SAND2018-0463J

Enhancing Model Predictability for a ScramJet Using Probabilistic Learning on Manifolds

Christian Soize[*]
Université Paris-Est Marne-la-Vallée, Marne-la-Vallée, 77454, France.

Roger Ghanem[†]
University of Southern California, Los Angeles, CA 90089, USA

Cosmin Safta,[‡] Xun Huan,[§] Zachary P. Vane,[¶] Joseph C. Oefelein,[‖] Guilhem Lacaze,[**] and Habib N. Najm[††]
Sandia National Laboratories, Livermore, CA 99551, USA

The computational burden of Large-eddy Simulation for reactive flows is exacerbated in the presence of uncertainty in flow conditions or kinetic variables. A comprehensive statistical analysis, with a sufficiently large number of samples, remains elusive. Statistical learning to suitably constrain the domain of the variables of interest carries the promise of extracting more information from fewer samples. Such procedures, if successful, would greatly enhance the predictability of models constrained by the size of the associated statistical samples. In this paper, we show how a recently developed procedure for probabilistic learning on manifolds can serve to improve the predictability of a scramjet simulation. The estimates of the probability density functions of the quantities of interest are improved together with estimates of the statistics of their maxima. We also demonstrate how the improved statistical model adds critical insight to the performance of the model.

[*] Corresponding author, Professor, Laboratoire Modélisation et Simulation Multi Echelle, MSME UMR 8208 CNRS, 5 bd Descartes, 77454 Marne-la-Vallée, France (christian.soize@u-pem.fr).

[†] Professor, Department of Civil and Environmental Engineering, 210 KAP Hall, Los Angeles, CA 90089, USA (ghanem@usc.edu).

[‡] Quantitative Modeling and Analysis, 7011 East Avenue, Mail Stop 9159, Livermore, CA 94551, USA, AIAA Senior Member (csafta@sandia.gov).

[§] Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Member (xhuan@sandia.gov).

[¶] Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Member (zvane@sandia.gov).

[‖] Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Associate Fellow (joefelein@sandia.gov).

[**] Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Member (glacaze@sandia.gov).

[††] Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Member (hnnajm@sandia.gov).

Nomenclature

$C_{\mathbf{w}}$	=	admissible set of $\mathbf{w}$
m_w	=	dimension of $\mathbf{w}$ or $\mathbf{W}$
N	=	number of data points
N_{sup}	=	maximum value of N
n	=	dimension of $\mathbf{x}$
n_q	=	number of QoI
ν_{sim}	=	number of additional realizations
$p_{\mathbf{Q}}$	=	pdf of $\mathbf{Q}$
p_{Q}	=	pdf of Q
$p_{\mathrm{Q}_{\max}}$	=	pdf of $\mathrm{Q}_{\max}$
$p_{\mathbf{W}}$	=	pdf of $\mathbf{W}$
$p_{\mathbf{X}}$	=	pdf of $\mathbf{X}$
QoI	=	Quantity of Interest
$\mathbf{Q}$	=	$(\mathrm{Q}_1,...,\mathrm{Q}_{n_q})$, random QoI
Q	=	any component of $\mathbf{Q}$
Q_k	=	component k of $\mathbf{Q}$
$\mathrm{Q}_{\max}$	=	maximum of Q
QoI	=	Quantity of interest
$\mathbf{q}$	=	$(q_1,\ldots,q_{n_q})$
$\mathbf{q}^\ell$	=	ℓ-th realization of $\mathbf{Q}$
$\mathbf{q}_{\text{ar}}^\ell$	=	ℓ-th additional realization of $\mathbf{Q}$
q_k	=	component k of $\mathbf{q}$
$q_{\max}^\alpha$	=	α-th realization of $\mathrm{Q}_{\max}$
$\mathbb{R}$	=	set of all the real numbers
$\mathbb{R}^{m_w}$	=	Euclidean space of dimension m_w
$\mathbb{R}^n$	=	Euclidean space of dimension n
$\mathbb{R}^{n_q}$	=	Euclidean space of dimension n_q
$\mathbf{w}$	=	$(w_1,\ldots,w_{m_w})$, vector of parameters
$\mathbf{w}^\ell$	=	ℓ-th realization of $\mathbf{W}$
$\mathbf{w}_{\text{ar}}^\ell$	=	ℓ-th additional realization of $\mathbf{W}$
w_j	=	component j of $\mathbf{w}$
$\mathbf{W}$	=	$(W_1,\ldots,W_{m_w})$, random parameters
W_j	=	component j of $\mathbf{W}$
$\mathbf{X}$	=	$(X_1,\ldots,X_n) = (\mathbf{W},\mathbf{Q})$
X_j	=	component j of $\mathbf{X}$
$\mathbf{x}$	=	$(x_1,\ldots,x_n) = (\mathbf{w},\mathbf{q})$
$\mathbf{x}^\ell$	=	ℓ-th realization of $\mathbf{X}$
$\mathbf{x}_{\text{ar}}^\ell$	=	ℓ-th additional realization of $\mathbf{X}$
x_j	=	component j of $\mathbf{x}$

A lower case letter such as y is a real deterministic variable.
A boldface lower case letter such as $\mathbf{y}$ is a real deterministic vector.
An upper case letter such as Y is a real random variable.
A boldface upper case letter such as $\mathbf{Y}$ is a real random vector.
A lower case letter between brackets such as $[y]$ is a real deterministic matrix.
A boldface upper case letter between brackets such as $[\mathbf{Y}]$ is a real random matrix.

I. Introduction

The performance of a scramjet engine is closely tied to the evolution of physical phenomena on scales ranging from the size of the fuel injector to the geometry of the combustion chamber. Capturing the interaction between these phenomena requires the resolution of mathematical models using very fine spatio-temporal discretizations that continue to challenge the most advanced computational resources. Integrating these simulations into a model-based design optimization or a parametric uncertainty propagation context significantly exacerbates the computational burden as they require multiple multiple numerical simulations under varying design and parameter conditions. The task of optimization under uncertainty remains elusive, requiring simplifying assumptions on the physics of the problem that put into question the optimality and even the feasibility of the computed solution. In general, predictions from mathematical models are grounded in conservation laws and can thus be expected to have an implicit structure that may be conducive to numerical simplifications. As indicated previously, given the multiscale nature of relevant phenomena, reductions that oversimplify the physics may lose sight of quantities of interest that are critical for design or safety. Alternative reduction formalisms, as pursued in the present paper, may be cast in the form of probabilistic learning schemes, where intrinsic structure is progressively learned until sufficient credibility in the inferred statements can be certified. In this manner, the spatio-temporal resolution required by the physics is always honored, while the mathematical structure representing the dependence of some quantity of interest (QoI) (itself a function of the solution) on design variables or parameters is learned from consecutive expensive simulations. The hope is that sufficient learning will be achieved from a few such simulations; far fewer than would typically be required for optimization under uncertainty. Clearly, the learning and the simulations from which it is synthesized are dependent on the QoI.

The objective of the present paper is to adapt a recent procedure for probabilistic learning on manifolds [1, 2] to the challenges presented by an LES-resolved simulation of a scramjet. The manifold in question is the geometric structure defining the key design objectives in the span of design and and uncertain parameters. The procedure permits the localization of the support of the probability measure of all available data to the manifold discovered through a Markov process on this data [3]. Available data refers here to numerically generated data that, as indicated above, will be limited in view of the expense associated with its generation. Sampling procedures are then put in place that can augment the initial dataset with statistically consistent samples. While the present paper focuses on this statistical augmentation step, the extension of the results to the design optimization problem are self-evident. They do, however, require special care that places them outside the scope of the present work.

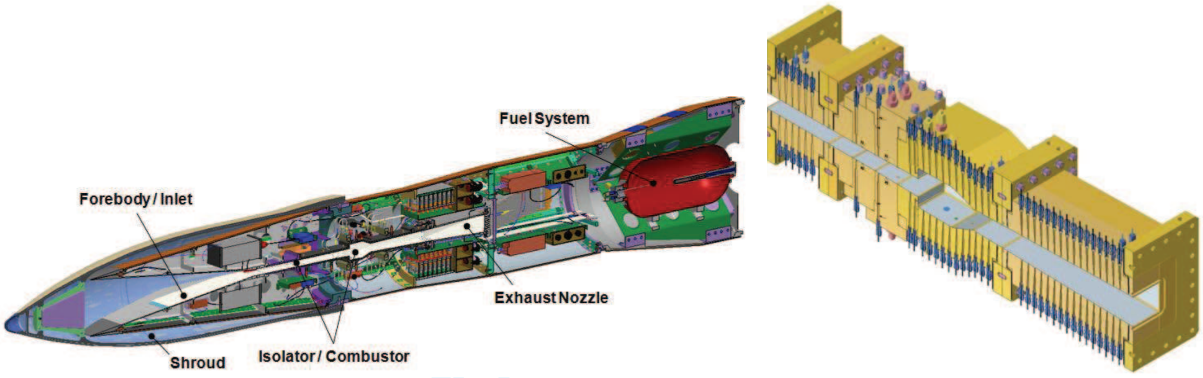
It should be noted that the statistical and probabilistic learning methods have been extensively developed [4–12]) and play an increasingly important role in computational science and engineering [13]), in particular for design optimization under uncertainties using large scale computational models and more generally, in artificial intelligence for extracting information from big data. In recent years, statistical learning methods have been developed in the form of surrogate models from which approximations of model-based function evaluations can easily be computed [14–17]. Although Gaussian process models are most commonly used in this context (see for instance [18, 19]), alternative approaches based on Bayesian methods such as Bayesian optimization have been proposed [14, 20, 21]. For the evaluations of expensive stochastic functions in presence of uncertainties, computational challenges remain currently significant enough to require relevant probabilistic approximations [16, 22–24]. There are many fields for which statistical and probabilistic learning methods are used. In the field of aeronautical engineering learning procedures have been used for over two decades with success for training neural networks [25, 26]. More recently, postprocessing of a given set of Monte Carlo realizations has been proposed for improving integral computation [27] and a machine-learning approach has been used [28] for improving predictive models of turbulence synthesized from limited experimental data. This last paper is certainly in the spirit of the work presented in this paper for which the objective is to enhance the knowledge extracted from limited data, but in using a non-Gaussian probabilistic learning process.

The probabilistic learning on manifold [1], which is used in this paper for enhancing model predictability, proposes a new methodology for generating additional realizations of a random vector whose non-Gaussian probability distribution is unknown and is presumed to be concentrated on an unknown manifold, for which the available information is only constituted of a dataset of independent realizations of this random vector. The probabilistic learning method consists (1) in discovering and in taking into account the geometrical structure of the dataset by using a diffusion maps technique in order to enrich the usual construction of the probability distribution based on a multidimensional Gaussian kernel-density estimation (nonparametric statistics), (2) in preserving the concentration of the additional realizations around the manifold, and (3) in constructing an associated Markov Chain Monte Carlo (MCMC) generator for generating additional realizations that follow the estimated probability distribution.

The paper is organized as follows. In Section II, we summarize the physical and computational model that is used for simulating the complex flow for a ScramJet by means of a large scale computational fluid dynamics model. This section allows also for defining the uncertain parameters of the computational fluid dynamics model (which are modeled as random variables), the random quantities of interest, the specifications of the computational model, and the simulations performed. Section III presents a brief summary of the probabilistic learning on manifold that is used for analyzing ScramJet data. The reader can find all the details of the algorithm in [1]. Section IV is devoted to the description of the ScramJet model representation, to the definition of the random parameters and the random quantities of interest that are retained for the ScramJet analysis, and finally, to the definition of the dataset used for the probabilistic learning. Section V presents the statistical estimation and analysis using the probabilistic learning on manifold that allows for generating additional realizations used for estimating the probability density functions of quantities of interest and of their maximum statistics. The numerical simulations and the analysis of the ScramJet database is presented in Section VI. In particular, we analyze the robustness of the probabilistic learning approach and we show how such an approach allows for enhancing model predictability.

II. Physical and Computational Model

We concentrate on a scramjet configuration studied under the HIFiRE (Hypersonic International Flight Research and Experimentation) program [29, 30], as depicted in Figure 1(a). A ground test rig, designated the HIFiRE Direct Connect Rig (HDCR) (Figure 1(b)), was developed to duplicate the isolator/combustor layout [31, 32]. Mirroring the HDCR setup, we aim to simulate and assess flow characteristics inside the isolator/combustor portion of the scramjet. The rig consists of a constant-area isolator (planar duct) attached to a combustion chamber. It includes four primary injectors mounted upstream of flame stabilization cavities on both the top and bottom walls. Four secondary injectors along both walls are positioned downstream of the cavities. Flow travels from left to right in the x-direction (streamwise), and the geometry is symmetric about the centerline in the y-direction. Numerical simulations take advantage of this symmetry by considering a domain that covers only the bottom half of this configuration. To further reduce the computational cost, we consider one set of primary/secondary injectors and impose periodic conditions in the z-direction (spanwise). The overall computational domain is highlighted by the red lines in Figure 2. JP-7



(a) HIFiRE Flight 2 payload

(b) HDCR

Fig. 1 HIFiRE Flight 2 payload and HDCR cut views.

surrogate fuel [33], composed of 36% methane and 64% ethylene by volume (mole), enters through these injectors. The combustion process is described by a reduced, three-step mechanism [34, 35]:

$$CH_4 + 2(O_2 + f_{NO}N_2) \rightarrow CO_2 + 2H_2O + 2f_{NO}N_2 \tag{1}$$

$$C_2H_4 + 3(O_2 + f_{NO}N_2) \rightarrow 2CO_2 + 2H_2O + 3f_{NO}N_2 \tag{2}$$

$$2CO + O_2 \rightarrow 2CO_2, \tag{3}$$

where $f_{NO} = 0.79/0.21$ is the ratio between the mole fractions of N_2 and O_2 in the oxidizer streams. Arrhenius kinetic parameters are tuned to match the heat release rate to a reference mechanism [36] and to retain robust/stable combustion in the current simulations.

Large eddy simulation (LES) calculations are then performed using the RAPTOR code framework developed by Oefelein [37, 38]. The theoretical framework solves the fully coupled conservation equations of mass, momentum, total-energy, and species for a chemically reacting flow. It is designed to handle high Reynolds number, high-pressure, real-gas and/or liquid conditions over a wide Mach operating range. It also accounts for detailed thermodynamics and transport processes at the molecular level. Noteworthy is that RAPTOR is designed specifically for LES using non-dissipative, discretely conservative, staggered, finite-volume differencing. This eliminates numerical contamination of the subfilter models due to artificial dissipation and provides discrete conservation of mass, momentum, energy, and species, which is imperative for high quality LES. Representative results and case studies using RAPTOR can be found in studies by Oefelein *et al.* [39–41].

In our numerical studies, we allow a total of 11 input parameters to be variable and uncertain, shown in Table 1 along with their uncertainty distributions. These distributions are assumed uniform across the ranges indicated. We focus on three quantities of interest (QoIs): (1) combustion efficiency (η_c) that is related to the burned equivalence ratio (ϕ_B), (2) stagnation pressure loss ratio ($R_{\bar{P}}$), and (3) wall-normal averaged turbulence kinetic energy (TKE) at various streamwise locations. The first two QoIs reflect the overall scramjet performance, while the third contains more localized descriptions that can offer insights for turbulence modeling. All QoIs are time-averaged variables. The data utilized in the current analysis are from 2D simulations of the scramjet computation, using grid resolutions where cell sizes are 1/8 and 1/16 of the injector diameter $d = 3.175$ mm.

- **Combustion efficiency** (η_c) is the combustion efficiency based on static enthalpy quantities [32, 42]:

$$\eta_c = \frac{H(T_{\text{ref}}, Y_{\text{e}}) - H(T_{\text{ref}}, Y_{\text{ref}})}{H(T_{\text{ref}}, Y_{\text{e,ideal}}) - H(T_{\text{ref}}, Y_{\text{ref}})}. \tag{4}$$

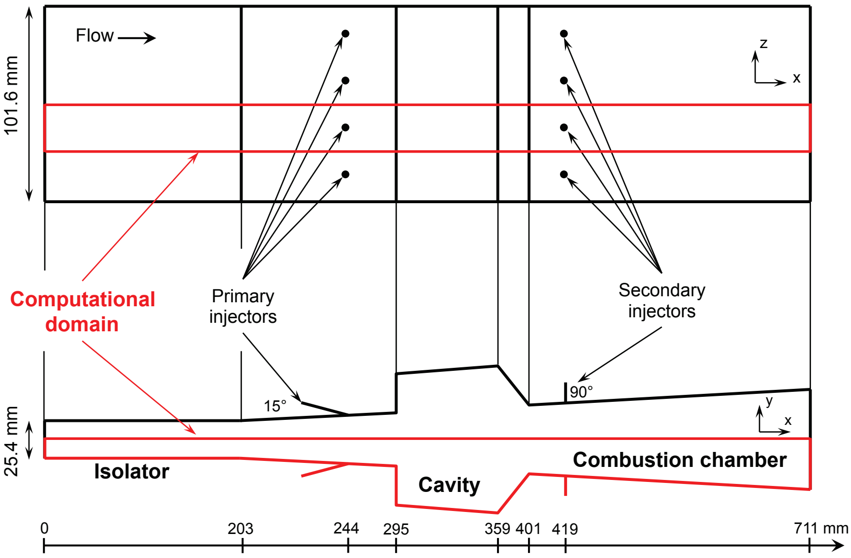


Fig. 2 The HDCR experimental setup and schematic of the full computational domain.

Here H is the total static enthalpy, the "ref" subscript indicates a reference condition derived from the inputs, the "e" subscript is for the exit, and the "ideal" subscript is for the ideal condition where all fuel is burnt to completion. The reference condition corresponds to that of a hypothetical non-reacting mixture of all inlet air and fuel at thermal equilibrium. The numerator, $H(T_{\text{ref}}, Y_{\text{e}}) - H(T_{\text{ref}}, Y_{\text{ref}})$, thus reflects the global heat released during the combustion, while the denominator represents the total heat release available in the fuel-air mixture.

- **Stagnation pressure loss ratio** ($R_{\bar{P}}$) is defined as

$$R_{\bar{P}} = 1 - \frac{P_{s,e}}{P_{s,i}}, \tag{5}$$

where $P_{s,e}$ and $P_{s,i}$ are the wall-normal-averaged stagnation pressure quantities at the exit and inlet planes, respectively.

- **Turbulence kinetic energy (TKE)** is characterized by the root-mean-square (RMS) velocity fluctuations at a given location:

$$\text{TKE} = \frac{1}{2}\left(u_{\text{rms}}^2 + v_{\text{rms}}^2 + w_{\text{rms}}^2\right), \tag{6}$$

where the RMS quantity is $u_{\text{rms}} = \sqrt{\overline{u^2} - \overline{u}^2}$, with $\overline{u}$ indicating time-averaged quantity. In the numerical investigations of this paper, we will look at TKE from multiple streamwise locations (i.e., different x locations).

III. Probabilistic Learning on Manifold for Analyzing ScramJet Data

In this section, we summarize the probabilistic learning methodology [1] that will be used throughout the paper for predicting the statistics and for performing model exploration to enhance model predictability of LES simulations of a ScramJet.

This probabilistic learning on manifold uses only a dataset of N data points $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ in $\mathbb{R}^n$, which are assumed to be N independent realizations of a random vector $\mathbf{X}$ with values in $\mathbb{R}^n$. The probability distribution of $\mathbf{X}$ is unknown and is assumed to be concentrated in a neighborhood of a subset of $\mathbb{R}^n$ (a manifold) that is also unknown and that has to be discovered. For the ScramJet database, vector $\mathbf{X}$ will be constituted of the 11 uncertain parameters of the computational model (modeled by random variables as explained in Section II) to which are added all the random quantities of interest (QoIs) that are outputs of the stochastic computational model. The objective of the probabilistic

Parameter	Range	Description
Inlet boundary conditions:		
p_0	$[1.406, 1.554] \times 10^6$ Pa	Stagnation pressure
T_0	$[1472.5, 1627.5]$ K	Stagnation temperature
M_0	$[2.259, 2.759]$	Mach number
L_i	$[0, 8] \times 10^{-3}$ m	Inlet turbulence length scale
I_i	$[0, 0.05]$	Turbulence intensity horizontal component
R_i	$[0.8, 1.2]$	Ratio of turbulence intensity vertical to horizontal components
Fuel inflow boundary conditions:		
I_f	$[0, 0.05]$	Turbulence intensity magnitude
L_f	$[0, 1] \times 10^{-3}$ m	Turbulence length scale
Turbulence model parameters:		
C_R	$[0.01, 0.06]$	Modified Smagorinsky constant
Pr_t	$[0.5, 1.7]$	Turbulent Prandtl number
Sc_t	$[0.5, 1.7]$	Turbulent Schmidt number

Table 1 Uncertain input parameters. The uncertainty distributions are assumed uniform across the ranges shown.

learning on manifold is to construct a probabilistic model of random vector $\mathbf{X}$ using only dataset $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, which allows for generating $\nu_{\text{sim}} \gg N$ additional independent realizations $\{\mathbf{x}^1_{\text{ar}}, \ldots, \mathbf{x}^{\nu_{\text{sim}}}_{\text{ar}}\}$ in $\mathbb{R}^n$ of random vector $\mathbf{X}$. The proposed method preserves the concentration of the additional realizations around the manifold. For the ScramJet analysis, we can then generate a very large number, $\nu_{\text{sim}} \gg N$, of additional realizations that allow for estimating the probability density functions of various QoIs, including the statistics of their maxima. The main steps of this methodology can be roughly summarized as follows.

1) A principal component analysis of $\mathbf{X}$ is carried out in order to normalize the dataset, which yields a new normalized dataset of N data points $\{\mathbf{y}^1, \ldots, \mathbf{y}^N\}$ in $\mathbb{R}^\nu$. This means that the random vector $\mathbf{Y}$ with values in $\mathbb{R}^\nu$ for which $\{\mathbf{y}^1, \ldots, \mathbf{y}^N\}$ are N independent realizations, has a zero empirical mean and an empirical covariance matrix that is the unity matrix.
2) Dataset $\{\mathbf{y}^1, \ldots, \mathbf{y}^N\}$ is rewritten as a $(\nu \times N)$ rectangular matrix $[y_d]$ that is construed as one realization of a $(\nu \times N)$ rectangular random matrix $[\mathbf{Y}] = [\mathbf{Y}^1 \ldots \mathbf{Y}^N]$ in which $\mathbf{Y}^1, \ldots, \mathbf{Y}^N$ are N independent random vectors. A modification [43] of the classical multidimensional Gaussian kernel-density estimation method [44, 45] is then used to construct and estimate the probability density function (pdf) $p_{[\mathbf{Y}]}([y])$ of random matrix $[\mathbf{Y}]$ with respect to the volume element $d[y]$ on the set of all the $(\nu \times N)$ real matrices.
3) A $(\nu \times N)$ matrix-valued Itô stochastic differential equation (ISDE), associated with the random matrix $[\mathbf{Y}]$, is constructed and corresponds to a stochastic nonlinear dissipative Hamiltonian dynamical system, for which $p_{[\mathbf{Y}]}([y])\, d[y]$ is the unique invariant measure. This construction is performed using the approach proposed in [43, 46] belonging to the class of Hamiltonian Monte Carlo methods [46–48], which is an MCMC algorithm [49].
4) The diffusion-map approach [3] is then used to discover and characterize the local geometry structure of the normalized dataset $[y_d]$. The subset of the diffusion-maps basis, represented by a $(N \times m)$ matrix $[g] = [\mathbf{g}^1 \ldots \mathbf{g}^m]$, are thus constructed with $m \ll N$. They are associated with the first m eigenvalues of the transition matrix of a Markov chain relative to the local geometric structure of the given normalized dataset $[y_d]$.
5) As proposed in [1], a reduced-order representation $[\mathbf{Y}] = [\mathbf{Z}]\,[g]^T$ is constructed in which $[\mathbf{Z}]$ is a $(\nu \times m)$ random matrix for which $m \ll N$. A reduced-ISDE, associated with random matrix $[\mathbf{Z}]$, is obtained by projecting the ISDE introduced in Step 3 onto the subspace spanned by the reduced-order vector basis represented by matrix $[g]^T$. It should be noted that such a projection corresponds to a reduction of the dataset dimension

and not to a reduction of the physical components of random vector $\mathbf{Y}$ that already results from a PCA applied to $\mathbf{X}$. Such a projection preserves the concentration of the generated realizations around the manifold. The constructed reduced ISDE is then used for generating additional realizations $[z_{\rm ar}^1],\ldots,[z_{\rm ar}^{n_{\rm MC}}]$ of random matrix $[\mathbf{Z}]$, and therefore, for deducing the additional realizations $[y_{\rm ar}^1],\ldots,[y_{\rm ar}^{n_{\rm MC}}]$ of random matrix $[\mathbf{Y}]$. Reshaping these $n_{\rm MC}$ matrices yields the $\nu_{\rm sim} = N \times n_{\rm MC}$ independent realizations $\{\mathbf{y}^1,\ldots,\mathbf{y}^{\nu_{\rm sim}}\}$ of random vector $\mathbf{Y}$. Using the PCA constructed in Step 1 allows for generating the $\nu_{\rm sim} \gg N$ additional independent realizations $\{\mathbf{x}_{\rm ar}^1,\ldots,\mathbf{x}_{\rm ar}^{\nu_{\rm sim}}\}$ in $\mathbb{R}^n$ of random vector $\mathbf{X}$.

IV. ScramJet Model Representation, Parameters, QoI, and Dataset for the Probabilistic Learning

A. ScramJet Model Representation

The ScramJet database is generated with the physical and computational model presented in Section II. The uncertain parameter of the computational model is a vector $\mathbf{w} = (w_1,\ldots,w_{m_w})$ that belongs to a subset $C_{\mathbf{w}}$ of $\mathbb{R}^{m_w}$ in which $m_w = 11$. This uncertain parameter $\mathbf{w}$ is modeled by a second-order $\mathbb{R}^{m_w}$-valued random variable $\mathbf{W} = (W_1,\ldots,W_{m_w})$ defined on a probability space $(\Theta,\mathcal{T},\mathcal{P})$ for which the support of the probability distribution is the set $C_{\mathbf{w}}$ that is defined in Table 1.

The vector-valued QoI that is deduced from the outputs of the computational model is denoted by $\mathbf{q} = (q_1,\ldots,q_{n_q}) \in \mathbb{R}^{n_q}$ in which $n_q = 10$. For $\mathbf{W} = \mathbf{w}$ fixed in $C_{\mathbf{w}}$, the QoI is modeled by a $\mathbb{R}^{n_q}$-valued random variable $\mathbf{F}(\mathbf{w})$, defined on probability space $(\Theta,\mathcal{T},\mathcal{P})$, for which any realization will be denoted by $\mathbf{F}(\mathbf{w};\theta)$ with $\theta \in \Theta$. It should be noted that, for representing the computational model, we could have considered an $\mathbb{R}^{n_q}$-valued deterministic variable, $\mathbf{q} = \mathbf{f}(\mathbf{w})$, but it is more realistic to consider other possible uncertainties that the one induced by parameter $\mathbf{w}$ due to the use of a very complex computational model that is run on a massively parallel computer. Consequently, the corresponding random QoI is the $\mathbb{R}^{n_q}$-valued random variable, $\mathbf{Q} = (\mathrm{Q}_1,\ldots,\mathrm{Q}_{n_q})$, defined on $(\Theta,\mathcal{T},\mathcal{P})$ and such that $\mathbf{Q} = \mathbf{F}(\mathbf{W})$. It is assumed that $\mathbf{Q}$ is a second-order random variable. The realizations of $\mathbf{W}$ and $\mathbf{Q}$ will be denoted $\mathbf{w}^\ell = \mathbf{W}(\theta_\ell)$ and $\mathbf{q}^\ell = \mathbf{Q}(\theta_\ell)$ with $\theta_\ell \in \Theta$. The probability distribution of $\mathbf{Q}$ is unknown.

B. Random Model Parameters and Random Quantities of Interest

For the ScramJet database, we have $m_w = 11$ and $n_q = 10$. The components of the random model parameters, represented by random vector $\mathbf{W}$, are (see Table 1):

W_1: Inlet stagnation pressure, p_0.
W_2: Inlet stagnation temperature, T_0.
W_3: Inlet Mach number, M_0.
W_4: Modified Smagorinsky constant, C_R.
W_5: Turbulent Prandtl number, Pr_t.
W_6: Turbulent Schmidt number, Sc_t.
W_7: Inlet turbulence intensity horizontal component, I_i.
W_8: Inlet turbulence length scale, L_i.
W_9: Inlet ratio of turbulence intensity vertical to horizontal components, R_i.
W_{10}: Fuel inflow turbulence intensity magnitude, I_f.
W_{11}: Fuel inflow turbulence length scale, L_f.

Subset $C_{\mathbf{w}}$ of $\mathbb{R}^{m_w}$ is written as the cartesian product $\mathcal{J}_1 \times \ldots \times \mathcal{J}_{n_w}$ of closed intervals $\mathcal{J}_j = [a_j, b_j] \subset \mathbb{R}$. The components of the random quantities of interest, represented by random vector $\mathbf{Q}$, are:

Q_1: Burned equivalence ratio
Q_2: Combustion efficiency
Q_3: Pressure stagnation loss ratio
Q_4: TKE at the inlet streamwise location
Q_5: TKE at streamwise location just before the primary injectors
Q_6: TKE at streamwise location after the primary injectors and before the cavity
Q_7: TKE at streamwise location inside the cavity
Q_8: TKE at streamwise location just after secondary injectors
Q_9: TKE at streamwise location inside the combustion chamber

Q_{10}: TKE at streamwise location at end of the combustion chamber

in which TKE is the wall-normal averaged turbulence kinetic energy at various streamwise locations for which the locations indicated in Figure 2).

For each considered dataset of the ScramJet database, the maximum number of data points that are available is denoted by N_{sup}. The current dimension of such a dataset that will be considered for the probabilistic learning is denoted by $N \leq N_{\text{sup}}$. A convergence analysis of the probabilistic learning analysis related to all the computed quantities will be performed with respect to the value of N when N will go to N_{sup}. For a given dataset of the ScramJet database, for fixed N such that $1 \leq N \leq N_{\text{sup}}$, and for $\ell = 1, \ldots, N$, the realizations $\mathbf{w}^\ell = \mathbf{W}(\theta_\ell) \in \mathbb{R}^{m_w}$ and the corresponding realizations $\mathbf{q}^\ell = \mathbf{Q}(\theta_\ell) \in \mathbb{R}^{n_q}$ of $\mathbf{Q}$ are such that

$$\mathbf{q}^\ell = \mathbf{F}(\mathbf{w}^\ell; \theta_\ell) \in \mathbb{R}^{n_q} \,. \tag{7}$$

C. Defining the Datasets for the Probabilistic Learning From the ScramJet Database

Three datasets are extracted from the ScramJet database. The first is defined as the d08 dataset and corresponds to the results generated with the computational model that is constructed with a grid resolution where cell size is 1/8 while the second one is defined as the d16 dataset and corresponds to a cell size of 1/16. The third one is the concatenated d08-d16 dataset that corresponds to the concatenation of the d08 dataset with the d16 dataset, obtained by interlacing the two datasets with respect to their data points. For each one of the three datadasets, the number N_{sup} of data points are $N_{\text{sup}} = 256$ for the d08 and d16 datasets, while $N_{\text{sup}} = 512$ for the concatenated d08-d16 dataset. For given $N \leq N_{\text{sup}}$, a dataset is made up of the N data points $\mathbf{x}^1, \ldots, \mathbf{x}^N$ in $\mathbb{R}^n$ with

$$n = m_w + n_q \,, \tag{8}$$

such that

$$\mathbf{x}^\ell = (\mathbf{w}^\ell, \mathbf{q}^\ell) \in \mathbb{R}^n = \mathbb{R}^{m_w} \times \mathbb{R}^{n_q} \quad , \quad \ell = 1, \ldots, N \,. \tag{9}$$

For fixed N, the probabilistic learning on manifold will be carried out using dataset $\{\mathbf{x}^\ell, \ell = 1, \ldots, N\}$. This dataset depends on N and as we have explained before, a convergence analysis of the probabilistic learning with respect to N will be performed for $1 \leq N \leq N_{\text{sup}}$. It should be noted that, for the concatenated d08-d16 dataset, if, for instance, $N = 200$, then there are the first 100 data points from the d08 dataset and the first 100 data points from the d16 dataset.

V. Statistical Estimation and Analysis Using Probabilistic Learning on Manifold

In all this section, N is fixed such that $1 \leq N \leq N_{\text{sup}}$. The probabilistic learning that will allow for generating $\nu_{\text{sim}} \gg N$ additional realizations of $\mathbf{X}$ will then depend on this value of N. For simplifying the notations, this dependence on N is removed when it is not necessary for the understanding.

A. Probability Distributions of Random Variables X, W, and Q

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a second-order random variable defined on probability space $(\Theta, \mathcal{T}, \mathcal{P})$ with values in $\mathbb{R}^n$, with $n = m_w + n_q$. Its probability distribution $P_{\mathbf{X}}(d\mathbf{x})$, that is assumed to be represented by a pdf $p_{\mathbf{X}}(\mathbf{x})$ (with respect to the Lebesgue measure $d\mathbf{x}$ on $\mathbb{R}^n$) is unknown but the N given data points $\mathbf{x}^1, \ldots, \mathbf{x}^N$ in $\mathbb{R}^n$, defined by Eq. (9), are assumed to be N given statistically independent realizations of $\mathbf{X}$. This means that the solely available information for estimating $p_{\mathbf{X}}$ is constituted of dataset $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ of N points in$\mathbb{R}^n$. Taking into account Eq. (9), random vector $\mathbf{X}$ can also be written as

$$\mathbf{X} = (\mathbf{W}, \mathbf{Q}) \,, \tag{10}$$

in which $\mathbf{W} = (W_1, \ldots, W_{m_w})$ and $\mathbf{Q} = (\mathrm{Q}_1, \ldots, \mathrm{Q}_{n_q})$ are the random vectors defined in Section IV. B for which the N realizations are $\mathbf{w}^\ell \in \mathbb{R}^{m_w}$ and $\mathbf{q}^\ell \in \mathbb{R}^{n_q}$. The pdf $\mathbf{x} \mapsto p_{\mathbf{X}}(\mathbf{x})$ on $\mathbb{R}^n$ of $\mathbf{X}$, with respect to $d\mathbf{x}$, can also be rewritten as the joint pdf $(\mathbf{w}, \mathbf{q}) \mapsto p_{\mathbf{W},\mathbf{Q}}(\mathbf{w}, \mathbf{q})$ on $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ of $\mathbf{W}$ and $\mathbf{Q}$, with respect to $d\mathbf{w}\, d\mathbf{q}$, in which $\mathbf{x} = (\mathbf{w}, \mathbf{q})$. As explained in Section III, for the considered fixed value of N, the probabilistic learning will allow for generating ν_{sim} additional realizations $\{\mathbf{x}_{\text{ar}}^1, \ldots, \mathbf{x}_{\text{ar}}^{\nu_{\text{sim}}}\}$ of $\mathbf{X}$, with $\nu_{\text{sim}} \gg N$, by using only dataset $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$. For estimating the statistics related to $\mathbf{Q}$, we will need to extract the corresponding ν_{sim} additional realizations $\{\mathbf{q}_{\text{ar}}^1, \ldots, \mathbf{q}_{\text{ar}}^{\nu_{\text{sim}}}\}$ for $\mathbf{Q}$ such that,

$$(\mathbf{w}_{\text{ar}}^\ell, \mathbf{q}_{\text{ar}}^\ell) = \mathbf{x}_{\text{ar}}^\ell \quad , \quad \ell = 1, \ldots, \nu_{\text{sim}} \,. \tag{11}$$

B. Selecting the Random QoI for the Statistical Estimates

Random vector $\mathbf{Q}$ is completely defined by its probability density function $\mathbf{q} \mapsto p_{\mathbf{Q}}(\mathbf{q})$ on $\mathbb{R}^{n_q}$, which can be estimated using nonparametric statistics with a large number, $\nu_{\rm sim}$, of additional realizations of $\mathbf{Q}$. In addition, we are interested in analyzing the maximum statistics of the random components of $\mathbf{Q}$. In order to limit the number of figures presented in the paper, we will not consider all the possible marginal probability density functions of random vector $\mathbf{Q}$, but we will only consider the probability density function of each random component Q_k of $\mathbf{Q}$ for which k is in $\{1,\ldots,n_q\}$ (marginal probability density function of order 1). In the following, in order to not complicate the notations, index k is removed and notation Q is used instead of Q_k (except if confusion is possible).

C. Defining the Maximum Statistics for the Selected Random QoI and Computing their Realizations

For the ScramJet application, since the real-valued random variables that are observed are positive almost surely, we are only interested in constructing their maximum statistics, but their minimum statistics could similarly be constructed although of low interest for this case. For a sufficiently large integer ν_s, the maximum of the real-valued random variable Q can classically be defined as the real-valued random variable $\mathrm{Q}_{\rm max}$ such that $\mathrm{Q}_{\rm max} = \max\{\mathrm{Q}^{(1)},\ldots,\mathrm{Q}^{(\nu_s)}\}$, in which $\mathrm{Q}^{(1)},\ldots,\mathrm{Q}^{(\nu_s)}$ are ν_s independent copies of real-valued random variable Q. Random variable $\mathrm{Q}_{\rm max}$ depends on ν_s, but in order to simplify the notations, the dependence on ν_s is removed. The realizations of $\mathrm{Q}_{\rm max}$ are computed as follows. For fixed N such that $N \leq N^{\rm max}$, for a given value $\nu_{\rm sim}$ of additional realizations $\{(\mathbf{w}^{\ell}_{\rm ar}, q^{\ell}_{\rm ar}) \in \mathbb{R}^{m_w} \times \mathbb{R}, \ell = 1,\ldots,\nu_{\rm sim}\}$ introduced in Section V. C and computed thanks to the probabilistic learning, and for ν_s sufficiently large such that $\nu_s \ll \nu_{\rm sim}$, we construct $\nu_\alpha = \nu_{\rm sim}/\nu_s$ independent realizations $\{q^1_{\rm max},\ldots,q^{\nu_\alpha}_{\rm max}\}$ of $\mathrm{Q}_{\rm max}$ such that, for $\alpha = 1,\ldots,\nu_\alpha$, $q^{\alpha}_{\rm max} = \max_{\ell\,\in\,\{\nu_s(\alpha-1)+1,\,\ldots\,,\,\alpha\nu_s\}}\, q^{\ell}_{\rm ar}$. For the Scramjet results presented in Section VI and for a fixed number $\nu_{\rm sim}$ of additional realizations (that is a finite number!), a convergence analysis of the estimated probability density function of $\mathrm{Q}_{\rm max}$ has been performed as a function of ν_s. We have found that, for the finite number of additional realizations that is considered, a reasonable convergence was obtained for $\nu_s = 100$, such a convergence being obviously only considered as sufficient in the framework for which the pdf of $\mathrm{Q}_{\rm max}$ is studied for the enhancing of the model prediction. Note that, since $\nu_{\rm sim}$ can arbitrarily be increased without significant computational cost, ν_s and ν_α could arbitrarily be increased in satisfying the equation $\nu_{\rm sim} = \nu_\alpha \times \nu_s$ with $\nu_s < \nu_\alpha$.

D. Estimates of the Second-order Moments and the pdf of Random Variables Q and $\mathrm{Q}_{\rm max}$

For a fixed value of N, $\nu_{\rm sim}$, and ν_s (and consequently, of $\nu_\alpha = \nu_{\rm sim}/\nu_s$), the standard deviations σ_{Q} and $\sigma_{\mathrm{Q}_{\rm max}}$ of the real-valued random variables Q and $\mathrm{Q}_{\rm max}$, and their probability density functions $q \mapsto p_{\mathrm{Q}}(q)$ and $q \mapsto p_{\mathrm{Q}_{\rm max}}(q)$ with respect to dq on $\mathbb{R}$, are estimated using the classical estimates (empirical estimates for the standard deviation and Gaussian kernel density estimation for the pdf) based on the use of the additional realizations $\{q^1_{\rm ar},\ldots,q^{\nu_{\rm sim}}_{\rm ar}\}$ for Q and of the realizations $\{q^1_{\rm max},\ldots,q^{\nu_\alpha}_{\rm max}\}$ for $\mathrm{Q}_{\rm max}$ (for 11 components). The convergence analysis of these quantities has been performed with respect to N (in order to analyze how the probabilistic learning approach learns from the dataset as a function of its dimension) and with respect to $\nu_{\rm sim}$ (in order to analyze the robustness of the estimates). Nevertheless, for limiting the number of figures, in Section VI, only the convergence with respect to N of the probability density functions $q \mapsto p_{\mathrm{Q}}(q)$ and $q \mapsto p_{\mathrm{Q}_{\rm max}}(q)$ are shown.

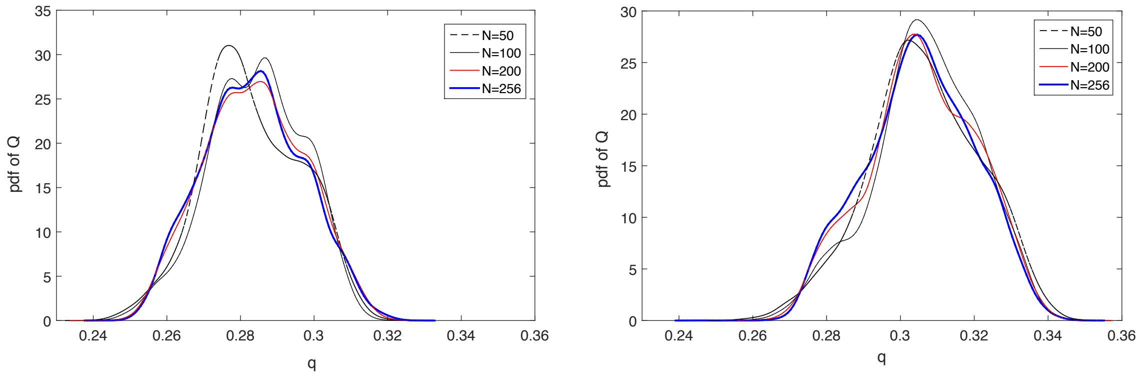
VI. Numerical Simulations and Statistical Analysis for the Datasets of the ScramJet Database

For the d08 and d16 datasets, and for the concatenated d08-d16 dataset, the probabilistic learning has been performed with the all the components of $\mathbf{W}$ (11 components) and with all the components of $\mathbf{Q}$ (10 components). The components, Q_k, of random vector $\mathbf{Q}$ for which the statistics are presented below are Q_2, Q_3, Q_6, Q_7, Q_8, Q_9, and Q_{10}.

A. Methodology Used for the Statistical Analysis

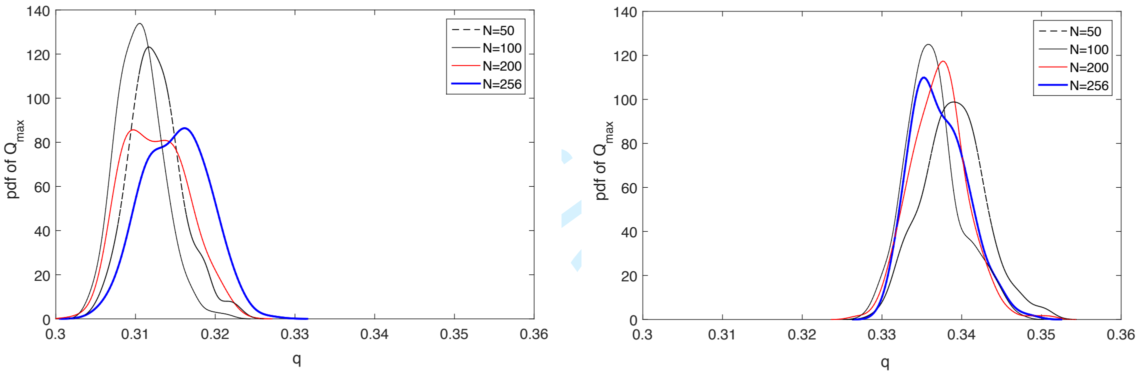
The methodology adopted for the statistical analysis is as follows:

1) For the d08 and d16 datasets, for Q_2 and Q_3, and for $\nu_{\rm sim} = 25{,}600$ additional realizations, an analysis of the robustness of the probabilistic learning is performed with respect to the number N of data points with $N = \{50, 100, 200, 256\}$. Note that $\nu_{\rm sim} = N \times n_{\rm MC}$ is maintained to 25,600 for each value of N (Section VI. B.1).
2) For the d08 and d16 datasets, the model predictability of TKE is performed at various streamwise locations corresponding to $\{\mathrm{Q}_k, k = 6,\ldots,10\}$, for $N = 256$ and for $\nu_{\rm sim} = 25{,}600$ additional realizations (Section VI. B.2).



(a) d08 dataset (b) d16 dataset

Fig. 3 Combustion efficiency Q_2: probability density functions $p_Q(q)$ of random variable Q for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 256$ (thick blue line) with $\nu_{\rm sim} = 25,600$.



(a) d08 dataset (b) d16 dataset

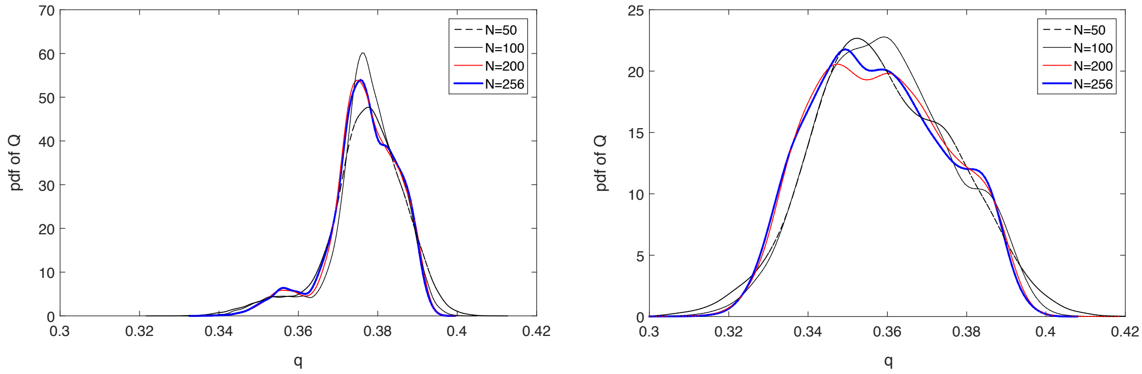
Fig. 4 Combustion efficiency Q_2: probability density functions $p_{Q_{\rm max}}(q)$ of random variable $Q_{\rm max}$ for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 256$ (thick blue line) with $\nu_{\rm sim} = 25,600$.

3) For the concatenated d08-d16 dataset, the analysis of the robustness of the probabilistic learning is again performed for Q_2 and Q_3 with respect to the number N of data points with $N = \{50, 100, 200, 450, 512\}$ and $\nu_{\rm sim} = N \times n_{\rm MC} = 51,200$ (Section VI. C.1).
4) Finally, for the concatenated d08-d16 dataset, the model predictability of TKE is again performed at the same streamwise locations corresponding to $\{Q_k, k = 6, \dots, 10\}$, for $N = 512$ and $\nu_{\rm sim} = 51,200$ additional realizations (Section VI. C.2).

B. Probabilistic Learning Approach for Analyzing the d08 and d16 Datasets

1. Robustness Analysis of the Probabilistic Learning Approach for the Combustion Efficiency and the Pressure Stagnation Loss Ratio

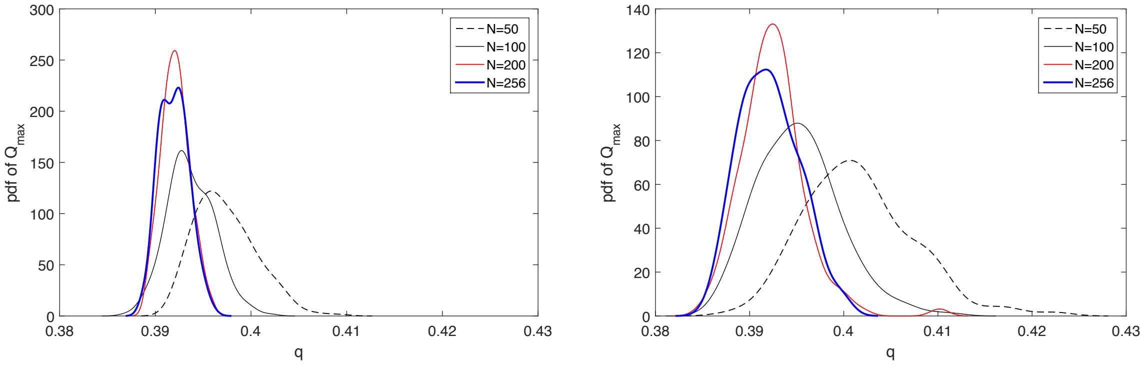
For each one of the d08 and d16 datasets, and for $\nu_{\rm sim} = 25,600$, an analysis has been carried out by studying, for Q_2 (combustion efficiency, Figures 3 and 4) and for Q_3 (pressure stagnation loss ratio, Figures 5 and 6), the evolution with respect to N of the probability density functions $p_Q(q)$ of random variable Q (Figures 3 and 5) and $p_{Q_{\rm max}}(q)$ of random variable $Q_{\rm max}$ (Figures 4 and 6).



(a) d08 dataset

(b) d16 dataset

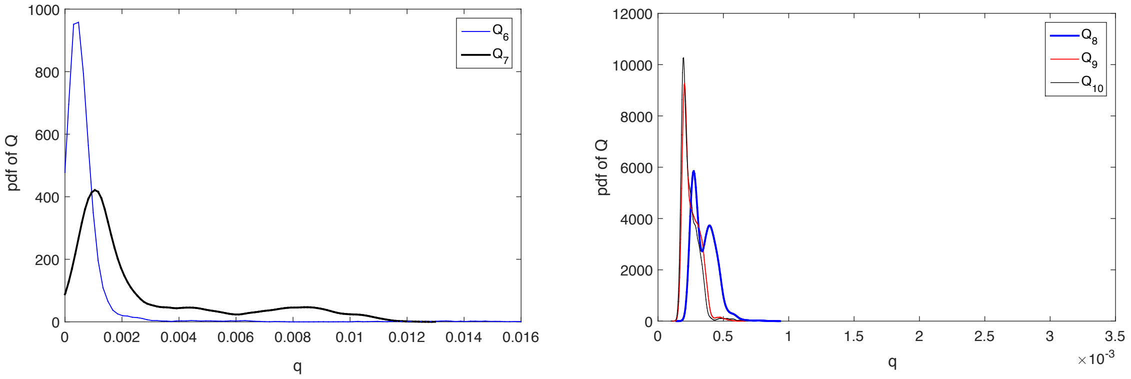
Fig. 5 Pressure stagnation loss ratio Q_3: probability density functions $p_{\mathbf{Q}}(q)$ of random variable Q for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 256$ (thick blue line) with $\nu_{\text{sim}} = 25,600$.



(a) d08 dataset

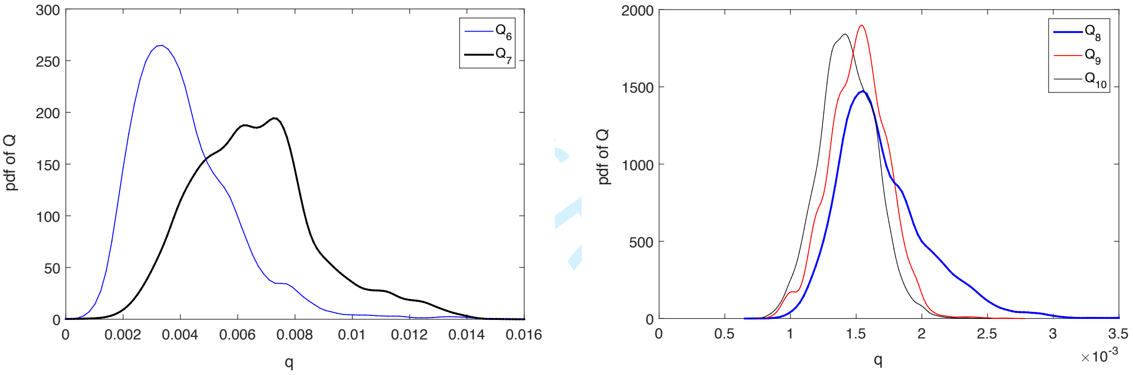
(b) d16 dataset

Fig. 6 Pressure stagnation loss ratio Q_3: probability density functions $p_{\mathbf{Q}_{\max}}(q)$ of random variable $\mathbf{Q}_{\max}$ for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 256$ (thick blue line) with $\nu_{\text{sim}} = 25,600$.



(a) (b)

Fig. 7 For the d08 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{\mathbf{Q}}(q)$ of TKE Q. (a): Q_6 (mid blue line) and Q_7 (thin black line). (b): Q_8 (thick blue line), Q_9 (mid red line), and Q_{10} (thin black line).



(a) (b)

Fig. 8 For the d16 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{\mathbf{Q}}(q)$ of TKE Q. (a): Q_6 (mid blue line) and Q_7 location (thin black line). (b): Q_8 (thick blue line), Q_9 (mid red line), and Q_{10} (thin black line).

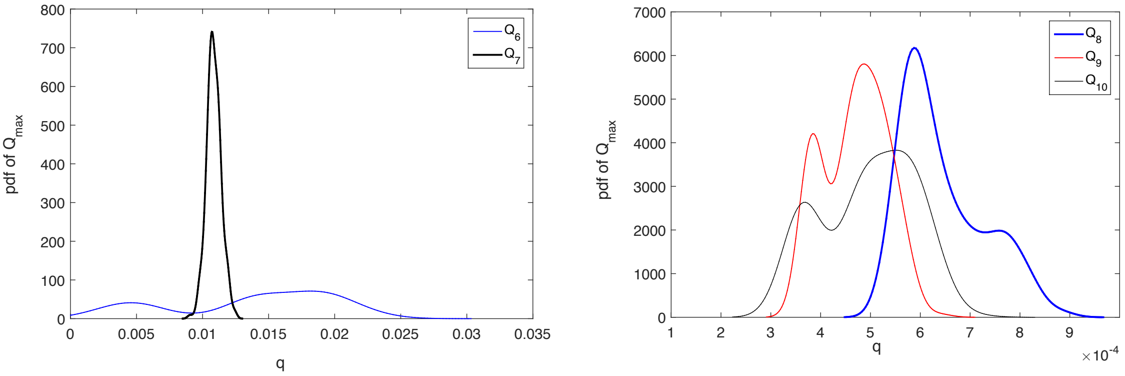
2. Model Predictability of the Wall-Normal averaged Turbulence Kinetic Energy Performed at Various Streamwise Locations Using the Probabilistic Learning Approach

From the convergence analyses presented in Section VI. B.1, it can be concluded that $N = 256$ and $\nu_{\text{sim}} = 25,600$ are good values for studying TKE at the various streamwise locations associated with Q_6, Q_7, Q_8, Q_9, and Q_{10}. For the d08 and d16 datasets, the analysis of the evolution of probability density functions $p_{\mathrm{Q}}(q)$ of random variable Q is shown in Figures 7 and 8 as a function of the location of the observations along the flow while the evolution of $p_{\mathrm{Q}_{\max}}(q)$ of random variable $\mathrm{Q}_{\max}$ is shown in Figure 9 and 10.

C. Probabilistic Learning Approach for Analyzing the Concatenated d08-d16 Dataset

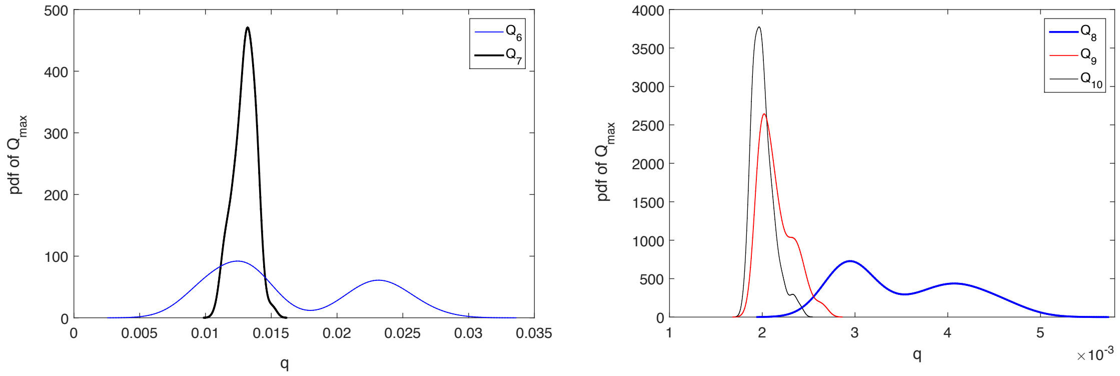
1. Robustness Analysis of the Probabilistic Learning Approach for the Combustion Efficiency and the Pressure Stagnation Loss Ratio

A similar analysis that the one presented in Section VI. B.1, has been performed for the concatenated d08-d16 dataset that is constructed in interlacing the data points of the d08 dataset with the d16 dataset. Therefore, there are $N_{\text{sup}} = 512$ data points in the concatenated d08-d16 dataset. Similarly to Section VI. B.2, for the concatenated d08-d16



(a) (b)

Fig. 9 For the d08 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{Q_{\text{max}}}(q)$ of TKE Q_{max}. (a): Q_6 (mid blue line) and Q_7 (thin black line). (b): Q_8 (thick blue line), Q_9 (mid red line), and Q_{10} (thin black line).



(a) (b)

Fig. 10 For the d16 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{Q_{\text{max}}}(q)$ of TKE Q_{max}. (a): Q_6 (mid blue line) and Q_7 (thin black line). (b): Q_8 (thick blue line), Q_9 (mid red line), and Q_{10} (thin black line).

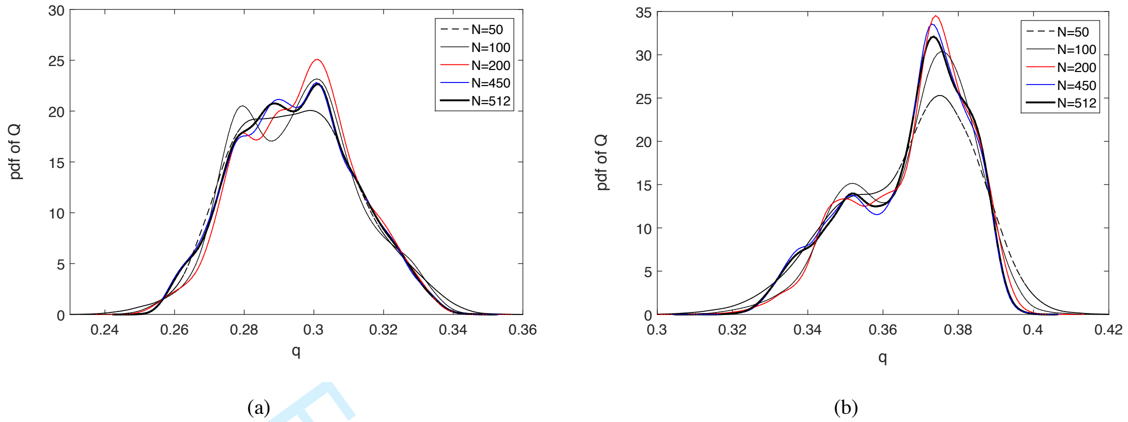


Fig. 11 d08-d16 dataset: probability density functions $p_{\mathrm{Q}}(q)$ of random variable Q (a) for combustion efficiency Q_2 and (b) for pressure stagnation loss ratio Q_3, for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 450$ (med blue line), $N = 512$ (thick black line) with $\nu_{\mathrm{sim}} = 51,200$.

dataset and for $\nu_{\mathrm{sim}} = 51,200$, an analysis has been carried out by studying the evolution with respect to $N \leq N_{\mathrm{sup}}$ of the probability density function $p_{\mathrm{Q}}(q)$ of random variable Q for Q = Q_2 (combustion efficiency, Figure 11(a)) and for Q = Q_3 (pressure stagnation loss ratio, Figure 11(b)), while Figures 12(a) and (b) display the evolution of the probability density function $p_{\mathrm{Q}_{\max}}(q)$ of random variable $\mathrm{Q}_{\max}$.

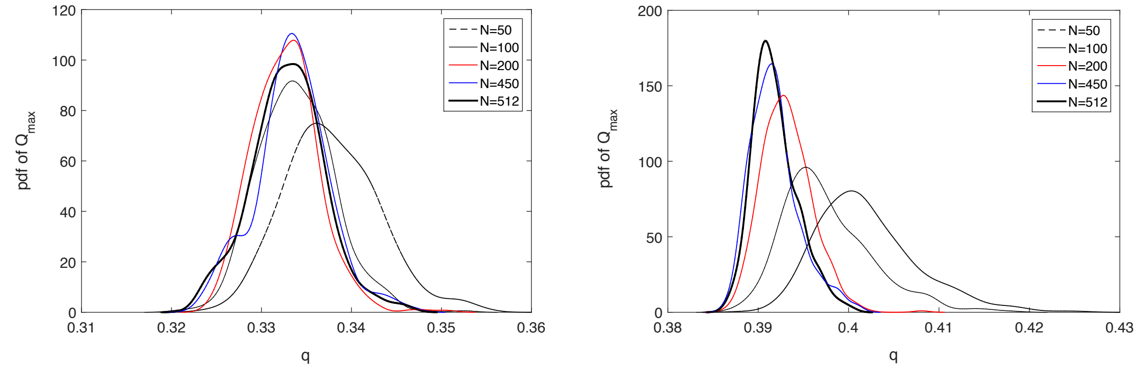
2. Model Predictability of the Wall-Normal Averaged Turbulence Kinetic Energy Performed at Several Streamwise Locations Using the Probabilistic Learning Approach With the Concatenated d08-d16 Dataset

From the convergence analyses presented in Section VI. C.1, it can be concluded that $N = 512$ and $\nu_{\mathrm{sim}} = 51,200$ are good values for studying TKE at various streamwise locations associated with Q_6, Q_7, Q_8, Q_9, and Q_{10}. For the concatenated d08-d16 dataset, Figure 13 displays the probability density function $p_{\mathrm{Q}}(q)$ of TKE associated with Q_6 to Q_{10}, while Figure 14 displays the probability density function $p_{\mathrm{Q}_{\max}}(q)$.

D. Analysis of the Results Obtained With the Probabilistic Learning

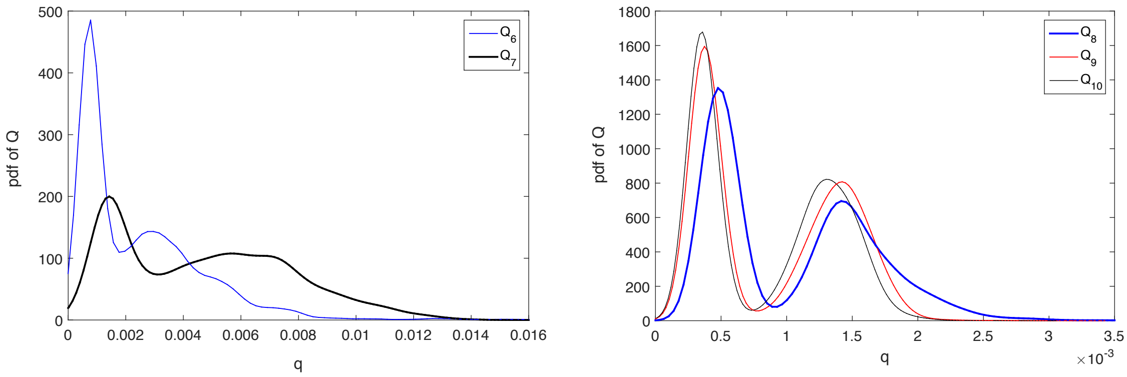
A few general observations can be made from inspecting Figures 3 to 14. Figures 3 and 5 show that combustion efficiency (Q_2) and pressure stagnation loss ratio (Q_3) are learned with minimal effort using $N = 50$ data points, while the maximum of these quantities requires about 200 data points (see Figures 4 and 6) of the learning process. It is also observed that with the d16 dataset, the learning process is significantly faster than for the d08 dataset indicating a stronger signature of the physics in the dataset. Furthermore, it is noted that learned d08 pdf for Q_3 exhibits a slightly bimodal behavior that may be is associated with a lack of combustion in a few data points of the d08 dataset.

The turbulent kinetic energy (TKE), on the other hand required all 256 data points for the convergence of the learning process, both for the d08 and d16 datasets, with distinctly behavior at different streamwise locations. For instance, as observed by inspecting Figures 7 and 8, for Q_6 (TKE after the primary injector and before cavity), the d08 dataset exhibits a much narrower variation than the corresponding the d16 dataset. On the other hand, the bimodal behavior observed for Q_7 (TKE inside the cavity) is present both in the d08 and d16 datasets, which could be explained by the mixing of two turbulence regimes.This bimodality persists in the pdf of the maximum statistics (see Figures 9 and 10) suggesting that each of these turbulent regimes contribute to extreme behavior. We also note that the TKE just after the secondary injectors, Q_8, inside the combustion chamber, Q_9, and at the end of the combustion camber, Q_{10}, exhibit distinct behaviors between the d08 and d16 datasets with Q_8 demonstrating bimodal behavior in both datasets. This bimodality is visible also in the extreme statistics of d08 (see Figures 9 and 10). The bimodality of Q_8, given exposition right after the secondary injectors, can again be attributed to the mixing of two turbulence regimes. At this point, we should note that the learning process for the extreme statistics of TKE Q_6, Q_8, and Q_9 are not converged for the d08 dataset. This suggests that this dataset does not capture sufficient features of the underlying physical processes



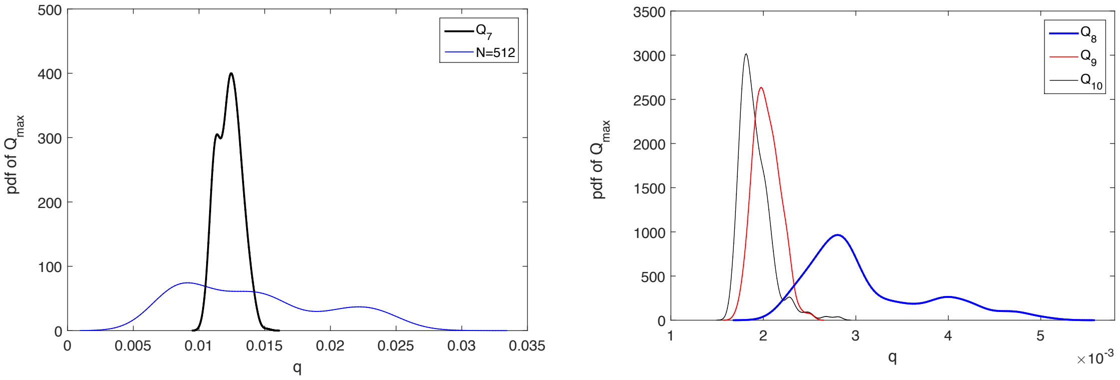
(a) (b)

Fig. 12 d08-d16 dataset: probability density functions $p_{\mathbf{Q}_{\max}}(q)$ of random variable $\mathbf{Q}_{\max}$ (a) for combustion efficiency $\mathbf{Q}_2$ and (b) for pressure stagnation loss ratio $\mathbf{Q}_3$ (right figure), for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 450$ (med blue line), $N = 512$ (thick black line) with $\nu_{\text{sim}} = 51,200$.



(a) (b)

Fig. 13 For the d08-d16 dataset and for $N = 512$ and $\nu_{\text{sim}} = 51,200$: probability density function $p_{\mathbf{Q}}(q)$ of TKE $\mathbf{Q}$. (a): $\mathbf{Q}_6$ (mid blue line) and $\mathbf{Q}_7$ (thin black line). (b): $\mathbf{Q}_8$ (thick blue line), $\mathbf{Q}_9$ (mid red line), and $\mathbf{Q}_{10}$ (thin black line).



(a)

(b)

Fig. 14 For the d08-d16 dataset and for $N = 512$ and $\nu_{\mathrm{sim}} = 51,200$: probability density function $p_{\mathrm{Q}_{\max}}(q)$ of TKE $\mathrm{Q}_{\max}$. (a): Q_6 (mid blue line) and Q_7 (thin black line). (b): Q_8 (thick blue line), Q_9 (mid red line), and Q_{10} (thin black line).

that may be responsible for extreme behavior. Indeed, the learning process for these same statistics is converged for the d16 dataset and with only 200 data points.

Figures 11 to 14 show the pdf of the QoIs for the concatenated d08-d16 dataset. It is observed that, while the learning process is improved by the presence of the d16 data, the width of the pdf is adversely affected by the presence of the d08 data. The bimodality of the extreme values of Q_8 (see Figure 14) is weakly affected by the d08 data. On the other hand, the bimodality of Q_6 to Q_{10} (see Figure 13) is an artifact of concatenating the d08 and d16 data and should not be interpreted as reflecting physical behavior.

VII. Conclusion

In this paper, we have delineated an implicit diffusion manifold and demonstrated its use for enhancing the predictability of complex flows within a scramjet. Leveraging this implicit structure, fewer statistical samples are required to accurately characterize the statistics of LES predictions induced by parametric variations. The analysis is based on a novel probabilistic "learning on manifolds" procedure that generates realizations of a random vector whose non-Gaussian probability distribution is unknown and is presumed to be concentrated on an unknown manifold to be characterized through a probabilistic learning process. Applied to the ScramJet database, the probability density functions of the quantities of interest and their associated maximum statistics are estimated even though the number of simulations available from the LES runs is not sufficient to obtain sufficiently converged estimates of these quantities. We have shown how the probabilistic learning method learns as a function of the size of the datasets. This type of analysis also serves to determine if the dimension of the initial dataset is sufficiently large for providing an assessment of the quality of the probabilistic learning. The analysis of these probability density functions allows for interpreting the physical behavior of the complex turbulent flow in relationship to the mesh size of the fluid domain and the time averaging that is used for constructing the quantities of interest, such as the turbulent kinetic energy at different streamwise locations of the flow.

Acknowledgments

Support for this research was provided by the Defense Advanced Research Projects Agency (DARPA) program on Enabling Quantification of Uncertainty in Physical Systems (EQUiPS).

References

[1] Soize, C., and Ghanem, R., “Data-driven probability concentration and sampling on manifold,” *Journal of Computational Physics*, Vol. 321, 2016, pp. 242–258. doi:10.1016/j.jcp.2016.05.044.

[2] Ghanem, R., and Soize, C., "Probabilistic nonconvex constrained optimization with fixed number of function evaluations," *International Journal for Numerical Methods in Engineering*, 2017, pp. 1–25. doi:10.1002/nme.5632.

[3] Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F., and Zucker, S., "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *PNAS*, Vol. 102, No. 21, 2005, pp. 7426–7431.

[4] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York, 2000.

[5] Aggarwal, C. C., and Zhai, C., *Mining Text Data*, Springer Science & Business Media, New York, 2012.

[6] Dalalyan, A. S., and Tsybakov, A. B., "Sparse regression learning by aggregation and Langevin Monte-Carlo," *Journal of Computer and System Sciences*, Vol. 78, No. 5, 2012, pp. 1423–1443. doi:10.1016/j.jcss.2011.12.023.

[7] Murphy, K. P., *Machine Learning: A Probabilistic Perspective*, MIT press, 2012.

[8] Balcan, M.-f. F., and Feldman, V., "Statistical active learning algorithms," *in: Advances in Neural Information Processing Systems*, 2013, pp. 1295–1303.

[9] James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning*, Vol. 112, Springer, 2013.

[10] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W., "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2014, pp. 601–610.

[11] Ghahramani, Z., "Probabilistic machine learning and artificial intelligence," *Nature*, Vol. 521, No. 7553, 2015, pp. 452–459. doi:10.1038/nature14541.

[12] Taylor, J., and Tibshirani, R. J., "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences*, Vol. 112, No. 25, 2015, pp. 7629–7634. doi:10.1073/pnas.1507583112.

[13] Ghanem, R., Higdon, D., and Owhadi, H., *Handbook of Uncertainty Quantification*, Springer, 2017.

[14] Jones, D., Schonlau, M., and Welch, W., "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455–492.

[15] Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, K., "Surrogate-based analysis and optimization," *Progress in Aerospace Science*, Vol. 41, No. 1, 2005, pp. 1–28. doi:10.1016/j.paerosci.2005.02.001.

[16] Byrd, R., Chin, G., Neveitt, W., and Nocedal, J., "On the use of stochastic Hessian information in optimization methods for machine learning," *SIAM Journal of Optimization*, Vol. 21, No. 3, 2011, pp. 977–995. doi:10.1137/10079923X.

[17] Homem-de Mello, T., and Bayraksan, G., "Monte Carlo sampling-based methods for stochastic optimization," *Surveys in Operations Researdh and Management Science*, Vol. 19, No. 1, 2014, pp. 56–85. doi:10.1016/j.sorms.2014.05.001.

[18] Keane, A. J., "Statistical improvement criteria for use in multiobjective design optimization," *AIAA Journal*, Vol. 44, No. 4, 2006, pp. 879–891. doi:10.2514/1.16875.

[19] Kleijnen, J., van Beers, W., and van Nieuwenhuyse, I., "Constrained optimization in expensive simulation: novel approach," *European Journal of Operational Research*, Vol. 202, No. 1, 2010, pp. 164–174. doi:10.1016/j.ejor.2009.05.002.

[20] Wang, Z., Zoghi, M., Hutter, F., Matheson, D., and de Freitas, N., "Bayesian optimization in a billion dimensions via random embeddings," *Journal of Artificial Intelligence Research*, Vol. 55, 2016, pp. 361–387. doi:10.1613/jair.4806, URL `http://jair.org/media/4806/live-4806-9131-jair.pdf`.

[21] Xie, J., Frazier, P., and Chick, S., "Bayesian optimization via simulation with pairwise sampling and correlated pair beliefs," *Operations Research*, Vol. 64, No. 2, 2016, pp. 542–559. doi:10.1287/opre.2016.1480.

[22] Du, X., and Chen, W., "Sequential optimization and reliability assessment method for efficient probabilistic design," *ASME Journal of Mechanical Design*, Vol. 126, No. 2, 2004, pp. 225–233. doi:10.1115/1.1649968.

[23] Eldred, M., "Design under uncertainty employing stochastic expansion methods," *International Journal for Uncertainty Quantification*, Vol. 1, No. 2, 2011, pp. 119–146. doi:10.1615/Int.J.UncertaintyQuantification.v1.i2.20.

[24] Yao, W., Chen, X., Luo, W., vanTooren, M., and Guo, J., "Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles," *Progress in Aerospace Sciences*, Vol. 47, 2011, pp. 450–479. doi:10.1016/j.paerosci.2011.05.001.

[25] Kodiyalam, S., and Gurumoorthy, R., “Neural network approximator with novel learning scheme for design optimization with variable complexity data,” *AIAA Journal*, Vol. 35, No. 4, 1997, pp. 736–739. doi:10.2514/2.166.

[26] Luo, H., and Hanagud, S., “Dynamic learning rate neural network training and composite structural damage detection,” *AIAA Journal*, Vol. 35, No. 9, 1997, pp. 1522–1527. doi:10.2514/2.7480.

[27] Tracey, B., Wolpert, D., and Alonso, J. J., “Using supervised learning to improve monte carlo integral estimation,” *AIAA Journal*, Vol. 51, No. 8, 2013, pp. 2015–2023. doi:10.2514/1.J051655.

[28] Singh, A. P., Medida, S., and Duraisamy, K., “Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils,” *AIAA Journal*, Vol. 55, No. 7, 2017, pp. 2215–2227. doi:10.2514/1.J055595.

[29] Dolvin, D. J., “Hypersonic international flight research and experimentation (HIFiRE),” *15th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, Dayton, OH, 2008. doi:10.2514/6.2008-2581.

[30] Dolvin, D. J., “Hypersonic international flight research and experimentation,” *16th AIAA/DLR/DGLR International Space Planes and Hypersonic Systems and Technologies Conference*, Bremen, Germany, 2009. doi:10.2514/6.2009-7228.

[31] Hass, N. E., Cabell, K. F., and Storch, A. M., “HIFiRE Direct-Connect Rig (HDCR), Phase I, Ground test results from the NASA Langley Arc-Heated Scramjet Test Facility,” Tech. rep., NASA, 2010.

[32] Storch, A. M., Bynum, M., Liu, J., and Gruber, M., “Combustor operability and performance verification for HIFiRE Flight 2,” *17th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, San Francisco, CA, 2011. doi:10.2514/6.2011-2249.

[33] Pellett, G. L., Vaden, S. N., and Wilson, L. G., “Opposed jet burner extinction limits: Simple mixed hydrocarbon scramjet fuels vs air,” *43rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*, Cincinnati, OH, 2007. doi:10.2514/6.2007-5664.

[34] Lu, T., and Law, C. K., “A directed relation graph method for mechanism reduction,” *Proceedings of the Combustion Institute*, Vol. 30, No. 1, 2005, pp. 1333–1341. doi:10.1016/j.proci.2004.08.145.

[35] Zambon, A. C., and Chelliah, H. K., “Explicit reduced reaction models for ignition, flame propagation, and extinction of C2H4/CH4/H2 and air systems,” *Combustion and Flame*, Vol. 150, No. 1-2, 2007, pp. 71–91. doi:10.1016/j.combustflame.2007.03.003.

[36] Lacaze, G., Vane, Z. P., and Oefelein, J. C., “Large eddy simulation of the HIFiRE direct connect rig scramjet combustor,” *55th AIAA Aerospace Sciences Meeting*, Grapevine, TX, 2017. doi:10.2514/6.2017-0142.

[37] Oefelein, J. C., “Large eddy simulation of turbulent combustion processes in propulsion and power systems,” *Progress in Aerospace Sciences*, Vol. 42, No. 1, 2006, pp. 2–37. doi:10.1016/j.paerosci.2006.02.001.

[38] Oefelein, J. C., “Simulation and analysis of turbulent multiphase combustion processes at high pressures,” Ph.D. thesis, The Pennsylvania State University, 1997.

[39] Oefelein, J. C., Schefer, R. W., and Barlow, R. S., “Toward validation of large eddy simulation for turbulent combustion,” *AIAA Journal*, Vol. 44, No. 3, 2006, pp. 418–433. doi:10.2514/1.16425.

[40] Oefelein, J. C., Lacaze, G., Dahms, R., Ruiz, A., and Misdariis, A., “Effects of real-fluid thermodynamics on high-pressure fuel injection processes,” *SAE International Journal of Engines*, Vol. 7, No. 3, 2014, pp. 1125–1136. doi:10.4271/2014-01-1429.

[41] Lacaze, G., Misdariis, A., Ruiz, A., and Oefelein, J. C., “Analysis of high-pressure Diesel fuel injection processes using LES with real-fluid thermodynamics and transport,” *Proceedings of the Combustion Institute*, Vol. 35, No. 2, 2015, pp. 1603–1611. doi:10.1016/j.proci.2014.06.072.

[42] Gruber, M. R., Jackson, K., and Liu, J., “Hydrocarbon-fueled scramjet combustor flowpath development for Mach 6-8 HIFiRE flight experiments,” Tech. rep., AFRL, 2008.

[43] Soize, C., “Polynomial chaos expansion of a multimodal random vector,” *SIAM/ASA Journal on Uncertainty Quantification*, Vol. 3, No. 1, 2015, pp. 34–60. doi:10.1137/140968495.

[44] Bowman, A., and Azzalini, A., *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford, UK, 1997.

[45] Scott, D., *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd ed., John Wiley and Sons, New York, 2015.

[46] Soize, C., “Construction of probability distributions in high dimension using the maximum entropy principle. Applications to stochastic processes, random fields and random matrices,” *International Journal for Numerical Methods in Engineering*, Vol. 76, No. 10, 2008, pp. 1583–1611. doi:10.1002/nme.2385.

[47] Girolami, M., and Calderhead, B., “Riemann manifold Langevin and Hamiltonian Monte Carlo methods,” *Journal of the Royal Statistics Society*, Vol. 73, No. 2, 2011, pp. 123–214. doi:10.1111/j.1467-9868.2010.00765.x.

[48] Neal, R., “MCMC using Hamiltonian dynamics,” *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. Meng, Chapman and Hall-CRC Press, Boca Raton, 2012.

[49] Spall, J., *Introduction to Stochastic Search and Optimization*, John Wiley and Sons, Hoboken, New Jersey, 2003.