) \mathrm{d}x_3 \right] d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f_{12}(x_1, x_2) \mathrm{E}[f_{23}(x_2, x_3) \mid x_2] d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f_{12}(x_1, x_2) \times 0 \; d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 \\
&= 0,
\end{aligned}
\tag{B.2.23}
$$

with the introduced zero resulting from the equality $\mathrm{E}[f_{23}(x_2, x_3) \mid x_2] = 0$ indicated in conjunction with Eq. (B.2.13). Similarly, inner products $\langle f_{12}, f_{13} \rangle$, $\langle f_{13}, f_{23} \rangle$, $\langle f_{12}, f_{123} \rangle$, $\langle f_{13}, f_{123} \rangle$ and $\langle f_{23}, f_{123} \rangle$ are also equal to zero.

As indicated in conjunction with Eqs. (B.2.21)-(B.2.23),

$$
\langle g, h \rangle = 0 \tag{B.2.24}
$$

when $g$ and $h$ are two different functions from $f_0$, $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ because the integral defining $\langle g,h\rangle$ will always involve an isolated integral with respect to a single component (i.e., $x_1$, $x_2$ or $x_3$) of $g$ or $h$ (see Eqs. (B.2.15) -(B.2.17)). Thus, $f_0$, $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ constitute a set of orthogonal functions with respect to the inner product in Eq. (B.2.18).

The variance decomposition for $f$ that derives from the representation in Eq. (B.2.1) is now considered. The variance for a function $g$ defined for $\mathbf{x} \in \mathcal{X}$ is defined by

$$\mathrm{V}(g) = \int_{\mathcal{X}} [g(\mathbf{x}) - \mathrm{E}(g)]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} = \langle g - g_0, g - g_0 \rangle, \tag{B.2.25}$$

where $g_0 = \mathrm{E}(g)$ and $\langle \dots, \dots \rangle$ is the inner product defined in Eq. (B.2.18). For notational convenience, $g_1, g_2, \dots, g_7$ are used to represent $f_1, f_2, \dots, f_{123}$. The variance V($f$) of $f$ is then given by

$$\begin{aligned}
\mathrm{V}(f) &= \int_{\mathcal{X}} [f(\mathbf{x}) - f_0]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \langle f - f_0, f - f_0 \rangle \\
&= \left\langle \sum_{i=1}^{7} g_i, \sum_{i=1}^{7} g_i \right\rangle \\
&= \sum_{i=1}^{7}\sum_{j=1}^{7} \langle g_i, g_j \rangle \\
&= \sum_{i=1}^{7} \langle g_i, g_i \rangle \quad \text{from Eq. (B.2.24)} \\
&= \sum_{i=1}^{7} \mathrm{V}(g_i),
\end{aligned} \tag{B.2.26}$$

with the last equality following since $\mathrm{E}(g_i) = 0$ (see Eqs. (B.2.15)-(B.2.17)). Equivalently,

$$\mathrm{V}(f) = \mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23}) + \mathrm{V}(f_{123}), \tag{B.2.27}$$

which is the desired Sobol' decomposition of V($f$) in terms of the variances of $f_1, f_2, \dots, f_{123}$.

Variance provides a measure of the variability about the expected value, with a small variance indicating little variability about the expected value and a large variance indicating a large variability about the expected value (see Eq, (B.2.25)).The significance of the decomposition in

Eq. (B.2.27) is that $V(f_i)$ provides a measure of the amount of this variability that is due solely to $x_i$, $V(f_{ij})$ provides a measure of the amount of this variability that is due to the interaction of $x_i$ and $x_j$ after a correction has been made to remove the individual effects of $x_i$ and $x_j$, and $V(f_{123})$ provides a measure of the amount of this variability that is due to interaction $x_1$, $x_2$ and $x_3$ after a correction has been made for the individual effects of $x_1$, $x_2$, $x_3$ and interactions between pairs of these variables.

Division of both sides of Eq. (B.2.27) by V(*f*) yields

$$1 = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123}, \tag{B.2.28}$$

with

$$S_i = V(f_i) / V(f) \text{ for } i = 1,2,3, \tag{B.2.29}$$

$$S_{ij} = V(f_{ij}) / V(f) \text{ for } 1 \le i < j \le 3, \tag{B.2.30}$$

$$S_{123} = V(f_{123}) / V(f). \tag{B.2.31}$$

The terms on the right-hand side of Eq. (B.2.28) equal the fractions of the variance V(*f*) of $f$ that can be accounted for by $f_1, f_2, ..., f_{123}$ and can be used as measures of sensitivity. An additional measure of sensitivity is given by

$$\begin{aligned} S_{1T} &= [V(f_1) + V(f_{12}) + V(f_{13}) + V(f_{123})] / V(f) \\ &= S_1 + S_{12} + S_{13} + S_{123}, \end{aligned} \tag{B.2.32}$$

which measures the effect on V(*f*) of $x_1$ and all interactions of $x_1$ with $x_2$ and $x_3$. Corresponding measures $S_{2T}$ and $S_{3T}$ for $x_2$ and $x_3$ are defined analogously. However, there is overlap in the effects measured by $S_{1T}$, $S_{2T}$ and $S_{3T}$.

The variances $V(f_1), V(f_2), ..., V(f_{123})$ have representations that contain integrals involving *f*. As an initial example,

$$\begin{aligned} V(f_1) &=_1 \langle f_1 - E(f_1), f_1 - E(f_1) \rangle \\ &=_2 \langle f_1, f_1 \rangle \text{ from } E(f_1) = 0; \text{ see Eq. (B.2.12)} \\ &=_3 \langle E[f(\mathbf{x}) \mid x_1] - f_0, E[f(\mathbf{x}) \mid x_1] - f_0 \rangle \text{ from Eq. (B.2.3)} \\ &=_4 V\{E[f(\mathbf{x}) \mid x_1]\} \text{ from } E\{E[f(\mathbf{x}) \mid x_1]\} = f_0. \end{aligned} \tag{B.2.33}$$

Further, continuation from Equality 3 in the preceding equation yields

$$
\begin{aligned}
\mathrm{V}(f_1) &= \int_{\mathcal{X}_1} \mathrm{E}[f(\mathbf{x}) \mid x_1]^2 d_1(x_1)\mathrm{d}x_1 - 2\int_{\mathcal{X}_1} f_0 \mathrm{E}[f(\mathbf{x}) \mid x_1] d_1(x_1)\mathrm{d}x_1 + \int_{\mathcal{X}_1} f_0^2 d_1(x_1)\mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} \left[ \int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1, x_2, x_3 \mid x_1) d_3(x_3) d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 \right]^2 d_1(x_1)\mathrm{d}x_1 - 2f_0^2 + f_0^2 \qquad \text{(B.2.34)} \\
&= \int_{\mathcal{X}_1} \left[ \int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1, x_2, x_3 \mid x_1) d_3(x_3) d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 \right]^2 d_1(x_1)\mathrm{d}x_1 - f_0^2 .
\end{aligned}
$$

In summary,

$$
\begin{aligned}
\mathrm{V}(f_1) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1]\} \\
&= \int_{\mathcal{X}_1} \left[ \int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1, x_2, x_3 \mid x_1) d_3(x_3) d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 \right]^2 d_1(x_1)\mathrm{d}x_1 - f_0^2 .
\end{aligned} \qquad \text{(B.2.35)}
$$

Similar derivations for $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ produce the representations

$$
\begin{aligned}
\mathrm{V}(f_2) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_2]\} \\
&= \int_{\mathcal{X}_2} \left[ \int_{\mathcal{X}_1}\int_{\mathcal{X}_3} f(x_1, x_2, x_3 \mid x_2) d_3(x_3) d_1(x_1)\mathrm{d}x_3\mathrm{d}x_1 \right]^2 d_2(x_2)\mathrm{d}x_2 - f_0^2
\end{aligned} \qquad \text{(B.2.36)}
$$

and

$$
\begin{aligned}
\mathrm{V}(f_3) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_3]\} \\
&= \int_{\mathcal{X}_3} \left[ \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f(x_1, x_2, x_3 \mid x_3) d_2(x_2) d_1(x_1)\mathrm{d}x_2\mathrm{d}x_1 \right]^2 d_3(x_3)\mathrm{d}x_3 - f_0^2 .
\end{aligned} \qquad \text{(B.2.37)}
$$

Th variance $\mathrm{V}(f_{12})$ of $f_{12}$ is given by

$$
\begin{aligned}
\mathrm{V}(f_{12}) &= \left\langle f_{12} - \mathrm{E}(f_{12}), f_{12} - \mathrm{E}(f_{12}) \right\rangle \\
&= \left\langle f_{12}, f_{12} \right\rangle \ \text{ from } \ \mathrm{E}(f_{12}) = 0; \ \text{see Eq. (B.2.16)} \\
&= \left\langle \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0 - f_1 - f_2, \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0 - f_1 - f_2 \right\rangle \\
&\quad \text{from Eq. (B.2.6)} \\
&= \left\langle \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0, \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0 \right\rangle \\
&\quad - 2\left\langle \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0, f_1 + f_2 \right\rangle + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle \\
&= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - 2\left\langle f_{12} + f_1 + f_2, f_1 + f_2 \right\rangle + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle \\
&\quad \text{from } \ \mathrm{E}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} = f_0, \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] - f_0 = f_{12} + f_1 + f_2 \\
&= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - 2\{\left\langle f_{12}, f_1 + f_2 \right\rangle + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle\} + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle \\
&= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - 2\{0 + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle\} + \left\langle f_1 + f_2, f_1 + f_2 \right\rangle \\
&= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - \left\langle f_1 + f_2, f_1 + f_2 \right\rangle \\
&= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - \mathrm{V}(f_1) - \mathrm{V}(f_2).
\end{aligned}
\tag{B.2.38}
$$

Further, continuation from the last equality in the preceding equation yields

$$
\begin{aligned}
\mathrm{V}(f_{12}) &= \int_{\mathcal{X}} \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]^2 \, d(\mathbf{x}) \mathrm{d}\mathbf{x} - 2\int_{\mathcal{X}} f_0 \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2] d(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{\mathcal{X}} f_0^2 d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\quad - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&= \int_{\mathcal{X}} \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]^2 \, d(\mathbf{x}) \mathrm{d}\mathbf{x} - 2f_0^2 + f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&= \int_{\mathcal{X}} \mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]^2 \, d(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \left[ \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 \right]^2 d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2).
\end{aligned}
\tag{B.2.39}
$$

In summary,

$$
\begin{aligned}
\mathrm{V}(f_{12}) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \left[ \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 \right]^2 d_2(x_2) d_1(x_1) \mathrm{d}x_2 \mathrm{d}x_1 - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2).
\end{aligned}
\tag{B.2.40}
$$

Similar derivations for $\mathrm{V}(f_{13})$ and $\mathrm{V}(f_{23})$ produce the representations

$$\begin{aligned} \mathrm{V}(f_{13}) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_3]\} - \mathrm{V}(f_1) - \mathrm{V}(f_3) \\ &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_3} \left[ \int_{\mathcal{X}_2} f(x_1, x_2, x_3) d_2(x_2) \mathrm{d}x_2 \right]^2 d_3(x_3) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_1 - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_3) \end{aligned} \tag{B.2.41}$$

and

$$\begin{aligned} \mathrm{V}(f_{23}) &= \mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_2, x_3]\} - \mathrm{V}(f_2) - \mathrm{V}(f_3) \\ &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} \left[ \int_{\mathcal{X}_1} f(x_1, x_2, x_3) d_1(x_1) \mathrm{d}x_1 \right]^2 d_2(x_2) d_3(x_3) \mathrm{d}x_3 \mathrm{d}x_2 - f_0^2 - \mathrm{V}(f_2) - \mathrm{V}(f_3). \end{aligned} \tag{B.2.42}$$

In addition, a rearrangement of the first equality in Eqs. (B.2.40)-(B.2.42) gives the representations

$$\mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_2]\} = \mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_{12}) \tag{B.2.43}$$

$$\mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_1, x_3]\} = \mathrm{V}(f_1) + \mathrm{V}(f_3) + \mathrm{V}(f_{13}) \tag{B.2.44}$$

and

$$\mathrm{V}\{\mathrm{E}[f(\mathbf{x}) \mid x_2, x_3]\} = \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{23}). \tag{B.2.45}$$

The variance $\mathrm{V}(f_{123})$ of $f_{123}$ is given by

$$
\begin{aligned}
\mathrm{V}(f_{123}) &= \langle f_{123} - \mathrm{E}(f_{123}), f_{123} - \mathrm{E}(f_{123}) \rangle \\
&= \langle f_{123}, f_{123} \rangle \quad \text{from } \mathrm{E}(f_{123}) = 0; \text{ see Eq. (B.2.17)} \\
&= \langle f - f_0 - f_1 - f_2 - f_3 - f_{12} - f_{13} - f_{23}, f - f_0 - f_1 - f_2 - f_3 - f_{12} - f_{13} - f_{23} \rangle \\
&\quad \text{from Eq. (B.2.9)} \\
&= \langle f - f_0, f - f_0 \rangle - 2\langle f - f_0, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&\quad + \langle f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&= \mathrm{V}(f) - 2\langle f_{123} + f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&\quad + \langle f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&= \mathrm{V}(f) - 2\langle f_{123}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&\quad - 2\langle f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&\quad + \langle f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&= \mathrm{V}(f) - 2\langle f_{123}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&\quad - \langle f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23}, f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} \rangle \\
&= \mathrm{V}(f) - 2 \times 0 - \{\langle f_1, f_1 \rangle + 0 + \langle f_2, f_2 \rangle + 0 + \langle f_3, f_3 \rangle + 0 + \langle f_{12}, f_{12} \rangle \\
&\quad + 0 + \langle f_{13}, f_{13} \rangle + 0 + \langle f_{23}, f_{23} \rangle\} \\
&= \mathrm{V}(f) - \{\mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23})\}.
\end{aligned}
\tag{B.2.46}
$$

Thus, $\mathrm{V}(f_{123})$ is the part of V($f$) that cannot be accounted for by $\mathrm{V}(f_1), \mathrm{V}(f_2), ..., \mathrm{V}(f_{23})$.

Given the preceding representation in Eq. (B.2.46) for $\mathrm{V}(f_{123})$, the sensitivity measure $S_{1T}$ defined in Eq. (B.2.32) can also be represented by

$$
\begin{aligned}
S_{1T} &= [\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{123})] / \mathrm{V}(f) \\
&= \{[\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13})] \\
&\qquad + [\mathrm{V}(f) - \{\mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23})\}]\} / \mathrm{V}(f) \\
&= [\mathrm{V}(f) - \mathrm{V}(f_2) - \mathrm{V}(f_3) - \mathrm{V}(f_{23})] / \mathrm{V}(f),
\end{aligned}
\tag{B.2.47}
$$

which clearly indicates the removal of isolated effects associated with $x_2$ and $x_3$. Similarly, $S_{2T}$ and $S_{3T}$ have the representations

$$
S_{2T} = [\mathrm{V}(f) - \mathrm{V}(f_1) - \mathrm{V}(f_3) - \mathrm{V}(f_{13})] / \mathrm{V}(f) \tag{B.2.48}
$$

and

$$
S_{3T} = [\mathrm{V}(f) - \mathrm{V}(f_1) - \mathrm{V}(f_2) - \mathrm{V}(f_{12})] / \mathrm{V}(f). \tag{B.2.49}
$$

## B.3 Approximation of Variance Decomposition for $f(x_1, x_2, x_3)$

In practice, the variances $\mathrm{V}(f)$, $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$, …, $\mathrm{V}(f_{123})$ defined in Sect. A.2 must be evaluated numerically. Due to the complexity of the function *f* in most real analyses, analytic evaluation of the integrals that define these variances is unlikely to be practicable. Rather, some type of numerical procedure will be required. Sampling-based procedures are likely candidates for these evaluations and are now described.

The evaluation of $f_0 = \mathrm{E}(f)$ and V(*f*) with sampling-based procedures is straightforward. A random or Latin hypercube sample

$$\mathbf{x}_s = [x_{1s}, x_{2s}, x_{3s}], s = 1, 2, ..., nS, \tag{B.3.1}$$

is generated from $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ in consistency with the probability spaces $(\mathcal{X}_1, \mathbb{X}_1, m_1)$, $(\mathcal{X}_2, \mathbb{X}_2, m_2)$ and $(\mathcal{X}_3, \mathbb{X}_3, m_3)$ and the associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$. Then, the following approximations result:

$$\hat{f}_0 = \hat{\mathrm{E}}(f) = \sum_{s=1}^{nS} f(\mathbf{x}_s) / nS \cong \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{B.3.2}$$

and

$$\begin{aligned} \hat{\mathrm{V}}(f) &= \sum_{s=1}^{nS} \left[ f(\mathbf{x}_s) - \hat{f}_0 \right]^2 / nS \\ &= \sum_{s=1}^{nS} f^2(\mathbf{x}_s) / nS - \hat{f}_0^2 \\ &\cong \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0^2, \end{aligned} \tag{B.3.3}$$

with the "hat" symbol used to represent an estimated quantity. When random sampling is employed, use of $nS - 1$ rather than *nS* in Eq, (B.3.3) will result in an unbiased estimate for V(*f*).

The variances $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ can also be estimated with sampling-based procedures. A popular approach is based on a reformulation of the integral representations for $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ in Eqs. (B.2.35)-(B.2.37). For $\mathrm{V}(f_1)$, this reformulation is given by

$$
\begin{aligned}
\mathrm{V}(f_1) &=_1 \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2\right]^2 d_1(x_1)\mathrm{d}x_1 - f_0^2 \\
&=_2 \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_{23}} f(x_1,x_2,x_3)d(\mathbf{x}_{23})\mathrm{d}\mathbf{x}_{23}\right]\left[\int_{\mathcal{X}_{23}} f(x_1,x_2,x_3)d(\mathbf{x}_{23})\mathrm{d}\mathbf{x}_{23}\right] d_1(x_1)\mathrm{d}x_1 - f_0^2 \\
&=_3 \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_{23}} f(x_1,x_2,x_3)d(\mathbf{x}_{23})\mathrm{d}\mathbf{x}_{23}\right]\left[\int_{\tilde{\mathcal{X}}_{23}} f(x_1,\tilde{x}_2,\tilde{x}_3)d(\tilde{\mathbf{x}}_{23})\mathrm{d}\tilde{\mathbf{x}}_{23}\right] d_1(x_1)\mathrm{d}x_1 - f_0^2 \qquad \text{(B.3.4)} \\
&=_4 \int_{\mathcal{X}_1}\int_{\mathcal{X}_{23}}\int_{\tilde{\mathcal{X}}_{23}} f(x_1,x_2,x_3)f(x_1,\tilde{x}_2,\tilde{x}_3)d(\tilde{\mathbf{x}}_{23})d(\mathbf{x}_{23})d_1(x_1)\mathrm{d}\tilde{\mathbf{x}}_{23}\mathrm{d}\mathbf{x}_{23}\mathrm{d}_1x_1 - f_0^2 \\
&=_5 \int_{\mathcal{X}}\int_{\tilde{\mathcal{X}}_{23}} f(x_1,x_2,x_3)f(x_1,\tilde{x}_2,\tilde{x}_3)d(\tilde{\mathbf{x}}_{23})d(\mathbf{x})\mathrm{d}\tilde{\mathbf{x}}_{23}\mathrm{d}\mathbf{x} - f_0^2,
\end{aligned}
$$

where (i) Equality 1 is a repeat of Eq. (B.2.35), (ii) Equality 2 is a notational restatement of Equality 1 with $\mathcal{X}_{23} = \mathcal{X}_2 \times \mathcal{X}_3$, $\mathbf{x}_{23} = [x_2, x_3]$, $d(\mathbf{x}_{23}) = d_2(x_2)d_3(x_3)$, $\mathrm{d}\mathbf{x}_{23} = \mathrm{d}x_2\mathrm{d}x_3$, and $[\sim]^2$ replaced by $[\sim]\times[\sim]$, (iii) Equality 3 involves the additional notational additions $\tilde{\mathcal{X}}_{23} = \mathcal{X}_2 \times \mathcal{X}_3$, $\tilde{\mathbf{x}}_{23} = [\tilde{x}_2, \tilde{x}_3]$, $d(\tilde{\mathbf{x}}_{23}) = d_2(\tilde{x}_2)d_3(\tilde{x}_3)$ and $\mathrm{d}\tilde{\mathbf{x}}_{23} = \mathrm{d}\tilde{x}_2\mathrm{d}\tilde{x}_3$ to emphasize the that the variable pairs $(x_2, x_3)$ in the two inner double integrals in Equality 2 can be viewed as pairs of independent, identically distributed variables (i.e., $\mathcal{X}_{23} = \tilde{\mathcal{X}}_{23} = \mathcal{X}_2 \times \mathcal{X}_3$; $x_2$ and $\tilde{x}_2$ have the same density function; and similarly, $x_3$ and $\tilde{x}_3$ have the same density function ), and (iv) Equality 4 results from combining the two multiplied double integrals over $\mathcal{X}_{23}$ and $\tilde{\mathcal{X}}_{23}$ with the outer integral over $\mathcal{X}_1$ to produce a 5-dimensional integral of the function

$$
F_1(x_1, \mathbf{x}_{23}, \tilde{\mathbf{x}}_{23}) = f(x_1,x_2,x_3)f(x_1,\tilde{x}_2,\tilde{x}_3) \qquad \text{(B.3.5)}
$$

over the subset $\mathcal{X}_1 \times \mathcal{X}_{23} \times \tilde{\mathcal{X}}_{23} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{X}_2 \times \mathcal{X}_3$ of $R^5$, and (v) Equality 5 is the result of combining the integrals over $\mathcal{X}_1$ and $\mathcal{X}_{23}$ into an integral over $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$.

For $\mathrm{V}(f_2)$, the equivalent reformulation is

$$
\mathrm{V}(f_2) = \int_{\mathcal{X}_2}\int_{\mathcal{X}_{13}}\int_{\tilde{\mathcal{X}}_{13}} f(x_1,x_2,x_3)f(\tilde{x}_1,x_2,\tilde{x}_3)d(\tilde{\mathbf{x}}_{13})d(\mathbf{x}_{13})d_2(x_2)\mathrm{d}\tilde{\mathbf{x}}_{13}\mathrm{d}\mathbf{x}_{13}\mathrm{d}x_2 - f_0^2 \qquad \text{(B.3.6)}
$$

with $\mathcal{X}_{13} = \tilde{\mathcal{X}}_{13} = \mathcal{X}_1 \times \mathcal{X}_3$, $\mathbf{x}_{13} = [x_1, x_3]$, $\tilde{\mathbf{x}}_{13} = [\tilde{x}_1, \tilde{x}_3]$, $d(\mathbf{x}_{13}) = d_1(x_1)d_3(x_3)$, $d(\tilde{\mathbf{x}}_{13}) = d_1(\tilde{x}_1)d_3(\tilde{x}_3)$, $\mathrm{d}\mathbf{x}_{13} = \mathrm{d}x_1\mathrm{d}x_3$ and $\mathrm{d}\tilde{\mathbf{x}}_{13} = \mathrm{d}\tilde{x}_1\mathrm{d}\tilde{x}_3$. The preceding reformulation results in $\mathrm{V}(f_2)$ being expressed in terms of an integral of function

$$F_2(x_2, \mathbf{x}_{13}, \tilde{\mathbf{x}}_{13}) = f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, \tilde{x}_3) \tag{B.3.7}$$

over the subset $\mathcal{X}_2 \times \mathcal{X}_{13} \times \tilde{\mathcal{X}}_{13} = \mathcal{X}_2 \times \mathcal{X}_1 \times \mathcal{X}_3 \times \mathcal{X}_1 \times \mathcal{X}_3$ of $R^5$. Similarly,

$$\mathrm{V}(f_3) = \int_{\mathcal{X}_3} \int_{\mathcal{X}_{12}} \int_{\tilde{\mathcal{X}}_{12}} f(x_1, x_2, x_3) f(\tilde{x}_1, \tilde{x}_2, x_3) d(\tilde{\mathbf{x}}_{12}) d(\mathbf{x}_{12}) d_3(x_3) \mathrm{d}\tilde{\mathbf{x}}_{12} \mathrm{d}\mathbf{x}_{12} \mathrm{d}x_3 - f_0^2 \tag{B.3.8}$$

with $\mathcal{X}_{12} = \tilde{\mathcal{X}}_{12} = \mathcal{X}_1 \times \mathcal{X}_2$, $\mathbf{x}_{12} = [x_1, x_2]$, $\tilde{\mathbf{x}}_{12} = [\tilde{x}_1, \tilde{x}_2]$, $d(\mathbf{x}_{12}) = d_1(x_1) d_2(x_2)$, $d(\tilde{\mathbf{x}}_{12}) = d_1(\tilde{x}_1) d_2(\tilde{x}_2)$, $\mathrm{d}\mathbf{x}_{12} = \mathrm{d}x_1 \mathrm{d}x_2$ and $\mathrm{d}\tilde{\mathbf{x}}_{12} = \mathrm{d}\tilde{x}_1 \mathrm{d}\tilde{x}_2$. The preceding reformulation results in $\mathrm{V}(f_3)$ being expressed in terms of an integral of the function

$$F_3(x_3, \mathbf{x}_{12}, \tilde{\mathbf{x}}_{12}) = f(x_1, x_2, x_3) f(\tilde{x}_1, \tilde{x}_2, x_3) \tag{B.3.9}$$

over the subset $\mathcal{X}_3 \times \mathcal{X}_{12} \times \tilde{\mathcal{X}}_{12} = \mathcal{X}_3 \times \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_1 \times \mathcal{X}_2$ of $R^5$.

Given the representations in Eqs. (B.3.4), (B.3.6) and (B.3.8), the variances $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ can be estimated with use of a random or Latin hypercube sample

$$[\mathbf{x}_s, \tilde{\mathbf{x}}_s] = [x_{1s}, x_{2s}, x_{3s}, \tilde{x}_{1s}, \tilde{x}_{2s}, \tilde{x}_{3s}], s = 1, 2, ..., nS, \tag{B.3.10}$$

from $\mathcal{X} \times \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 = \mathcal{X} \times \tilde{\mathcal{X}}$ generated in consistency with use of the density functions $d_1$, $d_2$ and $d_3$ for the corresponding elements of $\mathbf{x} = [x_1, x_2, x_3]$ and $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \tilde{x}_3]$. In turn,

$$\hat{\mathrm{V}}(f_1) \cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(x_{1s}, \tilde{x}_{2s}, \tilde{x}_{3s}) / nS - \hat{f}_0^2 = \hat{I}_1 - \hat{f}_0^2, \tag{B.3.11}$$

$$\hat{\mathrm{V}}(f_2) \cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{1s}, x_{2s}, \tilde{x}_{3s}) / nS - \hat{f}_0^2 = \hat{I}_2 - \hat{f}_0^2, \tag{B.3.12}$$

and

$$\hat{\mathrm{V}}(f_3) \cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{1s}, \tilde{x}_{2s}, x_{3s}) / nS - \hat{f}_0^2 = \hat{I}_3 - \hat{f}_0^2, \tag{B.3.13}$$

with (i) the approximation $\hat{f}_0$ to $f_0 = \mathrm{E}(f)$ defined in Eq. (B.3.2) and (ii) $\hat{I}_1$, $\hat{I}_2$, $\hat{I}_3$ introduced as compact representations for the associated integral approximations.

The sampling procedure indicated in Eqs. (B.3.10)-(B.3.13) requires $nS$ evaluations of each of the functions $f(x_{1s}, x_{2s}, x_{3s})$, $f(x_{1s}, \tilde{x}_{2s}, \tilde{x}_{3s})$, $f(\tilde{x}_{1s}, x_{2s}, \tilde{x}_{3s})$ and $f(\tilde{x}_{1s}, x_{2s}, \tilde{x}_{3s})$ for a total of 4 $nS$ evaluations of the function $f$. This number of functions evaluations for $\mathbf{x} = [x_1, x_2, x_3]$ can be cut in half by making efficient use of the sample indicated in Eq. (B.3.1). Specifically, a second sample

$$\tilde{\mathbf{x}}_s = [\tilde{x}_{1s}, \tilde{x}_{2s}, \tilde{x}_{3s}], s = 1, 2, ..., nS, \tag{B.3.14}$$

can be obtained by randomly rearranging the sampled values for each variable. For example, the sequence of values $\tilde{x}_{11}, \tilde{x}_{12}, ..., \tilde{x}_{1,nS}$ for $x_1$ in the sample in Eq. (B.3.14) is simply a random rearrangement of the sequence of values $x_{11}, x_{12}, ..., x_{1,nS}$ for $x_1$ in the sample in Eq. (B.3.1). Then, the sampled values $\tilde{\mathbf{x}}_s$ are reordered to produce 3 new samples:

$$\begin{aligned} S_1 &: \tilde{\mathbf{x}}_{1s} = [\tilde{x}_{11s}, \tilde{x}_{12s}, \tilde{x}_{13s}], s = 1, 2, ..., nS, \text{ with } \tilde{x}_{11s} = x_{1s} \\ S_2 &: \tilde{\mathbf{x}}_{2s} = [\tilde{x}_{21s}, \tilde{x}_{22s}, \tilde{x}_{23s}], s = 1, 2, ..., nS, \text{ with } \tilde{x}_{22s} = x_{2s} \\ S_3 &: \tilde{\mathbf{x}}_{3s} = [\tilde{x}_{31s}, \tilde{x}_{32s}, \tilde{x}_{33s}], s = 1, 2, ..., nS, \text{ with } \tilde{x}_{33s} = x_{3s}. \end{aligned} \tag{B.3.15}$$

Then, the sample in Eq. (B.3.1) and the samples $S_1$, $S_2$ and $S_3$ are used in place of the sample in Eq. (B.3.10) in the approximation of $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$:

$$\begin{aligned} \hat{\mathrm{V}}(f_1) &\cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{11s}, \tilde{x}_{12s}, \tilde{x}_{13s}) / nS - \hat{f}_0^2 = \hat{I}_1 - \hat{f}_0^2, \\ &= \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(x_{1s}, \tilde{x}_{12s}, \tilde{x}_{13s}) / nS - \hat{f}_0^2 = \hat{I}_1 - \hat{f}_0^2, \end{aligned} \tag{B.3.16}$$

$$\begin{aligned} \hat{\mathrm{V}}(f_2) &\cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{21s}, \tilde{x}_{22s}, \tilde{x}_{23s}) / nS - \hat{f}_0^2 = \hat{I}_2 - \hat{f}_0^2 \\ &= \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{21s}, x_{2s}, \tilde{x}_{23s}) / nS - \hat{f}_0^2 = \hat{I}_2 - \hat{f}_0^2 \end{aligned} \tag{B.3.17}$$

and

$$\begin{aligned} \hat{\mathrm{V}}(f_3) &\cong \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{31s}, \tilde{x}_{32s}, \tilde{x}_{33s}) / nS - \hat{f}_0^2 = \hat{I}_3 - \hat{f}_0^2 \\ &= \sum_{s=1}^{nS} f(x_{1s}, x_{2s}, x_{3s}) f(\tilde{x}_{31s}, \tilde{x}_{32s}, x_{3s}) / nS - \hat{f}_0^2 = \hat{I}_3 - \hat{f}_0^2. \end{aligned} \tag{B.3.18}$$

Use of the preceding sampling procedure in the approximation of $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ reduces the number of required evaluations of $f(x_1, x_2, x_3)$ from 4 *nS* to 2 *nS*. More significantly, if *n* uncertain variables are under consideration and approximations for $\mathrm{V}(f_i), i = 1, 2, ..., n,$ are to be determined, use of the preceding sampling procedure reduces the number of required function evaluations from $(n+1)nS$ to 2 *nS*.

The variances $\mathrm{V}(f_{12}), \mathrm{V}(f_{13})$ and $\mathrm{V}(f_{23})$ can also be estimated with sampling-based procedures in a manner similar to that shown for $\mathrm{V}(f_1)$, $\mathrm{V}(f_2)$ and $\mathrm{V}(f_3)$ in Eqs. (B.3.11)-(B.3.13). Specifically, reformulation of the final integral in Eq. (B.2.40) for $\mathrm{V}(f_{12})$ yields

$$
\begin{aligned}
\mathrm{V}(f_{12}) &=_1 \int_{\mathcal{X}_{12}} \int_{\mathcal{X}_3} \int_{\tilde{\mathcal{X}}_3} f(x_1, x_2, x_3) f(x_1, x_2, \tilde{x}_3) d_3(\tilde{x}_3) d_3(x_3) d(\mathbf{x}_{12}) \mathrm{d}\tilde{x}_3 \mathrm{d}x_3 \mathrm{d}\mathbf{x}_{12} \\
&\quad - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&\cong_2 \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(x_{1r}, x_{2r}, \tilde{x}_{3r}) / nS - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_2) \\
&\cong_3 \hat{I}_{12} - \hat{f}_0^2 - \hat{\mathrm{V}}(f_1) - \hat{\mathrm{V}}(f_2) \\
&\cong_4 \hat{I}_{12} - \hat{f}_0^2 - \left[\hat{I}_1 - \hat{f}_0^2\right] - \left[\hat{I}_2 - \hat{f}_0^2\right] \\
&=_5 \hat{I}_{12} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_2 \\
&=_6 \hat{\mathrm{V}}(f_{12}),
\end{aligned}
\tag{B.3.19}
$$

where (i) equality 1 uses the notations $\mathcal{X}_{12} = \mathcal{X}_1 \times \mathcal{X}_2$, $\mathbf{x}_{12} = [x_1, x_2]$, $d(\mathbf{x}_{12}) = d_1(x_1) d_2(x_2)$, $\mathrm{d}\mathbf{x}_{12} = \mathrm{d}x_1 \mathrm{d}x_2$ and is obtained by manipulations similar to those used in Eq. (B.3.4), (ii) approximate equality 2 for the integral in equality 1 is obtained with use of the sample in Eq. (B.3.10), (iii) equality 3 introduces the notation $\hat{I}_{12}$ for the integral approximation in equality 2 and also symbolic representations of the approximations for $f_0$, $\mathrm{V}(f_1)$ and $\mathrm{V}(f_2)$ defined in Eqs. (B.3.2), (B.3.11) and (B.3.12), (iv) equality 4 is obtained from equality 3 by replacing $\hat{\mathrm{V}}(f_1)$ and $\hat{\mathrm{V}}(f_2)$ by their representations in Eqs. (B.3.11) and (B.3.12), (v) equality 5 is obtained from equality 4 by simple algebra, and (vi) equality 6 introduces the notation $\hat{\mathrm{V}}(f_{12})$ for the derived sampling-based approximation for $\mathrm{V}(f_{12})$.

Derivations similar to the preceding derivation for $\hat{\mathrm{V}}(f_{12})$ produce the following approximations $\hat{\mathrm{V}}(f_{13})$ and $\hat{\mathrm{V}}(f_{23})$ for $\mathrm{V}(f_{13})$ and $\mathrm{V}(f_{23})$:

$$
\begin{aligned}
\mathrm{V}(f_{13}) =_1 & \int_{\mathcal{X}_{13}} \int_{\mathcal{X}_2} \int_{\tilde{\mathcal{X}}_2} f(x_1, x_2, x_3) f(x_1, \tilde{x}_2, x_3) d_2(\tilde{x}_2) d_2(x_2) d(\mathbf{x}_{13}) \mathrm{d}\tilde{x}_2 \mathrm{d}x_2 \mathrm{d}\mathbf{x}_{13} \\
& - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_3) \\
\cong_2 & \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(x_{1r}, \tilde{x}_{2r}, x_{3r}) / nS - f_0^2 - \mathrm{V}(f_1) - \mathrm{V}(f_3) \\
\cong_3 & \hat{I}_{13} - \hat{f}_0^2 - \hat{\mathrm{V}}(f_1) - \hat{\mathrm{V}}(f_3) \\
=_4 & \hat{I}_{13} - \hat{f}_0^2 - \left[\hat{I}_1 - \hat{f}_0^2\right] - \left[\hat{I}_3 - \hat{f}_0^2\right] \\
=_5 & \hat{I}_{13} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_3 \\
=_6 & \hat{\mathrm{V}}(f_{13})
\end{aligned}
\tag{B.3.20}
$$

and

$$
\begin{aligned}
\mathrm{V}(f_{23}) =_1 & \int_{\mathcal{X}_{23}} \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, x_3) d_1(\tilde{x}_1) d_1(x_1) d(\mathbf{x}_{23}) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1 \mathrm{d}\mathbf{x}_{23} \\
& - f_0^2 - \mathrm{V}(f_2) - \mathrm{V}(f_3) \\
\cong_2 & \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(\tilde{x}_{1r}, x_{2r}, x_{3r}) / nS - f_0^2 - \mathrm{V}(f_2) - \mathrm{V}(f_3) \\
\cong_3 & \hat{I}_{23} - \hat{f}_0^2 - \hat{\mathrm{V}}(f_2) - \hat{\mathrm{V}}(f_3) \\
=_4 & \hat{I}_{23} - \hat{f}_0^2 - \left[\hat{I}_2 - \hat{f}_0^2\right] - \left[\hat{I}_3 - \hat{f}_0^2\right] \\
=_5 & \hat{I}_{23} + \hat{f}_0^2 - \hat{I}_2 - \hat{I}_3 \\
=_6 & \hat{\mathrm{V}}(f_{23}),
\end{aligned}
\tag{B.3.21}
$$

with the numbered equalities in Eqs. (B.3.20) and (B.3.21) matching the numbered equalities in Eq. (B.3.19). The explanations for the equalities in Eq. (B.3.19) are also applicable to the equalities in Eqs. (B.3.20) and (B.3.21).

Sampling-based approximation of $\mathrm{V}(f_{123})$ is now considered. First, for consistency with the notational use of $\hat{I}_1$, $\hat{I}_2$,…, $\hat{I}_{23}$, it is noted that the approximation $\hat{\mathrm{V}}(f)$ in Eq. (B.3.3) can be expressed as

$$
\hat{\mathrm{V}}(f) = \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(x_{1r}, x_{2r}, x_{3r}) / nS - \hat{f}_0^2 = \hat{I}_{123} - \hat{f}_0^2 . \tag{B.3.22}
$$

In turn,

$$
\begin{aligned}
\mathrm{V}(f_{123}) &= \mathrm{V}(f) - \{\mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23})\} \\
&\cong \hat{\mathrm{V}}(f) - \{\hat{\mathrm{V}}(f_1) + \hat{\mathrm{V}}(f_2) + \hat{\mathrm{V}}(f_3) + \hat{\mathrm{V}}(f_{12}) + \hat{\mathrm{V}}(f_{13}) + \hat{\mathrm{V}}(f_{23})\} \\
&= \left\{\hat{I}_{123} - \hat{f}_0^2 .\right\} - \left\{\left[\hat{I}_1 - \hat{f}_0^2\right] + \left[\hat{I}_2 - \hat{f}_0^2\right] + \left[\hat{I}_3 - \hat{f}_0^2\right]\right. \\
&\qquad \left. + \left[\hat{I}_{12} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_2\right] + \left[\hat{I}_{13} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_3\right] + \left[\hat{I}_{23} + \hat{f}_0^2 - \hat{I}_2 - \hat{I}_3\right]\right\} \\
&= \left\{\hat{I}_{123} - \hat{f}_0^2\right\} + \left\{-\hat{I}_1 + \hat{f}_0^2 - \hat{I}_2 + \hat{f}_0^2 - \hat{I}_3 + \hat{f}_0^2\right. \\
&\qquad \left. - \hat{I}_{12} - \hat{f}_0^2 + \hat{I}_1 + \hat{I}_2 - \hat{I}_{13} - \hat{f}_0^2 + \hat{I}_1 + \hat{I}_3 - \hat{I}_{23} - \hat{f}_0^2 + \hat{I}_2 + \hat{I}_3\right\} \\
&= \hat{I}_{123} + \hat{I}_1 + \hat{I}_2 + \hat{I}_3 - \hat{I}_{12} - \hat{I}_{13} - \hat{I}_{23} - \hat{f}_0^2 \\
&= \hat{\mathrm{V}}(f_{123}).
\end{aligned}
\tag{B.3.23}
$$

Once the variance approximations $\hat{\mathrm{V}}(f)$, $\hat{\mathrm{V}}(f_1)$, $\hat{\mathrm{V}}(f_2)$, …, $\hat{\mathrm{V}}(f_{123})$ are available, they can be used to obtain approximations to the variance decomposition sensitivity measures defined in conjunction with Eqs. (B.2.28)-(B.2.32). For the sensitivity measure $S_{1T}$ defined in Eq. (B.2.32), the resulting approximation is

$$
\begin{aligned}
\hat{S}_{1T} &= \left[\hat{\mathrm{V}}(f_1) + \hat{\mathrm{V}}(f_{12}) + \hat{\mathrm{V}}(f_{13}) + \hat{\mathrm{V}}(f_{123})\right] / \hat{\mathrm{V}}(f) \\
&= \left\{\left[\hat{I}_1 - \hat{f}_0^2\right] + \left[\hat{I}_{12} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_2\right] + \left[\hat{I}_{13} + \hat{f}_0^2 - \hat{I}_1 - \hat{I}_3\right]\right. \\
&\qquad \left. + \left[\hat{I}_{123} + \hat{I}_1 + \hat{I}_2 + \hat{I}_3 - \hat{I}_{12} - \hat{I}_{13} - \hat{I}_{23} - \hat{f}_0^2\right]\right\} \left\{\hat{I}_{123} - \hat{f}_0^2\right\}^{-1} \\
&= \left\{\hat{I}_{123} - \hat{I}_{23}\right\} \left\{\hat{I}_{123} - \hat{f}_0^2\right\}^{-1}.
\end{aligned}
\tag{B.3.24}
$$

Similarly, the sensitivity measures $S_{2T}$ and $S_{3T}$ defined in conjunction with Eq. (B.2.32) can be approximated by

$$
\hat{S}_{2T} = \left\{\hat{I}_{123} - \hat{I}_{13}\right\} \left\{\hat{I}_{123} - \hat{f}_0^2\right\}^{-1} \quad \text{and} \quad \hat{S}_{3T} = \left\{\hat{I}_{123} - \hat{I}_{12}\right\} \left\{\hat{I}_{123} - \hat{f}_0^2\right\}^{-1}. \tag{B.3.25}
$$

The results leading to the approximations for $S_{1T}, S_{2T}$ and $S_{3T}$ in Eqs. (B.3.24) and (B.3.25) also provide a basis for the following closed from representations and approximations for $S_{1T}$, $S_{2T}$ and $S_{3T}$. The representation for $S_{1T}$ is given by

$$
\begin{aligned}
S_{1T} &=_1 [\mathrm{V}(f) - \mathrm{V}(f_2) - \mathrm{V}(f_3) - \mathrm{V}(f_{23})] / \mathrm{V}(f) \\
&=_2 \{I_{123} - I_{23}\} / \mathrm{V}(f) \\
&=_3 \left\{ \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f^2(x_1, x_2, x_3)\, d(\mathbf{x}) \mathrm{d}\mathbf{x} \right. \\
&\qquad \left. - \int_{\mathcal{X}_{23}} \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, x_3) d_1(\tilde{x}_1) d_1(x_1) d(\mathbf{x}_{23}) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1 \mathrm{d}\mathbf{x}_{23} \right\} / V(f) \\
&=_4 \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} \int_{\tilde{\mathcal{X}}_1} \left\{ f^2(x_1, x_2, x_3) - f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, x_3) \right\} d_1(\tilde{x}_1) d(\mathbf{x}) \mathrm{d}\tilde{x}_1 \mathrm{d}\mathbf{x} / V(f) \\
&\cong_5 \left\{ \sum_{r=1}^{nS} \left\{ f^2(x_{1r}, x_{2r}, x_{3r}) - f(x_{1r}, x_{2r}, x_{3r}) f(\tilde{x}_{1r}, x_{2r}, x_{3r}) \right\} / nS \right\} / V(f) \\
&=_6 \left\{ \sum_{r=1}^{nS} f^2(x_{1r}, x_{2r}, x_{3r}) / nS - \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(\tilde{x}_{1r}, x_{2r}, x_{3r}) / nS \right\} / V(f) \\
&=_7 \{\hat{I}_{123} - \hat{I}_{23}\} / \mathrm{V}(f),
\end{aligned}
\tag{B.3.26}
$$

where (i) equality 1 is from Eq. (B.2.47), (ii) the expressions $I_{123}$ and $I_{23}$ in equality 2 correspond to the integrals approximated by $\hat{I}_{123}$ and $\hat{I}_{23}$, (iii) the two integrals in equality 3 are obtained from Eqs. (B.3.3) and (B.3.21) and are the integrals approximated by $\hat{I}_{123}$ and $\hat{I}_{23}$, (iv) equality 4 is the result of combining the two integrals in equality 3 into a single integral, (v) approximant equality 5 results from use of the sample in Eq. (B.3.10), (vi) equality 6 is the result of expressing the single sum in equality 5 as two sums, and (vii) equality 7 results from the definitions of $\hat{I}_{123}$ and $\hat{I}_{23}$ in conjunction with Eqs. (B.3.22) and (B.3.21).

Similarly, the representations for $S_{2T}$ and $S_{3T}$ are

$$
\begin{aligned}
S_{2T} &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} \int_{\tilde{\mathcal{X}}_2} \left\{ f^2(x_1, x_2, x_3) - f(x_1, x_2, x_3) f(x_1, \tilde{x}_2, x_3) \right\} d_2(\tilde{x}_2) d(\mathbf{x}) \mathrm{d}\tilde{x}_2 \mathrm{d}\mathbf{x} / V(f) \\
&\cong \left\{ \sum_{r=1}^{nS} f^2(x_{1r}, x_{2r}, x_{3r}) / nS - \sum_{r=1}^{nS} f(x_{1r}, x_{2r}, x_{3r}) f(x_{1r}, \tilde{x}_{2r}, x_{3r}) / nS \right\} / V(f) \\
&= \{\hat{I}_{123} - \hat{I}_{13}\} / \mathrm{V}(f)
\end{aligned}
\tag{B.3.27}
$$

and

$$
\begin{aligned}
S_{3T} &= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3}\int_{\tilde{\mathcal{X}}_3}\left\{f^2(x_1,x_2,x_3)-f(x_1,x_2,x_3)f(x_1,x_2,\tilde{x}_3)\right\}d_3(\tilde{x}_3)d(\mathbf{x})\mathrm{d}\tilde{x}_3\mathrm{d}\mathbf{x}\,/\,V(f)\\
&\cong\left\{\sum_{r=1}^{nS}f^2(x_{1r},x_{2r},x_{3r})/nS-\sum_{r=1}^{nS}f(x_{1r},x_{2r},x_{3r})f(x_{1r},x_{2r},\tilde{x}_{3r})/nS\right\}/V(f) \qquad \text{(B.3.28)}\\
&=\{\hat{I}_{123}-\hat{I}_{12}\}/\mathrm{V}(f).
\end{aligned}
$$

The sensitivity measures $S_{1T}$, $S_{2T}$ and $S_{3T}$ are often defined as

$$
\begin{aligned}
S_{iT} &=_1 \mathrm{E}_{\mathcal{X}_{\sim i}}\left[\mathrm{V}_{\mathcal{X}_i}(f\,|\,\mathbf{x}_{\sim i})\right]/\mathrm{V}(f)\\
&=_2 \left\{\mathrm{V}(f)-\mathrm{V}_{\mathcal{X}_{\sim i}}\left[\mathrm{E}_{\mathcal{X}_i}(f\,|\,\mathbf{x}_{\sim i})\right]\right\}/\mathrm{V}(f) \qquad \text{(B.3.29)}\\
&=_3 1-\mathrm{V}_{\mathcal{X}_{\sim i}}\left[\mathrm{E}_{\mathcal{X}_i}(f\,|\,\mathbf{x}_{\sim i})\right]/\mathrm{V}(f),
\end{aligned}
$$

with

$$
\mathcal{X}_{\sim i}=\begin{cases}\mathcal{X}_2\times\mathcal{X}_3=\mathcal{X}_{23} & \text{for } i=1\\ \mathcal{X}_1\times\mathcal{X}_3=\mathcal{X}_{13} & \text{for } i=2\\ \mathcal{X}_1\times\mathcal{X}_2=\mathcal{X}_{12} & \text{for } i=3\end{cases}\quad\text{and}\quad \mathbf{x}_{\sim i}=\begin{cases}[x_2,x_3]=\mathbf{x}_{23} & \text{for } i=1\\ [x_1,x_3]=\mathbf{x}_{13} & \text{for } i=2\\ [x_1,x_2]=\mathbf{x}_{12} & \text{for } i=3.\end{cases} \qquad \text{(B.3.30)}
$$

The equivalence of $S_{1T}$ with the definition in equality 1 of Eq. (B.3.29) is established as follows. First,

$$
\begin{aligned}
\mathrm{V}_{\mathcal{X}_1}(f\,|\,\mathbf{x}_{\sim 1}) &= \int_{\mathcal{X}_1}\left(f(x_1,\mathbf{x}_{23})-\mathrm{E}[f(x_1,\mathbf{x}_{23})\,|\,\mathbf{x}_{23}]\right)^2 d_1(x_1)\mathrm{d}x_1\\
&=\int_{\mathcal{X}_1}\left(f^2(x_1,\mathbf{x}_{23})-2\mathrm{E}[f(x_1,\mathbf{x}_{23})\,|\,\mathbf{x}_{23}]f(x_1,\mathbf{x}_{23})+\mathrm{E}^2[f(x_1,\mathbf{x}_{23})\,|\,\mathbf{x}_{23}]\right)d_1(x_1)\mathrm{d}x_1\\
&=\int_{\mathcal{X}_1}f^2(x_1,\mathbf{x}_{23})d_1(x_1)\mathrm{d}x_1-\mathrm{E}^2[f(x_1,\mathbf{x}_{23})\,|\,\mathbf{x}_{23}] \qquad \text{(B.3.31)}\\
&=\int_{\mathcal{X}_1}f^2(x_1,\mathbf{x}_{23})d_1(x_1)\mathrm{d}x_1-\left(\int_{\mathcal{X}_1}f(x_1,\mathbf{x}_{23})d_1(x_1)\mathrm{d}x_1\right)^2\\
&=\int_{\mathcal{X}_1}f^2(x_1,\mathbf{x}_{23})d_1(x_1)\mathrm{d}x_1-\int_{\mathcal{X}_1}\int_{\tilde{\mathcal{X}}_1}f(x_1,\mathbf{x}_{23})f(\tilde{x}_1,\mathbf{x}_{23})d_1(\tilde{x}_1)d_1(x_1)\mathrm{d}\tilde{x}_1\mathrm{d}x_1.
\end{aligned}
$$

The notation $x_1$ and $\tilde{x}_1$ is introduced in the final equality of Eq. (B.3.31) to emphasize that, although $x_1$ and $\tilde{x}_1$ have the same distribution, they must be independent for the final double integral to equal the preceding squared integral. In turn,

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{X}_{\sim 1}}\left[\mathrm{V}_{\mathcal{X}_1}(f \mid \mathbf{x}_{\sim 1})\right] / \mathrm{V}(f) = \int_{\mathcal{X}_{23}} \mathrm{V}_{\mathcal{X}_1}(f \mid \mathbf{x}_{23}) d(\mathbf{x}_{23}) \mathrm{d}\mathbf{x}_{23} / \mathrm{V}(f) \\
&= \int_{\mathcal{X}_{23}} \left\{ \int_{\mathcal{X}_1} f^2(x_1, \mathbf{x}_{23}) d_1(x_1) \mathrm{d}x_1 \right. \\
&\qquad \left. - \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, \mathbf{x}_{23}) f(\tilde{x}_1, \mathbf{x}_{23}) d_1(\tilde{x}_1) d_1(x_1) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1. \right\} d(\mathbf{x}_{23}) \mathrm{d}\mathbf{x}_{23} / \mathrm{V}(f) \\
&= \left\{ \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f^2(x_1, x_2, x_3) d(\mathbf{x}) \mathrm{d}\mathbf{x} \right. \\
&\qquad \left. - \int_{\mathcal{X}_{23}} \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, x_3) d_1(\tilde{x}_1) d_1(x_1) d(\mathbf{x}_{23}) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1 \mathrm{d}\mathbf{x}_{23} \right\} / \mathrm{V}(f) \\
&= S_{1T},
\end{aligned}
\tag{B.3.32}
$$

with the last equality following from a match with equality 3 in Eq. (B.3.26). Similar derivations show that $S_{2T}$ and $S_{3T}$ are also consistent with the representation for $S_{iT}$ in equality 1 of Eq. (B.3.29).

The equivalence of $S_{1T}$ with the definition in equalities 2 and 3 of Eq. (B.3.29) is established as follows. First,

$$
\begin{aligned}
&\mathrm{V}_{\mathcal{X}_{\sim 1}}\left[\mathrm{E}_{\mathcal{X}_1}(f \mid \mathbf{x}_{\sim 1})\right] = \mathrm{V}_{\mathcal{X}_{\sim 1}}\left[ \int_{\mathcal{X}_1} f(x_1, \mathbf{x}_{23}) d_1(x_1) \mathrm{d}x_1 \right] \\
&= \int_{\mathcal{X}_{23}} \left[ \int_{\mathcal{X}_1} f(x_1, \mathbf{x}_{23}) d_1(x_1) \mathrm{d}x_1 - f_0 \right]^2 d_{23}(\mathbf{x}_{23}) \mathrm{d}\mathbf{x}_{23} \\
&= \int_{\mathcal{X}_{23}} \left\{ \left[ \int_{\mathcal{X}_1} f(x_1, \mathbf{x}_{23}) d_1(x_1) \mathrm{d}x_1 \right]^2 - 2 f_0^2 + f_0^2 \right\} d_{23}(\mathbf{x}_{23}) \mathrm{d}\mathbf{x}_{23} \\
&= \int_{\mathcal{X}_{23}} \left[ \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, \mathbf{x}_{23}) f(\tilde{x}_1, \mathbf{x}_{23}) d_1(\tilde{x}_1) d_1(x_1) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1 \right] d_{23}(\mathbf{x}_{23}) \mathrm{d}\mathbf{x}_{23} - f_0^2 \\
&= \int_{\mathcal{X}_{23}} \int_{\mathcal{X}_1} \int_{\tilde{\mathcal{X}}_1} f(x_1, x_2, x_3) f(\tilde{x}_1, x_2, x_3) d_1(\tilde{x}_1) d_1(x_1) d_{23}(\mathbf{x}_{23}) \mathrm{d}\tilde{x}_1 \mathrm{d}x_1 \mathrm{d}\mathbf{x}_{23} - f_0^2
\end{aligned}
\tag{B.3.33}
$$

In turn,

$$
\begin{aligned}
1-\mathrm{V}_{\mathcal{X}_{\sim i}}\left[\mathrm{E}_{\mathcal{X}_i}(f\mid\mathbf{x}_{\sim i})\right]/\mathrm{V}(f) &= \left\{\mathrm{V}(f)-\mathrm{V}_{\mathcal{X}_{\sim i}}\left[\mathrm{E}_{\mathcal{X}_i}(f\mid\mathbf{x}_{\sim i})\right]\right\}/\mathrm{V}(f) \\
&= \left\{\int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f^2(x_1,x_2,x_3)d(\mathbf{x})\mathrm{d}\mathbf{x} - f_0^2\right\} \\
&\quad -\left\{\int_{\mathcal{X}_{23}}\int_{\mathcal{X}_1}\int_{\tilde{\mathcal{X}}_1} f(x_1,x_2,x_3)f(\tilde{x}_1,x_2,x_3)d_1(\tilde{x}_1)d_1(x_1)d_{23}(\mathbf{x}_{23})\mathrm{d}\tilde{x}_1\mathrm{d}x_1\mathrm{d}\mathbf{x}_{23} - f_0^2\right\} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f^2(x_1,x_2,x_3)d(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\quad -\int_{\mathcal{X}_{23}}\int_{\mathcal{X}_1}\int_{\tilde{\mathcal{X}}_1} f(x_1,x_2,x_3)f(\tilde{x}_1,x_2,x_3)d_1(\tilde{x}_1)d_1(x_1)d_{23}(\mathbf{x}_{23})\mathrm{d}\tilde{x}_1\mathrm{d}x_1\mathrm{d}\mathbf{x}_{23} \\
&= S_{1T},
\end{aligned}
\tag{B.3.34}
$$

with (i) the second equality following from Eqs. (B.3.3) and (B.3.33) and (ii) the last equality following from a match with equality 3 in Eq. (B.3.26). Similar derivations show that $S_{2T}$ and $S_{3T}$ are also consistent with the representation for $S_{iT}$ in equalities 2 and 3 of Eq. (B.3.29).

## B.4 Variance Decomposition for $f(x_1, x_2, ..., x_n)$

The Sobol' variance decomposition for a function $f(\mathbf{x})$ of a vector

$$\mathbf{x} = [x_1, x_2, ..., x_n] \tag{B.4.1}$$

of arbitrary length *n* is now considered. As before, the elements $x_i$ of $\mathbf{x}$ are assumed to be independent and to have distributions defined by probability spaces $(\mathcal{X}_i, \mathbb{X}_i, m_i)$ and associated density functions $d_i(x_i)$ for $i = 1, 2, ..., n$. Further,

$$\mathcal{X} = \prod_{i=1}^{n}\mathcal{X}_i, \quad d(\mathbf{x}) = \prod_{i=1}^{n} d_i(x_i) \text{ and } \mathrm{d}\mathbf{x} = \prod_{i=1}^{n}\mathrm{d}_i x_i \tag{B.4.2}$$

for the product probability space that derives from the *n* evidence spaces $(\mathcal{X}_i, \mathbb{X}_i, m_i)$.

The Sobol' variance decomposition for $f(\mathbf{x})$ is based on the following representation for $f(\mathbf{x})$:

$$
\begin{aligned}
f(\mathbf{x}) = f_0 &+ \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{ij}(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n} f_{ijk}(x_i, x_j, x_k) \\
&+ \cdots + f_{12,...,n}(x_1, x_2, ..., x_n),
\end{aligned}
\tag{B.4.3}
$$

where

$$f_0 = \mathrm{E}(f) = \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{B.4.4}$$

$$f_i(x_i) = \mathrm{E}[f(\mathbf{x} \mid x_i)] - f_0 = \int_{\mathcal{X}_{\sim i}} f(\mathbf{x} \mid x_i) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} - f_0, \tag{B.4.5}$$

$$\begin{aligned} f_{ij}(x_i, x_j) &= \mathrm{E}[f(\mathbf{x} \mid x_i, x_j)] - f_0 - f_i(x_i) - f_j(x_j) \\ &= \int_{\mathcal{X}_{\sim ij}} f(\mathbf{x} \mid x_i, x_j) d_{\sim ij}(\mathbf{x}_{\sim ij}) \mathrm{d}\mathbf{x}_{\sim ij} - f_0 - f_i(x_i) - f_j(x_j), \end{aligned} \tag{B.4.6}$$

$$\begin{aligned} f_{ijk}(x_i, x_j, x_k) &= \mathrm{E}[f(\mathbf{x} \mid x_i, x_j, x_k)] - f_0 - f_i(x_i) - f_j(x_j) - f_k(x_k) \\ &\quad - f_{ij}(x_i, x_j) - f_{ik}(x_i, x_k) - f_{jk}(x_j, x_k) \\ &= \int_{\mathcal{X}_{\sim ijk}} f(\mathbf{x} \mid x_i, x_j, x_k) d_{\sim ijk}(\mathbf{x}_{\sim ijk}) \mathrm{d}\mathbf{x}_{\sim ijk} - f_0 - f_i(x_i) - f_j(x_j) - f_k(x_k) \\ &\quad - f_{ij}(x_i, x_j) - f_{ik}(x_i, x_k) - f_{jk}(x_j, x_k), \end{aligned} \tag{B.4.7}$$

and so on with the notations

$$\mathcal{X}_{\sim i}, \mathcal{X}_{\sim ij}, \mathcal{X}_{\sim ijk}, d_{\sim i}(\mathbf{x}_{\sim i}), d_{\sim ij}(\mathbf{x}_{\sim ij}), d_{\sim ijk}(\mathbf{x}_{\sim ijk}), \mathrm{d}\mathbf{x}_{\sim i}, \mathrm{d}\mathbf{x}_{\sim ij}, \mathrm{d}\mathbf{x}_{\sim ijk} \tag{B.4.8}$$

indicating that (i) $x_i$, (ii) $x_i$ and $x_j$, and (iii) $x_i$, $x_j$ and $x_k$, respectively, are omitted from the definitions of the corresponding expressions.

A general representation for the expressions in Eqs. (B.4.3)-(B.4.7) can be given with the introductions of the following notation:

$$\mathcal{I} = \{1, 2, ..., n\} \tag{B.4.9}$$

$$\mathcal{U} = \{\mathbf{u} : \mathbf{u} = [i_1, i_2, ..., i_r] \text{ with } \{i_1, i_2, ..., i_r\} \subset \mathcal{I}, i_1 < i_2 < \cdots < i_r\} \tag{B.4.10}$$

$$\mathcal{I}(\mathbf{u}) = \{i_1, i_2, ..., i_r\} \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \tag{B.4.11}$$

$$\mathbf{x}_{\mathbf{u}} = [x_{i_1}, x_{i_2}, ..., x_{i_r}] \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \tag{B.4.12}$$

$$\mathcal{X}_{\mathbf{u}} = \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \cdots \times \mathcal{X}_{i_r}] \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \tag{B.4.13}$$

$$d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) = d_{i_1}(x_{i_1}) d_{i_2}(x_{i_2}) \cdots d_{i_r}(x_{i_r})] \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \tag{B.4.14}$$

$$\mathrm{d}\mathbf{x}_{\mathbf{u}} = \mathrm{d}x_{i_1} \mathrm{d}x_{i_2} \cdots \mathrm{d}x_{i_r} \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \tag{B.4.15}$$

$$\mathcal{V}(\mathbf{u}) = \{\mathbf{v} : \mathbf{v} = [j_1, j_2, ..., j_s] \text{ with } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, j_1 < j_2 < \cdots < j_s, \\ s < r, \{j_1, j_2, ..., j_s\} \subset \mathcal{I}(\mathbf{u})\} \tag{B.4.16}$$

and

$$\tilde{\mathbf{u}} = [k_1, k_2, ..., k_{n-r}] \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}, \mathcal{I}(\tilde{\mathbf{u}}) = \{k_1, k_2, ..., k_{n-r}\}, \\ \mathcal{I}(\tilde{\mathbf{u}}) \cap \mathcal{I}(\mathbf{u}) = \varnothing \text{ and } \mathcal{I}(\tilde{\mathbf{u}}) \cup \mathcal{I}(\mathbf{u}) = \mathcal{I}. \tag{B.4.17}$$

In turn, the representation for $f(\mathbf{x})$ in Eq. (B.4.3) can be restated as

$$f(\mathbf{x}) = f_0 + \sum_{\mathbf{u} \in \mathcal{U}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) \tag{B.4.18}$$

with

$$\begin{aligned} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) &= \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}} - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \\ &= \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \end{aligned} \tag{B.4.19}$$

for $\mathbf{u} \in \mathcal{U}$.

The validity of the representation for $f(\mathbf{x})$ in Eqs. (B.4.3) and (B.4.18) follows by direct verification as shown below:

$$\begin{aligned} f_0 + \sum_{\mathbf{u} \in \mathcal{U}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) &=_1 f_0 + \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{i})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) + f_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}) \text{ for } \mathbf{i} = [1, 2, ..., n] \text{ and } \mathbf{x}_{\mathbf{i}} = [x_1, x_2, ..., x_n] \\ &=_2 f_0 + \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{i})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) + \left[ f(\mathbf{x}) - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{i})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right] \\ &=_3 f(\mathbf{x}) \end{aligned} \tag{B.4.20}$$

with the substitution for $f_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})$ in Equality 2 following from Eq. (B.4.19).

The expected value of

$$f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) = f_{i_1 i_2 \cdots i_r}(x_{i_1}, x_{i_2}, ..., x_{i_r}) \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U} \tag{B.4.21}$$

with respect to any of its components (i.e., $x_{i_1}, x_{i_2}, ..., x_{i_r}$) is zero. More formally, if $\mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}$ and $i \in \mathcal{I}(\mathbf{u})$, then

$$\int_{\mathcal{X}_i} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) d_i(x_i) \mathrm{d}x_i = \int_{\mathcal{X}_i} f_{i_1 i_2 \cdots i_r}(x_{i_1}, x_{i_2}, ..., x_{i_r}) d_i(x_i) \mathrm{d}x_i = 0. \tag{B.4.22}$$

This result can be established by induction for $r = 1, 2, ..., n$. For $r = 1$,

$$\begin{aligned}
\int_{\mathcal{X}_i} f_i(x_i) d_i(x_i) \mathrm{d}x_i &= \int_{\mathcal{X}_i} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) d_i(x_i) \mathrm{d}x_i \text{ for } \mathbf{u} = [i] \\
&= \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x} \mid x_i) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}} - f_0 \right] d_i(x_i) \mathrm{d}x_i \\
&= \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{\mathcal{X}_i} f_0 d_i(x_i) \mathrm{d}x_i \\
&= f_0 - f_0 \\
&= 0.
\end{aligned} \tag{B.4.23}$$

The result in Eq. (B.4.22) is now assumed to be true for integers from 1 to $r-1$ and established for *r*. Specifically,

$$\begin{aligned}
\int_{\mathcal{X}_i} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) d_i(x_i) \mathrm{d}x_i &=_1 \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}} - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right] d_i(x_i) \mathrm{d}x_i \\
&=_2 \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}} \right] d_i(x_i) \mathrm{d}x_i - \int_{\mathcal{X}_i} f_0 d_i(x_i) \mathrm{d}x_i - \int_{\mathcal{X}_i} \left[ \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right] d_i(x_i) \mathrm{d}x_i \\
&=_3 \int_{\mathcal{X}_{\tilde{\mathbf{w}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{w}}) d_{\tilde{\mathbf{w}}}(\mathbf{x}_{\tilde{\mathbf{w}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} - f_0 - \left[ \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{w})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) + f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) \right] \\
&=_4 \left[ \int_{\mathcal{X}_{\tilde{\mathbf{w}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{w}}) d_{\tilde{\mathbf{w}}}(\mathbf{x}_{\tilde{\mathbf{w}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{w})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right] - f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) \\
&=_5 f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) - f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) \\
&=_6 0,
\end{aligned} \tag{B.4.24}$$

with (i) Equality 1 following from the definition of $f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})$ in Eq. (B.4.19), (ii) Equality 2 obtained by splitting the single integral in Equality 1 into 3 integrals, (iii) Equality 3 obtained by introducing $\mathbf{w} = [j_1, j_2, ..., j_{r-1}]$ containing the same integers as $\mathbf{u} = [i_1, i_2, ..., i_r]$ except for the omission of *i* and then making the substitutions

$$\int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}} \right] d_i(x_i) \mathrm{d}x_i = \int_{\mathcal{X}_{\tilde{\mathbf{w}}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{w}}) d_{\tilde{\mathbf{w}}}(\mathbf{x}_{\tilde{\mathbf{w}}}) \mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} \tag{B.4.25}$$

based on $\mathcal{X}_i \times \mathcal{X}_{\tilde{\mathbf{u}}} = \mathcal{X}_{\tilde{\mathbf{w}}}$,

$$\int_{\mathcal{X}_i} f_0 d_i(x_i)\mathrm{d}x_i = f_0 \int_{\mathcal{X}_i} d_i(x_i)\mathrm{d}x_i = f_0, \tag{B.4.26}$$

and

$$\int_{\mathcal{X}_i} \left[ \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right] d_i(x_i)\mathrm{d}x_i = \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{w})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) + f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) \tag{B.4.27}$$

based on the induction hypothesis that Eq. (B.4.22) is true for integers from 1 to $r-1$, (iv) Equality 4 obtained by rearranging the terms in Equality 3, and (v) Equality 5 following from the definition of $f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}})$ in Eq. (B.4.19).

The equality

$$\mathrm{E}(f_{\mathbf{u}}) = 0 \quad \text{for } \mathbf{u} \in \mathcal{U} \tag{B.4.28}$$

is an immediate and important consequence of the result in Eq. (B.4.24).

As for the three-dimensional case (i.e., $n$ = 3) introduced in Sect. B.2, inner products and orthogonality play an important role in the development of Sobol' decomposition for arbitrary vectors (i.e., $\mathbf{x} = [x_1, x_2, ..., x_n]$).

Specifically, the inner product of functions $g$ and $h$ defined on $\mathcal{X}$ is defined by

$$\langle g, h \rangle = \int_{\mathcal{X}} g(\mathbf{x})h(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{B.4.29}$$

with $\mathcal{X}$, $d(\mathbf{x})$ and $\mathrm{d}\mathbf{x}$ introduced in conjunction with Eq. (B.4.2). The inner product $\langle g, h \rangle$ has the properties listed in Eqs. (B.2.19) and (B.2.20). The functions $g$ and $h$ are orthogonal if $\langle g, h \rangle = 0$. Further, the variance $V(g)$ is given by

$$\begin{aligned} V(g) &= \langle g - g_0, g - g_0 \rangle \\ &= \int_{\mathcal{X}} [g(\mathbf{x}) - g_0]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{X}} g^2(\mathbf{x})d(\mathbf{x})\mathrm{d}\mathbf{x} - g_0^2, \end{aligned} \tag{B.4.30}$$

where

$$g_0 = \mathrm{E}(g) = \int_{\mathcal{X}} g(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{B.4.31}$$

is the expected value for $g$.

Two functions $f_{\mathbf{u}}$ and $f_{\mathbf{v}}$ of the form defined in Eqs. (B.4.4)-(B.4.8) and more generally in Eq. (B.4.19) with $\mathbf{u} = [i_1, i_2, ..., i_r]$ and $\mathbf{v} = [j_1, j_2, ..., j_s]$ are orthogonal if

$$\mathbf{x}_{\mathbf{u}} = [x_{i_1}, x_{i_2}, ..., x_{i_r}] \text{ and } \mathbf{x}_{\mathbf{v}} = [x_{j_1}, x_{j_2}, ..., x_{j_s}] \tag{B.4.32}$$

differ in one or more elements of $\mathbf{x} = [x_1, x_2, ..., x_n]$ as shown below. For notational convenience, let (i) $\mathbf{a}$ represent the integers contained in both $\mathbf{u}$ and $\mathbf{v}$, (ii) $\mathbf{b}$ represent the integers contained in $\mathbf{u}$ but not $\mathbf{v}$, and (iii) $\mathbf{c}$ represent the integers contained in $\mathbf{v}$ but not $\mathbf{u}$ Then,

$$\begin{aligned}
\left\langle f_{\mathbf{u}}, f_{\mathbf{v}} \right\rangle &= \int_{\mathcal{X}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}_{\mathbf{a}}} \left[ \int_{\mathcal{X}_{\mathbf{b+c}}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{b+c}}) \mathrm{d}\mathbf{x}_{\mathbf{b+c}} \right] d(\mathbf{x}_{\mathbf{a}}) \mathrm{d}\mathbf{x}_{\mathbf{a}} \\
&= \int_{\mathcal{X}_{\mathbf{a}}} \left[ \int_{\mathcal{X}_{\mathbf{b}}} f_{\mathbf{u}}(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}}) d(\mathbf{x}_{\mathbf{b}}) \mathrm{d}\mathbf{x}_{\mathbf{b}} \times \int_{\mathcal{X}_{\mathbf{c}}} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{c}}) \mathrm{d}\mathbf{x}_{\mathbf{c}} \right] d(\mathbf{x}_{\mathbf{a}}) \mathrm{d}\mathbf{x}_{\mathbf{a}} \\
&= 0,
\end{aligned} \tag{B.4.33}$$

with

$$\mathcal{X}_{\mathbf{a}}, \mathcal{X}_{\mathbf{b}}, \mathcal{X}_{\mathbf{c}}, \mathcal{X}_{\mathbf{b+c}}, d_{\mathbf{a}}(\mathbf{x}_{\mathbf{a}}), d_{\mathbf{b}}(\mathbf{x}_{\mathbf{b}}), d_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}), d_{\mathbf{b+c}}(\mathbf{x}_{\mathbf{b+c}}), \mathrm{d}_{\mathbf{a}}\mathbf{x}_{\mathbf{a}}, \mathrm{d}_{\mathbf{b}}\mathbf{x}_{\mathbf{b}}, \mathrm{d}_{\mathbf{c}}\mathbf{x}_{\mathbf{c}}, \mathrm{d}_{\mathbf{b+c}}\mathbf{x}_{\mathbf{b+c}} \tag{B.4.34}$$

defined the same as in Eqs. (B.4.12)-(B.4.15) and equality to zero resulting because at least one of the two inner integrals must occur and be equal to zero due to the assumption that $\mathbf{u}$ and $\mathbf{v}$ differ in one or more elements.

The Sobol' variance decomposition follows from the orthogonality of the individual terms defined in in Eqs. (B.4.4)-(B.4.8) and more generally in Eq. (B.4.19). Specifically,

$$
\begin{aligned}
\mathrm{V}(f) &=_1 \left\langle f - f_0, f - f_0 \right\rangle \\
&=_2 \left\langle f_0 + \sum_{\mathbf{u}\in\mathcal{U}} f_{\mathbf{u}} - f_0, f_0 + \sum_{\mathbf{u}\in\mathcal{U}} f_{\mathbf{u}} - f_0 \right\rangle \\
&=_3 \left\langle \sum_{\mathbf{u}\in\mathcal{U}} f_{\mathbf{u}}, \sum_{\mathbf{u}\in\mathcal{U}} f_{\mathbf{u}} \right\rangle \\
&=_4 \sum_{\mathbf{u}\in\mathcal{U}} \sum_{\mathbf{v}\in\mathcal{U}} \left\langle f_{\mathbf{u}}, f_{\mathbf{v}} \right\rangle \\
&=_5 \sum_{\mathbf{u}\in\mathcal{U}} \left\langle f_{\mathbf{u}}, f_{\mathbf{u}} \right\rangle \\
&=_6 \sum_{\mathbf{u}\in\mathcal{U}} \left\langle f_{\mathbf{u}} - 0, f_{\mathbf{u}} - 0 \right\rangle \\
&=_7 \sum_{\mathbf{u}\in\mathcal{U}} \mathrm{V}(f_{\mathbf{u}}) \\
&=_8 \sum_{i=1}^{n} \mathrm{V}(f_i) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{V}(f_{ij}) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \mathrm{V}(f_{ijk}) + \cdots + \mathrm{V}(f_{12,\ldots,n}),
\end{aligned}
\tag{B.4.35}
$$

where (i) Equality 1 follows from the definition of variance in Eq. (B.4.30), (ii) Equality 2 follows from the representation for $f$ in Eq. (B.4.18), (iii) Equality 3 follows from $f_0 - f_0 = 0$, (iv) Equality 4 follows from Eq. (B.2.20), (v) Equality 5 follows from orthogonality (i.e., $\left\langle f_{\mathbf{u}}, f_{\mathbf{v}} \right\rangle = 0$ for $\mathbf{u} \neq \mathbf{v}$; see Eq. (B.4.33)), (vi) Equality 6 introduces $\mathrm{E}(f_{\mathbf{u}}) = 0$ from Eq. (B.4.28), (vii) Equality 7 follows from the definition of variance in Eq. (B.4.30), and (viii) Equality 8 is a restatement of the summation in Equality 7.

Division of both sides of Eq. (B.4.35) by V($f$) yields the representation

$$
\begin{aligned}
1 &= \sum_{\mathbf{u}\in\mathcal{U}} \mathrm{V}(f_{\mathbf{u}}) / \mathrm{V}(f) \\
&= \sum_{i=1}^{n} S_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} S_{ij} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} S_{ijk} + \cdots + S_{12,\ldots,n}
\end{aligned}
\tag{B.4.36}
$$

with

$$
S_i = \text{ fractional contribution of variable } x_i \text{ to } \mathrm{V}(f), \tag{B.4.37}
$$

$$
S_{ij} = \text{ fractional interaction contribution of variables } x_i \text{ and } x_j \text{ to } \mathrm{V}(f), \tag{B.4.38}
$$

$$
S_i + S_j + S_{ij} = \text{ combined fractional individual and interaction contributions of variables } x_i \text{ and } x_j \text{ to } \mathrm{V}(f), \tag{B.4.39}
$$

and, more generally,

$$S_{\mathbf{u}} = \mathrm{V}(f_{\mathbf{u}}) / \mathrm{V}(f) \text{ for } \mathbf{u} = [i_1, i_2, ..., i_r]$$
$$= \text{fractional interaction contribution of variables contained in } \mathbf{x}_{\mathbf{u}} = [x_{i_1}, x_{i_2}, ..., x_{i_r}] \text{ to } \mathrm{V}(f). \quad \text{(B.4.40)}$$

In addition, the total fractional contribution of variable $x_i$ to V(*f*) is given by

$$S_{iT} = \sum_{\mathbf{u} \in \mathcal{T}_i} S_{\mathbf{u}} = \sum_{\mathbf{u} \in \mathcal{T}_i} \mathrm{V}(f_{\mathbf{u}}) / \mathrm{V}(f) \quad \text{(B.4.41)}$$

with

$$\mathcal{T}_i = \{\mathbf{u} : \mathbf{u} = [i_1, i_2, ..., i_r], 1 \le r \le n, i_1 < i_2 < \cdots < i_r, i \in \{i_1, i_2, ..., i_r\}\} \quad \text{(B.4.42)}$$

The preceding summation contains the terms in Eq. (B.4.36) that do not have a subscript *i*. Thus, $S_{iT}$ equals the fraction of V(*f*) due to $x_i$ and all variable interactions involving $x_i$. The value of $S_{iT}$ provides a measure of the overall effect of $x_i$ on V(*f*). However, the $S_{iT}$ for all $x_i$'s typically do not sum to 1 because of overlapping effects (e.g., $S_{12}$ appears in each of the sums defining $S_{1T}$ and $S_{2T}$).

The variance $\mathrm{V}(f_i)$ of $f_i$ is given by

$$\begin{aligned} \mathrm{V}(f_i) &=_1 \langle f_i - \mathrm{E}_i(f_i), f_i - \mathrm{E}_i(f_i) \rangle \\ &=_2 \langle f_i - 0, f_i - 0 \rangle \\ &=_3 \langle f_i, f_i \rangle \\ &=_4 \langle \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] - f_0, \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] - f_0 \rangle \\ &=_5 \mathrm{V}_i \{\mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)]\}, \end{aligned} \quad \text{(B.4.43)}$$

with Equality 5 following from Equality 4 as a consequence of the equality

$$\begin{aligned} \mathrm{E}_i \{\mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)]\} &= \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\sim i}} f(\mathbf{x} \mid x_i) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} \right] d_i(x_i) \mathrm{d}x_i \\ &= \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} = f_0. \end{aligned} \quad \text{(B.4.44)}$$

Further, continuation from Equality 4 in Eq. (B.4.43) yields

$$
\begin{aligned}
\mathrm{V}(f_i) &= \left\langle \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] - f_0, \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] - f_0 \right\rangle \\
&= \left\langle \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)], \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] \right\rangle - 2f_0 \left\langle \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)], 1 \right\rangle + f_0^2 \\
&= \int_{\mathcal{X}_i} \left\{ \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] \right\}^2 d_i(x_i) \mathrm{d}x_i - 2f_0^2 + f_0^2 \\
&= \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\sim i}} f(\mathbf{x} \mid x_i) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} \right]^2 d_i(x_i) \mathrm{d}x_i - f_0^2 .
\end{aligned}
\tag{B.4.45}
$$

In turn, combining Eqs. (B.4.43) and (B.4.45) produces

$$
\begin{aligned}
\mathrm{V}(f_i) &= \mathrm{V}_i \left\{ \mathrm{E}_{\sim i}[f(\mathbf{x} \mid x_i)] \right\} \\
&= \int_{\mathcal{X}_i} \left[ \int_{\mathcal{X}_{\sim i}} f(\mathbf{x} \mid x_i) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} \right]^2 d_i(x_i) \mathrm{d}x_i - f_0^2 \\
&= \int_{\mathcal{X}_i} \int_{\mathcal{X}_{\sim i}} \int_{\tilde{\mathcal{X}}_{\sim i}} f(\mathbf{x}_{\sim i}, x_i) f(\tilde{\mathbf{x}}_{\sim i}, x_i) d_{\sim i}(\tilde{\mathbf{x}}_{\sim i}) d_{\sim i}(\mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}\tilde{\mathbf{x}}_{\sim i} \mathrm{d}\mathbf{x}_{\sim i} \mathrm{d}x_i - f_0^2 \\
&= \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_{\sim i}} f(\mathbf{x}) f(\tilde{\mathbf{x}}_{\sim i}, x_i) d_{\sim i}(\tilde{\mathbf{x}}_{\sim i}) d(\mathbf{x}) \mathrm{d}\tilde{\mathbf{x}}_{\sim i} \mathrm{d}\mathbf{x} - f_0^2 ,
\end{aligned}
\tag{B.4.46}
$$

with the final two equalities equality obtained as indicated in Eq. (B.3.4).

Similarly, the variance $\mathrm{V}(f_{ij})$ of $f_{ij}$ is given by

$$
\begin{aligned}
\mathrm{V}(f_{ij}) &=_1 \left\langle f_{ij} - \mathrm{E}_{ij}(f_{ij}), f_{ij} - \mathrm{E}_{ij}(f_{ij}) \right\rangle \\
&=_2 \left\langle f_{ij} - 0, f_{ij} - 0 \right\rangle \\
&=_3 \left\langle f_{ij}, f_{ij} \right\rangle \\
&=_4 \left\langle \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0 - f_i - f_j, \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0 - f_i - f_j \right\rangle \\
&=_5 \left\langle \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0, \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0 \right\rangle \\
&\quad - 2\left\langle \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0, f_i + f_j \right\rangle + \left\langle f_i + f_j, f_i + f_j \right\rangle \\
&=_6 \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - 2\left\langle f_{ij} + f_i + f_j, f_i + f_j \right\rangle + \left\langle f_i + f_j, f_i + f_j \right\rangle \\
&=_7 \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - 2\left\{ \left\langle f_{ij}, f_i + f_j \right\rangle + \left\langle f_i + f_j, f_i + f_j \right\rangle \right\} + \left\langle f_i + f_j, f_i + f_j \right\rangle \\
&=_8 \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - 2\left\langle f_{ij}, f_i + f_j \right\rangle - \left\langle f_i + f_j, f_i + f_j \right\rangle \\
&=_9 \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - 2\left\{ \left\langle f_{ij}, f_i \right\rangle + \left\langle f_{ij}, f_j \right\rangle \right\} - \left\{ \left\langle f_i, f_i \right\rangle - 2\left\langle f_i, f_j \right\rangle + \left\langle f_j, f_j \right\rangle \right\} \\
&=_{10} \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - 2\{0 + 0\} - \left\{ \mathrm{V}(f_i) - 2 \times 0 + \mathrm{V}(f_j) \right\} \\
&=_{11} \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - \mathrm{V}(f_i) - \mathrm{V}(f_j).
\end{aligned}
\tag{B.4.47}
$$

Further, continuation from the last equality in the preceding equation yields

$$
\begin{aligned}
\mathrm{V}(f_{ij}) &= \mathrm{V}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - \mathrm{E}_{ij}\left\{ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right\} \right]^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] - f_0 \right]^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] \right]^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - 2 f_0 \int_{\mathcal{X}_{ij}} \mathrm{E}_{\sim ij}[f(\mathbf{x} \mid \mathbf{x}_{ij})] d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} \\
&\qquad + \int_{\mathcal{X}_{ij}} f_0^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \int_{\mathcal{X}_{\sim ij}} f(\mathbf{x} \mid \mathbf{x}_{ij}) d_{\sim ij}(\mathbf{x}_{\sim ij}) \mathrm{d}\mathbf{x}_{\sim ij} \right]^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - 2 f_0^2 + f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \int_{\mathcal{X}_{\sim ij}} f(\mathbf{x} \mid \mathbf{x}_{ij}) d_{\sim ij}(\mathbf{x}_{\sim ij}) \mathrm{d}\mathbf{x}_{\sim ij} \right]^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j).
\end{aligned}
\tag{B.4.48}
$$

In summary, combining the final equalities in Eqs. (B.4.47) and (B.4.48) yields

$$
\begin{aligned}
\mathrm{V}(f_{ij}) &= \mathrm{V}_{ij}\left\{\mathrm{E}_{\sim ij}[f(\mathbf{x}\,|\,\mathbf{x}_{ij})]\right\} - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \left[ \int_{\mathcal{X}_{\sim ij}} f(\mathbf{x}\,|\,\mathbf{x}_{ij}) d_{\sim ij}(\mathbf{x}_{\sim ij}) \mathrm{d}\mathbf{x}_{\sim ij} \right]^2 d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\mathbf{x}_{ij} - f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j) \\
&= \int_{\mathcal{X}_{ij}} \int_{\mathcal{X}_{\sim ij}} \int_{\tilde{\mathcal{X}}_{\sim ij}} f(\mathbf{x}_{\sim ij}, \mathbf{x}_{ij}) f(\tilde{\mathbf{x}}_{\sim ij}, \mathbf{x}_{ij}) d_{\sim ij}(\tilde{\mathbf{x}}_{\sim ij}) d_{\sim ij}(\mathbf{x}_{\sim ij}) d_{ij}(\mathbf{x}_{ij}) \mathrm{d}\tilde{\mathbf{x}}_{\sim ij} \mathrm{d}\mathbf{x}_{\sim ij} \mathrm{d}\mathbf{x}_{ij} \quad \text{(B.4.49)} \\
&\qquad - f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j), \\
&= \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_{\sim ij}} f(\mathbf{x}) f(\tilde{\mathbf{x}}_{\sim ij}, \mathbf{x}_{ij}) d_{\sim ij}(\tilde{\mathbf{x}}_{\sim ij}) d(\mathbf{x}) \mathrm{d}\tilde{\mathbf{x}}_{\sim ij} \mathrm{d}\mathbf{x} - f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j),
\end{aligned}
$$

with the final two equalities obtained as indicated in Eq. (B.3.4).

More generally, if $\mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}$, then $\mathrm{V}(f_{\mathbf{u}})$ is defined by

$$
\begin{aligned}
\mathrm{V}(f_{\mathbf{u}}) &=_1 \left\langle f_{\mathbf{u}} - \mathrm{E}_{\mathbf{u}}(f_{\mathbf{u}}), f_{\mathbf{u}} - \mathrm{E}_{\mathbf{u}}(f_{\mathbf{u}}) \right\rangle \\
&=_2 \left\langle f_{\mathbf{u}} - 0, f_{\mathbf{u}} - 0 \right\rangle \\
&=_3 \left\langle f_{\mathbf{u}}, f_{\mathbf{u}} \right\rangle \\
&=_4 \left\langle \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_5 \left\langle \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0, \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0 \right\rangle - 2 \left\langle \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0, \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&\qquad + \left\langle \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_6 \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - 2 \left\langle f_{\mathbf{u}} + \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&\qquad + \left\langle \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_7 \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - 2 \left\{ \left\langle f_{\mathbf{u}}, \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle + \left\langle \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \right\} \\
&\qquad + \left\langle \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_8 \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - 2 \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} \left\langle f_{\mathbf{u}}, f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle - \left\langle \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_9 \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - 2 \times 0 - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} \sum_{\mathbf{w} \in \mathcal{V}(\mathbf{u})} \left\langle f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), f_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}}) \right\rangle \\
&=_{10} \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} \left\langle f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}), f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) \right\rangle \\
&=_{11} \mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} - \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}),
\end{aligned}
\tag{B.4.50}
$$

where (i) Equalities 2 and 3 follow from Eq. (B.4.28), (ii) Equality 4 follows from the definition of $f_{\mathbf{u}}$ in Eq. (B.4.19), (iii) Equality 6 follows from the equalities

$$
\mathrm{V}\left\{ \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] \right\} = \left\langle \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0, \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0 \right\rangle \tag{B.4.51}
$$

$$
f_{\mathbf{u}} + \sum_{\mathbf{v} \in \mathcal{V}(\mathbf{u})} f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) = \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x} \mid \mathbf{x}_{\mathbf{u}})] - f_0, \tag{B.4.52}
$$

(iv) Equalities 9 and 11 follow from Eq. (B.4.33), and (v) Equalities 5, 7, 8 and 10 follow from standard inner product manipulations.

Further, continuation from the last equality in Eq. (B.4.50) yields

$$
\begin{aligned}
\mathrm{V}(f_{\mathbf{u}}) &= \mathrm{V}\left\{\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})]\right\} - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})] - \mathrm{E}_{\mathbf{u}}\left\{\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})]\right\}\right]^2 - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})] - f_0\right]^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})]\right]^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - 2f_0 \int_{\mathcal{X}_{ij}} \mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})] d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} \\
&\qquad + \int_{\mathcal{X}_{\mathbf{u}}} f_0^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}})\mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}}\right]^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - 2f_0^2 + f_0^2 - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}})\mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}}\right]^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0^2 - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}})
\end{aligned}
\tag{B.4.53}
$$

with

$$
\begin{aligned}
\mathrm{E}_{\mathbf{u}}\left\{\mathrm{E}_{\sim\mathbf{u}}[f(\mathbf{x}_{\sim\mathbf{u}}\,|\,\mathbf{x}_{\mathbf{u}})]\right\} &= \int_{\mathcal{X}_{\mathbf{u}}} \left[\int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}})\mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}}\right] d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} \\
&= \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x})\mathrm{d}\mathbf{x} = f_0
\end{aligned}
\tag{B.4.54}
$$

in the second equality. Finally, combining the final equalities in Eqs. (B.4.50) and (B.4.53) yields

$$
\begin{aligned}
\mathrm{V}(f_{\mathbf{u}}) &= \mathrm{V}\left\{\mathrm{E}_{\tilde{\mathbf{u}}}[f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}})]\right\} - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \left[\int_{\mathcal{X}_{\tilde{\mathbf{u}}}} f(\mathbf{x}\,|\,\mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}})\mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}}\right]^2 d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0^2 - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}) \\
&= \int_{\mathcal{X}_{\mathbf{u}}} \int_{\mathcal{X}_{\tilde{\mathbf{u}}}} \int_{\tilde{\mathcal{X}}_{\tilde{\mathbf{u}}}} f(\mathbf{x}_{\tilde{\mathbf{u}}}, \mathbf{x}_{\mathbf{u}}) f(\tilde{\mathbf{x}}_{\tilde{\mathbf{u}}}, \mathbf{x}_{\mathbf{u}}) d_{\tilde{\mathbf{u}}}(\tilde{\mathbf{x}}_{\tilde{\mathbf{u}}}) d_{\tilde{\mathbf{u}}}(\mathbf{x}_{\tilde{\mathbf{u}}}) d_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\mathrm{d}\tilde{\mathbf{x}}_{\tilde{\mathbf{u}}}\mathrm{d}\mathbf{x}_{\tilde{\mathbf{u}}}\mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0^2 - \sum_{\mathbf{v}\in\mathcal{V}(\mathbf{u})} \mathrm{V}(f_{\mathbf{v}}),
\end{aligned}
\tag{B.4.55}
$$

with the final equality obtained as indicated in Eq. (B.3.4).

An integral-based representation for $S_{iT}$ is now determined. As a reminder, the notations $\mathcal{I} = \{1, 2, ..., n\}$ and $\mathcal{I}(\mathbf{u}) = \{i_1, i_2, ..., i_r\}$ for $\mathbf{u} = [i_1, i_2, ..., i_r] \in \mathcal{U}$ are introduced in Eqs. (B.4.9)-(B.4.11). The following considers variance decompositions for vectors

$$\mathbf{x}_\mathbf{u} \text{ with } \mathbf{u} = [i_1, i_2, ..., i_r] \text{ and } \mathbf{x}_\mathbf{v} \text{ with } \mathbf{v} = [j_1, j_2, ..., j_s] \tag{B.4.56}$$

with

$$\mathcal{I}(\mathbf{u}) \cup \mathcal{I}(\mathbf{v}) = \mathcal{I} = \{1, 2, ..., n\} \text{ and } \mathcal{I}(\mathbf{u}) \cap \mathcal{I}(\mathbf{v}) = \varnothing. \tag{B.4.57}$$

The purpose of the introduced vectors

$$\mathbf{x}_\mathbf{u} = [x_{i_1}, x_{i_2}, ..., x_{i_r}] \text{ and } \mathbf{x}_\mathbf{v} = [x_{j_1}, x_{j_2}, ..., x_{j_s}] \tag{B.4.58}$$

is to divide the variables associated with $\mathbf{x} = [x_1, x_2, ..., x_n]$ into two nonintersecting sets of variables and then construct a variance decomposition for these two sets of variables (i.e., for $\mathbf{x}_\mathbf{u}$ and $\mathbf{x}_\mathbf{v}$).

The following definitions are made to support the development of the indicated variance decomposition:

$$\begin{aligned} \overline{f}_\mathbf{u}(\mathbf{x}_\mathbf{u}) &= \mathrm{E}[f(\mathbf{x} \mid \mathbf{x}_\mathbf{u})] - \mathrm{E}[f(\mathbf{x})] \\ &= \int_{\mathcal{X}_\mathbf{v}} f(\mathbf{x} \mid \mathbf{x}_\mathbf{u}) d(\mathbf{x}_\mathbf{v}) \mathrm{d}\mathbf{x}_\mathbf{v} - f_0, \end{aligned} \tag{B.4.59}$$

$$\begin{aligned} \overline{f}_\mathbf{v}(\mathbf{x}_\mathbf{v}) &= \mathrm{E}[f(\mathbf{x} \mid \mathbf{x}_\mathbf{v})] - \mathrm{E}[f(\mathbf{x})] \\ &= \int_{\mathcal{X}_\mathbf{u}} f(\mathbf{x} \mid \mathbf{x}_\mathbf{v}) d(\mathbf{x}_\mathbf{u}) \mathrm{d}\mathbf{x}_\mathbf{u} - f_0, \end{aligned} \tag{B.4.60}$$

and

$$\begin{aligned} \overline{f}_{\mathbf{u},\mathbf{v}}(\mathbf{x}_\mathbf{u}, \mathbf{x}_\mathbf{v}) &= \mathrm{E}[f(\mathbf{x} \mid \mathbf{x}_\mathbf{u}, \mathbf{x}_\mathbf{v})] - \mathrm{E}[f(\mathbf{x})] - f_\mathbf{u}(\mathbf{x}_\mathbf{u}) - f_\mathbf{v}(\mathbf{x}_\mathbf{v}) \\ &= f(\mathbf{x}) - f_0 - f_\mathbf{u}(\mathbf{x}_\mathbf{u}) - f_\mathbf{v}(\mathbf{x}_\mathbf{v}), \end{aligned} \tag{B.4.61}$$

with the overbars added to distinguish the functions $\overline{f}_\mathbf{u}(\mathbf{x}_\mathbf{u})$, $\overline{f}_\mathbf{v}(\mathbf{x}_\mathbf{v})$ and $\overline{f}_{\mathbf{u},\mathbf{v}}(\mathbf{x}_\mathbf{u}, \mathbf{x}_\mathbf{v})$ from the functions $f_i(x_i)$, $f_{ij}(x_i, x_i)$, $f_{ijk}(x_i, x_j, x_k)$, … defined in Eqs. (B.4.5)-(B.4.8) and their often used notational representation $f_\mathbf{u}(\mathbf{x}_\mathbf{u})$ as indicated in Eqs. (B.4.18) and (B.4.19).

Given the preceding definitions for $\overline{f}_\mathbf{u}(\mathbf{x}_\mathbf{u})$, $\overline{f}_\mathbf{v}(\mathbf{x}_\mathbf{v})$ and $\overline{f}_{\mathbf{u},\mathbf{v}}(\mathbf{x}_\mathbf{u}, \mathbf{x}_\mathbf{v})$, the following equalities hold:

$$\mathrm{E}[\bar{f}_{\mathbf{u}}] = \mathrm{E}[\bar{f}_{\mathbf{v}}] = \mathrm{E}[\bar{f}_{\mathbf{u},\mathbf{v}}] = 0, \tag{B.4.62}$$

$$\left\langle \bar{f}_{\mathbf{u}}, \bar{f}_{\mathbf{v}} \right\rangle = \left\langle \bar{f}_{\mathbf{u}}, \bar{f}_{\mathbf{u},\mathbf{v}} \right\rangle = \left\langle \bar{f}_{\mathbf{v}}, \bar{f}_{\mathbf{u},\mathbf{v}} \right\rangle = 0 \tag{B.4.63}$$

and

$$f(\mathbf{x}) = f_0 + \bar{f}_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) + \bar{f}_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) + \bar{f}_{\mathbf{u},\mathbf{v}}(\mathbf{x}_{\mathbf{u}}, \mathbf{x}_{\mathbf{v}}). \tag{B.4.64}$$

In turn,

$$\mathrm{V}(f) = \mathrm{V}(\bar{f}_{\mathbf{u}}) + \mathrm{V}(\bar{f}_{\mathbf{v}}) + \mathrm{V}(\bar{f}_{\mathbf{u},\mathbf{v}}) \tag{B.4.65}$$

follows from Eqs. (B.4.62)-(B.4.64).

The variance representation in Eq. (B.4.65) provides a way to obtain $S_{iT}$ by assuming that $\mathbf{u}$ consists only of the integer $i$ (i.e., $\mathcal{I}(\mathbf{u}) = \{i\}$) and that $\mathbf{v}$ consists of the remaining integers contained in $\mathcal{I}$ (i.e., $\mathcal{I}(\mathbf{v}) = \{j : j \in \mathcal{I} = \{1, 2, ..., n\} \text{ and } j \neq i\}$ and $\mathcal{X}_{\mathbf{v}} = \mathcal{X}_{\sim i}$). Then, with the indicated definitions for $\mathbf{u}$ and $\mathbf{v}$,

$$S_{iT} = \left(\mathrm{V}(f) - \mathrm{V}(\bar{f}_{\mathbf{v}})\right) / \mathrm{V}(f) = 1 - \mathrm{V}(\bar{f}_{\mathbf{v}}) / \mathrm{V}(f). \tag{B.4.66}$$

The integral-based definition for $\mathrm{V}(\bar{f}_{\mathbf{v}})$ is given by

$$\begin{aligned}
\mathrm{V}(\bar{f}_{\mathbf{v}}) &=_1 \left\langle \bar{f}_{\mathbf{v}}, \bar{f}_{\mathbf{v}} \right\rangle \\
&=_2 \left\langle \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0 - 0, \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0 - 0 \right\rangle \\
&=_3 \left\langle \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0, \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0 \right\rangle \\
&=_4 \left\langle \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(x_i) \mathrm{d}x_i, \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(x_i) \mathrm{d}x_i \right\rangle - 2 f_0 \left\langle 1, \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(x_i) \mathrm{d}x_i \right\rangle + f_0^2 \\
&=_5 \int_{\mathcal{X}_{\mathbf{v}} = \mathcal{X}_{\sim i}} \left( \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d(x_i) \mathrm{d}x_i \times \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d(x_i) \mathrm{d}x_i \right) d(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} - 2 f_0^2 + f_0^2 \\
&=_6 \int_{\mathcal{X}_{\sim i}} \int_{\mathcal{X}_i} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}_{\sim i}, x_i) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(x_i) d(\mathbf{x}_{\sim i}) \mathrm{d}\tilde{x}_i \mathrm{d}x_i \mathrm{d}\mathbf{x}_{\sim i} - f_0^2 \\
&= \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x} - f_0^2 .
\end{aligned} \tag{B.4.67}$$

with (i) Equality 2 obtained from the definition of $\overline{f}_{\mathbf{v}}$ in Eq. (B.4.60) and the equality $\mathrm{E}(\overline{f}_{\mathbf{v}}) = 0$ as indicated in Eq. (B.4.62), (ii) the form of Equality 3 establishing the equality

$$\begin{aligned}
\mathrm{V}_{\mathcal{X}_{\sim i}}\left[\mathrm{E}_{\mathcal{X}_i}(f \mid \mathbf{x}_{\sim i})\right] &= \left\langle \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d(x_i) \mathrm{d}x_i - f_0, \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d(x_i) \mathrm{d}x_i - f_0 \right\rangle \\
&= \left\langle \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0, \int_{\mathcal{X}_{\mathbf{u}}} f(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}) d(\mathbf{x}_{\mathbf{u}}) \mathrm{d}\mathbf{x}_{\mathbf{u}} - f_0 \right\rangle \\
&= \mathrm{V}(\overline{f}_{\mathbf{v}})
\end{aligned} \tag{B.4.68}$$

(iii) Equality 6 obtained as indicated in Eq. (B.3.4), and (iv) the remaining equalities obtained by standard manipulations.

In turn, combining the representation for V($f$) in Eq. (B.4.30) with the representation for $\mathrm{V}(\overline{f}_{\mathbf{v}})$ in Eq. (B.4.67) produces

$$\begin{aligned}
S_{iT} &= \left\{\mathrm{V}(f) - \mathrm{V}(\overline{f}_{\mathbf{v}})\right\} / \mathrm{V}(f) \\
&= \left\{ \left[ \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0^2 \right] \right. \\
&\qquad \left. - \left[ \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x}. - f_0^2 \right] \right\} / \mathrm{V}(f) \\
&= \left\{ \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x} \right\} / \mathrm{V}(f)
\end{aligned} \tag{B.4.69}$$

if the integral representation for V($f$) is retained in the numerator or

$$\begin{aligned}
S_{iT} &= 1 - \mathrm{V}(\overline{f}_{\mathbf{v}}) / \mathrm{V}(f) \\
&= 1 - \left\{ \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x} - f_0^2 \right\} / \mathrm{V}(f)
\end{aligned} \tag{B.4.70}$$

if the integral representation for V($f$) is removed from the numerator by division.

Another equality that provides a representation for $S_{iT}$ is

$$S_{iT} = \mathrm{E}_{\sim i}[\mathrm{V}_i(f \mid \mathbf{x}_{\sim i})] / \mathrm{V}(f), \tag{B.4.71}$$

which can be established as follows. First,

$$
\begin{aligned}
\mathrm{V}_i(f \mid \mathbf{x}_{\sim i}) &= \left\langle f(\mathbf{x} \mid \mathbf{x}_{\sim i}) - \mathrm{E}_i[f(\mathbf{x} \mid \mathbf{x}_{\sim i})], f(\mathbf{x} \mid \mathbf{x}_{\sim i}) - \mathrm{E}_i[f(\mathbf{x} \mid \mathbf{x}_{\sim i})] \right\rangle \\
&= \left\langle f(\mathbf{x} \mid \mathbf{x}_{\sim i}) - \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i, f(\mathbf{x} \mid \mathbf{x}_{\sim i}) - \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right\rangle \\
&= \left\langle f(\mathbf{x} \mid \mathbf{x}_{\sim i}), f(\mathbf{x} \mid \mathbf{x}_{\sim i}) \right\rangle - 2 \left\langle f(\mathbf{x} \mid \mathbf{x}_{\sim i}), \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right\rangle \\
&\quad + \left\langle \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i, \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right\rangle \\
&= \int_{\mathcal{X}_i} f^2(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i - 2 \left[ \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right]^2 + \left[ \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right]^2 \\
&= \int_{\mathcal{X}_i} f^2(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i - \left[ \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right]^2 .
\end{aligned}
\tag{B.4.72}
$$

In turn,

$$
\begin{aligned}
\mathrm{E}_{\sim i}[\mathrm{V}_i(f \mid \mathbf{x}_{\sim i})] &= \int_{\mathcal{X}_{\sim i}} \left( \int_{\mathcal{X}_i} f^2(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i - \left[ \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right]^2 \right) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} \\
&= \int_{\mathcal{X}_{\sim i}} \left( \int_{\mathcal{X}_i} f^2(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} - \int_{\mathcal{X}_{\sim i}} \left( \left[ \int_{\mathcal{X}_i} f(\mathbf{x} \mid \mathbf{x}_{\sim i}) d_i(x_i) \mathrm{d}x_i \right]^2 \right) d_{\sim i}(\mathbf{x}_{\sim i}) \mathrm{d}\mathbf{x}_{\sim i} \\
&= \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{\mathcal{X}_{\sim i}} \int_{\mathcal{X}_i} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}_{\sim i}, x_i) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d_i(\tilde{x}_i) d_i(x_i) d_{\sim i}(\mathbf{x}_{\mathbf{v}}) \mathrm{d}\tilde{x}_i \mathrm{d}x_i \mathrm{d}\mathbf{x}_{\mathbf{v}} \\
&= \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x} \\
&= \mathrm{V}(f) - \mathrm{V}(\overline{f}_{\mathbf{v}}),
\end{aligned}
\tag{B.4.73}
$$

with the final equality following from a match with the numerator in the final equality in Eq. (B.4.69). Thus, from Eqs. (B.4.69) and (B.4.73) ,

$$
S_{iT} = \left\{ \mathrm{V}(f) - \mathrm{V}(\overline{f}_{\mathbf{v}}) \right\} / \mathrm{V}(f) = \mathrm{E}_{\sim i}[\mathrm{V}_i(f \mid \mathbf{x}_{\sim i})] / \mathrm{V}(f). \tag{B.4.74}
$$

with

$$
\mathrm{V}(\overline{f}_{\mathbf{v}}) = \mathrm{V}_{\mathcal{X}_{\sim i}} \left[ \mathrm{E}_{\mathcal{X}_i}(f \mid \mathbf{x}_{\sim i}) \right] \tag{B.4.75}
$$

from Eq. (B.4.68).

In summary, the representations

$$S_{iT} = \begin{cases} \mathrm{E}_{\sim i}[\mathrm{V}_i(f \mid \mathbf{x}_{\sim i})] / \mathrm{V}(f) \\ \left\{ \mathrm{V}(f) - \mathrm{V}_{\mathcal{X}_{\sim i}} \left[ \mathrm{E}_{\mathcal{X}_i}(f \mid \mathbf{x}_{\sim i}) \right] \right\} / \mathrm{V}(f) \\ 1 - \mathrm{V}_{\mathcal{X}_{\sim i}} \left[ \mathrm{E}_{\mathcal{X}_i}(f \mid \mathbf{x}_{\sim i}) \right] / \mathrm{V}(f) \end{cases} \tag{B.4.76}$$

follow from Eqs. (B.4.69), (B.4.73) and (B.4.68).

The variance decomposition associated with Eqs. (B.4.59)-(B.4.65) has potential uses in addition to facilitating the representation and numerical approximation of $S_{Ti}$. For example, an analysis might involve two models (i.e., Model A and Model B), where (i) $\mathbf{x}_\mathbf{u}$ and $\mathbf{x}_\mathbf{v}$ are vectors of uncertain variables that are inputs to Model A and Model B, respectively, and (ii)

$$y = f(\mathbf{x}_\mathbf{u}, \mathbf{x}_\mathbf{v}) = f(x_1, x_2, ..., x_n) \tag{B.4.77}$$

is the analysis result of interest produced by the combined operation of Models A and B. In this situation, (i) $\mathrm{V}(\overline{f}_\mathbf{u})$ and $\mathrm{V}(\overline{f}_\mathbf{v})$ would represent the contributions of Models A and B, respectively, to the uncertainty in $y$, and (ii) $\mathrm{V}(\overline{f}_{\mathbf{u},\mathbf{v}})$ would represent the contribution of interactions involving $\mathbf{x}_\mathbf{u}$ and $\mathbf{x}_\mathbf{v}$, and correspondingly Models A and B, to the uncertainty in $y$.

## B.5 Approximation of Variance Decomposition for $f(x_1, x_2, ..., x_n)$

As discussed in Sect. B.3 for $f(x_1, x_2, x_3)$, the variances $\mathrm{V}(f)$, $\mathrm{V}(f_i)$, $\mathrm{V}(f_{ij})$, … associated with $f(x_1, x_2, ..., x_n)$ will, most likely, need to be approximated with sampling-based procedures. The evaluation of $f_0 = \mathrm{E}(f)$ and V($f$) with sampling-based procedures is straight forward. A random or Latin hypercube sample

$$\mathbf{x}_s = [x_{s1}, x_{s2}, ..., x_{sn}], s = 1, 2, ..., nS, \tag{B.5.1}$$

is generated from $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ in consistency with the probability spaces $(\mathcal{X}_i, \mathbb{X}_i, m_i), i = 1, 2, ..., n,$ and the associated density functions $d_i(x_i), i = 1, 2, ..., n.$ Then, the approximations

$$\hat{f}_0 \cong f_0 \text{ and } \hat{\mathrm{V}}(f) \cong \mathrm{V}(f) \tag{B.5.2}$$

are determined as indicated in Eqs. (B.3.2) and (B.3.3).

As described in Eqs. (B.3.14)-(B.3.18), the variance $\mathrm{V}(f_i)$ can be approximated based on the representation

$$\mathrm{V}(f_i) = \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_{\sim i}} f(\mathbf{x}) f(\tilde{\mathbf{x}}_{\sim i}, x_i) d_{\sim i}(\tilde{\mathbf{x}}_{\sim i}) d(\mathbf{x}) \mathrm{d}\tilde{\mathbf{x}}_{\sim i} \mathrm{d}\mathbf{x} - f_0^2 \tag{B.5.3}$$

in Eq. (B.4.46) with use of the sample in Eq. (B.5.1) and one additional sample. This additional sample

$$\tilde{\mathbf{x}}_s = [\tilde{x}_{s1}, \tilde{x}_{s2}, ..., \tilde{x}_{sn}], s = 1, 2, ..., nS, \tag{B.5.4}$$

is obtained by randomly rearranging the sampled values for each variable. For example, the sequence of values $\tilde{x}_{11}, \tilde{x}_{21}, ..., \tilde{x}_{nS,1}$ for $x_1$ in the sample in Eq. (B.5.4) is simply a random rearrangement of the sequence of values $x_{11}, x_{21}, ..., x_{nS,1}$ for $x_1$ in the sample in Eq.(B.5.1). Then, for each variable $x_i$, a new sample $S_i$ is obtained by reordering the sample elements $\tilde{\mathbf{x}}_s, s = 1, 2, ..., nS,$ to produce a rearranged sequence

$$\tilde{\mathbf{x}}_{is} = [\tilde{x}_{is1}, \tilde{x}_{is2}, ..., \tilde{x}_{isn}], s = 1, 2, ..., nS, \text{ with } \tilde{x}_{isi} = \tilde{x}_{si}. \tag{B.5.5}$$

Then, $\mathrm{V}(f_i)$ is approximated by

$$\begin{aligned} \hat{\mathrm{V}}(f_i) &\cong \sum_{s=1}^{nS} f(\mathbf{x}_s) f(\tilde{\mathbf{x}}_{si}) / nS - \hat{f}_0^2 = \hat{I}_i - \hat{f}_0^2 \\ &= \sum_{s=1}^{nS} f(x_{s1}, ..., x_{si}, ..., x_{sn}) f(\tilde{x}_{is1}, ..., \tilde{x}_{isi}, ..., \tilde{x}_{isn}) / nS - \hat{f}_0^2 = \hat{I}_i - \hat{f}_0^2 \\ &= \sum_{s=1}^{nS} f(x_{s1}, ..., x_{si}, ..., x_{sn}) f(\tilde{x}_{is1}, ..., x_{si}, ..., \tilde{x}_{isn}) / nS - \hat{f}_0^2 = \hat{I}_i - \hat{f}_0^2. \end{aligned} \tag{B.5.6}$$

With this approach, 2 $nS$ evaluations of the function $f(x_1, x_2, ..., x_n)$ are required for approximation of all $n$ of the variances $\mathrm{V}(f_i)$. A less efficient approach is described in conjunction with Eqs. (B.3.10)-(B.3.13) that would require $(n+1)nS$ evaluations of $f(x_1, x_2, ..., x_n)$ for approximation of all $n$ of the variances $\mathrm{V}(f_i)$.

The variances $\mathrm{V}(f_{ij})$ can be approximated on the basis of the representation

$$\mathrm{V}(f_{ij}) = \int_{\mathcal{X}} \int_{\mathcal{X}_{\sim ij}} f(\mathbf{x}) f(\tilde{\mathbf{x}}_{\sim ij}, \mathbf{x}_{ij}) d_{\sim ij}(\tilde{\mathbf{x}}_{\sim ij}) d(\mathbf{x}) \mathrm{d}\tilde{\mathbf{x}}_{\sim ij} \mathrm{d}\mathbf{x} - f_0^2 - \mathrm{V}(f_i) - \mathrm{V}(f_j), \tag{B.5.7}$$

in Eq. (B.4.49). If all $p = n!/2!(n-2)!$ possibilities for the definition of $\mathrm{V}(f_{ij})$ are to be approximated, the starting point in a sampling-based approximation procedure is a random or Latin hypercube sample

$$[\mathbf{x}_s, \tilde{\mathbf{x}}_s] = [x_{s1}, x_{s2}, ..., x_{sn}, \tilde{x}_{s1}, \tilde{x}_{s2}, ..., \tilde{x}_{sn}], s = 1, 2, ..., nS \tag{B.5.8}$$

from $\mathcal{X} \times \tilde{\mathcal{X}}$, with $(\mathcal{X}, \mathbb{X}, m)$ and $(\tilde{\mathcal{X}}, \tilde{\mathbb{X}}, \tilde{m})$ assumed to be identical but independent probability spaces (i.e., both spaces have the same density function $d(\mathbf{x})$ but there are no correlations between the two spaces). Then, in consistency with Eq. (B.5.7),

$$\hat{\mathrm{V}}(f_{ij}) = \sum_{s=1}^{nS} f(\mathbf{x}_s) f(\tilde{\mathbf{x}}_{s\sim ij}, \mathbf{x}_{sij}) / nS - \hat{f}_0^2 - \hat{\mathrm{V}}(f_i) - \hat{\mathrm{V}}(f_j). \tag{B.5.9}$$

If a single (*i*,*j*) pair is under consideration, then the sample from $\mathcal{X} \times \tilde{\mathcal{X}}$ in Eq. (B.5.8) can be replaced by a sample from $\mathcal{X} \times \tilde{\mathcal{X}}_{\sim ij}$, where $\tilde{\mathcal{X}}_{\sim ij}$ corresponds to the space with $x_i$ and $x_j$ omitted.

If desired $\mathrm{V}(f_{\mathbf{u}})$ defined in Eq. (B.4.55) can be approximated in a manner similar that shown for $\mathrm{V}(f_{ij})$ in Eqs. (B.5.7)-(B.5.9). However, it is anticipated that a need for results of this detail (e.g., for $\mathbf{u} = [i, j, k]$) will rarely arise.

The variance contributions $S_{iT}$ can be approximated on the basis of the representation

$$S_{iT} = 1 - \left\{ \int_{\mathcal{X}} \int_{\tilde{\mathcal{X}}_i} f(\mathbf{x}) f(\mathbf{x}_{\sim i}, \tilde{x}_i) d(\tilde{x}_i) d(\mathbf{x}) \mathrm{d}\tilde{x}_i \mathrm{d}\mathbf{x} - f_0^2 \right\} / \mathrm{V}(f) \tag{B.5.10}$$

in Eq. (B.4.70). Given a sample from $\mathcal{X} \times \tilde{\mathcal{X}}$ as illustrated in Eq. (B.5.8), $S_{iT}$ can be approximated by

$$S_{iT} = 1 - \left\{ \sum_{s=1}^{nS} f(\mathbf{x}_s) f(\mathbf{x}_{s\sim i}, \tilde{x}_i) / nS - \hat{f}_0^2 \right\} / \hat{\mathrm{V}}(f). \tag{B.5.11}$$

If $S_{iT}$ is to be determined for a single variable $x_i$, then the sample from $\mathcal{X} \times \tilde{\mathcal{X}}$ in Eq. (B.5.8) can be replaced by a sample from $\mathcal{X} \times \tilde{\mathcal{X}}_i$.

## B.6 Additional Considerations

The required evaluations of a computationally demanding model can be a major challenge in the performance of an uncertainty and sensitivity analysis for this model based on Sobol' variance decomposition. A possible mitigation, but not complete removal, of this challenge can sometimes

be obtained by developing a surrogate model that approximates the model under consideration and then performing a Sobol' variance decomposition for this surrogate model.

A number of approaches to surrogate model development for use in conjunction with Sobol' variance decomposition calculations are described and illustrated in Ref. [11], including Multivariate Adaptive Regression Splines (MARS) [12], Random Forest (RF) procedures [13], Gradient Boosting Machine (GBM) procedures [14], Adaptive Component Selection and Smoothing Operator (ACOSSO) [15], and Gaussian Process (GP) models [16-18]. In addition to several notional examples, an example is presented in Ref. [11] for a result from the 1996 performance assessment for the Waste Isolation Pilot Plant (WIPP) [19].

The procedures presented in this appendix for calculating Sobol' variance decomposition are predicated on the assumption that the uncertain variables under consideration are independent. However, this is often not the case in many real analyses. Thus, it is desirable to have a procedure for including correlated variables in the determination of Sobol' variance decompositions.

Such a procedure has been introduced by A. Nataf [20] and is now commonly referred to as the Nataf transformation. With this transformation, correlated variables are transformed into uncorrelated, normally distributed variables that have structure that is connected to the correlation structure of the original correlated variables. The Nataf transformation is now widely used and an extensive literature exists (e.g., [21-29])

## B.7 References: Appendix B


1. Sobol' I.M. 1993. Sensitivity Analysis for Nonlinear Mathematical Models. *Mathematical Modeling and Computational Experiment* 1:407-414.
2. Homma T. and A. Saltelli. 1996. Importance Measures in Global Sensitivity Analysis of Nonlinear Models. *Reliability Engineering and System Safety* 52:1-17.
3. Sobol' I.M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* 55:271-280.
4. Sobol' I.M. 2003. Theorems and examples on high dimensional model representation. *Reliability Engineering and System Safety* 79:187-193.
5. Chan K., S. Tarantola and A. Saltelli. 2000. Variance-Based Methods. In *Sensitivity Analysis*, ed. A Saltelli, K Chan, EM Scott:167-197. New York, NY: Wiley.
6. Saltelli A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto and S. Tarantola. 2010. Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index. *Computer Physics Communications* 181:259-270.
7. Kucherenko S., S. Tarantola and P. Annoni. 2012. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications* 183:937-946.
8. Mara T.A. and S. Tarantola. 2012. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering and System Safety* 107:115-121.
9. Prieur C. and S. Tarantola. 2017. Variance-based sensitivity analysis: Theory and estimation algorithms. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1217-1239: Springer International Publishing Switzerland.
10. Kucherenko S. and S. Song. 2017. Different numerical estimators for main effect global sensitivity indices. *Reliability Engineering and System Safety* 165:222-238.

11. Storlie C.B., L.P. Swiler, J.C. Helton and C.J. Sallaberry. 2009. Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models. *Reliability Engineering and System Safety* 94:1735-1763.
12. Friedman J.H. 1991. Multivariate Adaptive Regression Splines (with discussion). *The Annals of Statistics* 19:1-141.
13. Breiman L. 2001. Random Forests. *Machine Learning* 45:5-32.
14. Friedman J.H. 2001. Greedy Function Aproximation: A Gradient Boosting Machine. *Annals of Statistics* 29:1189-1232.
15. Storlie C.B., H.D. Bondell, B.J. Reich and H.H. Zhang. 2011. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica* 21:679-705.
16. Sacks J., W.J. Welch, T.J. Mitchell and H.P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science* 4:409-423.
17. Kennedy M.C. and A. O'Hagan. 2001. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 63:425-464.
18. Oakley J.E. and A. O'Hagan. 2004. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 66:751-769.
19. Helton J.C., J.E. Bean, K. Economy, J.W. Garner, R.J. MacKinnon, J. Miller, J.D. Schreiber and P. Vaughn. 2000. Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Disturbed Conditions. *Reliability Engineering and System Safety* 69:263-304.
20. Nataf A. 1962. Determination des distribution don les marges sont donnees. *Comtes rendus de la academie des sciences* 225:42-43.
21. Liu P.L. and A. Der Kiureghian. 1986. Multivariate distribution models with prescribed marginals and covariances. *Probabilistic Engineering Mechanics* 1:105-112.
22. Lebrun R. and A. Dutfoy. 2009. Do Rosenblatt and Nataf isoprobabilistic transformations really differ? *Probabilistic Engineering Mechanics* 24:577-584.
23. Lebrun R. and A. Dutfoy. 2009. An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics* 24:312-320.
24. Lebrun R. and A. Dutfoy. 2009. A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics* 24:172-178.
25. Xiao Q. 2014. Evaluating correlation coefficient for Nataf transformation. *Probabilistic Engineering Mechanics* 37:1-6.
26. Williams M.M.R. 2015. Numerical solution of the Karhunen-Loeve integral equation with examples based on various kernels derived from the Nataf procedure. *Annals of Nuclear Energy* 76:19-26.
27. Jiang P., A. Basudhar and S. Missoum. 2011. Reliability assessment with correlated variables using support vector machines C3 - Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.
28. Bourinet J.M. 2017. FORM sensitivities to distribution parameters with the nataf transformation. *Springer Series in Reliability Engineering*:277-302.

29. Tarantola S. and T.A. Mara. 2017. Variance-based sensitivity indices of computer models with dependent inputs: The fourier amplitude sensitivity test. *International Journal for Uncertainty Quantification* 7:511-523.

# C. Polynomial Chaos

## C.1 Introduction

Function representations obtained with polynomial chaos procedures provide another way to obtain variance decompositions. Like Sobol' variance decomposition, polynomial chaos-based procedures for variance decomposition involve sets of orthonormal functions. However, an important distinction between the two approaches is that (i) the Sobol' approach constructs the set of orthonormal functions used to represent the function of interest from "scratch" (i.e., entirely from consideration of the function under study) and (ii) the polynomial chaos approach starts with sets of predefined orthonormal polynomials that are then used to construct the set of orthonormal functions used to represent the function of interest. Once the desired representation with orthonormal functions is constructed, both approaches are based on a variance decomposition of the form shown in Eqs. (B.2.27) and (B.4.35) of App. B.

As is the case for Sobol' variance decomposition, the notation needed to describe variance decomposition with procedures based on polynomial chaos expansions is complicated and often described in a very compact manner in the journal literature in order to meet the length requirements imposed on journal publications. As a result, it can be difficult for the uninitiated to develop an understanding of exactly what is the mathematical structure that underlies uncertainty and sensitivity analyses based on polynomial chaos procedures. The purpose of this appendix is to lead the interested reader through the details of this structure. To accomplish this, the structure of variance decomposition with polynomial chaos procedures is described separately for (i) the less complicated case involving a function $f(x_1, x_2, x_3)$ of only 3 uncertain variables and (ii) the notationally more complex case involving a function $f(x_1, x_2, ..., x_n)$ of $n$ uncertain variables.

The following topics are considered in this appendix: (i) Introduction to orthonormal polynomials in Sect. C.2, (ii) Variance decomposition for $f(x_1, x_2, x_3)$ with predefined orthonormal polynomials in Sect. C.3, (iii) Variance decomposition for $f(x_1, x_2, ..., x_n)$ with predefined orthonormal polynomials in Sect. C.4, (iv) Variance decomposition for $f(x_1, x_2, x_3)$ without predefined orthonormal polynomials in Sect. C.5, (v) Variance decomposition for $f(x_1, x_2, ..., x_n)$ without predefined orthonormal polynomials in Sect. C.6, and (vi) Approximation of polynomial chaos representations for $f(x_1, x_2, x_3)$ and $f(x_1, x_2, ..., x_n)$ in Sect. C.7.

Additional information on variance decomposition procedures employing polynomial chaos function representations is available in a number of publications (e.g., [1-17]).

## C.2 Introduction to Orthonormal Polynomials

Due to the central role that classical orthogonal polynomials have in the development of polynomial chaos representations for complicated functions and the decomposition of the variance associated with such functions, a brief introduction to orthogonal polynomials is given in this section. Orthogonal polynomials are extensively used in the modeling of many complex systems. As a result, an extensive literature exists where additional information on orthogonal polynomials can be obtained (e.g., Refs. [18-24]).

The polynomials in a set

$$\mathcal{P} = \left\{ p_n(x) : p_n(x) \text{ is a polynomial of degree } n \in \{0,1,2,...\} \text{ and } x \in \mathcal{X} = [\underline{x}, \overline{x}] \right\} \quad \text{(C.2.1)}$$

are orthogonal provided there exists a weight function $w(x)$ with the property that

$$\begin{aligned} \left\langle p_n(x), p_m(x) \right\rangle &= \int_{\mathcal{X}} p_n(x) p_m(x) w(x) \mathrm{d}x \\ &= \delta_{mn} \int_{\mathcal{X}} p_n^2(x) w(x) \mathrm{d}x \\ &= \begin{cases} \int_{\mathcal{X}} p_n^2(x) w(x) \mathrm{d}x & \text{if } m = n \\ 0 & \text{if } m \neq n, \end{cases} \end{aligned} \quad \text{(C.2.2)}$$

where

$$\delta_{mn} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n. \end{cases} \quad \text{(C.2.3)}$$

As a consequence of the linearity of the integral in Eq. (C.2.2), a positive multiple $cw(x)$, $c > 0$, of a weight function is also a weight function. In addition, the polynomials in the set $\mathcal{P}$ are orthonormal if there exists a weight function $\tilde{w}(x)$ with the property that

$$\left\langle p_n(x), p_m(x) \right\rangle = \int_{\mathcal{X}} p_n(x) p_m(x) \tilde{w}(x) \mathrm{d}x = \delta_{mn} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n. \end{cases} \quad \text{(C.2.4)}$$

Variance decomposition employing a polynomial chaos representation for a function (i.e., computer model) $f(x_1, x_2, ..., x_n)$ of variables $x_i, i = 1, 2, ..., n,$ with associated probability spaces $(\mathcal{X}_i, \mathbb{X}_i, m_i)$ and density functions $d_i(x_i)$ involves two initial steps:

Step 1: Determination of a set

$$\mathcal{P}_i = \left\{ p_{in}(x_i) : p_{in}(x_i) \text{ is a polynomial of degree } n \in \{0,1,2,...\} \text{ and } x_i \in \mathcal{X}_i = [\underline{x}_i, \overline{x}_i] \right\} \quad \text{(C.2.5)}$$

of orthogonal polynomials for each variable $x_i$ with the property that the density function $d_i(x_i)$ is a weight function for the polynomials in the set $\mathcal{P}_i$ (i.e., $d_i(x_i)$ satisfies the property indicated for $w(x)$ in Eq. (C.2.2)).

Step 2: Transform (i.e., linearly scale) the polynomials in $\mathcal{P}_i$ to obtain a set

$$\tilde{\mathcal{P}}_i = \left\{ \tilde{p}_{in}(x_i) : \tilde{p}_{in}(x_i) \text{ is a polynomial of degree } n \in \{0,1,2,...\} \text{ and } x_i \in \mathcal{X}_i = [\underline{x}_i, \overline{x}_i] \right\} \quad \text{(C.2.6)}$$

that is orthonormal with the use of $d_i(x_i)$ as a weight function (i.e., the polynomials $\tilde{p}_{in}(x_i)$ in $\tilde{\mathcal{P}}$ and the density function $d_i(x_i)$ satisfy the conditions in Eq. (C.2.4)).

The nature of Step 1 is illustrated in Table C.1. Specifically, Column 1 lists examples of probability distributions that might be considered in a variance decomposition analysis involving polynomial chaos calculations. Then, Columns 2 and 3 list the sample spaces and density functions for the indicated probability distributions. The goal is to determine a set of orthogonal functions for each probability distribution that has (i) the same domain of definition as the sample space of the probability distribution and (ii) a weight function that is linear scaling of the density function of the probability distribution. This goal is met by the orthogonal polynomials indicated in Columns 4, 5 and 6.

Table C.1 Illustration of probability distributions and associated sets of orthogonal polynomials.

| Column 1 Distribution | Column 2 Sample Space $\mathcal{X}$ | Column 3 Density Function $d(x)$ | Column 4 Associated Orthogonal Polynomials | Column5 Example | Column 6 Weight Function $w(x)$ |
|---|---|---|---|---|---|
| Uniform | $\mathcal{X}_1 = [-1,1]$ | $d_1(x) = 1/2$ | Legendre $P_n(x)$ | Table C.2a | $w_1(x) = 1$ on $[-1,1]$ |
| Uniform | $\mathcal{X}_2 = [0,1]$ | $d_2(x) = 1$ | Shifted Legendre $P_n^*(x)$ | Table C.2b | $w_1(x) = 1$ on $[0,1]$ |
| Normal with $\mu = 0, \sigma = 1$ | $\mathcal{X}_3 = (-\infty, \infty)$ | $d_3(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$ | Hermite $He_n(x)$ | Table C.2c | $w_3(x) = \exp(-x^2/2)$ on $(-\infty, \infty)$ |
| Exponential with $\lambda = 1$ | $\mathcal{X}_4 = [0,\infty)$ | $d_4(x) = \exp(-x)$ | Laguerre $L_n(x)$ | Table C.2d | $w_4(x) = \exp(-x)$ on $[0,\infty)$ |
| Gamma with $\alpha > -1, \lambda = 1$ | $\mathcal{X}_5 = [0,\infty)$ | $d_5(x) = \frac{x^\alpha \exp(-x)}{\Gamma(\alpha+1)}$ | Generalized Laguerre $L_n^\alpha(x)$ | Table C.2e | $w_5(x) = x^\alpha \exp(-x)$ on $[0,\infty)$ |
| Beta $(\alpha, \beta, a, b)$ with $\alpha > -1$, $\beta > -1$, $[a,b] = [-1,1]$ | $\mathcal{X}_6 = [-1,1]$ | $d_6(x) = (1-x)^\alpha (1+x)^\beta \times [2^{\alpha+\beta+1} \times B(\alpha+1, \beta+1)]^{-1}$ | Jacobi $P_n^{(\alpha,\beta)}(x)$ | Table C.2f | $w_6(x) = (1-x)^a (1+x)^b$ on $[-1,1]$ |

The general forms for the density functions $d_4(x)$, $d_5(x)$ and $d_6(x)$ are

$$d_4(x) = \lambda \exp(-\lambda x), \tag{C.2.7}$$

$$d_5(x) = \lambda^{\alpha+1} x^\alpha \exp(-\lambda x)/\Gamma(\alpha+1) \text{ with } \alpha > -1, \tag{C.2.8}$$

and

$$d_6(x) = \frac{(b-x)^\alpha (x-a)^\beta}{(b-a)^{\alpha+\beta+1} B(\alpha+1, \beta+1)} \text{ with } \alpha > -1, \beta > -1, a < b \text{ and } B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{C.2.9}$$

As indicated in Table C.1, some of the constants in the density functions in Eqs. (C.2.7)-(C.2.9) are replaced by specific values to match the corresponding orthogonal polynomials (i.e., $L_n(x)$, $L_n^\alpha(x)$ and $P_n^{(\alpha,\beta)}(x)$).

Table C.2 Examples of the orthogonal polynomials introduced in Table C.1.

| C.2a: Legendre $P_n(x)$ | C.2b: Shifted Legendre $P_n^*(x)$ |
|---|---|
| $P_0(x)=1$ | $P_0^*(x)=1$ |
| $P_1(x)=x$ | $P_1^*(x)=2x-1$ |
| $P_2(x)=(3x^2-1)/2$ | $P_2^*(x)=6x^2-6x+1$ |
| $P_3(x)=(5x^3-3x)/2$ | $P_3^*(x)=20x^3-30x^2+12x-1$ |
| $P_4(x)=(35x^4-30x^2+3)/8$ | $P_4^*(x)=70x^4-140x^3+90x^2-20x+1$ |
| $P_5(x)=(63x^5-70x^3+15x)/8$ | $P_5^*(x)=252x^5-630x^4+560x^3-210x^2+30x-1$ |

| C.2c: Hermite (0,1) $He_n(x)$ | C.2d: Laguerre $L_n(x)$ |
|---|---|
| $He_0(x)=1$ | $L_0(x)=1$ |
| $He_1(x)=x$ | $L_1(x)=-x+1$ |
| $He_2(x)=x^2-1$ | $L_2(x)=(x^2-4x+2)/2$ |
| $He_3(x)=x^3-3x$ | $L_3(x)=(-x^3+9x^2-18x+6)/6$ |
| $He_4(x)=x^4-6x^2+3$ | $L_4(x)=(x^4-16x^3+72x^2-96x+24)/24$ |
| $He_5(x)=x^5-10x^3+15x$ | $L_5(x)=(-x^5+25x^4-200x^3+600x^2-600x+120)/120$ |

| C.2e: Generalized Laguerre $L_n^{\alpha}(x)$ |
|---|
| $L_0^{\alpha}(x)=1$ |
| $L_1^{\alpha}(x)=-x+\alpha+1$ |
| $L_2^{\alpha}(x)=x^2/2-(\alpha+2)x+(\alpha+2)(\alpha+1)/2$ |
| $L_3^{\alpha}(x)=-x^3/6+(\alpha+3)x^2/2-(\alpha+2)(\alpha+3)x/2+(\alpha+1)(\alpha+2)(\alpha+3)/6$ |
| |
| C.2f: Jacobi $P_n^{(\alpha,\beta)}(x)$ |
| $P_0^{(\alpha,\beta)}(x)=1$ |
| $P_1^{(\alpha,\beta)}(x)=(\alpha+1)+(\alpha+\beta+2)(x-1)/2$ |
| $P_2^{(\alpha,\beta)}(x)=(\alpha+1)(\alpha+2)/2+(\alpha+2)(\alpha+\beta+3)(x-1)/2+(\alpha+\beta+3)(\alpha+\beta+4)(x-1)^2/8$ |

More generally, the polynomials in Table C.2 are defined by

$$P_n(x) = \frac{1}{2}\sum_{k=0}^{n}\binom{n}{k}^2 (x-1)^{n-k}(x+1)^k, \tag{C.2.10}$$

$$P_n^*(x) = P_n(2x-1) = (-1)^n \sum_{k=0}^{n}\binom{n}{k}\binom{n+k}{k}(-x)^k, \tag{C.2.11}$$

$$He_n(x) = n!\sum_{k=0}^{[n/2]}\frac{(-1)^k}{k!(n-2k)!}\frac{x^{n-2k}}{2^k} \quad \text{with } [n/2] = \begin{cases} n/2 & \text{for } n \text{ even} \\ (n-1)/2 & \text{for } n \text{ odd,} \end{cases} \tag{C.2.12}$$

$$L_n(x) = \sum_{k=0}^{n}\binom{n}{k}\frac{(-1)^k}{k!}x^k, \tag{C.2.13}$$

$$L_n^{\alpha}(x) = \sum_{k=0}^{n}(-1)^k\binom{n+\alpha}{n-k}\frac{x^k}{k!}, \tag{C.2.14}$$

$$P_n^{(\alpha,\beta)}(x) = \sum_{k=0}^{n}\binom{n+\alpha}{n-k}\binom{n+\beta}{k}\left(\frac{x-1}{2}\right)^k\left(\frac{x+1}{2}\right)^{n-k}. \tag{C.2.15}$$

Step 2 involves transforming the sets of orthogonal polynomials indicated in Table C.1 and Table C.2 into sets of orthonormal polynomials with respect to the density functions indicated in Table C.1. For notational purposes, the set $\mathcal{P}_i$ of orthogonal polynomials defined in Eq. (C.2.5) is assumed to correspond to a set of orthogonal polynomials associated with the probability space for variable $x_i$ with sample space $\mathcal{X}_i$ and density function $d_i(x_i)$. Then,

$$\langle p_{in}(x_i), p_{in}(x_i)\rangle = \int_{\mathcal{X}_i} p_{in}(x_i)p_{in}(x_i)d_i(x_i) = \int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i) \tag{C.2.16}$$

and the desired normalization of the polynomials in $\mathcal{P}_i$ is given by

$$\tilde{p}_{in}(x_i) = \frac{p_{in}(x_i)}{\sqrt{\langle p_{in}(x_i), p_{in}(x_i)\rangle}} = \frac{p_{in}(x_i)}{\sqrt{\int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i)}} \tag{C.2.17}$$

In turn, the desired normalization is verified by

$$\begin{aligned}\left\langle \tilde{p}_{in}(x_i), \tilde{p}_{in}(x_i) \right\rangle &= \int_{\mathcal{X}_i} \tilde{p}_{in}(x_i)\tilde{p}_{in}(x_i)d_i(x_i)\mathrm{d}x_i \\ &= \int_{\mathcal{X}_i} \frac{p_{in}(x_i)}{\sqrt{\int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i)}} \times \frac{p_{in}(x_i)}{\sqrt{\int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i)}} \times d_i(x_i)\mathrm{d}x_i \\ &= \frac{\int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i)}{\int_{\mathcal{X}_i} p_{in}^2(x_i)d_i(x_i)} \\ &= 1.\end{aligned} \tag{C.2.18}$$

For the polynomials introduced in Table C.1, the square of the normalization devisors to produce the desired sets of orthonormal polynomials are defined as follows:

$$\begin{aligned}\text{Legendre } P_n(x): \int_{-1}^{1} P_n(x)P_n(x)d_1(x)\mathrm{d}x &= \int_{-1}^{1} P_n(x)P_n(x)(1/2)\mathrm{d}x \\ &= (1/2)\int_{-1}^{1} P_n(x)P_n(x)(1)\mathrm{d}x \\ &= \frac{1}{2} \times \frac{2}{2n+1} = \frac{1}{2n+1},\end{aligned} \tag{C.2.19}$$

$$\text{Shifted Legendre } P_n^*(x): \int_0^1 P_n^*(x)P_n^*(x)d_1(x)\mathrm{d}x = \int_0^1 P_n^*(x)P_n^*(x)(1)\mathrm{d}x = \frac{1}{2n+1}, \tag{C.2.20}$$

$$\begin{aligned}\text{Normal } He_n(x): \int_{-\infty}^{\infty} He_n(x)He_n(x)d_3(x)\mathrm{d}x &= \int_{-\infty}^{\infty} He_n(x)He_n(x)\exp(-x^2/2)\big/\sqrt{2\pi}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} He_n(x)He_n(x)\exp(-x^2/2)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \times n!\sqrt{2\pi} = n!,\end{aligned} \tag{C.2.21}$$

$$\text{Laguerre } L_n(x): \int_0^{\infty} L_n(x)L_n(x)d_4(x)\mathrm{d}x = \int_0^{\infty} L_n(x)L_n(x)\exp(-x)\mathrm{d}x = 1. \tag{C.2.22}$$

$$\begin{aligned}\text{Generalized Laguerre } L_n^{\alpha}(x): \int_0^{\infty} L_n^{\alpha}(x)L_n^{\alpha}(x)d_5(x)\mathrm{d}x &= \int_0^{\infty} L_n^{\alpha}(x)L_n^{\alpha}(x)\frac{x^{\alpha}\exp(-x)}{\Gamma(\alpha+1)}\mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha+1)}\int_0^{\infty} L_n^{\alpha}(x)L_n^{\alpha}(x)x^{\alpha}\exp(-x)\mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha+1)} \times \frac{\Gamma(n+\alpha+1)}{n!} = \frac{\Gamma(n+\alpha+1)}{n!\Gamma(\alpha+1)},\end{aligned} \tag{C.2.23}$$

$$
\begin{aligned}
\text{Jacobi } P_n^{(\alpha,\beta)}(x) : & \int_{-1}^{1} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x) d_6(x)\mathrm{d}x \\
&= \int_{-1}^{1} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x) \frac{(1-x)^{\alpha}(1+x)^{\beta}}{2^{\alpha+\beta+1} B(\alpha+1,\beta+1)}\mathrm{d}x \\
&= \frac{1}{2^{\alpha+\beta+1} B(\alpha+1,\beta+1)} \int_{-1}^{1} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x)(1-x)^{\alpha}(1+x)^{\beta}\mathrm{d}x \\
&= \frac{1}{2^{\alpha+\beta+1} B(\alpha+1,\beta+1)} \times \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)n!} \\
&= \frac{1}{n!(2n+\alpha+\beta+1)} \frac{\Gamma(\alpha+\beta+2)}{\Gamma(n+\alpha+\beta+1)} \frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)} \frac{\Gamma(n+\beta+1)}{\Gamma(\beta+1)} \\
&= h_n^2 .
\end{aligned}
\tag{C.2.24}
$$

With use of the results in Eqs. (C.2.19)-(C.2.24), the desired orthonormal polynomials for use in conjunction with the density functions for the probability distributions in Table C.1 are

$$
\text{Uniform on } [-1,1] : \tilde{P}_n(x) = P_n(x) \Big/ \sqrt{1/(2n+1)} = P_n(x)\sqrt{2n+1}, \tag{C.2.25}
$$

$$
\text{Uniform on } [0,1] : \tilde{P}_n^*(x) = P_n^*(x) \Big/ \sqrt{1/(2n+1)} = P_n^*(x)\sqrt{2n+1}, \tag{C.2.26}
$$

$$
\text{Normal with } \mu = 0, \sigma = 1 \text{ on } [-\infty,\infty] : \tilde{He}_n(x) = He_n(x)/\sqrt{n!}, \tag{C.2.27}
$$

$$
\text{Laguerre with } \lambda = 1 \text{ on } [0,\infty] : \tilde{L}_n(x) = L_n(x) \Big/ \sqrt{1} = L_n(x), \tag{C.2.28}
$$

$$
\begin{aligned}
\text{Generalized Laguerre with } \lambda = 1 \text{ on } [0,\infty] : \tilde{L}_n^{\alpha}(x) &= \frac{L_n^{\alpha}(x)}{\sqrt{\Gamma(n+\alpha+1)/n!\Gamma(\alpha+1)}} \\
&= \frac{L_n^{\alpha}(x)\sqrt{n!\Gamma(\alpha+1)}}{\sqrt{\Gamma(n+\alpha+1)}},
\end{aligned}
\tag{C.2.29}
$$

$$
\text{Beta on } [a,b] = [-1,1] \text{ with } h_n \text{ from Eq. (5.24): } \tilde{P}_n^{(\alpha,\beta)}(x) = P_n^{(\alpha,\beta)}(x)/h_n . \tag{C.2.30}
$$

The following sections will describe the use of orthonormal polynomials is the determination of variance decomposition results.

## C.3 Variance Decomposition for $f(x_1, x_2, x_3)$ with Predefined Orthonormal Polynomials

The polynomial chaos approach to variance decomposition starts with sets of predefined orthonormal functions. As a motivational example, the previously introduced function $f(\mathbf{x})$ is used, with (i) $\mathbf{x} = [x_1, x_2, x_3]$, (ii) the epistemically uncertain variables $x_1$, $x_2$ and $x_3$ assumed to

be independent, and (iii) the uncertainty in $x_1$, $x_2$ and $x_3$ characterized by the probability spaces $(\mathcal{X}_1, \mathbb{X}_1, m_1)$, $(\mathcal{X}_2, \mathbb{X}_2, m_2)$ and $(\mathcal{X}_3, \mathbb{X}_3, m_3)$ with $\mathcal{X}_i = [\underline{x}_i, \overline{x}_i]$ and associated density functions $d_1(x_1)$, $d_2(x_2)$ and $d_3(x_3)$. Further, each uncertain variable $x_i, i = 1,2,3,$ is assumed to have an associated set

$$\mathcal{P}_i = \{p_{id}(x_i) : p_{id}(x_i) \text{ is a polynomial of degree } d \text{ for } x_i \in \mathcal{X}_i \text{ and } d = 0,1,2,...\} \quad \text{(C.3.1)}$$

of orthonormal polynomials with inner products defined by

$$\begin{aligned} \left\langle p_{ir}(x_i), p_{is}(x_i) \right\rangle &= \int_{\mathcal{X}_i} p_{ir}(x_i) p_{is}(x_i) d_i(x_i) \mathrm{d}x_i \quad \text{for } \mathcal{I} = \{0,1,2,3,....\} \text{ and } (r,s) \in \mathcal{I} \times \mathcal{I} \\ &= \begin{cases} 1 & \text{if } r = s \\ 0 & \text{if } r \neq s. \end{cases} \end{aligned} \quad \text{(C.3.2)}$$

As a reminder, the second equality in Eq. (C.3.2) is simply a statement of what it means for the polynomials in the set $\mathcal{P}_i$ to be orthonormal with respect to the inner product defined in the first equality in Eq. (C.3.2). Examples of possible definitions for the set $\mathcal{P}_i$ are shown in conjunction with Table C.1 and Table C.2.

As a consequence of the orthonormality of the polynomials contained in $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$, the multivariable polynomials contained in the set

$$\begin{aligned} \mathcal{P} = \{p_{rst}(\mathbf{x}) : p_{rst}(\mathbf{x}) = p_{1r}(x_1) p_{2s}(x_2) p_{3t}(x_3) \text{ for } p_{1r}(x_1) \in \mathcal{P}_1, p_{2s}(x_2) \in \mathcal{P}_2, \\ p_{3t}(x_3) \in \mathcal{P}_3, \text{ and } (r,s,t) \in \mathcal{I}_3 = \mathcal{I} \times \mathcal{I} \times \mathcal{I}\} \end{aligned} \quad \text{(C.3.3)}$$

are also orthonormal for the inner product

$$\begin{aligned} &\left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle = \int_{\mathcal{X}} p_{rst}(\mathbf{x}) p_{uvw}(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \quad \text{for } (r,s,t), (u,v,w) \in \mathcal{I}_3 = \mathcal{I} \times \mathcal{I} \times \mathcal{I} \\ &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} p_{1r}(x_1) p_{2s}(x_2) p_{3t}(x_3) \times p_{1u}(x_1) p_{2v}(x_2) p_{3w}(x_3) d_3(x_3) d_2(x_2) d_1(x_1) \mathrm{d}x_3 \mathrm{d}x_2 \mathrm{d}x_1 \\ &= \int_{\mathcal{X}_1} p_{1r}(x_1) p_{1u}(x_1) d_1(x_1) \mathrm{d}x_1 \times \int_{\mathcal{X}_2} p_{2s}(x_2) p_{2v}(x_2) d_2(x_2) \mathrm{d}x_2 \times \int_{\mathcal{X}_3} p_{3t}(x_3) p_{3w}(x_3) d_3(x_3) \mathrm{d}x_3 \\ &= \begin{cases} 1 \times 1 \times 1 = 1 & \text{if } p_{rst}(\mathbf{x}) = p_{uvw}(\mathbf{x}), \text{ i.e.; all integrals = 1 from normality} \\ 0 & \text{if } p_{rst}(\mathbf{x}) \neq p_{uvw}(\mathbf{x}), \text{ i.e., at least one integral = 0 from orthogonality.} \end{cases} \end{aligned} \quad \text{(C.3.4)}$$

The last equality in Eq. (C.3.4) establishes that the multivariable polynomials contained in the set $\mathcal{P}$ defined in Eq. (C.3.3) orthonormal with respect to the indicated inner product.

The core idea in the polynomial chaos approach to variance decomposition is to represent $f(\mathbf{x})$ by

$$\begin{aligned} f(\mathbf{x}) &= \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{rst}(\mathbf{x}) \quad \text{for } \mathcal{I}^3 = \mathcal{I}\times\mathcal{I}\times\mathcal{I} \\ &= \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{1r}(x_1)\, p_{2s}(x_2)\, p_{3t}(x_3) \\ &= f_0 + \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} c_{rst}\, p_{1r}(x_1)\, p_{2s}(x_2)\, p_{3t}(x_3), \end{aligned} \tag{C.3.5}$$

where (i) $c_{rst}$ is an appropriately determined coefficient as indicated in Eqs. (C.3.8) - (C.3.10), and (ii) the summation notation $(r,s,t)\neq(0,0,0)\in\mathcal{I}^3$ in the last equality indicates that the summation is over all $(r,s,t)\in\mathcal{I}^3$ except $(0,0,0)$. Specifically, the variance V(*f*) of $f$ is given by

$$\begin{aligned} \mathrm{V}(f) &= \mathrm{E}\left(\left[f(\mathbf{x}) - \mathrm{E}(f)\right]^2\right) \\ &= \int_{\mathcal{X}} \left[f(\mathbf{x}) - f_0\right]^2 d(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{X}} f^2(\mathbf{x}) d(\mathbf{x})\mathrm{d}\mathbf{x} - f_0^2 \\ &= \left\langle f(\mathbf{x}), f(\mathbf{x})\right\rangle - f_0^2 \\ &= \left\langle \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{rst}(\mathbf{x}), \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}\, p_{rst}(\mathbf{x}) \right\rangle - f_0^2 \\ &= \sum_{(r,s,t)\in\mathcal{I}^3} \sum_{(u,v,w)\in\mathcal{I}^3} c_{rst}\, c_{uvw} \left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x})\right\rangle - f_0^2 \\ &= \sum_{(r,s,t)\in\mathcal{I}^3} c_{rst}^2 - f_0^2 \\ &= \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} c_{rst}^2, \end{aligned} \tag{C.3.6}$$

with (i) the next to last equality following from the equalities

$$\left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x})\right\rangle = \begin{cases} 1 & \text{for } (r,s,t) = (u,v,w) \\ 0 & \text{for } (r,s,t) \neq (u,v,w) \end{cases} \tag{C.3.7}$$

and (ii) the final equality following from the equality $c_{000} = f_0$ as shown in Eq. (C.3.9).

The coefficients $c_{rst}$ in Eqs. (C.3.5) and (C.3.6) are defined by

$$
\begin{aligned}
\left\langle p_{rst}(\mathbf{x}), f(\mathbf{x}) \right\rangle &= \int_{\mathcal{X}} p_{rst}(\mathbf{x}) f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} p_{rst}(\mathbf{x}) \left( \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw} p_{uvw}(\mathbf{x}) \right) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw} \int_{\mathcal{X}} p_{rst}(\mathbf{x}) p_{uvw}(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \sum_{(u,v,w)\in\mathcal{I}^3} c_{uvw} \left\langle p_{rst}(\mathbf{x}), p_{uvw}(\mathbf{x}) \right\rangle \\
&= c_{rst} \left\langle p_{rst}(\mathbf{x}), p_{rst}(\mathbf{x}) \right\rangle \\
&= c_{rst},
\end{aligned}
\tag{C.3.8}
$$

with the final equality following from the equalities in Eq. (C.3.7). For $(r,s,t) = (0,0,0)$, the corresponding orthonormal polynomial of degree 0 is $p_{000}(\mathbf{x}) = 1$ with an associated coefficient $c_{000}$ defined by

$$
c_{000} = \left\langle p_{000}(\mathbf{x}), f(\mathbf{x}) \right\rangle = \int_{\mathcal{X}} p_{000}(\mathbf{x}) f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathrm{E}[f(\mathbf{x})] = f_0. \tag{C.3.9}
$$

As shown in Eq. (C.3.8), the remaining coefficients are defined by

$$
c_{rst} = \int_{\mathcal{X}} p_{rst}(\mathbf{x}) f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{C.3.10}
$$

with the indicated integral requiring numerical evaluation with quadrature or sampling-based methods. An alternate evaluation procedure based on regression analysis techniques is described in Sect.C.7.

The previously defined variances in Eqs. (B.2.35)-(B.2.46) of App. B for $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ can also be represented by appropriates sums of subsets of the squared coefficients $c_{rst}^2$ in the representation for V(*f*) in Eq. (C.3.6). The representation for $\mathrm{V}(f_1)$ is considered first. To start, $f_1(x_1)$ can be represented by

$$
f_1(x_1) = \sum_{r=0}^{\infty} c_r p_{1r}(x_1) = \sum_{r=1}^{\infty} c_r p_{1r}(x_1) = \sum_{r=1}^{\infty} c_{r00} p_{1r}(x_1), \tag{C.3.11}
$$

where (i) $p_{1r}(x_1)$, $r = 0,1,2,3,...,$ correspond to the set of orthonormal polynomials defined in conjunction with Eq. (C.3.1) for the variable $x_1$ and (ii) *r* rather than *d* is used to represent the degree of the polynomial $p_{1r}(x_1)$ to match with notational use of *r*, *s* and *t* to represent the degrees of the polynomials $p_{1r}(x_1)$, $p_{2s}(x_2)$ and $p_{3t}(x_3)$ in Eq. (C.3.5) and elsewhere. Specifically,

$$
\begin{aligned}
c_r &= \left\langle p_{1r}(x_1), f_1(x_1) \right\rangle \\
&= \int_{\mathcal{X}_1} p_{1r}(x_1) f_1(x_1) d_1(x_1) \mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1} p_{1r}(x_1) \left[ \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f_1(x_1, x_2, x_3) d_3(x_3) d_2(x_2) \mathrm{d}x_3 \mathrm{d}x_2 - f_0 \right] d_1(x_1) \mathrm{d}x_1 \\
&= \int_{\mathcal{X}} p_{1r}(x_1) f_1(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0 \int_{\mathcal{X}} p_{1r}(x_1) d_1(x_1) \mathrm{d}x_1 \\
&= \begin{cases} 0 \text{ for } r = 0 \text{ from } f_0 - f_0 = 0 \\ \int_{\mathcal{X}} p_{1r}(x_1) f_1(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \text{ for } r > 0 \text{ from } \int_{\mathcal{X}} p_{1r}(x_1) d_1(x_1) \mathrm{d}x_1 = 0 \end{cases} \\
&= c_{r00} \text{ in the notation of Eq. (C.3.5).}
\end{aligned}
\tag{C.3.12}
$$

In turn,

$$
\begin{aligned}
\mathrm{V}(f_1) &= \left\langle f_1 - \mathrm{E}(f_1), f_1 - \mathrm{E}(f_1) \right\rangle \\
&= \left\langle \sum_{r=1}^{\infty} c_r p_{1r}(x_1), \sum_{r=1}^{\infty} c_r p_{1r}(x_1) \right\rangle \text{ from Eq. (C.3.11) and } \mathrm{E}(f_1) = 0 \\
&= \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_r c_s \left\langle p_{1r}(x_1), p_{1s}(x_1) \right\rangle \\
&= \sum_{r=1}^{\infty} c_r^2 \text{ from orthonormality} \\
&= \sum_{r=1}^{\infty} c_{r00}^2,
\end{aligned}
\tag{C.3.13}
$$

with the final equality following from Eq. (C.3.12). Thus, the variance $\mathrm{V}(f_1)$ of $f_1(x_1)$ can be obtained from a sum of squared coefficients obtained in a polynomial expansion of $f(x_1, x_2, x_3)$. Similar derivations for $f_2(x_2)$ and $f_2(x_3)$ produce

$$
f_2(x_2) = \sum_{s=1}^{\infty} c_{0s0} p_{2s}(x_2) \text{ and } \mathrm{V}(f_2) = \sum_{s=1}^{\infty} c_{0s0}^2 \tag{C.3.14}
$$

and

$$
f_3(x_3) = \sum_{t=1}^{\infty} c_{00t} p_{3t}(x_3) \text{ and } \mathrm{V}(f_3) = \sum_{t=1}^{\infty} c_{00t}^2, \tag{C.3.15}
$$

with subscripts *s* and *t* used for consistency with Eqs. (C.3.5) and (C.3.6).

The variance $V(f_{12})$ of $f_{12}(x_1, x_2)$ obtained with polynomial chaos techniques is now considered. First, the following expansion for $f_{12}(x_1, x_2)$ is developed:

$$f_{12}(x_1, x_2) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} c_{rs} p_{1r}(x_1) p_{2s}(x_2). \tag{C.3.16}$$

Initially, $f_{12}(x_1, x_2)$ can be represented by

$$\begin{aligned} f_{12}(x_1, x_2) &= \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 - f_1(x_1) - f_2(x_2) - f_0 \\ &= \left[ \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 - f_0 \right] - \sum_{r=1}^{\infty} c_{r00} p_{1r}(x_1) - \sum_{s=1}^{\infty} c_{0s0} p_{2s}(x_2) \end{aligned} \tag{C.3.17}$$

with (i) the first equality corresponding to the definition of $f_{12}(x_1, x_2)$ in Eq. (B.2.6) and the second equality following from the representations

$$f_1(x_1) = \sum_{r=1}^{\infty} c_{r00} p_{1r}(x_1) \text{ and } f_2(x_2) = \sum_{r=1}^{\infty} c_{0s0} p_{2s}(x_2). \tag{C.3.18}$$

Next, the representation

$$\begin{aligned} \int_{\mathcal{X}_3} f(x_1, x_2, x_3) d_3(x_3) \mathrm{d}x_3 - f_0 &= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} c_{rs0} p_{1r}(x_1) p_{2s}(x_2) \\ &= \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} c_{rs0} p_{1r}(x_1) p_{2s}(x_2) + \sum_{r=1}^{\infty} c_{r00} p_{1r}(x_1) + \sum_{s=1}^{\infty} c_{0s0} p_{2s}(x_2) \end{aligned} \tag{C.3.19}$$

is obtained with

$$
\begin{aligned}
c_{rs} &= \left\langle p_{1r}(x_1)p_{2s}(x_2), \int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)\mathrm{d}x_3 - f_0 \right\rangle \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} p_{1r}(x_1)p_{2s}(x_2)\left[\int_{\mathcal{X}_3} f(x_1,x_2,x_3)d_3(x_3)\mathrm{d}x_3 - f_0\right] d_2(x_2)d_1(x_1)\mathrm{d}x_2\mathrm{d}x_1 \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} p_{1r}(x_1)p_{2s}(x_2)p_{30}(x_3)f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \\
&\qquad - f_0\int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} p_{1r}(x_1)p_{2s}(x_2)p_{30}(x_3)d_3(x_3)d_2(x_2)d_1(x_1)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \\
&= \begin{cases} c_{000} = f_0 - f_0 = 0 \ \text{ for } \ r = s = 0 \\ c_{r00} - f_0 \times 0 = c_{r00} \ \text{ for } \ r > 0, s = 0 \\ c_{0s0} - f_0 \times 0 = c_{0s0} \ \text{ for } \ r > 0, s = 0 \\ c_{rs0} - f_0 \times 0 = c_{rs0} \ \text{ for } \ r > 0, s > 0. \end{cases}
\end{aligned}
\tag{C.3.20}
$$

Finally, combining the results in Eqs. (C.3.17) and (C.3.19) produces

$$
\begin{aligned}
f_{12}(x_1,x_2) &= \left[\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}p_{1r}(x_1)p_{2s}(x_2) + \sum_{r=1}^{\infty} c_{r00}p_{1r}(x_1) + \sum_{s=1}^{\infty} c_{0s0}p_{2s}(x_2)\right] \\
&\qquad - \sum_{r=1}^{\infty} c_{r00}p_{1r}(x_1) - \sum_{s=1}^{\infty} c_{0s0}p_{2s}(x_2) \\
&= \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}p_{1r}(x_1)p_{2s}(x_2).
\end{aligned}
\tag{C.3.21}
$$

In turn, the variance $\mathrm{V}(f_{12})$ of $f_{12}(x_1,x_2)$ is given by

$$
\begin{aligned}
\mathrm{V}(f_{12}) &= \left\langle f_{12}(x_1,x_2) - \mathrm{E}[f_{12}(x_1,x_2)], f_{12}(x_1,x_2) - \mathrm{E}[f_{12}(x_1,x_2)] \right\rangle \\
&= \left\langle f_{12}(x_1,x_2), f_{12}(x_1,x_2) \right\rangle \ \text{ from } \ \mathrm{E}[f_{12}(x_1,x_2)] = 0 \\
&= \left\langle \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}p_{1r}(x_1)p_{2s}(x_2), \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}p_{1r}(x_1)p_{2s}(x_2) \right\rangle \\
&= \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}^2,
\end{aligned}
\tag{C.3.22}
$$

with the final equality following from the orthonormality of $p_{1r}(x_1)$ and $p_{2s}(x_2)$. By derivations similar to those in Eqs. (C.3.16)-(C.3.22), the variances for $f_{13}(x_1,x_3)$ and $f_{23}(x_1,x_3)$ are given by

$$f_{13}(x_1,x_3)=\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}\,p_{1r}(x_1)p_{3t}(x_3)\ \text{ and }\ \mathrm{V}(f_{13})=\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}^2 \tag{C.3.23}$$

and

$$f_{23}(x_2,x_3)=\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}\,p_{2s}(x_2)p_{3t}(x_3)\ \text{ and }\ \mathrm{V}(f_{23})=\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}^2. \tag{C.3.24}$$

The variance $\mathrm{V}(f_{123})$ of $f_{123}(x_1,x_2,x_3)$ obtained with polynomial chaos techniques is now considered. First, the following expansion for $f_{123}(x_1,x_2,x_3)$ is developed:

$$f_{123}(x_1,x_2,x_3)=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\sum_{t=0}^{\infty}c_{rst}\,p_{1r}(x_1)p_{2s}(x_2)p_{3t}(x_3). \tag{C.3.25}$$

Initially, $f_{123}(x_1,x_2,x_2)$ can be represented by

$$\begin{aligned}
f_{123}(x_1,x_2,x_3) &= f(x_1,x_2,x_3)-f_0-f_1(x_1)-f_2(x_2)-f_3(x_3)\\
&\quad -f_{12}(x_1,x_2)-f_{13}(x_1,x_3)-f_{23}(x_2,x_3)\\
&=\left[f_0+\sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3}c_{rst}\,p_{1r}(x_1)p_{2s}(x_2)p_{3t}(x_3)\right]-f_0\\
&\quad -\sum_{r=1}^{\infty}c_{r00}\,p_{1r}(x_1)-\sum_{s=1}^{\infty}c_{0s0}\,p_{2s}(x_2)-\sum_{s=1}^{\infty}c_{00t}\,p_{3t}(x_3)\\
&\quad -\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}c_{rs0}-\sum_{r=1}^{\infty}\sum_{t=1}^{\infty}c_{r0t}-\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{0st}\\
&=\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}c_{rst}\,p_{1r}(x_1)p_{2s}(x_2)p_{3t}(x_3),
\end{aligned} \tag{C.3.26}$$

with (i) the first equality corresponding to the definition of $f_{123}(x_1,x_2,x_3)$ in Eq. (B.2.9) of App. B and (ii) the second equality following from the representations for $f(x_1,x_2,x_3)$ in Eq. (C.3.5), $f_1(x_1)$, $f_2(x_2)$ and $f_3(x_3)$ in Eqs. (C.3.18), (C.3.14) and (C.3.15), and $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$ and $f_{23}(x_2,x_3)$ in Eqs.(C.3.21), (C.3.23) and (C.3.24), and (iii) third equality resulting from differences involving equal terms. Next, the variance $\mathrm{V}(f_{123})$ of $f_{123}(x_1,x_2,x_3)$ is given by

$$
\begin{aligned}
\mathrm{V}(f_{123}) &= \left\langle f_{123}(x_1,x_2,x_3) - \mathrm{E}(f_{123}), f_{123}(x_1,x_2,x_3) - \mathrm{E}(f_{123}) \right\rangle \\
&= \left\langle f_{123}(x_1,x_2,x_3), f_{123}(x_1,x_2,x_3) \right\rangle \quad \text{from} \quad \mathrm{E}(f_{123}) = 0 \\
&= \left\langle \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}\, p_{1r}(x_1)\, p_{2s}(x_2)\, p_{3t}(x_3), \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}\, p_{1r}(x_1)\, p_{2s}(x_2)\, p_{3t}(x_3) \right\rangle \\
&= \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}^2,
\end{aligned}
\tag{C.3.27}
$$

with the final equality following from the orthonormality of $p_{1r}(x_1)$, $p_{2s}(x_2)$ and $p_{3t}(x_3)$..

The polynomial chaos variance decomposition of $f(x_1,x_2,x_3)$ is based on the equality

$$
\begin{aligned}
\mathrm{V}(f) &=_1 \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} c_{rst}^2 \\
&=_2 \sum_{r=1}^{\infty} c_{r00}^2 + \sum_{s=1}^{\infty} c_{0s0}^2 + \sum_{t=1}^{\infty} c_{00t}^2 + \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}^2 \\
&\qquad + \sum_{r=1}^{\infty}\sum_{t=1}^{\infty} c_{r0t}^2 + \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{0st}^2 + \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}^2 \\
&=_3 \mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23}) + \mathrm{V}(f_{123}),
\end{aligned}
\tag{C.3.28}
$$

with (i) Equality 1 obtained from Eq. (C.3.6), (ii) Equality 2 obtained by a rearrangement of the summation in Equality 1, and (iii) Equality 3 obtained from the variances defined in Eqs. (C.3.13) , (C.3.14), (C.3.15), (C.3.22), (C.3.23), (C.3.24) and (C.3.27). In turn, as previously discussed in connection with Eqs. (3.28)-(3.32), division of both sides of Eq. (C.3.28) by V(*f*) yields

$$
1 = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123}, \tag{C.3.29}
$$

with

$$
\begin{aligned}
S_i &= \mathrm{V}(f_i)/\mathrm{V}(f) \quad \text{for} \quad i = 1,2,3 \\
&= \begin{cases}
\displaystyle\sum_{r=1}^{\infty} c_{r00}^2 / \mathrm{V}(f) & \text{for} \quad i = 1 \\
\displaystyle\sum_{s=1}^{\infty} c_{0s0}^2 / \mathrm{V}(f) & \text{for} \quad i = 2 \\
\displaystyle\sum_{t=1}^{\infty} c_{00t}^2 / \mathrm{V}(f) & \text{for} \quad i = 3,
\end{cases}
\end{aligned}
\tag{C.3.30}
$$

$$S_{ij} = \mathrm{V}(f_{ij}) / \mathrm{V}(f) \quad \text{for } 1 \le i < j \le 3$$
$$= \begin{cases} \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} c_{rs0}^2 / \mathrm{V}(f) & \text{for } (i,j) = (1,2) \\ \sum_{r=1}^{\infty}\sum_{t=1}^{\infty} c_{r0t}^2 / \mathrm{V}(f) & \text{for } (i,j) = (1,3) \\ \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{0st}^2 / \mathrm{V}(f) & \text{for } (i,j) = (2,3), \end{cases} \tag{C.3.31}$$

$$S_{123} = \mathrm{V}(f_{123}) / \mathrm{V}(f) = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}^2 / \mathrm{V}(f) \tag{C.3.32}$$

and, as previously indicated,

$$\mathrm{V}(f) = \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} c_{rst}^2 \tag{C.3.33}$$

in Eqs. (C.3.30)-(C.3.32). The terms on the right-hand side of Eq. (C.3.29) equal the fractions of the variance V(*f*) of $f$ that can be accounted for by $f_1, f_2, ..., f_{123}$ and can be used as measures of sensitivity. An additional measure of sensitivity is given by

$$\begin{aligned} S_{1T} &= [\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{123})] / \mathrm{V}(f) \\ &= S_1 + S_{12} + S_{13} + S_{123}, \end{aligned} \tag{C.3.34}$$

which measures the effect on V(*f*) of $x_1$ and all interactions of $x_1$ with $x_2$ and $x_3$. Corresponding measures $S_{2T}$ and $S_{3T}$ for $x_2$ and $x_3$ are defined analogously.

## C.4 Variance Decomposition for $f(x_1, x_2, ..., x_n)$ with Predefined Orthonormal Polynomials

This presentation has used $n$ = 3, i.e., $\mathbf{x} = [x_1, x_2, x_3]$, for illustration of the basic ideas in the chaos representation of a function and its associated variance. The ideas are similar but notationally more complex for a potentially larger value for $n$. For generality, a function $f(\mathbf{x})$ is now considered, with (i) $\mathbf{x} = [x_1, x_2, ..., x_n]$, (ii) the epistemically uncertain variables $x_1, x_2, ..., x_n$ assumed to be independent, and (iii) the uncertainty in each variable $x_i$ characterized by a probability space $(\mathcal{X}_i, \mathbb{X}_i, m_i)$ with $\mathcal{X}_i = [\underline{x}_i, \overline{x}_i]$ and an associated density function $d_i(x_i)$. Further, the distribution for each uncertain variable $x_i$ is assumed to have an associated set

$$\mathcal{P}_i = \{p_{id}(x_i) : p_{id}(x_i) \text{ is a polynomial of degree } d \text{ for } x_i \in \mathcal{X}_i \text{ and } d = 0, 1, 2, ...\} \tag{C.4.1}$$

of orthonormal polynomials with inner products defined as indicated in Eq. (C.3.2). Examples of possible definitions for the set $\mathcal{P}_i$ are shown in Table C.2.

The same type of manipulations that lead to the representations in Eqs. (C.3.5) and (C.3.6) for $f(\mathbf{x})$ and V($f$) with $n$ = 3 lead to similar, but somewhat more complicated, representations for $f(\mathbf{x})$ and V($f$) with an arbitrary value for $n$. The following additional notations are needed for the representations for $f(\mathbf{x})$ and V($f$) with an arbitrary value for $n$:

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n, \quad d(\mathbf{x}) = \prod_{i=1}^{n} d_i(x_i), \quad \mathrm{d}\mathbf{x} = \prod_{i=1}^{n} \mathrm{d}x_i, \tag{C.4.2}$$

$$f_0 = \mathrm{E}(f) = \int_{\mathcal{X}} f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{C.4.3}$$

$$\mathcal{I}^n = \mathcal{I}_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n \text{ with } \mathcal{I}_i = \{0,1,2,...\}, \tag{C.4.4}$$

$$p(\mathbf{x} \mid \mathbf{d}) = \prod_{i=1}^{n} p_{id_i}(x_i) \text{ for } \mathbf{d} = [d_1, d_2, ..., d_n] \in \mathcal{I}^n, \tag{C.4.5}$$

$$c_{\mathbf{d}} = \left\langle p(\mathbf{x} \mid \mathbf{d}), f(\mathbf{x}) \right\rangle = \int_{\mathcal{X}} p(\mathbf{x} \mid \mathbf{d}) f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \text{ for } \mathbf{d} \in \mathcal{I}^n, \tag{C.4.6}$$

and

$$\mathcal{D}^n_{i_1, i_2, ..., i_s} = \left\{ \mathbf{d} : \mathbf{d} = [d_1, d_2, ..., d_n] \in \mathcal{I}^n \text{ and } d_k > 0 \text{ iff } k \in \{i_1, i_2, ..., i_s\} \right\}, \tag{C.4.7}$$

where, in Eq. (C.4.7), (i) $i_1, i_2, ..., i_s$ is a sequence of integers with $1 \le i_1 < i_2 < \cdots < i_s \le n$ and (ii) iff is used as an abbreviation for "if and only if".

Given the preceding notations, manipulations similar to those that lead to the representations in Eqs. (C.3.5) and (C.3.6) for $f(\mathbf{x})$ and V($f$) with $n$ = 3 produce the following representations for $f(\mathbf{x})$ and V($f$) with an arbitrary value for $n$:

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, x_2, ..., x_n) \\ &= \sum_{\mathbf{d} \in \mathcal{I}^n} c_{\mathbf{d}} p(\mathbf{x} \mid \mathbf{d}) \\ &= f_0 + \sum_{i_1=1}^{n} \sum_{\mathbf{d} \in \mathcal{D}^n_{i_1}} c_{\mathbf{d}} p(\mathbf{x} \mid \mathbf{d}) + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}^n_{i_1, i_2}} c_{\mathbf{d}} p(\mathbf{x} \mid \mathbf{d}) + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}^n_{i_1, i_2, i_3}} c_{\mathbf{d}} p(\mathbf{x} \mid \mathbf{d}) \\ &\quad + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}^n_{i_1, i_2, ..., i_n}} c_{\mathbf{d}} p(\mathbf{x} \mid \mathbf{d}) \end{aligned} \tag{C.4.8}$$

and

$$
\begin{aligned}
\mathrm{V}(f) &= \sum_{\mathbf{d}\neq[0,0,\ldots,0]\in\mathcal{I}^n} c_{\mathbf{d}}^2 \\
&= \sum_{i_1=1}^{n}\sum_{\mathbf{d}\in\mathcal{D}_{i_1}^n} c_{\mathbf{d}}^2 + \sum_{i_1=1}^{n-1}\sum_{i_2=i_1+1}^{n}\sum_{\mathbf{d}\in\mathcal{D}_{i_1,i_2}^n} c_{\mathbf{d}}^2 + \sum_{i_1=1}^{n-2}\sum_{i_2=i_1+1}^{n-1}\sum_{i_3=i_2+1}^{n}\sum_{\mathbf{d}\in\mathcal{D}_{i_1,i_2,i_3}^n} c_{\mathbf{d}}^2 \\
&\qquad + \cdots + \sum_{i_1=1}^{1}\sum_{i_2=i_1+1}^{2}\sum_{i_3=i_2+1}^{3}\cdots\sum_{i_n=i_{n-1}+1`}^{n}\sum_{\mathbf{d}\in\mathcal{D}_{i_1,i_2,\ldots,i_n}^n} c_{\mathbf{d}}^2 \qquad \text{(C.4.9)} \\
&= \sum_{i_1=1}^{n}\mathrm{V}(f_{i_1}) + \sum_{i_1=1}^{n-1}\sum_{i_2=i_1+1}^{n}\mathrm{V}(f_{i_1 i_2}) + \sum_{i_1=1}^{n-2}\sum_{i_2=i_1+1}^{n-1}\sum_{i_3=i_2+1}^{n}\mathrm{V}(f_{i_1 i_2 i_3}) \\
&\qquad + \cdots + \sum_{i_1=1}^{1}\sum_{i_2=i_1+1}^{2}\sum_{i_3=i_2+1}^{3}\cdots\sum_{i_n=i_{n-1}+1`}^{n}\mathrm{V}(f_{i_1 i_2\cdots i_n}),
\end{aligned}
$$

with the final equality in Eq. (C.4.9) following from manipulations that are similar to the manipulations that lead to the final equality in Eq. (C.3.28).

With Eq. (C.4.9) established, normalized variances associated with $f(\mathbf{x})$ can be obtained as indicated in Eqs. (C.3.28)-(C.3.34) for $n = 3$. Specifically, division of both sides of Eq. (C.4.9) by V($f$) produces

$$
\begin{aligned}
1 &= \sum_{i_1=1}^{n} S_{i_1} + \sum_{i_1=1}^{n-1}\sum_{i_2=i_1+1}^{n} S_{i_1 i_2} + \sum_{i_1=1}^{n-2}\sum_{i_2=i_1+1}^{n-1}\sum_{i_3=i_2+1}^{n} S_{i_1 i_2 i_3} \\
&\qquad + \cdots + \sum_{i_1=1}^{1}\sum_{i_2=i_1+1}^{2}\sum_{i_3=i_2+1}^{3}\cdots\sum_{i_n=i_{n-1}+1}^{n} S_{i_1 i_2\cdots i_3},
\end{aligned}
\qquad \text{(C.4.10)}
$$

with

$$
S_{i_1} = \mathrm{V}(f_{i_1}) / \mathrm{V}(f) = \sum_{\mathbf{d}\in\mathcal{D}_{i_1}^n} c_{\mathbf{d}}^2 / \mathrm{V}(f) = \sum_{\mathbf{d}\in\mathcal{D}_{i_1}^n} c_{\mathbf{d}}^2 \Bigg/ \sum_{\mathbf{d}\neq[0,0,\ldots,0]\in\mathcal{I}^n} c_{\mathbf{d}}^2 \qquad \text{(C.4.11)}
$$

for $i_1 = 1, 2, \ldots, n$ and the other normalized variances $S_{i_1 i_2}$, $S_{i_1 i_2 i_3}$, …, $S_{i_1 i_2 \cdots i_3}$ defined similarly.

## C.5 Variance Decomposition for $f(x_1, x_2, x_3)$ without Predefined Orthonormal Polynomials

To this point, the polynomial chaos decomposition of $f(\mathbf{x}) = f(x_1, x_2, x_3)$ and associated variance decomposition have been derived with the assumptions indicated in conjunction with Eqs. (C.3.1) and (C.3.2). Central to these assumptions is that the distributions for $x_1$, $x_2$ and $x_3$ have

associated sets of orthonormal polynomials with the properties indicated in Eqs. (C.3.1) and (C.3.2). However, it may be the case that the desired sets of orthonormal polynomials are not available for the distributions for one or more of the variables under consideration. Fortunately, this complication can be handled with appropriate variable transformations as described below.

For starters, assume that a set of orthonormal polynomials with the properties indicated in Eqs. (C.3.1) and (C.3.2) is not available for variable $x_i$, $i$ = 1, 2 or 3, and its associated probability space $(\mathcal{X}_i, \mathbb{X}_i, m_i)$. In the following, it is shown that $x_i$ can be replaced in a polynomial chaos calculation by the transformed variable $T_i(u_i)$ defined by

$$C_{xi}[T_i(u_i)] = C_{ui}(u_i) \Rightarrow T_i(u_i) = C_{xi}^{-1}[C_{ui}(u_i)], \tag{C.5.1}$$

where (i) $C_{xi}(x_i)$ is the CDF associated with the probability space $(\mathcal{X}_i, \mathbb{X}_i, m_i)$ for $x_i$, (ii) the variable $u_i$ has an associated probability space $(\mathcal{U}_i, \mathbb{U}_i, m_{ui})$ with $\mathcal{U}_i = [\underline{u}_i, \overline{u}_i]$, (iii) $C_{ui}(u_i)$ and $d_{ui}(u_i)$ are the CDF and density function associated with $(\mathcal{U}_i, \mathbb{U}_i, m_{ui})$, and (iv) $(\mathcal{U}_i, \mathbb{U}_i, m_{ui})$ has an associated set

$$\mathcal{P}_{ui} = \{\tilde{p}_{id}(u_i) : \tilde{p}_{id}(u_i) \text{ is a polynomial of degree } d \text{ for } u_i \in \mathcal{U}_i \text{ and } d = 0,1,2,...\} \tag{C.5.2}$$

of orthonormal polynomials with an inner product defined by

$$\begin{aligned} \left\langle \tilde{p}_{ir}(u_i), \tilde{p}_{is}(u_i) \right\rangle &= \int_{\mathcal{U}_i} \tilde{p}_{ir}(u_i)\tilde{p}_{is}(u_i) d_{ui}(u_i) \mathrm{d}u_i \quad \text{for } \mathcal{I} = \{0,1,2,3,....\} \text{ and } (r,s) \in \mathcal{I} \times \mathcal{I} \\ &= \begin{cases} 1 & \text{if } r = s \\ 0 & \text{if } r \neq s. \end{cases} \end{aligned} \tag{C.5.3}$$

The relationships

$$\begin{aligned} d_{ui}(u_i) &= \mathrm{d}C_{ui}(u_i)/\mathrm{d}u_i \\ &= \mathrm{d}C_{xi}[T_i(u_i)]/\mathrm{d}u_i \\ &= \mathrm{d}C_{xi}(T)]/\mathrm{d}T\Big|_{T_i(u_i)} \times \mathrm{d}T_i(u_i)/\mathrm{d}u_i \\ &= d_{xi}[T_i(u_i)]T_i'(u_i) \quad \text{with } T_i'(u_i) = \mathrm{d}T_i(u_i)/\mathrm{d}u_i, \end{aligned} \tag{C.5.4}$$

$$T_i(\underline{u}_i) = C_{xi}^{-1}[C_{ui}(\underline{u}_i)] = C_{xi}^{-1}(0) = \underline{x}_i \tag{C.5.5}$$

and

$$T_i(\overline{u}_i) = C_{xi}^{-1}[C_{ui}(\overline{u}_i)] = C_{xi}^{-1}(1.0) = \overline{x}_i \tag{C.5.6}$$

follow from the definition of $T_i(u_i)$ in Eq. (C.5.1) and will be useful in later derivations.

For starters, it is noted that $f(x_1, x_2, x_3)$ and $\tilde{f}(u_1, u_2, u_3) = f[T_1(u_1), T_2(u_2), T_3(u_3)]$ have the same expected value and variance. Specifically,

$$
\begin{aligned}
\mathrm{E}(\tilde{f}) &= \mathrm{E}\left(f[T_1(u_1), T_2(u_2), T_3(u_3)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\int_{\underline{u}_3}^{\overline{u}_3} f[T_1(u_1), T_2(u_2), T_3(u_3)]\prod_{i=3}^{1} d_{ui}(u_i)\prod_{i=3}^{1}\mathrm{d}u_i \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\int_{\underline{u}_3}^{\overline{u}_3} f[T_1(u_1), T_2(u_2), T_3(u_3)]\prod_{i=3}^{1} d_{xi}[T_i(u_i)]T_i'(u_i)\prod_{i=3}^{1}\mathrm{d}u_i \quad \text{from Eq. (C.5.4)} \\
&= \int_{T_1(\underline{u}_1)}^{T_1(\overline{u}_1)}\int_{T_2(\underline{u}_2)}^{T_2(\overline{u}_2)}\int_{T_3(\underline{u}_3)}^{T_3(\overline{u}_3)} f(T_1, T_2, T_3)\prod_{i=3}^{1} d_{xi}(T_i)]\prod_{i=3}^{1}\mathrm{d}T_i \quad \text{from change of variables} \\
&= \int_{\underline{x}_1}^{\overline{x}_1}\int_{\underline{x}_2}^{\overline{x}_2}\int_{\underline{x}_3}^{\overline{x}_3} f(T_1, T_2, T_3)\prod_{i=3}^{1} d_{xi}(T_i)]\prod_{i=3}^{1}\mathrm{d}T_i \quad \text{from Eqs. (C.5.5)-(C.5.6)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f(x_1, x_2, x_3)\prod_{i=3}^{1} d_{xi}(x_i)]\prod_{i=3}^{1}\mathrm{d}x_i \quad \text{notation change: } T_i \Rightarrow x_i \\
&= \mathrm{E}[f(x_1, x_2, x_3)] \\
&= f_0
\end{aligned}
\tag{C.5.7}
$$

and

$$
\begin{aligned}
\mathrm{V}(\tilde{f}) &= \mathrm{V}\left(f[T_1(u_1), T_2(u_2), T_3(u_3)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\int_{\underline{u}_3}^{\overline{u}_3} f^2[T_1(u_1), T_2(u_2), T_3(u_3)]\prod_{i=3}^{1} d_{ui}(u_i)\prod_{i=3}^{1}\mathrm{d}u_i \\
&\qquad - \mathrm{E}^2\left(f[T_1(u_1), T_2(u_2), T_3(u_3)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\int_{\underline{u}_3}^{\overline{u}_3} f^2[T_1(u_1), T_2(u_2), T_3(u_3)]\prod_{i=3}^{1} d_{ui}(u_i)\prod_{i=3}^{1}\mathrm{d}u_i - f_0^2 \quad \text{from Eq. (C.5.7)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathcal{X}_3} f^2(x_1, x_2, x_3)\prod_{i=3}^{1} d_{xi}(x_i)]\prod_{i=3}^{1}\mathrm{d}x_i - f_0^2 \quad \text{same edits as in Eq. (C.5.7)} \\
&= \mathrm{V}[f(x_1, x_2, x_3)] \\
&= \mathrm{V}(f).
\end{aligned}
\tag{C.5.8}
$$

As indicated for $f(x_1, x_2, x_3)$ in Eq. (C.3.5), $f[T_1(u_1), T_2(u_2), T_3(u_3)]$ can be similarly represented by

$$
\begin{aligned}
f[T_1(u_1),T_2(u_2),T_3(u_3)] &= \sum_{(r,s,t)\in\mathcal{I}^3} \tilde{c}_{rst}\,\tilde{p}_{rst}(\mathbf{u}) \\
&= \sum_{(r,s,t)\in\mathcal{I}^3} \tilde{c}_{rst}\,\tilde{p}_{1r}(u_1)\tilde{p}_{2s}(u_2)\tilde{p}_{3t}(u_3) \\
&= f_0 + \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} \tilde{c}_{rst}\,\tilde{p}_{1r}(u_1)\tilde{p}_{2s}(u_2)\tilde{p}_{3t}(u_3),
\end{aligned}
\tag{C.5.9}
$$

with $\mathbf{u} = [u_1, u_2, u_3]$, $\tilde{p}_{rst}(\mathbf{u}) = \tilde{p}_{1r}(u_1)\tilde{p}_{2s}(u_2)\tilde{p}_{3t}(u_3)$ and $\mathcal{I}^3 = \mathcal{I}\times\mathcal{I}\times\mathcal{I}$. In turn, $\tilde{c}_{rst}$ is defined by

$$
\tilde{c}_{rst} = \left\langle \tilde{p}_{rst}(\mathbf{u}), f[T(\mathbf{u})] \right\rangle = \int_{\mathcal{U}} \tilde{p}_{rst}(\mathbf{u}) f[T(\mathbf{u})] d_u(\mathbf{u}) \mathrm{d}\mathbf{u} \tag{C.5.10}
$$

as indicated in Eq. (C.3.8) with $T(\mathbf{u}) = [T_1(u_1), T_2(u_2), T_3(u_3)]$ and $d_u(\mathbf{u}) = d_{u1}(u_1)d_{u2}(u_2)d_{u3}(u_3)$. In turn,

$$
\begin{aligned}
\mathrm{V}\left(f[T(\mathbf{u})]\right) &= \mathrm{V}\left(f[T_1(u_1),T_2(u_2),T_3(u_3)]\right) \\
&= \left\langle \sum_{(r,s,t)\in\mathcal{I}^3} \tilde{c}_{rst}\,\tilde{p}_{rst}(\mathbf{u}), \sum_{(r,s,t)\in\mathcal{I}^3} \tilde{c}_{rst}\,\tilde{p}_{rst}(\mathbf{u}) \right\rangle - \mathrm{E}^2\left(f[T(\mathbf{u})]\right) \\
&= \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} \tilde{c}_{rst}^2
\end{aligned}
\tag{C.5.11}
$$

as indicated in Eq. (C.3.8).

The previously defined variances defined in Eqs. (C.3.13)-(C.3.27) for $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$ can also be represented by appropriates sums of subsets of the squared coefficients $\tilde{c}_{rst}^2$ in the representation for $\mathrm{V}\left(f[T(\mathbf{u})]\right)$ in Eq. (C.5.11). To accomplish this, analogous functions $\tilde{f}_1(u_1)$, $\tilde{f}_2(u_2)$, $\tilde{f}_3(u_3)$, $\tilde{f}_{12}(u_1,u_2)$, $\tilde{f}_{13}(u_1,u_3)$, $\tilde{f}_{23}(u_2,u_3)$ and $\tilde{f}_{123}(u_1,u_2,u_3)$ need to be defined and shown to have the same variances as $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1,x_2)$, $f_{13}(x_1,x_3)$, $f_{23}(x_2,x_3)$ and $f_{123}(x_1,x_2,x_3)$.

The function $\tilde{f}_1(u_1)$ is considered first and is defined by

$$
\begin{aligned}
\tilde{f}_1(u_1) &= \mathrm{E}\left(f[T(\mathbf{u}:u_1)]\right) - f_0 \quad \text{for } u_1 \in \mathcal{U}_1 \\
&= \int_{\mathcal{U}_2}\int_{\mathcal{U}_3} f[T(\mathbf{u}:u_1)]d_{u3}(u_3)d_{u2}(u_2)\mathrm{d}u_3\mathrm{d}u_2 - f_0 \\
&= \int_{\mathcal{U}_2}\int_{\mathcal{U}_3} f[T_1(u_1),T_2(u_2),T_3(u_3)]d_{u3}(u_3)d_{u2}(u_2)\mathrm{d}u_3\mathrm{d}u_2 - f_0 \\
&= \int_{\mathcal{X}_2}\int_{X_3} f[T_1(u_1),x_2,x_3]d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0,
\end{aligned}
\tag{C.5.12}
$$

with the final equality following from the same edits (i.e., steps) used in Eq. (C.5.7). Next,

$$
\begin{aligned}
\mathrm{E}\left[\tilde{f}_1(u_1)\right] &= \int_{\mathcal{U}_1}\left[\int_{\mathcal{X}_2}\int_{X_3} f[T_1(u_1),x_2,x_3]d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0\right]d_{u1}(u_1)\mathrm{d}u_1 \\
&= \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_2}\int_{X_3} f[x_1,x_2,x_3]d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0\right]d_1(x_1)\mathrm{d}x_1 \\
&\qquad \text{same edits as in Eq. (C.5.7)} \\
&= f_0 - f_0 \\
&= 0,
\end{aligned}
\tag{C.5.13}
$$

$$
\begin{aligned}
\mathrm{V}\left[\tilde{f}_1(u_1)\right] &= \int_{\mathcal{U}_1}\left[\int_{\mathcal{X}_2}\int_{X_3} f[T_1(u_1),x_2,x_3]d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0\right]^2 d_{u1}(u_1)\mathrm{d}u_1 - 0^2 \\
&= \int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_2}\int_{X_3} f(x_1,x_2,x_3)d_3(x_3)d_2(x_2)\mathrm{d}x_3\mathrm{d}x_2 - f_0\right]^2 d_1(x_1)\mathrm{d}x_1 \\
&\qquad \text{same edits as in Eq. (C.5.8)} \\
&= \mathrm{V}\left[f_1(x_1)\right],
\end{aligned}
\tag{C.5.14}
$$

$$
\begin{aligned}
\tilde{c}_r &= \left\langle \tilde{p}_{1r}, \tilde{f}_1(u_1) \right\rangle \\
&= \left\langle \tilde{p}_{1r}, \int_{\mathcal{U}_2} \int_{\mathcal{U}_3} f[T(\mathbf{u}:u_1)] d_{u3}(u_3) d_{u2}(u_2) \mathrm{d}u_3 \mathrm{d}u_2 - f_0 \right\rangle \\
&= \int_{\mathcal{U}_1} \tilde{p}_{1r}(u_1) \left[ \int_{\mathcal{U}_2} \int_{\mathcal{U}_3} f[T(\mathbf{u}:u_1)] d_{u3}(u_3) d_{u2}(u_2) \mathrm{d}u_3 \mathrm{d}u_2 \right] d_{u1}(u_1) \mathrm{d}u_1 \\
&\qquad - f_0 \int_{\mathcal{U}_1} \tilde{p}_{1r}(u_1) d_{u1}(u_1) \mathrm{d}u_1 \\
&= \int_{\mathcal{U}_1} \int_{\mathcal{U}_2} \int_{\mathcal{U}_3} \tilde{p}_{1r}(u_1) f[T(\mathbf{u})] d_u(\mathbf{u}) \mathrm{d}\mathbf{u} - \begin{cases} f_0 \times 1 & \text{for } r = 0 \\ f_0 \times 0 & \text{for } r > 0 \end{cases} \\
&= \begin{cases} f_0 - f_0 = 0 & \text{for } r = 0 \\ \int_{\mathcal{U}_1} \int_{\mathcal{U}_2} \int_{\mathcal{U}_3} \tilde{p}_{1r}(u_1) f[T(\mathbf{u})] d_u(\mathbf{u}) \mathrm{d}\mathbf{u} & \text{for } r > 0 \end{cases} \\
&= \tilde{c}_{r00} \quad \text{for } r = 1, 2, \ldots,
\end{aligned}
\tag{C.5.15}
$$

$$
\tilde{f}_1(u_1) = \sum_{r=1}^{\infty} \tilde{c}_r \tilde{p}_{1r}(u_1) = \sum_{r=1}^{\infty} \tilde{c}_{r00} \tilde{p}_{1r}(u_1) \tilde{p}_{20}(u_2) \tilde{p}_{30}(u_3) \tag{C.5.16}
$$

$$
\mathrm{V}\left[ \tilde{f}_1(u_1) \right] = \left\langle \sum_{r=1}^{\infty} \tilde{c}_r \tilde{p}_{1r}(u_1) - 0, \sum_{r=1}^{\infty} \tilde{c}_r \tilde{p}_{1r}(u_1) - 0 \right\rangle = \sum_{r=1}^{\infty} \tilde{c}_r^2 = \sum_{r=1}^{\infty} \tilde{c}_{r00}^2 \tag{C.5.17}
$$

and in final summary

$$
\mathrm{V}\left[ f_1(x_1) \right] = \mathrm{V}\left[ \tilde{f}_1(u_1) \right] = \sum_{r=1}^{\infty} \tilde{c}_r^2 = \sum_{r=1}^{\infty} \tilde{c}_{r00}^2 . \tag{C.5.18}
$$

In similarity to the definition of $\tilde{f}_1(u_1)$ in Eq. (C.5.12), $\tilde{f}_2(u_2)$, $\tilde{f}_3(u_3)$, $\tilde{f}_{12}(u_1, u_2)$, $\tilde{f}_{13}(u_1, u_3)$, $\tilde{f}_{23}(u_2, u_3)$ and $\tilde{f}_{123}(u_1, u_2, u_3)$ are defined by

$$
\tilde{f}_i(u_i) = \mathrm{E}\left( f[T(\mathbf{u}:u_i)] \right) - f_0 \quad \text{for } u_i \in \mathcal{U}_i, i = 1, 2, 3, \tag{C.5.19}
$$

$$
\begin{aligned}
\tilde{f}_{ij}(u_i, u_j) = \mathrm{E}\left( f[T(\mathbf{u}:u_i, u_j)] \right) - \tilde{f}_i(u_i) - \tilde{f}_j(u_j) - f_0 \\
\text{for } (u_i, u_j) \in \mathcal{U}_i \times \mathcal{U}_j, 1 \le i < j \le 3
\end{aligned}
\tag{C.5.20}
$$

and

$$\tilde{f}_{123}(u_1,u_2,u_3) = \mathrm{E}\big(f[T(\mathbf{u}:u_1,u_2,u_3)]\big) - f_0 - \tilde{f}_1(u_1) - \tilde{f}_2(u_2) - \tilde{f}_3(u_3)$$
$$- \tilde{f}_{12}(u_1,u_2) - \tilde{f}_{13}(u_1,u_3) - f_{23}(u_2,u_3) \qquad \text{(C.5.21)}$$
$$\text{for } (u_1,u_2,u_3) \in \mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3.$$

Next, the following variance relationships follow by derivations similar to the derivation in Eqs. (C.5.12)-(C.5.17) that produced the equalities involving $\mathrm{V}\big[f_1(x_1)\big]$ and $\mathrm{V}\big[\tilde{f}_1(u_1)\big]$ in Eq. (C.5.18) :

$$\mathrm{V}\big[f_2(x_2)\big] = \mathrm{V}\big[\tilde{f}_2(u_2)\big] = \sum_{s=1}^{\infty} \tilde{c}_{0s0}^2 \qquad \text{(C.5.22)}$$

$$\mathrm{V}\big[f_3(x_3)\big] = \mathrm{V}\big[\tilde{f}_3(u_3)\big] = \sum_{t=1}^{\infty} \tilde{c}_{00t}^2, \qquad \text{(C.5.23)}$$

$$\mathrm{V}\big[f_{12}(x_1,x_2)\big] = \mathrm{V}\big[\tilde{f}_2(u_1,u_2)\big] = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \tilde{c}_{rs0}^2 \qquad \text{(C.5.24)}$$

$$\mathrm{V}\big[f_{13}(x_1,x_3)\big] = \mathrm{V}\big[\tilde{f}_{13}(u_1,u_3)\big] = \sum_{r=1}^{\infty}\sum_{t=1}^{\infty} \tilde{c}_{r0t}^2 \qquad \text{(C.5.25)}$$

$$\mathrm{V}\big[f_{23}(x_2,x_3)\big] = \mathrm{V}\big[\tilde{f}_{23}(u_2,u_3)\big] = \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \tilde{c}_{0st}^2 \qquad \text{(C.5.26)}$$

and

$$\mathrm{V}\big[f_{123}(x_1,x_2,x_3)\big] = \mathrm{V}\big[\tilde{f}_{123}(u_1,u_2,u_3)\big] = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \tilde{c}_{rst}^2. \qquad \text{(C.5.27)}$$

The polynomial chaos variance decomposition of $f(x_1,x_2,x_3)$ is based on the equalities

$$
\begin{aligned}
\mathrm{V}[f(x_1,x_2,x_3)] &=_1 \mathrm{V}\left(f[T_1(u_1),T_2(u_2),T_3(u_3)]\right) \\
&=_2 \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} \tilde{c}^2_{rst} \\
&=_3 \sum_{r=1}^{\infty}\tilde{c}^2_{r00} + \sum_{s=1}^{\infty}\tilde{c}^2_{0s0} + \sum_{t=1}^{\infty}\tilde{c}^2_{00t} + \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\tilde{c}^2_{rs0} \\
&\quad + \sum_{r=1}^{\infty}\sum_{t=1}^{\infty}\tilde{c}^2_{r0t} + \sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\tilde{c}^2_{0st} + \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\tilde{c}^2_{rst} \\
&=_4 \mathrm{V}(\tilde{f}_1) + \mathrm{V}(\tilde{f}_2) + \mathrm{V}(\tilde{f}_3) + \mathrm{V}(\tilde{f}_{12}) + \mathrm{V}(\tilde{f}_{13}) + \mathrm{V}(\tilde{f}_{23}) + \mathrm{V}(\tilde{f}_{123}) \\
&=_5 \mathrm{V}(f_1) + \mathrm{V}(f_2) + \mathrm{V}(f_3) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{23}) + \mathrm{V}(f_{123}),
\end{aligned}
\tag{C.5.28}
$$

with (i) Equality 1 following from Eq. (C.5.8), (ii) Equality 2 following from Eq. (C.5.11), (iii) Equality 3 following from the same manipulations that produced Equality 2 in Eq. (C.3.28), and Equalities 4 and 5 following from Eqs. (C.5.18) and (C.5.22)-(C.5.27).

Next, as previously discussed in connection with Eqs. (3.28)-(3.32), division of both sides of Eq. (C.5.28) by any one of the equivalent expressions

$$
\mathrm{V}(f) = \mathrm{V}[f(x_1,x_2,x_3)] = \mathrm{V}\left(f[T_1(u_1),T_2(u_2),T_3(u_3)]\right) = \mathrm{V}(\tilde{f}) \tag{C.5.29}
$$

yields

$$
1 = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123}, \tag{C.5.30}
$$

with

$$
\begin{aligned}
S_i &= \mathrm{V}(f_i)/\mathrm{V}(f) \quad \text{for} \quad i = 1,2,3 \\
&= \begin{cases}
\displaystyle\sum_{r=1}^{\infty}\tilde{c}^2_{r00}/\mathrm{V}(\tilde{f}) & \text{for} \quad i = 1 \\
\displaystyle\sum_{s=1}^{\infty}\tilde{c}^2_{0s0}/\mathrm{V}(\tilde{f}) & \text{for} \quad i = 2 \\
\displaystyle\sum_{t=1}^{\infty}\tilde{c}^2_{00t}/\mathrm{V}(\tilde{f}) & \text{for} \quad i = 3,
\end{cases}
\end{aligned}
\tag{C.5.31}
$$

$$
\begin{aligned}
S_{ij} &= \mathrm{V}(f_{ij}) / \mathrm{V}(f) \quad \text{for } 1 \le i < j \le 3 \\
&= \begin{cases}
\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \tilde{c}_{rs0}^2 / \mathrm{V}(\tilde{f}) & \text{for } (i,j) = (1,2) \\
\sum_{r=1}^{\infty}\sum_{t=1}^{\infty} \tilde{c}_{r0t}^2 / \mathrm{V}(\tilde{f}) & \text{for } (i,j) = (1,3) \\
\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \tilde{c}_{0st}^2 / \mathrm{V}(\tilde{f}) & \text{for } (i,j) = (2,3),
\end{cases}
\end{aligned}
\tag{C.5.32}
$$

$$
S_{123} = \mathrm{V}(f_{123}) / \mathrm{V}(f) = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} c_{rst}^2 / \mathrm{V}(\tilde{f}) \tag{C.5.33}
$$

and, as previously indicated,

$$
\mathrm{V}(\tilde{f}) = \sum_{(r,s,t)\neq(0,0,0)\in\mathcal{I}^3} \tilde{c}_{rst}^2 . \tag{C.5.34}
$$

The terms on the right-hand side of Eq. (C.5.30) equal the fractions of the variance V(*f*) of $f$ that can be accounted for by $f_1, f_2, ..., f_{123}$ and can be used as measures of sensitivity. An additional measure of sensitivity is given by

$$
\begin{aligned}
S_{1T} &= [\mathrm{V}(f_1) + \mathrm{V}(f_{12}) + \mathrm{V}(f_{13}) + \mathrm{V}(f_{123})] / \mathrm{V}(f) \\
&= [\mathrm{V}(\tilde{f}_1) + \mathrm{V}(\tilde{f}_{12}) + \mathrm{V}(\tilde{f}_{13}) + \mathrm{V}(\tilde{f}_{123})] / \mathrm{V}(\tilde{f}) \\
&= S_1 + S_{12} + S_{13} + S_{123},
\end{aligned}
\tag{C.5.35}
$$

which measures the effect on V(*f*) of $x_1$ and all interactions of $x_1$ with $x_2$ and $x_3$. Corresponding measures $S_{2T}$ and $S_{3T}$ for $x_2$ and $x_3$ are defined analogously.

The notation used in this section implies that the distribution for each variable $x_i$ does not have a conveniently defined associated set of orthonormal polynomials. However, the notation in use actually allows for the inclusion of the case where the distributions for one or more variables have known associated sets of orthonormal polynomials. In this case, these sets orthonormal polynomials and associated distributions can be used in the polynomial chaos expansion as indicated in this section without a notational change from use of the transformation $T_i(u_i)$ defined in Eq. (C.5.1). This result occurs because, if $x_i$ and $u_i$ have the same distributions, then $T_i(u_i)$ as defined in Eq. (C.5.1) simply becomes the identity function (i.e., $T_i(u_i) = u_i$), which means that the use of the transformation $T_i(u_i)$ in a polynomial chaos has no effect on the results of the expansion.

## C.6 Variance Decomposition for $f(x_1, x_2, ..., x_n)$ without Predefined Orthonormal Polynomials

Sect. C.5 has used $n = 3$, i.e., $\mathbf{x} = [x_1, x_2, x_3]$, for illustration of the basic ideas in the polynomial chaos representation of a function and its associated variance without the assumption that the distribution for each $x_i$ has an associated set of orthonormal polynomials. The ideas are similar but notationally more complex for a potentially larger value for $n$. For generality, a function $f(\mathbf{x})$ is now considered, with (i) $\mathbf{x} = [x_1, x_2, ..., x_n]$, (ii) the epistemically uncertain variables $x_1, x_2, ..., x_n$ assumed to be independent, and (iii) the uncertainty in each variable $x_i$ characterized by a probability space $(\mathcal{X}_i, \mathbb{X}_i, m_i)$ with $\mathcal{X}_i = [\underline{x}_i, \overline{x}_i]$ and a corresponding density function $d_i(x_i)$. An associated set of orthonormal polynomials is not assumed for the distribution assigned to each $x_i$. However, as described in conjunction with Eqs. (C.5.1)-(C.5.6), each $x_i$ is assigned a set $\mathcal{P}_{ui}$ orthonormal polynomials that, in general, is not related to the distribution assigned to $x_i$.

The same type of manipulations that lead to the equalities $\mathrm{E}(\tilde{f}) = \mathrm{E}(f)$ and $\mathrm{V}(\tilde{f}) = \mathrm{V}(f)$ in Eqs. (C.5.7) and (C.5.8) for $n = 3$ lead to the same equalities for an arbitrary value for $n$. For this development, the notational definitions in Eqs. (C.4.2)-(C.4.7) remain unchanged with the following similar notational definitions made for the orthonormal polynomials in the sets $\mathcal{P}_{ui}$ and associated distributions defined by the probability spaces $(\mathcal{U}_i, \mathbb{U}_i, m_{ui})$ and density functions $d_{ui}(u_i)$:

$$\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \cdots \times \mathcal{U}_n, \quad d_u(\mathbf{u}) = \prod_{i=1}^{n} d_{ui}(u_i), \quad \mathrm{d}\mathbf{u} = \prod_{i=1}^{n} \mathrm{d}u_i, \tag{C.6.1}$$

$$\tilde{f}(\mathbf{u}) = f[T(\mathbf{u})] = f[T_1(u_1), T_2(u_2), ..., T_n(u_n)] \tag{C.6.2}$$

$$\tilde{p}(\mathbf{u} \mid \mathbf{d}) = \prod_{i=1}^{n} \tilde{p}_{id_i}(u_i) \text{ for } \mathbf{d} = [d_1, d_2, ..., d_n] \in \mathcal{I}^n, \tag{C.6.3}$$

and

$$\tilde{c}_{\mathbf{d}} = \left\langle \tilde{p}(\mathbf{u} \mid \mathbf{d}), f\left[T(\mathbf{u})\right] \right\rangle = \int_{\mathcal{U}} \tilde{p}(\mathbf{x} \mid \mathbf{d}) f\left[T(\mathbf{u})\right] d_u(\mathbf{u}) \mathrm{d}\mathbf{u} \text{ for } \mathbf{d} \in \mathcal{I}^n. \tag{C.6.4}$$

Given the preceding notations, manipulations similar to those that lead to the equalities in Eqs. (C.5.7) and (C.5.8) for $n = 3$ produce corresponding equalities for an arbitrary value for $n$. Specifically,

$$
\begin{aligned}
\mathrm{E}(\tilde{f}) &= \mathrm{E}\left(f[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\cdots\int_{\underline{u}_n}^{\overline{u}_n} f[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\prod_{i=n}^{1} d_{ui}(u_i)\prod_{i=n}^{1}\mathrm{d}u_i \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\cdots\int_{\underline{u}_n}^{\overline{u}_n} f[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\prod_{i=n}^{1} d_{xi}[T_i(u_i)]T_i'(u_i)\prod_{i=n}^{1}\mathrm{d}u_i \quad \text{from Eq. (C.5.4)} \\
&= \int_{T_1(\underline{u}_1)}^{T_1(\overline{u}_1)}\int_{T_2(\underline{u}_2)}^{T_2(\overline{u}_2)}\cdots\int_{T_n(\underline{u}_n)}^{T_n(\overline{u}_n)} f(T_1, T_2, ..., T_n)\prod_{i=n}^{1} d_{xi}(T_i)]\prod_{i=n}^{1}\mathrm{d}T_i \quad \text{from change of variables} \qquad \text{(C.6.5)} \\
&= \int_{\underline{x}_1}^{\overline{x}_1}\int_{\underline{x}_2}^{\overline{x}_2}\cdots\int_{\underline{x}_n}^{\overline{x}_n} f(T_1, T_2, ..., T_n)\prod_{i=n}^{1} d_{xi}(T_i)]\prod_{i=n}^{1}\mathrm{d}T_i \quad \text{from Eqs. (C.5.5)-(C.5.6)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\cdots\int_{\mathcal{X}_n} f(x_1, x_2, ..., x_n)\prod_{i=n}^{1} d_{xi}(x_i)]\prod_{i=n}^{1}\mathrm{d}x_i \quad \text{notation change: } T_i \Rightarrow x_i \\
&= [f(x_1, x_2, ..., x_3)] = \mathrm{E}(f) = f_0.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathrm{V}(\tilde{f}) &= \mathrm{V}\left(f[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\cdots\int_{\underline{u}_n}^{\overline{u}_n} f^2[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\prod_{i=n}^{1} d_{ui}(u_i)\prod_{i=n}^{1}\mathrm{d}u_i \\
&\qquad - \mathrm{E}^2\left(f[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\right) \\
&= \int_{\underline{u}_1}^{\overline{u}_1}\int_{\underline{u}_2}^{\overline{u}_2}\cdots\int_{\underline{u}_3}^{\overline{u}_3} f^2[T_1(u_1), T_2(u_2), ..., T_n(u_n)]\prod_{i=n}^{1} d_{ui}(u_i)\prod_{i=n}^{1}\mathrm{d}u_i - f_0^2 \quad \text{from Eq. (C.5.7)} \qquad \text{(C.6.6)} \\
&= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\cdots\int_{\mathcal{X}_3} f^2(x_1, x_2, ..., x_n)\prod_{i=n}^{1} d_{xi}(x_i)]\prod_{i=n}^{1}\mathrm{d}x_i - f_0^2 \quad \text{same edits as in Eq. (C.5.7)} \\
&= \mathrm{V}[f(x_1, x_2, ..., x_n)] = \mathrm{V}(f).
\end{aligned}
$$

Next, similarly to $f$ and V($f$) in Eqs. (C.4.8) and (C.4.9), the polynomial chaos representations for $\tilde{f}$ and $\mathrm{V}(\tilde{f})$ with $\mathcal{D}^n_{i_1,i_2,...,i_s}$ and $\tilde{c}_{\mathbf{d}}$ defined in Eqs. (C.4.7) and (C.6.4) are given by

$$
\begin{aligned}
\tilde{f}(\mathbf{u}) &= f[T_1(u_1), T_2(u_2), ..., T_n(u_n)] \\
&= \sum_{\mathbf{d} \in \mathcal{I}^n} \tilde{c}_{\mathbf{d}} \tilde{p}(\mathbf{u} \mid \mathbf{d}) \\
&= f_0 + \sum_{i_1=1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1}^n} \tilde{c}_{\mathbf{d}} \tilde{p}(\mathbf{u} \mid \mathbf{d}) + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2}^n} \tilde{c}_{\mathbf{d}} \tilde{p}(\mathbf{u} \mid \mathbf{d})) + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2,i_3}^n} \tilde{c}_{\mathbf{d}} \tilde{p}(\mathbf{u} \mid \mathbf{d}) \\
&\quad + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2,...,i_n}^n} \tilde{c}_{\mathbf{d}} \tilde{p}(\mathbf{u} \mid \mathbf{d})
\end{aligned}
\tag{C.6.7}
$$

and

$$
\begin{aligned}
\mathrm{V}(\tilde{f}) &=_1 \sum_{\mathbf{d} \neq [0,0,...,0] \in \mathcal{I}^n} c_{\mathbf{d}}^2 \\
&=_2 \sum_{i_1=1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1 i_1}^n} \tilde{c}_{\mathbf{d}}^2 + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2}^n} \tilde{c}_{\mathbf{d}}^2 + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2,i_3}^n} \tilde{c}_{\mathbf{d}}^2 \\
&\quad + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1`}^{n} \sum_{\mathbf{d} \in \mathcal{D}_{i_1,i_2,...,i_n}^n} \tilde{c}_{\mathbf{d}}^2 \\
&=_3 \sum_{i_1=1}^{n} \mathrm{V}(\tilde{f}_{i_1}) + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} \mathrm{V}(\tilde{f}_{i_1 i_2}) + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} \mathrm{V}(\tilde{f}_{i_1 i_2 i_3}) \\
&\quad + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1`}^{n} \mathrm{V}(\tilde{f}_{i_1 i_2 \cdots i_n}), \\
&=_4 \sum_{i_1=1}^{n} \mathrm{V}(f_{i_1}) + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} \mathrm{V}(f_{i_1 i_2}) + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} \mathrm{V}(f_{i_1 i_2 i_3}) \\
&\quad + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1`}^{n} \mathrm{V}(f_{i_1 i_2 \cdots i_n}),
\end{aligned}
\tag{C.6.8}
$$

where (i) Equality 1 follows from the orthonormality of the polynomials in the sets $\mathcal{P}_{u_i}$, $i$ = 1, 2, …, $n$, (ii) Equality 2 follows from a rearrangement of the summation in equality 1, (iii) Equality 3 involves functions $\tilde{f}_i, \tilde{f}_{ij}, ..., \tilde{f}_{ij...n}$ defined the same as the functions $f_i, f_{ij}, ..., f_{ij...n}$ except with respect to $u_1, u_2, ..., u_n$ rather than $x_1, x_2, ..., x_n$ and with the equalities between the summations in Equality 2 and the variances $\mathrm{V}(\tilde{f}_i), \mathrm{V}(\tilde{f}_{ij}), ..., \mathrm{V}(\tilde{f}_{ij...n})$ obtained in a manner similar to that shown in Eqs. (C.3.13)-(C.3.28), and (iv) Equality 4 follows from the equalities $\mathrm{V}(\tilde{f}_i) = \mathrm{V}(f_i)$, $\mathrm{V}(\tilde{f}_{ij}) = \mathrm{V}(f_{ij}), \ldots, \mathrm{V}(\tilde{f}_{ij...n}) = \mathrm{V}(f_{ij...n})$ obtained in a manner similar to that shown in Eqs. (C.5.12)-(C.5.27).

With Eq. (C.6.8) established, normalized variances associated with $f(\mathbf{x})$ can be obtained as indicated in Eqs. (C.5.29)-(C.5.30) for $n$ = 3. Specifically, division of both sides of Eq. (C.6.8)by $\mathrm{V}(\tilde{f})$ produces

$$1 = \sum_{i_1=1}^{n} S_{i_1} + \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} S_{i_1 i_2} + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} S_{i_1 i_2 i_3} + \cdots + \sum_{i_1=1}^{1} \sum_{i_2=i_1+1}^{2} \sum_{i_3=i_2+1}^{3} \cdots \sum_{i_n=i_{n-1}+1}^{n} S_{i_1 i_2 \cdots i_3}, \tag{C.6.9}$$

with

$$\begin{aligned} S_{i_1} &= \mathrm{V}(f_{i_1}) / \mathrm{V}(f) \\ &= \mathrm{V}(\tilde{f}_{i_1}) / \mathrm{V}(\tilde{f}) \\ &= \sum_{\mathbf{d} \in D_{i_1}^{n}} \tilde{c}_{\mathbf{d}}^{2} \Big/ \sum_{\mathbf{d} \neq [0,0,\ldots,0] \in \mathcal{I}^{n}} \tilde{c}_{\mathbf{d}}^{2} \end{aligned} \tag{C.6.10}$$

for $i_1 = 1, 2, ..., n$ and the other normalized variances $S_{i_1 i_2}$, $S_{i_1 i_2 i_3}$, …, $S_{i_1 i_2 \cdots i_n}$ defined similarly.

Most analyses are likely to involve either (i) a mixture of variables some of which have a distribution associated with one of the commonly defined sets of orthonormal polynomials while the other variables do not have such an association or (ii) only variables that have distributions that are not associated with any of the commonly defined sets of orthonormal polynomials. For example, uncertainty and sensitivity analyses often make extensive use of distributions such as piecewise uniform, log uniform, triangular and truncated normal.

A simple way to design a polynomial chaos analysis structure is to assume that shifted Legendre polynomials are the only set of orthonormal polynomials to be used. Then, each variable and its associated distribution will undergo a transformation of the form indicated in Eq. (C.5.1), and then the analysis will have the conceptual structure described in this section. The shifted Legendre polynomials are particularly convenient to work with as their associated distribution is a uniform distribution on [0, 1] With this distribution (i.e., uniform on [0,1]), the transformation in Eq. (C.5.1) corresponds to what is usually done in the generation of random and Latin hypercube samples of size *nS* (i.e., for each variable *x* under consideration, sampled values $u_1, u_2, ..., u_{nS}$ from a uniform distribution on [0,1] are generated and then used obtain sampled values for *x* from the transform $x_i = CDF_x^{-1}(u_i)$, where $CDF_x$ is the CDF for *x*). Thus, the transformation in Eq. (C.5.1) is a commonly used transformation in sampling when shifted Legendre polynomials are used in a polynomial chaos analysis.

## C.7 Approximation of Polynomial Chaos Representations for $f(x_1, x_2, x_3)$ and $f(x_1, x_2, ..., x_n)$

The approximation of the polynomial chaos representation

$$f(\mathbf{x}) = f(x_1, x_1, ..., x_n) = \sum_{\mathbf{d} \in \mathcal{D}} c_{\mathbf{d}}\, p(\mathbf{x} \mid \mathbf{d}) \text{ with } \mathcal{D} = \{\mathbf{d} : \mathbf{d} \in \mathcal{I}^n\} \tag{C.7.1}$$

is now considered, with

$$\mathcal{P}_i = \{p_{id}(x_i) : p_{id}(x_i) \text{ is a polynomial of degree } d \text{ for } x_i \in \mathcal{X}_i \text{ and } d = 0,1,2,...\}, \tag{C.7.2}$$

$$p(\mathbf{x} \mid \mathbf{d}) = \prod_{i=1}^{n} p_{id_i}(x_i) \text{ for } \mathbf{d} = [d_1, d_2, ..., d_n] \in \mathcal{D} = \mathcal{I}^n, \tag{C.7.3}$$

$$c_{\mathbf{d}} = \langle p(\mathbf{x} \mid \mathbf{d}), f(\mathbf{x}) \rangle = \int_{\mathcal{X}} p(\mathbf{x} \mid \mathbf{d}) f(\mathbf{x}) d(\mathbf{x}) \mathrm{d}\mathbf{x} \text{ for } \mathbf{d} \in \mathcal{D} = \mathcal{I}^n, \tag{C.7.4}$$

$$\|\mathbf{d}\| = \sum_{i=1}^{n} d_i = \text{ degree of } p(\mathbf{x} \mid \mathbf{d}) \text{ for } \mathbf{d} \in \mathcal{D} = \mathcal{I}^n. \tag{C.7.5}$$

The indicated approximations are usually defined by truncating the infinite series in Eq. (C.7.1). This results in an approximation of the form

$$f(\mathbf{x}) = f(x_1, x_1, ..., x_n) = \sum_{\mathbf{d} \in \widehat{\mathcal{D}}} c_{\mathbf{d}}\, p(\mathbf{x} \mid \mathbf{d}), \tag{C.7.6}$$

where $\widehat{\mathcal{D}}$ is a suitably restricted subset of $\mathcal{D}$.

The manner in which $\widehat{\mathcal{D}}$ is defined is likely to be analysis specific. A possibility is to define $\widehat{\mathcal{D}}$ on the basis of a maximum value $d_{mx}$ for the degrees $\|\mathbf{d}\|$ of the polynomials $p(\mathbf{x} \mid \mathbf{d})$ that result in the coefficients $c_{\mathbf{d}}$ in Eq. (C.7.6). This manner of defining $\widehat{\mathcal{D}}$ results in

$$\widehat{\mathcal{D}} = \{\mathbf{d} : \mathbf{d} \in \mathcal{D} \text{ and } \|\mathbf{d}\| \le d_{mx}\}. \tag{C.7.7}$$

In turn, this definition for $\widehat{\mathcal{D}}$ results in

$$nT = \binom{n + d_{mx}}{d_{mx}} = \frac{(n + d_{mx})!}{n! d_{mx}!} \tag{C.7.8}$$

terms in the summation in Eq. (C.7.6) ([25], Eq. 17).

In concept, coefficients $c_{\mathbf{d}}, \mathbf{d} \in \widehat{\mathcal{D}}$, in Eq. (C.7.6) can be obtained by numerically evaluating the defining integral for each $c_{\mathbf{d}}$ shown in Eq. (C.7.4). However, this is not a workable approach in most analyses due to (i) the large number of coefficients that must be evaluated and (ii) the expense of evaluating the function $f(\mathbf{x})$, which in most analyses is a complex and computationally demanding computer program.

An alternate procedure is to use a regression-based procedure for the estimation of the coefficients $c_{\mathbf{d}}, \mathbf{d} \in \widehat{\mathcal{D}}$, in Eq. (C.7.6) [26]. This approach is based on the use of (i) a random or Latin hypercube sample

$$\mathbf{x}_s = [x_{1s}, x_{2s}, ..., x_{ns}], s = 1, 2, ..., nS, \tag{C.7.9}$$

generated from $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ in consistency with the density functions $d_i(x_i), i = 1, 2, ..., n$, and (ii) the associated model evaluations

$$f(\mathbf{x}_s) = f(x_{1s}, x_{2s} ..., x_{ns}), i = 1, 2, ..., nS. \tag{C.7.10}$$

In turn, the desire is to obtain estimates $\hat{c}_{\mathbf{d}}$ for the coeffects $c_{\mathbf{d}}$ in Eq. (C.7.6) that produce the smallest (i.e., minimum) value for the sum of squares

$$\sum_{s=1}^{nS} \left[ f(\mathbf{x}_s) - \sum_{\mathbf{d} \in \widehat{\mathcal{D}}} \hat{c}_{\mathbf{d}} p(\mathbf{x}_s \mid \mathbf{d}) \right]^2 = (\mathbf{f} - \mathbf{P}\hat{\mathbf{c}})'(\mathbf{f} - \mathbf{P}\hat{\mathbf{c}}), \tag{C.7.11}$$

with

$$\mathbf{f} = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_{nS}) \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} p(\mathbf{x}_1 \mid \mathbf{d}_1) & p(\mathbf{x}_1 \mid \mathbf{d}_2) & \cdots & p(\mathbf{x}_1 \mid \mathbf{d}_{nT}) \\ p(\mathbf{x}_2 \mid \mathbf{d}_1) & p(\mathbf{x}_2 \mid \mathbf{d}_2) & \cdots & p(\mathbf{x}_2 \mid \mathbf{d}_{nT}) \\ \vdots & \vdots & \vdots & \vdots \\ p(\mathbf{x}_{nS} \mid \mathbf{d}_1) & p(\mathbf{x}_{nS} \mid \mathbf{d}_2) & \cdots & p(\mathbf{x}_{nS} \mid \mathbf{d}_{nT}) \end{bmatrix}, \quad \hat{\mathbf{c}} = \begin{bmatrix} \hat{c}_{\mathbf{d}_1} \\ \hat{c}_{\mathbf{d}_1} \\ \vdots \\ \hat{c}_{\mathbf{d}_1} \end{bmatrix} = \begin{bmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \vdots \\ \hat{c}_{nT} \end{bmatrix}, \tag{C.7.12}$$

a prime (e.g., $\mathbf{P}'$) designating the transpose of a matrix, and the elements of $\widehat{\mathcal{D}}$ represented by $\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{nT}$ for notational convenience. The desired estimates are obtained from the minimization procedure

$$\mathbf{0} = \partial / \partial \hat{\mathbf{c}} \left[ (\mathbf{f} - \mathbf{P}\hat{\mathbf{c}})'(\mathbf{f} - \mathbf{P}\hat{\mathbf{c}}) \right] = -2\mathbf{P}'\mathbf{f} + 2(\mathbf{P}'\mathbf{P}\hat{\mathbf{c}}) \tag{C.7.13}$$

and the resultant equalities

$$(\mathbf{P}'\mathbf{P})\hat{\mathbf{c}} = \mathbf{P}'\mathbf{f} \text{ and } \hat{\mathbf{c}} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{f} \tag{C.7.14}$$

with

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}' \text{ and } \partial / \partial \hat{\mathbf{c}} = \begin{bmatrix} \partial / \partial c_1 & \partial / \partial c_2 & \cdots & \partial / \partial c_{nT} \end{bmatrix}. \qquad \text{(C.7.15)}$$

Thus, the estimated coefficients $\hat{c}_{\mathbf{d}}$ that minimize the sum of squares in Eq. (C.7.11) can be obtained from the second equality in Eq. (C.7.14) ([27], p.88). The justification for this result derives from a result known as the Gauss-Markov Theorem.

## C.8 References: Appendix C


1. Crestaux T., O. Le Maître and J.M. Martinez. 2009. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering and System Safety* 94:1161-1172.
2. Sudret B. 2008. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety* 93:964-979.
3. Blatman G. and B. Sudret. 2010. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering and System Safety* 95:1216-1229.
4. Le Gratiet L.L., S. Marelli and B. Sudret. 2017. Metamodel-based sensitivity analysis: Polynomial chaos expansions and gaussian processes. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:1289-1325: Springer International Publishing Switzerland.
5. Ghanem R. and J. Red-Horse. 2017. Polynomial chaos: Modeling, estimation, and approximation. In *Handbook of Uncertainty Quantification*, ed. R ghanem, Higdon, D., Owhadi, H.:521-551: Springer International Publishing Switzerland.
6. Soize C. and R. Ghanem. 2004. Physical Systems with Random Characterristics: Chaos Representations with Arbitrary Probability Measure. *SIAM J. Scientific Computing* 26:395-410.
7. Hadigol M. and A. Doostan. 2018. Least squares polynomial chaos expansion: A review of sampling strategies. *Computer Methods in Applied Mechanics and Engineering* 332:382-407.
8. Sudret B. and C.V. Mai. 2015. Computing derivative-based global sensitivity measures using polynomial chaos expansions. *Reliability Engineering and System Safety* 134:241-250.
9. Konakli K. and B. Sudret. 2016. Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions. *Journal of Computational Physics* 321:1144-1169.
10. Debusschere B. 2017. Intrusive polynomial chaos methods for forward uncertainty propagation. In *Handbook of Uncertainty Quantification*, ed. R Ghanem, Higdon, D., Owhadi, H.:617-636: Springer International Publishing Switzerland.
11. Eldred M.S. 2009. Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design C3 - Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.
12. Burnaev E., I. Panin and B. Sudret. 2017. Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions. *Annals of Mathematics and Artificial Intelligence* 81:187-207.

13. Garcia-Cabrejo O. and A. Valocchi. 2014. Global Sensitivity Analysis for multivariate output using Polynomial Chaos Expansion. *Reliability Engineering and System Safety* 126:25-36.
14. Eldred M.S. and J. Burkardt. 2009. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification C3 - 47th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition.
15. Blatman G. and B. Sudret. 2011. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics* 230:2345-2367.
16. Jakeman J.D., F. Franzelin, A. Narayan, M. Eldred and D. Plfüger. 2019. Polynomial chaos expansions for dependent random variables. *Computer Methods in Applied Mechanics and Engineering* 351:643-666.
17. Jakeman J.D., M.S. Eldred and K. Sargsyan. 2015. Enhancing ℓ1-minimization estimates of polynomial chaos expansions using basis selection. *Journal of Computational Physics* 289:18-34.
18. Szego G. 1939. *Orthogonal Polynomials*. New York, NY: American Mathematical society.
19. Abramowitz M. and I.A. Stegun, eds. 1983. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, NY: Dover Publications.
20. Chihara T.S. 1978. *An Introduction to Orthogonal Polynomials*. New York, NY: Gordon and Breach.
21. Jackson D. 2004. *Fourier Series and Orthogonal Polynomials*. New York, NY: Dover.
22. Ismail M.E.H. 2005. *Classical and Quantum Orthogonal Polynomials in One Variable*. Cambridge: Cambridge University Press.
23. Totik V. 2005. Orthogonal Polynomials. *Surveys in Approximation Theory* 1:70-125.
24. Oliver S.W.J., D.M. Lozier, R.F. Boisvert and C.W. Clark, eds. 2010. *NIST Handbook of Mathematical Functions*. Cambridge: Cambridge UniversityPress.
25. Le Gratiet L., Marelli, S., Sudret, B. 2017. Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expnasions and Gaussian Processes. In *Handbook of Uncertainty Quantifiction* ed. R Ghanem, Higdon, D., Owhadi, H.:1289-1326: Springer International Publishing Switzerland.
26. Berveiller M., B. Sudret and M. Lemaire. 2006. Stochastic finite element: A non intrusive approach by regression. *European Journal of Computational Mechanics* 15:81-92.
27. Myers R.H. 1990. *Classical and Modern Regression with Applications*. Boston: Duxbury Press.