Turbulence Modeling in the Age of Data

Karthik Duraisamy[1,⋆], Gianluca Iaccarino[2,⋆], and Heng Xiao[3,⋆]

[1]Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109; kdur@umich.edu
[2]Department of Mechanical Engineering, Stanford University, Stanford, CA 94305; jops@stanford.edu
[3]Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA 24060; hengxiao@vt.edu
⋆ These authors contributed equally to this article and are listed alphabetically

Annual Review of Fluid Mechanics 2019. 51:1–23

https://doi.org/10.1146/annurev-fluid-010518-040547

Keywords

turbulence modeling, statistical inference, machine learning, data-driven modeling, uncertainty quantification

Abstract

Data from experiments and direct simulations of turbulence have historically been used to *calibrate* simple engineering models such as those based on the Reynolds-averaged Navier–Stokes (RANS) equations. In the past few years, with the availability of large and diverse datasets, researchers have begun to explore methods to *systematically inform* turbulence models with data, with the goal of quantifying and reducing model uncertainties. This review surveys recent developments in bounding uncertainties in RANS models via physical constraints, in adopting statistical inference to characterize model coefficients and estimate discrepancy, and in using machine learning to improve turbulence models. Key principles, achievements and challenges are discussed. A central perspective advocated in this review is that by exploiting foundational knowledge in turbulence modeling and physical constraints, data-driven approaches can yield useful predictive models.

1. Introduction

Turbulence is a common physical characteristic of fluid flows. In wind turbine design, the knowledge of the turbulence in the incoming flow and in the blade boundary layers is important for performance; in internal combustion engines, vigorous turbulence increases fuel/air mixing, improving overall efficiency and reducing emissions; in airplane design, delaying the occurrence of turbulence in boundary layers over the wing surfaces leads to reduced fuel consumption. These examples, and a vast number of other applications, demonstrate the importance of determining the effect of turbulence on the performance of engineering devices, and justify the continuous interest in developing techniques to simulate and predict turbulent flows.

The representation of turbulent motions is challenging because of the broad range of active spatial and temporal scales involved and the strong chaotic nature of the phenomenon. Over the past half century, starting from the pioneering theoretical studies of Prandtl, Kolmogorov, and von Karman, (Darrigol 2005) many theoretical and computational approaches have been introduced to characterize turbulence. The continuous growth of computer power has enabled direct numerical simulations (DNS) of a number of turbulent flows and processes involving the physics of turbulence (Kim et al. 1987; Rastegari and Akhavan 2018). However, simplified engineering approximations continue to remain popular and widespread across different industries. Among these, RANS (Reynolds Averaged Navier-Stokes) and LES (Large Eddy Simulation) approaches are the most common, although there exist many alternatives (Girimaji 2006; Spalart 2009). RANS techniques rely completely on modeling assumptions to represent turbulent characteristics and, therefore, lead to considerably lower computational requirements than DNS. RANS models are constructed using a formal averaging procedure applied to the exact governing equations of motion and require *closures* to represent the turbulent stresses and scalar fluxes emerging from the averaging process. The discipline of turbulence modeling has evolved using a combination of intuition, asymptotic theories and empiricism, while constrained by practical needs such as numerical stability and computational efficiency. Single-point RANS models of turbulence, that are the focus of this review, are by far the most popular methods. These models implicitly assume an equilibrium spectrum and locally-defined constitutive relationships to *close* the averaged governing equations and express unclosed terms as a function of averaged, local flow quantities.

The title of this article is inspired by the recent book of Efron and Hastie (2016). It reflects the belief that recent advances in data sciences are offering new perspectives to the classical field of turbulence modeling.

LES techniques, on the other hand, directly represent a portion of the active scales and only require modeling to account for the unresolved turbulent motions. LES is gaining popularity in many industrial applications characterized by relatively small Reynolds numbers. As LES also involves modeling assumptions, some of the ideas outlined in this manuscript regarding RANS are amenable for use in LES (Gamahara and Hattori 2017; Jofre et al. 2018).

The inherent assumptions in the RANS approach and the process of formulating closure models introduce potential accuracy limitations and, consequently, reduced credibility in its predictive ability. Direct quantification of the errors introduced by closure models is intractable in general, but formal uncertainty quantification techniques have recently enabled the interpretation of RANS predictions in probabilistic terms, while characterizing the corresponding confidence levels. Experimental observations have routinely been used to *calibrate* the closures and attempt to improve the accuracy of the resulting computations. Statistical inference approaches enable a more comprehensive *fusion* of data and models, resulting in improved predictions. Furthermore, the introduction of modern machine learning

strategies brings fresh perspectives to the classic problem of turbulence modeling.

Lexicon of data-driven modeling

The elements of a data-driven model are:

- $\mathcal{M}$: the (computational) *model*, which is typically a function of an array of independent variables $\mathbf{w}$ and based on a set of algebraic or differential operators $\mathcal{P}$; the model includes a set of parameters $\mathbf{c}$ that are the *primary* target of the data modeling;
- $\boldsymbol{\theta}$: the *data*, which in general is accompanied by a quality estimate, or in other words, the uncertainty $\boldsymbol{\epsilon_\theta}$;
- $\boldsymbol{o}$: the *output* of the model corresponding to the data $\boldsymbol{\theta}$;
- $\boldsymbol{\delta}$: the *discrepancy* which describes the ability of the model to *represent* the data. In general, $\boldsymbol{\delta}$ is a function of the model and it is unknown; it is typically described in terms of $\boldsymbol{\theta}$ and a set of *features* $\boldsymbol{\eta}$ that are derived from prior knowledge, constraints or directly from data.

A general data driven model is written as

$$\widetilde{\mathcal{M}} \equiv \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \mathbf{c}; \boldsymbol{\theta}; \boldsymbol{\delta}; \boldsymbol{\epsilon_\theta})$$

and one is generally interested in predicting quantities of interest $\boldsymbol{q}(\widetilde{\mathcal{M}})$.

2. Turbulence closures and uncertainties

Assumptions are introduced in several stages whilst constructing a Reynolds-averaged model. Although fully justified under certain conditions, these hypotheses introduce potential inadequacies that limit the credibility of the overall predictions if not properly quantified. In this section, we illustrate the four layers of simplifications that are typically required to formulate a RANS closure.

L1: Uncertainties introduced by ensemble-averaging, which are fundamentally irrecoverable.

L1: The application of time- or ensemble-averaging operators $\langle\cdot\rangle$ combined with the nonlinearity of the Navier-Stokes equations (indicated hereafter as $\mathcal{N}(\cdot) = 0$) leads to an undetermined system of equations, that requires the introduction of modeling assumptions to *close* the system.

$$\langle\mathcal{N}(\cdot)\rangle \neq \mathcal{N}(\langle\cdot\rangle) \qquad 1.$$

At a given instant in time, there are infinite realizations of velocity fields (microscopic state) that are compatible with an averaged field (macroscopic state); however each of these realizations might evolve dynamically in different ways, leading to hysteresis-like phenomena and thus uncertainty. This inadequacy is unavoidable in RANS, due to the loss of information in the averaging process and is fundamentally irrecoverable regardless of the sophistication of the turbulence model. This situation is not limited to turbulence modeling – and in general terms, is referred to as *upscaling* or *coarse graining* (Rudd and Broughton 1998).

L2: Uncertainties in the functional and operational representation of Reynolds stress

L2: In the process of developing closures, a model representation is invoked to relate the macroscopic state to the microscopic state and formally remove the unknowns resulting from the averaging process.

$$\langle \mathcal{N}(\cdot) \rangle = \mathcal{N}(\langle \cdot \rangle) + \mathcal{M}(\cdot). \tag{2.}$$

For an incompressible fluid, the unclosed term is simply written as:

$$\mathcal{M}(\cdot) = \nabla \cdot \boldsymbol{\tau}, \tag{3.}$$

where $\boldsymbol{\tau}$ is the Reynolds stress tensor.

$\mathcal{M}$ is written in terms of a set of independent, averaged variables $\mathbf{w}$, defined either locally or globally, leading to one-point or two-point closures. For instance, with the assumption that the Reynolds stress tensor is only a function of the local, averaged velocity gradient tensor, the Cayley-Hamilton theorem can be applied to derive an exact expansion basis (Gatski and Jongen 2000). Linear eddy viscosity models and algebraic stress models are examples of L2-level assumptions.

L3: Uncertainties in functional forms within a model

L3: Once the independent variables are selected, a specific functional form is postulated. Either algebraic or differential equations, denoted here as $\mathcal{P}(\cdot)$, are typically used to represent physical processes or specific assumptions. Schematically, the model is now:

$$\mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w})). \tag{4.}$$

One and two-equation models are the most popular, although many different choices of the independent variables exist in literature (Wilcox 2006). Often $\mathcal{P}(\cdot)$ mimic the terms in the Navier-Stokes equations such as convection and diffusion; additional contributions and source terms are often included to represent known sensitivities, such as near wall dynamics, rotational corrections, etc. (Durbin 2017). Although it is possible to derive differential models formally through repeated applications of the Navier-Stokes operator ($\mathcal{N}(\boldsymbol{\tau}) = \cdot$), this leads to computationally intensive closures and numerically cumbersome and non-interpretable terms.

L4: Uncertainties in the coefficients within a model

L4: Finally, given a complete model structure and functional form, a set of coefficients $\mathbf{c}$ must be specified to calibrate the relative importance of the various contributions in the closure. Formally the closure is then:

$$\mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \mathbf{c}). \tag{5.}$$

It is common to use consistency between the closure and known asymptotic turbulence states to define some of the coefficients, although in many cases empirical evidence is more effective in producing realistic models (Durbin 2017). The choice of the C_μ coefficient in two-equation linear eddy viscosity models is a classical L4 closure issue.

A RANS prediction of a quantity of interest $\mathbf{q}$ is then in general

$$\mathbf{q} = \mathbf{q}\left(\mathcal{N}(\langle \cdot \rangle); \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \mathbf{c})\right) \tag{6.}$$

Together, these four modeling layers showcase the difficulty in assessing the true predictive nature of a turbulence closure, the inconsistency inherent in comparing different strategies, and the need for a careful and transparent process for quantifying model inadequacies and reducing them.

Uncertainty Quantification

In spite of the considerable popularity of simulations in science and engineering, the process of generating objective confidence levels in numerical predictions remains a challenge. The complexity arises from (a) the imprecision or natural variability in the inputs to any simulation of a real-world system (*aleatory* uncertainties), and (b) the limitations intrinsic in the physics models (*epistemic* and *model-form* uncertainties). Uncertainty Quantification (UQ) aims to rigorously measure and rank the effect of these uncertainties on prediction outputs.

The first step in UQ is the identification of the sources of uncertainty and the introduction of an appropriate description (typically in probabilistic terms) $\boldsymbol{\epsilon}$; this is then *propagated* through the model $\mathcal{M}$, resulting in predictions of a quantity of interest $\mathbf{q}(\mathcal{M}, \boldsymbol{\epsilon})$. The propagation step is typically computationally intensive and has received considerable attention in the last decade, leading to the development of extremely efficient UQ strategies. The result is a prediction that explicitly represents the impact of the uncertainty; if $\boldsymbol{\epsilon}$ is a stochastic quantity, the resulting prediction is the probability distribution $\mathbb{P}(\mathbf{q})$ and, therefore, a rigorous measure of the confidence interval can be extracted from the analysis. In some cases, only statistical moments of $\mathbf{q}$ are required, leading to more cost-effective UQ propagation strategies. Alternative descriptions of the uncertainties are also possible. For instance, when very limited observations are available to represent a specific uncertainty source, it is appropriate to introduce a range (an *interval* $[\boldsymbol{\epsilon}^-; \boldsymbol{\epsilon}^+]$) and consequently seek an interval on the model predictions $[\mathbf{q}^-; \mathbf{q}^+]$. In general, for non-linear models $\mathbf{q}^\pm \neq \mathbf{q}(\mathcal{M}, \boldsymbol{\epsilon}^\pm)$. In such cases, optimization techniques and bounding strategies are used instead of probabilistic approaches.

3. Models, data and calibration

The use of experimental observations to drive physical insight is a staple of the scientific method. The understanding of turbulent flows has benefited considerably from detailed measurement campaigns such as the famous isotropic turbulence experiments of Comte-Bellot and Corrsin (1966). The measured turbulence decay rates have been used to constrain the value of the (L4) constants $\mathbf{c}$ defined earlier.

In the last four decades, in addition to experimental datasets, direct simulations of the Navier-Stokes equations (DNS) have provided a further, invaluable source of data to gather modeling insights. The Summer Program of the Center for Turbulence Research at Stanford University had been established with the goal of *studying turbulence using numerical simulation databases* in 1987. There has been a concerted effort by the turbulence community to gather and archive data sets including the databases in US (Li et al. 2008) and Europe (Coupland 1993), among others. Until the past decade however, experimental and simulation data have been used mostly to obtain modeling insight and to aid in validation. Recently, data has been used towards the end of systematically informing turbulence models with the goal of quantifying and reducing model uncertainties.

In this section we describe how data is used in building new, calibrated models $\widetilde{\mathcal{M}} \neq \mathcal{M}$.

3.1. Naive calibration

The simplest calibration process typically involves the selection of an experimental configuration that is similar to that of the prediction target. Often, the measured data may be the same as the quantity of interest $\mathbf{q}$, but available in different flow conditions or configurations. Uncertainties in the measurements are typically ignored. Finally it is assumed that the model coefficients ($\mathbf{c}$) are the dominant source of uncertainty in the model and, therefore, the calibrated model is:

$$\widetilde{\mathcal{M}} = \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \tilde{\mathbf{c}}_{\mathbf{q}}), \tag{7.}$$

and the prediction accuracy is judged by the difference in $\mathbf{q}$ obtained when using $\mathbf{c}$ or $\tilde{\mathbf{c}}_{\mathbf{q}}$. This process has led to the proliferation of turbulence model variants and the inherent difficulty in assessing predictive capabilities.

3.2. Statistical inference

Statistical inference is the generalization of the calibration process described above; specifically, uncertainty in the experiments can be directly accounted for and a potential discrepancy (misfit) between the model prediction $\boldsymbol{\delta}$ and the data is also included. Furthermore, the calibration data can include evidence from different sources while the objective is simply to represent the data. The inference is formulated in a probabilistic setting, inspired by the Bayes theorem and the result is a calibrated, stochastic model:

$$\widetilde{\mathcal{M}} = \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \tilde{\mathbf{c}}_{\boldsymbol{\theta}}) + \boldsymbol{\delta} + \boldsymbol{\epsilon}_{\boldsymbol{\theta}}. \tag{8.}$$

Formally, stochasticity is a consequence of uncertainty in the measurements, the prior information on the calibration parameters (for example, the range or the most likely values of $\mathbf{c}$), and the discrepancy function. A prior for the discrepancy function is typically left to the intuition of the modeler and is typically represented in a simple mathematical form, for example using a Gaussian random field with parameters that are also estimated through the calibration process, i.e. $\boldsymbol{\delta}(\boldsymbol{\theta})$. The inference problem is typically solved using Markov-Chain Monte Carlo (MCMC) strategies, and the result is a stochastic description (the posterior probability) of the model $\widetilde{\mathcal{M}}$.

The resulting model prediction $\mathbf{q}(\widetilde{\mathcal{M}})$ is a random quantity. This approach is referred to in the literature as a Bayesian inversion strategy. In many cases, only the *mode* of the the posterior probability is used; this is referred to as the maximum a posteriori (MAP) estimate and therefore the corresponding calibrated model is deterministic.

3.3. Data–driven modeling

In the last two decades, the introduction of computationally efficient statistical inference algorithms has led to the possibility of assimilating large amounts of data (for example, those generated by DNS simulations). This has spurred interest in approaches that rely more on the available data than on traditional models; in other words in Eq. 8, the emphasis is on $\boldsymbol{\delta}$ rather than on $\mathcal{M}$. Different choices for the functional representation of $\boldsymbol{\delta}$ are available, with increasing focus on machine-learning strategies. In addition to the inference, further work has been devoted towards representing the discrepancy $\boldsymbol{\delta}$ in terms of features $\boldsymbol{\eta}$ selected from a potentially large set of candidates. This enables representation of the resulting model in terms of quantities, such as the mean velocity gradients, that are likely

Statistical inversion

Statistical inversion aims to identify parameters $\boldsymbol{c}$ of a model $\mathcal{M}(\boldsymbol{c})$ given data $\boldsymbol{\theta}$ with uncertainty $\boldsymbol{\epsilon_\theta}$. Mathematically, the solution is determined by minimizing the difference between $\boldsymbol{\theta}$ and the corresponding model output $\boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))$. In a Bayesian framework, the result of the inversion, i.e. the *posterior* probability of $\boldsymbol{c}$ given $\boldsymbol{\theta}$, is given as

$$\mathbb{P}[\boldsymbol{c}|\boldsymbol{\theta}] \propto \mathbb{P}[\boldsymbol{\theta}|\boldsymbol{c}] \times \mathbb{P}[\boldsymbol{c}],$$

where $\mathbb{P}[\boldsymbol{c}]$ is the *prior*, i.e. the probability of the model before any data (evidence) is used, and $\mathbb{P}[\boldsymbol{\theta}|\boldsymbol{c}]$ is the probability of the model being consistent with the data, referred to as the *likelihood*.

The prior and the posterior represent the modeler's belief on the probability of $\mathcal{M}$, before and after observing the data. In the case that all underlying probability distributions are assumed to be Gaussian, it can be shown (Aster et al. 2011) that the *maximum a posteriori* (MAP) estimate of $\boldsymbol{c}$ can be determined by solving a deterministic optimization problem

$$\boldsymbol{c}_{\text{MAP}} = \arg\min \frac{1}{2}\left[\big(\boldsymbol{\theta} - \boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))\big)^T \boldsymbol{Q}_{\boldsymbol{\theta}}^{-1}\big(\boldsymbol{\theta} - \boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))\big) + \big(\boldsymbol{c} - \boldsymbol{c}_{prior}\big)^T \boldsymbol{Q}_{\boldsymbol{c}}^{-1}\big(\boldsymbol{c} - \boldsymbol{c}_{prior}\big)\right],$$

where $\boldsymbol{Q_\theta}$ and $\boldsymbol{Q_c}$ are the observation and prior covariance matrices, respectively.

An alternative strategy to solve inverse problems is the Least Squares (LS) approach. The LS procedure involves the minimization of the discrepancy between $\boldsymbol{\theta}$ and the model output $\boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))$ by solving the optimization problem

$$\boldsymbol{c}_{\text{LS}} = \arg\min ||\boldsymbol{\theta} - \boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))||_2^2 + \gamma||\boldsymbol{c}||_2,$$

where the second contribution scaled by γ is a regularization term included to improve well-posedness and conditioning of the inversion process (γ is a user-specified parameter). Again, assuming that all distributions are Gaussian (and $\gamma \equiv \boldsymbol{Q_\theta Q_c^{-1}}$), $\boldsymbol{c}_{\text{LS}} = \boldsymbol{c}_{\text{MAP}}$.

to be *descriptive* in a more general context than the one characterized by the available data. Furthermore, constrains such as symmetry properties or Galilean invariance can be enforced in the definition of the candidate features.

In general, data–driven models can then be expressed as:

$$\widetilde{\mathcal{M}} = \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}); \mathbf{c}(\boldsymbol{\theta}); \boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{\eta}); \boldsymbol{\epsilon_\theta}). \tag{9.}$$

3.4. Calibration and prediction

The objective of the calibration process is the definition of a model that incorporates evidence from available data. In practical applications, the next step is to use the calibrated model to predict a quantity of interest. In the seminal work of Kennedy and O'Hagan (2001), model-form uncertainty is introduced to the prediction by adding a discrepancy term to the model output $\boldsymbol{o}(\mathcal{M}(\boldsymbol{c}))$. Typically, Gaussian Process models are assumed for the model discrepancy $\boldsymbol{\delta}$, and Bayesian inversion is used to derive a posterior distribution for the hyperparameters of the Gaussian process as well as the inferred model parameters. However, in this approach, the the entire mapping between the input and prediction is

treated as a black-box and thus physics-agnostic. In the present review, we focus on methods that embed the calibration inside the model. The propagation of the relevant stochastic information that defines the calibrated model endows the prediction with uncertainty that represents the ability of the calibrated model to represent the data.

4. Quantifying uncertainties in RANS models

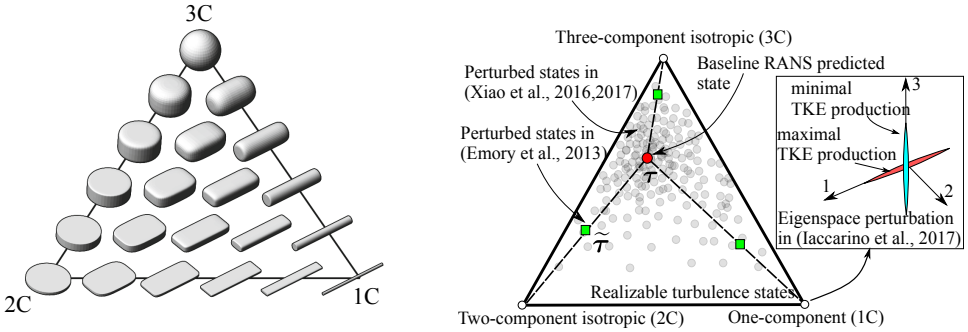
Predictions based on RANS models are affected by the assumptions invoked in the construction of the closure (L1–L4) and by the calibration process. A rigorous process is required to characterize the potential impact of these sources of uncertainties and the resulting confidence in the predictions. In this section, we review approaches that seek to derive $\widetilde{\mathcal{M}} = \mathcal{M} + \boldsymbol{\epsilon}_{\mathcal{M}}$ where $\boldsymbol{\epsilon}_{\mathcal{M}}$ is either obtained from theoretical arguments or $\boldsymbol{\epsilon}_{\mathcal{M}} = \boldsymbol{\epsilon}(\boldsymbol{\theta})$ from comparisons to existing data.

4.1. Uncertainties in the Reynolds stress tensor

We survey two strategies to characterize the uncertainty in the Reynolds stresses based on an interval and a probabilistic description of the uncertainty, respectively.

RANS modeling has traditionally aimed to approximate unclosed terms in the averaged equations with the goal of obtaining a *closed*, computable model. An alternative idea is to replace the unclosed terms with *bounds* that are based on theoretical arguments, leading to predictions that represent extreme but possible behaviors, rather than likely behaviors under specific assumptions. In other words, bounding aims to construct prediction intervals that can be proven to *contain* the true answers, as opposed to explicit estimates that might be inaccurate. In the 70s, pioneering work in turbulence analysis by Howard (1972), Busse (1970) and others, and more recently the work by Doering and coworkers (Doering and Constantin 1994), explored this idea; these authors applied variational approaches and formulated a generic framework for estimating bounds on physical quantities rigorously and directly from the equations of motion. Following this approach, named the *background flow method*, one manipulates the equations of motion relative to a steady trial background state and then decomposes the quantity of interest, for example the energy dissipation rate, into a background profile and a fluctuating component. If the fluctuation term satisfies a non-negativity condition, the background part admits an upper bound. A recent improvement of this approach (Seis 2015) led to quantified bounds on the energy dissipation rate for shear flow, channel flow, Rayleigh–Benard convection and porous medium convection. In spite of these encouraging results, it is difficult to envision that formal bounds can be derived for flow problems of practical engineering interest.

An alternative viewpoint on bounding is based on the concept of Reynolds stress realizability (Durbin and Speziale 1994; Lumley 1978; Pope 1985; Schumann 1977). Emory et al. (2011, 2013) proposed a scheme of introducing realizable, physics-constrained perturbations to the Reynolds stress tensor as a way of quantifying (L2) uncertainties in RANS models expressed as intervals. The starting point is the eigen-decomposition of the Reynolds stresses. Banerjee et al. (2007) defined a mapping based on the eigenvalues λ_1, λ_2, and λ_3 and corresponding Barycentric coordinates that enables a convenient visual representation of the realizability bounds in a two-dimensional coordinate system, depicted in Fig. 1.



(a) Turbulence componentality in Barycentric triangle (Jofre et al. 2018)

(b) Realizability-constrained perturbations

Figure 1: Geometric representation of the realizability constraint on a single-point Reynolds stress tensor.

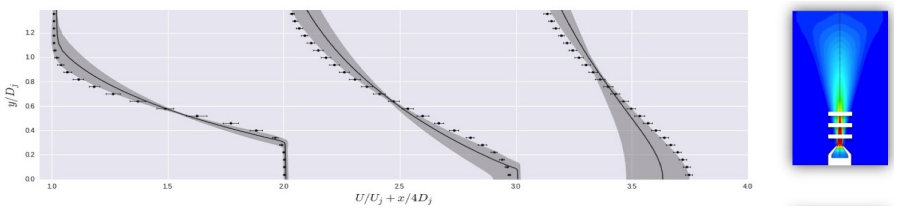
Formally a general expression for the Reynolds stresses can be written as:

$$\widetilde{\boldsymbol{\tau}} = \boldsymbol{\tau}^{\mathrm{RANS}} + \boldsymbol{\delta}_{\boldsymbol{\tau}} = 2\widetilde{k}\left(\frac{1}{3}\mathbf{I} + \widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{\top}\right). \qquad 10.$$

where $\widetilde{k}$, $\widetilde{\boldsymbol{\Lambda}}$, and $\widetilde{\boldsymbol{V}}$ are the perturbed counterparts to Reynolds stresses computed using a RANS model. For example, for the eigenvalues $\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^{\mathrm{RANS}} + \boldsymbol{\delta}_{\lambda}$.

With the objective of determining general bounds, the discrepancy $\boldsymbol{\delta}_{\boldsymbol{\tau}} = (\boldsymbol{\delta}_{\lambda}, \boldsymbol{\delta}_{k}, \boldsymbol{\delta}_{\mathbf{V}})$ has to be defined with appropriate physical constraints and without the use of calibration data. The reliazability condition of the Reynolds stress depicted in Fig. 1 provides clear constraints on the eigenvalues. In Emory et al.'s approaches, RANS predicted Reynolds stresses were perturbed towards three representative limiting states: one-component (1C), two-component (2C), and three-component (3C) turbulence, indicated by the vertices of the Barycentric triangle. In contrast to the strong constraint imposed by the realizability on the eigenvalues, the constraint on the turbulent kinetic energy is rather weak – it only has to be pointwise non-negative. Furthermore, the realizability condition does not give clear bounds on the eigenvectors. Iaccarino et al. (2017) used the two limiting states at which the turbulence kinetic energy (TKE) production $P = \tau_{ij}\partial U_i/\partial x_j$ achieves maximum and minimum values. These states are identified by specific alignments between Reynolds stress tensor τ_{ij} and mean velocity gradient tensor $\partial U_i/\partial x_j$. A prediction with a resulting interval-based uncertainty is obtained by performing simulations corresponding to the extreme componentality states, and the two set of eigenvectors providing minimum and maximum kinetic energy production. The computed bounds $[\mathbf{q}^{-}; \mathbf{q}^{+}]$ appear to provide a satisfactory representation on the uncertainty in the model as reported in Fig. 2 for a simple turbulent jet.

As an alternative to the physical approach based on eigen-decomposition, Xiao et al. (2017) pursued a probabilistic description of the Reynolds stress uncertainty. They modeled the true stress as a random matrix $\boldsymbol{T}$ defined on the set of symmetric positive semi-definite 3×3 matrices. The expectation of random matrix $\boldsymbol{T}$ is specified to be the RANS modeled Reynolds stress $\boldsymbol{\tau}^{\mathrm{RANS}}$, i.e. $\mathbb{E}[\boldsymbol{T}] = \boldsymbol{\tau}^{\mathrm{RANS}}$. They further defined a maximum entropy



.

Figure 2: Prediction bounds for the velocity profiles in a turbulent jet computed using the eigenspace perturbations (Mishra and Iaccarino 2017). The line represent the prediction using an eddy viscosity model, the symbols are the experimental measurements and the grey areas are the computed uncertainty bounds.

distribution for the Reynolds stress tensor, which is sampled to indicate the uncertainty of the Reynolds stresses.

The random matrix approach and the eigen-decomposition based approach are similar in the sense that they ensure realizability when perturbing the Reynolds stress tensor or sampling from the distribution thereof. The random matrix approach explores the uncertainty space of eigenvalues and eigenvector simultaneously (Wang et al. 2016a), with the correlation among them implicitly specified through the maximum entropy distribution defined on random matrix $\boldsymbol{T}$; it lacks, however, the clear interpretations of the limiting states as in the physics-based, eigen-decomposition approach.

Both approaches have focused on the error bound of the Reynolds stress at a *single point*. A potentially important source of uncertainty comes from the spatial variation of the Reynolds stress discrepancy as the *divergence* of the Reynolds stress field appears in the RANS equations; in other words, it is likely that $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{x})$. Emory et al. (2013) and Gorlé et al. (2014) specified a spatial field for the eigenvalue perturbations based on the assumed limitations of RANS models, while Wang et al. (2016b) and Xiao et al. (2017) used a non-stationary Gaussian process to encode the same empirical knowledge. Edeling et al. (2017) proposed a "return-to-eddy-viscosity model", which is a transport equation with a source term describing the departure of turbulence state from local equilibrium. Finally, Wu et al. (2018a) utilized the fundamental connection between the governing partial differential equations and the covariance of a discrepancy field to derive a physically consistent covariance structure. Xiao and Cinnella (2018) provided more detailed discussions on some of the approaches above.

4.2. Uncertainty in model parameters

The model parameters in turbulence models are often determined enforcing consistency in the prediction of fundamental flows (e.g., homogeneous isotropic turbulence, logarithmic layer). It is well known that these parameters are not universal and might require flow-specific adjustments. For example, Pope (2000) and Eisfeld (2017) list optimal parameters for several typical free shear flows (e.g., plane jet, round jet, wake). However, for the lack of better alternatives, the default parameters determined from the fundamental flows are still used in the simulation of complex turbulent flows: this lead to uncertainties (L4).

It is fairly straightforward to assess the impact of uncertainties in the choice of the coefficients in the models by using classical uncertainty propagation techniques. However, this exercise is fundamentally dependent on the choices made to describe the parameters, i.e. to define their range or their probabilistic representation. A more effective approach is to use data to infer the parameters and then propagate the resulting stochastic description through the simulations obtaining predictions with uncertainty intervals. This will be discussed in the next section because it is akin to a data-informed approach in which $\widetilde{\mathcal{M}} = \mathcal{M}(\boldsymbol{\theta}) + \boldsymbol{\epsilon_\theta}$.

4.3. Identifying regions of uncertainty

While the above discussion was focused on *quantifying* uncertainties, techniques have been developed to *identify* regions of potentially high uncertainties in RANS predictions. Gorlé et al. (2014) developed an analytical marker function to identify regions that deviate from parallel shear flow. This marker was found to correlate well with regions where the prediction of the Reynolds stress divergence was inaccurate. Ling and Templeton (2015) employed databases of DNS and RANS solutions and formulated the evaluation of potential adequacy of the RANS model as a *classification* problem in machine learning. The results include several fields of binary labels that indicate whether the specified model assumption is violated. Although these studies are useful to illustrate the failure of turbulence models there is no straightforward way to use the results for improving predictions.

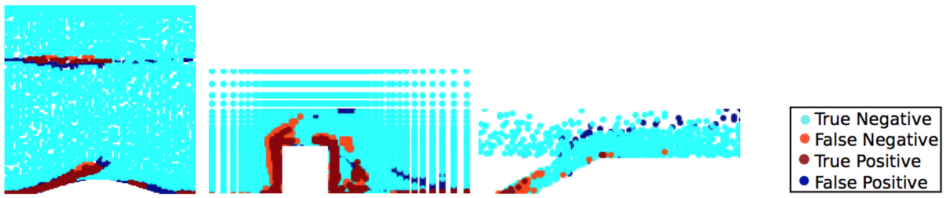


Figure 3: Random forest predicted markers (blue and maroon) showing regions with possible emerging of negative eddy viscosity in several flows, indicating likely failure of linear eddy viscosity models in such regions (Ling and Templeton 2015).

5. Predictive modeling using data-driven techniques

The efforts discussed in the previous section aim at providing confidence in the application of RANS closures by identifying and quantifying uncertainties in the models at various levels. In this section, we focus on approaches that attempts to improve the overall prediction accuracy by using data. In this section, we review strategies that seek to derive $\widetilde{\mathcal{M}} = \mathcal{M}(\boldsymbol{\theta})$, while explicitly introducing a discrepancy function $\boldsymbol{\delta}$.

5.1. Embedding inference-based discrepancy

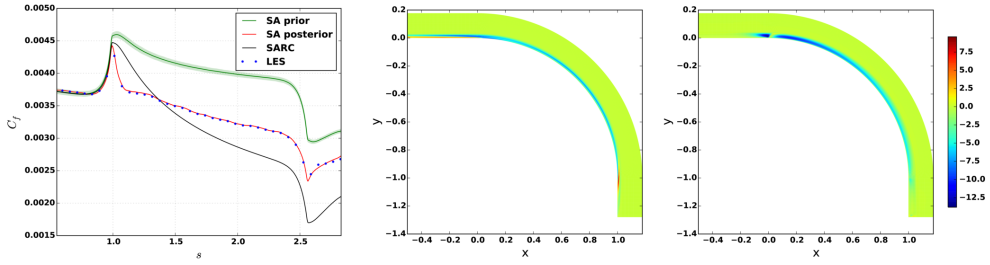
The use of rigorous statistical inference to determine model coefficients is only a relatively recent development. Oliver and Moser (2011) and Cheung et al. (2011) were the first to leverage DNS data from plane channel flows and Bayesian inference to assign posterior probability distributions to model parameters of several turbulence models. Edeling et al.

(2014a,b) further introduced the concept of Bayesian model scenario averaging in the calibration of model parameters to assess the effectiveness of diverse source of data in specific predictions. Their studies included data from a number of wall-bounded flows. Lefantzi et al. (2015) and Ray et al. (2016) used a similar approach to infer the RANS model coefficients and also investigated the likelihood of competing closures while focusing on jet-in-cross-flow, which is a canonical flow in film cooling in turbo-machinery applications. More recently, Ray et al. (2018) used experimental data and Bayesian inference to calibrate the parameters in a nonlinear eddy viscosity model.

The approaches listed above use statistical inference to construct posterior probability distribution of a quantity of interest based on data. Only calibration of the model coefficients is considered (an L4 uncertainty); typically a simple discrepancy function is defined in the process and often not directly used to make predictions. In contrast, Oliver and Moser (2009) introduced a Reynolds stress discrepancy tensor $\boldsymbol{\delta_\tau}$ to account for the uncertainty. $\boldsymbol{\delta_\tau}$ is a random field described by stochastic differential equations, which are structurally similar to, but simpler than the Reynolds stress transport equations (e.g., Launder et al. 1975). They demonstrated preliminary successes of their approach in plane channel flows at various Reynolds numbers. Their framework laid the foundation for many subsequent works in quantifying and reducing RANS model form uncertainties.

Dow and Wang (2011) used full-field DNS velocity data to infer the turbulent viscosity field in a plane channel, based on which they built Gaussian process models for the discrepancy field. The resulting stochastic turbulent viscosity field was then sampled to make predictions of the velocity field. Duraisamy et al. (2015); Singh and Duraisamy (2016) used limited measurements (such as surface pressures, skin friction) to extract the discrepancy field and applied it to channel flow, shock-boundary layer interactions, and flows with curvature and separation. Xiao et al. (2016) and Wu et al. (2016) used sparse velocity field data to infer the structure of the Reynolds stress magnitude and anisotropy. All these approaches involve large-scale statistical inference using adjoint-based algorithms (Giles et al. 2003) or derivative-free, iterative Ensemble Kalman method (Iglesias et al. 2013).

As a representative example of large-scale inversion, the methodology is illustrated (Singh and Duraisamy 2016) for the flow over a curved channel in Figure 4. In this case, the discrepancy function is defined as a multiplier to the production term of the Spalart-Allmaras (SA) turbulence model (Spalart and Allmaras 1992) and extracted by applying statical inversion using the skin friction C_f data obtained via LES on the lower (convex) wall. Figure 4 shows the baseline SA model (prior) and posterior (MAP) outputs alongside LES. In addition, a model which includes an analytically-defined correction to the production term of the SA model, namely the SARC model (Shur et al. 2000) is also reported. Also shown is the inferred correction term, $\boldsymbol{\delta}_{MAP}$, and the analytically-defined correction term from the corresponding SARC model. The trend in $\boldsymbol{\delta}_{MAP}$ is consistent with the expectation that the convex curvature reduces the turbulence intensity. Figure 5 shows the variation of the streamwise velocity with respect to the distance from the wall at various streamwise locations. The posterior velocity and Reynolds stresses (not shown) were observed to compare well with the LES counterparts and considerably improved compared to the SA and the SARC prediction, even though only the skin friction data was used in the inference. The results suggest that the SARC model requires improvements in the log layer; a modeler can use the result of the inference to further develop curvature corrections. An alternative viewpoint, is that this models aims at producing models in which the operators $\mathcal{P}(\mathbf{w})$ used in the (L3) modeling step, i.e. in Eq. 4, are informed by data resulting



(a) C_f (b) δ_{MAP} (c) δ_{SARC}

Figure 4: Skin-friction predictions and correction terms for a convex channel.

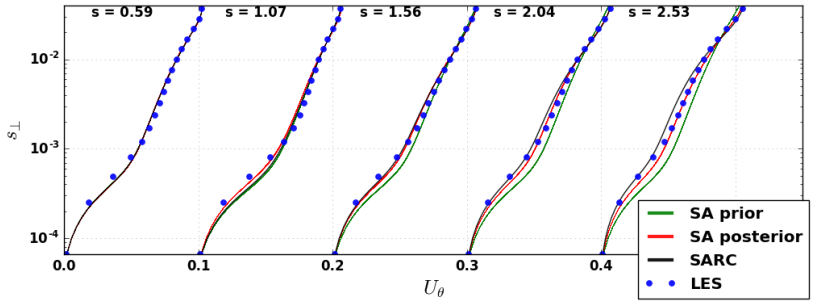


Figure 5: Predicted stream–wise velocity at various streamwise s locations for the convex channel. $s_\perp$ refers to the perpendicular distance from the lower wall.

in $\mathcal{P}(\mathbf{w};\boldsymbol{\theta})$.

5.2. Generalizing the embedded discrepancy

The studies reviewed above inferred the discrepancy as a spatially varying field using data that are directly relevant to the specific geometry and flow conditions of interest, and therefore are not easily generalizable. Although efforts such as the scenario averaging (Edeling et al. 2014b) provide some relief by incorporating evidence from different datasets, it is much more desirable to construct discrepancy functions that can be employed within a class of flows sharing similar features (e.g., separation, shock/boundary layer interaction, and jet/boundary layer interaction). Tracey et al. (2013) used machine learning to reconstruct discrepancies in the anisotropy tensor. Starting with the eigen-decomposition in Eq. 10, perturbations $\boldsymbol{\delta}_\lambda$ to the eigenvalues were derived at every spatial location $\boldsymbol{x}$ using a DNS dataset. At this point, $\boldsymbol{\delta}_\lambda(\boldsymbol{x})$ is mapped into a *feature* space, $\boldsymbol{\eta}(\boldsymbol{x})$, consisting of functions of relevant quantities such as mean velocity gradients and turbulent quantities. The mapping is *learned* via Gaussian process regression based on the DNS dataset, and the resulting discrepancy $\boldsymbol{\delta}_\lambda(\boldsymbol{\eta})$ was found to be relatively accurate. This work was followed by further developments in the field of machine learning and lead to a promising research av-

enue that combines turbulence modeling, inference, uncertainty quantification and learning strategies.

Machine Learning

Machine learning is an umbrella term for a wide range of techniques within the broader field of artificial intelligence and, it has been recently rejuvenated by algorithmic innovations, advances in computer hardware and the enormous growth in the availability of data.

Machine learning can be broadly categorized into unsupervised and supervised learning. In unsupervised learning, there are no specific targets to predict; the goal is to discover patterns and reduce the complexity of the data. Examples include clustering, i.e. grouping data points based on their similarity and dimension reduction, i.e. identifying a subset of dependent variables that describe the data. This is in contrast to supervised learning, where the objective is to construct a mapping of the inputs and the outputs. When the output is categorical, supervised machine learning strategies are also referred to as *classification* strategies; when the output is continuous these methods are referred to as *regression* approaches. The latter is of particular interest in the context of turbulence modeling. Example techniques commonly used in supervised learning (including both classification and regression) are random forests, support vector machines, and neural networks.

Neural networks are receiving considerable attention because of their capability to approximate complex functions in a flexible form. Typically, an extremely large number of calibration coefficients have to be determined to train a neural network, but efficient algorithms are available for implementation on modern computer architectures. From a mathematical perspective, Neural networks involve the composition of nonlinear functions. Starting with an example of a linear model, consider a data set $\boldsymbol{\theta}$ and a vector of inputs, or features, $\boldsymbol{\eta}$. A linear model for the output $\boldsymbol{\delta}(\boldsymbol{\eta})$ can be constructed considering $\boldsymbol{\delta}(\boldsymbol{\eta}) = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\beta}$, where the weight matrix $\boldsymbol{W}$ and the bias vector $\boldsymbol{\beta}$ are obtained by solving an optimization problem that minimizes the overall difference between $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$. This process is called model training or *learning*. However, such a simple model may lack the flexibility to represent a complex functional mapping, and therefore *intermediate* variables (layers) $\boldsymbol{\ell}$ are introduced:

$$\boldsymbol{\ell} = \sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{\eta} + \boldsymbol{\beta}^{(1)}\right) \quad \text{and} \quad \boldsymbol{\delta}(\boldsymbol{\eta}) = \boldsymbol{W}^{(2)}\boldsymbol{\ell} + \boldsymbol{\beta}^{(2)},$$

where σ is a user-specified *activation function* such as the hyperbolic tangent. The two-layer network written as composite function is

$$\boldsymbol{\delta}(\boldsymbol{\eta}) = \boldsymbol{W}^{(2)}\sigma\left(\boldsymbol{W}^{(1)}\boldsymbol{\eta} + \boldsymbol{\beta}^{(1)}\right) + \boldsymbol{\beta}^{(2)}.$$

The composition of several intermediate layers results in a *deep neural network*, which is capable of efficiently representing arbitrary complex functional forms.

5.3. Modeling using machine learning

Machine Learning (ML) provides effective strategies to construct mapping between large datasets and quantities of interest. ML can be applied directly as a *black-box* tool, or in combination with existing models to provide *a posteriori* corrections. Tracey et al. (2013) used supervised learning to represent perturbations to Barycentric coordinates. The perturbations are then reconstructed using a machine learning algorithm as a function of

local flow variables. While this is an activity at the L3 level (specific to a baseline model), the methodology is generally applicable to any turbulence model. In contrast, Wang et al. (2017) and Wu et al. (2018b) developed a more comprehensive perturbation strategy to predict discrepancies in the magnitude, anisotropy, and orientation of the Reynolds stress tensor. They demonstrated results for two sets of canonical flows, separated flows over periodic hills and secondary flows in a square duct. Representative results are presented in Figures 6 and 7, showing improved predictions of Reynolds stress anisotropy and mean velocities, respectively.

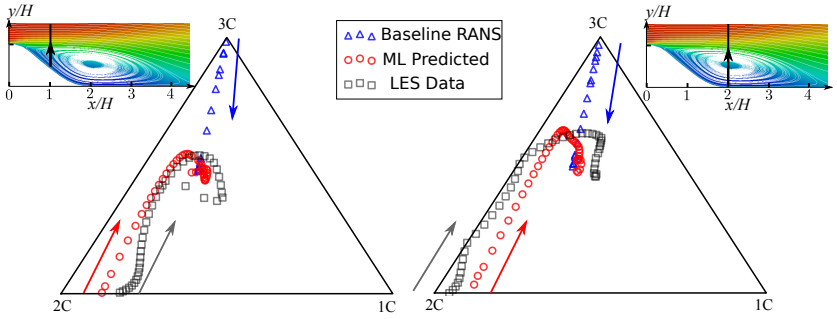


Figure 6: Anisotropy at locations (indicted in the insets) in the flow over periodic hills, predicted by using a random forest model trained on several separated flows in significant different *geometries* and configurations (Wang et al. 2017).

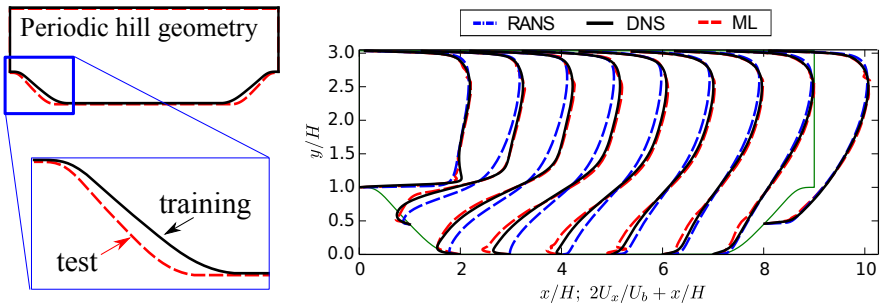


Figure 7: Velocities predicted by using machine-learning corrected Reynolds stresses. The training data is obtained from a flow over periodic hills in a slightly different geometry as shown on the left panel (Wu et al. 2018b).

An important aspect of applying machine learning techniques is to ensure the objectivity and the rotational invariance of the learned Reynolds stress models. Tracey et al. (2013),Wang et al. (2017), and Wu et al. (2018b) used tensor invariants based on the eigendecomposition of the Reynolds stresses, while for the representation of the stress orientation, both Euler angles and unit quaternions have been considered (Wu et al. 2017).

As discussed earlier a strategy to develop closures for Reynolds stresses is based on the formulation of a generalized expansion of the Reynolds stress tensor (Pope 1975). In the

assumption that the stresses only depend on the mean velocity gradient, one can write

$$\boldsymbol{\tau} = \sum_{n=1}^{10} c^{(n)} \boldsymbol{\mathcal{T}}^{(n)} \tag{11.}$$

where the coefficients $\boldsymbol{c}$ must be obtained from empirical information or further assumptions, while $\boldsymbol{\mathcal{T}}$ are known functions of the symmetric and anti-symmetric part of the velocity gradient tensor. In a machine learning framework, one rewrites the expansion as

$$\widetilde{\boldsymbol{\tau}} = \sum_{n=1}^{10} c^{(n)}(\boldsymbol{\theta}, \boldsymbol{\eta}) \boldsymbol{\mathcal{T}}^{(n)} \tag{12.}$$

Ling et al. (2016b) proposed a neural network architecture with embedded invariance properties to learn the coefficients $\boldsymbol{c}(\boldsymbol{\theta}, \boldsymbol{\eta})$ with good predictive capability but no explicit expression for the resulting model (i.e. any stress evaluation requires the use of the original, calibrated neural network). In a related effort Weatheritt and Sandberg (2016, 2017) used symbolic regression and gene expression programming for defining the coefficients $\boldsymbol{c}(\boldsymbol{\theta}, \boldsymbol{\eta})$ in the context of algebraic Reynolds stress models, resulting in an explicit model form that can be readily implemented in RANS solvers. DNS data of flow over a backward-facing step at a low Reynolds number is used for training, while the flow at a high Reynolds number is predicted. While the results are encouraging, a high level of uncertainty is observed by applying the resulting model to the flow over periodic hills.

The approaches discussed above use the same starting point, an L2 level assumption, and a different set of features ($\boldsymbol{\eta}$) and data ($\boldsymbol{\theta}$). The work of Ling and Templeton (2015) illustrated a scheme for crafting features based on flow physics and normalizing them using local quantities; later work has expanded this approach by using invariants of a tensorial set (Ling et al. 2016a). Wang et al. (2017) and Wu et al. (2018b) used such approaches to predict Reynolds stress discrepancies with a large feature set. These approaches only use local quantities to construct the set of features. In general, further work based on modeling non-local, non-equilibrium effects can expand the predictive capabilities of these approaches. For example, in a traditional eddy viscosity model Hamlington and Dahm (2008) used variables that account for non-local behavior through streamline integration, which provided an inspiring approach for choosing features in data-driven modeling.

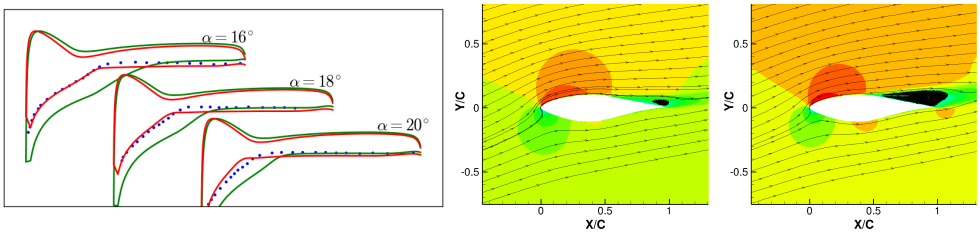
Tracey et al. (2015) performed a proof-of-concept study to learn known turbulence modeling terms from data using neural networks. The neural network based terms were then embedded within an iterative RANS solver and used for predictions, demonstrating the viability of using ML methods in a hybrid PDE/neural networks setting.

5.4. Combining inference and machine learning

If machine learning is applied directly to high-fidelity data, inconsistencies may arise between the training environment and the prediction environment. Since turbulence models are typically formulated to accurately represent first and second moments, many latent variables assume an operational rather than a physically precise role in the turbulence models. For instance, the role of the dissipation rate in a model is only to provide scale information, and the model is typically calibrated to provide accurate results for the moments, even if the values assumed by are different from the true value.

Statistical inference provides a rigorous framework to calibrate models using data while machine learning offers a flexible setting to formulate discrepancy functions in terms of a

vast number of general features. The combination of these two strategies yields the promise of deriving effective data-driven closures for turbulence models. Duraisamy et al. (2015) and Parish and Duraisamy (2016) have explored this avenue. In the first step, the spatial structure of the model discrepancy $\boldsymbol{\delta}(\boldsymbol{x})$ is extracted using statistical inference from datasets representative of the phenomena of interest. Machine learning is then used to reconstruct the functional form of the discrepancy in terms of the mean flow and turbulence variables, $\boldsymbol{\delta}(\boldsymbol{\eta})$.



(a) (b) (c)

Figure 8: Example of application of inference and learning to turbulent flows over airfoils (Singh et al. 2017b). (a) Pressure over airfoil surface (Green: Baseline model; Red: ML-augmented model; Blue: Experimental measurements).(b) Baseline flow prediction (pressure contours and streamlines). (c) Flow prediction using data-driven SA model.

The resulting discrepancy is then embedded in RANS solvers as a correction to traditional turbulence models results in convincing improvements of the predictions. In Singh et al. (2017a,b), this approach was used for the simulation of turbulent flow over airfoils. Figure 8 shows results from a data-driven SA model. Of particular interest, is that the inversion process does not require extensive datasets, and even very limited experimental measurements, such as the lift coefficient, provides useful information that lead to considerable improvements in the predictions with respect to the baseline models.

6. Challenges and perspectives

The concurrent enhancements in statistical inference algorithms, machine learning and uncertainty quantification approaches combined with the growth in available data is spurring renewed interest in turbulence modeling. We have surveyed efforts in the context of RANS modeling, however data-driven approaches are being pursued in a variety of contexts in fluid dynamics and for increasingly complex applications, such as multiphase flows. Promising activities in LES include the use on neural networks to model subgrid-scale stress (Gamahara and Hattori 2017; Vollant et al. 2014, 2017) and to represent the deconvolution of flow quantities from filtered flow fields (Maulik and San 2017). Ma et al. (2015, 2016) used machine learning to model the inter-phase mass and momentum fluxes in multiphase flow simulations.

In spite of recent successes, several challenges remain.

- *What data to use?* Databases of experimental measurements and DNS are readily available today, but they might have only limited relevance in specific problems of interest. The quantification of the *information content* in the data is a critical aspect

of data-driven models. Ideally, the process of calibration should provide direct indication of the need for additional data or potential overfitting. This is a classic setting to introduce formal design-of-experiments to drive further data-collection activities. In addition, the uncertainty present in the data must be accounted for during the inference and learning stages, and eventually propagated to the final prediction to set reasonable expectations in terms of prediction accuracy.

- *How to enforce data-model consistency?* If machine learning is applied directly on a dataset, a compounding problem is the consistency between the data and the models, i.e. the difference between the learning environment (DNS) and the injection environment (RANS). It is well known that even if DNS-computed quantities are used to completely replace specific terms in the RANS closure, the overall predictions will remain unsatisfactory (Poroseva et al. 2016; Thompson et al. 2016) due to the assumptions and approximations at various levels in models, compounded by the potential ill conditioning of the RANS equations (Wu et al. 2018c). Furthermore, scale-providing variables such as the turbulent dissipation rate will be very different in the RANS and DNS context. The addition of the inference step before the learning phase enforces consistency between the learning and prediction environment.
- *What to learn?* The blind application of learning techniques to predict a quantity of interest based on available data cannot be expected, in general, to produce credible results. A more realistic goal is to focus on learning discrepancy functions and an appropriate set of features that satisfy physical and mathematical constraints. But how many features are required? And what is the optimal choice for broad application of the resulting calibrated model to different problems? These remain open questions and the subject of on-going investigations. An alternative, promising approach is to focus on a specific component of a closure and introduce correction terms that can be learned from data, as in the example reported earlier corresponding to the curvature correction of the Spalart-Allmaras model (Singh and Duraisamy 2016). In this case, the learning strategy can provide direct insights to modelers.
- *What is the confidence in the predictions?* Calibrated models have potentially a limited applicability because of the unavoidable dependency on the data; furthermore, data-driven models might suffer from a lack of *interpretability*, i.e. the difficulty of explaining causal relationships between the data, the discrepancy and the corresponding prediction. The use of deep learning strategies and vast amount of data in the inference process exacerbates this issue.
- *What is the right balance between data and models?* Recent works (Brunton et al. 2016; Raissi et al. 2017; Schmidt and Lipson 2009) have explored the possibility of extracting models purely from data. It has been shown that the analytical form of *simple* dynamical systems can be extracted from data and solutions of the Navier–Stokes equations can be reconstructed on a given grid and flow condition. While these methods have focused on *rediscovering* known governing equations or *reconstructing* known solutions, the possibility of discovering unknown equations and deriving accurate predictive models purely from data remains an open question, though unknown closure terms have been extracted for simple dynamical systems (Pan and Duraisamy 2018). Ultimately, the decision to leverage existing model structures and incorporate/enforce prior knowledge and physics constraints is a *modeling choice*. In the limit of *infinite amounts of data*, machine learning could potentially identify universal closure model equations directly from data. On the other hand, relying on

machine learning alone, when dealing with large but finite amount of data, problem-specific/spurious laws might be discovered, resulting in very limited predictive value. Therefore, the modeler's choice is dictated by the *relative faith* in the available data and prior model structures, physical constraints, and the purpose of the model itself (e.g. whether the model will be used in reconstruction, or parametric prediction, or true prediction).

In general, a holistic approach that (1) leverages advances in statistical inference *and* learning, (2) combines the data-driven process with the assessment of the *information content*, (3) complies with physical and mathematical constraints, (4) acknowledges the assumptions intrinsic in the closures, and (5) rigorously quantifies uncertainties, has the potential to lead to credible and useful models.

In conclusion, we expect a pervasive growth of data-driven models fueled by advances in algorithms and accelerated by novel computer architectures. Moreover, we expect data to profoundly impact models in all aspects, i.e. through parameters $\mathbf{c}$, algebraic or differential operators $\mathcal{P}$ and discrepancy $\boldsymbol{\delta}$; this will result in general models written as:

$$\widetilde{\mathcal{M}} = \mathcal{M}(\mathbf{w}; \mathcal{P}(\mathbf{w}; \boldsymbol{\theta}); \mathbf{c}(\boldsymbol{\theta}); \boldsymbol{\delta}(\boldsymbol{\theta}, \boldsymbol{\eta}); \boldsymbol{\epsilon}_{\boldsymbol{\theta}}). \tag{13.}$$

and recommend the resulting predictions to be accompanied by explicit uncertainty estimates $\widetilde{\mathbf{q}} = \widetilde{\mathcal{M}} + \boldsymbol{\epsilon}_{\mathbf{q}}$.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

GI acknowledges support from the Advanced Simulation and Computing program of the US Department of Energy via the PSAAP II Center at Stanford under contract DE-NA-0002373. KD acknowledges NASA grant NNX15AN98A (tech. monitor: Dr. Gary Coleman) and NSF grant 1507928 (tech. monitor: Dr. Ron Joslin). GI and KD acknowledge support from the Defense Advanced Research Projects Agency under the Enabling Quantification of Uncertainty in Physical Systems (EQUiPS) project (tech. monitor: Dr Fariba Fahroo). HX acknowledges support and mentoring from the Department of Aerospace and Ocean Engineering at Virginia Tech and particularly Prof. C.J. Roy and Prof. E.G. Paterson. The authors would like to thank Dr. J.-L. Wu, Dr. J.-X. Wang, Dr. J. Ling, Dr. A. Singh and Dr. B. Tracey for their collaborations and Dr. A. A. Mishra for useful suggestions on the manuscript.

LITERATURE CITED

R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter estimation and inverse problems*, volume 90. Academic Press, 2011.

S. Banerjee, R. Krahl, F. Durst, and C. Zenger. Presentation of anisotropy properties of turbulence, invariants versus eigenvalue approaches. *Journal of Turbulence*, 8(32):1–27, 2007.

S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

F. Busse. Bounds for turbulent shear flow. *Journal of Fluid Mechanics*, 41(1):219–240, 1970.

S. H. Cheung, T. A. Oliver, E. E. Prudencio, S. Prudhomme, and R. D. Moser. Bayesian uncertainty analysis with applications to turbulence modeling. *Reliability Engineering & System Safety*, 96 (9):1137–1149, 2011.

G. Comte-Bellot and S. Corrsin. The use of a contraction to improve the isotropy of grid-generated turbulence. *Journal of fluid mechanics*, 25(4):657–682, 1966.

J. Coupland. ERCOFTAC classic database, 1993. `http://cfd.mace.manchester.ac.uk/ercoftac`.

O. Darrigol. *Worlds of flow: A history of hydrodynamics from the Bernoullis to Prandtl*. Oxford University Press, 2005.

C. R. Doering and P. Constantin. Variational bounds on energy dissipation in incompressible flows: Shear flow. *Physical Review E*, 49(5):4087, 1994.

E. Dow and Q. Wang. Quantification of structural uncertainties in the k–ω turbulence model. In *52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, pages 2011–1762, Denver, Colorado, 4-7 April 2011 2011. AIAA.

K. Duraisamy, Z. J. Zhang, and A. P. Singh. New approaches in turbulence and transition modeling using data-driven techniques. *AIAA Paper*, 1284:2015, 2015.

P. Durbin and C. Speziale. Realizability of second-moment closure via stochastic analysis. *Journal of Fluid Mechanics*, 280:395–407, 1994.

P. A. Durbin. Some recent developments in turbulence closure modeling. *Annual Review of Fluid Mechanics*, (0), 2017.

W. N. Edeling, P. Cinnella, and R. P. Dwight. Predictive RANS simulations via Bayesian model-scenario averaging. *Journal of Computational Physics*, 275:65–91, 2014a.

W. N. Edeling, P. Cinnella, R. P. Dwight, and H. Bijl. Bayesian estimates of parameter variability in the k–ε turbulence model. *Journal of Computational Physics*, 258:73–94, 2014b.

W. N. Edeling, G. Iaccarino, and P. Cinnella. Data-free and data-driven rans predictions with quantified uncertainty. *Flow, Turbulence and Combustion*, pages 1–24, 2017.

B. Efron and T. Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

B. Eisfeld. Reynolds stress anisotropy in self-preserving turbulent shear flows. 2017.

M. Emory, R. Pecnik, and G. Iaccarino. Modeling structural uncertainties in Reynolds-averaged computations of shock/boundary layer interactions. *AIAA paper*, 479:2011, 2011.

M. Emory, J. Larsson, and G. Iaccarino. Modeling of structural uncertainties in Reynolds-averaged Navier-Stokes closures. *Physics of Fluids (1994-present)*, 25(11):110822, 2013.

M. Gamahara and Y. Hattori. Searching for turbulence models by artificial neural network. *Physical Review Fluids*, 2(5):054604, 2017.

T. Gatski and T. Jongen. Nonlinear eddy viscosity and algebraic stress models for solving complex turbulent flows. *Progress in Aerospace Sciences*, 36(8):655–682, 2000.

M. B. Giles, M. C. Duta, J.-D. M-uacute, ller, and N. A. Pierce. Algorithm developments for discrete adjoint methods. *AIAA journal*, 41(2):198–205, 2003.

S. S. Girimaji. Partially-averaged navier-stokes model for turbulence: A Reynolds-averaged Navier–Stokes to direct numerical simulation bridging method. *Journal of Applied Mechanics*, 73(3): 413–421, 2006.

C. Gorlé, J. Larsson, M. Emory, and G. Iaccarino. The deviation from parallel shear flow as an indicator of linear eddy-viscosity model inaccuracy. *Physics of Fluids (1994-present)*, 26(5): 051702, 2014.

P. E. Hamlington and W. J. Dahm. Reynolds stress closure for nonequilibrium effects in turbulent flows. *Physics of Fluids*, 20(11):115101, 2008.

L. N. Howard. Bounds on flow quantities. *Annual Review of Fluid Mechanics*, 4(1):473–494, 1972.

G. Iaccarino, A. A. Mishra, and S. Ghili. Eigenspace perturbations for uncertainty estimation of single-point turbulence closures. *Physical Review Fluids*, 2(2):024605, 2017.

M. A. Iglesias, K. J. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems.

Inverse Problems, 29(4):045001, 2013.

L. Jofre, S. P. Domino, and G. Iaccarino. A framework for characterizing structural uncertainty in large-eddy simulation closures. *Flow, Turbulence and Combustion*, 100(2):341–363, 2018.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

J. Kim, P. Moin, and R. Moser. Turbulence statistics in fully developed channel flow at low reynolds number. *Journal of Fluid Mechanics*, 177:133–166, 1987.

B. Launder, G. J. Reece, and W. Rodi. Progress in the development of a reynolds-stress turbulence closure. *Journal of fluid mechanics*, 68(3):537–566, 1975.

S. Lefantzi, J. Ray, S. Arunajatesan, and L. Dechant. Estimation of k–ε parameters using surrogate models and jet-in-crossflow data. Technical report, Sandia National Laboratories, Livermore, CA, USA, 2015.

Y. Li, E. Perlman, M. Wan, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence*, (9):N31, 2008.

J. Ling and J. Templeton. Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty. *Physics of Fluids (1994-present)*, 27(8): 085103, 2015.

J. Ling, R. Jones, and J. Templeton. Machine learning strategies for systems with invariance properties. *Journal of Computational Physics*, 318:22–35, 2016a.

J. Ling, A. Kurzawski, and J. Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016b.

J. L. Lumley. Computational modeling of turbulent flows. *Advances in applied mechanics*, 18(123): 213, 1978.

M. Ma, J. Lu, and G. Tryggvason. Using statistical learning to close two-fluid multiphase flow equations for a simple bubbly system. *Physics of Fluids*, 27(9):092101, 2015.

M. Ma, J. Lu, and G. Tryggvason. Using statistical learning to close two-fluid multiphase flow equations for bubbly flows in vertical channels. *International Journal of Multiphase Flow*, 85: 336–347, 2016.

R. Maulik and O. San. A neural network approach for the blind deconvolution of turbulent flows. *Journal of Fluid Mechanics*, 831:151–181, 2017.

A. A. Mishra and G. Iaccarino. Uncertainty estimation for reynolds-averaged navier–stokes predictions of high–speed aircraft nozzle jets. *AIAA Journal*, 55:3999–4004, 2017.

T. Oliver and R. Moser. Uncertainty quantification for RANS turbulence model predictions. In *APS Division of Fluid Dynamics Meeting Abstracts*, Nov. 2009.

T. A. Oliver and R. D. Moser. Bayesian uncertainty quantification applied to RANS turbulence models. In *Journal of Physics: Conference Series*, volume 318, page 042032. IOP Publishing, 2011.

S. Pan and K. Duraisamy. Data-driven Discovery of Closure Models. *SIAM Journal on Applied Dynamical Systems, arXiv:1803.09318*, 2018.

E. J. Parish and K. Duraisamy. A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, 305:758–774, 2016.

S. Pope. A more general effective-viscosity hypothesis. *Journal of Fluid Mechanics*, 72(2):331–340, 1975.

S. Pope. PDF methods for turbulent reactive flows. *Progress in Energy and Combustion Science*, 11(2):119–192, 1985.

S. B. Pope. *Turbulent Flows*. Cambridge University Press, 2000.

S. Poroseva, J. D. Colmenares F, and S. Murman. On the accuracy of RANS simulations with DNS data. *Physics of Fluids*, 28(11):115102, 2016.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part ii): Data-driven discovery of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10566*,

2017.

A. Rastegari and R. Akhavan. The common mechanism of turbulent skin-friction drag reduction with superhydrophobic longitudinal microgrooves and riblets. *Journal of Fluid Mechanics*, 838: 68–104, 2018.

J. Ray, S. Lefantzi, S. Arunajatesan, and L. Dechant. Bayesian parameter estimation of a k–ε model for accurate jet-in-crossflow simulations. *AIAA Journal*, pages 2432–2448, 2016.

J. Ray, S. Lefantzi, S. Arunajatesan, and L. Dechant. Learning an eddy viscosity model using shrinkage and Bayesian calibration: A jet-in-crossflow case study. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 4(1):011001, 2018.

R. E. Rudd and J. Q. Broughton. Coarse-grained molecular dynamics and the atomic limit of finite elements. *Physical review B*, 58(10):R5893, 1998.

M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *science*, 324 (5923):81–85, 2009.

U. Schumann. Realizability of Reynolds-stress turbulence models. *Physics of Fluids (1958-1988)*, 20(5):721–725, 1977.

C. Seis. Scaling bounds on dissipation in turbulent flows. *Journal of Fluid Mechanics*, 777:591–603, 2015.

M. L. Shur, M. K. Strelets, A. K. Travin, and P. R. Spalart. Turbulence modeling in rotating and curved channels: assessing the spalart-shur correction. *AIAA journal*, 38(5):784–792, 2000.

A. P. Singh and K. Duraisamy. Using field inversion to quantify functional errors in turbulence closures. *Physics of Fluids (1994-present)*, 28(4):045110, 2016.

A. P. Singh, K. Duraisamy, and Z. J. Zhang. Augmentation of turbulence models using field inversion and machine learning. In *55th AIAA Aerospace Sciences Meeting*, page 0993, 2017a.

A. P. Singh, S. Medida, and K. Duraisamy. Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA Journal*, pages 1–13, 2017b.

P. Spalart and S. Allmaras. A one-equation turbulence model for aerodynamic flows. In *30th aerospace sciences meeting and exhibit*, page 439, 1992.

P. R. Spalart. Detached-eddy simulation. *Annual review of fluid mechanics*, 41:181–202, 2009.

R. L. Thompson, L. E. B. Sampaio, F. A. V. de Bragança A., L. Thais, and G. Mompean. A methodology to evaluate statistical errors in DNS data of plane channel flows. *Computers & Fluids*, 130:1–7, 2016.

B. Tracey, K. Duraisamy, and J. Alonso. Application of supervised learning to quantify uncertainties in turbulence and combustion modeling. In *51st AIAA Aerospace Sciences Meeting*, 2013. Dallas, TX, paper 2013-0259.

B. Tracey, K. Duraisamy, and J. J. Alonso. A machine learning strategy to assist turbulence model development. *AIAA Paper*, 1287:2015, 2015.

A. Vollant, G. Balarac, G. Geraci, and C. Corre. Optimal estimator and artificial neural network as efficient tools for the subgrid-scale scalar flux modeling. Technical report, Proceedings of Summer Research Program, Center of Turbulence Research, Stanford University, Stanford, CA, USA, 2014.

A. Vollant, G. Balarac, and C. Corre. Subgrid-scale scalar flux modelling based on optimal estimation theory and machine-learning procedures. *Journal of Turbulence*, pages 1–25, 2017.

J.-X. Wang, R. Sun, and H. Xiao. Quantification of uncertainties in turbulence modeling: A comparison of physics-based and random matrix theoretic approaches. *International Journal of Heat and Fluid Flow*, 62:577–592, 2016a.

J.-X. Wang, J.-L. Wu, and H. Xiao. Incorporating prior knowledge for quantifying and reducing model-form uncertainty in RANS simulations. *International Journal for Uncertainty Quantification*, 6(2), 2016b.

J.-X. Wang, J.-L. Wu, and H. Xiao. Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data. *Physical Review Fluids*, 2(3):034603, 2017.

J. Weatheritt and R. Sandberg. A novel evolutionary algorithm applied to algebraic modifications of the RANS stress–strain relationship. *Journal of Computational Physics*, 325:22–37, 2016.

J. Weatheritt and R. D. Sandberg. The development of algebraic stress models using a novel evolutionary algorithm. *International Journal of Heat and Fluid Flow*, 2017.

D. C. Wilcox. *Turbulence modeling for CFD*. DCW industries La Canada, CA, 2006.

J.-L. Wu, J.-X. Wang, and H. Xiao. A Bayesian calibration–prediction method for reducing model-form uncertainties with application in RANS simulations. *Flow, Turbulence and Combustion*, 97 (3):761–786, 2016.

J.-L. Wu, R. Sun, S. Laizet, and H. Xiao. Representation of Reynolds stress perturbations with application in machine-learning-assisted turbulence modeling. *Computer Methods in Applied Mechanics and Engineering [arXiv:1709.05683]*, 2018.

J.-L. Wu, C. M. Ströfer, and H. Xiao. PDE-informed construction of covariance kernel in uncertainty quantification of random fields. In preparation. Bibliography details to be available while this article is under review, 2018a.

J.-L. Wu, H. Xiao, and E. Paterson. Data-driven augmentation of turbulence models with physics-informed machine learning. *Physical Review Fluids*, 3:074602, 2018b.

J.-L. Wu, H. Xiao, R. Sun, and Q. Wang. RANS equations with Reynolds stress closure can be ill-conditioned. Submitted. arXiv:1803.05581, 2018c.

H. Xiao, J.-L. Wu, J.-X. Wang, R. Sun, and C. J. Roy. Quantifying and reducing model-form uncertainties in Reynolds-averaged Navier–Stokes simulations: A data-driven, physics-informed bayesian approach. *Journal of Computational Physics*, 324:115–136, 2016.

H. Xiao and P. Cinnella. 2018. Quantification of model uncertainty in RANS simulations: a review. *Progress in Aerospace Sciences. arXiv:1806.10434 [physics.flu-dyn]*, 2018

H. Xiao, J.-X. Wang, and R. G. Ghanem. A random matrix approach for quantifying model-form uncertainties in turbulence modeling. *Computer Methods in Applied Mechanics and Engineering*, 313:941–965, 2017.