ANNUAL REVIEWS

Annual Review of Fluid Mechanics

Learning Nonlinear Reduced Models from Data with Operator Inference

Boris Kramer,[1] Benjamin Peherstorfer,[2] and Karen E. Willcox[3]

[1]Department of Mechanical and Aerospace Engineering, University of California San Diego, La Jolla, California, USA

[2]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

[3]Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas, USA; email: kwillcox@oden.utexas.edu

Annu. Rev. Fluid Mech. 2024. 56:521–48

First published as a Review in Advance on November 1, 2023

The *Annual Review of Fluid Mechanics* is online at fluid.annualreviews.org

https://doi.org/10.1146/annurev-fluid-121021-025220

Keywords

data-driven modeling, scientific machine learning, structure preservation, nonlinear model reduction, Operator Inference

Abstract

This review discusses Operator Inference, a nonintrusive reduced modeling approach that incorporates physical governing equations by defining a structured polynomial form for the reduced model, and then learns the corresponding reduced operators from simulated training data. The polynomial model form of Operator Inference is sufficiently expressive to cover a wide range of nonlinear dynamics found in fluid mechanics and other fields of science and engineering, while still providing efficient reduced model computations. The learning steps of Operator Inference are rooted in classical projection-based model reduction; thus, some of the rich theory of model reduction can be applied to models learned with Operator Inference. This connection to projection-based model reduction theory offers a pathway toward deriving error estimates and gaining insights to improve predictions. Furthermore, through formulations of Operator Inference that preserve Hamiltonian and other structures, important physical properties such as energy conservation can be guaranteed in the predictions of the reduced model beyond the training horizon. This review illustrates key computational steps of Operator Inference through a large-scale combustion example.

1. INTRODUCTION

Surrogate model: approximates high-fidelity numerical simulations of physical phenomena; can be categorized into three types: statistical data-fit models, simplified models, and reduced models

Learning models from data via automated methods is an increasingly important component of computational science and engineering. We distinguish between two different broad problem settings of learning models from data: (*a*) The governing equations of the physical phenomena of interest are unknown and the goal is to discover them from data, and (*b*) the governing equations of the physical phenomena are known, high-fidelity numerical simulations are available but are prohibitively expensive for the task at hand, and the goal is to learn a computationally efficient surrogate model. This review addresses the second problem setting. For discussions about the different problem setting of learning governing equations from data, we refer the reader to Langley (1981), Schmidt & Lipson (2009), Brunton et al. (2016b), and Raissi & Karniadakis (2018).

The high cost of numerical simulations for complex physical phenomena is a major barrier to achieving optimization, design, control, data assimilation, and uncertainty quantification for scientific and engineering systems. These numerical tasks are all so-called outer-loop applications (Peherstorfer et al. 2018) that require repeated simulations for different inputs, parameters, and configurations. Surrogate models provide approximations of high-fidelity numerical simulations at greatly reduced costs and play a key role in making these outer-loop applications tractable. When it comes to approximating high-fidelity numerical simulations of physical phenomena, surrogate models can be categorized into three types: statistical data-fit models, simplified models, and reduced models. Statistical data-fit models approximate the input–output maps induced by high-fidelity numerical simulations. The maps are fitted with statistical methods from training data, with the surrogate model employing a generic functional form that does not explicitly reflect the structure of the physical governing equations underlying the numerical simulations. While many different parameterizations of the input–output map have been considered, Gaussian process models have been particularly successful because they are equipped with error indicators that can be used for adaptation (Rasmussen & Williams 2006, Forrester et al. 2008).

Surrogate models of the second type, simplified models, are obtained by simplifying the models underlying the high-fidelity numerical simulations to obtain approximations with reduced cost. For example, nonlinear terms can be linearized, physics can be simplified, iterative solvers can be terminated early, and coarser grids can be used for discretization. In-depth domain knowledge about the physics, governing equations, and numerical methods underlying the simulations is necessary to understand in which situations these simplifications can be made while maintaining sufficient accuracy.

Reduced models form the third category of surrogate models and encompass elements of both statistical data-fit and simplified models. Reduced modeling—also referred to as model reduction—learns patterns from training data of high-fidelity numerical simulations in order to identify low-dimensional structure, while embedding knowledge of the numerical models and governing equations in the form of the reduced model. In this sense, reduced modeling is an early example of scientific machine learning, because learning from data is combined with incorporating physical structure and insights obtained from the numerical models used in high-fidelity numerical simulations (for discussions about scientific machine learning and physics-informed machine learning, see Baker et al. 2019, Duraisamy et al. 2019, Swischuk et al. 2019, Brunton et al. 2020, and Karniadakis et al. 2021, as well as Coveney et al. 2016 and Willcox et al. 2021).

A large class of reduced models is based on projection, where low-dimensional subspaces that capture the most important dynamics are learned from data and the governing equations are solved in the subspaces via projection (for textbooks and surveys, see Antoulas 2005, Rozza et al. 2008, Quarteroni & Rozza 2014, Benner et al. 2015, Hesthaven et al. 2016). Reduced models based on

projection have a long history in fluid mechanics and represent some of the pioneering research in the field (Lumley 1967, Sirovich 1987, Holmes et al. 1996, Hall et al. 2000, Dowell & Hall 2001, Rowley & Dawson 2017). An advantage of reduced modeling techniques is the availability of rigorous theoretical guarantees in some settings, in particular via a posteriori error estimation (Prud'homme et al. 2001, Veroy et al. 2002, Hinze & Volkwein 2005, Haasdonk & Ohlberger 2011, Urban & Patera 2012). A drawback is that reduced models have traditionally been constructed by intrusive methods that compute the reduced operators by explicit projection of the governing equations onto the low-dimensional subspace. This process requires access to the numerical operators of the high-fidelity simulations either in assembled form or via routines that provide the action of the operators. As discussed by Ghattas & Willcox (2021), while this access is possible in many settings, it has been a barrier to practical adoption of model reduction, especially for legacy and commercial numerical tools.

Intrusive methods: require access to the operators of the high-fidelity numerical solver either in assembled form or via routines that provide the action of the operators on functions

Instead of constructing reduced models via intrusive procedures, we aim to learn reduced models nonintrusively from training data while maintaining some of the theoretical guarantees provided by intrusive methods. A major advantage of nonintrusive reduced modeling is its ease of implementation, because the high-fidelity simulators are used as data generators only. Structure, knowledge about governing equations, and other physical insights can still be embedded in a nonintrusive reduced model, but access to the high-fidelity operators in an intrusive sense is avoided. Furthermore, nonintrusive formulations offer the flexibility to learn a reduced model from both simulated and experimental data (e.g., Schmid 2010, Hemati et al. 2017). This is particularly relevant to fluid mechanics, where techniques such as particle image velocimetry can provide high-resolution spatiotemporal flow data.

Nonintrusive reduced modeling can be achieved in various ways. Some methods learn only parts of the dynamics of the high-dimensional numerical models (as in Gear et al. 2003). A range of methods learn physics-informed representations obtained from high-fidelity numerical simulations (Audouze et al. 2009, Hesthaven & Ubbiali 2018, Swischuk et al. 2019). In the systems and control community, nonintrusive model reduction has been widely studied and is related to system identification (Ljung 1987). The Loewner approach leverages the dynamical-system structure of high-fidelity models and fits rational functions to frequency-response data (Antoulas et al. 2021). The dynamic mode decomposition (DMD) (Rowley et al. 2009, Schmid 2010, Tu et al. 2014, Kutz et al. 2016) best-fits linear operators to state trajectories (see also the survey in Schmid 2022). Methods based on Koopman operators have been developed to extend DMD to nonlinear systems (Mezić 2005, Williams et al. 2015, Brunton et al. 2016a), where the challenge is selecting observables such that the dynamics become close to linear.

In this review, we discuss reduced modeling with Operator Inference, introduced by Peherstorfer & Willcox (2016a). Similarly to DMD, Operator Inference fits operators of reduced models to data; however, it allows for nonlinear terms and thus can capture nonlinear dynamics. Operator Inference explicitly embeds the underlying physics through the structured form of the reduced model it learns. In Section 2, we describe the basic Operator Inference approach. Section 3 demonstrates Operator Inference on the example of learning a surrogate model for a large-scale computational fluid dynamics (CFD) model of a combustion process. Section 4 discusses conditions under which Operator Inference recovers the same reduced models that would be obtained with intrusive projection-based model reduction, which provides a pathway to carry over the rich theory of intrusive methods to data-driven modeling with Operator Inference. In Section 5, we delve into structure preservation, noting that a major advantage of reduced models compared with statistical data-fit models is explicit preservation of the structure of the underlying physics and high-fidelity numerical models. Conclusions are drawn in Section 6.

2. OPERATOR INFERENCE

Nonintrusive methods: have access to simulated data and potentially additional information; have no access to the operators of the numerical models in an intrusive sense

This section reviews Operator Inference, a nonintrusive method for learning low-dimensional computationally efficient surrogate models that approximate large-scale, expensive numerical simulations.

2.1. High-Fidelity Physics-Based Numerical Models and Their Structure

This review considers the large class of scientific and engineering applications governed by partial differential equations (PDEs). In scientific computing, numerical models for such systems are typically obtained by discretizing the governing PDEs with numerical methods such as finite-volume, finite-difference, and finite-element schemes. The resulting numerical models describe the underlying systems of interest with high fidelity but often entail high computational costs when used for numerical simulations. In this subsection, we present a generic form for such models, with an emphasis on the structured model form that arises on the basis of the physical governing equations at hand. In the next subsection, we discuss how Operator Inference exploits this structure to derive data-driven reduced models.

2.1.1. Numerical models. To keep the discussion general, we consider a generic, spatially discretized form of the governing PDEs. That is, we consider numerical models of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t;\boldsymbol{\mu}) = \boldsymbol{f}(\boldsymbol{x}(t;\boldsymbol{\mu}), \boldsymbol{u}(t);\boldsymbol{\mu}), \qquad 1.$$

where $\boldsymbol{x}(t;\boldsymbol{\mu}) \in \mathbb{R}^N$ is the spatially discretized state vector. The dimension of the state is $N \in \mathbb{N}$, which scales with the (typically large) number of degrees of freedom in the spatial discretization. In the most general case, the state depends on time t and the d'-dimensional parameter $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^{d'}$. The d-dimensional input $\boldsymbol{u}(t) \in \mathbb{R}^d$ represents terms such as time-dependent boundary conditions and source terms. The dynamics are described by the function $\boldsymbol{f} : \mathbb{R}^N \times \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^N$.

2.1.2. Structure of numerical models. In numerical models of interest in science and engineering, the function $\boldsymbol{f}$ in Equation 1 typically has a particular structure that reflects the terms in the equations governing the modeled physical processes. In fluid mechanics, examples of these physical processes are convection, diffusion, and reaction, each of which gives rise to terms with varying, but known, structural forms. For example, consider the following mathematical model given by the Burgers equation for convection-diffusion flows:

$$\frac{\partial}{\partial t}w(t,\xi;\boldsymbol{\mu}) + w(t,\xi;\boldsymbol{\mu})\frac{\partial}{\partial \xi}w(t,\xi;\boldsymbol{\mu}) - \mu_1\frac{\partial^2}{\partial \xi^2}w(t,\xi;\boldsymbol{\mu}) = \mu_2 s(t,\xi), \qquad \xi \in \Omega, \qquad 2.$$

with solution field $w(t,\xi;\boldsymbol{\mu})$ at time t and spatial coordinate ξ in the spatial domain Ω. Equation 2 depends on the source term $s(t,\xi)$ and on the parameter $\boldsymbol{\mu} = [\mu_1, \mu_2] \in \mathbb{R}^2$, with the components of $\boldsymbol{\mu}$ corresponding to the viscosity, μ_1, and source-term parameter, μ_2.

Imposing appropriate boundary conditions, and discretizing Equation 2 in the spatial domain with, for example, a finite-difference method, yields a system of ordinary differential equations, as shown in Equation 1. In this case, the function $\boldsymbol{f}$ has linear-quadratic structure:

$$\boldsymbol{f}(\boldsymbol{x}(t;\boldsymbol{\mu}), \boldsymbol{u}(t);\boldsymbol{\mu}) = \boldsymbol{A}_1(\boldsymbol{\mu})\boldsymbol{x}(t;\boldsymbol{\mu}) + \boldsymbol{A}_2\boldsymbol{x}^2(t;\boldsymbol{\mu}) + \boldsymbol{B}(\boldsymbol{\mu})\boldsymbol{u}(t)\,.$$

Here, $\boldsymbol{x}(t;\boldsymbol{\mu}) \in \mathbb{R}^N$ is the spatially discretized approximation of the state w, which depends on time t and the parameter vector $\boldsymbol{\mu}$, and $\boldsymbol{u}(t)$ is the spatially discretized representation of the source term $s(t,\xi)$. The matrix $\boldsymbol{A}_1(\boldsymbol{\mu}) \in \mathbb{R}^{N\times N}$ arises from discretization of the linear diffusion term in Equation 2 [i.e., $\boldsymbol{A}_1(\boldsymbol{\mu})\boldsymbol{x}(t;\boldsymbol{\mu})$ corresponds to the numerical approximation of the term $\mu_1\partial_\xi^2 w(t,\xi;\boldsymbol{\mu})$]. The matrix $\boldsymbol{A}_2 \in \mathbb{R}^{N\times N_2}$ arises from discretizing the nonlinear convection term in

NOTE ON POLYNOMIAL FORM

Many numerical models arising in fluid mechanics naturally have polynomial form. Examples include the shallow-water equations, incompressible Navier-Stokes equations, and Euler equations (Hughes et al. 1986, Balajewicz et al. 2016, Qian et al. 2020). Furthermore, many nonpolynomial models can be written in polynomial form, as in Equation 3, after variable transformations are applied. For example, the Euler equations have quadratic structure when written in specific volume variables.

Equation 2, which has a quadratic dependence on the state. The squared vector $\boldsymbol{x}^2(t;\boldsymbol{\mu})$ contains all components of the Kronecker product $\boldsymbol{x}(t;\boldsymbol{\mu}) \otimes \boldsymbol{x}(t;\boldsymbol{\mu})$ except the duplicates due to commutativity of multiplication. For example, the squared vector of $\boldsymbol{x} = [x_1, x_2]^\top$ is $\boldsymbol{x}^2 = [x_1x_1, x_1x_2, x_2x_2]^\top$; the vector $\boldsymbol{x}^2$ is obtained by removing all duplicates due to the commutativity of multiplication from the Kronecker product: $\boldsymbol{x} \otimes \boldsymbol{x} = [x_1x_1, x_1x_2, x_2x_1, x_2x_2]^\top$. The squared vector $\boldsymbol{x}^2(t;\boldsymbol{\mu})$ has $N_2 = \binom{N+1}{2}$ components, which is also the number of columns of $\boldsymbol{A}_2$. The matrix $\boldsymbol{B}(\boldsymbol{\mu}) \in \mathbb{R}^{N\times d}$ maps the effects on the dynamics of the source term and constant terms such as those arising from discretization of boundary operators.

Just as discretizing the Burgers equation leads to a numerical model with linear-quadratic structure, the models of many processes and phenomena lead to a form of $\boldsymbol{f}$ in Equation 1 that has polynomial structure. For example, heat conduction described by linear diffusion models leads to linear time-invariant dynamics so that $\boldsymbol{f}(\boldsymbol{x}(t;\boldsymbol{\mu}), \boldsymbol{u}(t);\boldsymbol{\mu}) = \boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{x}(t;\boldsymbol{\mu}) + \boldsymbol{B}(\boldsymbol{\mu})\boldsymbol{u}(t)$, which is a polynomial of degree one in the state $\boldsymbol{x}(t;\boldsymbol{\mu})$. For more examples, see the sidebar titled Note on Polynomial Form. These examples motivate us to consider polynomial models,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t;\boldsymbol{\mu}) = \boldsymbol{f}(\boldsymbol{x}(t;\boldsymbol{\mu}), \boldsymbol{u}(t);\boldsymbol{\mu}) = \sum_{i=1}^{\ell} \boldsymbol{A}_i(\boldsymbol{\mu})\boldsymbol{x}^i(t;\boldsymbol{\mu}) + \boldsymbol{B}(\boldsymbol{\mu})\boldsymbol{u}(t), \qquad 3.$$

where $\ell \in \mathbb{N}$ is the degree of the polynomial and $\boldsymbol{x}^i(t;\boldsymbol{\mu}) \in \mathbb{R}^{N_i}$ contains the components of the i-times Kronecker product $\boldsymbol{x}(t;\boldsymbol{\mu}) \otimes \cdots \otimes \boldsymbol{x}(t;\boldsymbol{\mu})$ up to the duplicates due to commutativity of multiplication. Notice that the vector $\boldsymbol{x}^i(t;\boldsymbol{\mu})$ has $N_i = \binom{N+i-1}{i}$ components, whereas the i-times Kronecker product $\boldsymbol{x}(t;\boldsymbol{\mu}) \otimes \cdots \otimes \boldsymbol{x}(t;\boldsymbol{\mu})$ has N^i (N to the ith power) components. For $i = 1, \ldots, \ell$, the matrix $\boldsymbol{A}_i(\boldsymbol{\mu})$ has size $N \times N_i$, and $\boldsymbol{B}(\boldsymbol{\mu})$ has size $N \times d$. In the next subsection, we discuss the transformation of governing equations for an even broader class of systems into the polynomial form of Equation 3.

2.1.3. Lifting of nonpolynomial models. The polynomial model form in Equation 3 already encompasses a large portion of discretized processes in engineering and science; however, a number of mathematical models include nonpolynomial terms. The vast majority of these fall into the class of systems that can be written in polynomial form by leveraging variable transformations. To begin, we note that dynamical system models are not unique: The same process can be modeled mathematically with different variables, which can have a tremendous impact on computational modeling and analysis.

Lifting: refers to the process of rewriting a nonlinear model with different variables such that it becomes of polynomial form, and at best quadratic

The idea of variable transformations (referred to as lifting when extra variables are added) to promote model structure has a long history spanning different communities. In fluid dynamics, variable transformations have long been recognized as providing useful alternative representations, such as choosing particular variables to enhance stability properties (Hughes et al. 1986, Kalashnikova & Barone 2011, Balajewicz et al. 2016, Rezaian & Wei 2020). As another classical example, the well-known Cole–Hopf transformation turns the nonlinear Burgers equation into a linear equation (Hopf 1950, Cole 1951). In the dynamical systems field, DMD (Rowley et al.

2009, Schmid 2010) is often used to learn low-dimensional models from data. Employing variable transformations to use a different choice of variables (called observables) enables Koopman analysis via extended DMD (Williams et al. 2015, Netto et al. 2021), which leads to more accurate DMD models.

In controller design, feedback linearization uses a nonlinear state transformation to bring a general nonlinear system into a structured linear model (Jakubczyk & Respondek 1980, Khalil 2002) that can then be controlled with classical methods. Bringing nonlinear systems into canonical and abstract forms can further improve their numerical solution, analysis, and verification, as shown by Savageau & Voit (1987), Liu et al. (2015), Brenig (2018), and Guillot et al. (2019).

Quadratic model structure has garnered broad interest in model reduction due to the advantages of analyzing and simulating quadratic models (versus other nonlinear models). In the context of optimization, McCormick (1976) is credited with introducing variable substitutions to achieve quadratic structure so that nonconvex optimization problems can be recast as convex problems in the new variables. In the field of model reduction, lifting to a quadratic form was introduced by Gu (2011) and subsequently developed further for model reduction methods that are tailored to quadratic model form (Benner & Breiten 2015; Benner et al. 2018; Kramer & Willcox 2019, 2022; Liljegren-Sailer & Marheineke 2022). In the context of Operator Inference, the Lift & Learn method introduced by Qian et al. (2020) and in related research (Swischuk et al. 2020a, McQuarrie et al. 2021, Qian et al. 2022) uses lifting transformations to learn quadratic reduced models approximating complex nonlinear systems, such as combustion dynamics, from lifted data. The quadratic structure can be further exploited to equip these learned models with stability guarantees (Kramer 2021). These stability guarantees can also be derived and integrated into Operator Inference in the case of a model with cubic structure (as in Sawant et al. 2023).

How are these lifting transformations identified? In most cases, they are readily identified manually from the form of the governing PDEs or ordinary differential equations (see the sidebar titled Lifting a Nonpolynomial System). There is active research in algorithms, methods, and software for polynomialization (Hemery et al. 2021) and quadratization (Hemery et al. 2020, Bychkov & Pogudin 2021). In some cases, the lifting transformations result in differential algebraic equations, such as for the additive manufacturing solidification example presented by Khodabakhshi & Willcox (2022).

LIFTING A NONPOLYNOMIAL SYSTEM

Consider a simple ordinary differential equation with a nonpolynomial term and a linear term:

$$\dot{x}(t) = ax(t) + e^{-x(t)}.$$

Set $x_1 = x$ and take $x_2 = e^{-x_1}$ as the auxiliary variable, so that $\dot{x}_2 = -x_2(ax_1 + x_2) = -ax_1x_2 - x_2^2$. The lifted system is now in quadratic form:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} a & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -a & -1 \end{bmatrix} \left(\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \otimes \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \right).$$

Operator Inference does not simulate the lifted system but instead applies the lifting transformations to the original state snapshots that have been collected. The lifted snapshots are then used to assemble the data matrix $\boldsymbol{X}$. This postprocessing step generates snapshots of the lifted state $\boldsymbol{x} = [x_1,\ x_2]^\top$, from which Operator Inference learns a (here quadratic) model in the lifted variables.

In summary, variable transformations and lifting expose polynomial structure in a broad class of nonlinear systems. In the next subsections, we discuss how this polynomial structure is highly amenable to surrogate modeling via model reduction. Note that while the variable transformations are derived from the form of the governing PDEs, the transformations are not applied to the high-fidelity numerical model but rather to the data from which the reduced model is learned. As a result, variable transformations and lifting are a powerful yet broadly applicable set of tools for approximating complex nonlinear systems.

2.2. Nonintrusive Model Reduction with Operator Inference

Model reduction differs from other surrogate modeling approaches in that the reduced model explicitly accounts for the structure of the governing equations. Operator Inference nonintrusively constructs reduced models via a data-driven regression problem that learns reduced matrices from snapshot data. The data-driven nature of Operator Inference enables the use of variable transformations to expose structure in nonlinear systems.

2.2.1. Surrogate modeling via model reduction. Model reduction typically consists of two phases. In the offline (training) phase, training data are generated, low-dimensional structure is identified, and the reduced models are constructed. In the online (evaluation/deployment) phase, the reduced models are used to make rapid predictions, including at new initial conditions and parameter values that were not sampled during the offline phase. A wide range of model reduction techniques are surveyed by, for instance, Antoulas (2005), Rozza et al. (2008), Benner et al. (2015), Hesthaven et al. (2016), and Antoulas et al. (2021).

We focus on snapshot-based model reduction methods, which identify a low-dimensional coordinate system based on the analysis of sampled state solutions. We denote a trajectory of state solutions generated by solving Equation 1 for a parameter $\boldsymbol{\mu}$ as

$$\boldsymbol{X}(\boldsymbol{\mu}) = [\boldsymbol{x}(t_1; \boldsymbol{\mu}), \ldots, \boldsymbol{x}(t_K; \boldsymbol{\mu})] \in \mathbb{R}^{N \times K}, \tag{4.}$$

with the initial condition $\boldsymbol{x}(0; \boldsymbol{\mu})$. The number of time steps in each trajectory is K and the time steps are $0 = t_0 < t_1 < t_2 < \cdots < t_K$. We assume the time steps to be equidistant to ease exposition, with constant time-step size $\delta t > 0$, but this is not a requirement. The state solution $\boldsymbol{x}(t_j; \boldsymbol{\mu})$ is referred to as the jth snapshot in the trajectory $\boldsymbol{X}(\boldsymbol{\mu})$. Snapshots are generated for multiple trajectories $\boldsymbol{X}(\boldsymbol{\mu}_1), \ldots, \boldsymbol{X}(\boldsymbol{\mu}_M)$ by solving Equation 1 for training parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M$ and corresponding initial conditions.

2.2.2. Operator Inference basic algorithm. Operator Inference follows three steps. It is distinctive from other model reduction methods in Step 3, where the reduced model is constructed by solving a data-driven regression problem that learns reduced-order matrices from snapshot data.

Step 1: snapshot generation. Trajectories such as those given in Equation 4 are collected and then concatenated together with the initial conditions into a snapshot matrix:

$$\boldsymbol{X} = [\boldsymbol{x}(0; \boldsymbol{\mu}_1), \boldsymbol{X}(\boldsymbol{\mu}_1), \ldots, \boldsymbol{x}(0; \boldsymbol{\mu}_M), \boldsymbol{X}(\boldsymbol{\mu}_M)] \in \mathbb{R}^{N \times M(K+1)}. \tag{5.}$$

If lifting or variable transformations are employed, the transformations are applied to each snapshot. The snapshots are then concatenated together with the initial conditions (also transformed into the appropriate variables) into a snapshot matrix. To keep the exposition simple, we continue to use the symbol $\boldsymbol{X}$ to denote the snapshot matrix, but note that in some cases the physical variables contained within the snapshots may be transformed and differ from the native variables of the original nonlinear system in Equation 1.

Step 2: constructing a low-dimensional basis. A common approach is to apply proper orthogonal decomposition (POD), which has a long tradition in fluid mechanics (Sirovich 1987,

Holmes et al. 1996). Computing the POD entails constructing the first $n \ll N$ left-singular vectors of $\boldsymbol{X}$ corresponding to the largest singular values, and then collecting these singular vectors as columns in a basis matrix: $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]$. This basis is orthonormal, $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$, and spans an n-dimensional subspace $\mathcal{V} \subseteq \mathbb{R}^N$. In practice, the snapshot matrix is often first centered (by subtracting the snapshot mean from each snapshot), and possibly scaled, before the singular vectors are computed. There are various heuristics to choose the reduced dimension n; the most common are based on decay of the singular values (Benner et al. 2015). The basis matrix $\boldsymbol{V}$ defines a low-dimensional coordinate system in which we now construct a reduced model. A full-order state is approximated as $\boldsymbol{x}(t; \boldsymbol{\mu}) \approx \boldsymbol{V}\hat{\boldsymbol{x}}(t; \boldsymbol{\mu})$, where $\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}) \in \mathbb{R}^n$ is the reduced state of dimension n.

Reduced state: a full-order state $\boldsymbol{x}(t; \boldsymbol{\mu}) \in \mathbb{R}^N$ is approximated as $\boldsymbol{x}(t; \boldsymbol{\mu}) \approx \boldsymbol{V}\hat{\boldsymbol{x}}(t; \boldsymbol{\mu})$, where the columns of $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]$ form the POD basis of order n and $\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}) \in \mathbb{R}^n$ is the reduced state of dimension n

Step 3: reduced model construction via Operator Inference. The reduced model of Equation 1 takes the general form

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}) = \hat{\boldsymbol{f}}(\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}), \boldsymbol{u}(t); \boldsymbol{\mu}), \tag{6.}$$

with the reduced-order dynamics described by the function $\hat{\boldsymbol{f}} : \mathbb{R}^n \times \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^n$. As discussed above, we consider systems where the governing equations admit the polynomial form of Equation 3. The corresponding form of the reduced model is then

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}) = \sum_{i=1}^{\ell} \hat{\boldsymbol{A}}_i(\boldsymbol{\mu})\hat{\boldsymbol{x}}^i(t; \boldsymbol{\mu}) + \hat{\boldsymbol{B}}(\boldsymbol{\mu})\boldsymbol{u}(t), \tag{7.}$$

where, following the notation of Equation 3, the vector $\hat{\boldsymbol{x}}^i \in \mathbb{R}^{n_i}$ contains the unique components of the i-times Kronecker product $\hat{\boldsymbol{x}}(t; \boldsymbol{\mu}) \otimes \cdots \otimes \hat{\boldsymbol{x}}(t; \boldsymbol{\mu})$, with $n_i = \binom{n+i-1}{i}$. The reduced-order matrices $\hat{\boldsymbol{A}}_i(\boldsymbol{\mu})$ and $\hat{\boldsymbol{B}}(\boldsymbol{\mu})$ will be learned from snapshot data using Operator Inference, as described in the next three substeps.

Reduced model construction via Operator Inference: solves a linear least-squares problem to determine the reduced model operators that best fit the reduced snapshot trajectories $\breve{\boldsymbol{X}}(\boldsymbol{\mu}_j)$ in a minimum residual sense

Step 3a: dimension reduction. Project the (possibly lifted) trajectories $\boldsymbol{X}$ onto the reduced space $\mathcal{V}$ via

$$\breve{\boldsymbol{X}}(\boldsymbol{\mu}_j) = \boldsymbol{V}^\top \boldsymbol{X}(\boldsymbol{\mu}_j), \qquad j = 1, \ldots, M, \tag{8.}$$

where $\breve{\boldsymbol{X}}(\boldsymbol{\mu}_j) = [\breve{\boldsymbol{x}}(t_1; \boldsymbol{\mu}_j), \ldots, \breve{\boldsymbol{x}}(t_K; \boldsymbol{\mu}_j)]$ are the projected trajectories comprising snapshots of reduced dimension n. Note that we use the notation $\breve{\boldsymbol{x}} \in \mathbb{R}^n$ to denote a projected snapshot (i.e., $\breve{\boldsymbol{x}} = \boldsymbol{V}^\top \boldsymbol{x}$) in contrast to the notation $\hat{\boldsymbol{x}} \in \mathbb{R}^n$, which denotes a reduced state computed by solving a reduced model. The difference between the two becomes important in the theoretical analysis in Section 4. Also, compute the projected initial conditions, $\breve{\boldsymbol{x}}(0; \boldsymbol{\mu}_j) = \boldsymbol{V}^\top \boldsymbol{x}(0; \boldsymbol{\mu}_j)$. Additionally, approximate the time derivatives (if they are not given) to obtain, for instance, with fourth-order finite differences and time-step size δt,

$$\breve{\boldsymbol{x}}'(t_k; \boldsymbol{\mu}_j) = \frac{1}{12\delta t}\left(-\breve{\boldsymbol{x}}(t_{k-2}; \boldsymbol{\mu}_j) + 8\breve{\boldsymbol{x}}(t_{k-1}; \boldsymbol{\mu}_j) - 8\breve{\boldsymbol{x}}(t_{k+1}; \boldsymbol{\mu}_j) + \breve{\boldsymbol{x}}(t_{k+2}; \boldsymbol{\mu}_j)\right),$$

with $j = 1, \ldots, M$ and $k = 2, \ldots, K-2$. Here, $\breve{\boldsymbol{x}}'(t_k; \boldsymbol{\mu}_j)$ approximates the time derivative $(\mathrm{d}/\mathrm{d}t)\breve{\boldsymbol{x}}(t_k; \boldsymbol{\mu}_j)$. The approximate time derivative $\breve{\boldsymbol{x}}'(t_k; \boldsymbol{\mu}_j)$ can be computed with wide finite-difference stencils that involve many neighbors, corresponding to higher-order accuracy, because snapshots at all time steps are available.

Dimension reduction: a trajectory of snapshots $\boldsymbol{X}(\boldsymbol{\mu}_j)$ is represented in the reduced subspace as $\breve{\boldsymbol{X}}(\boldsymbol{\mu}_j) = \boldsymbol{V}^\top \boldsymbol{X}(\boldsymbol{\mu}_j)$

Step 3b: learning. For each training parameter $\boldsymbol{\mu}_j$ with $j = 1, \ldots, M$, Operator Inference fits the reduced matrices $\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j), \ldots, \hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j)$ and $\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)$ by minimizing the objective

$$J_j(\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j), \ldots, \hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j), \hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)) = \sum_{k=1}^{K} \left\| \sum_{i=1}^{\ell} \hat{\boldsymbol{A}}_i(\boldsymbol{\mu}_j)\breve{\boldsymbol{x}}^i(t_k; \boldsymbol{\mu}_j) + \hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)\boldsymbol{u}(t_k; \boldsymbol{\mu}_j) - \breve{\boldsymbol{x}}'(t_k; \boldsymbol{\mu}_j) \right\|_2^2, \tag{9.}$$

as in the optimization problem

$$\min_{\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j),\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)} J_j(\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j),\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j))\,. \tag{10.}$$

The optimization problem stated in Equation 10 is a linear least-squares problem that can be solved efficiently, as discussed in detail in Section 2.2.4.

Step 3c: reduced model assembly. For any of the training parameters $\boldsymbol{\mu}_j$ with $j = 1,\ldots,M$, the learned operators $\hat{\boldsymbol{A}}_i(\boldsymbol{\mu}_j)$, $i = 1,\ldots,\ell$, and $\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)$ define a reduced model that can then be used to issue predictions at new, unseen initial conditions that have not been used for training. For a different parameter $\boldsymbol{\mu} \in \mathcal{D} \setminus \{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_M\}$ that is not in the training set, the operators $\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu})$, and $\boldsymbol{B}(\boldsymbol{\mu})$ can be obtained via interpolation between the operators computed in Step 3b. In particular, the interpolation can be performed on matrix manifolds to preserve the structure of the inferred matrices in the interpolated matrices (Amsallem & Farhat 2008, Degroote et al. 2010, Panzer et al. 2010).

Regularization: adds a bias that guides the learning process to compensate for too few and noisy data samples as well as model misspecifications

2.2.3. Regularization and stability of Operator Inference models.

Regularization imposes a bias on the learning process to guide Operator Inference toward meaningful models that have predictive capabilities and that generalize well to unseen parameters, inputs, and initial conditions. Regularization is especially helpful when models are misspecified. Also, data are often polluted with numerical noise due to early stopping of iterative solvers, limited numerical precision, and other perturbations.

Tikhonov regularization has been proposed to enhance Operator Inference by preventing overfitting (Swischuk et al. 2020a, Jain et al. 2021, McQuarrie et al. 2021). Tikhonov regularization suggests modifying the optimization problem given in Equation 10 as

$$\min_{\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j),\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)} J_j(\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j),\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)) + \sum_{i=1}^{\ell} \lambda_i \|\hat{\boldsymbol{A}}_i(\boldsymbol{\mu}_j)\|_F^2 + \lambda_{\ell+1}\|\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)\|_F^2. \tag{11.}$$

There are several techniques to choose regularization coefficients $\lambda_1, \lambda_2,\ldots,\lambda_{\ell+1}$ in the context of Operator Inference (Swischuk et al. 2020a, McQuarrie et al. 2021). McQuarrie et al. (2021) determine the regularization coefficients by both a data-fit criterion during training and a state constraint that additionally enforces stability in the training and testing set. Regularizers may often be tailored to the physics of the problem. For example, bounds on the stability radius of polynomial dynamical-system models can be derived on the basis of Lyapunov stability theory (e.g., Tesi et al. 1994, Chesi 2007, Kramer 2021). Such bounds are computable in case of polynomial models and depend on the norms of the learned operators $\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j)$. This was the motivation for Sawant et al. (2023) to propose regularizers that penalize operators $\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j)$ with large norms to encourage learning models with large stability bounds. One of the key insights gained by applying Lyapunov stability theory to models learned with Operator Inference is that it is beneficial in terms of stability to regularize the quadratic and higher-order terms only and does not impose a regularizer on the linear term (for details, see Sawant et al. 2023).

2.2.4. Scalability of Operator Inference.

For any of the training parameters $\boldsymbol{\mu}_j$, $j = 1,\ldots,M$, the Operator Inference optimization problem given in Equation 11 can be written in the form

$$\min_{\boldsymbol{O}(\boldsymbol{\mu}_j)} \|\boldsymbol{D}(\boldsymbol{\mu}_j)\boldsymbol{O}(\boldsymbol{\mu}_j) - \boldsymbol{R}(\boldsymbol{\mu}_j)\|_F^2 + \|\boldsymbol{\Gamma}(\boldsymbol{\mu}_j)\boldsymbol{O}(\boldsymbol{\mu}_j)\|_F^2, \tag{12.}$$

where the inferred operators are given as blocks in the matrix $\boldsymbol{O}(\boldsymbol{\mu}_j) = [\hat{\boldsymbol{A}}_1(\boldsymbol{\mu}_j),\ldots,\hat{\boldsymbol{A}}_\ell(\boldsymbol{\mu}_j),\hat{\boldsymbol{B}}(\boldsymbol{\mu}_j)]^\top$, which has $\bar{n} = \sum_{i=1}^{\ell} n_i + d$ rows and n columns. The data matrix is

$$\boldsymbol{D}(\boldsymbol{\mu}_j) = [\check{\boldsymbol{X}}(\boldsymbol{\mu}_j), \check{\boldsymbol{X}}^2(\boldsymbol{\mu}_j),\ldots,\check{\boldsymbol{X}}^\ell(\boldsymbol{\mu}_j), \boldsymbol{U}(\boldsymbol{\mu}_j)]^\top \in \mathbb{R}^{K\times\bar{n}}, \tag{13.}$$

A PYTHON PACKAGE FOR OPERATOR INFERENCE

Operator Inference for learning polynomial reduced models of dynamical systems is available as a Pypi Python package (see **https://pypi.org/project/opinf**) under the MIT License. The documentation for the package includes a discussion of all functionalities such as regularization and postprocessing features, tutorials, an installation guide, and references.

where the matrix $\check{\boldsymbol{X}}^{i}(\boldsymbol{\mu}_j) = [\check{\boldsymbol{x}}^{i}(t_1;\boldsymbol{\mu}_j),\ldots,\check{\boldsymbol{x}}^{i}(t_K;\boldsymbol{\mu}_j)] \in \mathbb{R}^{n_i \times K}$ for $i = 1,\ldots,\ell$ collects the unique components of i-times Kronecker products of the projected snapshots and the matrix $\boldsymbol{U}(\boldsymbol{\mu}_j) = [\boldsymbol{u}(t_1;\boldsymbol{\mu}_j),\ldots,\boldsymbol{u}(t_K;\boldsymbol{\mu}_j)] \in \mathbb{R}^{d \times K}$ collects the input trajectories that were used to generate the snapshots. The right-hand-side matrix, $\boldsymbol{R}(\boldsymbol{\mu}_j) = [\check{\boldsymbol{x}}'(t_1;\boldsymbol{\mu}_j),\ldots,\check{\boldsymbol{x}}'(t_K;\boldsymbol{\mu}_j)]^{\top} \in \mathbb{R}^{K \times n}$, collects the approximate time derivatives of the projected snapshots. The matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu}_j) \in \mathbb{R}^{\bar{n} \times \bar{n}}$ corresponds to the regularization weights of the regularized problem given in Equation 11. It has zero entries only if no regularization is applied.

Each step of Operator Inference is scalable. The reduction steps (Steps 2 and 3a) typically entail computing a singular value decomposition, for which highly efficient and parallel implementations exist. Furthermore, studies have shown that randomized methods can scale the reduction step to large data sets (e.g., Swischuk et al. 2020a; Farcaş et al. 2022, 2023a). Randomized methods typically rely on a low-rank structure, which can reasonably be expected to be present in trajectory data when learning low-dimensional models is the goal. The learning step (Step 3b) solves the optimization problem defined in Equation 12, which can be decomposed into n independent linear least-squares problems with $\bar{n}$ unknowns each. Each of the independent problems can be solved in parallel. The optimization problem is a standard least-squares problem, which can again be solved with scalable implementations of the singular value decomposition. The assembly step (Step 3c) assembles the reduced model based on the inferred operators, which are low dimensional; therefore, computational costs are typically low. A scalable and fast implementation of operator inference is available in Python (see the sidebar titled A Python Package for Operator Inference).

3. OPERATOR INFERENCE IN ACTION

We now walk through the computational procedure of applying Operator Inference to learn a surrogate model of a large-scale combustion simulation. The results that we show are drawn from Qian et al. (2022).

3.1. Numerical Model of the Continuously Variable Resonance Combustor

The Continuously Variable Resonance Combustor (CVRC) is an experimental combustor at Purdue University (Harvazinski et al. 2015). We consider a numerical model of the CVRC with three spatial dimensions, as used by Qian et al. (2022), and refer the reader to Wang et al. (2019), Peherstorfer (2020a), Swischuk et al. (2020a), and McQuarrie et al. (2021) for other surrogate-modeling techniques applied to one- and two-dimensional models of the CVRC. Models of the CVRC can be simulated with the General Equation and Mesh Solver (GEMS) code introduced by Harvazinski et al. (2015). Details about the GEMS setup and reduced modeling for the CVRC are given by Huang et al. (2019, 2020).

The combustor is cylindrically symmetric around the first coordinate direction and approximately 28 cm in the axial direction (Harvazinski et al. 2015). A forcing in the form of a 10% fluctuation in the back pressure at the combustor outlet drives the dynamics. The forcing frequency is 2,000 Hz, and the baseline back pressure is 1.1 MPa. The governing equations are

the three-dimensional Navier–Stokes equations with a flamelet/progress variable chemical model. The GEMS model contains seven state variables, $\boldsymbol{w} = [\rho, \rho u_1, \rho u_2, \rho u_3, \rho e, \rho Z_\mathrm{m}, \rho C]^\top$, where ρ is the density; u_1, u_2, and u_3 are the velocities in the three spatial directions; and e is the specific energy. The variables Z_m and C are the mixture mean and reaction progress variable, respectively, of the chemical model. The governing equations in conservative form are written as

$$\frac{\partial}{\partial t}\boldsymbol{w}(t,\boldsymbol{\xi}) + \nabla \cdot (\boldsymbol{F}(t,\boldsymbol{\xi},\boldsymbol{w}) - \boldsymbol{F}_v(t,\boldsymbol{\xi},\boldsymbol{w})) = \boldsymbol{s}(t,\boldsymbol{\xi},\boldsymbol{w}), \qquad 14.$$

where $\boldsymbol{\xi} \in \mathbb{R}^3$ is the spatial coordinate; t is time; and $\boldsymbol{F}$ and $\boldsymbol{F}_v$ are the inviscid and viscous flux, respectively. The function $\boldsymbol{s}$ is the source term (for details, see Qian et al. 2022). The GEMS code uses the finite-volume method to discretize Equation 14 into a system of ordinary differential equations, as given in Equation 1, with state dimension $N \approx 18.5 \times 10^6$ (i.e., there are approximately 18.5 million equations and unknowns in the large-scale CFD model). The scenario considered simulates 5 ms of combustion dynamics, which takes more than 45,000 CPU hours with GEMS.

3.2. Applying Operator Inference to Learn a Reduced Model of the Continuously Variable Resonance Combustor

In the Qian et al. (2022) paper, Operator Inference learns a reduced model from $K = 3{,}000$ simulated snapshots collected over a time interval from $t = 15$ ms to $t = 17.999$ ms. The learned reduced models are used to issue predictions a further 2 ms beyond the training horizon, until $t = 20$ ms. For this example, no parametric variation is considered, so the reduced model does not have a functional dependence on $\boldsymbol{\mu}$.

Step 1: snapshot generation. GEMS is simulated over the training horizon to generate a trajectory as in Equation 4, comprising $K = 3{,}000$ snapshots. The solutions output by GEMS contain (at each spatial location in the CFD mesh) the primitive flow variables ρ, u_1, u_2, u_3, and p (where p is pressure), the flamelet/progress variables Z_m and C, the temperature T, and the enthalpy. There are many nonlinear terms in the governing equations given in Equation 14; therefore, a variable lifting, as outlined in Section 2.1.3, is applied to the snapshots. The lifting transforms many—but not all—nonlinear terms into quadratic form. The lifting transformation is inspired by the quadratic representation of the compressible Euler equations in the specific volume variables, as described by Qian et al. (2020). Specifically, the GEMS snapshots are postprocessed to the lifted variables $\zeta = 1/\rho$, u_1, u_2, u_3, p, ρZ_m, ρC, and T. Note that the lifting is applied to the snapshots obtained from simulating Equation 14 in the original variables. Thus, the lifting is nonintrusive and does not require modification of the numerical solver.

Step 2: constructing a low-dimensional basis. After centering and scaling the lifted snapshots, POD is applied to construct a basis. The singular values $\sigma_1, \ldots, \sigma_{3{,}000}$ of the scaled snapshots indicate the relative fraction of the energy retained by the n-dimensional POD basis as $\eta_n = 1 - \sum_{i=n+1}^{3{,}000} \sigma_i^2 / \sum_{i=1}^{3{,}000} \sigma_i^2$ (**Figure 1*a***). This metric provides an empirical heuristic for determining the number of basis vectors n. We keep up to $n = 100$ basis vectors and thus retain up to 95% of the snapshot energy. Below, we consider POD bases with dimensions $n \in \{50, 75, 100\}$ to demonstrate how varying the dimension affects the reduced model prediction errors.

Step 3a: dimension reduction. The lifted snapshots are projected onto the reduced space defined by the POD basis as in Equation 8.

Step 3b: learning. After the lifting, most of the variables enter linearly and quadratically. Therefore, we learn a polynomial reduced model as in Equation 7, with $\ell = 2$ and with an additional term $\hat{\boldsymbol{A}}_0 \in \mathbb{R}^{n \times 1}$ that is constant in the state variables:

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{x}}(t) = \hat{\boldsymbol{A}}_1\hat{\boldsymbol{x}}(t) + \hat{\boldsymbol{A}}_2\hat{\boldsymbol{x}}^2(t) + \hat{\boldsymbol{A}}_0 + \hat{\boldsymbol{B}}\boldsymbol{u}(t).$$

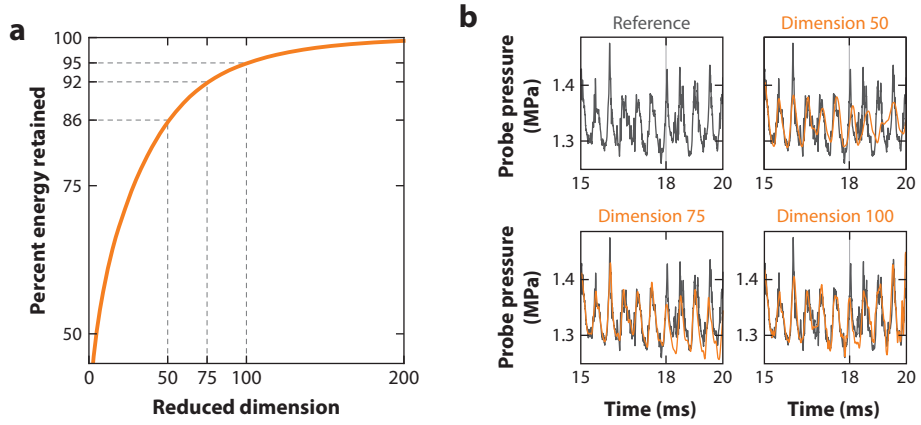

Figure 1

(*a*) The percent of energy retained by POD modes provides a heuristic for selecting the dimension n of Operator Inference models. We keep up to $n = 100$ basis vectors to retain up to 95% of the energy. (*b*) The learned Operator Inference models accurately predict the pressure within only a few seconds of simulation run time, whereas the reference numerical model solved with GEMS requires tens of thousands of CPU hours. The vertical black line indicates where the training horizon ends and the prediction horizon begins. Abbreviations: GEMS, General Equation and Mesh Solver; POD, proper orthogonal decomposition. Figure adapted from Qian et al. (2022) with permission from the authors.

Here, the input $\boldsymbol{u}(t)$ corresponds to the back pressure forcing. Tikhonov regularization (as discussed in Section 2.2.3) is applied, leading to an optimization problem, as given in Equation 11. The regularization coefficients are selected via sampling such that the prediction error over the training period is minimized. The optimization problem can be decomposed into n independent least-squares problems and efficiently solved with scalable numerical linear algebra packages (as discussed in Section 2.2.4).

Step 3c: reduced model assembly. Once the model operators are inferred, they are assembled into a reduced model, which can then be integrated forward in time with a second-order Runge–Kutta time-stepping scheme to predict the combustion dynamics.

3.3. Predictive Capabilities of Operator Inference Models on the Continuously Variable Resonance Combustor Example

Figure 1*b* shows the pressure at a probe obtained with the learned models for dimensions $n = 50$, 75, and 100. The probe is located at the downstream combustor boundary. The predicted pressure is in good agreement in terms of frequency and amplitude with the reference obtained from the high-dimensional numerical model solved with GEMS. The error in the prediction is around 2–3%. **Figure 2** shows spatial fields of the reaction progress variable C. The Operator Inference models have orders-of-magnitude-fewer degrees of freedom than the reference numerical model and thus smear out some of the fine-scale features, but they provide reasonable predictions of coarse flow features. These results also highlight some of the difficulties in achieving accurate surrogate-based predictions for complex fluid flows, especially when the dynamics are transport dominated, as they are here.

This example demonstrates the potential of Operator Inference in learning predictive models of complex systems. The high-dimensional GEMS numerical model with which snapshots are generated has more than 18 million unknowns and requires tens of thousands of CPU hours to simulate. The simulation run time of the learned Operator Inference model is only a few seconds,

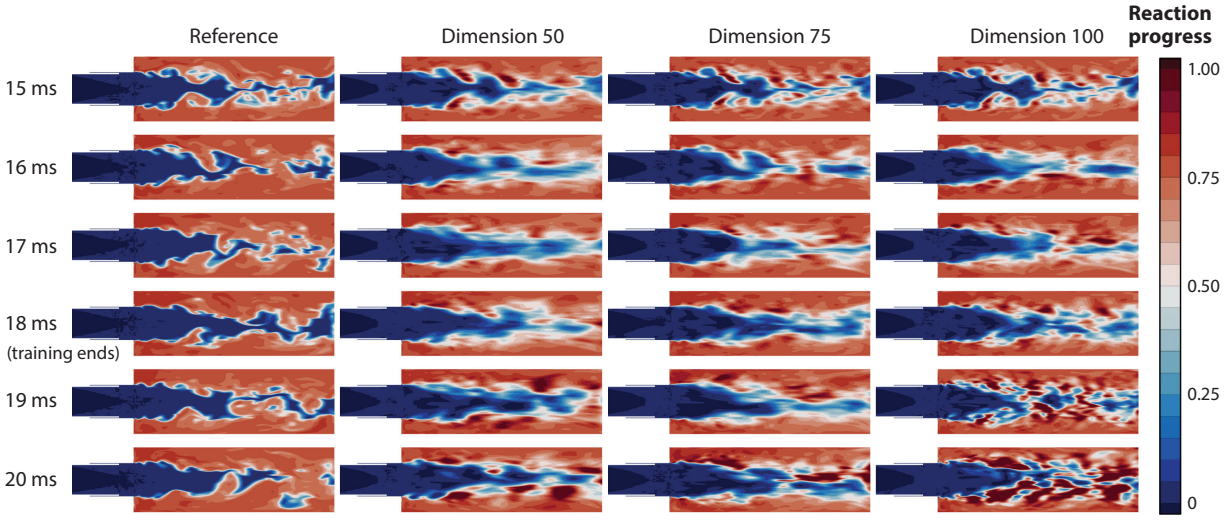

Figure 2

The reaction progress variable (C) field of the CVRC, as predicted by GEMS ($N = 18,500,000$ degrees of freedom) and Operator Inference reduced models ($n = 50, 75, 100$). The Operator Inference models provide model run time reduction of up to nine orders of magnitude. They smear out some of the fine-scale features but provide accurate predictions of the coarser flow features. Abbreviations: CVRC, Continuously Variable Resonance Combustor; GEMS, General Equation and Mesh Solver. Figure adapted from Qian et al. (2022) with permission from the authors.

but it provides sufficiently accurate predictions for use in outer-loop applications such as design and uncertainty quantification.

4. PREDICTION GUARANTEES FOR STRUCTURED OPERATOR INFERENCE MODELS

We now show that some of the rich theory of intrusive model reduction can be applied to models learned from data with Operator Inference. For ease of exposition, in this section we drop the dependence on the parameter $\boldsymbol{\mu}$ and the input $\boldsymbol{u}(t)$.

4.1. Recovering Projection-Based Reduced Models with Reprojection Schemes

Data sampling with reprojection (Peherstorfer 2020b) collects state trajectories of the high-dimensional model that are equivalent to the trajectories of reduced models obtained with intrusive model reduction. Learning from reprojected trajectories with Operator Inference guarantees recovery of the very same models that would be obtained with intrusive model reduction, under certain assumptions. This recovery of the intrusively reduced models allows one to carry over theoretical results from intrusive model reduction to data-driven modeling with Operator Inference, including a posteriori error estimation for certifying predictions.

4.1.1. Closure error and Markovian dynamics. Recall the numerical model given in Equation 3 and a trajectory $\boldsymbol{X}$ as defined in Equation 4. Given a basis matrix $\boldsymbol{V}$ that spans a space $\mathcal{V}$ of dimension n, intrusive model reduction provides the Galerkin reduced model $(\mathrm{d}/\mathrm{d}t)\tilde{\boldsymbol{x}}(t) = \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}(t))$, where $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}(t)) = \boldsymbol{V}^\top \boldsymbol{f}(\boldsymbol{V}\tilde{\boldsymbol{x}}(t))$ (Benner et al. 2015). Note that the learned model given in Equation 6 with state $\hat{\boldsymbol{x}}(t)$ is not necessarily the Galerkin reduced model. We denote a trajectory of the Galerkin reduced model as $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}(t_1), \ldots, \tilde{\boldsymbol{x}}(t_K)] \in \mathbb{R}^{n \times K}$, where the kth column is the state

$\tilde{\boldsymbol{x}}(t_k)$ of the Galerkin reduced model at time step t_k. Notice that the state of the Galerkin reduced model $\tilde{\boldsymbol{x}}(t_k)$ is different from the projected state $\check{\boldsymbol{x}}(t_k) = \boldsymbol{V}^\top \boldsymbol{x}(t_k)$; below, we refer to the difference, $\tilde{\boldsymbol{x}}(t_k) - \check{\boldsymbol{x}}(t_k)$, as the closure error.

We demonstrate the difference between the projected state trajectories and the trajectories obtained with intrusive Galerkin reduced models on an example with an autonomous linear model, which means we consider the model given in Equation 3 with $\ell = 1$ and no input, so that $(\mathrm{d}/\mathrm{d}t)\boldsymbol{x}(t) = \boldsymbol{A}_1\boldsymbol{x}(t)$. We split the space $\mathbb{R}^N$ into $\mathcal{V}$, which is spanned by the columns of the orthonormal basis matrix $\boldsymbol{V}$, and its orthogonal complement $\mathcal{V}^\perp$ with orthonormal basis matrix $\boldsymbol{V}^\perp$, so that $\mathbb{R}^N = \mathcal{V} \oplus \mathcal{V}^\perp$. Correspondingly, we split the representation of a state $\boldsymbol{x}(t)$ into $\boldsymbol{x}(t) = \boldsymbol{V}\check{\boldsymbol{x}}(t) + \boldsymbol{V}^\perp\boldsymbol{x}^\perp(t)$ with the projected state $\check{\boldsymbol{x}}(t) = \boldsymbol{V}^\top\boldsymbol{x}(t)$ and its orthogonal complement $\boldsymbol{x}^\perp(t) = (\boldsymbol{V}^\perp)^\top\boldsymbol{x}(t)$. After transformations described in detail by Peherstorfer (2020b), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\check{\boldsymbol{x}}(t) = \underbrace{\boldsymbol{A}_1^{\parallel\parallel}\check{\boldsymbol{x}}(t)}_{\text{Markovian term}} + \underbrace{\int_0^t \boldsymbol{A}_1^{\parallel\perp} \mathrm{e}^{(t-s)\boldsymbol{A}_1^{\perp\perp}} \boldsymbol{A}_1^{\perp\parallel}\check{\boldsymbol{x}}(s)\,\mathrm{d}s}_{\text{non-Markovian term}} + \boldsymbol{A}_1^{\parallel\perp}\mathrm{e}^{t\boldsymbol{A}_1^{\perp\perp}}\boldsymbol{x}^\perp(0), \qquad 15.$$

with the matrices $\boldsymbol{A}_1^{\parallel\parallel} = \boldsymbol{V}^\top\boldsymbol{A}_1\boldsymbol{V}, \boldsymbol{A}_1^{\parallel\perp} = \boldsymbol{V}^\top\boldsymbol{A}_1\boldsymbol{V}^\perp, \boldsymbol{A}_1^{\perp\parallel} = (\boldsymbol{V}^\perp)^\top\boldsymbol{A}_1\boldsymbol{V}$, and $\boldsymbol{A}_1^{\perp\perp} = (\boldsymbol{V}^\perp)^\top\boldsymbol{A}_1\boldsymbol{V}^\perp$. Equation 15 shows that the projected state $\check{\boldsymbol{x}}(t)$ corresponds to non-Markovian dynamics in the sense that its dynamics depend on $\check{\boldsymbol{x}}(t)$ as well as on the states of all earlier times than t via the non-Markovian (memory) term. This insight is well known under the Mori–Zwanzig formalism (Givon et al. 2004, Chorin & Stinis 2006). Thus, Operator Inference as formulated in Section 2.2.2 fits a Markovian model to data that represent non-Markovian dynamics. This works well in many cases but can prevent recovery guarantees for dimensions $n < N$.

4.1.2. Sampling Markovian dynamics with reprojection for recovering intrusive reduced models from data. Peherstorfer (2020b) introduced a sampling scheme to generate trajectories from numerical simulations with a model with high-dimensional states so that the projected state trajectories describe Markovian dynamics in reduced spaces. **Figure 3** depicts the reprojection scheme, which alternates between querying the high-dimensional model and projecting the collected state at the current time step onto $\mathcal{V}$. This process leads to the reprojected trajectory $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}(t_1), \ldots, \bar{\boldsymbol{x}}(t_K)]$, which is the same trajectory that a traditional Galerkin reduced model from intrusive model reduction generates. Thus, reprojected trajectories can be described with zero residual by the Markovian models we seek in Operator Inference, as given in Equation 7. Consider the case in which the model has polynomial structure, as in Equation 3. If sufficiently many reprojected states and their time derivatives are collected, such that the data matrix defined in Equation 13 has full rank, then fitting an Operator Inference model to reprojected trajectories recovers the reduced model that is obtained via intrusive projection-based model reduction. As stated formally and in more detail by Peherstorfer (2020b), if the data matrix $\boldsymbol{D}$ defined in

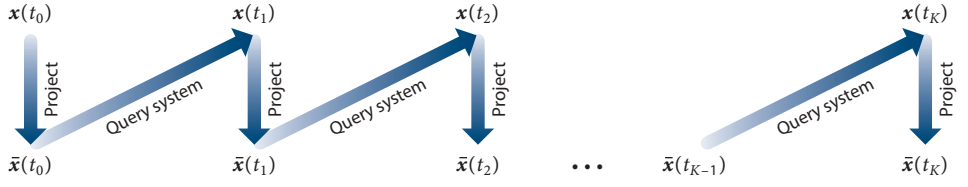

Figure 3

Data sampling with reprojection alternates between collecting data and projecting data onto the subspace $\mathcal{V}$ to generate a state trajectory that describes Markovian dynamics in $\mathcal{V}$. Fitting an Operator Inference model to reprojected trajectories recovers the same models that are obtained with intrusive model reduction, under mild conditions.

MOTIVATING THE STRUCTURAL FORM OF THE LEARNED REDUCED MODEL

Why is it that the high-dimensional numerical model of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t;\boldsymbol{\mu}) = \sum_{i=1}^{\ell} \boldsymbol{A}_i(\boldsymbol{\mu})\boldsymbol{x}^i(t;\boldsymbol{\mu}) + \boldsymbol{B}(\boldsymbol{\mu})\boldsymbol{u}(t) \qquad \text{SB1.}$$

leads us to learn a reduced-order model of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{x}}(t;\boldsymbol{\mu}) = \sum_{i=1}^{\ell} \hat{\boldsymbol{A}}_i(\boldsymbol{\mu})\hat{\boldsymbol{x}}^i(t;\boldsymbol{\mu}) + \hat{\boldsymbol{B}}(\boldsymbol{\mu})\boldsymbol{u}(t)\,? \qquad \text{SB2.}$$

The answer lies in classical intrusive projection-based model reduction methods. If we were to construct an intrusive projection-based reduced model, we would first approximate the high-dimensional state $\boldsymbol{x}(t;\boldsymbol{\mu}) \in \mathbb{R}^N$ in the n-dimensional subspace $\mathcal{V}$ spanned by the orthonormal basis $\boldsymbol{V} \in \mathbb{R}^{N\times n}$: $\boldsymbol{x}(t;\boldsymbol{\mu}) \approx \boldsymbol{V}\hat{\boldsymbol{x}}(t;\boldsymbol{\mu})$, where $\hat{\boldsymbol{x}}(t;\boldsymbol{\mu}) \in \mathbb{R}^n$ are the coordinates in the reduced-order representation. Substituting this approximation into the full-order model would yield a residual. The Galerkin reduced model would then be obtained by imposing the condition that the residual be orthogonal to the subspace $\mathcal{V}$, which leads to the reduced dynamical system, whose reduced-order matrices are the projections of the corresponding full-order matrices onto the subspace $\mathcal{V}$. That is, in this special case, we have $\hat{\boldsymbol{B}}(\boldsymbol{\mu}) = \boldsymbol{V}^\top\boldsymbol{B}(\boldsymbol{\mu}), \hat{\boldsymbol{A}}_1(\boldsymbol{\mu}) = \boldsymbol{V}^\top\boldsymbol{A}_1(\boldsymbol{\mu})\boldsymbol{V}$, and so on.

Operator Inference does not compute the reduced operators via projection; however, the intrusive projection framework reveals how projection preserves the structural polynomial form in the reduced model. Recall that this structural form in turn reflects the terms in the governing PDEs. Thus, it is the PDEs—the governing laws of physics—that dictate the particular structured form of the reduced model that we learn.

Equation 13 has full rank, then $\tilde{\boldsymbol{A}}_i = \hat{\boldsymbol{A}}_i$ and $\tilde{\boldsymbol{B}} = \hat{\boldsymbol{B}}$ for $i = 1,\ldots,\ell$. Note that $\hat{\boldsymbol{A}}_1,\ldots,\hat{\boldsymbol{A}}_\ell$ and $\hat{\boldsymbol{B}}$ are the matrices learned with Operator Inference for assembling the model given in Equation 7, whereas $\tilde{\boldsymbol{A}}_1,\ldots,\tilde{\boldsymbol{A}}_\ell$ and $\tilde{\boldsymbol{B}}$ are the reduced matrices of the Galerkin reduced model obtained with intrusive model reduction. In this sense, reprojection connects classical, intrusive model reduction that constructs $\tilde{\boldsymbol{A}}_1,\ldots,\tilde{\boldsymbol{A}}_\ell$ and $\tilde{\boldsymbol{B}}$ via an intrusive projection step (see the sidebar titled Motivating the Structural Form of the Learned Reduced Model) with nonintrusive Operator Inference that learns $\hat{\boldsymbol{A}}_1,\ldots,\hat{\boldsymbol{A}}_\ell$ and $\hat{\boldsymbol{B}}$ from data. Thus, it enables one to carry over theory from intrusive model reduction to nonintrusive model reduction.

4.2. Reprojection in Action

We now show how Operator Inference with reprojection can be put to use for error estimation (Section 4.2.1) as well as for learning from noisy (Section 4.2.2) and partially observed state trajectories (Section 4.2.3).

4.2.1. Error estimation for an end-to-end certification of predictions from data. On the basis of the recovery guarantee obtained with reprojection, Uy & Peherstorfer (2021b) introduce an a posteriori error estimator for models with linear dynamics. For a wide range of error estimators developed for intrusive model reduction, the key components are an efficient computation of the norm of the residual as well as a bound on the largest singular value of the operators of the high-dimensional models. As Uy & Peherstorfer (2021b) show, building on the error estimator introduced by Grepl & Patera (2005) and Haasdonk & Ohlberger (2011) from intrusive model reduction allows the same principles to be applied to Operator Inference models. The result is an end-to-end certification of predictions made with models obtained with Operator Inference

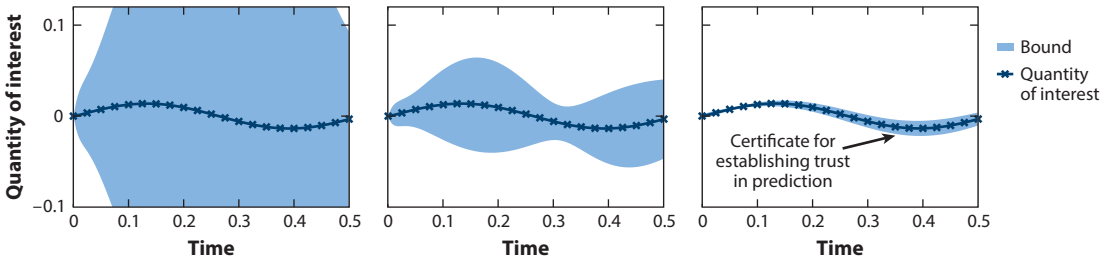

Figure 4

Models learned with Operator Inference from data of certain classes of systems can be equipped with a posteriori error bounds. The bounds provide accuracy certificates for predictions made with the learned models. In this example, the certificates (*blue*) become tighter about the prediction with increasing dimension, which means that predictions can be increasingly trusted as the dimension of the reduced models is increased.

from data. For example, based on the experiment reported by Uy & Peherstorfer (2021b), **Figure 4** shows an error bound for predicting the average outflow at boundary segments of the spatial domain of a convection-diffusion problem. As the dimension n of the learned Operator Inference model is increased, the certificates show that the predictions of the quantity of interest with the learned model can be increasingly trusted.

Other studies have considered error estimation and uncertainty quantification for Operator Inference. Guo et al. (2022) introduced a Bayesian Operator Inference approach. Given the posterior distribution, the predictions can be equipped with confidence intervals that reflect the uncertainties in the predictions.

4.2.2. Learning from noisy state observations with active data collection. Uy et al. (2023) investigated Operator Inference with reprojection in the context of sampling state trajectories with Gaussian noise using the reprojection scheme. They show that when collecting state observations with reprojection from systems polluted with Gaussian noise, the mean-squared error of the inferred operators can be bounded as

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}}_i - \tilde{\boldsymbol{A}}_i\|_F^2\right] \lesssim \frac{s^2}{\sigma_{\min}^2(\boldsymbol{D})}, \qquad 16.$$

where s is the standard deviation of the Gaussian noise; $\sigma_{\min}(\boldsymbol{D})$ denotes the smallest singular value of the data matrix $\boldsymbol{D}$; and $\hat{\boldsymbol{A}}_i$ and $\tilde{\boldsymbol{A}}_i$ are the learned and intrusive operators, respectively, for $i = 1, \ldots, \ell$. The bound given in Equation 16 holds because of Operator Inference's recovery guarantee with reprojection, which again lets the learned model be related to the projection-based Galerkin reduced model from intrusive model reduction; it also guarantees the unbiasedness of the learned operators. The ratio $s/\sigma_{\min}(\boldsymbol{D})$ can be interpreted as the noise-to-signal ratio, where s corresponds to the noise and $\sigma_{\min}(\boldsymbol{D})$ represents the strength of the signal in the observed states. On the basis of this interpretation as noise-to-signal ratio, Uy et al. (2023) introduce active Operator Inference, which selects from a dictionary initial conditions at which to query the high-dimensional model for data; the initial conditions are selected so that the noise-to-signal ratio is greedily minimized. The mathematical task underlying the active Operator Inference scheme is subselecting columns of a matrix to maximize the smallest singular value, which is widely studied in numerical linear algebra with numerous applications in model reduction (Astrid et al. 2008, Drmač & Gugercin 2016, Zimmermann & Willcox 2016, Peherstorfer et al. 2020).

4.2.3. Learning from partially observed state trajectories. When learning from partially observed state trajectories, the information that is lost due to partial observations can be

described with a similar concept of Markovian versus non-Markovian dynamics. To see that this is the case, consider a high-dimensional state $\boldsymbol{x}(t_k)$ at time t_k and consider the observation operator $\boldsymbol{T} \in \{0,1\}^{r\times N}$ that selects r components of $\boldsymbol{x}(t_k)$. The operator $\boldsymbol{T}$ is an orthonormal matrix that leads to a projection matrix $\boldsymbol{T}^\top\boldsymbol{T}$, which in turn projects a state $\boldsymbol{x}(t_k)$ onto the corresponding subspace $\mathcal{T}$. Analogously to Section 4.1.1, the states can then be decomposed into parts that are either in $\mathcal{T}$ or in the orthogonal complement of $\mathcal{T}$, corresponding to observed and unobserved components, respectively. This means that the observed states play an analogous role to the projected states in Section 4.1.1 and thus correspond to dynamics that are non-Markovian in the subspace $\mathcal{T}$. On the basis of these insights, Uy & Peherstorfer (2021a) propose to compensate for the lost information due to partially observed states in Operator Inference by learning reduced models with memory terms so that future-state predictions depend on the current state and the history of previous states. The result is a reduced model that includes Markovian and non-Markovian terms. Related concepts based on the Mori–Zwanzig formalism and time-delay embeddings to account for non-Markovian dynamics are widely used (Chorin et al. 2002, Li et al. 2014, Le Clainche & Vega 2017, Pan & Duraisamy 2018, Thiede et al. 2019). Uy & Peherstorfer (2021a) also introduce a reprojection scheme for partially observed states and show that non-Markovian terms, as in Equation 15, are recovered under certain assumptions.

Hamiltonian model: a model in which the state equation is derived from a Hamiltonian scalar function (an energy) via Hamilton's principle

4.3. Convergence of Learned Operators to Projected Operators

Peherstorfer & Willcox (2016a) and Benner et al. (2020) show that structure in the learning process can be leveraged to derive guarantees from data without reprojection. These guarantees are of an asymptotic nature, however. Under assumptions that can be established if the time-integration scheme used in the numerical model in Equation 3 is convergent, one can show that for all $\epsilon > 0$ there exists a dimension $n \leq N$ and a time-step size $\delta t > 0$ such that the difference between the learned operators $\hat{\boldsymbol{A}}_1, \ldots, \hat{\boldsymbol{A}}_\ell$ and $\hat{\boldsymbol{B}}$ from time-discrete states and the Galerkin operators $\tilde{\boldsymbol{A}}_1, \ldots, \tilde{\boldsymbol{A}}_\ell$ and $\tilde{\boldsymbol{B}}$ obtained via intrusive projection can be bounded as $\|\hat{\boldsymbol{A}}_i - \tilde{\boldsymbol{A}}_i\|_F \leq \epsilon$ and $\|\hat{\boldsymbol{B}} - \tilde{\boldsymbol{B}}\|_F \leq \epsilon$ for $i = 1, \ldots, \ell$. The result shows that the models learned with Operator Inference converge with $n \to N$ to the reduced models of intrusive model reduction. It does not imply a convergence rate, and empirical evidence suggests that the error does not decay monotonically with n.

5. STRUCTURED SOLUTION SPACES: HAMILTONIAN SYSTEMS

Thus far, we have considered general dynamical systems with polynomial structure, as in Equation 3. However, there is often more structure in the problem that one can exploit to obtain predictive reduced models. A key tenet of fluid mechanics is the conservation of derived quantities. For instance, the shallow-water equations and a large class of other wave-type problems can be formulated as Hamiltonian systems that conserve a system's energy over time. Hamiltonian models are derived from Hamilton's principle, so that their governing equations possess physical, mechanical, and mathematical structures in the form of symmetries, symplecticity, Casimirs, and energy conservation. The conservative nature and the underlying symplectic structure of Hamiltonian systems are fundamental to their discretization and numerical treatment and allow simulations to remain long-term predictive and stable, a major advantage of using such methods. The abovementioned structures constrain the solution space, and the behavior of these quantities in numerical simulation provides an important measure of accuracy of the discretized model. Special numerical approximations that satisfy these constraints are discussed by, for instance, Leimkuhler & Reich (2004), Hairer et al. (2006), and Sharma et al. (2020).

This section demonstrates how Hamiltonian structure can be enforced in the Operator Inference framework (for more details, see Sharma et al. 2022). Section 5.1 briefly describes this

structure, and Section 5.2 outlines how such structures can be preserved during learning. For ease of exposition, in this section we drop the dependence on $\boldsymbol{\mu}$.

5.1. Hamiltonian Models

In canonical Hamiltonian dynamical systems, the state is separated via $\boldsymbol{x}(t) = [\boldsymbol{q}(t)^\top, \boldsymbol{p}(t)^\top]^\top \in \mathbb{R}^{2N}$, where $\boldsymbol{q}(t) \in \mathbb{R}^N$ is the generalized position vector and $\boldsymbol{p}(t) \in \mathbb{R}^N$ is the generalized momentum vector. This separation of the state is due to distinct physical interpretations of the components of the state, and their relationship induces the canonical Hamiltonian structure. Consider a canonical nonlinear Hamiltonian system of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\boldsymbol{q}(t)\\ \boldsymbol{p}(t)\end{bmatrix} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ -\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\nabla_{\boldsymbol{q}} H(\boldsymbol{q}(t), \boldsymbol{p}(t))\\ \nabla_{\boldsymbol{p}} H(\boldsymbol{q}(t), \boldsymbol{p}(t))\end{bmatrix}.$$

In almost all cases, the Hamiltonian can be decomposed as

$$H(\boldsymbol{q}(t), \boldsymbol{p}(t)) = H_{\text{quad}}(\boldsymbol{q}(t), \boldsymbol{p}(t)) + H_{\text{nl}}(\boldsymbol{q}(t), \boldsymbol{p}(t)), \tag{17.}$$

where H_{quad} is quadratic in the states $\boldsymbol{q}$ and $\boldsymbol{p}$ and H_{nl} contains the remaining nonlinear terms. For a Hamiltonian system, its flow map preserves the canonical symplectic form, and the Hamiltonian is conserved; that is, $(\mathrm{d}/\mathrm{d}t)H(\boldsymbol{q}(t), \boldsymbol{p}(t)) = 0$ for all $t > 0$.

5.2. Hamiltonian Operator Inference

We consider a situation in which we are given the symbolic form of a canonical Hamiltonian PDE model and we have simulated data thereof. Our goal is to learn a Hamiltonian reduced model from the data of that system, so that the learned reduced model (*a*) is a canonical Hamiltonian system, (*b*) retains the physical interpretation of the state variables (generalized positions and momenta) and the coupling structure, and (*c*) respects the symmetric property of structure-preserving space discretizations. Let $\boldsymbol{q}(t_1), \ldots, \boldsymbol{q}(t_K)$ and $\boldsymbol{p}(t_1), \ldots, \boldsymbol{p}(t_K)$ be the solutions of the Hamiltonian high-dimensional numerical model computed with a structure-preserving numerical integration scheme. We define the snapshot matrices

$$\boldsymbol{Q} = \Big[\boldsymbol{q}(t_1), \ldots, \boldsymbol{q}(t_K)\Big] \in \mathbb{R}^{N\times K}, \qquad \boldsymbol{P} = \Big[\boldsymbol{p}(t_1), \ldots, \boldsymbol{p}(t_K)\Big] \in \mathbb{R}^{N\times K}. \tag{18.}$$

Assuming that we know the functional form of H_{nl}, we define the nonlinear forcing $\boldsymbol{f}_q$ and $\boldsymbol{f}_p$ as

$$\boldsymbol{f}_q(\boldsymbol{q}, \boldsymbol{p}) = \left[\frac{\partial H_{\text{nl}}}{\partial p_1}(q_1, p_1), \ldots, \frac{\partial H_{\text{nl}}}{\partial p_N}(q_N, p_N)\right]^\top, \boldsymbol{f}_p(\boldsymbol{q}, \boldsymbol{p}) = \left[\frac{\partial H_{\text{nl}}}{\partial q_1}(q_1, p_1), \ldots, \frac{\partial H_{\text{nl}}}{\partial q_N}(q_N, p_N)\right]^\top,$$

which allows us to compute the forcing snapshot matrices:

$$\boldsymbol{F}_q = \Big[\boldsymbol{f}_q(\boldsymbol{q}(t_1), \boldsymbol{p}(t_1)), \ldots, \boldsymbol{f}_q(\boldsymbol{q}(t_K), \boldsymbol{p}(t_K))\Big], \boldsymbol{F}_p = \Big[\boldsymbol{f}_p(\boldsymbol{q}(t_1), \boldsymbol{p}(t_1)), \ldots, \boldsymbol{f}_p(\boldsymbol{q}(t_K), \boldsymbol{p}(t_K))\Big].$$

Note that these forcing snapshot matrices are computed via postprocessing of the state snapshot data, so the approach remains nonintrusive but exploits knowledge of the functional form of the governing equations. Next, we compute time-derivative approximations $\boldsymbol{q}'(t_i)$ and $\boldsymbol{p}'(t_i)$ from the state trajectory data using a finite-difference scheme to build the snapshot matrices of the time-derivative data:

$$\boldsymbol{Q}' = \Big[\boldsymbol{q}'(t_1), \ldots, \boldsymbol{q}'(t_K)\Big] \in \mathbb{R}^{N\times K}, \qquad \boldsymbol{P}' = \Big[\boldsymbol{p}'(t_1), \ldots, \boldsymbol{p}'(t_K)\Big] \in \mathbb{R}^{N\times K}. \tag{19.}$$

To learn the reduced operators, we project the high-dimensional state trajectories onto low-dimensional symplectic subspaces and then fit operators to the projected trajectories in a

structure-preserving way. For the symplectic projection step, we choose the cotangent lift algorithm (see Peng & Mohseni 2016, where other options are also discussed) to generate a symplectic basis matrix $\boldsymbol{\Phi}$ that is used to approximate both the generalized position and the momenta. The following block structure of the symplectic basis matrix retains the physical interpretation of the reduced state variables:

$$\begin{bmatrix} \boldsymbol{q}(t) \\ \boldsymbol{p}(t) \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi} \end{bmatrix} \begin{bmatrix} \breve{\boldsymbol{q}}(t) \\ \breve{\boldsymbol{p}}(t) \end{bmatrix}.$$

We obtain projections of the trajectory snapshot data, the forcing functions, and the time-derivative data via the projections onto the symplectic basis matrix, namely $\breve{\boldsymbol{Q}} = \boldsymbol{\Phi}^\top \boldsymbol{Q}$, $\breve{\boldsymbol{P}} = \boldsymbol{\Phi}^\top \boldsymbol{P}$, $\breve{\boldsymbol{F}}_q = \boldsymbol{\Phi}^\top \boldsymbol{F}_q$, $\breve{\boldsymbol{F}}_p = \boldsymbol{\Phi}^\top \boldsymbol{F}_p$, $\breve{\boldsymbol{Q}}' = \boldsymbol{\Phi}^\top \boldsymbol{Q}'$, and $\breve{\boldsymbol{P}}' = \boldsymbol{\Phi}^\top \boldsymbol{P}'$, which are all real $n \times K$ data matrices. Inspired by the separation of the Hamiltonian functional in Equation 17, we define the following reduced Hamiltonian H_r in terms of the reduced operators $\hat{\boldsymbol{D}}_q \in \mathbb{R}^{n \times n}$ and $\hat{\boldsymbol{D}}_p \in \mathbb{R}^{n \times n}$ as

$$H_\mathrm{r}(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{p}}(t)) = \frac{1}{2} \hat{\boldsymbol{q}}(t)^\top \hat{\boldsymbol{D}}_q \hat{\boldsymbol{q}}(t) + \frac{1}{2} \hat{\boldsymbol{p}}(t)^\top \hat{\boldsymbol{D}}_p \hat{\boldsymbol{p}}(t) + H_\mathrm{nl}(\Phi \hat{\boldsymbol{q}}(t), \Phi \hat{\boldsymbol{p}}(t)). \tag{20.}$$

The equations of motion of the reduced model are then derived from Hamilton's principle:

$$\dot{\hat{\boldsymbol{q}}}(t) = \frac{\partial H_\mathrm{r}}{\partial \hat{\boldsymbol{p}}}(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{p}}(t)) = \hat{\boldsymbol{D}}_p \hat{\boldsymbol{p}}(t) + \boldsymbol{\Phi}^\top \boldsymbol{f}_q(\boldsymbol{\Phi} \hat{\boldsymbol{q}}(t), \boldsymbol{\Phi} \hat{\boldsymbol{p}}(t)),$$

$$\dot{\hat{\boldsymbol{p}}}(t) = -\frac{\partial H_\mathrm{r}}{\partial \hat{\boldsymbol{q}}}(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{p}}(t)) = -\hat{\boldsymbol{D}}_q \hat{\boldsymbol{q}}(t) - \boldsymbol{\Phi}^\top \boldsymbol{f}_p(\boldsymbol{\Phi} \hat{\boldsymbol{q}}(t), \boldsymbol{\Phi} \hat{\boldsymbol{p}}(t)),$$

which ensures that the reduced system is a canonical Hamiltonian system. Thus, the reduced model states are physically interpretable in that they retain the coupling structure and mechanical meaning of the states. We solve for $\hat{\boldsymbol{D}}_q$ and $\hat{\boldsymbol{D}}_p$ via the constrained optimization

$$\min_{\substack{\hat{\boldsymbol{D}}_q = \hat{\boldsymbol{D}}_q^\top, \\ \hat{\boldsymbol{D}}_p = \hat{\boldsymbol{D}}_p^\top}} \left\| \begin{bmatrix} \breve{\boldsymbol{Q}}' - \breve{\boldsymbol{F}}_q(\breve{\boldsymbol{Q}}, \breve{\boldsymbol{P}}) \\ \breve{\boldsymbol{P}}' + \breve{\boldsymbol{F}}_p(\breve{\boldsymbol{Q}}, \breve{\boldsymbol{P}}) \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \hat{\boldsymbol{D}}_p \\ -\hat{\boldsymbol{D}}_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \breve{\boldsymbol{Q}} \\ \breve{\boldsymbol{P}} \end{bmatrix} \right\|_F. \tag{21.}$$

The optimization problem for $\hat{\boldsymbol{D}}_q$ and $\hat{\boldsymbol{D}}_p$ can be broken down into separate symmetric linear least-squares problems; subsequently, the symmetric reduced operators can be obtained by solving the Lyapunov equations

$$(\breve{\boldsymbol{Q}} \breve{\boldsymbol{Q}}^\top) \hat{\boldsymbol{D}}_q + \hat{\boldsymbol{D}}_q (\breve{\boldsymbol{Q}} \breve{\boldsymbol{Q}}^\top) = \breve{\boldsymbol{Q}} \hat{\boldsymbol{R}}_p^\top + \hat{\boldsymbol{R}}_q \breve{\boldsymbol{Q}}^\top, \; (\breve{\boldsymbol{P}} \breve{\boldsymbol{P}}^\top) \hat{\boldsymbol{D}}_p + \hat{\boldsymbol{D}}_p (\breve{\boldsymbol{P}} \breve{\boldsymbol{P}}^\top) = \breve{\boldsymbol{P}} \breve{\boldsymbol{R}}_p^\top + \breve{\boldsymbol{R}}_p \breve{\boldsymbol{P}}^\top, \tag{22.}$$

where $\breve{\boldsymbol{R}}_p = \breve{\boldsymbol{Q}}' - \breve{\boldsymbol{F}}_q(\breve{\boldsymbol{Q}}, \breve{\boldsymbol{P}})$ and $\breve{\boldsymbol{R}}_p = \breve{\boldsymbol{P}}' + \breve{\boldsymbol{F}}_p(\breve{\boldsymbol{Q}}, \breve{\boldsymbol{P}})$. The symmetry constraints on $\hat{\boldsymbol{D}}_q$ and $\hat{\boldsymbol{D}}_p$ ensure that the learned operators yield structure-preserving reduced models that are Hamiltonian systems.

Figure 5 depicts a numerical example of the structure-preserving properties that Hamiltonian Operator Inference can achieve. The test problem is a nonlinear Schrödinger equation, which, in addition to preserving the nonlinear Hamiltonian energy, possesses quadratic mass and momentum invariants (for details, see Sharma et al. 2022, section 4.3). **Figure 5** plots the errors in the conservation of energy, which is evaluated via the Hamiltonian of the high-dimensional model as $|H(\boldsymbol{\Phi} \hat{\boldsymbol{q}}(t), \boldsymbol{\Phi} \hat{\boldsymbol{p}}(t)) - H(\boldsymbol{\Phi} \hat{\boldsymbol{q}}(t_0), \boldsymbol{\Phi} \hat{\boldsymbol{p}}(t_0))|$. Conservation of mass and momentum are similarly evaluated; these errors are also plotted in the figure. As shown in **Figure 5*a***, the Hamiltonian Operator Inference reduced model exhibits bounded energy error of the high-dimensional Hamiltonian $H(\cdot)$, which is a key indicator of the structure-preserving property. **Figure 5*b*** shows that the learned reduced model also exhibits bounded error for the mass invariant, and

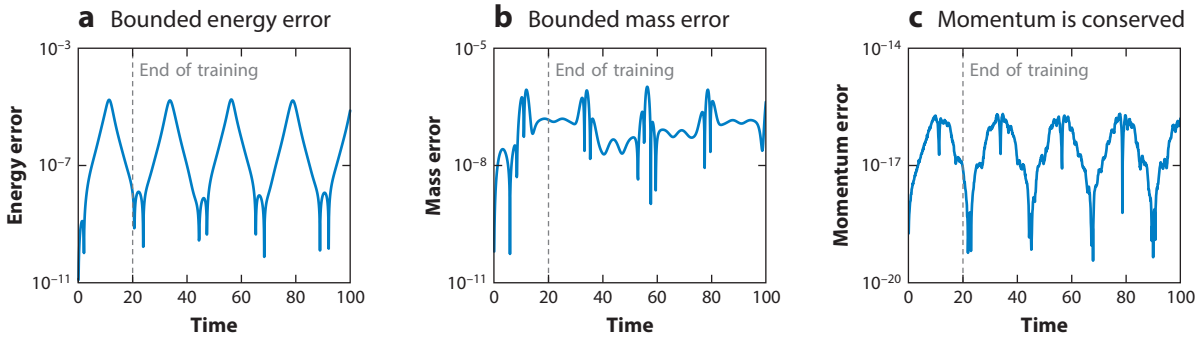

Figure 5

Hamiltonian Operator Inference reduced models for the nonlinear Schrödinger equation. (*a*) The reduced model exhibits bounded energy error. (*b*) The learned reduced model also exhibits bounded error for the mass invariant. (*c*) The reduced model conserves the momentum invariant to machine precision. These are key indicators of structure preservation. Figure adapted with permission from Sharma et al. (2022).

Figure 5*c* shows that the model exactly (to machine precision) conserves the momentum invariant. These invariants are evaluated as defined by Sharma et al. (2022, section 4.3). We further observe that the structure-preserving Hamiltonian reduced models show excellent long-term predictive capabilities, as the models provide accurate results 400% outside the training interval.

6. DISCUSSION AND OUTLOOK

Surrogate modeling continues to play an increasingly important role in achieving optimization, design, control, data assimilation, and uncertainty quantification for complex physical systems. Even as computing capabilities continue to increase, the demands for increased resolution, higher fidelity, and rapid turnaround times will necessitate surrogate models. For example, surrogate modeling is an essential enabler for digital twins, which require high-fidelity yet rapid and lightweight computations to achieve real-time data assimilation and control. While automated learning of surrogate models from data is becoming increasingly popular, it remains critical to rigorously incorporate information about the physics that underlie the systems of interest. This information, provided by governing equations, first principles, and theoretical insights, is imperative to derive surrogate models that generalize well to new, unseen parameter ranges and that provide physically meaningful predictions even in edge cases and limit states.

In this review, we have discussed Operator Inference, a nonlinear reduced modeling approach that incorporates the physics by defining a structured form for the reduced model, and then learns the corresponding reduced operators from simulated training data. We demonstrate that this structured approach to learning surrogate models connects data-driven modeling with systems and control theoretic concepts, such as Lyapunov stability theory, to validate models beyond mere empirical evaluations on test data sets. The polynomial structure of the Operator Inference models also plays a key role in carrying over the rich theory of traditional intrusive model reduction to data-driven modeling with Operator Inference. In particular, we discuss conditions under which Operator Inference can provide error bounds for model predictions, and we derive insights to explain and improve predictions from noisy data and partial observations. We also show that the fundamental concepts of Operator Inference are flexible, so that energy-preserving models based on Hamiltonian structure can be learned to predict far into the future (i.e., beyond a training horizon) while retaining the physical meaning of states.

We emphasize that Operator Inference was designed with ease of use in mind; having a low barrier to adoption is critical for the success of a computational method in practice. First, Operator Inference is a nonintrusive approach, which sets it apart from traditional model reduction methods that are predominantly intrusive and thus typically require rewriting solvers. The intrusive nature of many model reduction methods has limited their scope and use in application-driven science and engineering—it is often infeasible to develop reduced model solvers from scratch when large-scale codes are involved that have grown over years, if not decades. Second, the training process in Operator Inference relies on only a few hyperparameters that can be selected via cross-validation in a principled way. Third, all steps in training Operator Inference reduced models are scalable to snapshots with millions of state components via scalable linear algebra (e.g., randomized algorithms), as demonstrated by Swischuk et al. (2020a) and Farcaş et al. (2022, 2023a). Fourth, the polynomial structure of Operator Inference models strikes a balance between, on one hand, being sufficiently expressive to cover a wide range of dynamics found in science and engineering and, on the other hand, being manageable for efficient computations. In contrast, other nonlinear approaches, such as deep neural network architectures, tend to be more expressive in general but also allow almost arbitrary architectures, with little guidance on how to choose them.

We highlight three avenues for future research that apply to Operator Inference and, more generally, to learning low-dimensional models from data. First, this review has focused on reduction methods that exploit low-rank structure by constructing low-dimensional linear subspaces in which to evolve the reduced model state; however, dynamics that are dominated by transport, such as strongly advecting flows and particle systems, are affected by the so-called Kolmogorov barrier, which states that the error of linear approximations in low-dimensional spaces decays slowly (for a survey, see Peherstorfer 2022). In the context of nonintrusive modeling with Operator Inference, researchers have proposed adapting models during the online phase (Peherstorfer & Willcox 2015, 2016b; Kramer et al. 2017), constructing multiple local subspaces instead of a single global one (Geelen & Willcox 2022), learning on quadratic manifolds (Geelen et al. 2023), and learning and embedding spatial shifts when formulating the Operator Inference problem (Issan & Kramer 2022) to obtain nonlinear reductions for circumventing the Kolmogorov barrier. More remains to be done, especially regarding stability, online efficiency, and structure preservation.

A second important avenue of future research is establishing trust in predictions. This review touches on studies that derive probabilistic error bounds and Bayesian formulations of Operator Inference in Section 4.2.1, but much more research is required to give scientists and engineers tools that they can trust for high-consequence decisions. An opportunity is that models are often used within outer-loop applications such as design, control, and data assimilation, meaning that it may be sufficient to certify the outer-loop result (e.g., the optimal design, the constructed controller) rather than to provide bounds for the model predictions. This approach is taken by multifidelity methods that leverage computationally efficient approximate models while keeping limited recourse to the expensive, high-fidelity models to establish accuracy guarantees (Peherstorfer et al. 2018). In Section 5, we establish trust in models by retaining their physical interpretability and long-term stability. Lagrangian models, another important class of model structures, arise by applying the Euler–Lagrange equations to a Lagrangian function, which is the difference between the kinetic and potential energy. Lagrangian models are second order in time, and many wave-type problems exhibit Lagrangian structure and have physically interpretable quantities such as momentum, energy, or vorticity. A famous result obtained by Noether (1971) states that there exists an invariant of the motion corresponding to each symmetry of the system Lagrangian. While Lagrangian structure-preserving models can be learned via Operator Inference (Sharma & Kramer 2022), more research needs to be done to include identification of

nonlinear terms, more efficient basis constructions (e.g., via online updating), and extensions to the parametric case.

A third future research direction is the development of methods that incorporate the downstream design/control task into the learning. This is closely related to goal-oriented intrusive modeling where, for example, meshes and models are adapted so that quantities of interest are accurately approximated rather than the high-dimensional states (Prudhomme & Oden 1999, Becker & Rannacher 2001) and reduced models are constructed with specific goals in mind (Willcox et al. 2005; Bui-Thanh et al. 2007; Lieberman & Willcox 2013, 2014; Spantini et al. 2017). When it comes to learning from data, the simplicity of a task can be measured with sample complexity, that is, how many data samples are required to achieve a certain accuracy or success. For example, Werner & Peherstorfer (2023a,b) show that inferring state-feedback controllers for the task of stabilization requires provably fewer samples than learning models of the system dynamics and subsequently applying control strategies on the learned models. Similarly, when models are used in specific contexts such as within multifidelity computations for solving outer-loop applications, taking this context into account during model training helps reduce the number of training samples that are required (e.g., Alsup & Peherstorfer 2023, Farcaş et al. 2023b). Given that data are scarce and physics is complex in science and engineering applications, it is important to optimally identify what parts of the system dynamics need to be learned for solving downstream tasks and exploit this information in the training phase.

SUMMARY POINTS

1. Operator Inference is a reduced modeling approach that incorporates physics by defining a structured polynomial form for the reduced model, and then learns the corresponding reduced operators from simulated training data.
2. Operator Inference is nonintrusive and therefore has a low barrier to adoption. In contrast, traditional model reduction is intrusive and so requires access to operators of the high-fidelity numerical models.
3. The polynomial model form of Operator Inference strikes a balance between being sufficiently expressive to cover a wide range of nonlinear dynamics found in science and engineering and providing efficient reduced model computations.
4. Under some conditions, the theory of traditional intrusive model reduction can be carried over to data-driven modeling with Operator Inference, which provides error estimators for model predictions, yields insights to improve predictions from noisy data and partial observations, and makes it possible to encode Hamiltonian and other structures for physically meaningful predictions beyond the training horizon.

FUTURE ISSUES

1. Nonintrusive reduced modeling methods with nonlinear approximations are needed to circumvent the Kolmogorov barrier of physical phenomena dominated by transport, such as strongly advecting flows and wave problems. Early studies on nonlinear manifold model reduction methods exist, but much more remains to be done regarding stability, online efficiency, and structure preservation.

2. It remains challenging to provide guarantees and establish trust in predictions of data-driven models, which are required for high-consequence decisions. Active research directions include certifying model predictions directly with error bounds, Bayesian approaches, and multifidelity methods to certify outer-loop results using limited recourse to expensive, high-fidelity models.
3. Given that data are scarce and physics is complex in science and engineering, it is important to understand how to optimally identify what parts of system dynamics need to be learned to solve downstream decision tasks and to exploit this information in the training phase.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Elizabeth Qian and Ionuţ-Gabriel Farcaş for giving permission to reprint results from Qian et al. (2022). B.K. has been supported in part by the US Office of Naval Research (award N000142212624); the National Science Foundation (NSF; award CMMI-2144023); the Ministry of Trade, Industry and Energy and the Korea Institute for Advancement of Technology (award P0019804); and the Isaac Newton Institute for Mathematical Sciences (Engineering and Physical Sciences Research Council grant EP/R014604/1). B.P. has been supported by the US Department of Energy (DOE; award DE-SC0019334) and the NSF (awards DMS-2012250 and IIS-1901091). K.E.W. has been supported in part by the DOE (awards DE-SC0019303 and DE-SC0021239) and the Air Force Office of Scientific Research (awards FA9550-17-1-0195, FA9550-21-1-0084, and FA9550-22-1-0419).

LITERATURE CITED

Alsup T, Peherstorfer B. 2023. Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quant.* 11(1):285–319

Amsallem D, Farhat C. 2008. Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA J.* 46(7):1803–13

Antoulas AC. 2005. *Approximation of Large-Scale Dynamical Systems.* Philadelphia: SIAM

Antoulas AC, Beattie CA, Gugercin S. 2021. *Interpolatory Methods for Model Reduction.* Philadelphia: SIAM

Astrid P, Weiland S, Willcox K, Backx T. 2008. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Trans. Autom. Control* 53(10):2237–51

Audouze C, De Vuyst F, Nair PB. 2009. Reduced-order modeling of parameterized PDEs using time–space-parameter principal component analysis. *Int. J. Numer. Methods Eng.* 80(8):1025–57

Baker N, Alexander F, Bremer T, Hagberg A, Kevrekidis Y, et al. 2019. *Workshop report on basic research needs for scientific machine learning: core technologies for artificial intelligence.* Tech. Rep., Off. Sci., Dep. Energy, Washington, DC

Balajewicz M, Tezaur I, Dowell E. 2016. Minimal subspace rotation on the Stiefel manifold for stabilization and enhancement of projection-based reduced order models for the compressible Navier–Stokes equations. *J. Comput. Phys.* 321:224–41

Becker R, Rannacher R. 2001. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.* 10:1–102

Benner P, Breiten T. 2015. Two-sided projection methods for nonlinear model order reduction. *SIAM J. Sci. Comput.* 37(2):B239–60

Benner P, Goyal P, Gugercin S. 2018. $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM J. Matrix Anal. Appl.* 39(2):983–1032

Benner P, Goyal P, Kramer B, Peherstorfer B, Willcox K. 2020. Operator inference for non-intrusive model reduction of systems with non-polynomial nonlinear terms. *Comput. Methods Appl. Mech. Eng.* 372:113433

Benner P, Gugercin S, Willcox K. 2015. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.* 57(4):483–531

Brenig L. 2018. Reducing nonlinear dynamical systems to canonical forms. *Philos. Trans. R. Soc. A* 376(2124):20170384

Brunton SL, Brunton BW, Proctor JL, Kutz JN. 2016a. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PLOS ONE* 11(2):e0150171

Brunton SL, Noack BR, Koumoutsakos P. 2020. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* 52:477–508

Brunton SL, Proctor JL, Kutz JN. 2016b. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS* 113(15):3932–37

Bui-Thanh T, Willcox K, Ghattas O, van Bloemen Waanders B. 2007. Goal-oriented, model-constrained optimization for reduction of large-scale systems. *J. Comput. Phys.* 224(2):880–96

Bychkov A, Pogudin G. 2021. Optimal monomial quadratization for ODE systems. In *Combinatorial Algorithms*, ed. P Flocchini, L Moura, pp. 122–36. Cham, Switz.: Springer

Chesi G. 2007. Estimating the domain of attraction via union of continuous families of Lyapunov estimates. *Syst. Control Lett.* 56(4):326–33

Chorin A, Hald OH, Kupferman R. 2002. Optimal prediction with memory. *Physica D* 166(3):239–57

Chorin A, Stinis P. 2006. Problem reduction, renormalization, and memory. *Commun. Appl. Math. Comput. Sci.* 1(1):1–27

Cole JD. 1951. On a quasi-linear parabolic equation occurring in aerodynamics. *Q. Appl. Math.* 9(3):225–36

Coveney PV, Dougherty ER, Highfield RR. 2016. Big data need big theory too. *Philos. Trans. R. Soc. A* 374(2080):20160153

Degroote J, Vierendeels J, Willcox K. 2010. Interpolation among reduced-order matrices to obtain parameterized models for design, optimization and probabilistic analysis. *Int. J. Numer. Methods Fluids* 63(2):207–30

Dowell EH, Hall KC. 2001. Modeling of fluid-structure interaction. *Annu. Rev. Fluid Mech.* 33:445–90

Drmač Z, Gugercin S. 2016. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM J. Sci. Comput.* 38(2):A631–48

Duraisamy K, Iaccarino G, Xiao H. 2019. Turbulence modeling in the age of data. *Annu. Rev. Fluid Mech.* 51:357–77

Farcaş I, Gundevia R, Munipalli R, Willcox KE. 2023a. *Parametric non-intrusive reduced-order models via operator inference for large-scale rotating detonation engine simulations*. Paper presented at 2023 AIAA SCITECH Forum, National Harbor, MD, pap. AIAA 2023-0172

Farcaş I, Munipalli R, Willcox KE. 2022. *On filtering in non-intrusive data-driven reduced-order modeling.* Paper presented at AIAA AVIATION 2022 Forum, Chicago, pap. AIAA 2022-2111

Farcaş IG, Peherstorfer B, Neckel T, Jenko F, Bungartz HJ. 2023b. Context-aware learning of hierarchies of low-fidelity models for multi-fidelity uncertainty quantification. *Comput. Methods Appl. Mech. Eng.* 406:115908

Forrester AIJ, Sobester A, Keane AJ. 2008. *Engineering Design via Surrogate Modelling: A Practical Guide*. Hoboken, NJ: Wiley

Gear CW, Hyman JM, Kevrekidid PG, Kevrekidis IG, Runborg O, Theodoropoulos C. 2003. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Commun. Math. Sci.* 1(4):715–62

Geelen R, Willcox K. 2022. Localized non-intrusive reduced-order modelling in the operator inference framework. *Philos. Trans. R. Soc. A* 380(2229):20210206

Geelen R, Wright S, Willcox K. 2023. Operator inference for non-intrusive model reduction with quadratic manifolds. *Comput. Methods Appl. Mech. Eng.* 403:115717

Ghattas O, Willcox K. 2021. Learning physics-based models from data: perspectives from inverse problems and model reduction. *Acta Numer.* 30:445–554

Givon D, Kupferman R, Stuart A. 2004. Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity* 17(6):R55–127

Grepl MA, Patera AT. 2005. A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *ESAIM Math. Model. Numer. Anal.* 39(1):157–81

Gu C. 2011. QLMOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 30(9):1307–20

Guillot L, Cochelin B, Vergez C. 2019. A generic and efficient Taylor series–based continuation method using a quadratic recast of smooth nonlinear systems. *Int. J. Numer. Methods Eng.* 119(4):261–80

Guo M, McQuarrie SA, Willcox KE. 2022. Bayesian operator inference for data-driven reduced-order modeling. *Comput. Methods Appl. Mech. Eng.* 402:115336

Haasdonk B, Ohlberger M. 2011. Efficient reduced models and a posteriori error estimation for parametrized dynamical systems by offline/online decomposition. *Math. Comput. Model. Dyn. Syst.* 17(2):145–61

Hairer E, Lubich C, Wanner G. 2006. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Ser. Comput. Math. Vol. 31. Berlin: Springer

Hall KC, Thomas JP, Dowell EH. 2000. Proper orthogonal decomposition technique for transonic unsteady aerodynamic flows. *AIAA J.* 38(10):1853–62

Harvazinski ME, Huang C, Sankaran V, Feldman TW, Anderson WE, et al. 2015. Coupling between hydrodynamics, acoustics, and heat release in a self-excited unstable combustor. *Phys. Fluids* 27:045102

Hemati MS, Rowley CW, Deem EA, Cattafesta LN. 2017. De-biasing the dynamic mode decomposition for applied Koopman spectral analysis of noisy datasets. *Theor. Comput. Fluid Dyn.* 31:349–68

Hemery M, Fages F, Soliman S. 2020. On the complexity of quadratization for polynomial differential equations. In *Computational Methods in Systems Biology*, ed. A Abate, T Petrov, V Wolf, pp. 120–40. Cham, Switz.: Springer

Hemery M, Fages F, Soliman S. 2021. Compiling elementary mathematical functions into finite chemical reaction networks via a polynomialization algorithm for ODEs. In *Computational Methods in Systems Biology*, ed. E Cinquemani, L Paulevé, pp. 74–90. Cham, Switz.: Springer

Hesthaven JS, Rozza G, Stamm B. 2016. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. Cham, Switz.: Springer

Hesthaven JS, Ubbiali S. 2018. Non-intrusive reduced order modeling of nonlinear problems using neural networks. *J. Comput. Phys.* 363:55–78

Hinze M, Volkwein S. 2005. Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control. In *Dimension Reduction of Large-Scale Systems*, ed. P Benner, V Mehrmann, D Sorensen, pp. 261–306. Lect. Notes Comput. Appl. Math. Vol. 45. Berlin: Springer

Holmes P, Lumley J, Berkooz G. 1996. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge, UK: Cambridge Univ. Press

Hopf E. 1950. The partial differential equation $u_t + uu_x = \mu_{xx}$. *Commun. Pure Appl. Math.* 3(3):201–30

Huang C, Duraisamy K, Merkle CL. 2019. Investigations and improvement of robustness of reduced-order models of reacting flow. *AIAA J.* 57(12):5377–89

Huang C, Gejji R, Anderson W, Yoon C, Sankaran V. 2020. Combustion dynamics in a single-element lean direct injection gas turbine combustor. *Combust. Sci. Technol.* 192(12):2371–98

Hughes TJ, Franca L, Mallet M. 1986. A new finite element formulation for computational fluid dynamics. I. Symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comput. Methods Appl. Mech. Eng.* 54(2):223–34

Issan O, Kramer B. 2022. Predicting solar wind streams from the inner-heliosphere to Earth via shifted operator inference. *J. Comput. Phys.* 3473:111689

Jain P, McQuarrie S, Kramer B. 2021. *Performance comparison of data-driven reduced models for a single-injector combustion process*. Paper presented at 2021 AIAA Propulsion and Energy Forum and Exposition, online, pap. AIAA 2021-3633

Jakubczyk B, Respondek W. 1980. On linearization of control systems. *Bull. Acad. Polon. Sci.* 28:517–22
Kalashnikova I, Barone M. 2011. *Stable and efficient Galerkin reduced order models for non-linear fluid flow*. Paper presented at 6th AIAA Theoretical Fluid Mechanics Conference, Honolulu, HI, pap. AIAA 2011-3110
Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. 2021. Physics-informed machine learning. *Nat. Rev. Phys.* 3(6):422–40
Khalil HK. 2002. *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall. 3rd ed.
Khodabakhshi P, Willcox KE. 2022. Non-intrusive data-driven model reduction for differential–algebraic equations derived from lifting transformations. *Comput. Methods Appl. Mech. Eng.* 389:114296
Kramer B. 2021. Stability domains for quadratic-bilinear reduced-order models. *SIAM J. Appl. Dyn. Syst.* 20(2):981–96
Kramer B, Peherstorfer B, Willcox K. 2017. Feedback control for systems with uncertain parameters using online-adaptive reduced models. *SIAM J. Appl. Dyn. Syst.* 16(3):1563–86
Kramer B, Willcox K. 2019. Nonlinear model order reduction via lifting transformations and proper orthogonal decomposition. *AIAA J.* 57(6):2297–307
Kramer B, Willcox K. 2022. Balanced truncation model reduction for lifted nonlinear systems. In *Realization and Model Reduction of Dynamical Systems: A Festschrift in Honor of the 70th Birthday of Thanos Antoulas*, ed. C Beattie, P Benner, M Embree, S Gugercin, S Lefteriu, pp. 157–74. Cham, Switz.: Springer
Kutz JN, Brunton SL, Brunton BW, Proctor JL. 2016. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Philadelphia: SIAM
Langley P. 1981. Data-driven discovery of physical laws. *Cogn. Sci.* 5(1):31–54
Le Clainche S, Vega J. 2017. Higher order dynamic mode decomposition. *SIAM J. Appl. Dyn. Syst.* 16(2):882–925
Leimkuhler B, Reich S. 2004. *Simulating Hamiltonian Dynamics*. Cambridge, UK: Cambridge Univ. Press
Li Z, Bian X, Caswell B, Karniadakis GE. 2014. Construction of dissipative particle dynamics models for complex fluids via the Mori–Zwanzig formulation. *Soft Matter* 10(43):8659–72
Lieberman C, Willcox K. 2013. Goal-oriented inference: approach, linear theory, and application to advection diffusion. *SIAM Rev.* 55(3):493–519
Lieberman C, Willcox K. 2014. Nonlinear goal-oriented Bayesian inference: application to carbon capture and storage. *SIAM J. Sci. Comput.* 36(3):B427–49
Liljegren-Sailer B, Marheineke N. 2022. Input-tailored system-theoretic model order reduction for quadratic-bilinear systems. *SIAM J. Matrix Anal. Appl.* 43(1):1–39
Liu J, Zhan N, Zhao H, Zou L. 2015. Abstraction of elementary hybrid systems by variable transformation. In *FM15: Formal Methods. International Symposium on Formal Methods*, pp. 360–77. Berlin: Springer
Ljung L. 1987. *System Identification*. Upper Saddle River, NJ: Prentice Hall
Lumley J. 1967. The structures of inhomogeneous turbulent flow. In *Atmospheric Turbulence and Radio Wave Propagation*, ed. AM Yaglom, VI Tartarsky, pp. 166–78. Fort Belvoir, VA: Defense Tech. Inf. Cent.
McCormick GP. 1976. Computability of global solutions to factorable nonconvex programs. Part I: Convex underestimating problems. *Math. Progr.* 10(1):147–75
McQuarrie SA, Huang C, Willcox K. 2021. Data-driven reduced-order models via regularised operator inference for a single-injector combustion process. *J. R. Soc. N. Z.* 51(2):194–211
Mezić I. 2005. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dyn.* 41(1–3):309–25
Netto M, Susuki Y, Krishnan V, Zhang Y. 2021. On analytical construction of observable functions in extended dynamic mode decomposition for nonlinear estimation and prediction. In *2021 American Control Conference*, pp. 4190–95. Piscataway, NJ: IEEE
Noether E. 1971. Invariant variation problems. *Transp. Theory Stat. Phys.* 1(3):186–207
Pan S, Duraisamy K. 2018. Data-driven discovery of closure models. *SIAM J. Appl. Dyn. Syst.* 17(4):2381–413
Panzer H, Mohring J, Eid R, Lohmann B. 2010. Parametric model order reduction by matrix interpolation. *Automatisierungstechnik* 58(8):475–84
Peherstorfer B. 2020a. Model reduction for transport-dominated problems via online adaptive bases and adaptive sampling. *SIAM J. Sci. Comput.* 42(5):A2803–36
Peherstorfer B. 2020b. Sampling low-dimensional Markovian dynamics for preasymptotically recovering reduced models from data with operator inference. *SIAM J. Sci. Comput.* 42(5):A3489–515

Peherstorfer B. 2022. Breaking the Kolmogorov barrier with nonlinear model reduction. *Not. Am. Math. Soc.* 69:725–33

Peherstorfer B, Drmač Z, Gugercin S. 2020. Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points. *SIAM J. Sci. Comput.* 42:A2837–64

Peherstorfer B, Willcox K. 2015. Dynamic data-driven reduced-order models. *Comput. Methods Appl. Mech. Eng.* 291:21–41

Peherstorfer B, Willcox K. 2016a. Data-driven operator inference for nonintrusive projection-based model reduction. *Comput. Methods Appl. Mech. Eng.* 306:196–215

Peherstorfer B, Willcox K. 2016b. Dynamic data-driven model reduction: adapting reduced models from incomplete data. *Adv. Model. Simul. Eng. Sci.* 3:11

Peherstorfer B, Willcox K, Gunzburger M. 2018. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.* 60(3):550–91

Peng L, Mohseni K. 2016. Symplectic model reduction of Hamiltonian systems. *SIAM J. Sci. Comput.* 38(1):A1–27

Prud'homme C, Rovas DV, Veroy K, Machiels L, Maday Y, et al. 2001. Reliable real-time solution of parametrized partial differential equations: reduced-basis output bound methods. *J. Fluids Eng.* 124(1):70–80

Prudhomme S, Oden J. 1999. On goal-oriented error estimation for elliptic problems: application to the control of pointwise errors. *Comput. Methods Appl. Mech. Eng.* 176(1):313–31

Qian E, Farcaş IG, Willcox K. 2022. Reduced operator inference for nonlinear partial differential equations. *SIAM J. Sci. Comput.* 44(4):A1934–59

Qian E, Kramer B, Peherstorfer B, Willcox K. 2020. Lift & Learn: physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D* 406:132401

Quarteroni A, Rozza G, eds. 2014. *Reduced Order Methods for Modeling and Computational Reduction*. Berlin: Springer

Raissi M, Karniadakis GE. 2018. Hidden physics models: machine learning of nonlinear partial differential equations. *J. Comput. Phys.* 357:125–41

Rasmussen C, Williams C. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press

Rezaian E, Wei M. 2020. *Impact of symmetrization on the robustness of POD-Galerkin ROMs for compressible flows*. Paper presented at 2020 AIAA SCITECH Forum, Orlando, FL, pap. AIAA 2020-1318

Rowley CW, Dawson ST. 2017. Model reduction for flow analysis and control. *Annu. Rev. Fluid Mech.* 49:387–417

Rowley CW, Mezić I, Bagheri S, Schlatter P, Henningson DS. 2009. Spectral analysis of nonlinear flows. *J. Fluid Mech.* 641:115–27

Rozza G, Huynh DBP, Patera AT. 2008. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Arch. Comput. Methods Eng.* 15(3):229–75

Savageau MA, Voit EO. 1987. Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Math. Biosci.* 87(1):83–115

Sawant N, Kramer B, Peherstorfer B. 2023. Physics-informed regularization and structure preservation for learning stable reduced models from data with operator inference. *Comput. Methods Appl. Mech. Eng.* 404:115836

Schmid PJ. 2010. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* 656:5–28

Schmid PJ. 2022. Dynamic mode decomposition and its variants. *Annu. Rev. Fluid Mech.* 54:225–54

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

Sharma H, Kramer B. 2022. Preserving Lagrangian structure in data-driven reduced-order modeling of large-scale mechanical systems. arXiv:2203.06361 [math.NA]

Sharma H, Patil M, Woolsey C. 2020. A review of structure-preserving numerical methods for engineering applications. *Comput. Methods Appl. Mech. Eng.* 366:113067

Sharma H, Wang Z, Kramer B. 2022. Hamiltonian operator inference: Physics-preserving learning of reduced-order models for Hamiltonian systems. *Physica D* 431:133122

Sirovich L. 1987. Turbulence and the dynamics of coherent structures. Part 1. Coherent structures. *Q. Appl. Math.* 45(3):561–71

Spantini A, Cui T, Willcox K, Tenorio L, Marzouk Y. 2017. Goal-oriented optimal approximations of Bayesian linear inverse problems. *SIAM J. Sci. Comput.* 39(5):S167–96

Swischuk R, Kramer B, Huang C, Willcox K. 2020a. Learning physics-based reduced-order models for a single-injector combustion process. *AIAA J.* 58(6):2658–72

Swischuk R, Mainini L, Peherstorfer B, Willcox K. 2019. Projection-based model reduction: formulations for physics-based machine learning. *Comput. Fluids* 179:704–17

Tesi A, Villoresi F, Genesio R. 1994. On stability domain estimation via a quadratic Lyapunov function: convexity and optimality properties for polynomial systems. In *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, Vol. 2, pp. 1907–12. Piscataway, NJ: IEEE

Thiede EH, Giannakis D, Dinner AR, Weare J. 2019. Galerkin approximation of dynamical quantities using trajectory data. *J. Chem. Phys.* 150:244111

Tu JH, Rowley CW, Luchtenburg DM, Brunton SL, Kutz JN. 2014. On dynamic mode decomposition: theory and applications. *J. Comput. Dyn.* 1(2):391–421

Urban K, Patera AT. 2012. A new error bound for reduced basis approximation of parabolic partial differential equations. *C. R. Math.* 350(3/4):203–7

Uy WIT, Peherstorfer B. 2021a. Operator inference of non-Markovian terms for learning reduced models from partially observed state trajectories. *J. Sci. Comput.* 88(3):91

Uy WIT, Peherstorfer B. 2021b. Probabilistic error estimation for non-intrusive reduced models learned from data of systems governed by linear parabolic partial differential equations. *ESAIM Math. Model. Numer. Anal.* 55(3):735–61

Uy WIT, Wang Y, Wen Y, Peherstorfer B. 2023. Active operator inference for learning low-dimensional dynamical-system models from noisy data. *SIAM J. Sci. Comput.* 45(4):A1462–90

Veroy K, Rovas DV, Patera AT. 2002. A posteriori error estimation for reduced-basis approximation of parametrized elliptic coercive partial differential equations: "convex inverse" bound conditioners. *ESAIM Control Optim. Calc. Var.* 8:1007–28

Wang Q, Hesthaven JS, Ray D. 2019. Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. *J. Comput. Phys.* 384:289–307

Werner SWR, Peherstorfer B. 2023a. Context-aware controller inference for stabilizing dynamical systems from scarce data. *Proc. R. Soc. A* 479(2270):20220506

Werner SWR, Peherstorfer B. 2023b. On the sample complexity of stabilizing linear dynamical systems from data. *Found. Comput. Math.* **https://doi.org/10.1007/s10208-023-09605-y**

Willcox KE, Ghattas O, Heimbach P. 2021. The imperative of physics-based modeling and inverse theory in computational science. *Nat. Comput. Sci.* 1:166–68

Willcox K, Ghattas O, van Bloemen Waanders B, Bader B. 2005. An optimization framework for goal-oriented, model-based reduction of large-scale systems. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 2265–71. Piscataway, NJ: IEEE

Williams MO, Kevrekidis IG, Rowley CW. 2015. A data–driven approximation of the Koopman operator: extending dynamic mode decomposition. *J. Nonlinear Sci.* 25(6):1307–46

Zimmermann R, Willcox K. 2016. An accelerated greedy missing point estimation procedure. *SIAM J. Sci. Comput.* 38(5):A2827–50

Annual Review of Fluid Mechanics

Volume 56, 2024

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Fluid Mechanics* articles may be found at http://www.annualreviews.org/errata/fluid