SANDIA REPORT

SAND2017-9373
Unlimited Release
Printed August 2017

SAChES: Scalable Adaptive Chain-Ensemble Sampling

J. Ray and L. Swiler and M. Ebeida and M. Huang and Z. Hou and J. Bao and H. Ren

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2017-9373
Unlimited Release
Printed August 2017

SAChES: Scalable Adaptive Chain-Ensemble Sampling

J. Ray,
Sandia National Laboratories, P. O. Box 969, Livermore CA 94551

L. Swiler and M. Ebeida
Sandia National Laboratories, P. O. Box 5800, Albuquerque NM 87185-1318

and

M. Huang, Z. Hou, J. Bao, and H. Ren
Pacific Northwest National Laboratory, PO Box 999, Richland, WA 99352

{jairay, lpswile, msebeid}@sandia.gov
{maoyi.huang, zhangshuan.hou, jie.bao, huiying.ren}@pnnl.gov

Abstract

We present the development of a parallel Markov Chain Monte Carlo (MCMC) method called SAChES, Scalable Adaptive Chain-Ensemble Sampling. This capability is targed to Bayesian calibration of computationally expensive simulation models. SAChES involves a hybrid of two methods: Differential Evolution Monte Carlo followed by Adaptive Metropolis. Both methods involve parallel chains. Differential evolution allows one to explore high-dimensional parameter spaces using loosely coupled (i.e., largely asynchronous) chains. Loose coupling allows the use of large chain ensembles, with far more chains than the number of parameters to explore. This reduces per-chain sampling burden, enables high-dimensional inversions and the use of computationally expensive forward models. The large number of chains can also ameliorate the impact of silent-errors, which may affect only a few chains. The chain ensemble can also be sampled to provide an initial condition when an aberrant chain is re-spawned. Adaptive Metropolis takes the best points from the differential evolution and efficiently hones in on the posterior density. The multitude of chains in SAChES is leveraged to (1) enable efficient exploration of the parameter space; and (2) ensure robustness to silent errors which may be unavoidable in extreme-scale computational platforms of the future.

This report outlines SAChES, describes four papers that are the result of the project, and discusses some additional results.

Acknowledgment

This work was funded the U.S. Department of Energy, via the Office of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830.

Contents

This page intentionally left blank

Figures

Tables

1 Introduction

This report describes the results of a project called SAChES: Scalable Adaptive Chain-Ensemble Sampling. The goal of this project is to deliver a multi-chain Markov chain Monte Carlo (MCMC) method that can be used for solving inverse problems. The focus has been on Bayesian calibration of computationally expensive PDE-based simulators with O(10) - O(100) parameters, typical for scientific and engineering models. The multitude of chains is leveraged to (1) enable efficient exploration of the parameter space; and (2) ensure robustness to silent errors which may be unavoidable in extreme-scale computational platforms of the future. Note that MCMC methods for parameter estimation are one of the main ways of implementing uncertainty quantification in inverse problems.

The outline of this report is as follows: Section 2 provides the background on MCMC approaches and Bayesian calibration of computer models. This is followed by four sections, 3 - 6, which describe journal articles that were produced as part of this work. Section 7 outlines some test problems which we have used to address correctness of the posterior, efficiency of multiple chains, and scalability in terms of number of parameters. Section 8 outlines a new approach for calibration based on ideas from computational geometry. Section 9 has some concluding thoughts.

1.1 Publications and Presentations

Under this project, we have four journal articles as well as a number of conference presentations and posters. The journal publications are as follows:

1. Bayesian Calibration of the Community Land Model using Surrogates. by J. Ray, Z. Hou, M. Huang, K. Sargsyan, and L. Swiler. SIAM/ASA Journal of Uncertainty Quantification 3(1). pp. 199-233. DOI. 10.1137/140957998. [31]

2. On the Applicability of Surrogate-based Markov Chain Monte-Carlo Bayesian inversion to the Community Land Model: Case studies at flux tower sites. by M. Huang, J. Ray, Z. Hou, H. Ren, Y. Liu, and L. Swiler. Journal of Geophysical Research: Atmospheres. 121(13). pp. 7548-7563. DOI. 10.1002/2015JD024339. [22]

3. Soil Moisture Estimation Using Tomographic Ground Penetrating Radar in a MCMC-Bayesian Framework. by J. Bao, Z. Hou, J. Ray, M. Huang, L. Swiler, and H. Ren. submitted to Stochastic Environmental Research and Risk Assessment, in revision.

4. Bayesian Inversion of Seismic and Electromagnetic Data for Marine Gas Reservoir Characterization Using Multi-chain Markov Chain Monte Carlo Sampling. by H. Ren, J. Ray, Z. Hou, M. Huang, J. Bao, and L. Swiler. Submitted to Journal of Applied Geophysics, in revision.

The conference presentations are as follows:

1. SIAM Conference on Uncertainty Quantification. Entropy-Bayesian inversion of hydrological parameters in the Community Land Model using heat flux and runoff data. by Z. Hou, M. Huang, J. Ray, and L. Swiler. SAND2014-2609C.

2. SIAM Conference on Uncertainty Quantification. Estimation of structural error in the Community Land Model using latent heat observations. by J. Ray, M. Huang, Z. Hou, and L. Swiler. SAND2014-2431C.

3. SIAM Conference on Computational Science and Engineering. Estimating a model discrepancy term for the Community Land Model using latent heat and runoff observations. by J. Ray, L. Swiler, M. Huang, and Z. Hou. SAND2015-1458C.

4. SIAM Conference on Computational Science and Engineering. Use of parallel MCMC with the Community Land Model. by L. Swiler, J. Ray, M. Huang, and Z. Hou. SAND2015-1854C.

5. SIAM Conference on Computational Science and Engineering. Quantification of structural uncertainty in a land surface model. by Z. Hou, M. Huang, J. Ray, and L. Swiler. SAND2015-1972C.

6. ASME Verification and Validation Conference. A parallel Markov chain Monte carlo method for calibrating computationally expensive models. by J. Ray, L. Swiler, M. Huang, and Z. Hou. SAND2016-4662C.

7. Probabilistics Mechanics Conference (PMC) 2016. A parallel MCMC method. by L. Swiler and J. Ray. SAND2016-4866C.

The posters are as follows:

1. American Geophysical Union 2014. Using Run-off Data to calibrate the Community Land Model. by J. Ray, Z. Hou, M. Huang, and L. Swiler. SAND2014-20410C.

2. American Geophysical Union 2014. Applicability of Surrogate-based MCMC Bayesian inversion of CLM at flux tower sites with various field conditions. by Z. Hou, J. Ray, M. Huang, L. Swiler, H. Ren, and Y. Liu. SAND2015-0706C.

3. CoDA 2014. Conference on Data Analysis, sponsored by LANL and DoE. Bayesian Calibration of Hydrological Parameters in the Community Land model. by J. Ray, K. Sargsyan, L. Swiler, Z. Hou, and M. Huang. SAND2014-1440C.

4. American Geophysical Union 2015. Parameter Estimation and Transferability Study across Major U.S. Watersheds. by H, Ren, Z. Hou, M. Huang, J. Bao, J. Ray, Y. Sun, T. Tesfa, R. Leung, and L. Swiler. SAND2015-10842C.

5. American Geophysical Union 2015. Bayesian Calibration of the Community Land Model using a Multi-Chain MCMC Method. by L. Swiler, J. Ray, M. Huang, and Z. Hou. SAND2015-10732C.

6. American Geophysical Union 2015. A Scalable Multi-chain Markov Chain Monte Carlo Method for Inverting Subsurface Hydraulic and Geological Properties by J. Bao, H. Ren, Z. Hou, M. Huang, J. Ray, and L. Swiler. SAND2015-10841C.

7. American Geophysical Union 2016. Using a Multi-chain Markov Chain Monte Carlo method to calibrate Geoscientific models. by J. Ray, L. Swiler, J. Bao, M. Huang, and Z. Hou. SAND2016-12361C.

2 Background

Model calibration is central to predictive simulations. Often, due to shortcomings of the models and/or sparsity of calibration data ("observations"), the model parameters can only be estimated with a large degree of uncertainty. Markov chain Monte Carlo methods enable Bayesian calibration of the model (estimation of parameters with quantified uncertainty) without imposing any distributional constraints. However, they invoke the forward model/simulator repeatedly and thus can hardly be applied to computationally expensive models or when a large number of parameters have to be estimated.

A few MCMC methods [27], [33] have attempted to address this limitation by distributing the sampling burden over a large number of interacting chains that jointly explore the parameter space. The chains are periodically synchronized, which impairs their scalability to large numbers of processors. Furthermore, synchronization makes the entire ensemble susceptible to individual chain failures, due to aphysical parameter combinations and/or resilience issues on extreme-scale computing platforms (e.g., sporadic node failures or bit-flips/"silent-errors"). These scalability and resilience issues make current multi-chain MCMC approaches of limited use in challenging calibration problems on extreme-scale hardware.

Our new method combines Differential Evolution - Markov chain (DE-MC, a genetic algorithm modified for Markov chains [35]), adaptive Metropolis, and ensemble sampling. A core hypothesis underlying our new approach is that the sampling efficiency of current MCMC methods with synchronized chains can be replicated using asynchronous, loosely-coupled Markov chains. Differential evolution allows one to explore high-dimensional parameter spaces using loosely coupled (i.e., largely asynchronous) chains. Loose coupling allows the use of large chain ensembles, with far more chains than the number of parameters/dimensions to explore. This reduces per-chain sampling burden, enabling high-dimensional inversions and the use of computationally expensive forward models. The large number of chains can also ameliorate the impact of silent-errors, which may affect only a few chains. The chain ensemble can also be sampled to provide an initial condition when an aberrant chain is re-spawned.

After Differential Evolution is used to explore the space, we use an adaptive Metropolis approach to refine the posterior density. The particular Differential Evolution approach we use is called DREAM, and the adaptive Metropolis approach is AM. Both DREAM and AM involve parallel chains. In DREAM, there are rules about using information from multiple chains to generate the next proposed sample location (these rules are similar to crossover rules in genetic algorithms). In AM, the chains communicate at a particular frequency, such as every n evaluations, and then update their individual proposal densities based on information from all of the chains. We call our method SAChES - Scalable Adaptive Chain-Ensemble Sampling.

The construction of SAChES required the following advances:

1. Algorithmic design using asynchronous communications: We allow for communication across chains in Snooker and parallel updates (used in DREAM). These updates combine past states of multiple chains and are an efficient means of pooling information across the ensemble of chains. Further, they can potentially be implemented via asynchronous reads and thus be very scalable. They may, however, not generate high-quality MCMC proposals by themselves. We hybridize the DREAM chains with adaptive Metropolis sampling (using a multivariate Gaussian proposal distribution), so that SAChES sampling retains its sparse communication and synchronization patterns while gaining the sampling efficiency of an adaptive Metropolis sampler.
2. Fault-detection and pruning/re-spawning of aberrant chains: Chains, driven to unpromising regions of the parameter space by silent error corruption, or stalled by aphysical parameter combinations, have

to be re-spawned using initial conditions drawn from the posterior distribution. We have implemented fault detection in SAChES.

The SAChES code is in a software repository at Sandia. Please contact Jaideep Ray or Laura Swiler if you are interested in this code.

2.1 Markov Chain Monte Carlo

This section describes Markov Chain Monte Carlo methods in more detail, specifically with their use in using a Bayesian approach to calibrate parameters of computational simulation models.

Estimation of parameters from observations can be cast as a Bayesian inverse problem. Let $\mathbf{y} = \mathbf{m}(\mathbf{p})$ be a model with parameters $\mathbf{p}$. The model outputs are related to observations $\mathbf{y}^{(obs)}$ as

$$\mathbf{y}^{(obs)} = \mathbf{y} + \varepsilon = \mathbf{m}(\mathbf{p}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Gamma) \tag{1}$$

where ε is a combination of measurement and structural error and $\mathcal{N}(0,\Gamma)$ denotes a multivariate Gaussian distribution with zero mean and Γ as the covariance matrix. Let $\pi(\mathbf{p},\Gamma)$ be the prior belief regarding the distribution of the parameters and the structural error. By Bayes' theorem, the posterior PDF $P(\mathbf{p},\Gamma|\mathbf{y}^{(obs)})$ of the parameters, conditioned on observations, can be given by

$$P(\mathbf{p},\Gamma|\mathbf{y}^{(obs)}) \propto \underbrace{|\Gamma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{y}^{(obs)} - \mathbf{m}(\mathbf{p})\right)^T \Gamma^{-1} \left(\mathbf{y}^{(obs)} - \mathbf{m}(\mathbf{p})\right)\right]}_{Likelihood, \mathcal{L}(\mathbf{y}^{(obs)}|\mathbf{p},\Gamma)} \pi(\mathbf{p},\Gamma) \tag{2}$$

This is the post-calibration or posterior distribution of the parameters $\mathbf{p}$. It can be constructed by sampling from the right hand side of (2) and generating a histogram of the samples. Markov chain Monte Carlo (MCMC) methods [15] allow the sampling to be performed efficiently. In MCMC, one starts with a guess of the parameter $\mathbf{p}_0$. Using this as the base, a proposal $\mathbf{p}'$ is chosen from a proposal PDF (often, but not necessarily, a multivariate Gaussian) $Q(\mathbf{p}'|\mathbf{p}_0)$. $\mathbf{p}'$ is retained according to certain acceptance criteria, which ensure that the chain is ergodic (so that a chain of infinite length visits all parts of the parameter space) and satisfies detailed balance (i.e., high-probability parameters are visited more often than the low probability ones). The mixing of the MCMC chain in the parameter space is largely dependent on $Q(:)$. Adaptive MCMC methods [19] seek to tune an optimal Q i.e., estimate its covariance periodically using samples $\mathbf{p}_i$ that have already been collected by the MCMC chain. Multichain MCMC methods [33, 6] that use multiple concurrent chains to explore the parameter space have been used in the estimation of climate model parameters [23].

The MCMC chain is stopped when the samples it collects results in a stationary posterior distribution $P(\mathbf{p},\Gamma|\mathbf{y}^{(obs)})$. An efficient MCMC method can require $O(10^4)$ samples to represent a posterior distribution for 3-4 parameters; for complex-shaped distributions, far more samples may be required. The convergence of a MCMC chain can be judged using the Raftery-Lewis [30] or Brooks-Gelman-Rubin [2] statistics. An unconverged MCMC chain usually leads to parameter PDFs that are too narrow i.e., it underestimates parametric uncertainty, and provides erroneous estimates of high-order moments of the distribution such as inter-parameter correlations. The quality of a Bayesian calibration is gauged by posterior predictive tests (PPTs; chapter on "Model Checking and Improvement" in [14]). Samples of $(\mathbf{p},\Gamma)$ are drawn from the posterior distribution and used to replicate observations via an ensemble of model simulations using (1). The

predictive skill of the ensemble is gauged by metrics such as the cumulative rank predictive score (CRPS), verification rank histogram (VRH), mean absolute error (MAE) etc. [17, 16].

Note that the formulation used for Bayesian calibration in (1) and (2) is a standard one. However, often the model $\mathbf{m}(\mathbf{p})$ used in the calculation of the likelihood is not the "true" simulation but a surrogate constructed for cheaper evaluation of the likelihood function. There are a variety of surrogate models used to "emulate" the expensive computational simulation. These include Gaussian process models (GPs) and Polynomial Chaos Expansions (PCEs). NEED: reference. We have used both in our research. One goal for SAChES was to reduce the need for surrogate models with the use of parallel chains. In some cases, surrogates still may be needed. We do not focus on surrogates for the purpose of this work, but wanted to mention their use in MCMC approaches.

2.2 Parallel Treatment in SAChES

As mentioned, there are two phases of SAChES: a differential evolution phase followed by an adaptive metropolis phase. Note that the differential evolution phase is optional: one may run only the adaptive metropolis phase. Both phases uses parallel chains. This section outlines how the chains are updated in each case.

1. **Differential Evolution.** We follow the differential evolution approach presented by Vrugt et al. in [36]. There are j chains in parallel. At each iteration i, a chain jumps to another position that is a sum of its current position plus a fixed multiple of the difference or two randomly chosen other chain members, k and m, where $j \neq k \neq m$. Given chain j currently is at parameter value $\theta_{j,i}$, the next step in the chain is given as follows:

$$\theta_{j,i+1} = \theta_{j,i} + \gamma(\theta_{k,i} - \theta_{m,i}) + \varepsilon \tag{3}$$

2. **Adaptive Metropolis.** For a single chain MCMC, adaptive metropolis works by updating the proposal covariance periodically. Since the proposal covariance governs how large the proposed steps in the chain are, if the proposal covariance is too big, the chain will be bouncing around the space and have trouble honing in on a high likelihood area. In contrast, if the proposal covariance is too small, the proposed steps will be tiny and it will take the chain a long time to converge. Haario et al. proposed an adaptive updating of the proposal covaraince that depends on the covariance between the accepted points in the chain. [19]. The proposal covariance is updated every p steps of the chain, where the proposal covariance at a particular iteration i is a weighted sum of the proposal covariance at the previous update, $i-p$, along with a term that involves the covariance of the p accepted values since the last proposal covariance update.

   With parallel chains, we follow a procedure outlined in [33]. Each chain j has a proposal covariance that is a Gaussian distribution centered at the parameter values at the particular iteration value i and a covariance that is the same across chains. That is, the proposal for chain j at iteration i is:

$$\mathcal{N}(\theta_{j,i}, \Sigma_i) \tag{4}$$

   The Σ_i values are only a function of iteration and not the chain. That is, all chains start at different parameter positions but with the same proposal covariance. The chains proceed for a certain number of steps p, then they combine ALL of the accepted values across chains (p*j accepted samples) and

use those samples to update the proposal covariance. After the update, the proposal covariance for each chain is updated but these covariances have the same values across the chain.

The proposal covariance update for iteration i is then:

$$\Sigma_i = \frac{i-p}{i}\Sigma_{i-p} + \frac{p}{i}(\Theta_{j,p} - \overline{\Theta})(\Theta_{j,p} - \overline{\Theta})^T) \tag{5}$$

where $\Theta_{j,p}$ represents the vector of the p accepted points from all j chains (e.g. it contains $j*p$ values), and $\overline{\Theta}$ represents the mean of that vector.

2.3 Fault Detection and Respawning of Aberrant Chains

The SAChES multichain sampler consists of N individual "samplers" on N processors, each driving a physics model (Community Land Model in our case). Each "sampler" implements DrEAM and AM. Inter-processor communication occur over an MPI intra-communicator. One of the 'samplers" is a "head-sampler" that is identical to other samplers except that it is responsible for any input/output to screen. Each "sampler" checkpoints its chain to disk periodically. The "head-sampler"' also checkpoints the proposal covariance matrix.

The "fault-tolerant" version of SAChES allows the killing and respawning of "samplers" if a fault occurs. A fault could imply a memory corruption that stalls the forward problem. It could also imply that a particular instance of a chain is exploring a nonphysical part of the parameter space, causing convergence issues in the forward problem, and considerably slowing it down. It is expected that for physically relevant parameter combinations, forward models will take approximately the same time to execute. On the detection of a fault, all "samplers" are killed and respawned using `MPI_Spawn` from their checkpoint files. The control architecture is implemented via a master-slave architecture.

The master-slave implementation covers $(N+1)$ processors. One process, the master, spawns N samplers via an `MPI_Spawn`, returning an intercommunicator over, over which all master-slave communications occur. The slaves commence sampling and at the end of each generation, return the generation number to the master. The master contains an array, N long, where each slave writes its generation count. Periodically the master analyzes the array of generation counts, and based on a user-defined threshold, decides if a slave is significantly lagging behind. This is interpreted as a fault. At the end of the analysis, the master communicates back to the slaves to indicate if they should terminate or proceed sampling.

The communication between master and slaves has to be specially designed since some of the slaves may have stalled and may not communicate back to the master. The communication is performed using passive target synchronization, using asynchronized one-sided MPI calls. We define two window objects, where the master's window contains allocated memory to hold variables, and slaves' do not. In the first (master's) window we allocate an array, N long, where the slaves report their generation number using `MPI_Put`. The master fetches the array (via `MPI_Get`) and analyzes it. The second window object, has one boolean variable (in the master's window) where the master deposits its decision to continue for another generation or terminate. This decision is fetched by the slaves via `MPI_Get`. `MPI_Win_lock`/`MPI_Win_unlock` are used for passive synchronization.

3 Bayesian Calibration of the Community Land Model using Surrogates

The first paper published in this project focused on Bayesian calibration of the Community Land Model (CLM, [29]) using surrogates. A statistical inverse problem is formulated for three hydrological parameters, conditioned on observations of latent heat surface fluxes over 48 months. Our calibration method uses polynomial and Gaussian process surrogates of the CLM, and solves the parameter estimation problem using a Markov chain Monte Carlo sampler.

Posterior probability densities for the parameters are developed for two sites with different soil and vegetation covers. Our method also allows us to examine the structural error in CLM under two error models. We find that accurate surrogate models could be created for CLM in three out of the four cases we investigated. The posterior distributions lead to better prediction than the default parameter values in CLM. Climatologically averaging the observations does not modify the parameters' distributions significantly. The structural error model reveals a correlation time-scale which can potentially be used to identify physical processes that could be contributing to it. While the calibrated CLM has a higher predictive skill, the calibration is under-dispersive.

The Community Land Model (CLM, [29]), the land component of the Community Earth System Model (CESM, [11]), is used to simulate terrestrial water, energy, and biogeochemical processes in offline and coupled climate simulations. The CLM contains a large number of parameters that govern its behavior, many of which are not directly measurable. They are estimated from indirect measurements, and are therefore subject to great uncertainty. Further, many parameters are site-dependent i.e., they vary within certain ranges [18, 20, 21]. In addition, due to difficulties in estimating such parameters at a global scale, CLM is released with default values for these parameters obtained by benchmarking its simulations against global datasets using simple statistics [12]. The predictive accuracy of CLM is, to a large degree, dependent on obtaining "correct" values of these parameters, and calibrating to site-specific observational data is the best means of doing so.

Model calibration, to date, has meant optimizing parameter values to reduce the discrepancies between historical observations and their corresponding model predictions (e.g., from CLM). This leads to a number of practical challenges. For example, gradient-descent optimization methods e.g., L-BFGS-B [4] are sensitive to their starting guesses and can yield multiple "optimal" parameter combinations. More seriously, due to the limited amount of observational data, the measurement errors in observations, and the modeling shortcomings/simplifications in CLM, parameters cannot be estimated with a high degree of accuracy. As a result, the parameter estimates are uncertain, but such parametric uncertainty has not been well quantified. Consequently, CLM is not distributed with "error bounds" that reflect parametric uncertainty after calibration.

The problem of parametric uncertainty can be addressed using Bayesian calibration. It develops parameter estimates as probability density functions (PDFs). The PDFs can be general i.e. we do not have to stipulate a canonical family of distributions like Gaussian, log-normal etc. or make any approximations in the numerical scheme if the Bayesian calibration problem is solved using a Markov chain Monte Carlo (MCMC) method. The PDF captures parametric uncertainty and the correlation between parameter estimates concisely. Further, such a calibration also improves the predictive skill of CLM; instead of attempting to predict observations with one "optimal" parameter combination, one samples the PDF and constructs an ensemble of CLM predictions. Simple statistical measures [17, 16] can be used to summarize the "goodness of fit"; further, the statistical measures also reveal other aspects of the fit (e.g., over-/under-dispersive calibrations) that provide specific directions to pursue to improve CLM. However, Bayesian calibration poses two

technical challenges. Firstly, like contemporary optimization methods, Bayesian calibration minimizes the model-observation discrepancy. In addition, it also requires one to specify a statistical model for the discrepancy (henceforth called the structural error model). The sensitivity of calibration to this choice then has to be gauged. Secondly, MCMC can require many ($O(10^4)$-$O(10^5)$) CLM evaluations to reach converged posterior estimates, which is prohibitive. Thus, while Bayesian calibration holds much promise for CLM calibration, its use has been rare [34, 37].

In this paper, we describe a method that can allow MCMC calibration of CLM. The method is based on surrogates of CLM - inexpensive polynomial or Gaussian process representations of the mapping between CLM parameters being calibrated and the CLM outputs for which we have measurements. We therefore build on, and extend, recent developments on the use of surrogates to calibrate computationally expensive models [28, 25] and MCMC calibration of complex (e.g., those based on partial differential equations) models including structural errors (i.e., the fundamental inability of the model to reproduce observations due to modeling simplifications) [8, 10]. Our method is general, but we will demonstrate it in the estimation of three hydrological parameters using observations from two sites, US-ARM, located in Oklahoma, and US-MOz, located in Missouri. The method will also yield an approximation of CLM's structural error. Our method is dependent on an accurate surrogate model; in its absence, our calibration method does not work. We will also present an example of this shortcoming.

The novel contributions of this paper are:

1. *Procedure for building CLM surrogates:* While the idea of building surrogates for computationally expensive models is not new [26], the particular form chosen for the surrogate is problem dependent. We describe the practical details of sampling the space of calibration parameters, performing the runs (which, in our case, produce a time-series of outputs), and the process of constructing surrogates while simultaneously simplifying them using sparsity. In particular, we will exploit a sparse reconstruction method, Bayesian compressive sensing [1], to perform model simplification.

2. *Choice of error model and their ramifications:* Bayesian calibration requires one to specify an error model. If competing models exist (as they do in our case), there has to be a systematic way of selecting one. We present an illustration of how to select an error model.

3. *Gauging the post-calibration predictive skill of CLM:* When one has a "point" estimate of parameters (the defaults or optimal values obtained from deterministic optimization), the predictive skill of a model is estimated by calculating bias and root-mean-square-error (RMSE) with respect to observations. When parameters are estimated as PDFs, a different set of error metrics can be used. Further, some of them can reveal how the model needs to be improved. We will compute these error metrics as a demonstration of the usefulness of Bayesian calibration beyond just parameter-estimation-with-uncertainty-quantification.

4 On the Applicability of Surrogate-based MCMC-Bayesian Inversion to the Community Land Model: Case Studies at Flux Tower Sites

This paper extended the approach of the previous paper 3 in two ways (1) It examined the variation in the hydrological parameters of the Community Land Model across a wide variety of sites and (2) the surrogate models for these sites were often poor and so we developed a procedure for eliminating combinations of the input parameter domain which contribute to poor surrogate performance.

In more detail: the Community Land Model (CLM) has been widely used in climate and Earth system modeling. Accurate estimation of model parameters is needed for reliable model simulations and predictions under current and future conditions, respectively. In our previous work, a subset of hydrological parameters has been identified to have significant impact on surface energy fluxes at selected flux tower sites based on parameter screening and sensitivity analysis, which indicate that the parameters could potentially be estimated from surface flux observations at the towers. To date, such estimates do not exist.

In this paper, we assess the feasibility of applying a Bayesian model calibration technique to estimate CLM parameters at selected flux tower sites under various site conditions. The parameters are estimated as a joint probability density function (PDF) that provides estimates of uncertainty of the parameters being inverted, conditional on climatologically-average latent heat fluxes derived from observations. We find that the simulated mean latent heat fluxes from CLM using the calibrated parameters are generally improved at all sites when compared to those obtained with CLM simulations using default parameter sets. Further, our calibration method also results in credibility bounds around the simulated mean fluxes which bracket the measured data. The modes (or maximum a posteriori values) and 95

An aspect of this paper that is of general interest is the input screening approach. There are some combinations of inputs which often do not make physical sense, but which will be generated if one is sampling points uniformly from a hypercube to obtain data to create a surrogate model. Specifically, we excise the inappropriate portions of parameter domain $\mathcal{P}$ to obtain $\mathcal{R}$, which contains physically realistic parameters. For example, in this study, we generated 282 samples from the 3 hydrological parameters $\mathcal{P}$ using a space-filling, quasi Monte Carlo method. We used them to generate an ensemble of LH predictions with CLM. We calculated the RMSE at each particular parameter combination p. That is, the $RMSE(p)$ are calculated for each realization and then we specified a threshold $RMSE$ quartile $QRMSE$ to identify p whose predictions are close to observations at a given site. The selected samples of p discretely define $\mathcal{R}$. We defined an improper, informative prior $\pi(p)$ with support $\mathcal{R}$ such that the prior density is one inside $\mathcal{R}$ and zero outside. We constructed our surrogates using only parameter combinations p that reside inside $\mathcal{R}$. Usually setting $QRMSE = 0.7$ has allowed us remove the non-physical part of $\mathcal{P}$ and to construct accurate surrogates. Since Bayesian parameter calibration is so computationally intensive and requires thousands of model evaluations, having accurate surrogates is critical to the process. The approach developed in this paper allowed us to ensure the accuracy of our surrogate models.

5 Soil Moisture Estimation Using Tomographic Ground Penetrating Radar in a MCMC-Bayesian Framework

This study used SAChES without surrogates: the simulation model was executed directly when evaluating the likelihood function. In this paper, we focus on a hydrogeological inverse problem specifically targeting monitoring soil moisture variations using tomographic ground penetrating radar (GPR) travel time data. Technical challenges exist in the inversion of GPR tomographic data for handling non-uniqueness, nonlinearity and high-dimensionality of unknowns. We have developed a new method for estimating soil moisture fields from crosshole GPR data. It uses a pilot-point method to provide a low-dimensional representation of the relative dielectric permittivity field of the soil, which is the primary object of inference: the field can be converted to soil moisture using a petrophysical model.

We integrate a multi-chain Markov chain Monte Carlo (MCMC) - Bayesian inversion framework with the pilot point concept, a curved-ray GPR travel time model, and a sequential Gaussian simulation algorithm for estimating the dielectric permittivity at pilot point locations distributed within the tomogram, as well as the corresponding geostatistical parameters (i.e., spatial correlation range). We infer the dielectric permittivity as a probability density function, thus capturing the uncertainty in the inference. The multi-chain MCMC enables addressing high-dimensional inverse problems as required in the inversion setup. The method is scalable in terms of number of chains and processors, and is useful for computationally demanding Bayesian model calibration in scientific and engineering problems. The proposed inversion approach can successfully approximate the posterior density distributions of the pilot points, and capture the true values. The computational efficiency, accuracy, and convergence behaviors of the inversion approach were also systematically evaluated, by comparing the inversion results obtained with different levels of noises in the observations, increased observational data, as well as increased number of pilot points.

The full paper is in submission and available upon request. Below we provide some example results.

Figure 1(a) shows the synthetic relative dielectric permittivity field between two boreholes, and is considered as the true field in this study. The study area is 4 m wide and 15 m deep. The true ε_r field is created using a pilot point method using 8 pilot points and a variogram that has a range of 2 and 20 meters in the vertical and horizontal directions respectively. The base case considers 30 equally spaced source locations on the left side of the field (x=0 m), and for each source location, GPR arrival time data is collected at 30 evenly spaced receiver locations on the right side of the field (x=4 m) for a total of 900 observations. The forward GPR model computes the 900 first-arrival-travel times as shown in Figure 1(b), which are considered to be the observational data. The symbols + and numbers in Figure 1 indicate the positions and indices of 24 pilot points. Pilot points 1-8 are the ones used to generate the true ε_r field.

We explore the usefulness of parallel MCMC in inverting the ε_r field, which we model using multivariate Gaussians placed at the first 8 pilot points. The Gaussians are governed by the same variogram, whose range is also estimated from the ε_r field data. Thus our inversion contains nine parameters including ε_r at 8 pilot points and the variogram's range. They are treated as random variables in our statistical formulation of the inverse problem and their nine-dimensional joint PDF is inferred via MCMC. Since the pilot points used to construct the random field model are the ones used to generate the true ε_r field, it incurs no model-form errors. Any systematic under- or over-estimation of the parameters being estimated, conditional on error-free observations would indicate a problem with our likelihood expression or our implementation.

As a first step, we solve the inverse problem with perfect but limited observations. 30 sources and 30 receivers are used to calculate the first-arrival-travel time to compare against the 900 observations, as shown

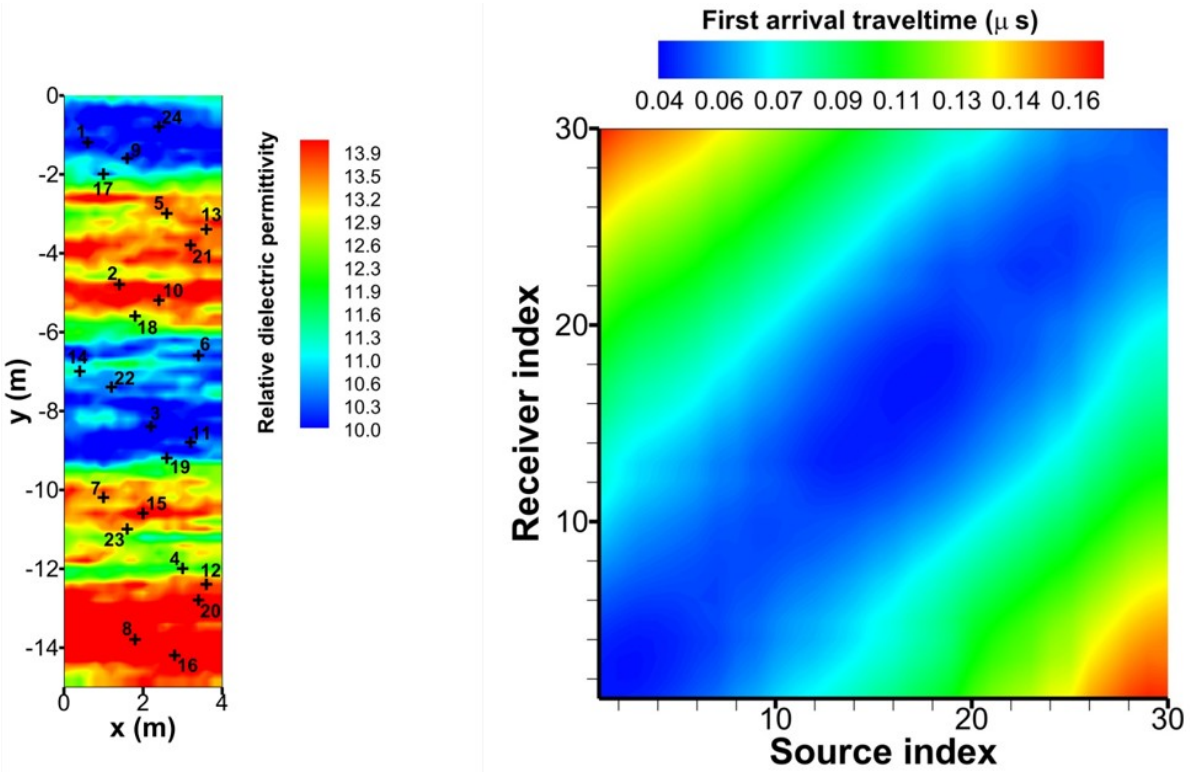


Figure 1: Synthetic experiment setups. (a) Relative dielectric permittivity of the synthetic field, the symbols + and numbers indicate the position and indices of the 24 pilot points. (b) Forward GPR simulated first-arrival-travel time between the each source and receiver for the synthetic field.

in Figure 2. No noise is added to the observations. In this study, we assume the horizontal range of the variogram is 10 times larger than the vertical one. Prior distributions for relative dielectric permittivity at the pilot points is U[4, 18] and U[1, 3] for range. 20 MCMC chains were used, and Figure 2 shows the posterior density distribution after 50000 iterations per chain i.e., a total of 1 million parameters samples were explored for constructing the posterior density distribution. The red vertical lines are the true value. The density distribution encapsulates the true values for all the parameters, except the parameter range. A possible reason is that the locations of the 8 pilot points already impose a length-scale for the ε_r field, which may conflict with the 9th parameter (spatial correlation range). Further, there is no consistent over- or underestimation of ε_r at the 8 pilot points. The MAP (maximum a posteriori) estimate, e.g. the peak of the marginalized PDF, at pilot points 4 and 7 are overestimates, whereas the estimate at pilot point 8 and the range are underestimates. There is no substantial difference in the MAP estimates and true values for the rest of the parameters. Thus our formulation and implementation seem to be correct and do not introduce bias in the results.

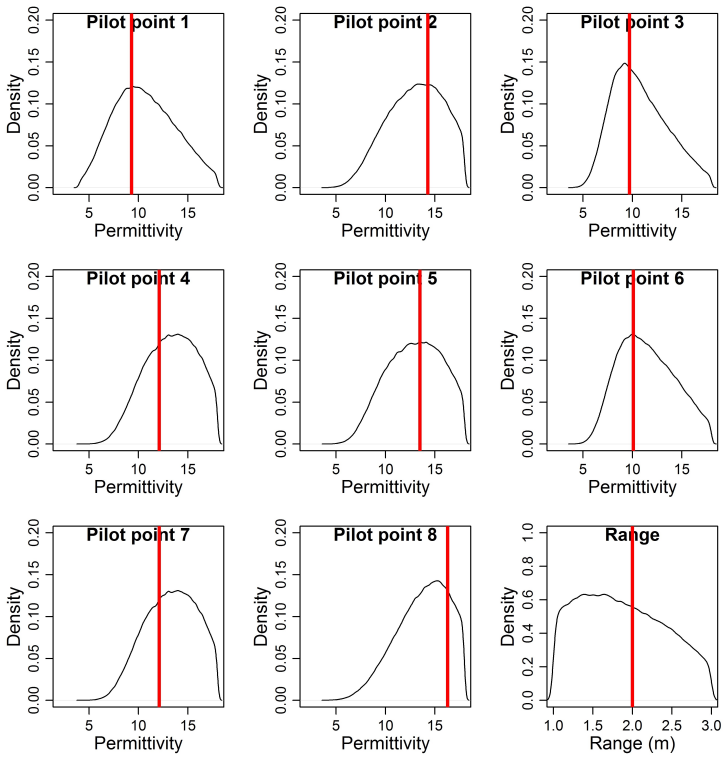


Figure 2: Posterior Densities for the GPR Problem based on 8 pilot points

Figure 3 shows the mean dielectric permittivity field for the case with 8 pilot points controlling the field. The mean dielectric permittivity field is the average of 200,000 realizations of the ε_r fields generated using samples randomly picked from the MCMC chains. Compared to the true field in Figure 1(a), the inferred mean field based on 8 pilot points can capture the main spatial variation of the field.

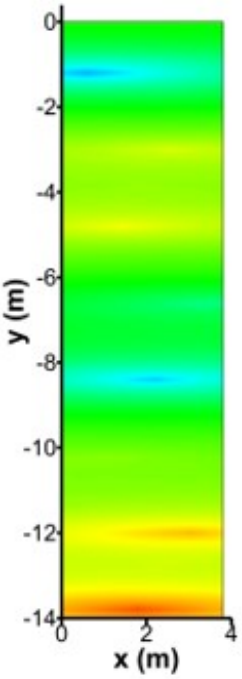


Figure 3: Mean of the inverted random field based on 8 pilot points and 200K MCMC samples

6 Bayesian Inversion of Seismic and Electromagnetic Data for Marine Gas Reservoir Characterization Using Multi-chain Markov Chain Monte Carlo Sampling

In this study, we developed an efficient Bayesian inversion framework for interpreting marine seismic amplitude versus angle (AVA) and controlled source electromagnetic (CSEM) data for marine reservoir characterization. The framework uses a multi-chain Markov-chain Monte Carlo (MCMC) sampler, which is a hybrid of DiffeRential Evolution Adaptive Metropolis and Adaptive Metropolis samplers. The inversion framework is tested by estimating reservoir-fluid saturations and porosity based on marine seismic and controlled source electromagnetic data.

In typical geophysical characterization, the existence of noise and the inadequacy (e.g., spatial and temporal coverage and resolution) of the data imply that the problem is ill constrained and therefore, geophysical characterization is a good target for statistical inference. Since there is usually an infinite number of models that can fit the data, it is useful to employ stochastic approaches (e.g., Bayesian), where unknowns are inferred in the form of a posterior probability density function (PDF), thus automatically quantifying the uncertainty in the estimates of the unknowns. The estimation problem is posed as a statistical inverse problem, which provides an expression for the posterior density of the unknowns of interest. The PDF is realized by drawing samples using a method such as Markov chain Monte Carlo (MCMC). The SAChES code was used to provide the multi-chain (i.e., parallel) MCMC framework for computational efficiency. Our MCMC procedure starts with 4 chains running DREAM (DiffeRential Evolution Adaptive Metropolis; Vrugt et al, 2009). When a sufficient number of samples have been collected by DREAM to make a useful proposal distribution, the MCMC method transitions to a parallel (4 chains) Adaptive Metropolis [19], implemented in a manner identical to Solonen's method [33]. In this work, we evaluated the efficiency and scalability of the multi-chain MCMC.

The study considered a five-layer reservoir model, similar to the synthetic model setup in (Hoversten et al., 2006), to demonstrate the accuracy and efficiency of the newly developed multi-chain MCMC-Bayesian approach. The unknowns include gas saturation and porosity in each layer in the reservoir. The results indicate that the seismic AVA and CSEM joint inversion provides better estimation of reservoir saturations than the seismic AVA-only inversion, especially for the parameters in deep layers. The performance of the inversion approach for various levels of noise in observational data was evaluated - reasonable estimates can be obtained with noise levels up to 25

The marginalized posterior PDFs of gas saturation and porosity using seismic-only inversion are shown in Figure 4. The PDFs indicate accurate estimation of porosity with relatively low uncertainty (i.e., narrow bounds) in each layer. For gas saturation estimation, seismic-only inversion infers the upper layers (e.g., layer 1) accurately, with relatively small uncertainty. The predicted modes of the PDFs are close to the actual values for the rest of the layers, but the uncertainty level of layer 3, 4 and 5 are large. This is expected because seismic AVA responses are less sensitive to gas saturation changes. Note that Figure 4 is a case with no noise.

As an example of some of the analysis we performed, Figure 5 shows the marginalized posterior PDFs of porosity using observations with various levels of noise. The porosity estimation is acceptable when the noise level is less than 25%; further, the uncertainty levels are comparable. When the noise level exceeds 25%, the multi-chain MCMC does not converge to the true parameter value and the chains cannot mix well.

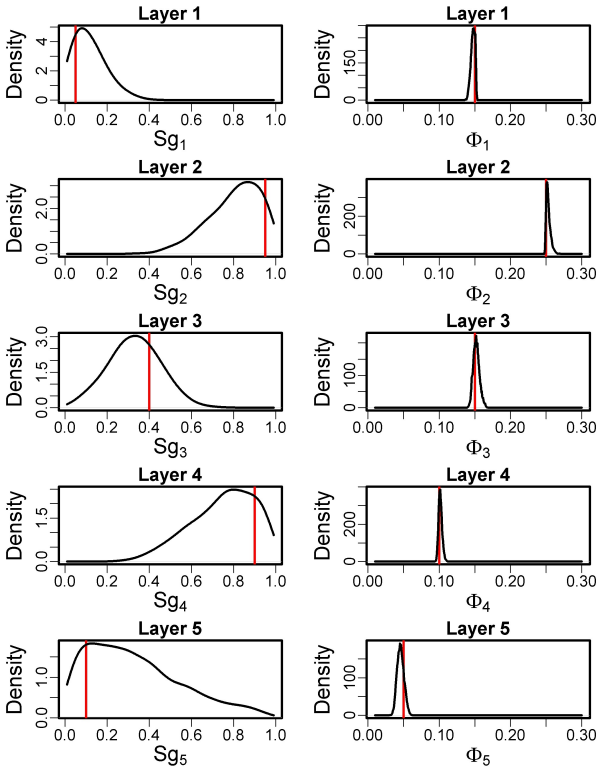


Figure 4: Posterior distribution of gas saturation parameter, Sg, and porosity, Φ, per layer based on seismic data only.

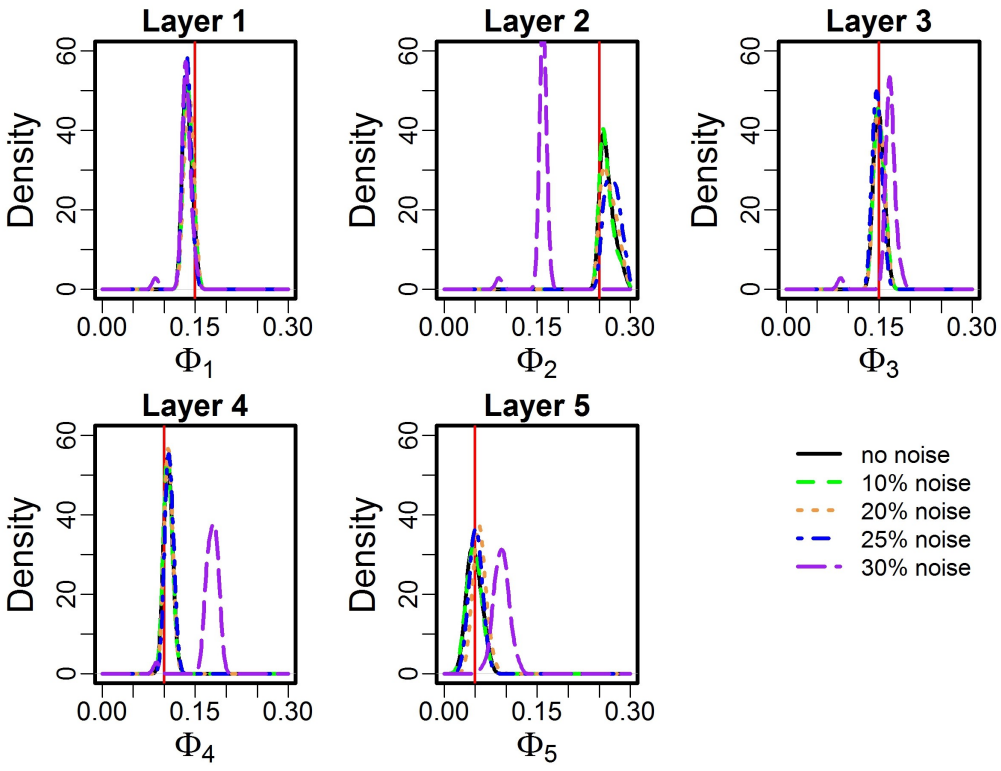


Figure 5: Posterior distribution of porosity with increasing observational noise.

7 Numerical Examples

7.1 Convergence

We constructed some simple numerical examples to test the SAChES approach. As we tested the performance of SAChES, we wanted to develop formal metrics for understanding the convergence with respect to parallel chains. This section describes some common approaches used to assess convergence of MCMC and also outlines the challenges with respect to convergence.

We have the following premise: unless the posterior is pathological, most of the chain is spent in refining the posterior. That is, after the initial exploration of the space with DREAM, we already have good estimates for regions around which to refine, and the rest of the MCMC portion of SAChES using AM involves sampling around those good posterior regions. If we need a number of samples Q to obtain accurate statistical estimators on the posterior, m chains should allow us to be approximately m times more efficient (e.g. split up the Q samples into Q/m).

There are a variety of metrics to test for chain convergence, but they measure different things and one must be careful in deciding which convergence metric to use. Here is a list of some common metrics [5], [3]:

1. Raftery-Lewis: This provides a way of bounding the variances of estimates of the quantiles. R-L is posed as "what number of samples are needed to obtain a quantile q to within an accuracy of +/- r with probability p?
2. McGibbsit: by Gregory Warnes, is a multi-chain extension of Raftery-Lewis. [13]
3. Gelman-Rubin: Start with multiple chains which begin at overdispersed starting points. The "shrink" factor approaches 1 when the within-chain variances dominates the between-chain variance, so that all chains have escaped the influence of their starting points.
4. Geweke: This approach uses spectral analysis to assess convergence where the intent is to understand changes in the mean of some function of the calibration parameters. This approach treats the Markov chain as a time series and uses spectral analysis to obtain variance estimates of subsequences via the spectral density function.

The problem with these metrics is that they all assume one long MCMC chain or multiple independent chains. We have neither in SAChES because our parallel chains do communicate. So, there are issues of how to aggregate the information from multiple chains. We cannot simply append the chains and treat that as one long chain. However, we can pool the information from multiple chains and examine statistics such as means, quantiles, etc. from the pooled chains. In practice, the pooled information is based on the appended chain, but we are not using it in a particular order but just to obtain statistics. For example, if we have 5 chains of length 1000 each, we have a pooled chain of 5000 which we can use to obtain the 5th and 95th percentiles.

To demonstrate how we might use pooled information across chains, we use a bootstrap approach. Bootstrapping is a resampling method with replacement that allows one to assign measures of accuracy (e.g. variance, confidence intervals) for complex estimators of a distribution, such as percentiles, proportions, etc. [9] For example, we might generate an MCMC chain with 10K samples. What is the uncertainty in the 95th percentile? With the bootstrap, we draw 500 samples, where each of the 500 samples is a resampling

of the 10K with replacement. For each of the 500 samples, we generate the 95th percentile, then examine the statistics over the 500. Our proposed convergence metric is then:

$$\text{Bootstrap metric} = \frac{\sigma(QoI)}{\mid QoI \mid} \tag{6}$$

This metric gives an indication of the size of the relative error in the QoI. As we increase the number of generations in the chain, this ratio can be used as a stopping criterion for an m-chain MCMC, regardless of what m is. The bootstrap metric is valid whether we have 1-chain or a pooled m-chain.

7.2 Convergence Results on a Gaussian Problem

In this section, we present convergence results for a 2-D problem, where the prior on each of the two variables, P1 and P2, is a uniform distribution from [-5,8] and the likelihood is a correlated bivariate Gaussian with mean of [1,3] and a correlation of 0.8.

Figure 6 shows the posterior samples from four SAChES chains. With uniform priors and a Gaussian likelihood, we expect the posterior to be Gaussian with a mean of [1,3] and a correlation of 0.8. The actual posterior mean is [0.99,2.99] with the correct correlation. The marginals of the two variables are shown in Figure 7 and Figure 8.

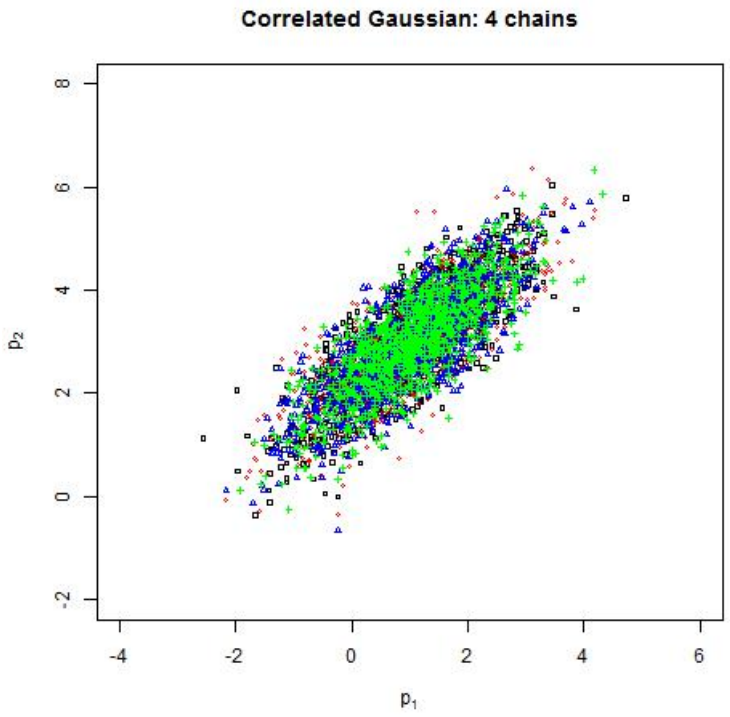


Figure 6: Posterior Samples of Four Chains shown in blue, black, red, and green for a problem with a uniform prior on 2 parameters (P1 and P2) and a Gaussian likelihood centered at (1,3) with correlation of 0.8.

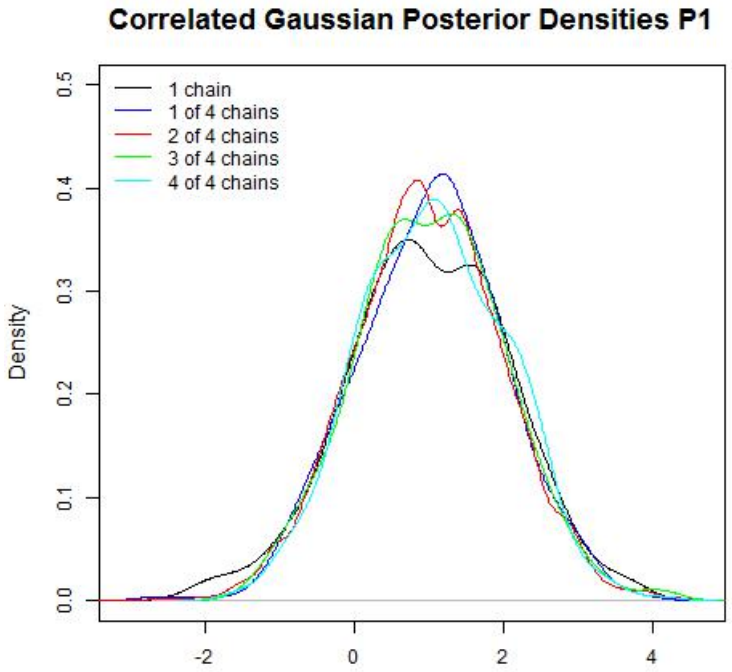


Figure 7: Posterior Density of parameter P1 comparison of 1 single chain vs. 4 chains run in SAChES.

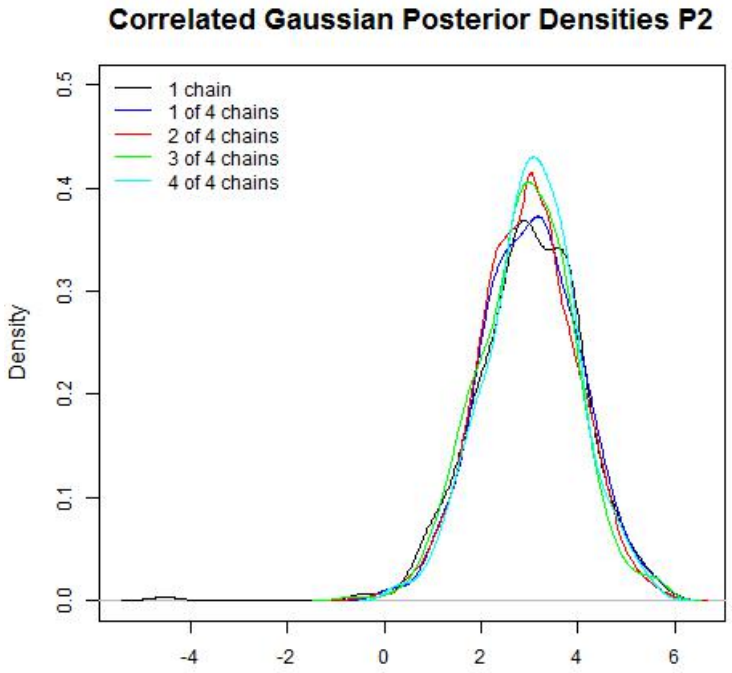


Figure 8: Posterior Density of parameter P2 comparison of 1 single chain vs. 4 chains run in SAChES.

Figures 9 and 10 show convergence of the percentiles of the chains as we increase the number of parallel chains to 16 chains. Note that these figures have a log scale on the x-axis which shows the number of generations of the MCMC. Figure 11 shows the bootstrap metric defined above for the interquartile range (75th percentile minus 25t percentile). This figure indicates that more chains do have more accurate statistics on the quantity of interest (in this case, the interquartile range), according to this bootstrap metric.

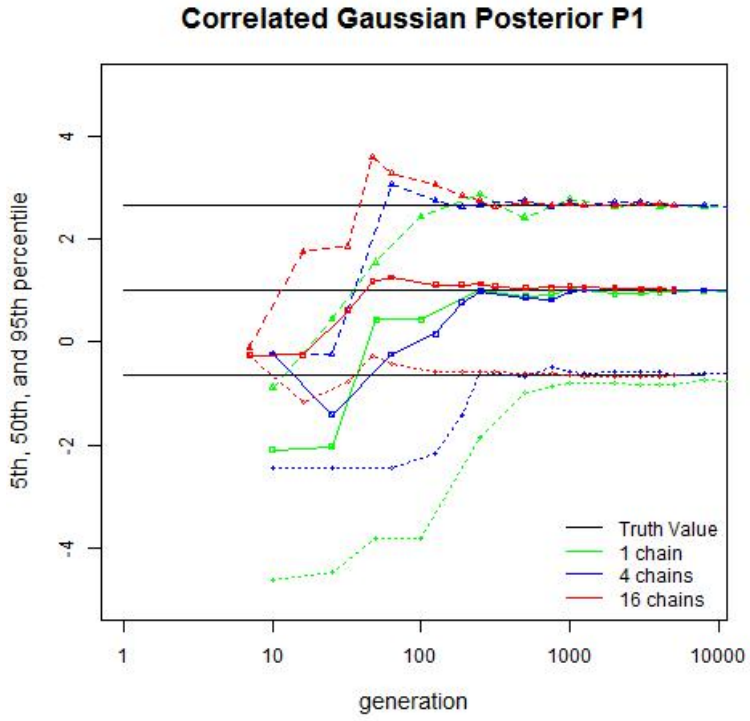


Figure 9: Chain Percentiles as a function of generation for P1

.

Finally, Table 1 shows some results of the typical convergence metrics. There are a few points to note from this table. The Raftery-Lewis statistics indicate that one would need only about 10K iterations if you wanted to obtain estimates of the fifth quantile q=0.05 at an accuracy of r=+/-0.01. However, this increases to around 40K iterations for each variable if the accuracy requirement is increased to 0.005. Finally, the Gelman-Rubin convergence metric is not as stringent, indicating fewer iterations.

7.3 Convergence Results on the Rosenbrock Problem

In this section, we present convergence results for a more challenging 2-D problem, where the prior on each of the two variables, P1 and P2, is a uniform distribution from [-2,3] and the likelihood is based on the Rosenbrock function. The Rosenbrock function is a "banana-shaped" function commonly used to test optimization algorithms because of the challenges its banana-shaped contours pose. In this problem, the likelihood function is:

$$Likelihood = \exp\left(-1 * ((1 - x_1)^2 + 100.0 * ((x_2 - {x_1}^2)^2)\right) \tag{7}$$

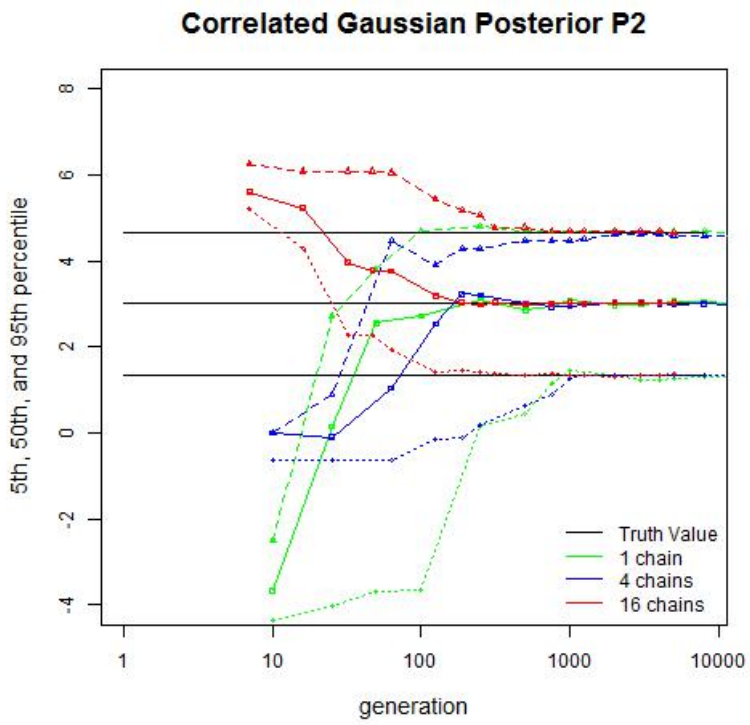


Figure 10: Chain Percentiles as a function of generation for P2

.

Table 1: Convergence Metrics, Correlated Gaussian 2-D Problem.

Number of chains	Iterations run	Convergence Metric	Samples Required [P1][P2]
1	100K	Raftery-Lewis q=0.05, r=0.01	[10K][10K]
4	20K/chain	Raftery-Lewis q=0.05, r=0.01	[10K][10K]
16	5K	Raftery-Lewis q=0.05, r=0.01	[9K][9K]
4	20K	Gelman-Rubin	[6K][6K]
16	5K	Gelman-Rubin	[3K][3K]
1	100K	Raftery-Lewis q=0.05, r=0.001	[41K][41K]
4	20K/chain	Raftery-Lewis q=0.05, r=0.001	[39K][40K]
16	5K	Raftery-Lewis q=0.05, r=0.001	[36K][39K]

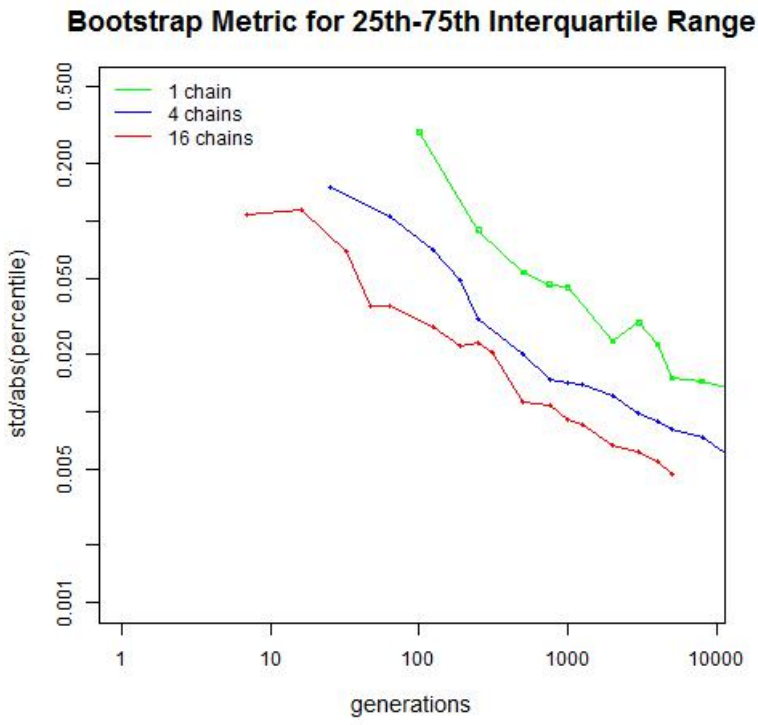


Figure 11: Bootstrap Convergence Metric for Interquartile Range (75th - 25th percentile) for Gaussian Problem

.

Figure 12 shows the posterior samples from four SAChES chains. One can visually see that with uniform priors and a likelihood governed by the Rosenbrock function, the posterior takes the banana shape of the Rosenbrock function. The marginals of the two variables are shown in Figure 13 and Figure 14.

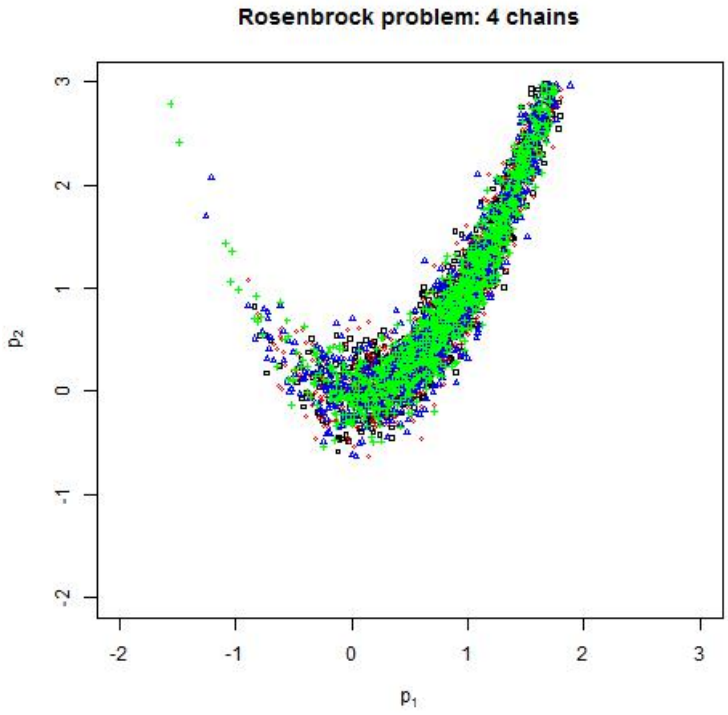


Figure 12: Posterior Samples of Four Chains shown in blue, black, red, and green for a problem with a uniform prior on 2 parameters (P1 and P2) and a likelihood based on the Rosenbrock function.

Figures 15 and 16 show convergence of the percentiles of the chains as we increase the number of parallel chains to 16 chains. Note that these figures have a log scale on the x-axis which shows the number of generations of the MCMC.Also note that in comparison to the Gaussian convergence plts shown in 9 and 10 which apppear to converge in 1000 iterations, these take longer and are not necessarily converged at 5000 iterations. Figure 17 shows the bootstrap metric defined above for the interquartile range (75th percentile minus 25t percentile). This figure indicates that more chains do have more accurate statistics on the quantity of interest (in this case, the interquartile range), according to this bootstrap metric.

Finally, Table 2 shows some results of the typical convergence metrics. While this table is similar to 1, there are some important differences. The convergence metrics indicate many more iterations are needed to converge parameter P1 vs P2, often a factor of 3 or 4. The Raftery-Lewis statistics indicate that one would need 14K iterations if you wanted to obtain estimates of the fifth quantile of P1 at an accuracy of r=+/-0.01, whereas only 4K iterations would be necessary to obtain a similar accuracy on the fifth quantile of P2 with one chain only. This increases to around 82K vs. 22K for P1 and P2 with a single chain, again to converge the fifth quantile but to an accuracy of r=+/-0.001. The multi-chain convergence metrics show similar trends. Again we see that the Gelman-Rubin metric is not nearly as stringent, indicating fewer iterations for convergence.

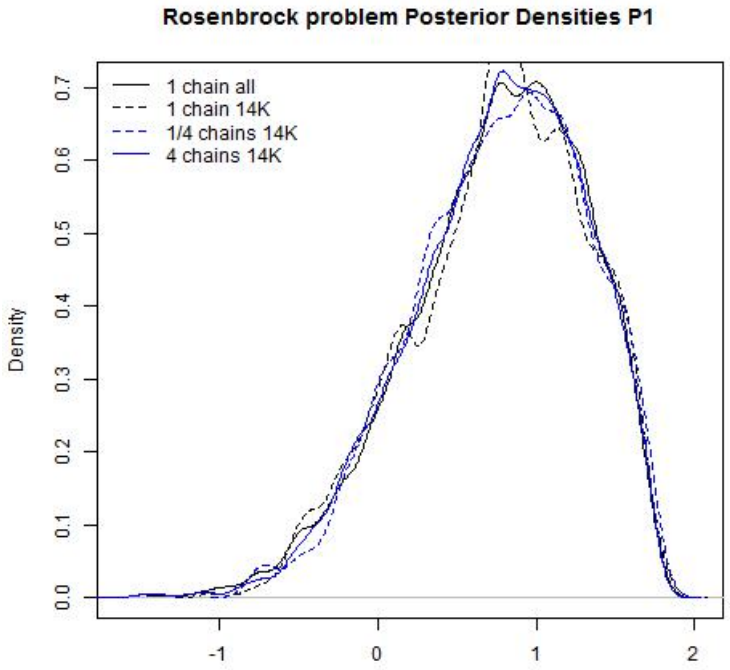


Figure 13: Posterior Density of parameter P1 comparison of 1 single chain vs. 4 chains for Rosenbrock problem.

Table 2: Convergence Metrics, Rosenbrock 2-D Problem.

Number of chains	Iterations run	Convergence Metric	Samples Required [P1][P2]
1	100K	Raftery-Lewis q=0.05, r=0.01	[14K][4K]
4	20K/chain	Raftery-Lewis q=0.05, r=0.01	[10.5K][4K]
16	5K	Raftery-Lewis q=0.05, r=0.01	[13K][4K]
4	20K	Gelman-Rubin	[5K][3K]
16	5K	Gelman-Rubin	[3K][3K]
1	100K	Raftery-Lewis q=0.05, r=0.001	[82K][22K]
4	20K/chain	Raftery-Lewis q=0.05, r=0.001	[62K][22K]
16	5K	Raftery-Lewis q=0.05, r=0.001	[79K][23K]

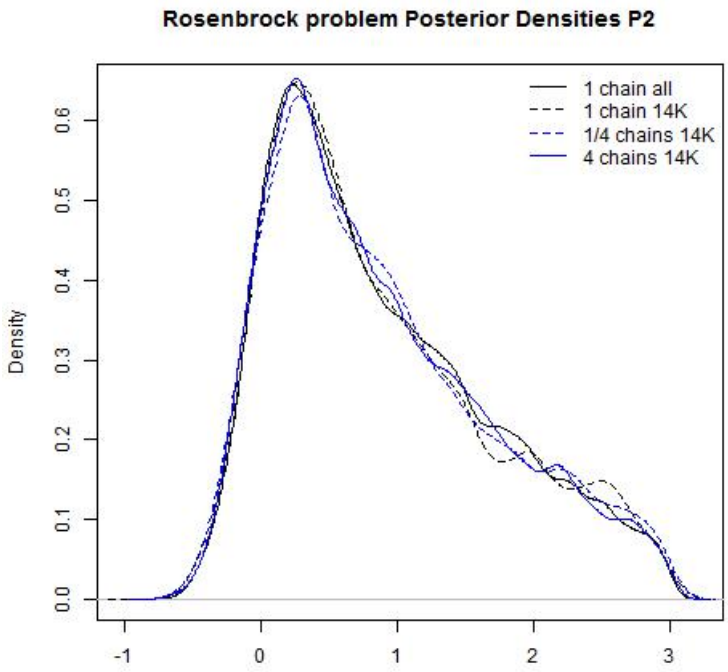


Figure 14: Posterior Density of parameter P2 comparison of 1 single chain vs. 4 chains for Rosenbrock problem.

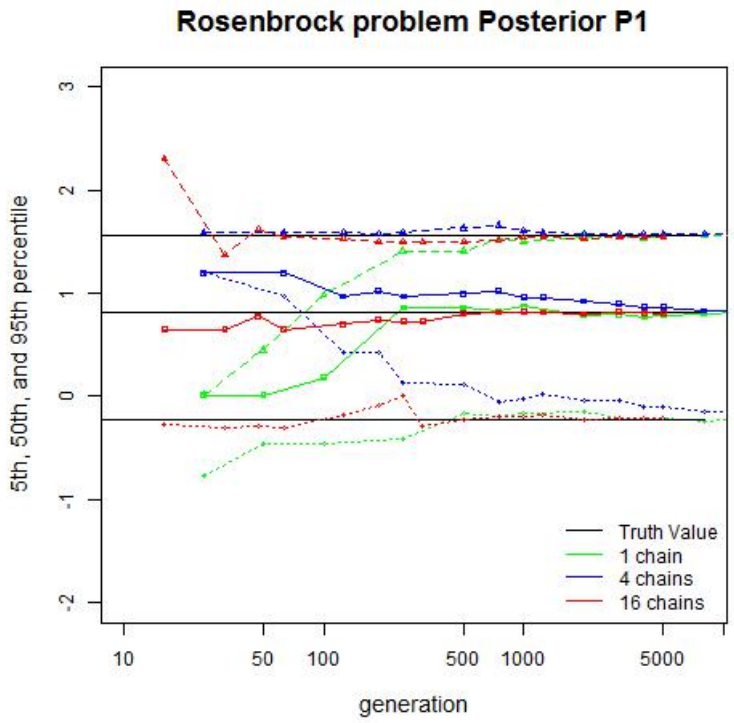


Figure 15: Chain Percentiles as a function of generation for P1, Rosenbrock likelihood

.

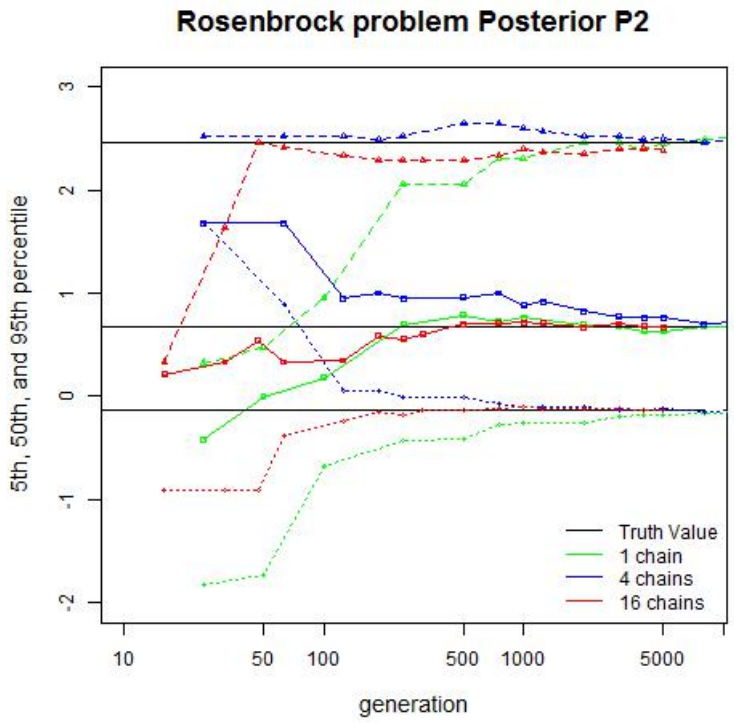


Figure 16: Chain Percentiles as a function of generation for P2, Rosenbrock likelihood

.

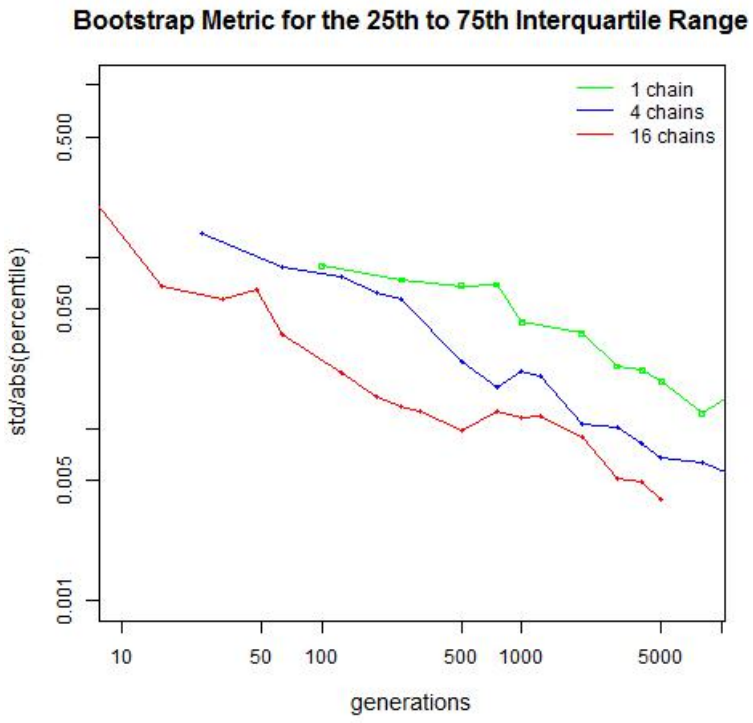


Figure 17: Bootstrap Convergence Metric for Interquartile Range (75th - 25th percentile) for Rosenbrock Problem

.

7.4 Convergence analysis using quantiles

In this subsection we address the question whether an adaptive MCMC method, with mutiple chains, provides an advantage over a single adaptive chain. The hypothesis is that a multitude of Markov chains can explore a posterior distribution far better than a single chain; further, if the chains pool their samples occasionally to construct a proposal distribution that reflects their random walks, it may be possible to arrive an "good" proposal distribution faster than if the proposal distribution were to be limited to the samples collected by a single chain. However, clearly, the behavior of the ensemble of such chains will be affected by the starting points of the individual chains (a multitude of chains, starting at poor positions in the posterior distribution, could destablize the evolution of a few that start at good positions). Further, the acceleration in convergence to the true posterior distribution might become apparent only later in the evolution when the proposal distribution begins to resemble the posterior distribution. We examine these questions below, using the reciprocal of a 5D Rosenbrock function masquerading as a posterior distribution. The posterior distribution will be sampled by an 8-chain, adaptive MCMC methods (SAChES), and tracked by computing the 5^{th} and 95^{th} percentiles, and the median of each of the five parameters. For completeness, we write the analytical form of the distribution being sampled as:

$$\log\left(P(\mathbf{x})\right) = -\sum_{i=1}^{N-1}\{100(x_{i+1}-x_i^2)^2+(1-x_i)^2\}, \qquad -2 \le x_i \le 3. \tag{8}$$

The first step consits of determining a set of good and bad starting points for the chains. We define an equispaced grid in the parameter space, 15 points per dimension, and evaluate the log-posterior distribution. The 0.01 and 0.001 quantiles of the 15^5 log posterior density values are determined and we choose 8 points of each type; we call them the $q_{0.1}$ and $q_{0.001}$ set of starting points in the $\mathbf{x}$ space. We first check whether communication between chains affects the convergence of the distribution Eq. 8.

In Fig. 18, we plot the 5^{th} and 95^{th} quantiles, as well as the median of x_i, $i \ldots 5$ of the posterior distribution, as a function of the number of samples collected my the MCMC method. The chains are started from the $q_{0.001}$ points. The blue line indicates the results from SAChES (i.e., adaptive, communicating chains) while the red lines are from 8 separate, 1-chain runs of the Adaptive Metropolis sampler that underlies SAChES. The green dashed line is the final converged value of the quantile. We see that SAChES (adaptive comunicating chains) has an advantage only when we desire the MCMC chain to provide a posterior distribution very close to the converged value; for coarse levels of convergence, the single-chain MCMC does as well as the 8-chain method. Repeating this test using $q_{0.1}$ starting points showed that there was no difference between the 8-chain and 1-chain methods; the starting points were sufficiently good that the pooling of information across 8-chains provided no better estimate of the posterior (for constructing a proposal) than a single-chain MCMC itself.

We next check if the starting point affect the rate of convergence of SAChES (adaptive communicating chains). In Fig. 19, we plot the medians of the marginal distributions for x_i, $i \ldots 5$ for 2 8-chain SAChES runs, one starting from $q_{0.001}$ points (blue line) and the other from $q_{0.1}$ points (red line). Dashed green lines denote the converged values of the median. We see that the $q_{0.001}$ starting points lead to Markov chains that converge somewhat slower. This leads to the question whether a set of chains starting from a mixture of good and bad starting points (the case most likely to occour in realistic settings) would be destabilized by the bad starting points. This is investigated in Fig. 20, where we start off a 8-chain SAChEs run from a set of 4 $q_{0.1}$ and 4 $q_{0.001}$ starting points (red line). It is compared with a run started from 8 $q_{0.001}$ points (blue line). Only the medians of the marginal distributions for $x_i, i = 1 \ldots 5$ are plotted. are plotted. We see that the chains starting from the $q_{0.001}$ points are slower to converge i.e., the "good" chains starting from $q_{0.1}$

points help accelerate the convergence of the "bad" chains that started from $q_{0.001}$ points.

To summarize, communicating chains accelerated convergence to the posterior distribution only if tight convergence tolerances are desired. Also, commuicating chains are robust i.e., if a subset of chains in SAChES start from bad starting points, the rest of the chains bring them back into higher density region of the parameter space.

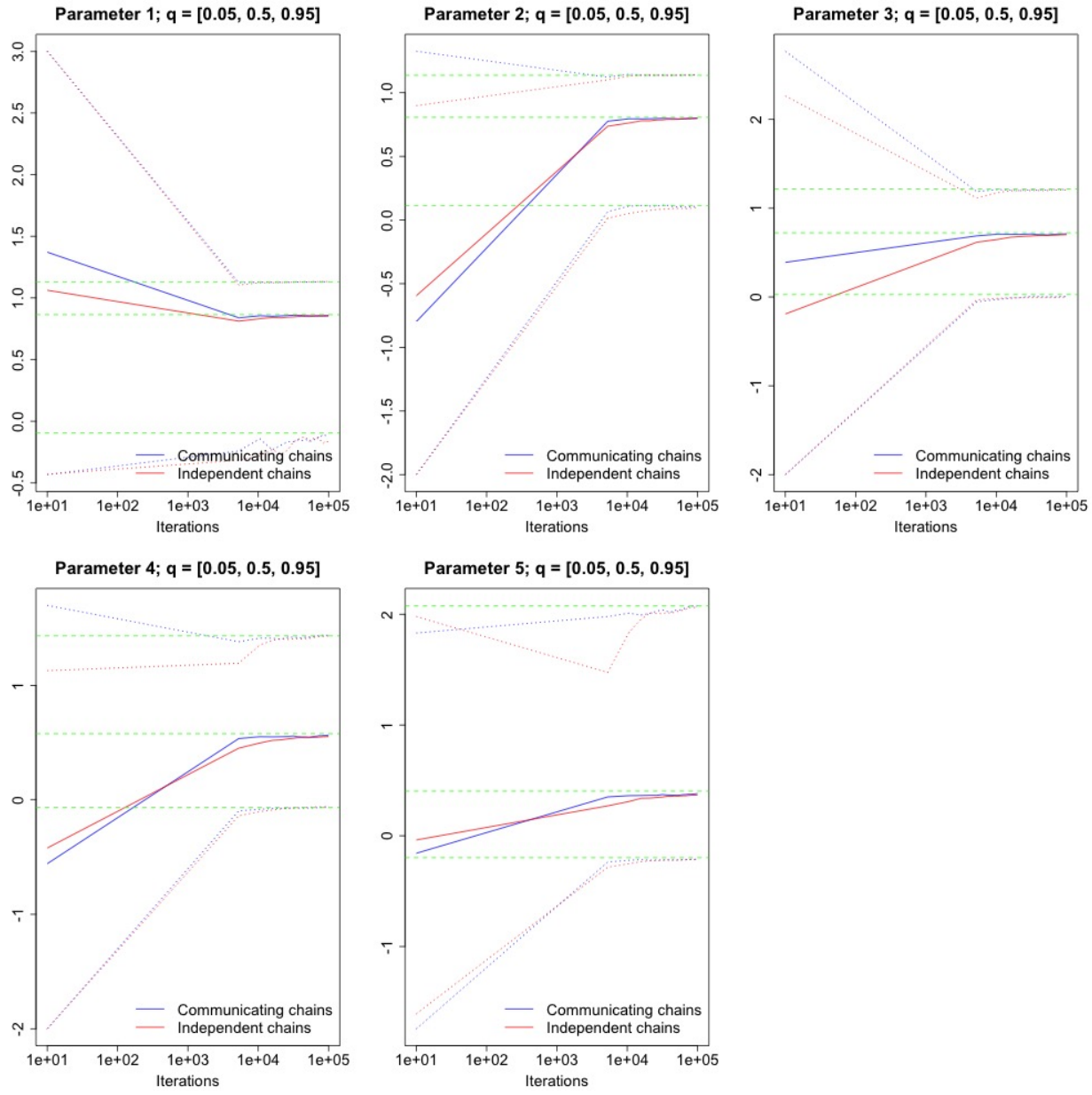


Figure 18: Plot of the convergence of 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles (dashed lines) and the median (solid line) of the five independent parameters $x_i, i = 1 \ldots 5$ of the 5D Rosenbrock distribution (Eq. 8), plotted as a function of the number of samples collected by the MCMC chain.. The blue lines are for the distriution obtained SAChES (8 communicating adaptive chains). The red lines are for 8 adaptive chains which do *not* communicate. The dashed green horizontal lines are the converged values of the quantiles. Note that the horizontal axes are log-transformed.

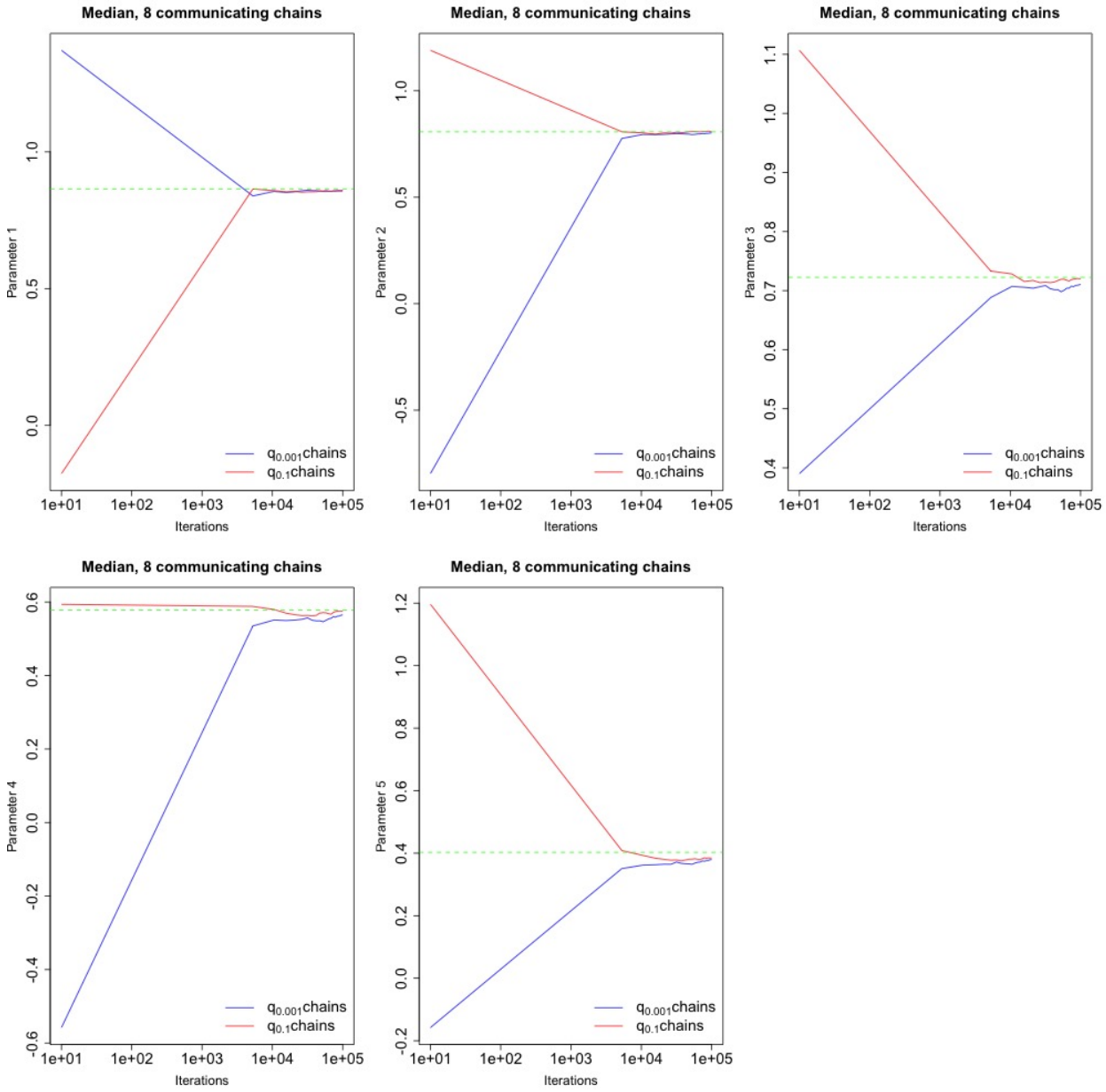


Figure 19: Convergence of the median of the marginal distributions of $x_i, i = 1 \ldots 5$, computed using SAChES with 8 chains. The blue lines plot chains started from $q_{0.001}$ starting points ("bad points") and the red chains start from $q_{0.1}$ starting points ("good points"). The green dashed lines are the converged values. The horizontal axis is log-transformed and shows the number of the samples collected by the MCMC chain.

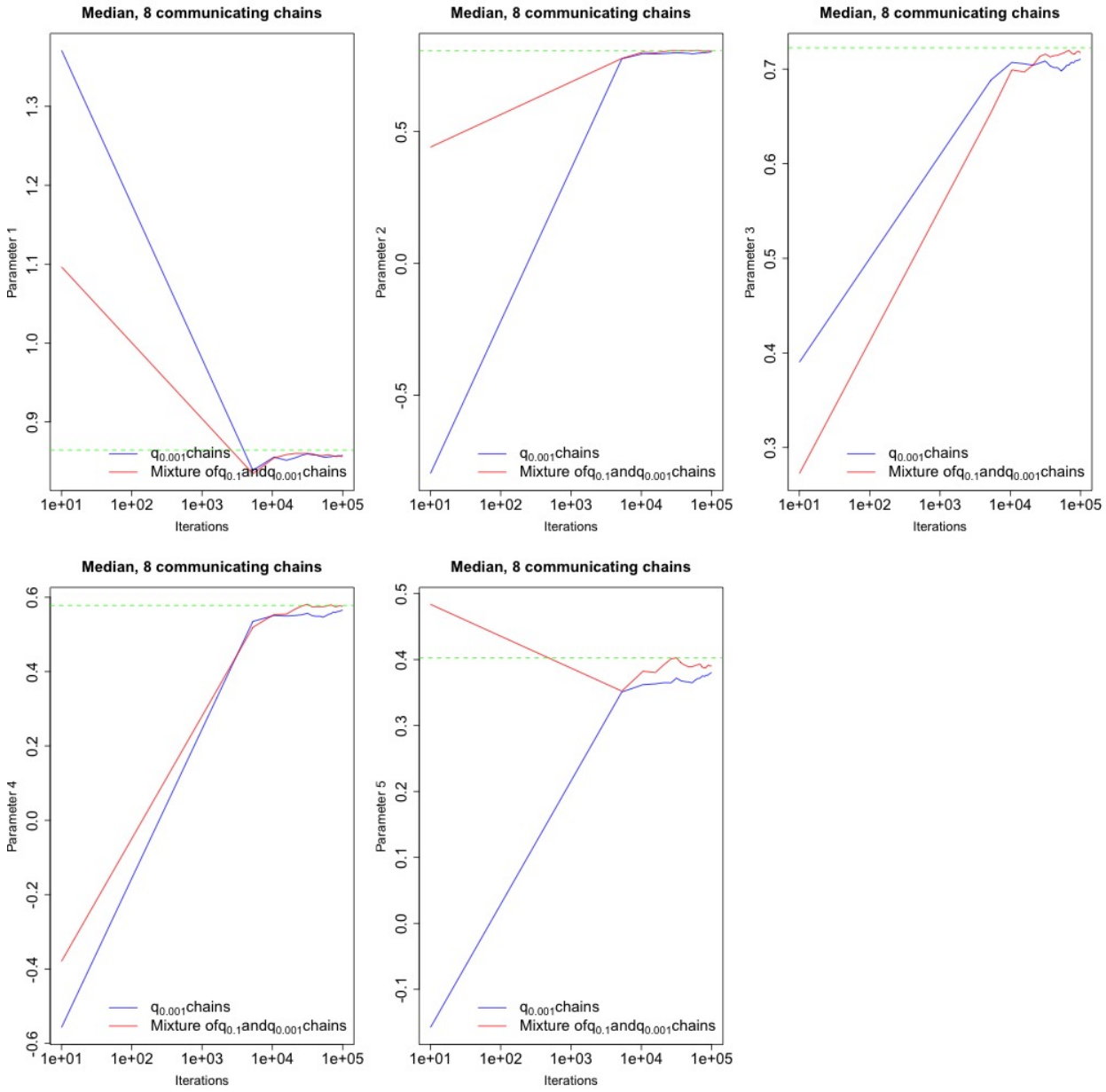


Figure 20: Convergence of the median of the marginal distributions of $x_i, i = 1 \ldots 5$, computed using SAChES with 8 chains. The blue lines plot chains started from $q_{0.001}$ starting points ("bad points") and the red chains start from an equal mixture of $q_{0.001}$ and $q_{0.1}$ starting points. The green dashed lines are the converged values. The horizontal axis is log-transformed and shows the number of the samples collected by the MCMC chain.

8 A new Bayesian calibration approach using Voronoi Global Optimization (VGOpt)

8.1 Motivation

This study presents a novel solution to Bayesian parameter calibration. Bayesian parameter calibration requires an efficient mechanism to sample from a multivariate, typically high-dimensional, posterior probability distribution π^o given knowledge or assumptions about a prior probability distribution π^i and a likelihood function ℓ as given by the Bayesian framework:

$$\pi^o = \pi^i \times \ell \tag{9}$$

Our goal here is to produce samples from the posterior distribution of model parameters that are calibrated in a Bayesian framework without using a Markov Chain Monte Carlo (MCMC) sampling method. There are several disadvantages of MCMC: it is slow to converge, often requiring hundreds of thousands of samples, where a "sample" involves running a computational model and using the result to calculate the likelihood of accepting this sample. MCMC methods also have a high rejection rate, meaning that majority of the sample points explored are not even included in the posterior. Finally, because of the computational cost, surrogate models are often used instead of the actual computational simulation to calculate the likelihood. This can make the calibration process less accurate and computationally expensive.

To avoid the complexity and slow convergence disadvantages of MCMC, we present a new Voronoi Global Optimization (VGOpt) technique, that only uses a small amount of simulation samples. It starts with a very small sample number such as $2d$, where d is the number of parameters being inferred. Then, samples are adaptively added up to the user-specified budget N, where N might be as low as 100 or 500 samples, for example. The process by which samples are added involves a global optimization procedure that relies on the Voronoi domain decomposition of the domain. It is similar to the DIRECT domain decomposition method, but more efficient. The Voronoi-based approach picks the best Voronoi vertex for refinement, rather than splitting cells as DIRECT does.

The entire approach relies on the Voronoi cells *implicitly* defined by the sample points. Within each Voronoi cell, the probability density function of the posterior is taken to be constant at the value given by the PDF at the cell seed. At the end of the sampling process, the user is presented with a list of the N sample points and their densities, along with an approach that can easily generate M new samples from the posterior without rejection as MCMC does.

8.2 Voronoi Global Optimization (VGOpt) Algorithm

In this section, we describe the details of the proposed VGOpt algorithm and illustrate its adaptive sampling mechanism and explain how to quickly sample from the approximated posterior.

We reformulate the posterior estimation problem in a global optimization sense: samples are adaptively added where the highest probability is predicted. To achieve that, consider the Voronoi domain decomposition, implicitly defined around existing samples and use a consistent global optimization metric to decide where to sample next. This approach mimics the DIRECT algorithm [24] which recursively subdivides the parameter space into hyperrectangles and iteratively refines one cell into three based on the function value at the cell's sampled center, the distance from the center to a farthest point of the rectangle, and an unknown

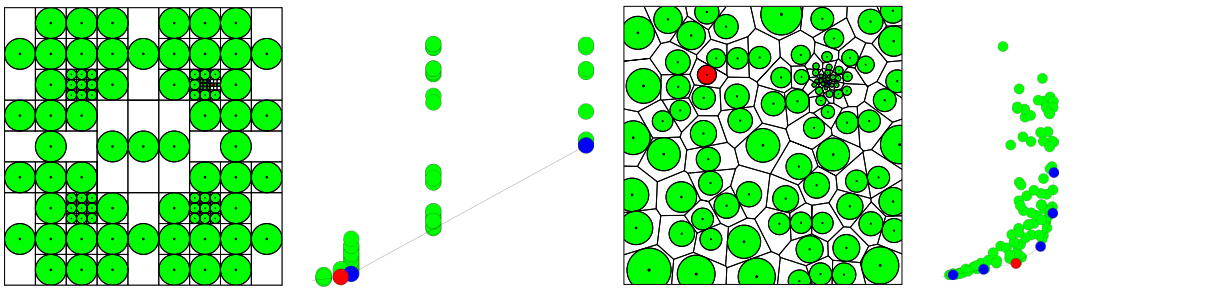
Lipschitz constant. This approach often gets stuck in refining a cell further and further, if it neighbors another cell with the true global optimal point.

Alternatively, VGOpt subdivides the parameter space into approximate Voronoi cells (instead of rectangles) around each sample, and points are added one at a time. However, to avoid the no-vertex case of the very first single point, we start with $2d$ well-spaced points (e.g., 4 points in 2-d). Then, to choose a point for addition, a Voronoi cell is picked using Monte Carlo sampling, and a candidate Voronoi vertex is sampled from the boundaries of that cell. We can also sample a point from a Voronoi facet of that cell (or higher order elements) if vertices were not detected. Sampling Voronoi vertices in high-dimensional spaces is enabled via the Recursive Spoke-Darts (RSD) capability [7]. A candidate Voronoi vertex v is then associated with a distance h from those $d+1$ samples $\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}$ that are equidistant to it, and a probability evaluation f_v, where

$$f_v = \min_{\mathbf{x}_1, \ldots \mathbf{x}_{d+1}} (-f_{\mathbf{x}}). \tag{10}$$

We then construct the convex hull that lower-bounds the $h - f$ behavior and choose a random vertex to be the new sample. This operation terminates when a maximum number of successive misses m is reached, where a "miss" is defined as choosing a vertex that does not change the convex hull construction.

Figure 21 illustrates the basic comparison between DIRECT and VGOpt operations.



(a) Direct sampling (b) h-f convex hull (Direct) (c) VGOpt sampling (d) h-f convex hull (VGOpt)

Figure 21: DIRECT domain decomposition technique versus VGOpt sampling mechanism at Voronoi vertices for $N = 100$ samples.

Sampling From the Posterior

Once the sample budget in exhausted, a piecewise-constant surrogate model is implicitly constructed over the final sample set and their associated probability density values. Each Voronoi cell picks the same probability value at its seed. Note that higher order surrogate models can also be used over different Voronoi cells [32]. To enable future sampling from the approximated posterior PDF, we estimate the volumes of the Voronoi cells using exhaustive Monte Carlo sampling and adjust the values at the seeds by a constant to ensure a probability sum of 1 over the entire domain. We then construct a CDF over Voronoi cells of the probability values at their seeds multiplied by their volume ratios of the total domain volume. We pick a random value $r \in [0, 1]$, choose the cell whose CDF contribution added up to r, and randomly sample from that cell. Larger cells are expected to have very low PDF values so they are less probable to be picked. Cells with larger PDF values are more probable to be picked, especially if they have larger volumes. VGOpt

operations can be easily repeated to result in several experimental realizations with a few number of samples for each iteration. This sampling process is bias-free. The only source of error stems from the accuracy of estimating the posterior.

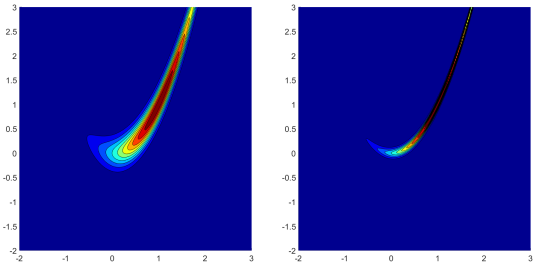
8.3 Experimental Results

For illustration, we apply the basic VGOpt operations to the Rosenbrock function. This function is commonly used to test optimization algorithms because of the challenges its banana-shaped contours pose. In this problem, the likelihood function ℓ is given by the exponential of the Rosenbrock function:

$$\ell = \exp\left(-((1-x_1)^2 + c*((x_2 - {x_1}^2)^2)\right) \tag{11}$$

where x_1 and x_2 are uniformly distributed parameters, over the domain $[-2,3]^2$, and c is a constant that determines how narrow the banana shape is, as demonstrated in Figure 22. In this experiement, we assume a uniform prior distribution, at which the posterior to be sampled is given by:

$$\pi^o = \ell \tag{12}$$



(a) $c = 10$ (b) $c = 100$

Figure 22: 2d isocontours of the Rosenbrock function in (11) for different values of c.

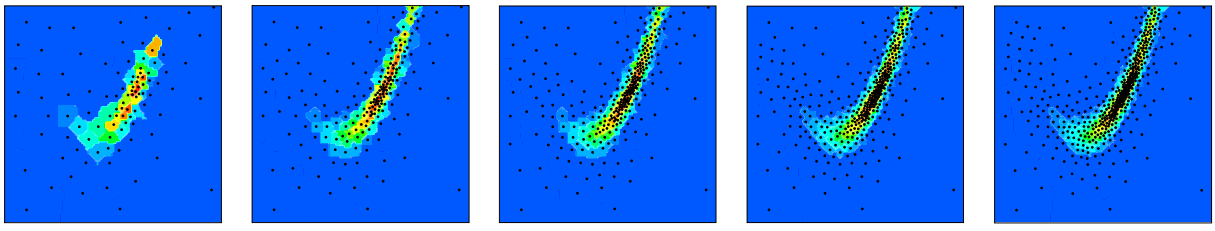
Figures 23 and 24 show experimental results when testing the Rosenbrock function with $c = 10$ and $c = 100$, respectively.

Comparison to Non-Adaptive Sampling

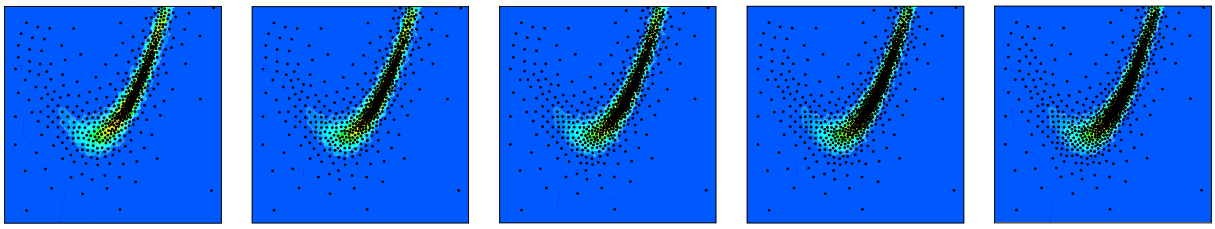
For comparison purposes, we consider non-adaptive sampling using random well-spaced points. Figure 25 shows posterior PDF estimation using increasing number of well-spaced points. Clearly, the performance of this sampling approach lacks the accuracy that VGOpt provides. Further illustration of this conclusion is provided via convergence studies next.

8.4 Convergence Studies

Figure 26 shows the performance of VGOpt in 2d for $c = 10$ and $c = 100$, as well as in 5d for $c = 10$. For comparison purposes, we also show the convergence behavior of well-spaced random sampling. For the

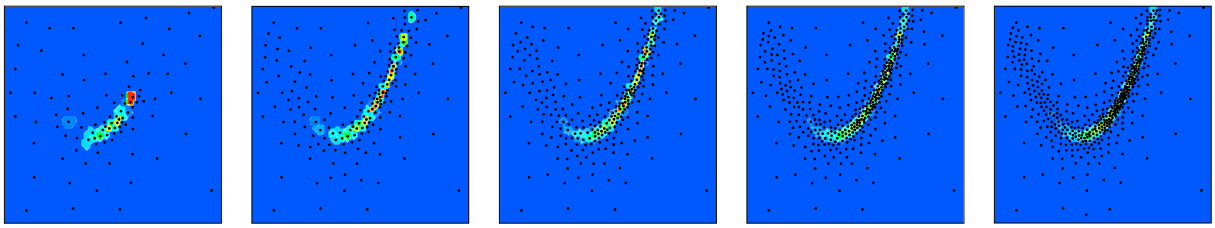


(a) $N = 100$ (b) $N = 200$ (c) $N = 300$ (d) $N = 400$ (e) $N = 500$

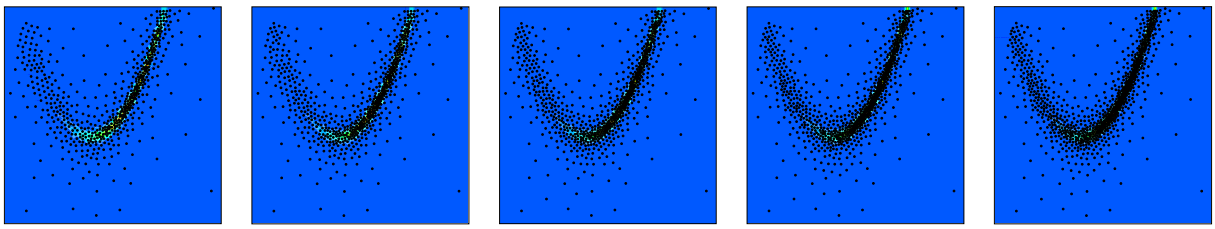


(f) $N = 600$ (g) $N = 700$ (h) $N = 800$ (i) $N = 900$ (j) $N = 1000$

Figure 23: 2d illustration of VGOpt using the Rosenbrock function with $c = 10$.

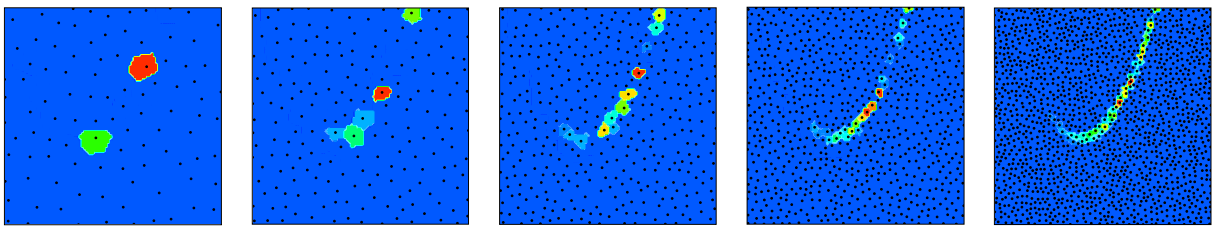


(a) $N = 100$ (b) $N = 200$ (c) $N = 300$ (d) $N = 400$ (e) $N = 500$



(f) $N = 600$ (g) $N = 700$ (h) $N = 800$ (i) $N = 900$ (j) $N = 1000$

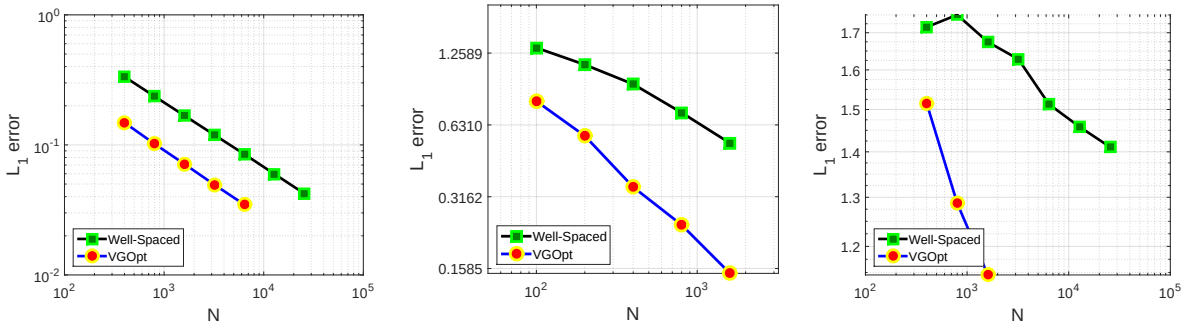
Figure 24: 2d illustration of VGOpt using the Rosenbrock function with $c = 100$.



(a) $N = 100$ (b) $N = 200$ (c) $N = 400$ (d) $N = 800$ (e) $N = 1600$

Figure 25: 2d illustration of sampling the posterior with random well-spaced points.

same dimension, a steeper PDF (e.g., higher value of c for the Rosenbrock test function) was captured more accurately using VGOpt compared to uniform random sampling. It is also important to observe that for the same constant c, VGOpt provides more gain for higher dimension at the same sample count.



(a) $2d, c = 10$ (b) $2d, c = 100$ (c) $5d, c = 10$

Figure 26: Error comparisons of VGOpt versus well-spaced random sampling at different values of c and d. Each data point is the average of 20 experiments.

8.5 Conclusion

We introduced a novel approach to solve the calibration problem without relying on costly MCMC sampling. Using implicit Voronoi domain decompositions, we broke down the posterior sampling problem into two pieces: 1) piecewise posterior estimation using a novel global optimization technique that relies on Voronoi partitioning, and 2) a bias-free method to directly sample for the estimated piecewise posterior.

9 Conclusions

The scientific achievement of this work was the development of a scalable Markov chain Monte Carlo method called SAChES to solve statistical inverse problems. The SAChES method is based on a combination of genetic algorithms, machine learning and multichain Markov chain Monte Carlo sampling. The differential evolution approach was used to get an efficient starting configuration for multi-chain MCMC. We also developed a novel method to encode efficient prior distributions using support vector machine classifiers, and we have a framework for streamlined and automated construction of surrogate models. Our method is designed for calibrating complex, computationally expensive models, conditional on sparse data. The SAChES method estimates parameters as a joint probability density function, so that the estimation uncertainty is also captured; in the presence of limited data and structural errors, these can be large. Methodological details are provided in [31].

The method was tested on parameter estimation (using the Community Land Model, CLM) and inference of fields (estimation of soil moisture and gas saturation using measurements of seismic and EM waves). For the CLM model, the inference process using SAChES enabled accurate site-scale predictions and revealed a dependency between CLM hydrologic parameters and intrinsic land properties such as soil type and vegetation. We estimated CLM parameters using observations from twelve different sites. The solution also contained a multivariate Gaussian model for the model-data mismatch. Calibration allowed CLM, typically configured to be accurate only at global scales, to provide improved forecasts at the local scale. Further, the data-model mismatch, conditional on data, provided information on processes modeled in CLM that contribute most to the structural error. [22]

The inference method has been tested with 20 chains, and up to 24-dimensional inverse problems (inference of a 24-pilot-point random field, using ground penetrating radar data). The method is implemented using MPI-2, to enable one-sided communication and dynamic process spawning, and is resilient to processor failures.

References

[1] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Signal Processing*, 19(1), 2010.

[2] S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–445, 1998.

[3] S. P. Brooks and G. O. Roberts. Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing*, 8(4):319–335, 1998.

[4] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound-constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1995.

[5] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[6] R. Craiu, J. Rosental, and C. Yang. Learn from thy neighbor: Parallel-chain regional adaptive MCMC. *Journal of the Americal Statistical Association*, (104):1454–1466, 2009.

[7] Mohamed S Ebeida and Ahmad A Rushdi. Recursive spoke darts: Local hyperplane sampling for delaunay and voronoi meshing in arbitrary dimensions. *Procedia Engineering*, 163:110–122, 2016.

[8] W. N. Edeling, P. Cinnella, R. P. Dwight, and H. Bijl. Bayesian estimates of parameter variability in the k-ε turbulence model. *Journal of Computational Physics*, 258:73–94, 2013.

[9] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, FL, 1993.

[10] M. Emory, R. Pecnik, and G. Iaccarino. Modeling structural uncertainties in Reynolds-Averaged computations of shock/boundary layer interactions. In *49th AIAA Aerospace Sciences Meeting*, 2011.

[11] J. W. Hurrell et. al. The Community Earth System model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360, 2013.

[12] Y. Q. Luo et al. A framework for benchmarking land models. *Biogeosciences*, 9(10):3857–3874, 2012.

[13] R. Burrows G. R. Warnes. Warnes and raftery's mcgibbsit mcmc diagnostic. *R Documentation*, 2013.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, chapter Model checking and improvement. 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] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[17] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[18] M. Gohler, J. Mai, and M. Cuntz. Use of eigendecomposition in a parameter sensitivity analysis of the community land model. *Journal of Geophysical Research: Biogeosciences*, 118(2):904–921, 2013.

[19] Heikki Haario, Marko Laine, Antoinietta Mira, and Eero Saksman. DRAM-Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

[20] Z. Hou, M. Huang, L. R. Leung, G. Lin, and D. M. Ricciuto. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. *Journal of Geophysical Research*, 117, 2012. D15108.

[21] M. Huang, Z. Hou, L. R. Leung, Y. Ke, Y. Liu, Z. Fang, and Y. Sun. Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model âĂŞ Evidence from MOPEX basins. *Journal of Hydrometeorology*, 2013.

[22] Maoyi Huang, Jaideep Ray, Zhangshuan Hou, Huiying Ren, Ying Liu, and Laura 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.

[23] H. Järvinen, P. Räisänen, M. Laine, J. Tamminen, A. Lin, E. Oja, A. Solonen, and H. Haario. Estimation of ECHAM5 climate model closure parameters with adaptive MCMC. *Atmospheric Chemistry and Physics*, 10:9993–10002, 2010.

[24] Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of optimization Theory and Applications*, 79(1):157–181, 1993.

[25] J.-C. Jouhaud, P. Sagaut, B. Enaux, and J. Laurenceau. Sensitivity analysis and multiobjective optimization for LES numerical parameters. *Journal of Fluid Engineering*, 130:021401, 2008.

[26] 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.

[27] Eric Laloy, Bart Rogiers, Jasper A. Vrugt, Dirk Mallants, and Diederik Jacques. Efficient posterior exploration of a high-dimensional groundwater model from two-stage markov chain monte carlo simulation and polynomial chaos expansion. *Water Resources Research*, 49(5):2664–2682, 2013.

[28] J. Laurenceau and P. Sagaut. Building efficient response surfaces of aerodynamic functions with kriging and cokriging. *AIAA Journal*, 46(2):498–507, 2008.

[29] K. W. Oleson, D. M. Lawrence, G B. Bonan, M. G. Flanner, E. Kluzek, P. J. Lawrence, S. Levis, S. C. Swenson, and P. E. Thornton. Technical description of version 4.0 of the Community Land Model (CLM), 2010.

[30] A. Raftery and Steven M. Lewis. Implementing MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 115–130. Chapman and Hall, 1996.

[31] J. Ray, Z. Hou, M. Huang, K. Sargsyan, and L. Swiler. Bayesian calibration of the community land model using surrogates. *SIAM/ASA Journal on Uncertainty Quantification*, 3:199–233, 2015.

[32] Ahmad A Rushdi, Laura P Swiler, Eric T Phipps, Marta D'Elia, and Mohamed S Ebeida. Vps: Voronoi piecewise surrogate models for high-dimensional data fitting. *International Journal for Uncertainty Quantification*, 7(1), 2017.

[33] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen. Efficient mcmc for climate model parameter estimation: Parallel adaptive chains and early rejection. *Bayesian Analysis*, 7(3):715–736, 09 2012.

[34] Y. Sun, Z. Hou, M. Huang, F. Tian, and L. Ruby Leung. Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model. *Hydrology and Earth System Sciences*, 17:4995–5011, 2013.

[35] Cajo J. F. ter Braak and Jasper A. Vrugt. Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446, 2008.

[36] J. A. Vrugt, C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290, 2009.

[37] X. Zeng, B. A. Drewniak, and E. M. Constantinescu. Calibration of the crop model in the Community Land Model. *Geosciences Model Development Discussions*, 6:379–398, 2013.

DISTRIBUTION:

1 Steven Lee
Advanced Scientific Computing Research
U.S. Department of Energy
SC-21, Germantown Building
1000 Independence Ave., SW
Washington, DC 20585-1290

1 Jie Bao
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352

1 Maoyi Huang
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352

1 Zhangshuan Hou
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352

1 Huiying Ren
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352

1 MS 1318 J.R. Stewart, 01440

1 MS 1318 M.S. Ebeida, 01441

1 MS 1318 L.P. Swiler, 01441

1 MS 1323 D.Z. Turner, 01441

1 MS 9158 G.E. Thayer, 08959

1 MS 9159 J. Ray, 08959

1 MS 0899 Technical Library, 8944 (electronic copy)

Sandia National Laboratories