# Enhancing model predictability for a scramjet using probabilistic learning on manifolds

Christian Soize, Roger Ghanem, Cosmin Safta, Xun Huan, Zachary P Vane, Joseph C Oefelein, Guilhem Lacaze, Habib N Najm

▶ **To cite this version:**

Christian Soize, Roger Ghanem, Cosmin Safta, Xun Huan, Zachary P Vane, et al.. Enhancing model predictability for a scramjet using probabilistic learning on manifolds. AIAA Journal, 2019, 57 (1), pp.365-378. 10.2514/1.J057069 . hal-02052839

**HAL Id: hal-02052839**
**https://hal.science/hal-02052839v1**

Submitted on 28 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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 is an approach that allows for extracting more information using fewer samples. Such procedures, if successful, would greatly enhance the predictability of models in the sense of improving exploration and characterization of uncertainty due to model error and input dependencies, all while being 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 in a probabilistic framework 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@alumni.stanford.edu).

‖Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Associate Fellow (joseph.oefelein@aerospace.gatech.edu).

**Combustion Research Facility, 7011 East Avenue, Mail Stop 9051, Livermore, CA 94551, USA, AIAA Member (guilhem.lacaze@gmail.com).

††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 $\mathbf{w}$ | $\mathbf{q}_{\text{ar}}^{\ell}$ | = | $\ell$-th additional realization of $\mathbf{Q}$ |
| $\eta_c$ | = | combustion efficiency | $q_k$ | = | component $k$ of $\mathbf{q}$ |
| $m_w$ | = | dimension of $\mathbf{w}$ or $\mathbf{W}$ | $q_{\text{max}}^{\alpha}$ | = | $\alpha$-th realization of $\text{Q}_{\text{max}}$ |
| $N$ | = | number of data points | $R_{\bar{P}}$ | = | stagnation pressure loss ratio |
| $N_{\text{sup}}$ | = | maximum value of $N$ | $\mathbb{R}$ | = | set of all the real numbers |
| $n$ | = | dimension of $\mathbf{x}$of $\mathbf{X}$ | $\mathbb{R}^{m_w}$ | = | Euclidean space of dimension $m_w$ |
| $n_q$ | = | number of QoI | $\mathbb{R}^{n}$ | = | Euclidean space of dimension $n$ |
| $\nu$ | = | dimension of $\mathbf{y}$ or $\mathbf{Y}$ | $\mathbb{R}^{n_q}$ | = | Euclidean space of dimension $n_q$ |
| $\nu_{\text{sim}}$ | = | number of additional realizations | $\mathbb{R}^{\nu}$ | = | Euclidean space of dimension $\nu$ |
| $p_{\mathbf{Q}}$ | = | pdf of $\mathbf{Q}$ | TKE | = | wall-normal averaged turbulence kinetic energy |
| $p_{\text{Q}}$ | = | pdf of Q | $\mathbf{w}$ | = | $(w_1, \ldots, w_{m_w})$, vector of parameters |
| $p_{\text{Q}_{\text{max}}}$ | = | pdf of $\text{Q}_{\text{max}}$ | $\mathbf{w}^{\ell}$ | = | $\ell$-th realization of $\mathbf{W}$ |
| $p_{\mathbf{W}}$ | = | pdf of $\mathbf{W}$ | $\mathbf{w}_{\text{ar}}^{\ell}$ | = | $\ell$-th additional realization of $\mathbf{W}$ |
| $p_{\mathbf{X}}$ | = | pdf of $\mathbf{X}$ | $w_j$ | = | component $j$ of $\mathbf{w}$ |
| QoI | = | Quantity of Interest | $\mathbf{W}$ | = | $(W_1, \ldots, W_{m_w})$, random parameters |
| $\mathbf{Q}$ | = | $(\text{Q}_1, ..., \text{Q}_{n_q})$, random QoI | $W_j$ | = | component $j$ of $\mathbf{W}$ |
| Q | = | any component of $\mathbf{Q}$ | $\mathbf{X}$ | = | $(X_1, \ldots, X_n) = (\mathbf{W}, \mathbf{Q})$ |
| $\text{Q}_k$ | = | component $k$ of $\mathbf{Q}$ | $X_j$ | = | component $j$ of $\mathbf{X}$ |
| $\text{Q}_{\text{max}}$ | = | maximum of Q | $\mathbf{x}$ | = | $(x_1, \ldots, x_n) = (\mathbf{w}, \mathbf{q})$ |
| QoI | = | Quantity of interest | $\mathbf{x}^{\ell}$ | = | $\ell$-th realization of $\mathbf{X}$ |
| $\mathbf{q}$ | = | $(q_1, \ldots, q_{n_q})$ | $\mathbf{x}_{\text{ar}}^{\ell}$ | = | $\ell$-th additional realization of $\mathbf{X}$ |
| $\mathbf{q}^{\ell}$ | = | $\ell$-th realization of $\mathbf{Q}$ | $x_j$ | = | 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 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. The hope is that sufficient learning be achieved from a relatively small number of simulations, in anyway 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 use the recent approach devoted to probabilistic learning on manifolds [1] to the challenges presented by large-eddy simulations (LES) of reactive flows inside a scramjet combustor. While investigations adopting probabilistic approaches for scramjet applications are growing in recent years [2–8], substantial challenges remain in characterizing and predicting combustion properties for turbulent flows under extreme conditions especially in conjunction with uncertainty quantification. We are particularly interested in employing and enabling probabilistic methods with LES, since these simulations, while computationally more demanding, can allow us to access some turbulence details and features often not available through models involving additional simplifications, such as with Reynolds-averaged Navier-Stokes (RANS). Indeed, enabling uncertainty quantification with LES involves pushing the limits of computational science and engineering, and is recognized as one of the grand challenges of scramjets computations [9]. More precisely, this paper is a first stage for enhancing the probabilistic predictability of the computational model. The second stage could be a design optimization under uncertainty using the approach detailed in [10], but which is not presented in this paper.

Available data refers here to numerically generated data that, as indicated above, will be limited in view of the expense associated with its generation. These generated data correspond to realizations of a random vector $\mathbf{X} = (\mathbf{W}, \mathbf{Q})$

that is constituted of the vector $\mathbf{W}$ of the uncertain parameters of the computational model to which is added the vector $\mathbf{Q}$ of all the random quantities of interest that are the outputs of the computational model. Consequently, there exists an unknown mapping $\mathbf{f}$, characterizing the computational model such that $\mathbf{Q} = \mathbf{f}(\mathbf{W})$ and defining a manifold (that is unknown). The unknown probability measure of $\mathbf{X}$ will be estimated using the generated data that correspond to realizations of $\mathbf{X}$. The support of this probability measure is this manifold. The procedure permits to discover this unknown support (the manifold) through a Markov process constructed using only the generated data [11]. A sampling procedure is then put in place for augmenting the initial dataset with additional samples generated with the probability measure whose support is the manifold. 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 methodology for solving stochastic nonconvex optimization problems using the probabilistic learning on manifolds, which is used in this paper, has been developed and validated on simple examples (see [10]). This methodology is being developed for very complex optimization problems.

It should be noted that the statistical and probabilistic learning methods have been extensively developed [12–20]) and play an increasingly important role in computational science and engineering [21]), 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 [22–25]. Gaussian process models are most commonly used in this context (see for instance [26, 27]), as well as the approaches based on Bayesian methods including the Bayesian optimization as proposed in [22, 28, 29]. For the evaluations of expensive stochastic functions in presence of uncertainties, computational challenges remain currently significant enough to require relevant probabilistic approximations [24, 30–32]. 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 [33, 34]. More recently, postprocessing of a given set of Monte Carlo realizations has been proposed for improving integral computation [35] and a machine-learning approach has been used [36] 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 , in the sense of improving exploration and characterization of uncertainty due to model error and input dependencies within a probabilistic framework, 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 involves (1) discovering and 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) preserving the concentration of the additional realizations around the manifold, and (3) constructing an associated Markov Chain Monte Carlo (MCMC) generator for generating additional realizations that follow the estimated probability distribution.

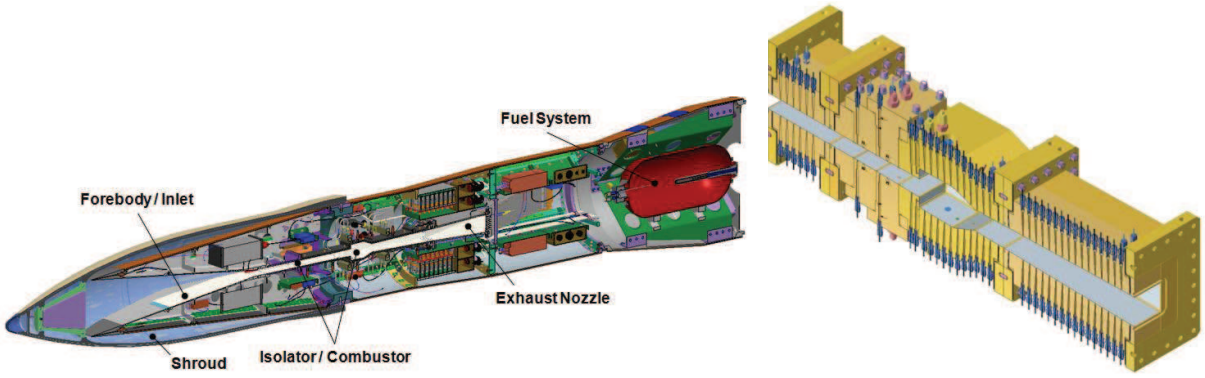
The paper is organized as follows. In Section II, we summarize the physical and computational model that is used for simulating the complex flow for a ScramJet by means of a large scale computational fluid dynamics model. This section allows also for defining the uncertain parameters of the computational 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 (which are extreme value 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 [37, 38]. One of its flight tests, the HIFiRE Flight 2 (HF2) project [39–41], involved a payload depicted in Figure 1(a) and was tested under flight conditions of Mach 6–8+. The configuration consists of a cavity-based hydrocarbon-fueled dual-mode scramjet. A ground test rig, designated the HIFiRE Direct Connect Rig (HDCR) (Figure 1(b)), was developed to duplicate the isolator/combustor layout of the flight test hardware, and to provide ground-based measurements for comparisons with flight test data, verifying engine performance and operability, and designing fuel delivery schedule [42, 43]. Since the HDCR ground test data are publicly available [42, 44], we aim to simulate and assess reactive flows inside the HDCR with the intention of leveraging existing experimental datasets to drive future modeling developments.

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 (stream-

wise), 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. The consequence of this approximation is to exclude any asymmetric modes of the flow dynamics from the present modeling framework. 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 surrogate fuel [45], composed of 36% methane and 64% ethylene by volume, enters through these injectors. A




(a) HIFiRE Flight 2 payload (b) HDCR

**Fig. 1 HIFiRE Flight 2 payload [40] and HDCR cut views [42].**

reduced, three-step mechanism [46, 47] is initially adopted to describe the combustion process:

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

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

$$2CO + O_2 \rightleftharpoons 2CO_2, \quad (3)$$

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 selected to retain robust/stable combustion in the current simulations.

LES calculations are then performed using the RAPTOR code framework developed by Oefelein [48, 49]. The theoretical framework solves the fully coupled conservation equations of mass, momentum, total-energy, and species for a chemically reacting flow while accounting for detailed thermodynamics and transport processes at the molecular level. It is designed to handle high Reynolds number, high-pressure, real-gas and/or liquid conditions over a wide Mach operating range. 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

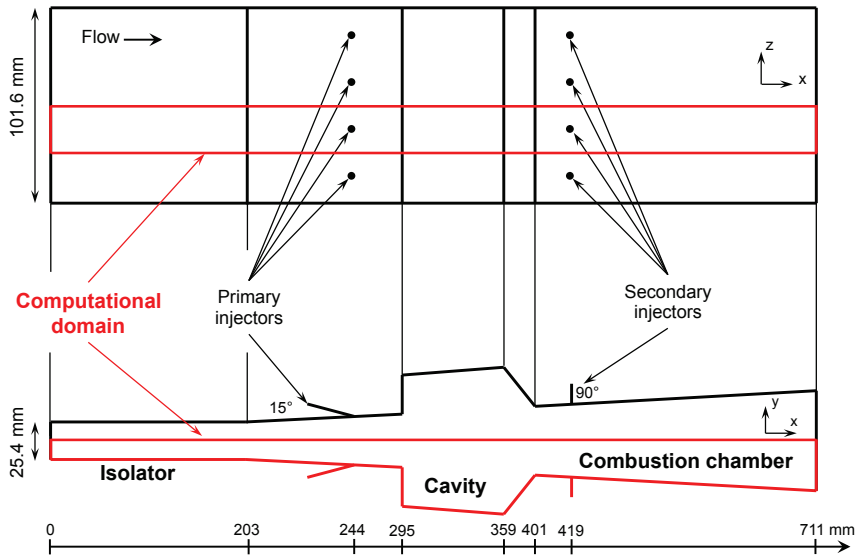



**Fig. 2 The HDCR experimental setup and schematic of the full computational domain.**

Oefelein *et al.* [50–52].

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. The uncertain parameters reflect uncertainty in inlet and fuel inflow boundary conditions as well as turbulence model parameters from utilizing the Smagorinsky model. The nominal values of the operating conditions for our combustor domain correspond to the Mach 5.84 flight condition, and these values were calculated in past CFD analysis and employed for the HDCR ground tests [42]. With lower and upper bounds suggested by domain experts, we invoke the maximum entropy principle [53, 54] and endow the parameters with uninformative uniform "prior" distributions across the ranges indicated in the table. Our simulation data are from two-dimensional scramjet computations, employing grid resolutions where cell sizes are $1/8$ (referred as "d08" grids) and $1/16$ (referred as "d16" grids) of the injector diameter $d = 3.175$ mm. Calculations are performed on the two grid levels from their respective warm-start solutions that were engineered from a quasi-steady state nominal condition simulation. A run length of $10^5$ time steps is selected to balance between washing out transient start-up behavior and operating under practical constraints of limited computational resources. The last halves of these runs time-histories are used for time-averaging. Timestep sizes are determined adaptively based on guidance from the Courant-Friedrichs-Lewy (CFL) condition. A total number of 256 simulations is performed for both the d08 and d16 grids in establishing our database, and their average CPU times per run are roughly 743 hours and 2160 hours, respectively.

We would also like to point out some limitations of our numerical results in the current paper stemming from additional simplifications necessitated by practical considerations. In particular, constraints on computational resources both encouraged and compelled a current investigation involving simulations in a two-dimensional geometry, where we placed a single cell in the $z$-direction at a $x$-$y$ plane intersecting the injectors. We fully acknowledge the decreased

fidelity of these runs as a result of the reduced geometric description as well as the relatively simple chemical model in Eq.(1-3). Indeed, certain physical features and phenomenon are eroded or otherwise not representable in a two-dimensional setting. Nonetheless, given the scale of computations demanded by any form of statistical assessments, enabling computational methods under a probabilistic framework even with these emulatory settings has not been achieved previously. At the same time, fully three-dimensional simulations are computationally possible but only for relatively coarse grids and where only a very small number of runs can be completed under the present computational budget; they are thus not ready to support a meaningful demonstration of the centerpiece of this paper—the manifold learning technique. While certainly desirable under ideal situations, seeking higher-fidelity scramjet LES datasets for the purpose of this study would be practically impossible to achieve at this time. Nonetheless, we emphasize the high degree of information, and enhanced fidelity, available from the present LES computations of this flow, in terms of both flow/flame structure and dynamics, as opposed to, say RANS simulations. The present results highlight what is indeed currently achievable in the context of UQ for scramjet LES computations emplolying state of the art UQ methods.

We focus on three quantities of interest (QoIs): (1) combustion efficiency ($\eta_c$) that is related to the burned equivalence ratio ($\phi_B$), (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.

- **Combustion efficiency ($\eta_c$)** is the combustion efficiency based on static enthalpy quantities [43, 55]:

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

  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).

| 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 probability distributions are assumed uniform across the ranges defined.**

## 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 and uncertainty quantification 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 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 [56] of the classical multidimensional Gaussian kernel-density estimation method [57, 58] 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 [56, 59] belonging to the class of Hamiltonian Monte Carlo methods [59–61], which is an MCMC algorithm [62].
4) The diffusion-map approach [11] 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 $[z^1_{\text{ar}}],\ldots,[z^{n_{\text{MC}}}_{\text{ar}}]$ of random matrix $[\mathbf{Z}]$, and therefore, for deducing the additional realizations $[y^1_{\text{ar}}],\ldots,[y^{n_{\text{MC}}}_{\text{ar}}]$ of random matrix $[\mathbf{Y}]$. Reshap-

ing these $n_{\text{MC}}$ matrices yields the $\nu_{\text{sim}} = N \times n_{\text{MC}}$ independent realizations $\{\mathbf{y}^1, \ldots, \mathbf{y}^{\nu_{\text{sim}}}\}$ of random vector $\mathbf{Y}$. Using the PCA constructed in Step 1 allows for generating the $\nu_{\text{sim}} \gg N$ additional independent realizations $\{\mathbf{x}_{\text{ar}}^1, \ldots, \mathbf{x}_{\text{ar}}^{\nu_{\text{sim}}}\}$ in $\mathbb{R}^n$ of random vector $\mathbf{X}$.

*Remark 1.* For the general case of the probabilistic learning on manifolds, the pdf of $p_{\mathbf{X}}$ of $\mathbf{X}$ is unknown. As explained in Section I, $\mathbf{X}$ is written as $\mathbf{X} = (\mathbf{W}, \mathbf{Q})$ in which $\mathbf{W}$ is the random vector of the uncertain parameters of the computational model and $\mathbf{Q}$ is the random vector of the quantities of interest that are the outputs of the computational model. Consequently, the pdf $p_{\mathbf{W}}(\mathbf{w})$ of $\mathbf{W}$ is not supposed to be known. However, for certain applications for which the probabilistic learning on manifolds is used, $p_{\mathbf{W}}(\mathbf{w})$ is known and is used for generating the realizations of $\mathbf{W}$ in order to generate the dataset $\{\mathbf{x}^\ell = (\mathbf{w}^\ell, \mathbf{q}^\ell), \ell = 1, \ldots, N\}$ in which $\mathbf{q}^\ell$ is the realization of $\mathbf{Q}$ computed with the computational model for $\mathbf{W} = \mathbf{w}^\ell$ (this point will be detailed in Section IV for the ScramJet application). One could then wonder if the knowledge of $p_{\mathbf{W}}$ could be useful in the probabilistic learning on manifolds in addition to its use to generate the $N$ points of the dataset. In fact, it is not so for the following reasons. Starting from the dataset of $N$ points constituted of realizations of random vector $\mathbf{X} = (\mathbf{W}, \mathbf{Q})$, we want to generate $\nu_{\text{ar}} \gg N$ additional realizations $\{\mathbf{x}_{\text{ar}}^\ell\}_\ell$ (by the probabilistic learning) in order to improve the estimate of the probability distribution of $\mathbf{X}$ whose support is the manifold defined by the mapping that maps $\mathbf{W}$ in $\mathbf{Q}$. For such a characterization, the joint probability distribution $p_{\mathbf{W},\mathbf{Q}}(\mathbf{w}, \mathbf{q})\, d\mathbf{w}\, d\mathbf{q}$ is required, which means that we need the probability distribution $p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$ of $\mathbf{X}$. The pdf $p_{\mathbf{W}}(\mathbf{w})$ of $\mathbf{W}$ is only used for generate the initial dataset constituted of $N$ points. The use of $p_{\mathbf{W}}(\mathbf{w})$ in the learning process would require to introduce the conditional pdf $p_{\mathbf{Q}|\mathbf{W}}(\mathbf{q}|\mathbf{w})$ of $\mathbf{Q}$ given $\mathbf{W} = \mathbf{w}$ such that $p_{\mathbf{W},\mathbf{Q}}(\mathbf{w}, \mathbf{q}) = p_{\mathbf{Q}|\mathbf{W}}(\mathbf{q}|\mathbf{w})\, p_{\mathbf{W}}(\mathbf{w})$. Formulated in terms of such a conditional pdf, the probabilistic learning process would be strictly equivalent to the probabilistic learning formulated in terms of $p_{\mathbf{W},\mathbf{Q}}(\mathbf{w}, \mathbf{q})\, d\mathbf{w}\, d\mathbf{q} = p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$, because the number of additional realizations of $\mathbf{Q}|\mathbf{w}$ given $\mathbf{W} = \mathbf{w}$ must be the same that the number of additional realizations of $\mathbf{W}$ in order to correctly represents the manifold. In addition, each realization of $\mathbf{Q}$ should then be associated with the corresponding realization of $\mathbf{W}$. It would be a nontrivial time-consuming problem because it would require an additional smoothing step. Consequently, there would be a loss of efficiency without improving the learning procedure that is proposed.

*Remark 2.* The transition kernel of the homogeneous Markov chain of the Markov Chain Monte Carlo (MCMC) method can be constructed using the Metropolis-Hastings algorithm (that requires the definition of a good proposal distribution), the Gibbs sampling (that requires the knowledge of the conditional distribution) or the slice sampling (that can exhibit difficulties related to the general shape of the probability distribution, in particular for multimodal distributions). In general, these algorithms are efficient, but can also be not efficient if there exist attraction regions which do not correspond to the invariant measure under consideration and tricky even in high dimension. These cases cannot easily be detected and are time consuming. The MCMC method used for constructing the probabilistic learning on manifolds [1], which is based on a nonlinear Itô stochastic differential equation (ISDE) formulated for a dissipative

Hamiltonian dynamical system (first introduced in [59]), has been used for the following reasons:

(i) This Hamiltonian MCMC method is very robust. It looks similar to the Gibbs approach but corresponds to a more direct construction of a random generator of realizations for random matrix $[\mathbf{Y}]$ that can be in very high dimension and for which its probability distribution $p_{[\mathbf{Y}]}([y])\, d[y]$ can have a support that is not a connected set and that is multimodal. The difference between the Gibbs algorithm and the proposed algorithm is that the convergence properties in the proposed method can be studied with all the mathematical results concerning the existence and uniqueness of Itô stochastic differential equation. In addition, a parameter is introduced, which allows the transient part of the response to be killed in order to get more rapidly the stationary solution corresponding to the invariant measure. The construction of the transition kernel by using the detailed balance equation is replaced by the construction of an Itô Stochastic Differential Equation (ISDE), which admits $p_{[\mathbf{Y}]}])\, d[y]$ as a unique invariant measure.

(ii) The second fundamental reason is the possibility to take into account the local geometry structure of the dataset by projecting the nonlinear ISDE on the subspace spanned by the diffusion-maps basis. This aspect is the main innovation introduced in the construction of the probabilistic learning on manifolds [1], thanks to this choice of the MCMC generator.

*Remark 3.* The methodology of the probabilistic learning on manifolds (see [1]) introduces two hyperparameters. The construction of the diffusion-maps basis involves an isotropic-diffusion kernel with a first hyperparameter $\varepsilon$. The second hyperparameter in the dimension, $m \ll N$, of the diffusion-maps basis for projecting the ISDE. In a recent work, an entropy-based closure has been developed for identifying optimal values of $\varepsilon$ and $m$ using only the $N$ points of the dataset. This entropy argument ensures that out of all possible models, this is the one that is the most uncertain beyond any specified constraints, which is selected. The presentation of this complement of the methodology is outside the scope of the present paper.

## 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 its known probability distribution $P_{\mathbf{X}}(d\mathbf{x})$ 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 $n_q = 10$. There is an unknown measurable mapping $\mathbf{w} \mapsto \mathbf{f}(\mathbf{w})$ from $C_{\mathbf{W}} \subset \mathbb{R}^{m_w}$ into $\mathbb{R}^{n_q}$ such that $\mathbf{q} = \mathbf{f}(\mathbf{w})$. Consequently, the random QoI is the $\mathbb{R}^{n_q}$-valued random variable defined on $(\Theta, \mathcal{T}, \mathcal{P})$, which is such that $\mathbf{Q} = \mathbf{f}(\mathbf{W})$. The probability distribution $P_{\mathbf{Q}}(d\mathbf{q})$ of $\mathbf{Q}$ is the image of $P_{\mathbf{W}}(d\mathbf{w})$ under mapping $\mathbf{f}$. 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.

*Remark.* The probabilistic learning on manifolds can be used for a more general case for which $\mathbf{Q} = \mathbf{F}(\mathbf{W})$ in which $\mathbf{F}$ is a random mapping that can be written as $\mathbf{F}(\mathbf{W}) = \mathbf{f}(\mathbf{W}, \mathbf{U})$ where $(\mathbf{w}, \mathbf{u}) \mapsto \mathbf{f}(\mathbf{w}, \mathbf{u})$ is a measurable mapping from $C_{\mathbf{W}} \times C_{\mathbf{U}} \subset \mathbb{R}^{m_w} \times \mathbb{R}^{m_u}$ into $\mathbb{R}^{n_q}$, and where the joint probability distribution of random variables $(\mathbf{W}, \mathbf{U})$ is $P_{\mathbf{W},\mathbf{U}}(d\mathbf{w}, d\mathbf{u})$ whose support is $C_{\mathbf{W}} \times C_{\mathbf{U}}$.

**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 $N_{\text{sup}}$. 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 probability distribution of the quantities of interest with respect to the value of $N$ when $N$ will go to $N_{\text{sup}}$ will be carried out. In the following, the terminology "convergence analysis of the probabilistic learning" or simply "convergence of the learning" will refer to this definition. 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

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

in which the meaning of the symbols used are detailed in Section IV.A.

*Remark.* As explain in Section II, the probability distribution of random vector $\mathbf{W}$ has been chosen as a uniform distribution on $C_{\mathbf{w}}$ for focusing the analysis on the uncertainty propagation. This probabilistic model corresponds to the use of the Maximum Entropy principle from Information Theory, for which the only available information is the support $C_{\mathbf{w}}$ of the unknown pdf of random vector $\mathbf{W}$.

**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 $N_{\text{sup}}$ 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 $P_{\mathbf{X}}(d\mathbf{x})$ 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 probability distribution $P_{\mathbf{X}}(d\mathbf{x})$ on $\mathbb{R}^n$ of $\mathbf{X} = (\mathbf{W}, \mathbf{Q})$ can also be rewritten as the joint probability distribution $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ on $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ of $\mathbf{W}$ and $\mathbf{Q}$. It should be noted that since mapping $\mathbf{f}$ is deterministic, probability distribution $P_{\mathbf{X}}(d\mathbf{x})$ cannot be represented by a pdf $p_{\mathbf{X}}(\mathbf{x})$ with respect to the Lebesgue measure $d\mathbf{x}$ on $\mathbb{R}^n$ (see Appendix).

As explained in Section III, for the considered fixed value of $N$, the probabilistic learning will allow for generating $\nu_{\text{sim}}$ 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 $\nu_{\text{sim}}$ additional realizations $\{\mathbf{q}_{\text{ar}}^1, \ldots, \mathbf{q}_{\text{ar}}^{\nu_{\text{sim}}}\}$ for $\mathbf{Q}$ such that,

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

*Remark*. As explained before, the probabilistic learning on manifolds only uses a data set of $N$ points that are constructed using the computational model and allows for constructing $\nu_{\text{sim}} \gg N$ additional realizations without using the computational model, in order to better estimate the probability distribution of the random QoI. A natural question can then be asked. Are the additional realizations that are generated satisfy the computational model? The answer is yes in the sense of probabilities. Since the proposed approach aims to improve the predictability of the model in a probabilistic framework, the objective of the proposed method is achieved. Detailed explanations are given in Appendix on this point.

**B. Selecting the Random QoI for the Statistical Estimates**

The random vector $\mathbf{Q}$ is completely defined by its probability distribution $P_{\mathbf{Q}}(d\mathbf{q})$ that is assumed to have a density $\mathbf{q} \mapsto p_{\mathbf{Q}}(\mathbf{q})$ on $\mathbb{R}^{n_q}$ with respect to the Lebesgue measure $d\mathbf{q}$, which can be estimated using nonparametric statistics with a large number, $\nu_{\text{sim}}$, 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 $\mathrm{Q}_k$ 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 $\mathrm{Q}_k$ (except if confusion is possible).

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

In this paragraph, we define the maximum statistics for the selected random QoI, which allow us to explore the probability distribution of a random variable $\mathrm{Q}_{\max}$ whose realizations are in the tail of the probability distribution of random variable Q. These statistics also make it possible to well analyze the convergence of learning, that is, the convergence with respect to N. 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 $\nu_s$, the maximum of the real-valued random variable Q can classically be defined as the real-valued random variable $\mathrm{Q}_{\max}$ such that $\mathrm{Q}_{\max} = \max\{\mathrm{Q}^{(1)},\ldots,\mathrm{Q}^{(\nu_s)}\}$, in which $\mathrm{Q}^{(1)},\ldots,\mathrm{Q}^{(\nu_s)}$ are $\nu_s$ independent copies of real-valued random variable Q. Random variable $\mathrm{Q}_{\max}$ depends on $\nu_s$, but in order to simplify the notations, the dependence on $\nu_s$ is removed. The realizations of $\mathrm{Q}_{\max}$ are computed as follows. For fixed $N$ such that $N \leq N^{\max}$, for a given value $\nu_{\text{sim}}$ of additional realizations $\{(\mathbf{w}_{\text{ar}}^{\ell}, q_{\text{ar}}^{\ell}) \in \mathbb{R}^{m_w} \times \mathbb{R}, \ell = 1,\ldots,\nu_{\text{sim}}\}$ introduced in Section V. C and computed thanks to the probabilistic learning, and for $\nu_s$ sufficiently large such that $\nu_s \ll \nu_{\text{sim}}$, we construct $\nu_\alpha = \nu_{\text{sim}}/\nu_s$ independent realizations $\{q_{\max}^1,\ldots,q_{\max}^{\nu_\alpha}\}$ of $\mathrm{Q}_{\max}$ such that, for $\alpha = 1,\ldots,\nu_\alpha$, $q_{\max}^{\alpha} = \max_{\ell \in \{\nu_s(\alpha-1)+1,\ \ldots,\ \alpha\nu_s\}} q_{\text{ar}}^{\ell}$. For the Scramjet results presented in Section VI and for a fixed number $\nu_{\text{sim}}$ of additional realizations (that is a finite number!), a

convergence analysis of the estimated probability density function of $Q_{max}$ has been performed as a function of $\nu_s$. 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 $Q_{max}$ is studied for the enhancing of the model prediction. Note that, since $\nu_{sim}$ can arbitrarily be increased without significant computational cost, $\nu_s$ and $\nu_\alpha$ could arbitrarily be increased in satisfying the equation $\nu_{sim} = \nu_\alpha \times \nu_s$ with $\nu_s < \nu_\alpha$.

### D. Estimates of the Second-order Moments and the pdf of Random Variables Q and $Q_{max}$

For a fixed value of $N$, $\nu_{sim}$, and $\nu_s$ (and consequently, of $\nu_\alpha = \nu_{sim}/\nu_s$), the standard deviations $\sigma_Q$ and $\sigma_{Q_{max}}$ of the real-valued random variables Q and $Q_{max}$, and their probability density functions $q \mapsto p_Q(q)$ and $q \mapsto p_{Q_{max}}(q)$ 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 $\{q_{ar}^1, \ldots, q_{ar}^{\nu_{sim}}\}$ for Q and of the realizations $\{q_{max}^1, \ldots, q_{max}^{\nu_\alpha}\}$ for $Q_{max}$ (for 10 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 $\nu_{sim}$ (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 $q \mapsto p_Q(q)$ and $q \mapsto p_{Q_{max}}(q)$ are shown.

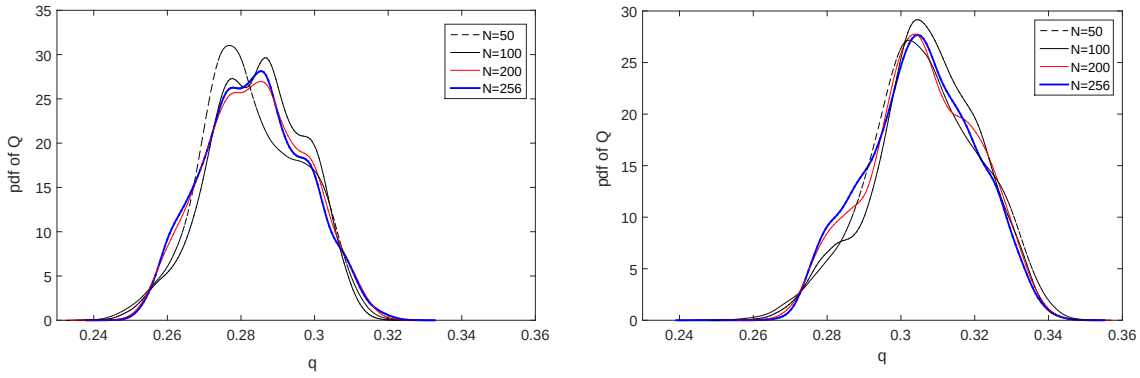
## VI. Numerical Simulations and Statistical Analysis for the Datasets of the ScramJet Database

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 **W** (11 components) and with all the components of **Q** (10 components). The components, $Q_k$, of random vector **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 $\nu_{sim} = 25,600$ 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 $\nu_{sim} = N \times n_{MC}$ 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 $\{Q_k, k = 6, \ldots, 10\}$, for $N = 256$ and for $\nu_{sim} = 25,600$ additional realizations (Section VI. B.2).
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 $\nu_{sim} = N \times n_{MC} = 51,200$ (Section VI. C.1).




(a) d08 dataset

(b) d16 dataset

**Fig. 3 Combustion efficiency $Q_2$: 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 black line) with $\nu_{\text{sim}} = 25,600$.**

4) Finally, for the concatenated d08-d16 dataset, the model predictability of TKE is again performed at the same streamwise locations corresponding to $\{Q_k, k = 6,\ldots,10\}$, for $N = 512$ and $\nu_{\text{sim}} = 51,200$ additional realizations (Section VI. C.2).

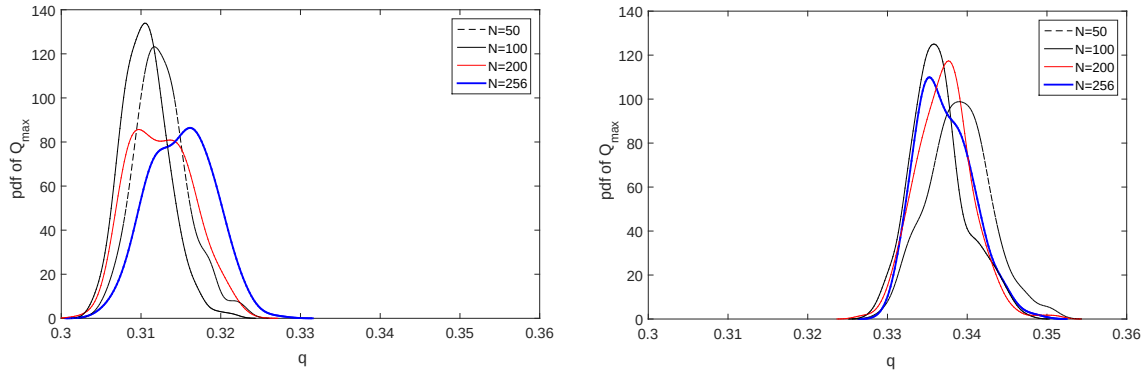
### B. Probabilistic Learning Approach for Analyzing the d08 and d16 Datasets

*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 $\nu_{\text{sim}} = 25,600$, 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 $p_{\text{Q}}(q)$ of random variable Q (Figures 3 and 5) and $p_{\text{Q}_{\text{max}}}(q)$ of random variable $Q_{\text{max}}$ (Figures 4 and 6).

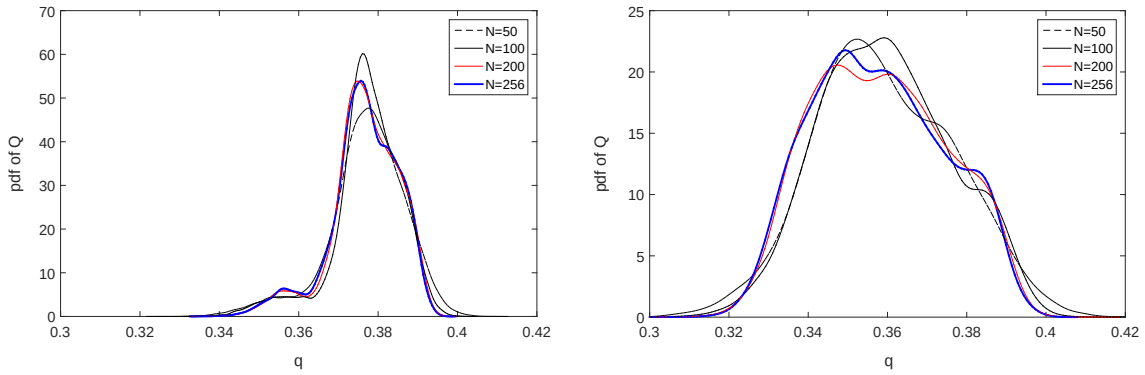
*2. Model Predictability of the Wall-Normal averaged Turbulence Kinetic Energy Performed at Various Streamwise Locations Using the Probabilistic Learning Approach*

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 $p_{\text{Q}}(q)$ 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 $p_{\text{Q}_{\text{max}}}(q)$ of random variable $Q_{\text{max}}$ is shown in Figure 9 and 10.



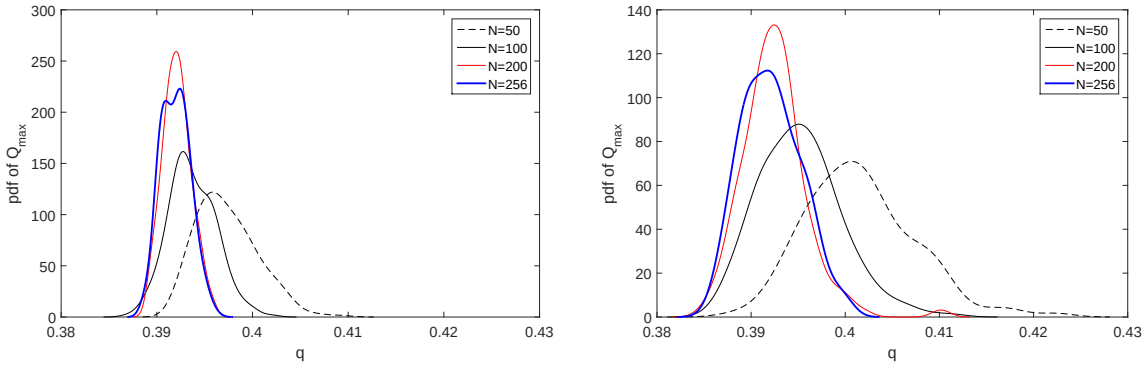

(a) d08 dataset (b) d16 dataset

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



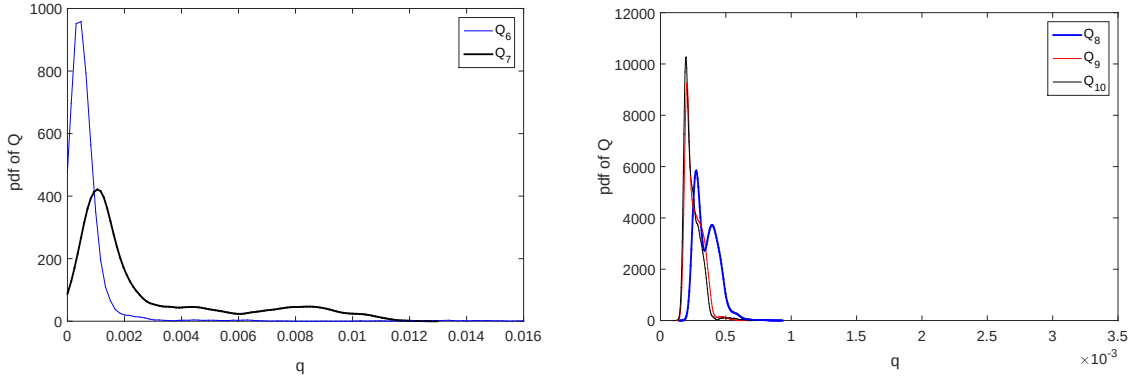

(a) d08 dataset (b) d16 dataset

**Fig. 5 Pressure stagnation loss ratio $Q_3$: 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 black line) with $\nu_{sim} = 25,600$.**



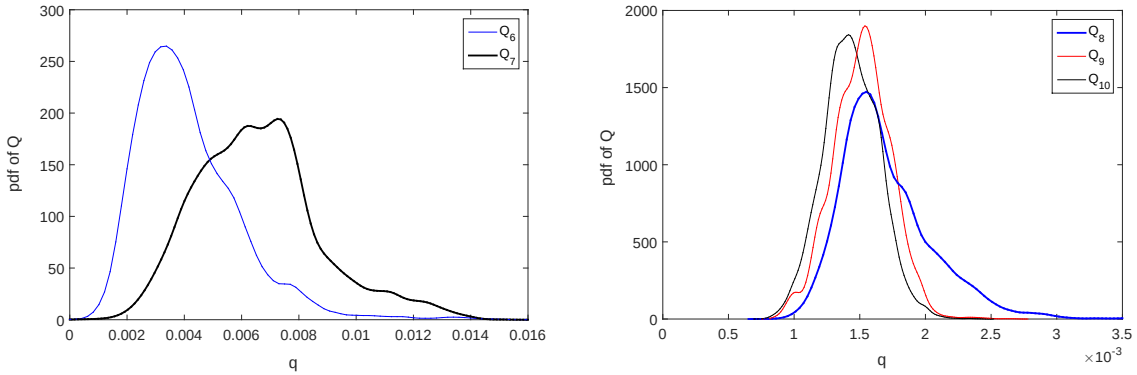

(a) d08 dataset (b) d16 dataset

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



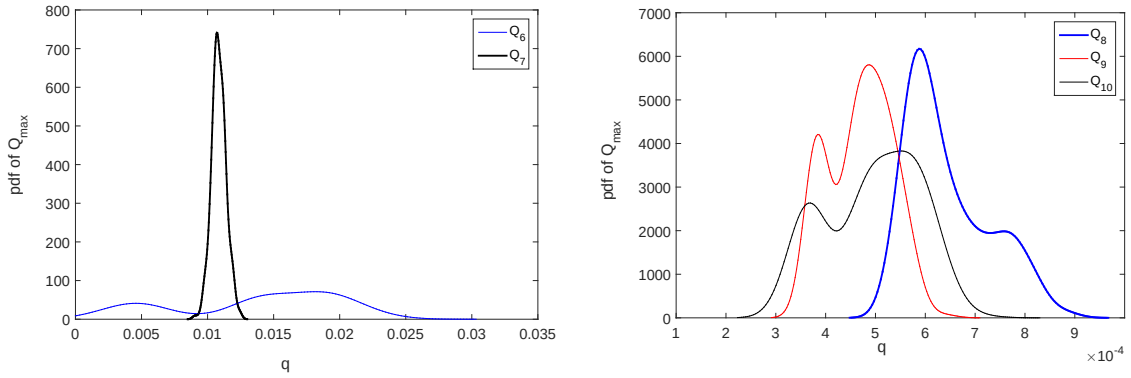

(a) (b)

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



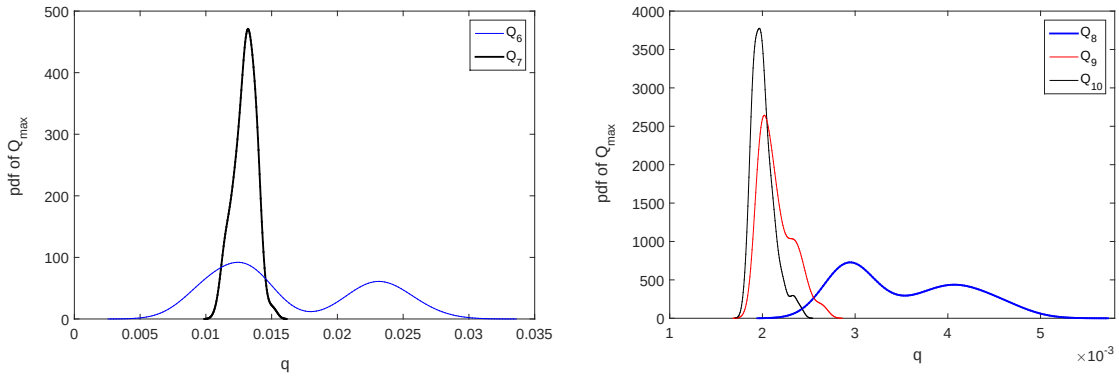

(a) (b)

**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): $\mathbf{Q}_6$ (mid black line) and $\mathbf{Q}_7$ location (thin black line). (b): $\mathbf{Q}_8$ (thick black line), $\mathbf{Q}_9$ (mid red line), and $\mathbf{Q}_{10}$ (thin black line).**




(a) (b)

**Fig. 9 For the d08 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{\mathbf{Q}_{\max}}(q)$ of TKE $\mathbf{Q}_{\max}$. (a): $\mathbf{Q}_6$ (mid black line) and $\mathbf{Q}_7$ (thin black line). (b): $\mathbf{Q}_8$ (thick black line), $\mathbf{Q}_9$ (mid red line), and $\mathbf{Q}_{10}$ (thin black line).**




(a) (b)

**Fig. 10 For the d16 dataset, for $N = 256$ and $\nu_{\text{sim}} = 25,600$: probability density function $p_{\mathbf{Q}_{\text{max}}}(q)$ of TKE $\mathbf{Q}_{\text{max}}$. (a): $\mathbf{Q}_6$ (mid black line) and $\mathbf{Q}_7$ (thin black line). (b): $\mathbf{Q}_8$ (thick black line), $\mathbf{Q}_9$ (mid red line), and $\mathbf{Q}_{10}$ (thin black line).**

### 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*

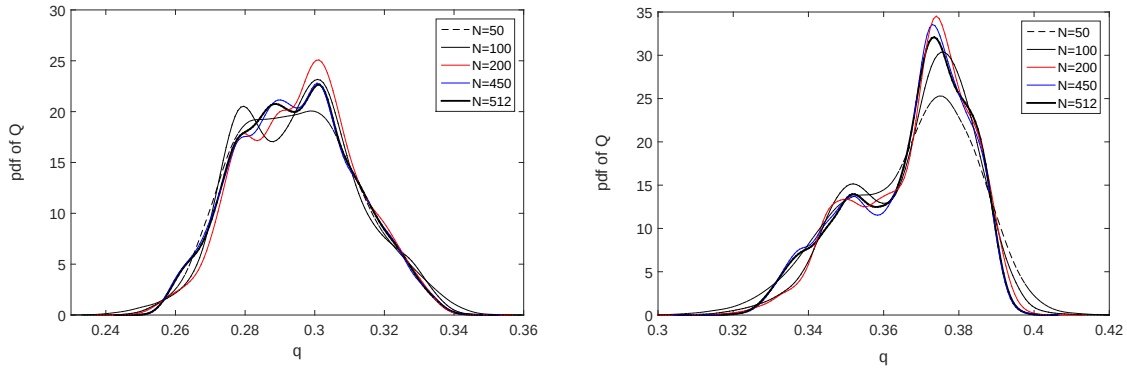
A similar analysis that the one presented in Section VI. B.1, has been performed for the concatenated d08-d16 dataset that is constructed in interlacing the data points of the d08 dataset with the d16 dataset. Therefore, there are $N_{\text{sup}} = 512$ data points in the concatenated d08-d16 dataset. Similarly to Section VI. B.2, for the concatenated d08-d16 dataset and for $\nu_{\text{sim}} = 51,200$, an analysis has been carried out by studying the evolution with respect to $N \leq N_{\text{sup}}$ of the probability density function $p_{\text{Q}}(q)$ of random variable Q for $\text{Q} = \text{Q}_2$ (combustion efficiency, Figure 11(a)) and for $\text{Q} = \text{Q}_3$ (pressure stagnation loss ratio, Figure 11(b)), while Figures 12(a) and (b) display the evolution of the probability density function $p_{\text{Q}_{\text{max}}}(q)$ of random variable $\text{Q}_{\text{max}}$.

*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 $\nu_{\text{sim}} = 51,200$ are good values for studying TKE at various streamwise locations associated with $\text{Q}_6$, $\text{Q}_7$, $\text{Q}_8$, $\text{Q}_9$, and $\text{Q}_{10}$. For the concatenated d08-d16 dataset, Figure 13 displays the probability density function $p_{\text{Q}}(q)$ of TKE associated with $\text{Q}_6$ to $\text{Q}_{10}$, while Figure 14 displays the probability density function $p_{\text{Q}_{\text{max}}}(q)$.

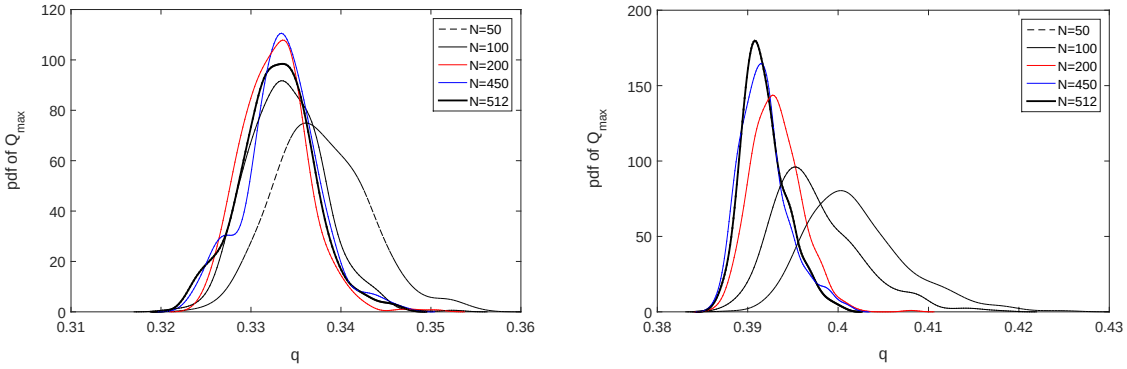
### D. Analysis of the Results Obtained With the Probabilistic Learning

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



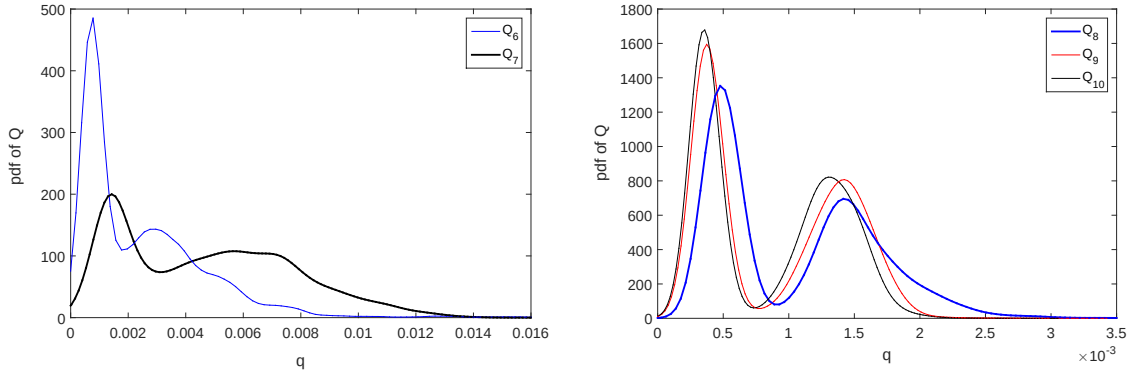

(a) (b)

**Fig. 11 d08-d16 dataset: probability density functions $p_{\mathbf{Q}}(q)$ of random variable Q (a) for combustion efficiency $\mathbf{Q}_2$ and (b) for pressure stagnation loss ratio $\mathbf{Q}_3$, for $N = 50$ (dashed black line), $N = 100$ (thin black line), $N = 200$ (med red line), $N = 450$ (med black line), $N = 512$ (thick black line) with $\nu_{\text{sim}} = 51,200$.**



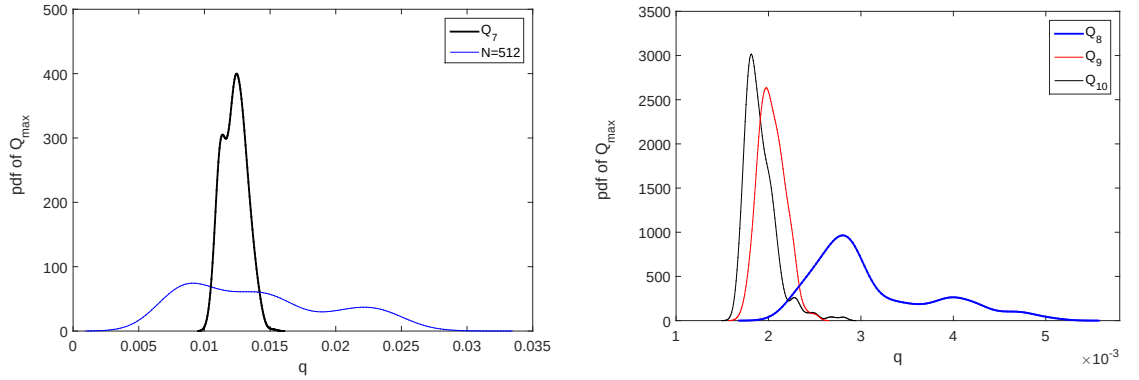

(a) (b)

**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 black 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 Q. (a): $\mathbf{Q}_6$ (mid black line) and $\mathbf{Q}_7$ (thin black line). (b): $\mathbf{Q}_8$ (thick black 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 $\nu_{\text{sim}} = 51,200$: probability density function $p_{\mathbf{Q}_{\max}}(q)$ of TKE $\mathbf{Q}_{\max}$. (a): $\mathbf{Q}_6$ (mid black line) and $\mathbf{Q}_7$ (thin black line). (b): $\mathbf{Q}_8$ (thick black line), $\mathbf{Q}_9$ (mid red line), and $\mathbf{Q}_{10}$ (thin black line).**

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 could 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 chamber, $Q_{10}$, 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, could 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 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 in a probabilistic framework 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 proposing reasonable interpretations of 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.

## Appendix

The objective of this Appendix is to explain in which sense the additional realizations generated by the probabilistic learning on manifold satisfy the computational model.

(i) In order to simplify the explanations, we will assume that $\mathbf{Q} = \mathbf{f}(\mathbf{W})$ in which $\mathbf{f}$ is a measurable mapping from $C_{\mathbf{W}} \subset \mathbb{R}^{m_w}$ into $\mathbb{R}^{n_q}$ (therefore $\mathbf{f}$ is a deterministic mapping). Let $P_{\mathbf{W}}(d\mathbf{w})$ be the probability measure (probability distribution) of $\mathbb{R}^{m_w}$-valued random variable $\mathbf{W}$, defined on $\mathbb{R}^{m_w}$ for which its support is $C_{\mathbf{W}}$. This measure is assumed to be given (known). Let $\mathbf{Q}$ be the $\mathbb{R}^{n_q}$-valued random variable such that $\mathbf{Q} = \mathbf{f}(\mathbf{W})$. Let $P_{\mathbf{Q}}(d\mathbf{q})$ be the probability measure on $\mathbb{R}^{n_q}$ of random variable $\mathbf{Q}$. Then $P_{\mathbf{Q}}(d\mathbf{q})$ is the image of measure $P_{\mathbf{W}}(d\mathbf{w})$ under mapping $\mathbf{f}$.

Let $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ be the joint probability measure on $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ of random variables $\mathbf{W}$ and $\mathbf{Q}$. Therefore, the support $\mathcal{S}$ of $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ is the manifold in $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ defined by $\mathcal{S} = \{(\mathbf{w}, \mathbf{q}) \in C_{\mathbf{W}} \times \mathbb{R}^{n_q}, \mathbf{q} = \mathbf{f}(\mathbf{w})\}$. It can be seen that $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ does not admits a density with respect to the Lebesgue measure on $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$. It is equivalent to give probability measure $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ or to give mapping $\mathbf{f}$ with probability measure $P_{\mathbf{W}}(d\mathbf{w})$ whose support is $C_{\mathbf{W}}$. Consequently, if $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ was known, then $\mathbf{f}$ would be known. This means that if we assumed that $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ was known and if $\{(\mathbf{w}^\ell, \mathbf{q}^\ell), \ell = 1, \ldots, N\}$ were $N$ independent realizations in $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ of random variable $(\mathbf{W}, \mathbf{Q})$ taken from $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$, then the realizations would be such that $\mathbf{q}^\ell = \mathbf{f}(\mathbf{w}^\ell)$ because the support of $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ is manifold $\mathcal{S}$.

(ii) Let us now assume that $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ is unknown. Let $\{(\mathbf{w}^\ell, \mathbf{q}^\ell), \ell = 1, \ldots, N\}$ be $N$ points such that $\{\mathbf{w}^\ell, \ell = 1, \ldots, N\}$ are $N$ independent realizations of $\mathbf{W}$ taken from the known probability measure $P_{\mathbf{W}}(d\mathbf{w})$ and such that $\{\mathbf{q}^\ell = \mathbf{f}(\mathbf{w}^\ell), \ell = 1, \ldots, N\}$ are the $N$ points computed using the computational model. Consequently, the values of $\mathbf{f}$ are known for these $N$ points $\mathbf{w}^\ell$ but are unknown for any point $\mathbf{w}$ in $C_{\mathbf{W}}$ different from that $N$ points. Let $P^{(N)}_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ be a nonparametric statistical estimate of $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ using dataset $\{(\mathbf{w}^\ell, \mathbf{q}^\ell), \ell = 1, \ldots, N\}$ and the multidimensional Gaussian kernel-density estimation method. Consequently, $P^{(N)}_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$ is represented by a

probability density function $(\mathbf{w},\mathbf{q}) \mapsto p_{\mathbf{W},\mathbf{Q}}^{(N)}(\mathbf{w},\mathbf{q})$ on $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$ with respect to the Lebesgue measure $d\mathbf{w}\, d\mathbf{q}$. For $N \to +\infty$, the sequence of probability measures $P_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q}) = p_{\mathbf{W},\mathbf{Q}}^{(N)}(\mathbf{w},\mathbf{q})\, d\mathbf{w}\, d\mathbf{q}$ tends to probability measure $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$. For $N$ finite, the support $\mathcal{S}^{(N)}$ of density $p_{\mathbf{W},\mathbf{Q}}^{(N)}$ is not the manifold $\mathcal{S}$, which means that, in general, mapping $\mathbf{f}$ can only be identified for $N \to +\infty$.

(iii) Let $\{(\mathbf{w}_{\text{ar}}^{\ell}, \mathbf{q}_{\text{ar}}^{\ell}), \ell = 1, \ldots, \nu_{\text{sim}}\}$ be $\nu_{\text{sim}}$ independent realizations taken from the probability measure $\widehat{P}_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ that is constructed with the probabilistic learning on manifolds. This probability measure is deduced from measure $P_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ but is a better estimated than $P_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ because the information is enriched using diffusion maps as demonstrated in [1]. Since the support $\widehat{\mathcal{S}}^{(N)}$ of $\widehat{P}_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ is not $\mathcal{S}$, $\mathbf{q}_{\text{ar}}^{\ell}$ is not equal to $\mathbf{f}(\mathbf{w}_{\text{ar}}^{\ell})$ for all $\ell = 1, \ldots, \nu_{\text{sim}}$, which means that the points $\{(\mathbf{w}_{\text{ar}}^{\ell}, \mathbf{q}_{\text{ar}}^{\ell}), \ell = 1, \ldots, \nu_{\text{sim}}\}$ do not belongs, in general, to $\mathcal{S}$ almost surely. Nevertheless, if $N$ is sufficiently large, the points $\{(\mathbf{w}_{\text{ar}}^{\ell}, \mathbf{q}_{\text{ar}}^{\ell}), \ell = 1, \ldots, \nu_{\text{sim}}\}$ generated by the probabilistic learning on manifolds are concentrated in a subset of $\mathbb{R}^{m_w} \times \mathbb{R}^{n_q}$, localized in the neighborhood of manifold $\mathcal{S}$. In addition, if $N$ goes to infinity, $\widehat{\mathcal{S}}^{(N)}$ goes to $\mathcal{S}$ (due to the convergence of $\widehat{P}_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ towards $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$).

(iv) The probabilistic learning on manifolds is used for enhancing the model predictability in a probabilistic framework. This means that we are not interested in characterizing deterministic mapping $\mathbf{f}$ by generating $\nu_{\text{sim}}$ points $\{(\mathbf{w}_{\text{ar}}^{\ell}, \mathbf{q}_{\text{ar}}^{\ell}), \ell = 1, \ldots, \nu_{\text{sim}}\}$ such that $\mathbf{q}_{\text{ar}}^{\ell} = \mathbf{f}(\mathbf{w}_{\text{ar}}^{\ell})$ for all $\ell = 1, \ldots, \nu_{\text{sim}}$ (that is not possible for a finite value of $N$). We are interested in characterizing the unknown approximation $\mathbf{f}^{(N)}$ of $\mathbf{f}$ by a known probability measure $\widehat{P}_{\mathbf{W},\mathbf{Q}}^{(N)}(d\mathbf{w}, d\mathbf{q})$ that correctly approximate $P_{\mathbf{W},\mathbf{Q}}(d\mathbf{w}, d\mathbf{q})$, and for which the convergence is assured for $N \to +\infty$. For given $N$, the probability density function $\mathbf{q} \mapsto p_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}(\mathbf{q})$ of random QoI $\mathbf{Q}$ is estimated using the multivariate Gaussian kernel-density estimation method with the $\nu_{\text{sim}} \gg N$ additional realizations $\{\mathbf{q}_{\text{ar}}^{\ell}, \ell = 1, \ldots, \nu_{\text{sim}}\}$. This estimate can be obtained with any accuracy because $\nu_{\text{sim}}$ can be chosen arbitrarily large without using the computational model and therefore, the convergence with respect to $\nu_{\text{sim}}$ can be controlled without any problem.

(v) A question is related to the convergence of the sequence of probability measures $P_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}(d\mathbf{q}) = p_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}(\mathbf{q})\, d\mathbf{q}$ towards the unknown measure $P_{\mathbf{Q}}(d\mathbf{q})$ introduced in point (i) before. The construction proposed is such that the convergence is guaranteed, which means that $P_{\mathbf{Q}}(d\mathbf{q}) = \lim_{N\to+\infty, \nu_{\text{sim}}\to+\infty} P_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}(d\mathbf{q})$ in the space of probability measures on $\mathbb{R}^{n_q}$. Because $P_{\mathbf{Q}}(d\mathbf{q})$ is unknown the convergence can be analyzed, for instance, by studying the convergence of the sequences of probability density functions $\{p_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}\}_{N,\nu_{\text{sim}}}$. Let us assumed that for each fixed value of $N$, the value of $\nu_{\text{sim}}(N)$ is identified in order to obtain a given accuracy of the convergence of the sequence of functions $\{p_{\mathbf{Q}}^{(N,\nu_{\text{sim}})}\}_{\nu_{\text{sim}}}$ in the space of integrable functions. The analysis of the convergence of the sequence of probability measures $\{P_{\mathbf{Q}}^{(N,\nu_{\text{sim}}(N))}(d\mathbf{q})\}_N$ with $P_{\mathbf{Q}}^{(N,\nu_{\text{sim}}(N))}(d\mathbf{q}) = p_{\mathbf{Q}}^{(N,\nu_{\text{sim}}(N))}(\mathbf{q})\, d\mathbf{q}$ towards $P_{\mathbf{q}}(d\mathbf{q})$ can be done and corresponds to the convergence of the learning. In practice, if the maximum available value of $N$ is $N_{\text{max}}$, the convergence of the

family of functions $p_{\mathbf{Q}}^{(N,\nu_{\text{sim}}(N))}$ for $N \in [1, N_{\max}]$ is analyzed. If convergence of the learning is not obtained for a value of $N$ smallest than or equal to $N_{\max}$, this means that $N_{\max}$ is not sufficiently large and that additional calculations have to be performed with the computational model.

## 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). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This work was supported in part by a grant of computer time from the DoD High Performance Computing Modernization Program at Navy DSRC. 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-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department Of Energy or the United States Government.

## 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] Mantis, G. C., and Mavris, D. N., "A Bayesian Approach to Non-Deterministic Hypersonic Vehicle Design," *SAE Aerospace Congress and Exhibition*, Technical Paper 2001-01-3033, Seattle, WA, 2001. doi:10.4271/2001-01-3033.

[3] Witteveen, J., Duraisamy, K., and Iaccarino, G., "Uncertainty Quantification and Error Estimation in Scramjet Simulation," *17th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2011-2283, San Francisco, CA, 2011. doi:10.2514/6.2011-2283.

[4] Constantine, P. G., Emory, M., Larsson, J., and Iaccarino, G., "Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet," *Journal of Computational Physics*, Vol. 302, 2015, pp. 1–20. doi:10.1016/j.jcp.2015.09.001.

[5] Geraci, G., Eldred, M. S., and Iaccarino, G., "A multifidelity multilevel Monte Carlo method for uncertainty propagation in aerospace applications," *19th AIAA Non-Deterministic Approaches Conference*, AIAA Paper 2017-1951, Grapevine, TX, 2017. doi:10.2514/6.2017-1951.

[6] Huan, X., Geraci, G., Safta, C., Eldred, M. S., Sargsyan, K., Vane, Z. P., Oefelein, J. C., and Najm, H. N., "Multifidelity Statistical Analysis of Large Eddy Simulations in Scramjet Computations," *20th AIAA Non-Deterministic Approaches Conference*, AIAA Paper 2018-1180, Kissimmee, FL, 2018. doi:10.2514/6.2018-1180.

[7] Huan, X., Safta, C., Sargsyan, K., Geraci, G., Eldred, M. S., Vane, Z. P., Lacaze, G., Oefelein, J. C., and Najm, H. N., "Global Sensitivity Analysis and Estimation of Model Error, Toward Uncertainty Quantification in Scramjet Computations," *AIAA Journal*, Vol. 56, No. 3, 2018, pp. 1170–1184. doi:10.2514/1.J056278.

[8] Feil, M., and Staudacher, S., "Uncertainty quantification of a generic scramjet engine using a probabilistic collocation and a hybrid approach," *CEAS Aeronautical Journal*, 2018. doi:10.1007/s13272-018-0303-6.

[9] Urzay, J., "Supersonic Combustion in Air-Breathing Propulsion Systems for Hypersonic Flight," *Annual Review of Fluid Mechanics*, Vol. 50, No. 1, 2018, pp. 593–627. doi:10.1146/annurev-fluid-122316-045217.

[10] 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.

[11] 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.

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

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

[14] 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.

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

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

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

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Dolvin, D. J., "Hypersonic International Flight Research and Experimentation (HIFiRE)," *15th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2008-2581, Dayton, OH, 2008. doi:10.2514/6.2008-2581.

[38] Dolvin, D. J., "Hypersonic International Flight Research and Experimentation," *16th AIAA/DLR/DGLR International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2009-7228, Bremen, Germany, 2009. doi:10.2514/6.2009-7228.

[39] Jackson, K. R., Gruber, M. R., and Barhorst, T. F., "The HIFiRE Flight 2 Experiment: An Overview and Status Update," *45th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*, AIAA Paper 2009-5029, Denver, CO, 2009. doi:10.2514/6.2009-5029.

[40] Jackson, K. R., Gruber, M. R., and Buccellato, S., "HIFiRE Flight 2 Overview and Status Uptate 2011," *17th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2011-2202, San Francisco, CA, 2011. doi:10.2514/6.2011-2202.

[41] Jackson, K. R., Gruber, M. R., and Buccellato, S., "Mach 6–8+ Hydrocarbon-Fueled Scramjet Flight Experiment: The HIFiRE Flight 2 Project," *Journal of Propulsion and Power*, Vol. 31, No. 1, 2015, pp. 36–53. doi:10.2514/1.B35350.

[42] 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. LF99-8888, NASA, 2010.

[43] 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*, AIAA Paper 2011-2249, San Francisco, CA, 2011. doi:10.2514/6.2011-2249.

[44] Cabell, K. F., Hass, N. E., Storch, A. M., and Gruber, M., "HIFiRE Direct-Connect Rig (HDCR) Phase I Scramjet Test Results from the NASA Langley Arc-Heated Scramjet Test Facility," *17th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2011-2248, San Francisco, CA, 2011. doi:10.2514/6.2011-2248.

[45] 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*, AIAA Paper 2007-5664, Cincinnati, OH, 2007. doi:10.2514/6.2007-5664.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

[53] Jaynes, E. T., "Information theory and statistical mechanics," *Physical Review*, Vol. 106, No. 4, 1957, pp. 620–630. doi: 10.1103/PhysRev.106.620.

[54] Jaynes, E. T., "Information Theory and Statistical Mechanics. II," *Physical Review*, Vol. 108, No. 2, 1957, pp. 171–190. doi:10.1103/PhysRev.108.171.

[55] Gruber, M. R., Jackson, K., and Liu, J., "Hydrocarbon-fueled scramjet combustor flowpath development for Mach 6-8 HIFiRE flight experiments," Tech. Rep. AFRL-RZ-WP-TP-2010-2243, AFRL, 2008.

[56] 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.

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

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

[59] 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.

[60] 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.

[61] 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.

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