Physica D 449 (2023) 133743

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

A causality-based learning approach for discovering the underlying dynamics of complex systems from partial observations with stochastic parameterization

Nan Chen, Yinling Zhang *

Department of Mathematics, University of Wisconsin-Madison, 480 Lincoln Dr., Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 17 August 2022
Received in revised form 11 February 2023
Accepted 29 March 2023
Available online 3 April 2023
Communicated by R. Kuske

Keywords:
Partial observations
Causality-based learning
Data assimilation
Parameter estimation
Localization
Physics constraints

ABSTRACT

Discovering the underlying dynamics of complex systems from data is an important practical topic. Constrained optimization algorithms are widely utilized and lead to many successes. Yet, such purely data-driven methods may bring about incorrect physics in the presence of random noise and cannot easily handle the situation with incomplete data. In this paper, a new iterative learning algorithm for complex turbulent systems with partial observations is developed that alternates between identifying model structures, recovering unobserved variables, and estimating parameters. First, a causality-based learning approach is utilized for the sparse identification of model structures, which takes into account certain physics knowledge that is pre-learned from data. It has unique advantages in coping with indirect coupling between features and is robust to stochastic noise. A practical algorithm is designed to facilitate causal inference for high-dimensional systems. Next, a systematic nonlinear stochastic parameterization is built to characterize the time evolution of the unobserved variables. Closed analytic formula via efficient nonlinear data assimilation is exploited to sample the trajectories of the unobserved variables, which are then treated as synthetic observations to advance a rapid parameter estimation. Furthermore, the localization of the state variable dependence and the physics constraints are incorporated into the learning procedure. This mitigates the curse of dimensionality and prevents the finite time blow-up issue. Numerical experiments show that the new algorithm identifies the model structure and provides suitable stochastic parameterizations for many complex nonlinear systems with chaotic dynamics, spatiotemporal multiscale structures, intermittency, and extreme events.

1. Introduction

Complex turbulent dynamical systems appear in many areas, such as geophysics, neural science, engineering, and atmosphere and ocean science [1–3]. These complex systems are characterized by strong nonlinearity, high dimensionality, and multiscale structures. The nonlinear interactions between different scales transfer energy throughout the system, which triggers a large dimension of strong instabilities. As a result, many non-Gaussian features, such as extreme and rare events, intermittency, and fat-tailed probability density functions (PDFs), are observed in these systems [4–9]. In addition to improving the description of the phenomena, accurate modeling of these complex systems is the prerequisite of state estimation, uncertainty quantification, and prediction [10–14].

Due to the incomplete physical understanding of nature and the inadequate model resolution resulting from the limited computing power, model error is often inevitable in inferring these complex systems utilizing the purely knowledge-based modeling approaches that aim at revealing the entire model structure based on explicit physical laws [15–19]. To this end, learning the underlying dynamics of these complex systems with the help of data is of practical significance. Appropriately exploiting the data may facilitate the discovery of additional physical structures beyond those obtained from the available partial knowledge. It also provides useful information for developing effective parameterizations that compensate for the inadequate model resolution. The data-driven learning methods can be divided into two categories depending on the amount of available physical understanding of the problem of interest. On the one hand, if there is little prior knowledge of the underlying dynamics, then the model structure and the model parameters have to be learned almost entirely from the data. In such a situation, the learning algorithm typically starts with a large library of candidate functions and then a certain sparse identification technique, such as the least absolute shrinkage and selection operator (LASSO) regression [20,21], or dynamical constraints [22], is incorporated into the optimization procedure for the function selection to obtain a parsimonious

* Corresponding author.
E-mail address: zhang2447@wisc.edu (Y. Zhang).

https://doi.org/10.1016/j.physd.2023.133743

model [23–25]. Note that sparse identification can be combined with other methods, such as the ensemble Kalman inversion [26], to facilitate the learning process. Learning the underlying dynamics from undersampled data utilizing compressive sensing has also been studied in [27]. On the other hand, when part of the underlying dynamics is known, the task becomes more straightforward as the learning algorithm is primarily utilized to discover the residual part. In general, simple closure terms or parameterizations are developed from data to approximate the residual such that the complexity of the discovered model does not increase significantly. These closure methods include using data to fit specified ansatz motivated from fluids or other dynamical systems [28–31], data-driven reduced order models [32–34], closure models with physics constraints [35,36], conditional Gaussian nonlinear systems [37,38], non-Markovian closure models [39–41], statistical or stochastic closure models [42,43], etc. If the available partial dynamics are inaccurate, then systematic learning algorithms can be developed to either correct the model error explicitly from data [44] or introduce additional judicious model errors to offset the existing bias [45–47]. It is also worth mentioning that the learning output is sometimes represented in non-parametric forms. Such an alternative is particularly useful when the primary goal is to forecast the system instead of reaching the explicit physical formulation. This type of learning approach includes developing non-parametric closure models [48–51], subgrid parameterizations [52–54], and machine learning forecast models [55–57].

These data-driven methods have led to many successes in various contexts, especially in building approximate models and forecasting time series. Yet, several challenges still exist in exploiting data-driven approaches to discover the underlying dynamics of complex turbulent systems robustly. First, satisfying the model parsimony is only a necessary but not sufficient condition for identifying the true dynamics. The model identification based on purely data-driven constrained optimization algorithms may not be the best choice to characterize the physical or causal dependence between state variables. As a result, in the presence of even slight random noise, both the covariate selection accuracy and the fraction of zero entries may decrease significantly [58], leading to a significant bias in discovering the underlying dynamics. Second, it is often the case that only the observations of a subset of the state variables are available in practice, known as partial observations. In such a situation, the state estimation of the unobserved variables and the parameterization of these processes have to be carried out simultaneously as the discovery of the underlying dynamics of the observed variables and the parameter estimation. This significantly increases the computational cost since the uncertainty quantification of the estimated unobserved states has to be incorporated into the optimization procedure. In addition, as the dimension of the system becomes large, the number of candidate functions in the library often shoots up. The curse of dimensionality prevents an efficient selection of the most relevant functions.

In this paper, a causality-based iterative learning algorithm is developed, which aims at overcoming the above difficulties in discovering the dynamics of complex turbulent systems from only partial observations. The algorithm alternates between identifying model structures, recovering unobserved variables, and estimating parameters. First, the model identification procedure differs from the LASSO regression and many other straightforward constrained optimizations, where data is directly utilized to compute the loss function that involves a regularizer for the model sparsity. In the proposed approach, a causality-based sparse identification of the model structure is adopted, which takes into account certain physics knowledge that is pre-learned from data. Specifically, in light of the observational data, an information measurement called the causation entropy [58,59] is exploited to detect the possible causal relationship between each candidate function and the time evolution of the associated state variable. The model structure is then determined by retaining those candidate functions demonstrated to be crucial to the underlying dynamics based on the causal inference. Notably, using both linear and nonlinear test models, it has been shown that the causality-based sparse identification approach can have a higher selection accuracy than LASSO regression or elastic net [59]. It also indicates robust results in the presence of indirect coupling between features and stochastic noise [60], which are crucial features of complex turbulent systems. In addition, with the predetermined model structure from the causal relationship, the parameter estimation remains a quadratic optimization problem. Closed analytic formulae are available to efficiently and accurately solve the parameter estimation problem. This is very different from the traditional sparse identification based on a constrained optimization that involves an L1 regularization, which requires more expensive numerical solvers. Second, a systematic nonlinear stochastic parameterization is built to characterize the time evolution of the unobserved state variables, aiming at capturing their statistical feedback to the observed ones [37]. Stochastic parameterization is essential for the identified model, as completely ignoring the contribution from the unobserved variables may lead to a significant bias. To effectively learn the details of the stochastic parameterization, which is often a computationally expensive task utilizing direct optimization algorithms, an efficient nonlinear data assimilation method is developed that exploits closed analytic formulae to sample the trajectories of the unobserved variables [61]. These sampled trajectories are then treated as synthetic observations that allow the entire system to be fully observed, which facilitates the parameter estimation based on a simple maximum likelihood criterion. Finally, the localization of the state variable dependence and the physics constraints with energy-conserving nonlinearity are incorporated into the learning procedure [35,62,63], which overcome the curse of dimensionality and prevent the finite time blow-up issue of the complex systems.

The rest of the paper is organized as follows. The new causality-based learning algorithm with a stochastic parameterization for complex systems with partial observations is developed in Section 2. The quantitative measurements of assessing the learning algorithm are presented in Section 3. Examples of learning prototype complex systems are included in Section 4. The paper is concluded in Section 5.

2. The causality-based data-driven learning approach

2.1. Overview of the method

Let us start with the general formulation of complex nonlinear systems [8,10,64,65],

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = \mathbf{\Phi}(\mathbf{Z}(t)) + \boldsymbol{\sigma}\dot{\mathbf{W}}(t), \tag{1}$$

where $\mathbf{\Phi}(\mathbf{Z}(t))$ consists of any linear and nonlinear functions of the state variable $\mathbf{Z} \in \mathbb{R}^N$, $\boldsymbol{\sigma} \in \mathbb{R}^{N\times d}$ is the noise amplitude and $\dot{\mathbf{W}}(t) \in \mathbb{R}^{d\times 1}$ is a white noise. For the simplicity of presentation, d is assumed to be the same as N and $\boldsymbol{\sigma}$ is assumed to be a constant diagonal matrix, which occurs in many situations. For complex systems in geophysics and fluids, $\mathbf{\Phi}(\mathbf{Z}(t))$ usually contains linear dispersion and dissipation, external forcing, and energy-conserving quadratic nonlinear terms. More complicated and higher order nonlinearity can be included in $\mathbf{\Phi}(\mathbf{Z}(t))$ in other applications.

Next, denote by $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})^{\mathrm{T}}$ a decomposition of the state variables, where both $\mathbf{X}$ and $\mathbf{Y}$ are multivariate with $\mathbf{X} \in \mathbb{R}^{N_1}$,

$\mathbf{Y} \in \mathbb{R}^{N_2}$ and $N_1 + N_2 = N$. In general, $\mathbf{X}$ stands for the large-scale or resolved variables while $\mathbf{Y}$ contains a collection of medium- to small-scale variables or unresolved components. Assume one realization of the time series generated from the true underlying system is available for $\mathbf{X}$ serving as the observations while there is no observational data for $\mathbf{Y}$. To learn the underlying dynamics, a library of linear and nonlinear functions that represent different combinations of the components of $\mathbf{X}$ and $\mathbf{Y}$ is pre-developed. Given an initial guess of the model structure of the observed state variables (hereafter "model structure"), the stochastic parameterizations of the unobserved variables (hereafter "stochastic parameterization"), and the model parameters, the learning algorithm includes an iterative procedure that alternates between three steps until the solution converges.

1. Conditioned on the observed time series of $\mathbf{X}$, apply a conditional sampling to obtain a time series of the unobserved state variables $\mathbf{Y}$. The conditional sampling of $\mathbf{Y}$ can be achieved utilizing closed analytic formulae and is computationally inexpensive.
2. Treating the sampled trajectory of $\mathbf{Y}$ as the artificial "observations", compute the causality-based information transfer from each candidate function in the library to the time evolution of the given state variable. Determine the model structure and the form of the stochastic parameterization based on such a causal inference.
3. Utilize a simple maximum likelihood estimation to compute the coefficients of the above-selected functions.

It is worthwhile to highlight that the library of candidate functions is often chosen subjectively based on the empirical knowledge. The unknown true underlying dynamics may include terms that are outside the library. To this end, it is essential to supplement the governing equations of the observed variables with additional stochastic parameterizations. The nonlinear interactions between the stochastic parameterizations with the observed state variables provide additional features beyond those characterized by candidate functions in the library.

For high-dimensional systems, the localization of the state variable dependence is incorporated into the causal inference such that the information transfer from only a small number of the candidate functions needs to be computed, which can mitigate the curse of dimensionality. On the other hand, physics constraints with energy-conserving nonlinearity are added to the parameter estimation step, which allows the resulting model to capture the fundamental behavior of complex turbulent systems and prevents finite-time blow-up of the solutions.

An overview of the proposed causality-based data-driven learning algorithm with partial observations is summarized in Fig. 1.

2.2. Stochastic parameterization and conditional sampling of the unobserved state variables

The partial observations lead to one of the fundamental challenges in efficiently learning the underlying dynamics, as the uncertainty due to the lack of observations impedes the use of simple closed formulae for the identification of the model structure and the estimation of the model parameters. It is worthwhile to highlight that learning the exact underlying dynamics of these unobserved variables is intrinsically very challenging, if not entirely infeasible, for most complex turbulent systems since the random noise and chaotic behavior of the signal will largely affect the observability of the system. Therefore, it is natural to build stochastic parameterizations for characterizing the unobserved variables that can provide crucial feedback to the observed variables. The nonlinear interactions between the observed variables and these stochastic parameterizations also compensate for the effects that cannot be explicitly represented by the library candidate functions in the governing equations of the observed variables. To facilitate the computational efficiency of the learning process, the following stochastic parameterization structure $\mathbf{Y}$ is incorporated into the general nonlinear process of the observed state variable $\mathbf{X}$ [37,66,67],

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \Big[\mathbf{A_0}(\mathbf{X}, t) + \mathbf{A_1}(\mathbf{X}, t)\mathbf{Y}(t)\Big] + \mathbf{B_1}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{1}}(t), \tag{2a}$$

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = \Big[\mathbf{a_0}(\mathbf{X}, t) + \mathbf{a_1}(\mathbf{X}, t)\mathbf{Y}(t)\Big] + \mathbf{b_2}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{2}}(t). \tag{2b}$$

Recall that $\mathbf{X} \in \mathbb{R}^{N_1}$ and $\mathbf{Y} \in \mathbb{R}^{N_2}$. On the right hand side of (2), $\mathbf{A_1} \in \mathbb{R}^{N_1 \times N_2}$ and $\mathbf{a_1} \in \mathbb{R}^{N_2 \times N_2}$ are matrices, and $\mathbf{A_0} \in \mathbb{R}^{N_1 \times 1}$ and $\mathbf{a_0} \in \mathbb{R}^{N_2 \times 1}$ are vectors. The two independent noise coefficients $\mathbf{B_1} \in \mathbb{R}^{N_1 \times d_1}$ and $\mathbf{b_1} \in \mathbb{R}^{N_2 \times d_2}$ are also vectors while the two white noises $\dot{\mathbf{W}}_{\mathbf{1}} \in \mathbb{R}^{d_1 \times 1}$ and $\dot{\mathbf{W}}_{\mathbf{2}} \in \mathbb{R}^{d_2 \times 1}$ are vectors. The matrices or vectors $\mathbf{A_0}, \mathbf{a_0}, \mathbf{A_1}, \mathbf{a_1}, \mathbf{B_1}$ and $\mathbf{b_2}$ depend nonlinearly on the observed state variable $\mathbf{X}$ and time t. One key feature of the parameterization of the unobserved variable $\mathbf{Y}$ in (2) is that its governing equation (2b) is overall highly nonlinear and can produce strongly non-Gaussian statistics, but the process is conditionally linear with respect to $\mathbf{Y}$ once $\mathbf{X}$ is given. Such a family of stochastic parameterization is widely used in geophysics, climate, atmosphere, and ocean science, such as the stochastic superparameterization, dynamical super-resolution, and various stochastic forecast models in data assimilation [68–72]. Since $\mathbf{Y}$ often denotes the fast, small, and subgrid-scale components of the system, the terms corresponding to the nonlinear self-interaction of $\mathbf{Y}$ mostly involve high frequencies and rapid fluctuations [73]. Thus, these terms can often be effectively characterized by either simple stochastic noise [74–77] or suitable approximations that are nonlinear functions of $\mathbf{X}$ and conditionally linear functions of $\mathbf{Y}$ [38]. The resulting stochastic parameterization in (2b) can successfully capture the dominant dynamics and provide similar statistics feedback to $\mathbf{X}$ as the true system. Another justification of the parameterization in (2) is that many complex nonlinear systems already fit into this coupled modeling framework [37,67], including many physics-constrained nonlinear stochastic models (e.g., the noisy versions of Lorenz models, Charney–DeVore flows, and the paradigm model for topographic mean flow interactions), a large number of stochastically coupled reaction–diffusion models in neuroscience and ecology (e.g., the FitzHugh–Nagumo models and the SIR epidemic models), and a wide class of multiscale models in turbulence and geophysical flows (e.g., the spectrum representations of the Boussinesq equations and the rotating shallow water equation). Note that the feedback from $\mathbf{Y}$ to $\mathbf{X}$ can either be through an additive function or a multiplicative one, with the prefactor being an arbitrary nonlinear function of $\mathbf{X}$.

One desirable feature of the system (2) is that its mathematical structure facilitates an efficient conditional sampling of the trajectory of $\mathbf{Y}$ via a closed analytic formula, which avoids sampling errors from using the particle methods and greatly accelerates the calculation.

Proposition 1 (*Conditional Sampling [61]*)**.** *For the nonlinear system (2), conditioned on one realization of the observed variable $\mathbf{X}(s)$ for $s \in [0, T]$, the optimal strategy of sampling the trajectories associated with the unobserved variable $\mathbf{Y}$ within the same time interval satisfies the following explicit formula,*

$$\frac{\overleftarrow{\mathrm{d}\mathbf{Y}}}{\mathrm{d}t} = \frac{\overleftarrow{\mathrm{d}\boldsymbol{\mu}_{\mathbf{s}}}}{\mathrm{d}t} - \big(\mathbf{a_1} + (\mathbf{b_2}\mathbf{b_2}^{\mathrm{T}})\mathbf{R}_{\mathbf{f}}^{-1}\big)(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{s}}) + \mathbf{b_2}\dot{\mathbf{W}}_{\mathbf{Y}}(t), \tag{3}$$

where $\dot{\mathbf{W}}_{\mathbf{Y}}(t) \in \mathbb{R}^{d_2 \times 1}$ is a Gaussian random noise that is independent from $\dot{\mathbf{W}}_{\mathbf{2}}(t)$ in (2). The variables $\mathbf{R}_{\mathbf{f}} \in \mathbb{R}^{d_2 \times d_2}$ and $\boldsymbol{\mu}_{\mathbf{s}} \in \mathbb{R}^{d_2 \times 1}$ are the filtering covariance, and smoother mean, where the

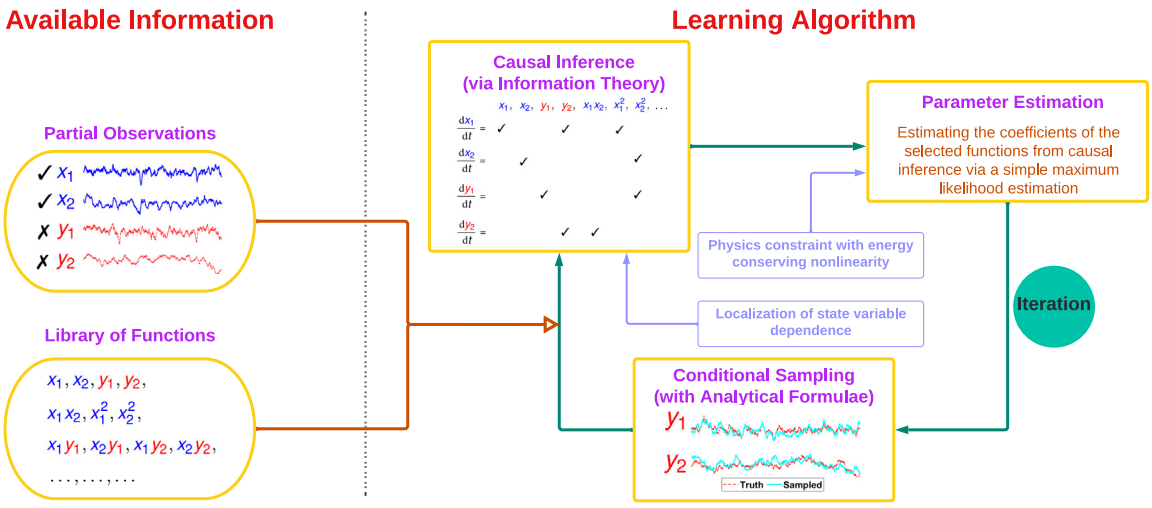

Fig. 1. Schematic diagram of the learning algorithm. Here, $\mathbf{X} = (x_1, x_2)$ and $\mathbf{Y} = (y_1, y_2)$.

filtering and smoothing of $\mathbf{Y}$ *are defined as the following conditional distributions*

$$\begin{aligned} p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t) &\sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{f}(t), \mathbf{R}_\mathbf{f}(t)), \\ p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0, T]) &\sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}(t), \mathbf{R}_\mathbf{s}(t)), \end{aligned} \tag{4}$$

and the associated evolution equations are given explicitly by

$$\frac{\mathrm{d}\boldsymbol{\mu}_\mathbf{f}}{\mathrm{d}t} = (\mathbf{a_0} + \mathbf{a_1}\boldsymbol{\mu}_\mathbf{f}) + (\mathbf{R_f}\mathbf{A_2}^{\mathrm{T}})(\mathbf{B_1}\mathbf{B_1}^{\mathrm{T}})^{-1}\left(\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} - (\mathbf{A_0} + \mathbf{A_1}\boldsymbol{\mu}_\mathbf{f})\right), \tag{5a}$$

$$\frac{\mathrm{d}\mathbf{R_f}}{\mathrm{d}t} = \mathbf{a_1}\mathbf{R_f} + \mathbf{R_f}\mathbf{a_1}^{\mathrm{T}} + \mathbf{b_2}\mathbf{b_2}^{\mathrm{T}} - (\mathbf{R_f}\mathbf{A_1}^{\mathrm{T}})(\mathbf{B_1}\mathbf{B_1}^{\mathrm{T}})^{-1}(\mathbf{A_1}\mathbf{R_f}), \tag{5b}$$

$$\frac{\overleftarrow{\mathrm{d}}\boldsymbol{\mu}_\mathbf{s}}{\mathrm{d}t} = -\mathbf{a_0} - \mathbf{a_1}\boldsymbol{\mu}_\mathbf{s} + (\mathbf{b_2}\mathbf{b_2}^{\mathrm{T}})\mathbf{R_f}^{-1}(\boldsymbol{\mu}_\mathbf{f} - \boldsymbol{\mu}_\mathbf{s}), \tag{5c}$$

$$\frac{\overleftarrow{\mathrm{d}}\mathbf{R_s}}{\mathrm{d}t} = -(\mathbf{a_1} + (\mathbf{b_2}\mathbf{b_2}^{\mathrm{T}})\mathbf{R_f}^{-1})\mathbf{R_s} - \mathbf{R_s}(\mathbf{a_1}^{\mathrm{T}} + (\mathbf{b_2}\mathbf{b_2}^{\mathrm{T}})\mathbf{R_f}) + \mathbf{b_2}\mathbf{b_2}^{\mathrm{T}}. \tag{5d}$$

The notation $\overleftarrow{\mathrm{d}}\cdot/\mathrm{d}t$ *in* (5c)–(5d) *corresponds to the negative of the usual derivative, which means that both the equations are solved backward over* $[0, T]$ *with* $(\boldsymbol{\mu}_\mathbf{s}(T), \mathbf{R_s}(T)) = (\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R_f}(T))$ *after* (5a)–(5b) *have been solved forward over* $[0, T]$. *The starting value of the nonlinear smoother* $(\boldsymbol{\mu}_\mathbf{s}(T), \mathbf{R_s}(T))$ *is the same as the endpoint value of the filter estimate* $(\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R_f}(T))$.

Therefore, the conditional sampling formula in (3) allows us to recover the time series of $\mathbf{Y}$, which, together with the observed trajectory of $\mathbf{X}$ forms a complete set of the time series for the entire system.

2.3. Causal inference for discovering the model dynamics via information theory

Given the observational time series of $\mathbf{X}$ and the sampled trajectories of $\mathbf{Y}$ from the previous step, the next task is to determine the functions in the library that are crucial to the time evolution of each state variable. A small subset of the functions is preferred to be retained to guarantee the sparsity of the identified model. Meanwhile, the causal relationship between different variables is incorporated into the identification process to make the resulting system physically explainable. Recall the collection of the state variables $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})^{\mathrm{T}}$. Denote the components of $\mathbf{Z}$ by $\mathbf{Z} = (z_1, z_2, \ldots, z_N)^{\mathrm{T}}$. Further denote by $\mathbf{F} = (f_1, f_2, \ldots, f_M)^{\mathrm{T}}$ the candidate functions in the pre-determined library. As a starting point, assume all these functions are possible candidates for the dynamics of each z_i, $i = 1, \ldots, N$. Then, after applying a forward Euler temporal discretization scheme, the deterministic part of the starting model of $\mathbf{Z}$ has the following form:

$$\begin{aligned} \begin{bmatrix} z_1(t+\Delta t) \\ z_2(t+\Delta t) \\ \vdots \\ z_N(t+\Delta t) \end{bmatrix} &= \begin{bmatrix} \xi_{1,1} & \cdots & \xi_{1,M} \\ \xi_{2,1} & \cdots & \xi_{2,M} \\ \vdots & \ddots & \vdots \\ \xi_{N,1} & \cdots & \xi_{N,M} \end{bmatrix} \begin{bmatrix} f_1(z_1(t), \ldots, z_N(t), t) \\ f_2(z_1(t), \ldots, z_N(t), t) \\ \vdots \\ f_M(z_1(t), \ldots, z_N(t), t) \end{bmatrix} \Delta t \\ &\quad + \begin{bmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_N(t) \end{bmatrix} \\ \Longrightarrow \mathbf{Z}(t+\Delta t) &= \boldsymbol{\Xi} \times \mathbf{F}(\mathbf{Z}(t), t)\, \Delta t + \mathbf{Z}(t), \end{aligned} \tag{6}$$

where $\boldsymbol{\Xi}$ is the coefficient matrix to be estimated. In general, the size of the matrix $\boldsymbol{\Xi}$ is quite large since M is often a big number. Therefore, physics-informed sparse identification is essential to force most of the entries to be zero.

To incorporate certain physical evidence into this identification process, the following causal inference is utilized [58]. Denote by $C_{f_m \to z_n | [\mathbf{F} \setminus f_m]}$ the causation entropy of $f_m(t)$ on $z_n(t+\Delta t)$. It is defined as the difference between two conditional entropies. One represents the information transfer from the entire $\mathbf{F}(t)$ to $z_n(t+\Delta t)$, and the other one stands for that from all the functions in $\mathbf{F}(t)$ excluding $f_m(t)$. Therefore, $C_{f_m \to z_n | [\mathbf{F} \setminus f_m]}$ allows to explore the composition of $z_n(t+\Delta t)$ that comes solely from $f_m(t)$. If such a causation entropy is zero (or practically nearly zero), then $f_m(t)$ does not have a contribution to $z_n(t+\Delta t)$ or its contribution is indirect and has already been effectively characterized by the other candidate functions. In such a case, the associated parameter $\xi_{n,m}$ is set to be zero. By computing such a causation entropy for different $m = 1, \ldots, M$ and $n = 1, \ldots, N$, a sparse causation entropy matrix is reached, which indicates if each entry of $\boldsymbol{\Xi}$ should be estimated. Note that f_m is supposed to be a non-constant function, and the constant terms are always assumed to exist. After the function selection, a simple maximum likelihood estimation based on a quadratic optimization can be easily applied to determine the actual values of those nonzero entries in $\boldsymbol{\Xi}$, which can often be solved via closed analytic formulae. This is very different from the LASSO regression, where the sparse model identification and parameter estimation have to be carried out simultaneously with a suitable numerical algorithm.

The causation entropy $C_{f_m \to z_n | [\mathbf{F} \backslash f_m]}$ is defined as follows,

$$\begin{aligned} C_{f_m \to z_n | [\mathbf{F} \backslash f_m]} &= H(z_n | [\mathbf{F} \backslash f_m]) - H(z_n | [\mathbf{F} \backslash f_m], f_m) \\ &= H(z_n | [\mathbf{F} \backslash f_m]) - H(z_n | \mathbf{F}). \end{aligned} \tag{7}$$

In (7), $H(\cdot|\cdot)$ is the conditional entropy, which is related to the Shannon's entropy $H(\cdot)$ and the joint entropy $H(\cdot,\cdot)$. They are defined as:

$$\begin{aligned} H(X) &= -\int_x p(x)\log(p(x))\,\mathrm{d}x, \\ H(Y|X) &= -\int_x\int_y p(x,y)\log(p(y|x))\,\mathrm{d}y\,\mathrm{d}x, \\ H(X,Y) &= -\int_x\int_y p(x,y)\log(p(x,y))\,\mathrm{d}y\,\mathrm{d}x, \end{aligned}$$

where p is the associated PDF that is typically determined by a histogram from the given time series (assuming the ergodicity). The difference on the right-hand side of (7) naturally represents the contribution of f_m to z_n.

Note that other information measurements have also been applied for the causality detection, such as the transfer entropy [78,79] and the directed information [80,81]. These methods, in general, work well. Nevertheless, the causation entropy in (7) has its advantages in identifying model structure in the presence of indirect coupling between features and stochastic noise [60], which are crucial features of complex turbulent systems. The former is detected through the conditional entropy while the use of the PDF in the conditional entropy is a more robust measurement with the noise.

The calculation of the causation entropy.

The direct calculation of the causation entropy in (7) is nontrivial and computationally expensive. Reconstructing the exact PDFs from a given time series is challenging. The kernel density estimation (KDE) [82], the box-counting algorithm [83], and many other direct estimation methods suffer from the curse of dimensionality. Some alternative methods have been proposed, such as the k-nearest neighbors [84,85], which can mitigate the issue to some extent but may remain to be complicated. Since determining the model structure only depends on if the causation entropy is zero or not rather than its exact value, the following properties will facilitate the calculation of the causation entropy in high dimensions.

Proposition 2 (*Chain Rule*)**.** *The conditional entropy can be represented by the Shannon's entropy and the joint entropy via the following chain rule:*

$$H(Y|X) = H(X,Y) - H(X). \tag{8}$$

Proposition 3 (*Gaussian Approximation*)**.** *If $p \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ satisfies a s-dimensional Gaussian distribution, then the Shannon's entropy has the following explicit form*

$$H(p) = \frac{s}{2}(1 + \ln(2\pi)) + \frac{1}{2}\ln(\det(\mathbf{R})), \tag{9}$$

where 'det' is the matrix determinant.

With these properties in hand, the practical calculation of the causation entropy can be the following.

Proposition 4 (*Practical Calculation of the Causation Entropy*)**.** *By approximating all the joint and marginal distributions as Gaussians, the causation entropy can be computed in the following way:*

$$\begin{aligned} C_{Z\to X|Y} &= H(X|Y) - H(X|Y,Z) \\ &= H(X,Y) - H(Y) - H(X,Y,Z) + H(Y,Z) \\ &= \frac{1}{2}\ln(\det(\mathbf{R}_{XY})) - \frac{1}{2}\ln(\det(\mathbf{R}_Y)) \\ &\quad - \frac{1}{2}\ln(\det(\mathbf{R}_{XYZ})) + \frac{1}{2}\ln(\det(\mathbf{R}_{YZ})), \end{aligned} \tag{10}$$

where $\mathbf{R}_{XYZ}$ denotes the covariance matrix of the state variables $(X,Y,Z)^{\mathrm{T}}$ and similar for other covariances.

The Gaussian approximation allows us to efficiently determine the structure of the sparse matrix of $\boldsymbol{\Xi}$, where the exact values of the nonzero entries will then be determined via a simple maximum likelihood estimation. Note that the Gaussian approximation is only applied to post-process the calculation of the causation entropy by approximating the resulting non-Gaussian distribution utilizing the first two moments. It does not require linearizing the underlying dynamics to obtain the time series that satisfies a Gaussian distribution. Thus, the results using the Gaussian approximation still reflect the nonlinear nature of the underlying dynamics. The Gaussian approximation has been widely applied to compute various information measurements and lead to reasonably accurate results [43,86–88]. The Gaussian approximation may lead to certain inaccuracies in computing the exact value of the causation entropy if the true distribution is highly non-Gaussian. Nevertheless, it often suffices to detect all the index pairs (m,n) in $\boldsymbol{\Xi}$, associated with which the causation entropy $C_{f_m \to z_n | [\mathbf{F} \backslash f_m]}$ is nonzero (or practically above a small threshold value). Note that such a threshold value is an analog to the level of the regularity in LASSO in determining the model sparsity [23].

Similar to (1), the system by retaining only the functions corresponding to the nonzero causation entropy entries can be written as

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = \widetilde{\boldsymbol{\Phi}}(\mathbf{Z}(t)) + \boldsymbol{\sigma}\dot{\mathbf{W}}(t). \tag{11}$$

Further denote by $\boldsymbol{\Theta}$ the collection of the parameters to be estimated, which correspond to the nonzero entries in $\boldsymbol{\Xi}$.

2.4. Parameter estimation via a simple maximum likelihood estimation

Consider a temporal discretization of (11) using the Euler–Maruyama scheme [89],

$$\mathbf{Z}^{j+1} = \mathbf{Z}^j + \widetilde{\boldsymbol{\Phi}}(\mathbf{Z}^j)\Delta t + \boldsymbol{\sigma}\boldsymbol{\varepsilon}^j\sqrt{\Delta t}, \tag{12}$$

where j is the index in time, Δt is a small time step, and $\boldsymbol{\varepsilon}^j$ is an independent and identically distributed (i.i.d.) standard multidimensional Gaussian random number. For the convenience of presentation, let the dimensions $d = N$ and therefore $\boldsymbol{\sigma} \in \mathbb{R}^{N\times N}$. Further assume $\boldsymbol{\sigma}$ is a diagonal matrix. Denote by $\mathbf{z}^j$ the given numerical value of $\mathbf{Z}^j$ from observations. Further denote by $\mathbf{M}^j\boldsymbol{\Theta} + \mathbf{s}^j := \mathbf{z}^j + \widetilde{\boldsymbol{\Phi}}(\mathbf{z}^j)\Delta t$, namely the deterministic part on the right hand side of (12) evaluated at $\mathbf{z}^j$. Here the nonlinear candidate functions in the library are included in $\mathbf{M}^j$ and $\mathbf{s}^j$, where the former appears as the multiplicative prefactor of the parameters $\boldsymbol{\Theta}$ while the latter appears on its own such as the first term $-\mathbf{Z}^j-$ on the right-hand side of (12). Due to the Euler–Maruyama approximation, the one-step time evolution from $\mathbf{Z}^j$ to $\mathbf{Z}^{j+1}$ is approximated by a linear function within such a short time interval. Therefore, the likelihood can be computed based on a Gaussian distribution,

$$\mathcal{N}(\boldsymbol{\mu}^j, \boldsymbol{\Sigma}) = C|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{z}^{j+1} - \boldsymbol{\mu}^j)^{\mathrm{T}}(\boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}^{j+1} - \boldsymbol{\mu}^j)\right), \tag{13}$$

where the mean and the covariance are given by $\boldsymbol{\mu}^j = \mathbf{M}^j\boldsymbol{\Theta} + \mathbf{s}^j$ and $\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}\Delta t$, respectively. Note that $\boldsymbol{\sigma}$ has been assumed to be a diagonal constant matrix, and therefore $\boldsymbol{\Sigma}$ does not depend on j. Taking a logarithm operation to cancel the exponential function and summing up the likelihood over the entire time period yield

$$\mathcal{L} = \frac{1}{2}\sum_{j=1}^{J}(\mathbf{z}^{j+1}-\mathbf{M}^j\boldsymbol{\Theta}-\mathbf{s}^j)^{\mathrm{T}}(\boldsymbol{\Sigma})^{-1}(\mathbf{z}^{j+1}-\mathbf{M}^j\boldsymbol{\Sigma}-\mathbf{s}^j)-\frac{J}{2}\log|\boldsymbol{\Sigma}|, \quad (14)$$

where $J+1 = \lfloor T/\Delta t \rfloor$ with $\lfloor\cdot\rfloor$ being the floor function that rounds down the result to the nearest integer. To find the minimum of $\mathcal{L}$, it is sufficient to find the zeros of $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Theta}} = 0$ and $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma}} = 0$, which leads to

$$\boldsymbol{\Sigma} = \frac{1}{J}\sum_{j=1}^{J}(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j)(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j)^{\mathrm{T}}, \quad (15a)$$

$$\boldsymbol{\Theta} = \mathbf{D}^{-1}\mathbf{c}, \quad (15b)$$

where

$$\mathbf{D} = \sum_{j=1}^{J}(\mathbf{M}^j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{M}^j \quad \text{and} \quad \mathbf{c} = \sum_{j=1}^{J}(\mathbf{M}^j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}^{j+1} - \mathbf{s}^j). \quad (16)$$

The equations in (15) are solved by first setting $\boldsymbol{\Theta} = \mathbf{0}$ in finding $\boldsymbol{\Sigma}$ in (15a) via essentially the quadratic variation, and then plugging in the result into (16) and (15b) to obtain $\boldsymbol{\Theta}$. The analytic solution in (15)–(16) significantly facilitates the estimation of the parameter values compared with a numerical solver for a non-quadratic optimization problem, for example, in the standard LASSO regression methods. The computational cost in (15)–(16) is proportional to the square of the number of functions in the identified model (in computing $\mathbf{D}$) and to the total number of observational points in time. Since the goal is to find a parsimonious model, the number of the candidate functions remaining in the model is expected to be small, which leads to a relatively low computational cost in solving (15)–(16).

2.5. Physics constraints

Physics constraints, meaning the conservation of energy in the quadratic nonlinear terms, are important properties in many complex turbulent systems and appear in most of the classical geophysics and fluid models [35,36]. The physics constraints prevent the finite-time blow-up of the solutions and facilitate a skillful medium- to long-range forecast. Therefore, taking into account the physics constraints and other constraints is crucial for the learning algorithm, especially in the parameter estimation step. These constraints, together with other constraints, can, in general, be represented in the following way:

$$\mathbf{H}\boldsymbol{\Theta} = \mathbf{g}, \quad (17)$$

where $\mathbf{H}$ and $\mathbf{g}$ are constant matrices. To incorporate these constraints, the Lagrangian multiplier method is applied, which modifies the objective function in (14),

$$\mathcal{L} = \frac{1}{2}\sum_{j=1}^{J}(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j)^{\mathrm{T}}(\boldsymbol{\Sigma})^{-1}(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j) - \frac{J}{2}\log|\boldsymbol{\Sigma}^{-1}| + \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{H}\boldsymbol{\Theta} - \mathbf{g}), \quad (18)$$

where $\boldsymbol{\lambda}$ is the Lagrangian multiplier. The solution to the minimization problem with the new objective function (18) is given as follows,

$$\boldsymbol{\Sigma} = \frac{1}{J}\sum_{j=1}^{J}(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j)(\mathbf{z}^{j+1} - \mathbf{M}^j\boldsymbol{\Theta} - \mathbf{s}^j)^{\mathrm{T}} \quad (19a)$$

$$\boldsymbol{\lambda} = \left(\mathbf{H}\mathbf{D}^{-1}\mathbf{H}^{\mathrm{T}}\right)^{-1}(\mathbf{H}\mathbf{D}^{-1}\mathbf{c} - \mathbf{g}), \quad (19b)$$

$$\boldsymbol{\Theta} = \mathbf{D}^{-1}\left(\mathbf{c} - \mathbf{H}^{\mathrm{T}}\boldsymbol{\lambda}\right), \quad (19c)$$

where $\mathbf{D}$ and $\mathbf{c}$ are defined in (16).

2.6. Localization of the state variable dependence

Localization of the state variable dependence is typical in many complex turbulent systems for modeling large-scale dynamics and stochastic parameterizations. On the one hand, high-dimensional stochastic ordinary differential equations (SDEs) are usually obtained due to the spatial discretization of a stochastic partial differential equation. The advection, diffusion, and dispersion are all local operators, which implies that each state variable in the SDEs interacts with only the nearby few states [64,90]. On the other hand, the stochastic parameterizations of the states at the subgrid scales also depend only on the nearby corresponding large-scale state variables [91–93]. Besides, the idea of localization is widely utilized in data assimilation, and prediction [94–96].

The localization of the state variables is incorporated into the proposed learning algorithm at both the causal detection and the conditional sampling steps. The necessity of localization in causal detection is to mitigate the curse of dimensionality. When the dimension of the system becomes large, the number of functions in the library that includes different combinations of the state variables increases exponentially. As a result, the cost of computing the causation entropy for all these functions also shoots up. The localization, which requires computing the causation entropy of only those functions that involve the local interactions of the state variables, can overcome the curse of dimensionality. Next, the stochastic parameterizations of each component of the subgrid variable $\mathbf{Y}$ in (2b), denoted by $y_{i,j}$, depend only on the associated large-scale observed state variable of $\mathbf{X}$, namely x_i and $x_{i\pm1}, \ldots, x_{i\pm s}$ with s being a small positive integer. This leads to a block covariance matrix of $\mathbf{R}_{\mathbf{f}}$ in (3) when carrying out the conditional sampling to recover the trajectory of $\mathbf{Y}$. In other words, the giant covariance matrix not only becomes sparse but can be divided into several low-dimensional blocks, which are then solved in a parallel way. This significantly facilitates the learning algorithm in applying to high-dimensional systems, which are otherwise difficult to handle due to the heavy computational burden of storing and solving the full covariance matrix.

3. Quantitative assessment of the learning skill

Recall that the learning algorithm exploits stochastic parameterization to compensate for the contribution from the unobserved state variables. This implies that the exact dynamics of these variables can hardly be identified due to the lack of observations. Therefore, focusing on the skill of recovering the dynamics of the observed variables is a natural choice for assessing the learning algorithm. It is also worth noting that stochastic parameterization may affect the dynamics of the observed variables since each parameterized process may involve the combined contribution from several variables in the original system. In addition, due to the chaotic features and the random noise, computing the path-wise error in the identified model related to the truth is not a suitable strategy, as the trajectory from the identified model will diverge from the truth within a short time even starting from the same initial conditions. Because of these reasons, the assessment of the learning algorithm is based on evaluating the error in reproducing the following two crucial statistical quantities of the observed variables. The first quantity is the equilibrium PDF, which represents the long-term statistical behavior of the

system. The second measurement is the temporal autocorrelation function (ACF), which is the correlation of a signal with a delayed copy of itself as a function of delay. It reflects the overall temporal structure of the system. These two assessments are widely used to measure the overall dynamical and statistical behavior of the identified model. Nevertheless, it is worth noting that small errors based on these two statistical assessments are only the necessary conditions for validating learning skills.

3.1. The information distance between two PDFs

Denote by $p(\mathbf{u})$ and $p^M(\mathbf{u})$ the equilibrium PDF of the truth and that of the identified model, where $\mathbf{u}$ is the state variable. A natural way to quantify the error in q compared with p is through the relative entropy $\mathcal{P}(p, p^M)$, [87,97,98],

$$\mathcal{P}(p, p^M) = \int p(\mathbf{u}) \log\left(\frac{p(\mathbf{u})}{p^M(\mathbf{u})}\right) \mathrm{d}\mathbf{u}, \tag{20}$$

which is also known as Kullback–Leibler divergence or information divergence [99–101]. Despite the lack of symmetry, the relative entropy has two attractive features. First, $\mathcal{P}(p, p^M) \geq 0$ with equality if and only if $p = p^M$. A significant value of the relative entropy means a large difference between the two PDFs. Second, $\mathcal{P}(p, p^M)$ is invariant under general nonlinear changes of variables. These provide an attract framework for assessing model errors in many applications [102–106].

3.2. The information distance between two ACFs

Denote by z a scalar variable. The ACF of the time series $u(t)$ is defined as

$$R(t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{u(t+\tau)u^*(\tau)}{\mathrm{Var}(u)} \mathrm{d}\tau, \tag{21}$$

where $\mathrm{Var}(u)$ is the variance of u. The ACF starts from 1 and decays to 0 with possible oscillations. Directly computing the relative error between the two ACFs may not be the best choice. A small phase shift between two ACFs may lead to a big pathwise error, but the associated dynamics may remain similar. Information theory can provide a rigorous and practical way to quantify the difference between two ACFs. The method is based on the fact that the ACF is related to the power spectrum of the time series. Once the ACF is mapped to the spectrum, such a spectral representation facilitates the use of relative entropy to compute the information distance between the two ACFs. The detailed procedure can be found in [43,107]. A summary of the method is included in the following.

According to Khinchin's formula [108], if the autocorrelation function $R(t)$ is smooth and rapid-decay, which is the typical property for most systems, then there exists a non-negative function $E(\lambda) \geq 0$ such that

$$R(t) = \int_{-\infty}^{\infty} e^{i\lambda t} \mathrm{d}F(\lambda), \tag{22}$$

with $\mathrm{d}F(\lambda) = E(\lambda)\mathrm{d}\lambda$ a non-decreasing function. Therefore the spectral representation of the stationary process of u can be constructed as

$$u(t) = \int_{-\infty}^{\infty} e^{i\lambda t} \hat{Z}(d\lambda). \tag{23}$$

Applying the theory to the spectral representation of stationary processes, a one-to-one correspondence between the ACF $R(t)$ and non-negative energy spectrum $E(\lambda)$ can be found. Consider approximating this random process with second-order statistics by a lattice random field with spacing $\Delta\lambda$. By independence, the true increment $\hat{Z}(\Delta\lambda_j) = \hat{Z}(\lambda_j + \Delta\lambda) - \hat{Z}(\lambda_j)$ has the second order Gaussian probability density function approximation

$$\hat{Z}(\Delta\lambda) \sim p_G(x;\lambda)\Delta\lambda = \mathcal{N}(0, E(\lambda)\Delta\lambda),$$

and the corresponding spectral representation from the identified model also has the density function

$$\hat{Z}^M(\Delta\lambda) \sim p_G^M(x;\lambda)\Delta\lambda = \mathcal{N}(0, E^M(\lambda)\Delta\lambda),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian random variable with mean μ and variance σ^2. Since the spectral measure has an independent increment, the truth and the approximated model Gaussian random fields can be approximated by

$$p_G = \prod_j \mathcal{N}(0, E(\lambda_j)\Delta\lambda), \qquad p_G^M = \prod_j \mathcal{N}(0, E^M(\lambda_j)\Delta\lambda).$$

Then the normalized relative entropy between these two Gaussian fields becomes

$$\begin{aligned}\mathcal{P}(p_G, p_G^M) &= \sum_j \mathcal{P}\big(p_G(x;\lambda_j), p_G^M(x;\lambda_j)\big)\Delta\lambda, \\ &\to \int_{-\infty}^{\infty} \mathcal{P}\big(p_G(x;\lambda), p_G^M(x;\lambda)\big)d\lambda, \qquad \text{as } \Delta\lambda \to 0.\end{aligned}$$

In the following, the information distance between the two PDFs (and the two ACFs) is always computed for each one-dimensional component of the observed variables.

4. Test examples

In this section, three nonlinear chaotic or turbulent systems are utilized for testing the learning algorithm developed in the previous section. The first test model is a three-dimensional low-order system, mainly used as a proof-of-concept and to display the detailed procedure of the method. The other two models are spatially-extended multiscale systems, which are adopted to understand the skill of model identification and stochastic parameterizations.

In all these experiments, physics constraints are incorporated into the learning algorithm. Localization is adopted in the second and third experiments, which facilitates the reduction of the computational cost. In all the experiments, the total length of the observation is 500 units while the numerical integration time step is $\Delta t = 0.001$ such that there are in total of 500,000 points in each observational time series.

4.1. Lorenz 1984 model

The first test example is a low-dimensional chaotic system, known as the Lorenz 1984 (L-84) model, which is a simple analog of the global atmospheric circulation [64,65]. It has the following form [109,110]:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= -(y^2+z^2) - a(x-f) + \sigma_x \dot{W}_x, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -bxz + xy - y + g + \sigma_y \dot{W}_y, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= bxy + xz - z + \sigma_z \dot{W}_z.\end{aligned} \tag{24}$$

In (24), the zonal flow x represents the intensity of the mid-latitude westerly wind current, and a wave component exists with y and z representing the cosine and sine phases of a chain of vortices superimposed on the zonal flow. Relative to the zonal flow, the wave variables are scaled so that $x^2 + y^2 + z^2$ is the total scaled energy. These equations can be derived as a Galerkin truncation of the two-layer quasigeostrophic potential vorticity equations in a channel. The additional stochastic noise represents

the interactions between these resolved scale variables and the unresolved ones.

The following parameters are utilized for the test here:

$$a = \frac{1}{4}, \quad b = 4, \quad f = 8, \quad g = 1, \quad \text{and} \quad \sigma_x = \sigma_y = \sigma_z = 0.1, \tag{25}$$

which are the standard parameters that create chaotic behavior [109]. In fact, $f > 1$ is a necessary condition for the zonal flow becoming unstable, forming steadily progressing vortices, while $g > 0$ triggers the chaotic behavior of the entire system.

4.1.1. The experiment setup

In this experiment, y and z are taken as the observed variables, while the observation of x is not directly available. Note that the L-84 model (24) automatically fits into the framework (2) with $\mathbf{X} = (y, z)^{\mathrm{T}}$ and $\mathbf{Y} = x$. This fact, together with the small system noises $\sigma_x = \sigma_y = \sigma_z = 0.1$ utilized here, allows the learning algorithm to have a potential to fully recover the system, including the unobserved process, as the contribution from the deterministic part of the dynamics is only weakly polluted by the system noises.

It is natural to incorporate all the linear and quadratic nonlinear functions of y and z, namely y, z, y^2, z^2, and yz, into the library of the candidate functions. This mimics the general form of the geophysical flows, the nonlinearity of which is dominated by the quadratic terms. In addition, the linear and conditional linear functions of x, namely x, xy, xz, as well as the constant forcing term, are included in the library. All the quadratic nonlinear functions of the three state variables, except x^2 that breaks the structure of (2), are contained in the library of the candidate functions. To further increase the complexity of the library, the cubic terms that satisfy (2) are also added, which contain the quadratic terms of y or z multiplying x but not the quadratic or cubic functions of x itself.

A random and complicated initial model structure is utilized to start the iterative algorithm,

$$\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}t} &= y^2 - z^2 + 2 + (y^2 - z^2)x + \sigma_x \dot{W}_x, \\
\frac{\mathrm{d}y}{\mathrm{d}t} &= -y - 2y^2 + z^2 + 1 + (-y - 8z - yz)x + \sigma_y \dot{W}_y, \\
\frac{\mathrm{d}z}{\mathrm{d}t} &= -z + z^2 - yz + (8y + z + z^2)x + \sigma_z \dot{W}_z.
\end{aligned} \tag{26}$$

The initial values of the noise coefficients in the observed processes, namely σ_y and σ_z, are chosen to be 1. The initial values of these two parameters will not affect the learning algorithm, as they will converge to the truth within one iteration step based on the quadratic variation (19a). On the other hand, σ_x is not uniquely determined from the algorithm since the effect due to the increase of σ_x can be completely offset by decreasing the coefficients in front of x in the observed processes. Therefore, if the primary goal is to learn the dynamics of the observed variables with a reasonable parameterization of the unobserved ones, then an arbitrary value of σ_x can be used. For the simplicity of the study here, $\sigma_x = 0.1$ is set to be known. It is also worthwhile to remark that, as the quadratic variation of the unobserved variable, in general, cannot be directly updated by the conditional sampling, a change of variable to normalize such a diffusion coefficient is often adopted to update such a parameter if the coefficients in front of x in the observed processes are known [111].

Table 1
Comparison of the parameters in the true system (24) and those in the identified model. Since the Frobenius norm of $\mathbf{C} - \mathbf{C}_{true}$ converges to zero, the identified model has exactly the same structure and the truth. They are both rewritten in the form of (27) for the convenience of comparing the model parameters.

	θ_x^x	θ_y^y	θ_z^z	θ_{yy}^x	θ_{zz}^x	θ_{xz}^y
Truth	−0.2500	−1.0000	−1.0000	−1.0000	−1.0000	−4.0000
Identified	−0.2680	−0.9987	−1.0076	−0.9993	−1.0061	−3.9956
	θ_{xy}^y	θ_{xy}^z	θ_{xz}^z	θ_1^x	θ_1^y	θ_1^z
Truth	1.0000	4.0000	1.0000	2.0000	1.0000	0.0000
Identified	0.9993	3.9956	1.0061	2.0223	0.9939	0.0053

4.1.2. Results

Fig. 2 displays the detailed procedure of the learning algorithm. Panel (a) shows the sampled trajectory of the unobserved variable x at the 1st, the 5th, the 50th, and the 110th iterations. It is seen that the sampled trajectory converges to the truth as the number of iterations increases, indicating that the learning process eventually recovers the unobserved trajectory and identifies the model structure.

To better understand the iterative procedure, Panel (b) of Fig. 2 shows the convergence of the model structure towards the truth. Here, a causation entropy matrix indicator $\mathbf{C}$ is introduced, which is of size $N \times M$, where $N = 3$ is the dimension of the system while M is the total number of candidate functions. The matrix $\mathbf{C}$ has the same structure as Ξ in (6) except that $\mathbf{C}$ is a logical matrix with entries being either 0 or 1. If the causation entropy associated with a specific term exceeds the pre-determined threshold (which is 10^{-3} here), then the corresponding entry in $\mathbf{C}$ is set to be 1, meaning that the term should be maintained in the dynamics. Then the Frobenius norm of $\mathbf{C} - \mathbf{C}_{\text{true}}$ is computed, where $\mathbf{C}_{\text{true}}$ is the causation entropy matrix indicator corresponding to the true model (24). It is seen that despite the large gap in the initial random guess of the model structure, there are only 5 terms (corresponding to Frobenius being 2.2361) that are mismatched after 1 iteration (the first point in the curve) and the correct structure is reached after merely 5 iterations. Note that, at the 5th iteration, the sampled trajectory of x (green) in Panel (a) is still far from the truth. Although the model structure is already perfectly identified, additional iterations are still required for the parameters to converge. One desirable feature observed in Fig. 2 is that the model structure does not change after the 5-th iteration, but only the parameters are updated that simultaneously provide the improved sampled trajectory of x. The final parameter values after 120 iterations are shown in Table 1, which are almost indistinguishable from the truth. Here, the model (24) is rewritten in the following form for the convenience of comparing the parameters displayed in Table 1,

$$\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}t} &= \theta_{yy}^x y^2 + \theta_{zz}^x z^2 + \theta_x^x x + \theta_1^x + \sigma_x \dot{W}_x, \\
\frac{\mathrm{d}y}{\mathrm{d}t} &= \theta_{xz}^y xz + \theta_{xy}^y xy + \theta_y^y y + \theta_1^y + \sigma_y \dot{W}_y, \\
\frac{\mathrm{d}z}{\mathrm{d}t} &= \theta_{xy}^z xy + \theta_{xz}^z xz + \theta_z^z z + \theta_1^y + \sigma_z \dot{W}_z.
\end{aligned} \tag{27}$$

Fig. 3 shows the model simulations and the associated statistics using the identified model together with the estimated parameters. Here, the random number generators in the true system and the identified model are set to be the same when generating the time series. In addition, the same initial conditions are applied to simulate the two systems. It is seen that the trajectories from the identified model coincide with the truth quite well up to time $t = 11$, indicating a good path-wise consistency, at least for the short-term behavior. Yet, due to the chaotic nature of the system, the trajectories from the true and the identified models do not

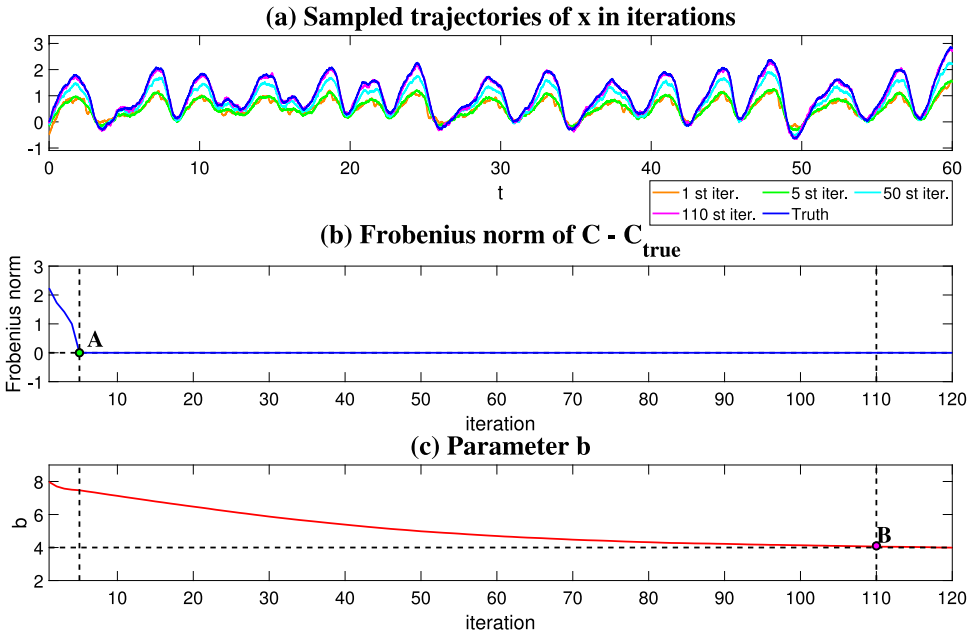

Fig. 2. Iterative procedure of learning the L-84 model with partial observations $(y, z)^{\mathrm{T}}$. Panel (a): The sampled trajectory of the unobserved variable x at the 1st, the 5th, the 50th, and the 110th iterations, where there are in total 120 iterations. The random initial guess of the model structure is shown in (26). Panel (b): The Frobenius norm of $\mathbf{C} - \mathbf{C}_{\text{true}}$ as a function of iterations, where $\mathbf{C}$ is the causation entropy matrix indicator at each iteration step with entries being either 0 or 1 and $\mathbf{C}_{\text{true}}$ is the causation entropy matrix indicator corresponding to the true model (24). Panel (d): The updates of the parameter b. The points **A** and **B** show the iteration at 5 and 110 steps. The model structure is identified correctly after the 5th iteration step, while the parameters converge to the truth at around the 110th step.

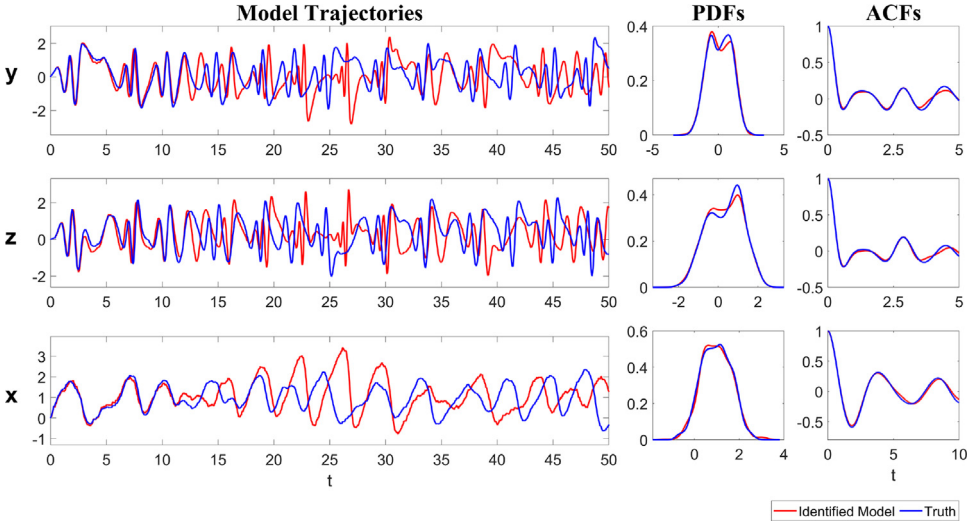

Fig. 3. Comparison of the model trajectories and the associated statistics using the true model (blue) and the identified model (red). The two statistics utilized are the PDF and the temporal ACF. Note that due to the chaotic nature of the system, the two curves do not expect to have a one-to-one point-wise match between each other. Instead, a qualitative similarity between the trajectories from the truth and the identified model is evidence to show the accuracy of the identified model.

expect to have a one-to-one point-wise match between each other for long-term behavior. Nevertheless, a qualitative similarity between the trajectories from the truth and the identified model is observed, which indicates the accuracy of the identified model. This is further confirmed by the nearly perfect recovery of the two statistics: the PDF and the temporal ACF. These facts conclude the skill of the algorithm based on this simple chaotic example.

Finally, a cross-validation test is carried out. Here, the identified model is utilized to generate a time series with the same length as the training signal from the true system for the purpose of model identification. It is found that the model learned from such a time series is very close to the identified model. In addition, starting from the same initial conditions as this time series, the trajectories running forward using the true model coincide with such a time series for a short term, similar to the behavior shown in Fig. 3.

4.2. Two-layer Lorenz 1996 models

The two-layer Lorenz 1996 (L-96) model [112,113] is a conceptual representation of geophysical turbulence that is commonly used as a testbed for data assimilation and parameterization in numerical weather forecasting [69,114–116]. The model mimics a coarse discretization of atmospheric flow on a latitude circle. It

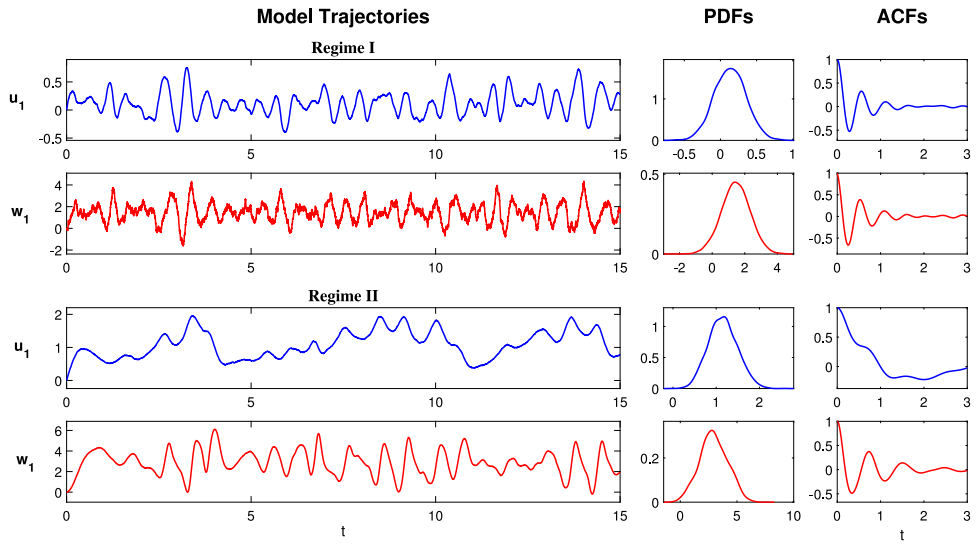

Fig. 4. Dynamical regimes of the Two-layer L-96 model (28), including the model trajectories, PDFs and ACFs of variables u_i and w_i at $i = 1$. Here w_i is defined as $w_i = \sum_{j=1}^{J} v_{i,j}$.

supports complex wave-like and chaotic behavior, and the two-layer structure schematically depicts the interactions between small-scale fluctuations and large-scale motions. The stochastic version of the model subject to additive noise forcing reads

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f - \frac{hc_i}{J}\sum_{j=1}^{J} v_{i,j} + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1, \ldots, I, \tag{28a}$$

$$\frac{\mathrm{d}v_{i,j}}{\mathrm{d}t} = -bc_i v_{i,j+1}\left(v_{i,j+2} - v_{i,j-1}\right) - c_i v_{i,j} + hc_i u_i + \sigma_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J, \tag{28b}$$

where I denotes the total number of large-scale variables, and J is the number of small-scale variables corresponding to each large-scale variable. In (28), f, h, c_i, b, σ_{u_i} and $\sigma_{v_{i,j}}$ are given scalar parameters while $\dot{W}_{u_i}$ and $\dot{W}_{v_{i,j}}$ are independent white noises. The large-scale variables u_i are periodic in i with $u_{i+I} = u_{i-I} = u_i$. The corresponding small-scale variables $v_{i,j}$s are periodic in i with $v_{i+I,j} = v_{i-I,j} = v_{i,j}$ and satisfy the following conditions in j: $v_{i,j+J} = v_{i+1,j}$, and $v_{i,j-J} = v_{i-1,j}$.

Two dynamical regimes are considered here as the truth. They share most of the parameters:

$$I = 20, \quad J = 4, \quad c_i = 2 + 0.7\cos(2\pi i/I), \quad b = 2, \quad f = 4, \quad \sigma_{u_i} = 0.05, \tag{29}$$

but they are differed by h and $\sigma_{v_{i,j}}$:

$$\begin{aligned} &\text{Regime I:} \quad h = 4.0 \quad \text{and} \quad \sigma_{v_{i,j}} = 1.00 \\ &\text{Regime II:} \quad h = 1.5 \quad \text{and} \quad \sigma_{v_{i,j}} = 0.05. \end{aligned} \tag{30}$$

The model trajectories and statistics of these two dynamical regimes are shown in Fig. 4. For the convenience of discussing the behavior of the two layers, a new single variable $w_i = \sum_{j=1}^{J} v_{i,j}$ is introduced, which describes the total variabilities in the second layer. It can be seen that there is no scale separation between u_i and v_i in regime I since the ACFs oscillate and decay in a similar fashion. On the other hand, u_i tends to occur in a slower time scale compared with v_i in Regime II, leading to multiscale features. It is also worth mentioning that the coefficient c_i is spatially varying, giving an inhomogeneous spatial pattern of the system.

4.2.1. The experiment setup

Here the u_i for $i = 1, 2, \ldots, I$ are the observed variables, and all the $v_{i,j}$s for different i and j are the unobserved ones. Since only the time series of u_i are provided while the structure of the true model is unavailable, the number J is unknown to us. Therefore, a natural way to build a suitable model is to incorporate stochastic parameterizations into the processes of u_i. Each of such a stochastic parameterization w_i takes into account the total contributions of the associated $v_{i,j}$ to a specific u_i, which is effectively $w_i = \sum_{j=1}^{J} v_{i,j}$. Therefore, the target model has $2I$ dimensions, with I state variables being the observed ones and the remaining I variables representing the stochastic parameterizations.

Due to the high dimensionality of the problem, the size of the library consisting of the candidate functions will become huge if all possible linear and nonlinear functions up to a certain order are considered. Nevertheless, since the main components of the dynamics, such as the advection and diffusion, involve only local interactions, it is natural to consider such localizations in building the library of the candidate functions. To this end, for each u_i, only the terms involving its adjacent variables u_{i-1}, u_{i-2}, u_{i+1} and u_{i+2} are utilized to construct the candidate functions. The nonlinearity considered here is up to the quadratic terms. In addition, the contribution from the stochastic parameterization needs to be included. One of the simplest choices is to augment the library with one additive term w_i and one multiplicative term $u_i w_i$, where w_i itself is driven by a hidden process representing the stochastic parameterization. Note that other more complicated nonlinear interactions between the state variables u_i and the stochastic parameterizations can be easily included in the library. But for the parsimony of the model, only these two related terms are utilized here. The set of the candidate functions for u_i is then given by a vector $\mathbf{F}_{u_i}$, which includes 23 terms:

$$\begin{aligned} &u_i,\ u_{i-1},\ u_{i-2},\ u_{i+1},\ u_{i+2},\ u_i^2,\ u_{i-1}^2,\ u_{i-2}^2,\ u_{i+1}^2,\\ &u_{i+2}^2,\ u_i u_{i-1},\ u_i u_{i-2},\ u_i u_{i+1},\ u_i u_{i+2},\ u_{i-1}u_{i-2},\ u_{i-1}u_{i+1},\\ &u_{i-1}u_{i+2},\ u_{i-2}u_{i+1},\ u_{i-2}u_{i+2},\ u_{i+1}u_{i+2},\ 1,\ w_i,\ u_i w_i. \end{aligned} \tag{31}$$

The candidate functions in the library allow rich features to appear in the dynamics, such as the diffusion and other quadratic nonlinear interactions that were not in the true system. On the other hand, only 4 terms are included in the library for each w_i, given by another vector $\mathbf{F}_{w_i}$,

$$u_i,\ u_i^2,\ 1,\ w_i. \tag{32}$$

This allows for a simple form of stochastic parameterization. Nevertheless, the nonlinear terms u_i^2 in $\mathbf{F}_{w_i}$ and $u_i w_i$ in $\mathbf{F}_{u_i}$ satisfy the physics constraints.

The initial guess of the model is constructed as follows:

$$\begin{aligned}\frac{\mathrm{d}u_i}{\mathrm{d}t} &= -u_{i-1}\left(u_{i-2}-u_{i+1}\right)+\left(u_{i+1}^2-u_{i-1}u_i\right)\\ &\quad+\left(u_i u_{i+1}-u_{i-1}^2\right)-u_i+f\\ &\quad-\frac{hc_i}{J}w_i+\sigma_{u_i}\dot{W}_{u_i},\quad i=1,\ldots,I,\end{aligned}\tag{33a}$$

$$\frac{\mathrm{d}w_i}{\mathrm{d}t} = hc_i u_i - w_i + \sigma_{w_i}\dot{W}_{w_i},\quad i=1,\ldots,I,\tag{33b}$$

where the terms $u_{i+1}^2-u_{i-1}u_i$ and $u_i u_{i+1}-u_{i-1}^2$ are two pairs of local quadratic advection added beyond those in the true system. Both the pairs of local quadratic advection satisfy the physical constraints.

In the following, a visualization diagram is utilized to represent the identified model structure and parameters. Fig. 5 includes an illustration of the visualization diagram for the starting model. Panels (a)–(c) show the general representation of the sparse coefficient matrix. Panel (d) corresponds to the starting model (33). In the big coefficient matrix, the ith row represents the right-hand side of the equations of u_i and w_i, where u_i and w_i contain 23 and 4 terms, respectively. The order of these terms in the figure is the same as that in (31) and (32). The colors indicate the parameter values.

4.2.2. Results

Let us start with Regime I, where u_i and v_i lie on the same time scale. A threshold of 0.001 is utilized in determining if each entry in the causation entropy matrix should be retained. After 50 iterations, the results converge, where the identified model is shown in Panel (c) of Fig. 6. It is seen that the identified model is qualitatively similar to the truth (Panel (a)). In particular, the inhomogeneous spatial structure in u_i is recovered. In addition, despite the chaotic behavior, the identified model captures the weakly eastward propagation of the individual waves and the westward propagation of the wave envelope in the spatiotemporal pattern of u_i. Fig. 7 compares the model trajectories and statistics. It is clear that the trajectories of u_i generated from the identified model (which uses the same random number generator as the truth) are qualitatively similar to the truth, and the statistics are much more accurate than the initial guess.

Next, it is essential to understand the role of stochastic parameterization. To this end, the so-called bare truncation model (BTM) is adopted for comparison, which only retains the dynamics of u_i but completely omits the equations of w_i. That is, only (28a) is utilized, where all $v_{i,j}$s are set to be zero. Thus, the BTM has a dimension of I. It is shown in Panel (d) of Fig. 6 that if the same parameters as in the true system are adopted for the BTM, then the wave patterns become much more regular than the truth due to the lack of perturbations from the small scales. Even by incorporating a parameter estimation into the BTM, the spatiotemporal pattern of the BTM is different from the truth (Panel (e)). This indicates the critical role of w_i in the original system, especially in such a case that there is no clear scale separation in the true system. Therefore, incorporating stochastic parameterization is essential to characterize its effect. One important finding is that the stochastic parameterization w_i recovers the combined contribution of all the $v_{i,j}$ for $j=1,\ldots,J$, according to Figs. 6–7. This is good evidence that shows the role of the one-dimensional stochastic parameterization in replacing the J-dimensional small-scale features in the true model. Note that the coefficients of $w_i u_i$ are zero in the identified model, which implies that the single additive term w_i is sufficient to parameterize the total contribution of the small-scale feedback.

Yet, this is one undesirable feature in the identified model. That is, several additional terms are remaining in the identified model (e.g., $\mathbf{T_7}$, $\mathbf{T_9}$, $\mathbf{T_{11}}$ and $\mathbf{T_{13}}$), which do not appear in the perfect system. This is because of the specific threshold used here to determine if each candidate function should be kept. The threshold value is $r=0.001$, which is relatively low. Therefore, it is natural to repeat the learning process but increase the threshold. To this end, a higher threshold of 0.01 is utilized, and the results are shown in Fig. 8. It is seen that not only those additional terms but also the advection terms disappear with this high threshold. The reason the advection terms, rather than the damping u_i, the forcing f, and the feedback from w_i, are chosen to be eliminated by the learning algorithm is because of its relatively weak role in the original dynamics in this special regime. In fact, according to Panel (c) of Fig. 8, the spatiotemporal pattern of u_i remains similar to the truth by a glance. In particular, the spatial inhomogeneity, the strengths of the signal, and the frequency at each fixed spatial grid point all resemble the truth. Yet, by a careful comparison with the truth, the weakly eastward propagation of the wave no longer exists, which is obviously due to the ignorance of the advection. Therefore, the comparison here indicates that the threshold value helps determine the importance of different terms in the identified system.

Fig. 9 shows the truth and the identified model in Regime II. Different from Regime I, where u_i and w_i lie in the same time scale and the signal of u_i is quite chaotic, a more precise wave propagation pattern is observed in the spatiotemporal pattern in Regime II. This indicates the more significant role played by the advection, as the feedback from the small-scale variables w_i becomes less dominant. The identified model with the threshold being $r=0.001$ again reproduces most features of the underlying dynamics. On the other hand, if a higher threshold $r=0.01$ is utilized, then the advection is again eliminated by the learning algorithm. However, in this dynamical regime, the spatiotemporal pattern of u_i without the advection becomes quite distinct from the truth due to its missing wave propagations. Fig. 10 shows the model trajectories, the ACFs, and the PDFs, which confirm the skill of the identified model in recovering the dynamical and statistical features of the observed variables u_i. Although there are certain gaps between the truth and the stochastic parameterized process w_i, all that is important for w_i in the identified model is its feedback to u_i but not its exact dynamics.

Finally, Fig. 11 provides a quantitative assessment of the errors in the PDFs and the ACFs of the state variables u_i in the two regimes. See Section 3 for the details of these measurements. The errors in the PDFs are almost negligible while those in the ACFs are also overall small. The quantitative results presented here are consistent with the qualitative ones shown above.

4.3. A stochastically coupled FitzHugh–Nagumo (FHN) model

The last test model for the learning algorithm is the following stochastically coupled FitzHugh–Nagumo (FHN) model. The FHN model is a prototype of an excitable system, which describes the activation and deactivation dynamics of a spiking neuron [117]. Stochastic versions of the FHN model with noise-induced limit cycles were widely studied and applied in the context of stochastic resonance [118–121]. Its spatially extended version has also attracted much attention as a noisy excitable medium [122–125]. By exploiting a finite difference discretization to the diffusion term, the stochastically coupled FHN model in the lattice form is given by

$$\begin{aligned}\epsilon\frac{\mathrm{d}u_i}{\mathrm{d}t} &= \left(d_u(u_{i+1}+u_{i-1}-2u_i)+u_i-\frac{1}{3}u_i^3-v_i\right)+\sqrt{\epsilon}\delta_1\dot{W}_{u_i},\\ \frac{\mathrm{d}v_i}{\mathrm{d}t} &= \left(u_i+a\right)+\delta_2\dot{W}_{v_i},\qquad i=1,\ldots,N,\end{aligned}\tag{34}$$

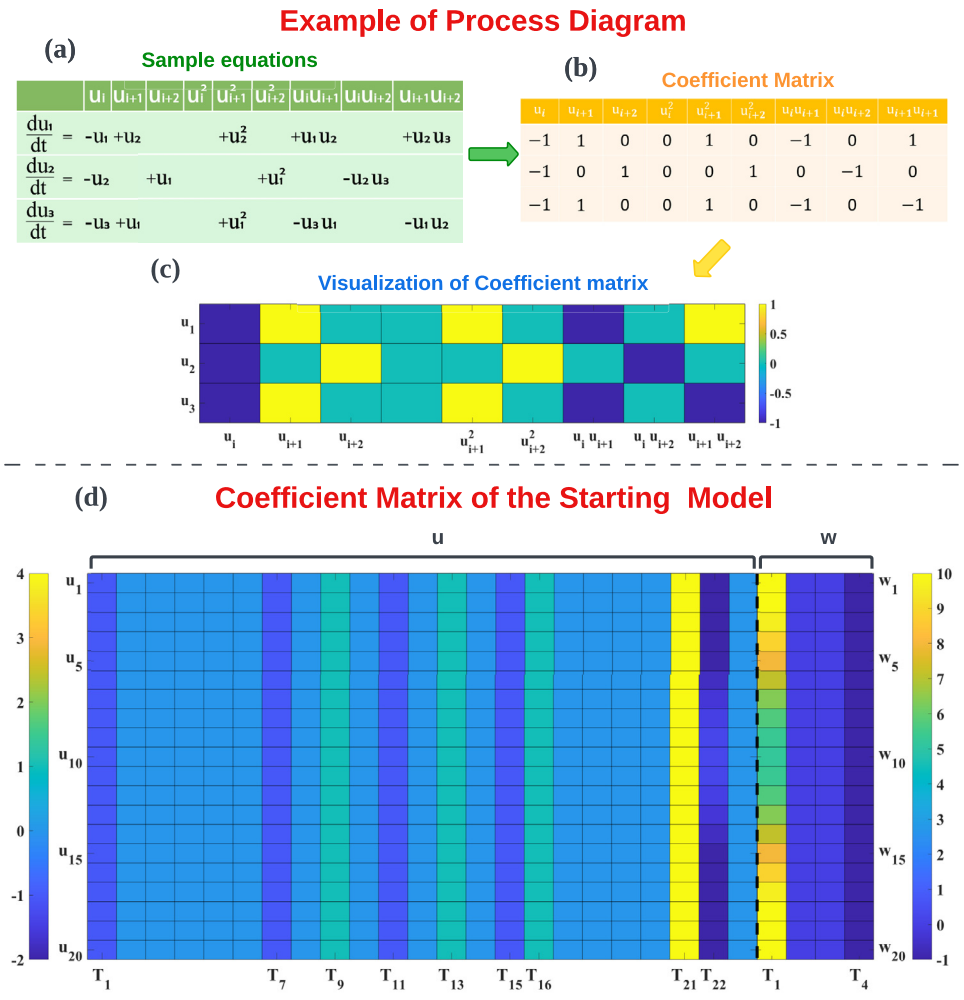

Fig. 5. Using a visualization diagram to represent the identified model structure and parameters. Panels (a)–(c): a low-dimensional ODE system as a simple example to illustrate the visualization of the coefficient matrix. Panel (d): the coefficient matrix of the two-layer L-96 model corresponding to the starting model (33). In the big coefficient matrix, the ith row represents the right hand side of the equations of u_i and w_i, where u_i and w_i contain 24 and 4 terms, respectively. The order of these terms in the figure is the same as that in (31) and (32). The parameter values are indicated by the colors. The $\mathbf{T}_j$ under the u and w parts stands for the jth term in (31) and (32), namely $\mathbf{F}_{u_i}$ and $\mathbf{F}_{w_i}$, respectively.

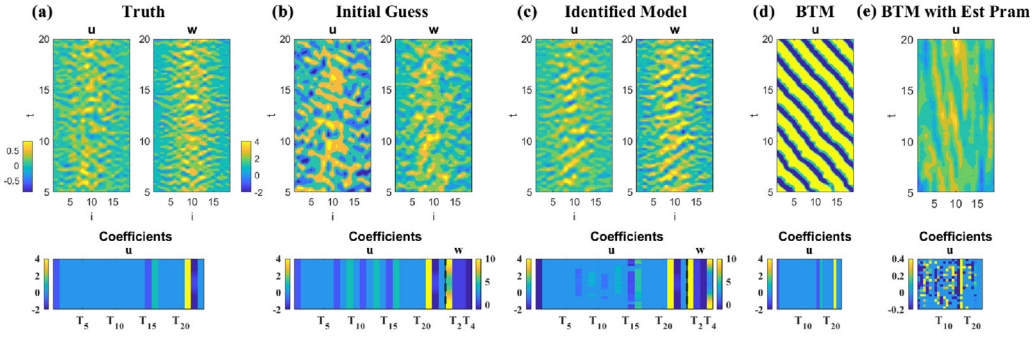

Fig. 6. Identifying the two-layer L-96 model in Regime I. Different columns show the truth, the initial guess, the identified model, and the bare truncation models (BTMs). The first row displays the spatiotemporal pattern of both u_i and w_i. Note that different random number seeds are utilized in different columns, and there is no point-wise correspondence between different patterns. The focus is only on the overall structure. The second row shows the coefficient matrix in each model, as was described in Fig. 5. The threshold value in determining the causation matrix for the identified model with stochastic parameterization is 0.001 and that for the BTM model (Panel (e)) is 0.0001 since the same threshold as the former leads to an even worse result.

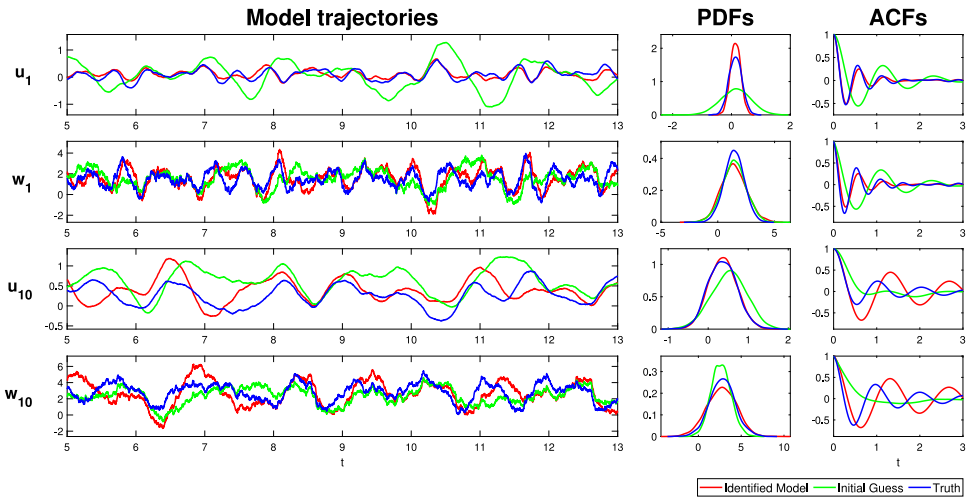

Fig. 7. Comparison of the model trajectories, the ACFs and the PDFs of the truth, the initial guess of the model, and the identified model in Regime I, at $i = 1$ and $i = 10$. Similar to Fig. 6, different random number seeds are utilized in different models, and there is no point-wise correspondence between the trajectories from different models. Only the qualitative features are the useful information obtained from these trajectories. The threshold value in determining the causation matrix is 0.001.

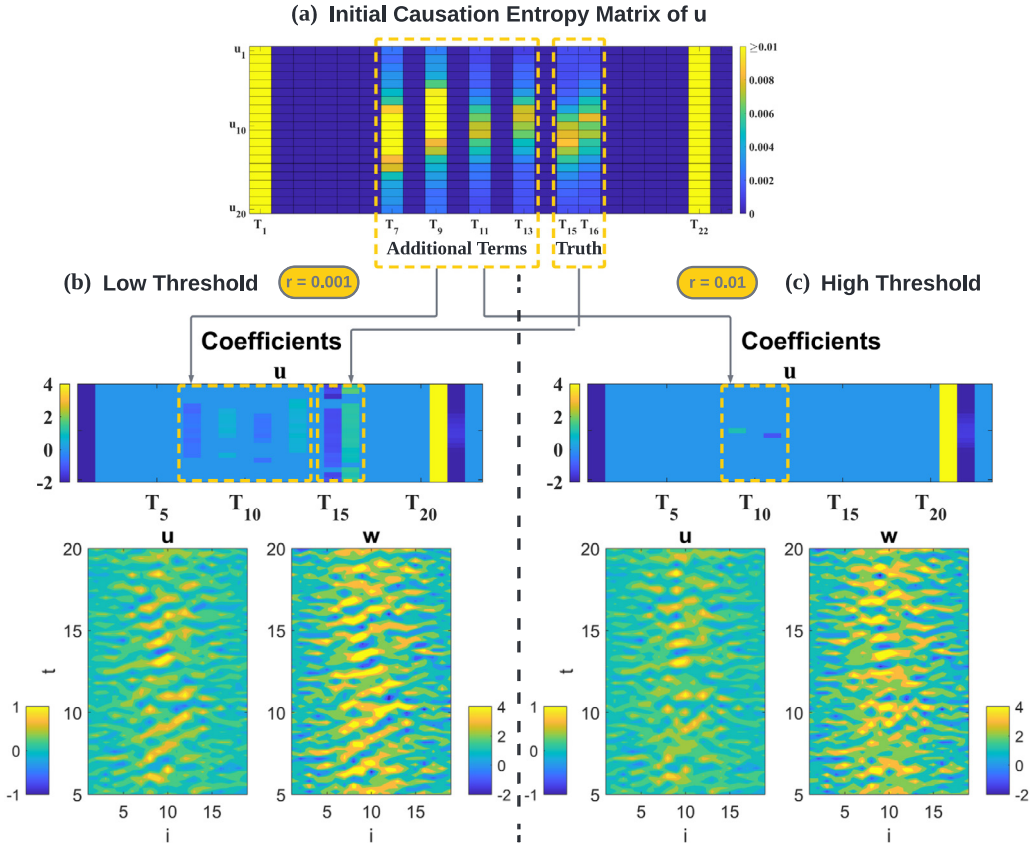

Fig. 8. Comparison of the identified model of the two-layer L96 model in Regime I using different threshold values for determining the causation entropy matrix. Panel (a): the causation entropy matrix of u_i corresponding to the initially guessed structure except for column $\mathbf{T_{21}}$ representing trivial constant terms. Panel (b): the coefficient matrix for u_i based on the low threshold $r = 0.001$ and the corresponding spatial–temporal patterns. Panel (c): The coefficient matrix for u based on the high threshold $r = 0.01$ and the corresponding spatial–temporal patterns.

where the parameter $a > 1$ is required to guarantee that the system has a global attractor in the absence of noise and diffusion. The random noise can drive the system above the threshold level of global stability and triggers limit cycles intermittently. The

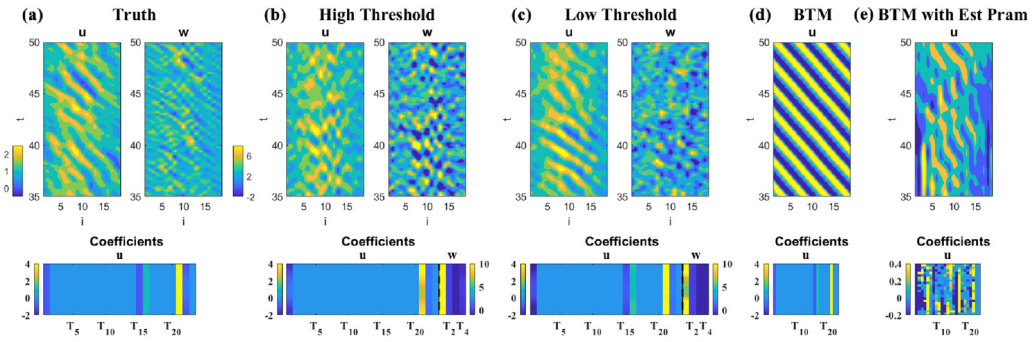

Fig. 9. Similar to Fig. 6 but for Regime II. Note that instead of repeating the column for the initial guess, the column for the high threshold case $r = 0.01$ is shown instead.

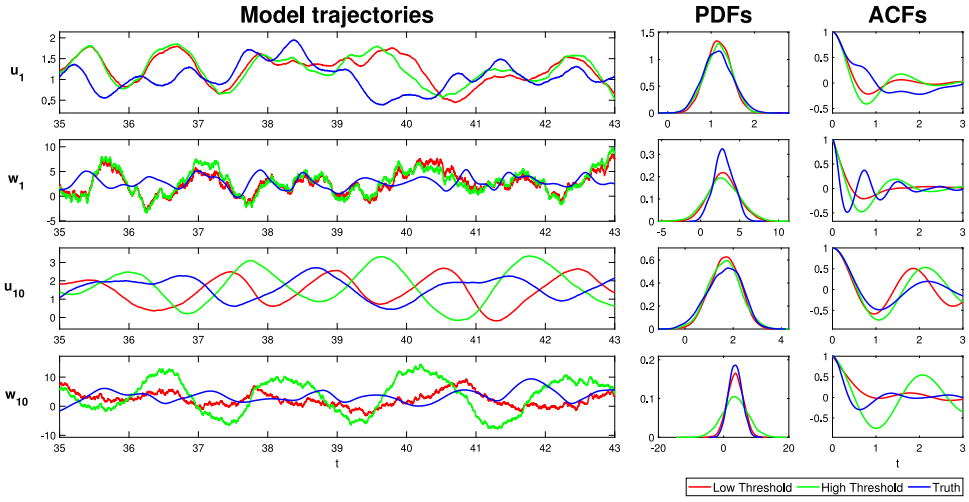

Fig. 10. Similar to Fig. 7 but for Regime II. Note that instead of repeating the curves for the initial guess, the curves for the high threshold case $r = 0.01$ are shown instead.

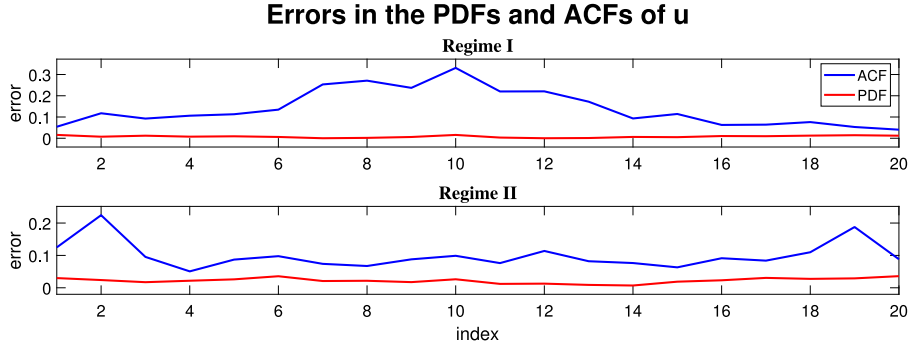

Fig. 11. The relative entropy for PDFs and relative error for ACFs of the state variable u_i, $i = 1, 2, \ldots, I$. In Regime I, the errors are computed between the truth and the identified model in Fig. 6. In Regime II, the errors are computed between the truth and the identified model under low threshold in Fig. 7.

time scale ratio ϵ is much smaller than one, implying that the u_i are fast and the v_i are the slow variables. The model in (34) is equipped with the spatial periodic boundary conditions.

The following parameters are utilized to generate the observational time series:

$$\epsilon = 0.01, \quad \delta_1 = 0.2, \quad \delta_2 = 0.1, \quad d_u = 10, \quad \text{and} \quad a = 1.05. \tag{35}$$

The number of the spatial grid points is $N = 40$. With these parameters, the structure of the solution exhibits strong spatial coherent structures in u_i. See Panel (a) of Fig. 12.

4.3.1. The experiment setup

The time series of u_i for $i = 1, 2, 3, \ldots, N$ are observed while there is no direct observation for v_i. Similar to the setup in the L-96 model, define a vector $\mathbf{F}_{u_i}$ that contains 28 candidates functions for each u_i:

$$u_i,\ u_{i-1},\ u_{i-2},\ u_{i+1},\ u_{i+2},\ u_i^2,\ u_{i-1}^2,\ u_{i-2}^2,\ u_{i+1}^2,\ u_{i+2}^2,\ u_iu_{i-1},\ u_iu_{i-2},\ u_iu_{i+1},\ u_iu_{i+2},\ u_{i-1}u_{i-2},\ u_{i-1}u_{i+1},\ u_{i-1}u_{i+2},\ u_{i-2}u_{i+1},\ u_{i-2}u_{i+2},\ u_{i+1}u_{i+2},\ u_i^3,\ u_{i-1}^3,\ u_{i-2}^3,\ u_{i+1}^3,\ u_{i+2}^3,\ 1,\ v_i,\ u_iv_i, \tag{36}$$

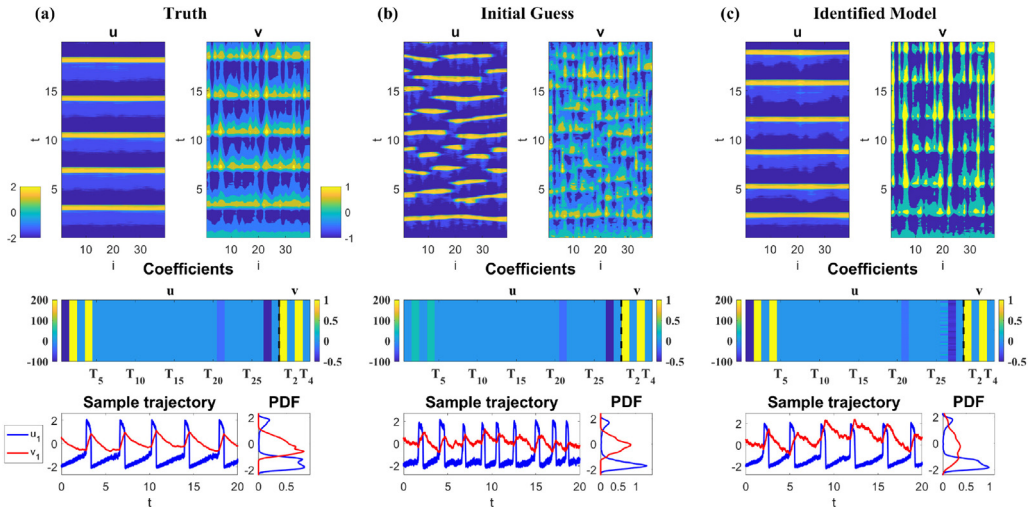

Fig. 12. Identification of the FHN model (34) for the first experiment. Panel (a): the true model structure. Panel (b): the initial guess. Panel (c): the identified model. In each panel, the first row shows the spatiotemporal patterns of the u_i and the v_i. The second row shows the model structure and the associated parameters, where the 28 columns of each u_i correspond to the 28 candidate functions in (36) ordered in the same way and similar for the 4 columns of each v_i. The third row shows the time series of u_1 and v_1.

and another vector $\mathbf{F}_{v_i}$ that includes 4 candidate functions for each v_i:

$$u_i,\ u_i^2,\ 1,\ v_i. \tag{37}$$

When these candidate functions are defined, localization has been utilized as the process of each u_i only depends on those terms within the nearby grid points. Nevertheless, compared with the true system, many more additional advection, diffusion and other quadratic nonlinear terms appear in the candidate functions, which are not in the stochastically coupled FHN system (34). The study here contains the following two experiments.

In the first experiment, the same model structure is taken in the initial guess as in the truth (34). However, very different parameters $d_u = 0.5$ and $\delta_2 = 0.4$ are utilized in the initial guess, which lead to a completely distinct spatiotemporal structure compared with the truth. In fact, as is shown in Panel (b) of Fig. 12, the coherent structure of the model corresponding to the initial guess only appears intermittently and happens in local regions. In addition, the time series of u_i at a fixed spatial location i has a much higher frequency than that of the truth.

In the second experiment, two kinds of the additional terms are added to the starting model: (i) the local quadratic advection satisfying the physical constraints such as u_{i+1}^2 and u_iu_{i-1}, u_{i-1}^2 and u_iu_{i+1}, and (ii) $-v_iu_i$ in the equations of u_i and the quadratic term of u_i^2 in the equations of v_i. In addition, $d_u = 0.5$ and $\delta_2 = 0.1$ are adopted for the initial guess of the model. The initial model reads:

$$\begin{aligned}\epsilon\frac{\mathrm{d}u_i}{\mathrm{d}t} &= \frac{1}{2}(u_{i+1}+u_{i-1}) - \frac{1}{3}u_i^2 - v_i \\ &\quad + \left(u_{i+1}^2 - u_{i-1}u_i\right) + \left(u_iu_{i+1} - u_{i-1}^2\right) \\ &\quad + \left(u_iu_{i+2} - u_{i-2}^2\right) + \epsilon u_iv_i + \sqrt{\epsilon}\delta_1\dot{W}_{u_i}, \quad i = 1,\dots,I,\end{aligned} \tag{38a}$$

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = \left(u_i - u_i^2 + a\right) + \delta_2\dot{W}_{v_i}, \qquad i = 1,\dots,N, \tag{38b}$$

where the terms $\left(u_{i+1}^2 - u_{i-1}u_i\right)$, $\left(u_iu_{i+1} - u_{i-1}^2\right)$ and $\left(u_iu_{i+2} - u_{i-2}^2\right)$ are the three pairs of local quadratic advection added in the starting model satisfying the physics constraints. Panel (b) of Fig. 13 shows the spatiotemporal patterns of such a starting model. It has much faster temporal frequencies than the truth despite the coherent structures.

4.3.2. Results

The results of the two experiments are shown in Fig. 12 and Fig. 13, respectively, where Panel (c) in each figure displays the identified model. In both experiments, the algorithm converges after 10 iterations. The second row depicts the structures and the associated parameter values of the perfect model, the initial guess, and the identified model, where the 28 columns of each u_i correspond to the 28 candidate functions in (36) ordered in the same way and similar for the 4 columns of each v_i.

From Fig. 12, it is seen that although the noise coefficient $\delta_2 = 0.4$ in the unobserved process is fixed and is chosen to be different from the truth $\delta_2 = 0.1$ that leads to a distinct spatiotemporal pattern of the v_i, the spatiotemporal pattern of the u_i generated from the identified model is almost the same as the truth. Specifically, the identified model recovers the strong coherent spatiotemporal structure, which is completely missed in the initial guess of the model. The parameters in the identified model also have similar values to the truth. In particular, the large error in the initial guess of the diffusion coefficient d_u, corresponding to coefficients #1, #2, and #4 in the figure (see also (36)), is almost eliminated in the identified model. This is the crucial mechanism that leads to a strong coherent structure. Note that, as the v_i are not observed, they are treated as the stochastic parameterizations in the identified model. Therefore, the v_i in the identified model is not necessarily the same as the truth. Still, their statistical feedback to each u_i is the crucial component that leads to the correct pattern of the latter, which the identified model is captured.

Fig. 13 illustrates similar results. Despite the significant difference in the model structure and the model parameters in the initial guess, the identified model recovers the truth accurately. Note that, as the true noise coefficient δ_2 is utilized in the identified model, the pattern of v_i is also clearly identified in addition to the recovery of the strong coherent structure of the u_i.

Finally, Fig. 14 shows the quantitative comparison of errors in PDFs and the ACFs of state variable u_i in the two experiments. Overall, the errors remain low, confirmed by the qualitative results shown in Figs. 12–13.

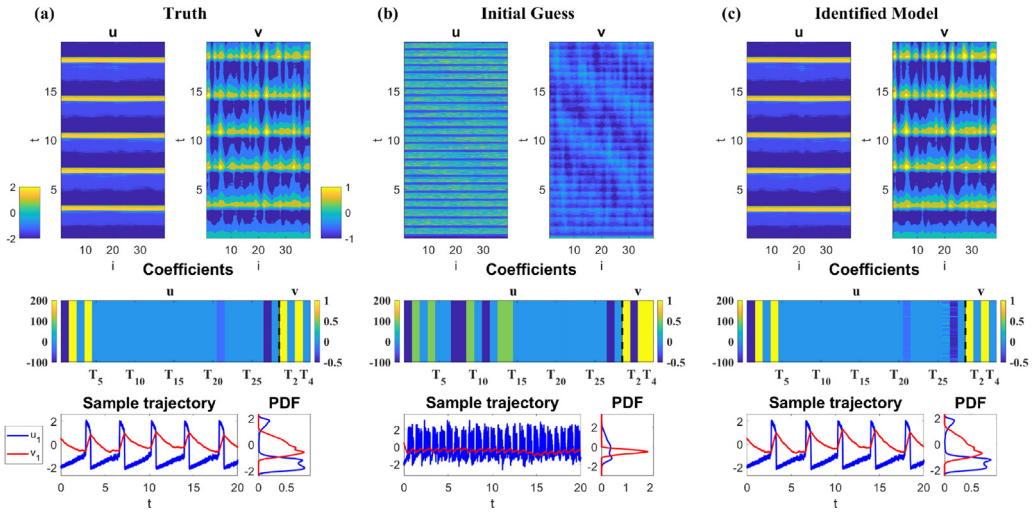

Fig. 13. Similar to Fig. 12 but for the second experiment.

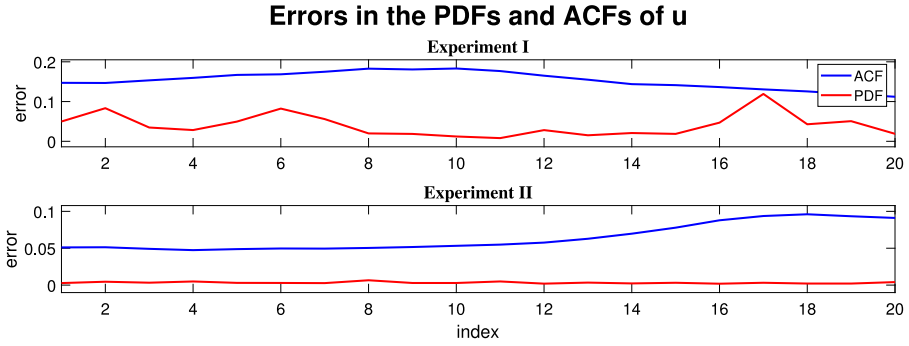

Fig. 14. The relative entropy for PDFs and relative error for ACFs of the state variable u_i, $i = 1, 2, \ldots, I$. The errors are computed between the truth and the identified model of two experiments in Figs. 12 and 13.

5. Conclusions and discussion

5.1. Summary of the work

In this paper, a causality-based learning algorithm is developed that alternatives between model structure identification, conditional sampling of unobserved state variables, and parameter estimation to learn the underlying dynamics of the resolved scale variables and help develop suitable stochastic parameterization for the unresolved variables. Unlike constrained optimization with an L1 regularizer, the method developed here exploits the causation entropy to pre-determine the candidate functions that have potential contributions to the dynamics, which retains a quadratic optimization problem for parameter estimation via maximum likelihood estimates. The closed analytic formula of conditional sampling allows efficient recovery of the unobserved trajectories that facilitates the calculation of the causation entropy for the time evolutions of both the observed and the unobserved state variables. Physics constraints and localization techniques are further incorporated into the learning algorithm to include the basic physical properties in the data-driven models and reduce the computational cost.

A hierarchy of chaotic and turbulent systems is adopted as test models. It is shown that the new learning algorithm effectively reproduces the dynamical and statistical features of the observed variables and provides suitable stochastic parameterizations with parsimonious structures. Many related topics are studied when implementing the numerical tests, including showing the necessity of stochastic parameterization, understanding the effect of choosing different thresholds of the causation matrix selection, and detecting the importance of various terms in the original systems.

Although all the examples in this paper involve stochasticity, the algorithm applies to situations where the observed turbulent signal is generated from a nonlinear deterministic model. Nevertheless, stochasticity plays an essential role in the identified model. The stochastic parameterization $\mathbf{Y}$ is utilized to compensate for the small-scale or unresolved-scale features induced by the nonlinearity that may not be explicitly included in the governing equation of $\mathbf{X}$. It also accounts for the uncertainty and model error. Note that the conditional sampling formulas (3)–(5) require the noise coefficients $\mathbf{B_1}$ in the observed process $\mathbf{X}$ to be nonzero otherwise the sampled trajectories of the additional parameterizations $\mathbf{Y}$ become deterministic.

5.2. Comparison with other methods

Adding stochastic parameterization into the identified model is one of the fundamental differences in the method developed here compared with many traditional data-driven learning algorithms. It has been shown in Fig. 6 that stochastic parameterization facilitates model identification, where a much more significant barrier is found between the truth and the optimal model without stochastic parameterization. As the library of candidate functions is chosen objectively and the true dynamics of many complex systems in practice are very complicated, the true dynamics may not be fully characterized by the functions from the library. Therefore, the additional stochastic parameterization

plays a vital role in compensating for the effect of the missing dynamics.

Another main feature of the proposed method is that the sparse identification is based on causation entropy instead of a direct L1 regularization via LASSO regression. The motivation to adopt the causation entropy is that it can have a higher selection accuracy than LASSO regression or elastic net [59]. Thus, the identified more is more physically explainable. The causality-based learning algorithm also shows robust results in the presence of indirect coupling between features and stochastic noise [60].

The proposed framework also benefits the parameter estimation procedure. With the pre-determined model structure from the causal relationship, the parameter estimation remains a quadratic optimization problem. Closed analytic formulae are available to efficiently and accurately solve the parameter estimation problem, even in the presence of constraints (19). This differs from the LASSO regression, which often requires more expensive numerical solvers than simple closed analytic expressions.

5.3. Potential improvement of the proposed method and future work

5.3.1. Building model closures

In addition to the library of the candidate functions, the work here does not assume any known model structure based on prior knowledge. One natural extension of the current study is to learn the statistical closure of a turbulent system, where part of the model information is given. In such a situation, causation entropy can play an essential role in justifying the given prior model and then identifying the structure of the closure. Another potential extension of the current framework is to combine the parametric form of the dynamics with non-parametric (or machine learning) structures to further improve the results. As was seen in Figs. 6–7, the stochastic parameterization improves the outcomes to a large extent. Yet, the results from the identified model with stochastic parameterization still differ from the truth. Imposing additional machine learning closure is expected to build such a gap.

5.3.2. Optimal solution

The iterative algorithm designed here only aims at finding a local optimum. This is sufficient for many applications. Yet, if the initial guess is very far from the truth or suitable local optimums, then the identified model may not fully capture the dynamical and statistical features of nature. Therefore, additional criteria can be incorporated into the learning procedure to facilitate the convergence of the algorithm, at least towards an appropriate local optimum. For example, the learning algorithm can start with multiple initial guesses. After a few iterations, the associated results are compared to decide the suitable ones for carrying out further iterations.

5.3.3. Extension to discrete-in-time observations

All the numerical test results here assumed continuous-in-time observations. In other words, the observational time step is set to be the same as the numerical integration time step in the numerical implementation. This assumption facilitates using the conditional sampling algorithm in Proposition 1, which was derived for a given continuous time series of $\mathbf{X}$.

The model identification framework developed here also applies to discrete-in-time observations. But additional procedure is needed. There are two possible approaches. One is to adopt a set of difference equations as the governing equations. Then closed analytic formulae for the conditional sampling, the causal inference, and the parameters are all available as analogs to the continuous-in-time case. In particular, the conditional linear difference equations can be utilized for the stochastic parameterizations, which lead to analytically solvable difference equations for the conditionally sampled trajectories [66]. Another potential method for handling discrete-in-time observations is to implement conditional sampling for not only recovering the unobserved variables but also filling in the gap between the two nearby observations, also known as data augmentation [126]. This allows the identified underlying dynamics to be differential equations instead of difference equations. The ensemble Kalman smoother [127] can be adopted to achieve this goal. Note that a large observational time step may deteriorate the skill of the identified model as a significant amount of information can be missed. Typically, the observational time step is expected to be smaller than the decorrelation time of the underlying dynamics.

5.3.4. Sequential model identification

A fixed observed trajectory of $\mathbf{X}$ is utilized when the learning algorithm was applied to the above test examples. Nevertheless, the algorithm can be carried out sequentially when new observations of $\mathbf{X}$ become available. One advantage of sequential learning is that it can naturally be combined with data assimilation to handle noisy observations. Such a sequential learning idea has been developed and applied to several geophysical systems [128,129]. When sequential learning is applied to the current framework, the identified model serves as the forecast model in data assimilation to reduce the noise in observations. The resulting time series up to the current observational time instant is then utilized to compute the causation entropy. Another topic related to sequential model identification is to detect of regime-switching and the change of model structure. The proposed framework here may have the potential to efficiently recognize these changes. Recall that the model structure identification here is separated from parameter estimation. Therefore, only the causation entropies need to be repeatedly calculated sequentially each time when a small batch of new observations arrives. In contrast, the parameter estimation only requires to be carried out once the regime-switching is confirmed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research of N.C. is partially funded by Army Research Office (Grant No. W911NF-23-1-0118), Office of Naval Research (Grant No. N00014-21-1-2904), and National Science Foundation (Grant No. 2118399). Y.Z. is supported as a PhD research assistant under the last two grants. The authors wish to thank the two anonymous reviewers for their constructive comments. All authors approved the version of the manuscript to be published.

References

[1] S.H. Strogatz, Nonlinear Dynamics and Chaos with Student Solutions Manual: With Applications to Physics, Biology, Chemistry, and Engineering, CRC Press, 2018.

[2] S.A. Sheard, A. Mostashari, Principles of complex systems for systems engineering, Syst. Eng. 12 (4) (2009) 295–311.

[3] M. Ghil, S. Childress, Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics, Springer Science & Business Media, 2012.

[4] T.P. Sapsis, Statistics of extreme events in fluid flows and waves, Annu. Rev. Fluid Mech. 53 (2021) 85–111.

[5] V. Lucarini, D. Faranda, J.M.M. de Freitas, M. Holland, T. Kuna, M. Nicol, M. Todd, S. Vaienti, et al., Extremes and Recurrence in Dynamical Systems, John Wiley & Sons, 2016.

[6] C.L. Franzke, T.J. O'Kane, J. Berner, P.D. Williams, V. Lucarini, Stochastic climate theory and modeling, Wiley Interdiscip. Rev. Clim. Change 6 (1) (2015) 63–78.

[7] D.C. Wilcox, Multiscale model for turbulent flows, AIAA J. 26 (11) (1988) 1311–1320.

[8] A.J. Majda, Introduction to Turbulent Dynamical Systems in Complex Systems, Springer, 2016.

[9] W.-K. Tao, J.-D. Chern, R. Atlas, D. Randall, M. Khairoutdinov, J.-L. Li, D.E. Waliser, A. Hou, X. Lin, C. Peters-Lidard, et al., A multiscale modeling system: Developments, applications, and critical issues, Bull. Am. Meteorol. Soc. 90 (4) (2009) 515–534.

[10] E. Kalnay, Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, 2003.

[11] W. Lahoz, B. Khattatov, R. Ménard, Data assimilation and information, in: Data Assimilation, Springer, 2010, pp. 3–12.

[12] A.J. Majda, J. Harlim, Filtering Complex Turbulent Systems, Cambridge University Press, 2012.

[13] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, Springer Science & Business Media, 2009.

[14] K. Law, A. Stuart, K. Zygalakis, Data Assimilation, Vol. 214, Springer, Cham, Switzerland, 2015.

[15] T.N. Palmer, A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models, Q. J. R. Meteorol. Soc. 127 (572) (2001) 279–304.

[16] A.J. Majda, M. Branicki, Lessons in uncertainty quantification for turbulent dynamical systems, Discrete Contin. Dyn. Syst.-A 32 (9) (2012) 3133–3221.

[17] D. Orrell, L. Smith, J. Barkmeijer, T. Palmer, Model error in weather forecasting, Nonlinear Processes Geophys. 8 (6) (2001) 357–371.

[18] X.-M. Hu, F. Zhang, J.W. Nielsen-Gammon, Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: A real-data study, Geophys. Res. Lett. 37 (8) (2010).

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

[20] F. Santosa, W.W. Symes, Linear inversion of band-limited reflection seismograms, SIAM J. Sci. Stat. Comput. 7 (4) (1986) 1307–1330.

[21] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[22] P. Goyal, P. Benner, Learning dynamics from noisy measurements using deep learning with a Runge-Kutta constraint, 2021, arXiv preprint arXiv:2109.11446.

[23] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.

[24] L. Boninsegna, F. Nüske, C. Clementi, Sparse learning of stochastic dynamical equations, J. Chem. Phys. 148 (24) (2018) 241723.

[25] A. Cortiella, K.-C. Park, A. Doostan, Sparse identification of nonlinear dynamical systems via reweighted $\ell 1$-regularized least squares, Comput. Methods Appl. Mech. Engrg. 376 (2021) 113620.

[26] T. Schneider, A.M. Stuart, J.-L. Wu, Learning stochastic closures using ensemble Kalman inversion, Trans. Math. Appl. 5 (1) (2021) tnab003.

[27] H. Schaeffer, G. Tran, R. Ward, Extracting sparse high-dimensional dynamics from limited data, SIAM J. Appl. Math. 78 (6) (2018) 3279–3295.

[28] S.E. Ahmed, S. Pawar, O. San, A. Rasheed, T. Iliescu, B.R. Noack, On closures for reduced order models–A spectrum of first-principle to machine-learned avenues, Phys. Fluids 33 (9) (2021) 091301.

[29] F. Smarra, A. Jain, T. De Rubeis, D. Ambrosini, A. D'Innocenzo, R. Mangharam, Data-driven model predictive control using random forests for building energy optimization and climate control, Appl. Energy 226 (2018) 1252–1272.

[30] B. Peherstorfer, K. Willcox, Dynamic data-driven reduced-order models, Comput. Methods Appl. Mech. Engrg. 291 (2015) 21–41.

[31] M.D. Chekroun, D. Kondrashov, Data-adaptive harmonic spectra and multilayer Stuart-Landau models, Chaos 27 (9) (2017) 093110.

[32] C. Mou, B. Koc, O. San, L.G. Rebholz, T. Iliescu, Data-driven variational multiscale reduced order models, Comput. Methods Appl. Mech. Engrg. 373 (2021) 113470.

[33] S. Hijazi, G. Stabile, A. Mola, G. Rozza, Data-driven POD-Galerkin reduced order model for turbulent flows, J. Comput. Phys. 416 (2020) 109513.

[34] N. Chen, H. Liu, F. Lu, Shock trace prediction by reduced models for a viscous stochastic Burgers equation, Chaos 32 (4) (2022) 043109.

[35] A.J. Majda, J. Harlim, Physics constrained nonlinear regression models for time series, Nonlinearity 26 (1) (2012) 201.

[36] J. Harlim, A. Mahdi, A.J. Majda, An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models, J. Comput. Phys. 257 (2014) 782–812.

[37] N. Chen, A. Majda, Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification, Entropy 20 (7) (2018) 509.

[38] C. Mou, N. Chen, T. Iliescu, An efficient data-driven multiscale stochastic reduced order modeling framework for complex turbulent systems, 2022, arXiv preprint arXiv:2203.11438.

[39] D. Kondrashov, M.D. Chekroun, M. Ghil, Data-driven non-Markovian closure models, Physica D 297 (2015) 33–55.

[40] K.K. Lin, F. Lu, Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism, J. Comput. Phys. 424 (2021) 109864.

[41] J. Wouters, V. Lucarini, Multi-level dynamical systems: Connecting the Ruelle response theory and the Mori-Zwanzig approach, J. Stat. Phys. 151 (5) (2013) 850–860.

[42] M.D. Chekroun, H. Liu, J.C. McWilliams, Stochastic rectification of fast oscillations on slow manifold closures, Proc. Natl. Acad. Sci. 118 (48) (2021) e2113650118.

[43] A.J. Majda, N. Chen, Model error, information barriers, state estimation and prediction in complex multiscale systems, Entropy 20 (9) (2018) 644.

[44] R. Mojgani, A. Chattopadhyay, P. Hassanzadeh, Closed-form discovery of structural errors in models of chaotic systems by integrating Bayesian sparse regression and data assimilation, 2021, arXiv preprint arXiv:2110.00546.

[45] B. Gershgorin, J. Harlim, A.J. Majda, Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation, J. Comput. Phys. 229 (1) (2010) 32–57.

[46] G.A. Gottwald, J. Harlim, The role of additive and multiplicative noise in filtering complex dynamical systems, Proc. R. Soc. A 469 (2155) (2013) 20130096.

[47] M. Branicki, N. Chen, A.J. Majda, Non-Gaussian test models for prediction and state estimation with model errors, Chinese Ann. Math. Ser. B 34 (1) (2013) 29–64.

[48] X. Xie, C. Webster, T. Iliescu, Closure learning for nonlinear model reduction using deep residual neural network, Fluids 5 (1) (2020) 39.

[49] O. San, R. Maulik, Extreme learning machine for reduced order modeling of turbulent geophysical flows, Phys. Rev. E 97 (4) (2018) 042322.

[50] M. Santos Gutiérrez, V. Lucarini, M.D. Chekroun, M. Ghil, Reduced-order models for coupled dynamical systems: Data-driven methods and the Koopman operator, Chaos 31 (5) (2021) 053116.

[51] S. Pawar, S.E. Ahmed, O. San, A. Rasheed, Data-driven recovery of hidden physics in reduced order modeling of fluid flows, Phys. Fluids 32 (3) (2020) 036602.

[52] A. Chattopadhyay, A. Subel, P. Hassanzadeh, Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning, J. Adv. Modelling Earth Syst. 12 (11) (2020) e2020MS002084.

[53] A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, K. Kashinath, Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence, in: Proceedings of the 10th International Conference on Climate Informatics, 2020, pp. 106–112.

[54] A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, E. Bach, K. Kashinath, Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving spatial transformers in a case study with ERA5, Geosci. Model Dev. Discuss. (2021) 1–23.

[55] N. Chen, D. Qi, A physics-informed data-driven algorithm for ensemble forecast of complex turbulent systems, 2022, arXiv preprint arXiv:2204.08547.

[56] A. Moosavi, R. Stefanescu, A. Sandu, Efficient construction of local parametric reduced order models using machine learning techniques, 2015, arXiv preprint arXiv:1511.02909.

[57] N. Chen, Y. Li, BAMCAFE: A Bayesian machine learning advanced forecast ensemble method for complex turbulent systems with partial observations, Chaos 31 (11) (2021) 113114.

[58] J. Elinger, Information Theoretic Causality Measures For Parameter Estimation and System Identification (Ph.D. thesis), Georgia Institute of Technology, 2020.

[59] J. Elinger, J. Rogers, Causation entropy method for covariate selection in dynamic models, in: 2021 American Control Conference, ACC, IEEE, 2021, pp. 2842–2847.

[60] C.J. Quinn, N. Kiyavash, T.P. Coleman, Directed information graphs, IEEE Trans. Inform. Theory 61 (12) (2015) 6887–6909.
[61] N. Chen, A.J. Majda, Efficient nonlinear optimal smoothing and sampling algorithms for complex turbulent nonlinear dynamical systems with partial observations, J. Comput. Phys. (2020) 109381.
[62] N. Chen, A.J. Majda, Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems, Proc. Natl. Acad. Sci. 114 (49) (2017) 12864–12869.
[63] N. Chen, Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics, J. Comput. Phys. 418 (2020) 109635.
[64] G.K. Vallis, Atmospheric and Oceanic Fluid Dynamics, Cambridge University Press, 2017.
[65] R. Salmon, Lectures on Geophysical Fluid Dynamics, Oxford University Press, 1998.
[66] R.S. Liptser, A.N. Shiryaev, Statistics of Random Processes II: Applications, Vol. 6, Springer Science & Business Media, 2013.
[67] N. Chen, A.J. Majda, Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics, Mon. Weather Rev. 144 (12) (2016) 4885–4917.
[68] B. Gershgorin, J. Harlim, A.J. Majda, Test models for improving filtering with model errors through stochastic parameter estimation, J. Comput. Phys. 229 (1) (2010) 1–31.
[69] I.G. Grooms, A.J. Majda, Stochastic superparameterization in a one-dimensional model for wave turbulence, Commun. Math. Sci. 12 (3) (2014) 509–525.
[70] M. Branicki, A.J. Majda, K.J. Law, Accuracy of some approximate Gaussian filters for the Navier–Stokes equation in the presence of model error, Multiscale Model. Simul. 16 (4) (2018) 1756–1794.
[71] N. Chen, Y. Li, H. Liu, Conditional Gaussian nonlinear system: A fast preconditioner and a cheap surrogate model for complex nonlinear systems, Chaos 32 (5) (2022) 053122.
[72] N. Chen, A.J. Majda, C. Sabeerali, R. Ajayamohan, Predicting monsoon intraseasonal precipitation using a low-order nonlinear stochastic model, J. Clim. 31 (11) (2018) 4403–4427.
[73] A.J. Majda, I. Timofeyev, E. Vanden-Eijnden, Systematic strategies for stochastic mode reduction in climate, J. Atmos. Sci. 60 (14) (2003) 1705–1722.
[74] A.J. Majda, I. Timofeyev, E. Vanden Eijnden, A mathematical framework for stochastic climate models, Comm. Pure Appl. Math. 54 (8) (2001) 891–974.
[75] R. Plant, G.C. Craig, A stochastic parameterization for deep convection based on equilibrium statistics, J. Atmos. Sci. 65 (1) (2008) 87–105.
[76] P.P. Mana, L. Zanna, Toward a stochastic parameterization of ocean mesoscale eddies, Ocean Model. 79 (2014) 1–20.
[77] J. Berner, U. Achatz, L. Batte, L. Bengtsson, A.d.l. Cámara, H.M. Christensen, M. Colangeli, D.R. Coleman, D. Crommelin, S.I. Dolaptchiev, et al., Stochastic parameterization: Toward a new view of weather and climate models, Bull. Am. Meteorol. Soc. 98 (3) (2017) 565–588.
[78] L. Barnett, A.B. Barrett, A.K. Seth, Granger causality and transfer entropy are equivalent for Gaussian variables, Phys. Rev. Lett. 103 (23) (2009) 238701.
[79] R. Vicente, M. Wibral, M. Lindner, G. Pipa, Transfer entropy—A model-free measure of effective connectivity for the neurosciences, J. Comput. Neurosci. 30 (1) (2011) 45–67.
[80] J. Massey, et al., Causality, feedback and directed information, in: Proc. Int. Symp. Inf. Theory Applic., ISITA-90, 1990, pp. 303–305.
[81] G. Kramer, Directed Information for Channels with Feedback, Citeseer, 1998.
[82] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2) (2000) 461.
[83] M. Paluš, V. Komárek, Z. Hrnčíř, K. Štěrbová, Synchronization as adjustment of information rates: Detection from bivariate time series, Phys. Rev. E 63 (4) (2001) 046211.
[84] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[85] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.
[86] M.K. Tippett, R. Kleeman, Y. Tang, Measuring the potential utility of seasonal climate predictions, Geophys. Res. Lett. 31 (22) (2004).
[87] R. Kleeman, Information theory and dynamical system predictability, Entropy 13 (3) (2011) 612–649.
[88] M. Branicki, A.J. Majda, Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency, Nonlinearity 25 (9) (2012) 2543.
[89] C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, in: vol. 13 of Springer Series in Synergetics, Springer-Verlag, Berlin, 2004.
[90] A. Majda, Introduction to PDEs and Waves for the Atmosphere and Ocean, Vol. 9, American Mathematical Soc., 2003.
[91] W.W. Grabowski, An improved framework for superparameterization, J. Atmos. Sci. 61 (15) (2004) 1940–1952.
[92] D.J. Gagne, H.M. Christensen, A.C. Subramanian, A.H. Monahan, Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model, J. Adv. Modelling Earth Syst. 12 (3) (2020) e2019MS001896.
[93] A. Chattopadhyay, P. Hassanzadeh, D. Subramanian, Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: Reservoir computing, artificial neural network, and long short-term memory network, Nonlinear Processes Geophys. 27 (3) (2020) 373–389.
[94] K. Bergemann, S. Reich, A localization technique for ensemble Kalman filters, Q. J. R. Meteorol. Soc. 136 (648) (2010) 701–707.
[95] J.L. Anderson, Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, Physica D 230 (1–2) (2007) 99–111.
[96] T. Janjić, L. Nerger, A. Albertella, J. Schröter, S. Skachko, On domain localization in ensemble-based Kalman filter algorithms, Mon. Weather Rev. 139 (7) (2011) 2046–2060.
[97] A.J. Majda, B. Gershgorin, Quantifying uncertainty in climate change science through empirical information theory, Proc. Natl. Acad. Sci. 107 (34) (2010) 14958–14963.
[98] A. Majda, R.V. Abramov, M.J. Grote, Information Theory and Stochastics for Multiscale Nonlinear Systems, Vol. 25, American Mathematical Soc., 2005.
[99] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[100] S. Kullback, Letter to the Editor: The Kullback-Leibler Distance, American Statistician, 1987.
[101] S. Kullback, Statistics and Information Theory, J Wiley Sons, New York, 1959.
[102] G. Branstator, H. Teng, Two limits of initial-value decadal predictability in a CGCM, J. Clim. 23 (23) (2010) 6292–6311.
[103] T. DelSole, Predictability and information theory. Part I: Measures of predictability, J. Atmos. Sci. 61 (20) (2004) 2425–2440.
[104] T. DelSole, Predictability and information theory. Part II: Imperfect forecasts, J. Atmos. Sci. 62 (9) (2005) 3368–3381.
[105] D. Giannakis, A.J. Majda, Quantifying the predictive skill in long-range forecasting. Part II: Model error in coarse-grained Markov models with application to ocean-circulation regimes, J. Clim. 25 (6) (2012) 1814–1826.
[106] R. Kleeman, Measuring dynamical prediction utility using relative entropy, J. Atmos. Sci. 59 (13) (2002) 2057–2072.
[107] D. Qi, A.J. Majda, Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory, Commun. Math. Sci. 14 (6) (2016) 1687–1722.
[108] A.M. Yaglom, An Introduction to the Theory of Stationary Random Functions, Courier Corporation, 2004.
[109] E.N. Lorenz, Formulation of a low-order model of a moist general circulation, J. Atmos. Sci. 41 (12) (1984) 1933–1945.
[110] E.N. Lorenz, Irregularity: A fundamental property of the atmosphere, Tellus A 36 (2) (1984) 98–110.
[111] A. Beskos, G.O. Roberts, Exact simulation of diffusions, Ann. Appl. Probab. 15 (4) (2005) 2422–2444.
[112] D.S. Wilks, Effects of stochastic parametrizations in the Lorenz'96 system, Q. J. R. Meteorol. Soc. 131 (606) (2005) 389–407.
[113] E.N. Lorenz, Predictability: A problem partly solved, in: Proc. Seminar on Predictability, Vol. 1, 1996, p. 1.
[114] A.J. Majda, M.J. Grote, Mathematical test models for superparametrization in anisotropic turbulence, Proc. Natl. Acad. Sci. 106 (14) (2009) 5470–5474.
[115] I. Grooms, A.J. Majda, Stochastic superparameterization in quasi-geostrophic turbulence, J. Comput. Phys. 271 (2014) 78–98.
[116] A.J. Majda, I. Grooms, New perspectives on superparameterization for geophysical turbulence, J. Comput. Phys. 271 (2014) 60–77.
[117] B. Lindner, J. Garcıa-Ojalvo, A. Neiman, L. Schimansky-Geier, Effects of noise in excitable systems, Phys. Rep. 392 (6) (2004) 321–424.
[118] H. Treutlein, K. Schulten, Noise induced limit cycles of the Bonhoeffer-Van der Pol model of neural pulses, Ber. Bunsenges. Phys. Chem. 89 (6) (1985) 710–718.

[119] B. Lindner, L. Schimansky-Geier, Coherence and stochastic resonance in a two-state system, Phys. Rev. E 61 (6) (2000) 6103.
[120] A. Longtin, Stochastic resonance in neuron models, J. Stat. Phys. 70 (1–2) (1993) 309–327.
[121] K. Wiesenfeld, D. Pierson, E. Pantazelou, C. Dames, F. Moss, Stochastic resonance on a circle, Phys. Rev. Lett. 72 (14) (1994) 2125.
[122] A. Neiman, L. Schimansky-Geier, A. Cornell-Bell, F. Moss, Noise-enhanced phase synchronization in excitable media, Phys. Rev. Lett. 83 (23) (1999) 4896.
[123] H. Hempel, L. Schimansky-Geier, J. Garcia-Ojalvo, Noise-sustained pulsating patterns and global oscillations in subexcitable media, Phys. Rev. Lett. 82 (18) (1999) 3713.
[124] B. Hu, C. Zhou, Phase synchronization in coupled nonidentical excitable systems and array-enhanced coherence resonance, Phys. Rev. E 61 (2) (2000) R1001.
[125] P. Jung, A. Cornell-Bell, K.S. Madden, F. Moss, Noise-induced spiral waves in astrocyte syncytia show evidence of self-organized criticality, J. Neurophysiol. 79 (2) (1998) 1098–1101.
[126] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation, J. Amer. Statist. Assoc. 82 (398) (1987) 528–540.
[127] G. Evensen, P.J. Van Leeuwen, An ensemble Kalman smoother for nonlinear dynamics, Mon. Weather Rev. 128 (6) (2000) 1852–1867.
[128] G.A. Gottwald, S. Reich, Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation, Physica D 423 (2021) 132911.
[129] A. Wikner, J. Pathak, B.R. Hunt, I. Szunyogh, M. Girvan, E. Ott, Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components, Chaos 31 (5) (2021) 053114.