Nonlinear Dyn (2025) 113:11327–11351
https://doi.org/10.1007/s11071-024-10824-3

RESEARCH

Minimum reduced-order models via causal inference

Nan Chen · **Honghu Liu**

Received: 4 July 2024 / Accepted: 20 December 2024 / Published online: 28 December 2024

Abstract Constructing sparse, effective reduced-order models (ROMs) for high-dimensional dynamical data is an active area of research in applied sciences. In this work, we study an efficient approach to identifying such sparse ROMs using an information-theoretic indicator called causation entropy. Given a feature library of possible building block terms for the sought ROMs, the causation entropy ranks the importance of each term to the dynamics conveyed by the training data before a parameter estimation procedure is performed. It thus allows for an efficient construction of a hierarchy of ROMs with varying degrees of sparsity to effectively handle different tasks. This article examines the ability of the causation entropy to identify skillful sparse ROMs when a relatively high-dimensional ROM is required to emulate the dynamics conveyed by the training dataset. We demonstrate that a Gaussian approximation of the causation entropy still performs exceptionally well even in presence of highly non-Gaussian statistics. Such approximations provide an efficient way to access the otherwise hard to compute causation entropies when the selected feature library contains a large number of candidate functions. Besides recovering long-term statistics, we also demonstrate good performance of the obtained ROMs in recovering unobserved dynamics via data assimilation with partial observations, a test that has not been done before for causation-based ROMs of partial differential equations. The paradigmatic Kuramoto–Sivashinsky equation placed in a chaotic regime with highly skewed, multimodal statistics is utilized for these purposes.

Keywords Causation entropy · Data assimilation · Parameter estimation · Kuramoto–Sivashinsky equation · Chaos

N. Chen
Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53705, USA
e-mail: chennan@math.wisc.edu

H. Liu (✉)
Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA
e-mail: hhliu@vt.edu

Contents

1 Introduction

Complex dynamical systems appear in many scientific areas, including climate science, geophysics, engineering, neuroscience, plasma physics, and material science [45,97,105,114,117]. They play a vital role in describing the underlying physics and facilitating the study of many important issues, such as state estimation and forecasting. However, direct numerical simulation is often quite expensive due to the high dimensionality and the multiscale nature of these systems. The situation becomes even more computationally prohibitive when ensemble methods, such as ensemble data assimilation and statistical forecast, are applied to these systems. Therefore, developing appropriate reduced-order models (ROMs) becomes essential not only to reduce the computational cost but also to discover the dominant dynamics of the underlying system.

There exists a vast literature on ROM techniques [2]. On the one hand, when the complex nonlinear governing equations of the full system are given, one systematic approach to developing ROMs is to project the full governing model to a few leading energetic modes through a Galerkin method. At the core of such Galerkin projections are empirical basis functions constructed with various techniques such as the proper orthogonal decomposition (POD) [51], the dynamic mode decomposition (DMD) [88,93], and the principal interaction patterns (PIPs) [48,66]. Once the projection is implemented, supplementing the resulting equations with closure terms is often essential to compensate for the truncation error [19,25,37,80,102,109,121]. On the other hand, many data-driven methods have been developed to learn the dynamics directly from the observed or simulated data associated with the large-scale features or dominant modes of the full systems [2,18,27,34,50,71,78,84,101]. The full system does not necessarily need to be known in such a case. It is worth mentioning that there are also many recent developments in non-parametric ROMs or surrogate models, including those resulting from machine learning methods [21–23,28,30,31,35,77,79,83,90,103,116]. Many of these developments focus primarily on providing efficient forecast results, while less effort is put into developing a systematic way of quantifying the importance of each constitutive term in the utilized ROMs.

When the ROMs are given by parametric forms, one would hope that the model is not only skillful in describing the underlying dynamics but also simple enough to facilitate efficient computations. While dynamical models arising from real-world applications usually have a parsimonious structure [110], data-driven ROMs derived from these full models typically do not inherit such parsimony. On the contrary, as the underlying full system is typically nonlinear, ROMs obtained from projection methods often contain a large number of nonlinear terms. This is because nonlinear interactions among different spatial modes usually cannot be confined to a small subspace spanned by a few spatial modes unless special cancellation properties exist, taking (14) below as an example. Similarly, starting with a comprehensive nonlinear model ansatz, applying a standard regression technique to the observed time series typically leads to a high percentage of the terms with non-zero coefficients. Eliminating the terms with weak contributions to dynamics is crucial in balancing the complexity and accuracy of the resulting ROMs.

For this purpose, it is important to quantify and rank the contribution of each potential constitutive term in a ROM. By the very fabric of chaotic systems, the dynamical contribution of a given term is not simply characterized by the amplitude of its coefficient or even the energy carried by the term. Indeed, removing terms with even a tiny amount of energy can cause a drastic change in the ROM's dynamics, especially when the underlying full model admits sudden transitions such as bursting behaviors [8,9,40]. Due to these reasons, an intuitive approach of ranking the model's coefficients and removing the terms with small coefficients can fail. Instead, constrained linear regression subject to ℓ_1-norm regularization, called the least absolute shrinkage and selection operator (LASSO) regression [91,111], has been widely used to discover sparse models from data [15,18,38,87,92,94]. By adjusting the degree of shrinkage (i.e., the level of the ℓ_1 regularization), the LASSO regression can achieve a varying degree of sparsity since the ℓ_1 constraint penalizes the absolute values of the model coefficients, pushing some of them to become precisely zero. However, while the importance of each candidate constitutive term in a ROM can be ranked by repeatedly applying the LASSO regression with different degrees of shrinkage, the total computation in such an approach is costly when the

total number of candidate terms is large, a situation typically faced when the dimension of the ROM is not small. It is also known that LASSO does not handle severe multicollinearity well [49].

In this paper, we take instead an information-theoretic approach based on the concept of causal inference [5,6,42–44,57,106] to efficiently construct skillful sparse ROMs. The use of information-theoretic metrics for identifying dynamical models already has a long history, and some pioneer works can be traced back to the 1970s due to H. Akaike [3,4]. Over the years, several entropy-based metrics have been studied to quantify information flow or the potential causal influence of one set of data on another, and their scopes of usage span far beyond model identification [58,70,72,75]. For our purpose here, we utilize the concept of *causation entropy* recalled in (2) below, and explore its ability to rank the relative importance of each potential constitutive term in a ROM.

Initially introduced in [106] in the context of network inference, causation entropy extends the concept of *transfer entropy* [95] to allow for the conditioning on all other candidate terms' influence when the influence of a particular term is considered. Previous studies illustrated that causation entropy, as well as other related variants, are appropriate surrogates for quantifying the dynamical relevance of each candidate term [6,36,42,57]. Additionally, as demonstrated in [6], under certain circumstances such as robustness toward noise and outliers, a causation-based criterion can lead to better sparse ROMs than those constructed from purely cost-function-based minimization techniques including least squares, LASSO, and SINDy (sparse identification of nonlinear dynamics).

Essentially, causation entropy quantifies the information brought by a given term to the underlying dynamics beyond the information already contained in all the other terms. A larger causation entropy implies that the associated term has a more significant influence/causal effect on the model concerned. Once a feature library of building block functions is chosen as the candidate constitutive terms, the causation entropy associated with each term in the library is computed based on training data using (2); see Sect. 2.2. Then, different cutoff thresholds can be used to construct ROMs with different complexity. From a ROM identification perspective, causation entropy offers a natural way of ranking the importance of each potential term in the ROM before a parameter estimate step is invoked to learn their corresponding optimal model coefficients (see Step 3 and Step 4 in Sect. 2.1). This separation of the model structure identification (Step 3 in Sect. 2.1) from the parameter estimation (Step 4 in Sect. 2.1) is a key feature of the causal inference framework that distinguishes it from purely cost-function-based minimization techniques. As we demonstrate in Sect. 3, it allows for an efficient construction of a hierarchy of ROMs with varying degrees of sparsity to effectively handle different tasks such as recovering long-term statistics and inferring unobserved dynamics via data assimilation.

Despite the attractive features of causal inference, previous studies on sparse identification using entropy-based metrics focused mainly on either relatively low-dimensional problems or partial differential equation (PDE) models placed in parameter regimes exhibiting chaotic but nearly quasi-periodic dynamics. This article aims to take a modest next step by examining the ability of causation entropy to identify skillful sparse ROMs when a relatively high-dimensional ROM is required to emulate the PDE's dynamics and statistics. By doing so, we also demonstrate that a Gaussian approximation of the causation entropy can still perform exceptionally well even in presence of highly non-Gaussian statistics, which is encouraging since a direct computation of the causation entropy is infeasible when the feature library contains a large number of candidate functions. We also examine the performance of the obtained ROMs for data assimilation, a test that has not been done before for PDEs using causation-based ROMs. We carry out our experiments for the Kuramoto–Sivashinsky equation (KSE) [65,99], which is a paradigmatic chaotic system with rich dynamical features.

As shown in Sect. 3, the corresponding causation-based ROMs can indeed successfully reproduce both key dynamical features and crucial statistics of the studied KSE model. We also illustrate that to maintain good modeling skills, the level of sparsity achieved depends both on the type of orthogonal basis used in constructing the ROMs and the goals of the ROMs. In that respect, ROMs built from two types of spatial bases are investigated, including the analytical Fourier and data-driven bases built from POD. Due to the particular form of nonlinearity in the KSE, its Galerkin projections under the Fourier basis already exhibit a very sparse structure. This is not the case for Galerkin projections constructed on a POD basis. It is shown in Sect. 3.2 that the causation-based ROMs can recover

almost perfectly the sparse structure in the Fourier–Galerkin projections and also reproduce their dynamical and statistical properties, while the ROMs built from the POD basis require far more terms to achieve a similar level of dynamical and statistical performance as reported in Sect. 3.3. We also show within the POD setting that when the goal is switched to performing state estimation using data assimilation, we can use a much sparser ROM even if we only observe the time evolution of a few relatively large-scale POD modes; see Sect. 3.4.

The rest of the article is organized as follows. We first outline in Sect. 2 the general procedure to determine the causation-based ROMs, with in particular the core concept of causation entropy and a computationally efficient approximation of this concept recalled in Sect. 2.2. The usefulness of this ROM framework is then illustrated on the Kuramoto–Sivashinsky in Sect. 3 in the context of deriving sparse data-driven ROMs to capture either the long-term statistics of the solution or to recover unobserved dynamics using data assimilation. Some additional remarks and potential future directions are then provided in Sect. 4.

2 A causation-based ROM framework

2.1 Overview

With the background and motivations of using causal inference to construct sparse ROMs clarified in the Introduction, we now describe the procedure to determine causation-based ROMs from training datasets. We break it into four steps, with details for the causal inference step and the parameter estimation step provided in Sects. 2.2 and 2.3, respectively.

As a starting point, we assume that the training data correspond to the time evolution of a dynamical quantity for which we aim to build a ROM. The training data can come from either observation or the simulation of a finite- or infinite-dimensional full model. For simplicity, we assume that the data are collected at equally spaced time instants, with the time step size denoted by Δt. Once such a dataset is available, we adopt the following steps to construct a causation-based ROM.

Step 1. *Determining the state vector of the ROM via data compression*. Usually, the dimension of the state space for the training data is much higher than the dimension of the sought ROM. In such cases, a data compression procedure is performed to learn a set of empirical basis functions from the training data and then project the training data onto the subspace spanned by the identified basis functions. This data compression procedure is a standard integral part of any projection-based ROMs, and different techniques are available for constructing these basis functions, as reviewed in the Introduction. The output of this process is a set of scalar-valued time series, $\{a_i^j : i = 1, \ldots, n, j = 1, \ldots, N_t\}$, where n denotes the number of empirical basis functions utilized for the projection, and N_t denotes the total number of time instants at which the data are recorded. Viewing a_i^j as the value of a state variable a_i at time $j\Delta t$, the state vector for the sought ROM is then taken to be $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$. While the choice of the dimension, n, of the ROM depends apparently on the dynamical nature of the modeled quantity as well as the goals of the ROM, one can also draw insights from existing works on rigorous ROM error estimates to assist with this task; see Remark 1. We also note that there are situations where the state vector consists simply of the variables for which the data are provided, taking the paleoclimate proxy records from Greenland ice cores [14] and certain physiological time series [86] as particular examples. Then, data compression is not needed, and Step 1 is skipped. Namely, in these situations, the reduction aspect is reflected only in reducing the possible number of terms in the ROM's right-hand side without lowering the dimension of the state space.

Step 2. *Constructing a feature library $\mathbb{F}$ of (scalar) candidate functions, from which the constitutive terms of the sought ROM for $\boldsymbol{a}$ are selected in Step 3*. Generally, one should use prior knowledge about the possible model behind the training data to inform a judicious choice of the candidate functions. But in case no prior knowledge is available, some typical choices are monomials in terms of the components $a_1, \ldots, a_n$ of $\boldsymbol{a}$ up to a certain degree, as well as trigonometric functions or any other common elementary functions in these variables. Time-dependent forcing terms can also be included

in the library in case a non-autonomous vector field is expected. At the same time, the computational cost to perform Step 3 below increases when the number of functions in the library increases. Note also that to avoid degeneracy when computing the covariance matrices in Step 3, $\mathbb{F}$ should not contain constant functions. Instead, a constant forcing term can be added afterwards in Step 4 at the stage of parameter estimation; see Sect. 2.3.

Step 3. *Computing the causation entropy for each function in the feature library $\mathbb{F}$ to determine the ROM's model structure*. This step ranks the importance (in the sense of direct causal relation [106]) of all the functions in $\mathbb{F}$ for the time evolution of each derivative $\mathrm{d}a_i/\mathrm{d}t$; see Sect. 2.2 for details. Subsequently, we can select different cutoff thresholds for the computed causation entropies to construct a hierarchy of ROMs with different sparsity levels. This separation of the model structure identification step from the parameter estimation step is a salient feature of the causal inference framework that distinguishes it from purely cost-function-based minimization techniques.

Step 4. *Estimating both the model parameters for the model structure identified in Step 3 and the associated noise amplitude matrix for the model residual, using, e.g., the maximum likelihood method.* See Sect. 2.3.

We end this subsection with the following remark that provides a couple of more comments about the causal inference framework.

Remark 1 When the training data is noisy, one may need to de-noise the data before performing Step 1 above. Various noise reduction techniques are available for this purpose, including for instance low-pass filtering [55], singular-spectrum analysis [115], and total variation based regularization [89].

In general, it is difficult to determine the appropriate dimension of the sought ROM in advance. But typically, the state vector identified in Step 1 needs to capture a significant amount of energy contained in the training data. In certain situations, one can also benefit from *a priori* ROM error estimates to get an idea about how large the dimension should be; see, e.g., [59,64].

Like other data-driven methods, this causation-based approach also requires data for the time derivative of the state vector $\boldsymbol{a}$. As a result, the time step Δt used in recording the training data should not be too large, unless the time derivative data can be collected through other means rather than by applying a finite difference scheme to the data of $\boldsymbol{a}$ obtained in Step 1.

It is also worth noting that when constructing ROMs for highly chaotic systems, one typically needs to include closure terms to take into account the impact of the orthogonal dynamics not resolved by the ROMs. Additionally, the ROM vector field may also need to respect certain symmetry or energy conservation constraints to ensure stability and accuracy. See Sect. 4 for further discussion along these lines.

2.2 Determining model structure using causation inference

We now provide details about how to carry out Step 3 in the previous subsection, once the state vector $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ and a feature library $\mathbb{F}$ are identified according to Step 1 and Step 2, respectively. For this purpose, we assume that $\mathbb{F}$ contains a total of M candidate functions, $f_1, \ldots, f_M$, for some $M > 0$:

$$\mathbb{F} = \{f_1, \ldots, f_{m-1}, f_m, f_{m+1}, \ldots, f_M\}. \tag{1}$$

2.2.1 The definition of causation entropy

The importance of a candidate function f_m in $\mathbb{F}$ to the dynamics of a_i is measured here by the concept of causation entropy. To introduce this concept, let us denote the time derivative of a_i by $\dot{a}_i$. For any f_m in $\mathbb{F}$, let also $\mathbb{F}\backslash f_m$ be the set of all functions in $\mathbb{F}$ excluding f_m. The causation entropy, $C_{f_m \to \dot{a}_i | [\mathbb{F}\backslash f_m]}$, measures new information provided to $\dot{a}_i$ by f_m in additional to the information already provided to $\dot{a}_i$ by all the other terms in the library $\mathbb{F}\backslash f_m$. Namely, $C_{f_m \to \dot{a}_i | [\mathbb{F}\backslash f_m]}$ quantifies to what extent the candidate function f_m contributes to the right-hand side of the equation for a_i. Its precise definition is given by [106,107]:

$$\begin{aligned} C_{f_m \to \dot{a}_i | [\mathbb{F}\backslash f_m]} &\stackrel{\text{def}}{=} H(\dot{a}_i \,|\, [\mathbb{F}\backslash f_m]) - H(\dot{a}_i \,|\, [\mathbb{F}\backslash f_m]\,, f_m) \\ &= H(\dot{a}_i \,|\, [\mathbb{F}\backslash f_m]) - H(\dot{a}_i \,|\, \mathbb{F}), \end{aligned} \tag{2}$$

where the term $H(\cdot|\cdot)$ is the conditional entropy, which is related to the Shannon entropy $H(\cdot)$ and the joint Shannon entropy $H(\cdot,\cdot)$ as follows. For two multi-dimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ (with the corresponding states being $\boldsymbol{x}$ and $\boldsymbol{y}$), the following identity

(known as the chain rule) holds [39, Theorem 2.2.1]:

$$H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{X}), \tag{3}$$

where the involved quantities are defined by

$$\begin{aligned} H(\boldsymbol{X}) &= -\int_{\boldsymbol{x}} p(\boldsymbol{x})\log(p(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x},\\ H(\boldsymbol{Y}|\boldsymbol{X}) &= -\int_{\boldsymbol{x}}\int_{\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y})\log(p(\boldsymbol{y}|\boldsymbol{x}))\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x},\\ H(\boldsymbol{X},\boldsymbol{Y}) &= -\int_{\boldsymbol{x}}\int_{\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y})\log(p(\boldsymbol{x},\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{4}$$

with $p(\boldsymbol{x})$ being the probability density function (PDF) of $\boldsymbol{x}$, $p(\boldsymbol{y}|\boldsymbol{x})$ the conditional PDF of $\boldsymbol{y}$ given $\boldsymbol{x}$, and $p(\boldsymbol{x},\boldsymbol{y})$ the joint PDF of $\boldsymbol{x}$ and $\boldsymbol{y}$. Regarding the logarithm function log() involved in (4), commonly used bases are 2, 10, and Euler's number e. For our purpose, the choice of the base is not essential as long as the same base is used for the calculation of all causation entropies since one can convert from base a to base b with the inclusion of a common conversion factor $\log_a b$ [39, Lemma 2.1.2]. This factor is precisely the conversion factor between different units used to measure entropies. For instance, the unit associated with base 2 is called *bit* and the one associated with base e called *nat*, and 1 nat $= \log_2(e)$ bits. To fix ideas, we use base 2 in Sect. 3.

On the right-hand side of (2), the difference between the two conditional entropies indicates the information in $\dot{a}_i$ contributed by the specific function f_m given the contributions from all the other functions. Thus, it tells if f_m provides additional information to $\dot{a}_i$ conditioned on the other potential terms in the dynamics.

Note that the causation entropy $C_{f_m\to\dot{a}_i|[\mathbb{F}\setminus f_m]}$ actually coincides with the conditional mutual information of $\dot{a}_i$ and f_m given $\mathbb{F}\setminus f_m$, usually denoted by $I(\dot{a}_i; f_m|\mathbb{F}\setminus f_m)$, which is always non-negative [39, Section 2.5]. Interestingly, even though conditional mutual information [120] is introduced much earlier than causation entropy, its usage for model identification does not seem to be explored until very recently [12,72]. Additionally, by using (3) in (2) (see also the second line in (5)), we get that $C_{f_m\to\dot{a}_i|[\mathbb{F}\setminus f_m]} = C_{\dot{a}_i\to f_m|[\mathbb{F}\setminus f_m]}$. Namely, given $\mathbb{F}\setminus f_m$, the new information provided to $\dot{a}_i$ by f_m is the same as the new information provided to f_m by $\dot{a}_i$.

It is also worthwhile to highlight that the causation entropy in (2) is fundamentally different from directly computing the correlation between $\dot{a}_i$ and f_m, as the causation entropy also considers the influence of the other library functions. If both $\dot{a}_i$ and f_m are caused by another function $f_{m'}$, then $\dot{a}_i$ and f_m can be highly correlated. Yet, in such a case, the causation entropy $C_{f_m\to\dot{a}_i|[\mathbb{F}\setminus f_m]}$ will be close to zero as f_m is not the causation of $\dot{a}_i$.

2.2.2 An efficient approximation of causation entropy.

We need to compute the causation entropy $C_{f_m\to\dot{a}_i|[\mathbb{F}\setminus f_m]}$ for each of the M candidate functions in $\mathbb{F}$ and for each component a_i of the n-dimensional state vector $\boldsymbol{a}$. Thus, there are in total nM causation entropies to be computed, which can be organized into an $n\times M$ matrix, called the causation entropy matrix. Note that the dimension of $\boldsymbol{X}$ in (4) is either $M-1$ (corresponding to $\mathbb{F}\setminus f_m$) or M (corresponding to $\mathbb{F}$) in the context of calculating the causation entropy given in (2). This implies that a direct calculation of causation entropies involves both the estimation of high-dimensional PDFs and high-dimensional numerical integrations when the number of library functions M is large, which is known to be computationally challenging. As an alternative, we approximate the joint and marginal distributions involved in (4) using Gaussians. In such a way, the causation entropy can be approximated as follows [1]:

$$\begin{aligned} C_{\boldsymbol{Z}\to\boldsymbol{X}|\boldsymbol{Y}} &= H(\boldsymbol{X}|\boldsymbol{Y}) - H(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})\\ &= H(\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) + H(\boldsymbol{Y},\boldsymbol{Z})\\ &\approx \frac{1}{2}\Big(\log(\det(\boldsymbol{R}_{\boldsymbol{XY}}))\\ &\quad - \log(\det(\boldsymbol{R}_{\boldsymbol{Y}})) - \log(\det(\boldsymbol{R}_{\boldsymbol{XYZ}}) + \log(\det(\boldsymbol{R}_{\boldsymbol{YZ}}))\Big), \end{aligned} \tag{5}$$

where $\det(\cdot)$ denotes the determinant of a matrix, $\boldsymbol{R}_{\boldsymbol{XYZ}}$ denotes the covariance matrix of the state variables $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$, and the other covariance matrices $\boldsymbol{R}_{\boldsymbol{Y}}$, $\boldsymbol{R}_{\boldsymbol{XY}}$ and $\boldsymbol{R}_{\boldsymbol{YZ}}$ are defined in the same way.

Thus, by assuming that all the involved PDFs follow multivariate Gaussian distributions, the computation of causation entropies boils down to estimating covariance matrices and computing the logarithm of the determinants (log-determinants) of these covariance matrices. This is a much more manageable task when the number of library functions M is too large for other entropy-estimation techniques [41,61,62,95] to operate effectively while ensuring accuracy and data efficiency.

Admittedly, when the concerned data exhibit highly non-Gaussian statistics, the use of Gaussian approximations to compute the associated causation entropy

can lead to errors. At the same time, as shown in Sect. 3, even though the statistics of the full model's dynamics are highly skewed and multimodal (as revealed in the PDF of the kinetic energy shown in Fig. 6), the Gaussian approximation (5) still performs exceptionally well as verified using a setting in which the true sparsity structure is known; see Fig. 3. In that respect, note also that Gaussian approximations have been widely applied to compute various information measurements in the literature and reasonably accurate results have been reported [17,58,75,112]. Finally, we would like to point out that even the computation of the log-determinants can be expensive when M is too large, and some further discussion about this is provided in Sect. 4.

2.2.3 Determining the model structure of the ROM.

With the $n \times M$ causation entropy matrix in hand, the next step is determining the model structure. This can be done by setting up a threshold value for the causation entropies and retaining only those candidate functions with the causation entropies exceeding the threshold. When there is a visible gap in the causation entropies (see, e.g., Fig. 2), it can serve as a strong indication to set the threshold within this gap. Otherwise, the threshold can be chosen to enforce that a given percentage of the terms in the feature library is kept in the ROM, allowing thus for a hierarchy of ROMs with varying degrees of sparsity by changing the cutoff threshold accordingly; see Sect. 3.3. It should also be emphasized that determining the importance of the terms using causation entropy fundamentally differs from that by first ranking the absolute values of the ROM's model coefficients learned from regression and then removing the terms with small coefficients. The latter does not explicitly quantify statistical significance of the eliminated terms and can lead to ROMs with much less accurate results as shown in Sect. 3.4.

2.3 Parameter estimation

The final step is to estimate the parameters in the resulting model. For this purpose, we denote the total number of terms in the identified model structure from Step 3 by s, and we use a column vector $\boldsymbol{\Theta} \in \mathbb{R}^s$ to denote the corresponding s model coefficients to be estimated. We introduce next an $n \times s$ matrix function of the state vector $\boldsymbol{a}$, denoted by $\boldsymbol{M}$, whose entries are built from the s terms identified in Step 3 as follows. We first put all the s identified terms into each row of $\boldsymbol{M}$, arranged in the same order, then for the ith row ($i = 1, \ldots, n$), we replace those terms that do not appear in the equation of a_i by 0. As a result, the model of $\boldsymbol{a}$ can be written in the following vector form:

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \boldsymbol{\Phi}(\boldsymbol{a}) + \boldsymbol{\sigma}\dot{\boldsymbol{W}}(t), \quad \text{with} \quad \boldsymbol{\Phi}(\boldsymbol{a}) = \boldsymbol{M}(\boldsymbol{a})\boldsymbol{\Theta}. \tag{6}$$

In the above model, $\boldsymbol{\sigma}\dot{\boldsymbol{W}}(t)$ is a stochastic term, with $\dot{\boldsymbol{W}}(t) \in \mathbb{R}^{d\times 1}$ being a white noise for some $d > 0$, and $\boldsymbol{\sigma} \in \mathbb{R}^{n\times d}$ being the noise amplitude matrix. This term $\boldsymbol{\sigma}\dot{\boldsymbol{W}}(t)$ aims to model the residual $\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} - \boldsymbol{\Phi}(\boldsymbol{a})$ since usually there does not exist a $\boldsymbol{\Theta}$ for which $\boldsymbol{\Phi}(\boldsymbol{a})$ fits perfectly the training data $\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t}$ except in some overfitting scenarios. Typically, the dimension of the noise, d, is the same as the dimension of the state variable n. However, in rare situations when the residual $\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} - \boldsymbol{\Phi}(\boldsymbol{a})$ computed from the training data associated with an estimated $\boldsymbol{\Phi}(\boldsymbol{a})$ has a degenerate covariance matrix, the dimension of $\boldsymbol{W}$ would be lower than n.

The parameter vector $\boldsymbol{\Theta}$ and the noise coefficient matrix $\boldsymbol{\sigma}$ in (6) can be determined using, e.g., the maximum likelihood estimation (MLE) [20]; see [29] for the technical details. Notably, the entire parameter estimation can be solved via closed analytic formulae, making the procedure highly efficient. Note also that the optimal parameter $\boldsymbol{\Theta}$ estimated from the MLE is the same as that obtained from the standard linear regression in the special case when the noise amplitude matrix $\boldsymbol{\sigma}$ is an $n \times n$ diagonal matrix [96, Chapter 3]. MLE can handle constrained minimizations too. Algebraic constraints among certain model parameters can be important depending on specific applications. For instance, the quadratic nonlinearity in many fluid systems represents advection and is a natural energy-conserved quantity. To explicitly enforce energy conservation for the corresponding ROM's quadratic nonlinearity can help prevent the finite time blowup of the solution and also enhance accuracy [47,76]. Remarkably, closed analytic formulae are still available for parameter estimation in the presence of such constraints; see, e.g., [36, Section 2.5].

As mentioned in Step 2 of Sect. 2.1, constant functions are excluded in the feature library $\mathbb{F}$ to prevent degeneracy. To add such a constant forcing vector $\boldsymbol{b} \in \mathbb{R}^n$ to (6), we just need to add n additional entries

to the parameter vector $\boldsymbol{\Theta}$, say after the last entry of the original $\boldsymbol{\Theta}$, and also augment $\boldsymbol{M}$ with an $n \times n$ identity matrix appended to the right of the last column in the original $\boldsymbol{M}$. Usually, if the mean of the residual $\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} - \boldsymbol{\Phi}(\boldsymbol{a})$ computed from the training data is not close to zero, it can be beneficial to add such a constant forcing vector to the ROM.

Note also that if one has precise prior knowledge about a portion of the full model's vector field, it can be advantageous to include it in the ROM. Denoting the projection of this known portion of the vector field onto the ROM subspace by $\boldsymbol{Q}$, the drift term $\boldsymbol{\Phi}(\boldsymbol{a})$ in (6) takes then the following form

$$\boldsymbol{\Phi}(\boldsymbol{a}) = \boldsymbol{M}(\boldsymbol{a})\boldsymbol{\Theta} + \boldsymbol{Q}(\boldsymbol{a}), \tag{7}$$

where $\boldsymbol{Q} \in \mathbb{R}^n$ is a column vector that can depend on $\boldsymbol{a}$ but does not involve any free parameters to be estimated. Of course, in this case, we should compute instead the following causation entropies $C_{f_m \to \dot{a}_i | ([\mathbb{F} \setminus f_m], \boldsymbol{Q})}$ for all i and m when determining the model structure in Step 3.

With the causation-based ROM framework outlined above, we now turn to a concrete application to illustrate its efficiency in identifying effective parsimonious ROMs for the Kuramoto–Sivashinsky equation.

3 Reduced-order models for the Kuramoto–Sivashinsky equation based on causal inference

3.1 Preliminaries and background

The Kuramoto–Sivashinsky equation (KSE) [65,99] is a fourth-order dissipative partial differential equation (PDE) that can exhibit intricate spatiotemporal chaotic patterns. It is a prototypical model for long-wave instabilities, which has been derived in various contexts of extended non-equilibrium systems that include unstable drift waves in plasmas [68], laminar flame fronts [99], pattern formation in reaction-diffusion systems [65], and long wave fluctuations in thin film [11,100]. Due to its rich dynamical features, the KSE has served as a test ground for various model reduction methods as well as data assimilation techniques in recent years; see, e.g., [6,25,54,69,73,74,81,104].

We consider the one-dimensional KSE:

$$\partial_t u = -\nu u_{xxxx} - D u_{xx} - \gamma u u_x, \tag{8}$$

which is posed on a bounded interval, $\mathcal{D} = (0, L)$, and subject to periodic boundary conditions. In (8), ν, D and γ are positive parameters.

Under the given boundary conditions, since the spatial average is a conserved quantity for the solution $u(x, t)$ of Eq. (8), for simplicity, we restrict to initial data with mean zero by imposing

$$\int_0^L u(x, t)\, \mathrm{d}x = 0, \quad \forall\, t \geq 0. \tag{9}$$

To build up understanding, throughout the numerical experiments reported below, we will also utilize Galerkin truncations of Eq. (8) constructed using either the eigenbasis of the linear operator in (8) or an empirically constructed basis built from the POD.

3.1.1 *Galerkin projections of the KSE under the Fourier basis*

Due to the assumed periodic boundary conditions, the eigenfunctions of the linear operator $\mathcal{A} = -\nu\partial_{xxxx} - D\partial_{xx}$ in Eq. (8) consist of sine and cosine functions. Thus, the eigenbasis coincides with the Fourier basis. The corresponding Galerkin approximations of Eq. (8) can be determined analytically as given below.

First note that the eigenvalues of the linear operator $\mathcal{A}$ subject to the periodic boundary conditions and the additional mean-zero condition (9) are given by

$$\beta_n = -\frac{16\nu\pi^4 n^4}{L^4} + \frac{4D\pi^2 n^2}{L^2}, \quad n \in \mathbb{N}, \tag{10}$$

where $\mathbb{N}$ denotes the set of all positive integers. Each eigenvalue is associated with two eigenfunctions (labeled by a superscript ℓ):

$$e_n^\ell(x) = \begin{cases} \sqrt{\frac{2}{L}} \cos\left(\frac{2\pi n x}{L}\right), & \text{if } \ell = 0, \\ \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi n x}{L}\right), & \text{if } \ell = 1. \end{cases} \tag{11}$$

These eigenfunctions are normalized so that their $L^2(\mathcal{D})$-norm equal to 1.

Since the eigenfunctions occur in a sine and cosine pair for each wave frequency, we consider Galerkin approximations of Eq. (8) of even dimensions. Denote the $2N$-dimensional Galerkin approximation of the KSE solution u under the Fourier basis by

$$u_G(x, t) = \sum_{n=1}^{N} \sum_{\ell=0}^{1} a_n^\ell(t) e_n^\ell(x). \tag{12}$$

Then the amplitudes, a_n^ℓ's, satisfy the following $2N$-dimensional ODE system

$$\frac{\mathrm{d}a_n^\ell}{\mathrm{d}t} = \beta_n a_n^\ell + \sum_{p,q=1}^{N} \sum_{\ell_1,\ell_2=0}^{1} \left\langle B(e_p^{\ell_1}, e_q^{\ell_2}), e_n^\ell \right\rangle a_p^{\ell_1} a_q^{\ell_2},$$
$$1 \le n \le N, \quad \ell \in \{0, 1\}, \tag{13}$$

where $B(u, v) = -\gamma u v_x$ denotes the quadratic nonlinear term in Eq. (8), and $\langle \cdot, \cdot \rangle$ denotes the L^2-inner product for the underlying Hilbert state space. By direct calculation, we have

$$\langle B(e_p^0, e_q^0), e_n^0 \rangle = \langle B(e_p^0, e_q^1), e_n^1 \rangle = \langle B(e_p^1, e_q^0), e_n^1 \rangle$$
$$= \langle B(e_p^1, e_q^1), e_n^0 \rangle = 0, \quad \forall\, p, q, n, \tag{14}$$

$$\langle B(e_p^0, e_q^1), e_n^0 \rangle = \langle B(e_q^1, e_p^0), e_n^0 \rangle$$
$$= \begin{cases} -\frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = p + q, \\ \frac{\gamma\pi(p-q)}{\sqrt{2}L^{3/2}}, & \text{if } n = |p - q|, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

and

$$\langle B(e_p^\ell, e_q^\ell), e_n^1 \rangle$$
$$= \begin{cases} (-1)^\ell \frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = p + q, \ \ell \in \{0, 1\}, \\ \frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = |p - q|, \ \ell \in \{0, 1\}, \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

Formulas (14)–(16) reveal that most of the nonlinear interaction coefficients $\langle B(e_p^{\ell_1}, e_q^{\ell_2}), e_n^\ell \rangle$ in (13) are zero. The resulting Galerkin system (13) has thus a sparse structure. We will show below in Sect. 3.2 that the causal inference criterion presented in Sect. 2 can be used in a data-driven modeling framework to recover this sparse structure with high fidelity.

3.1.2 Galerkin projections of the KSE under the POD basis

In many applications, empirically computed orthogonal bases can be a more favorable choice than analytic bases due e.g. to their data-adaptive features. We will thus also assess the skill of the causation inference approach when an empirical basis is used instead. Among the most common choices are the POD [46,52,98] and its variants [108]. Of demonstrated relevance for the reduction of nonlinear PDEs are also the PIPs [40,48,66,67] that find a compromise between minimizing tendency error with maximizing explained variance in the resolved modes. In the last decade, related promising techniques such as the DMD [88,93,118] have also emerged; see [113] for a discussion on the relationships between PIPs, DMD, and the linear inverse modeling [85].

Table 1 System parameters for the KSE (8)

ν	D	L	γ	Δt	N_x
8	1	20π	1	0.01	128

To fix ideas, we use the POD modes to construct the data-driven Galerkin approximations. Given a cut-off dimension N, we denote the POD basis by $\{\varphi_j : j = 1, \ldots, N\}$, where the basis functions are ranked by their energy content. Recall that the basis functions are orthonormal, i.e., $\langle \varphi_j, \varphi_k \rangle = \delta_{jk}$ for all j and k. The corresponding N-dimensional POD-Galerkin system of Eq. (8) reads

$$\frac{\mathrm{d}a_n^{\mathrm{POD}}}{\mathrm{d}t} = \sum_{j=1}^{N} A_{nj} a_j^{\mathrm{POD}} + \sum_{i,j=1}^{N} B_{ij}^n a_i^{\mathrm{POD}} a_j^{\mathrm{POD}},$$
$$1 \le n \le N, \tag{17}$$

where

$$A_{nj} = \langle \mathcal{A}\varphi_j, \varphi_n \rangle, \qquad B_{ij}^n = \langle B(\varphi_i, \varphi_j), \varphi_n \rangle, \tag{18}$$

with $\mathcal{A} = -\nu\partial_{xxxx} - D\partial_{xx}$ and $B(u, v) = -\gamma u v_x$ as before. Once (17) is solved, the corresponding spatiotemporal field that approximates the KSE solution u can be reconstructed via

$$u_G^{\mathrm{POD}}(x, t) = \sum_{n=1}^{N} a_n^{\mathrm{POD}}(t)\varphi_n(x). \tag{19}$$

3.1.3 Parameter regime and numerical setup

Throughout Sect. 3, we consider the KSE (8) in the parameter regime given by Table 1. For the chosen regime, there are six unstable eigen directions associated with the linear part of the KSE, and the KSE solution is chaotic. As revealed by the "bifurcation tree" shown in panel A of Fig. 1, the selected regime (with $\nu = 8$) is very close to the borderline in the parameter space where a transition from chaos back to steady state solutions occur as the diffusion coefficient ν of the stabilizing biharmonic term $-u_{xxxx}$ is further decreased. This interlace between chaotic and non-chaotic dynamics by varying certain model parameter is well documented in the literature, and is not limited to KSE

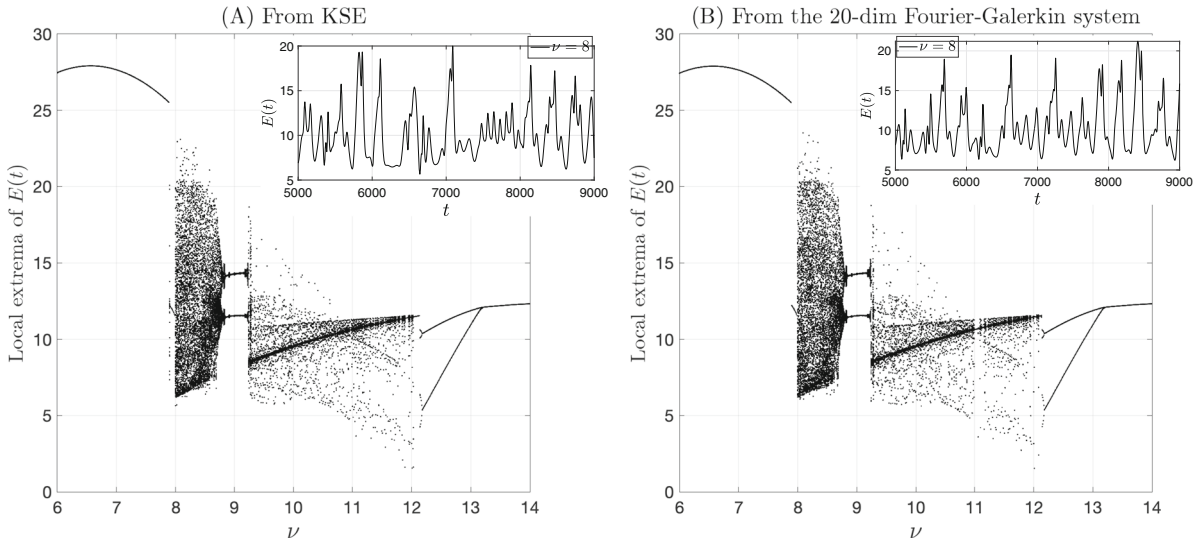

Fig. 1 Local extrema of the kinetic energy E for the KSE (8) (panel A) and its 20-dimensional Fourier–Galerkin approximation (panel B) as the diffusion parameter ν in (8) is varied, while the other parameters in (8) are fixed to be those given in Table 1. The kinetic energy E for the KSE's solution $u(x,t)$ is defined to be $E(t) = \int_0^L u^2(x,t)\,\mathrm{d}x$, and for the Galerkin system we have $E(t) = \sum_{n=1}^{10}\sum_{\ell=0}^{1}(a_n^\ell(t))^2$, where a_n^ℓ's are the state variables of the 20-dimensional Galerkin system of the form (13). Note that the linear operator $\mathcal{A} = -\nu\partial_{xxxx} - D\partial_{xx}$ of the KSE has more unstable eigen directions as ν is decreased; see (10). Thus, in principle a regime with a smaller ν value would require a higher-dimensional Galerkin system to reproduce the KSE's dynamics. We can see that the 20-dimensional Galerkin system recovers vey well the KSE's dynamics for the range of ν values shown here

[24,53]. We argue that setting ν close to this transition borderline creates a challenging scenario to test the performance of the causation-based ROMs, since the discrepancies between the identified ROMs and the full model can easily push the ROM's dynamics into the steady state regime.

On the implementation side, we solve the KSE by a pseudo-spectral code [56], in which the resulting stiff ODE system is integrated using the exponential time-differencing fourth-order Runge–Kutta (ETDRK4) method. We use $N_x = 128$ pairs of Fourier modes (hence 128 equally spaced grid points for the spatial domain) to perform the pseudo-spectral discretization, and we use a time step $\Delta t = 0.01$ to integrate the resulting ODE system with the ETDRK4; see Table 1.

It has been checked that a Fourier–Galerkin system with dimension 20 is sufficient to reproduce the dynamical features in the solution of the KSE for the chosen regime; see Fig. 1, in which we show that the 20-dimensional Fourier–Galerkin system reproduces accurately the bifurcation tree of the KSE as the diffusion parameter ν is varied in an interval where the KSE dynamics make transitions among steady states, periodic dynamics, and chaos. While generating a similar bifurcation tree for the 20-dimensional POD-Galerkin system is challenging, since this type of data-driven systems trained for a fixed parameter regime typically cannot be used to infer the dynamics in another parameter regime. Instead, we have checked that the leading Lyapunov exponents for the 20-dimensional Fourier–Galerkin system and the 20-dimensional POD-Galerkin system for the regime given in Table 1 are close to each other, with the leading three exponents for each of these systems given by 0.0087, 0.0066, 0.0049 and 0.0088, 0.0067, 0.0048, respectively. Instead of estimating from generated solution time series, these Lyapunov exponents are computed by evolving the corresponding nonlinear ODE system and their associated variational equations over a long time window [0, 6E5], following [119, Section 3].[1]

The above dynamical insights suggest that the 20-dimensional Galerkin systems (either constructed from the Fourier basis or the POD basis) approximate reason-

[1] We did not compute the Lyapunov exponents for the full KSE system due to high computational cost, since the method of [119, Section 3] would require to solve an ODE system of dimension $(128 + 128^2)$ over long time interval.

ably well the full KSE model's dynamics for the chosen parameter regime. We will then test 20-dimensional causation-based ROMs with different sparsity percentages in the contexts of both Fourier basis and POD basis. We intentionally choose a parameter regime in which the dimension of a high-fidelity Galerkin approximation is not too large in order to not inflate too much the number of candidate functions in the learning library used for computing the causation entropy; see again Sect. 2.2. See also Sect. 4 for some discussions about applying the framework to obtain larger causation-based ROMs with dimensions in the hundreds when needed.

All the Galerkin approximations of the KSE (8), either constructed from the Fourier basis or the POD basis, are solved using the fourth-order Runge–Kutta method. The causation-based ROMs as well as the thresholded POD-Galerkin systems to be introduced later are simulated with their drift parts approximated using the fourth-order Runge–Kutta method and the stochastic terms approximated using the Euler–Maruyama scheme. The system (27) involved in the data assimilation experiment below is simulated using the Euler–Maruyama scheme for simplicity, which turns out to be sufficient since the involvement of observational data helps alleviate the stiffness of its drift part. The time step size Δt for all these models is the same as the one used for solving the PDE itself.

The initial data for the KSE is taken to be $u_0 = \cos(2\pi x/L)$, and the computed solution over the time window $[10^4, 5 \times 10^4]$ is used for learning the POD basis as well as for training the related ROMs used in Sects. 3.3 and 3.4. To compute the coefficients involved in the POD-Galerkin system (17), we approximate each POD basis function using 64 pairs of Fourier modes and then perform the differentiation and integration involved in (18) analytically. For reasons explained in Sect. 3.2, the causation-based ROMs used in this subsection are trained using the solution for the Fourier–Galerkin systems, still over the time window $[10^4, 5 \times 10^4]$.

3.2 Data-driven inverse models under the Fourier basis

As pointed out in Sect. 3.1.1, the Galerkin approximations of the KSE (8) under the Fourier basis have a sparse structure. Such systems thus provide a good first proof of concept testbed to check whether the causal inference criterion presented in Sect. 2 can differentiate monomials appearing in the Fourier–Galerkin systems from those that do not when all possible linear and quadratic terms are included in the function library used for model identification.

For this purpose, we place the KSE in the parameter regime given in Sect. 3.1.3 and use the 20-dimensional Fourier–Galerkin system of the form (13) as the true model to generate the training data. We then follow the four-step procedure given in Sect. 2.1 to construct the sought ROM, as detailed below. Since we aim to check whether the constructed ROM can recover the sparse structure in the chosen Galerkin system, the state vector of the ROM is the same as that of the Galerkin system. Thus, Step 1 of Sect. 2.1 for state vector identification is not needed here. For Step 2, we include all possible linear and quadratic monomials in the feature library $\mathbb{F}$. There are thus a total of 230 candidate functions, consisting of 20 linear terms and 210 quadratic terms. To facilitate later discussions, we adopt the following ordering to arrange the 230 library functions. We arrange the 20 unknowns a_n^ℓ ($\ell = 0, 1$, $n = 1, \ldots, 10$) of the Galerkin system into a vector $\boldsymbol{a} = (a_1^0, \ldots, a_{10}^0, a_1^1, \ldots, a_{10}^1)^{\mathrm{T}}$, and denote the ith entry of $\boldsymbol{a}$ by a_i. The functions in the feature library $\mathbb{F}$ are arranged in the order of

$$\{a_1, \ldots, a_{20}\} \quad \text{followed by} \quad \{a_j a_k \mid 1 \le j \le k \le 20\}, \tag{20}$$

where the following lexicographic order for (j, k) is adopted for the quadratic terms: $(j_1, k_1) < (j_2, k_2)$ if $(j_1 < j_2)$ or $j_1 = j_2$ and $k_1 < k_2$.

Regarding Step 3 of Sect. 2.1 for identifying the ROM's model structure, following the description in Sect. 2.2, we compute the causation entropy $C_{f_m \to \dot{a}_i | [\mathbb{F} \setminus f_m]}$ for each library function f_m in $\mathbb{F}$ and each component a_i of $\boldsymbol{a}$, according to the approximation formula (5). In total, there are $230 \times 20 = 4600$ causation entropy values to compute. These values are shown in Fig. 2, which are grouped by equation, with the first 230 values (from left) for the first equation and the next 230 values for the second equation, etc. There is a visible gap in Fig. 2 that separates large causation entropy values (such as those above the red dashed horizontal line) from the smaller ones (below the blue dashed horizontal line), with only very few exceptions falling in between. One is then tempted to suspect that the cutoff threshold for the causation entropy value used

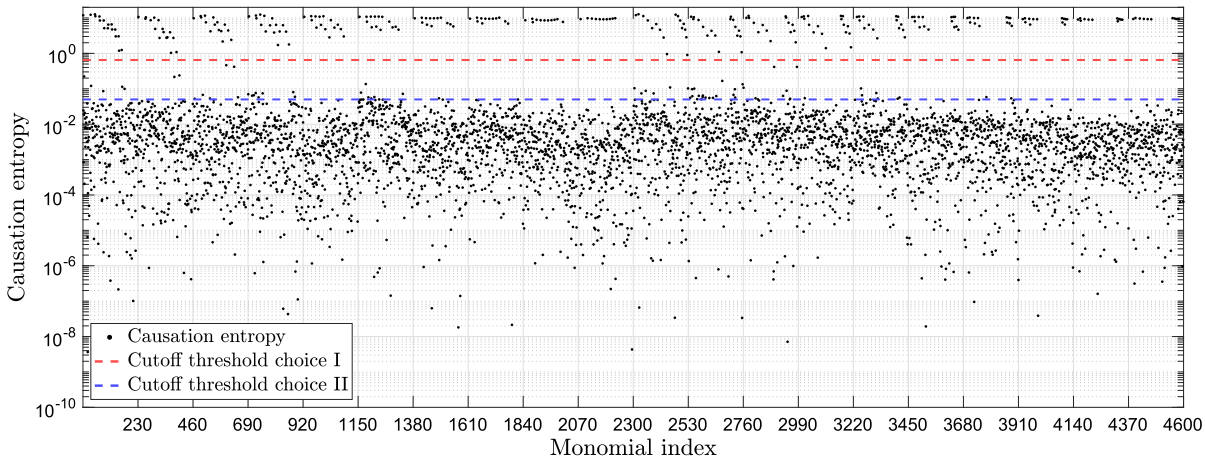

Fig. 2 Causation entropy that ranks the importance of the library functions for learning a data-driven quadratic inverse model of the 20-dimensional Fourier Galerkin system of the form (13). The candidate function library $\mathbb{F}$ includes all of the 230 linear and quadratic monomials constructed from the 20 components of the unknown $\boldsymbol{a}$ as listed in (20). The causation entropy from each library function f_m to the ith equation, $C_{f_m \to \dot{a}_i | [\mathbb{F} \setminus f_m]}$, is computed according to the approximation formula (5) given in Sect. 2.2. The parameter regime is the one given in Sect. 3.1.3. The causation entropy values are grouped by equation, with the first 230 values (from left) for the first equation, and the next 230 values for the second equation, etc. Also shown are two cutoff thresholds, 0.65 (red line) and 0.05 (blue line). It has been checked that the causation entropy values for all the terms appearing in the true Galerkin system are above the blue line, confirming thus the relevance of this casual inference criterion in identifying constituent terms in the data-driven model

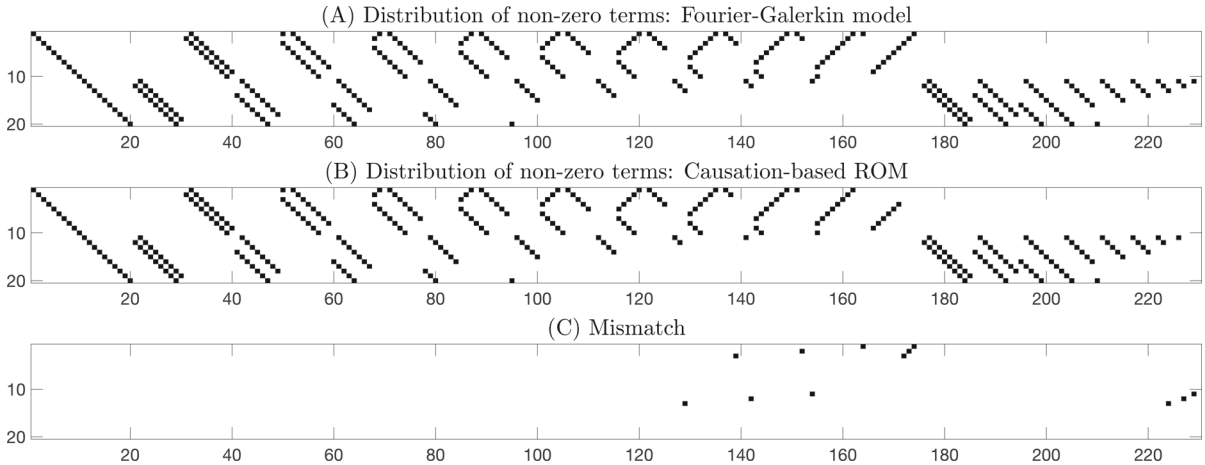

Fig. 3 Visualization of the distribution of the constituent terms in the 230 function library for the true Galerkin model (panel A) and the learned model (panel B). The learned model is constructed with the cutoff threshold for the causation entropy taken to be 0.65 (red line in Fig. 2). In each panel, the vertical axis consists of 20 rows with each row corresponding to one equation, and the horizontal axis consists of 230 columns with each column corresponding to one function in the learning library. They form thus a 20×230 mesh. A black square in the (i, j)-th grid indicates the jth monomial in the library is present in the ith equation. The functions in the library are ordered in the way given by (20). In particular, the linear terms are placed before the quadratic terms. For instance, the left-most diagonal block in panel A shows that the linear part of the true Galerkin model is diagonal. The mismatches between the learned model and the true model are shown in panel C. There are 12 mismatched terms, all of which are quadratic terms. It has been checked that all these 12 terms are present in the true model, but are missing in the learned model

for the model structure identification should fall within this gap.

In the remainder of this section, we use the more severe cutoff threshold marked by the red dashed line in Fig. 2 as the cutoff threshold, which corresponds to a numerical value of 0.65 in contrast to 0.05 for the blue dashed line. There are 283 terms whose causation entropy values are above this threshold of 0.65. It turns out all these 283 identified terms are present in the true Galerkin model, which itself has 295 terms. Figure 3 (panel B) shows the distribution of the identified 283 terms, while that for the true 20-dimensional Galerkin model (13) is shown in panel A of this figure. The 12 terms in the true model not identified with this cutoff threshold are shown in panel C of Fig. 3.

As can be seen in Fig. 3, the causation entropy criterion is remarkably successful in identifying the true sparse model. In particular, it has a negligible mismatch rate of $12/(20 \times 230) = 0.26\%$ for the total 4600 possible terms to be sifted through. It turns out that all these 12 mismatch terms are quadratic terms, while all the linear terms of the model are correctly identified. Indeed, the linear part of the true model is a diagonal matrix with eigenvalues on the diagonal (see Eq. 13). This is represented by the left-most diagonal block in panel A of Fig. 3 due to the way that the library functions are arranged; see again (20). These linear terms are fully captured using the chosen cutoff threshold as shown in panel B of this figure.

With the constituent terms of the ROM identified now, we follow Step 4 of Sect. 2.1 and use a standard MLE to determine the model coefficients; see Sect. 2.3. For this purpose, we use the same training solution data used in the previous causal inference step. Among the 283 coefficients in $\boldsymbol{\Theta}$, 20 coefficients on the diagonal of the matrix are to be learned for the linear part, and the remaining 263 coefficients are for the nonlinear terms. The numerical values of these coefficients are graphically shown in Fig. 4. For the linear part (top panels of Fig. 4), the learned model coefficients recover these for the true model with high precision. We have checked that the relative error is below 0.06% for all the 20 coefficients. For the nonlinear terms (bottom panels of Fig. 4), the largest differences between the learned model coefficients and the true ones occur at the 12 mismatched terms, as expected. Outside of these mismatched terms, the error is one-order smaller (on the scale of 10^{-3}) compared with those shown in panel F of Fig. 4. Note also that the causation entropy for the mismatched terms all fall below the red dashed line in Fig. 2 and thus being filtered out in the learned model for this chosen cutoff threshold. We then expect that they play a less important role than the other 283 terms in "orchestrating" the dynamics in the true model. We also note that the learned noise amplitude matrix $\boldsymbol{\sigma}$ associated with the Gaussian noise term (see $\boldsymbol{\sigma}\dot{\boldsymbol{W}}(t)$ in (6)) comes with very small entries on the scale of 10^{-7}. Thus, the noise term is essentially negligible in the resulting causation-based ROMs. It turns out that the learned model with this cutoff threshold can already capture faithfully the true dynamics as shown in Figs. 5 and 6.

In particular, we show in Fig. 5 the reconstructed spatiotemporal field for the true Galerkin model defined by (12) (left panel) and its analog from the causation-based ROM (right panel). The solutions are shown in a time window that is far beyond the training window $[10^4, 5 \times 10^4]$. The chaotic dynamics from the learned model are essentially indistinguishable from those in the true model, with local maxima (reddish patches) and local minima (bluish patches) progressing in a zigzag way as time evolves, forming a rich variety of local patterns. In that respect, we also point out that the long thin reddish strip observed in the left panel formed in the time window $[2.96 \times 10^5, 2.97 \times 10^5]$, which propagates from the left side of the domain ($x = 0$) all the way up to almost the right side of the domain, has also been observed in other time windows for the learned model. This good reproduction of the dynamics is further confirmed at the statistical level as shown in Fig. 6 for the energy spectrum E_k (top panels) as well as the PDF and the autocorrelation function (ACF) of the kinetic energy E (bottom panels); see the caption of this figure for further details.

Going back to Fig. 2, when the more conservative cutoff threshold 0.05 (corresponding to the blue dashed line) is used, the corresponding causation-based ROM contains a total of 354 terms, which includes all of the 295 terms appearing in the true Galerkin model. The performance of this new causation-based ROM is similar to those shown in Figs. 5B and 6. This indicates that the terms whose causation entropy values fall in between the gap marked by the red and blue dashed lines in Fig. 2 already play very little role in determining the dynamics of the learned model.

We also note that the existence of a clear gap to separate the larger and smaller causation entropy values, such as shown in Fig. 2 seems to be tied to the

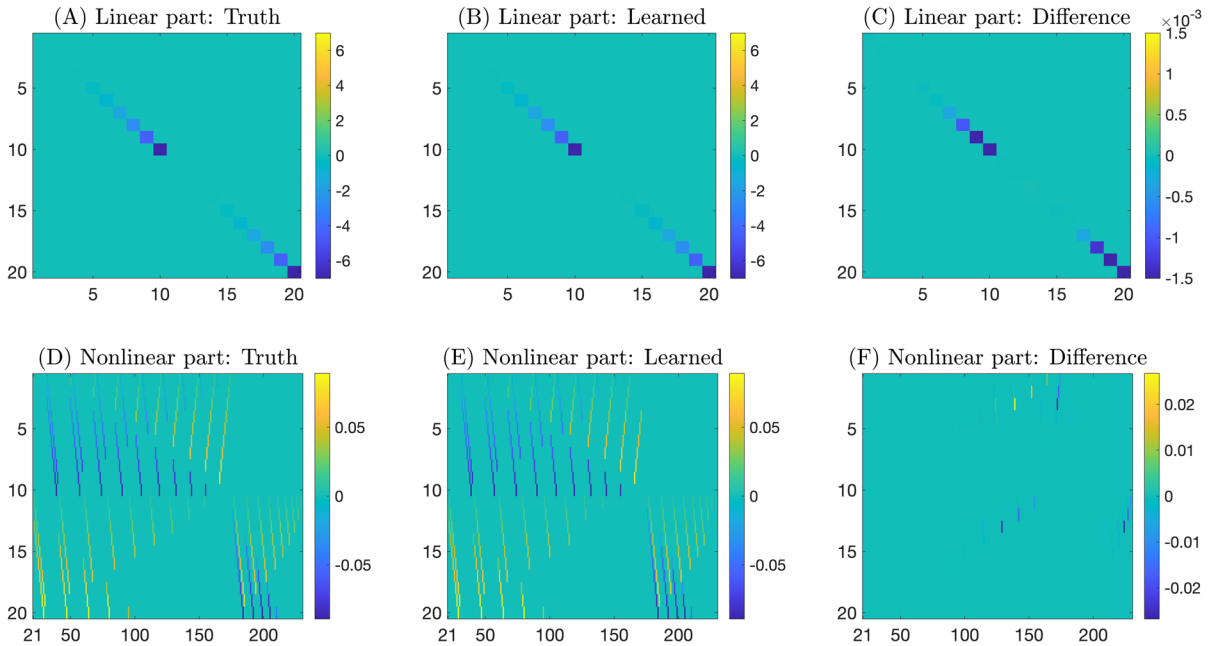

Fig. 4 Visualization of the model coefficients for the 20-dimensional Fourier–Galerkin system (true model) and the causation-based ROM (learned model). For the true model, all non-zero coefficients occur for the terms marked by black squares shown in panel A of Fig. 3. We separated the linear terms (panel A here) from the nonlinear terms (panel D here) for a better visualization, since the coefficients for some of the linear terms are two-order larger than those for the nonlinear terms. The sparsity structure for the learned model is the one shown in panel B of Fig. 3

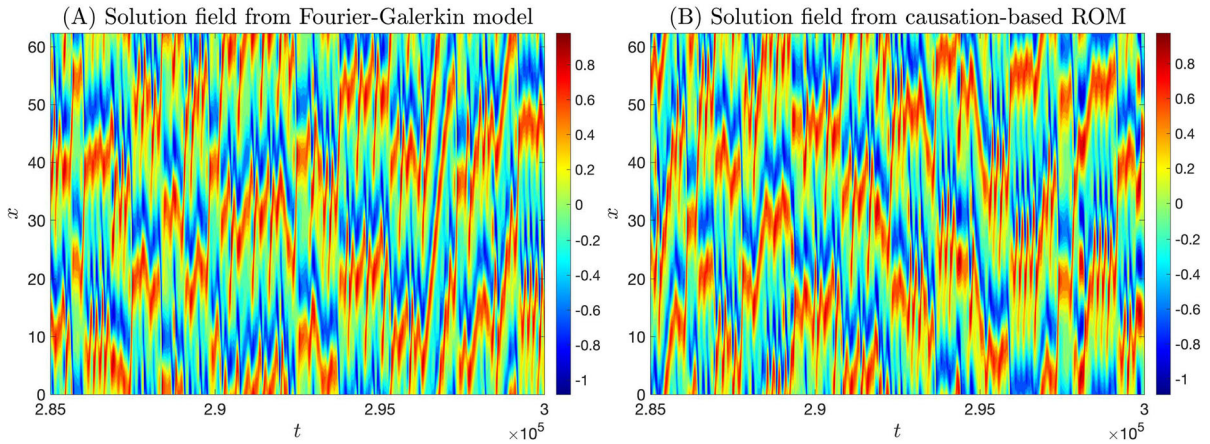

Fig. 5 Comparison of the reconstructed solution fields obtained using (12) for the 20-dimensional Fourier–Galerkin system (13) (left panel) and for the causation-based ROM of the same dimension (right panel). The results are shown here in a time window well beyond the training window $[10^4, 5 \times 10^4]$

fact that the Fourier–Galerkin systems (13) themselves admit a sparse structure; see again (14)–(16). When other (global) basis functions are used, the corresponding Galerkin system may no longer be sparse. As such, one should no longer expect a clear gap to present in the causation entropy plot. However, as shown below using the POD basis, the ranking of the library terms provided by the causation entropy still offers a compelling way to obtain skillful yet significantly sparsified models.

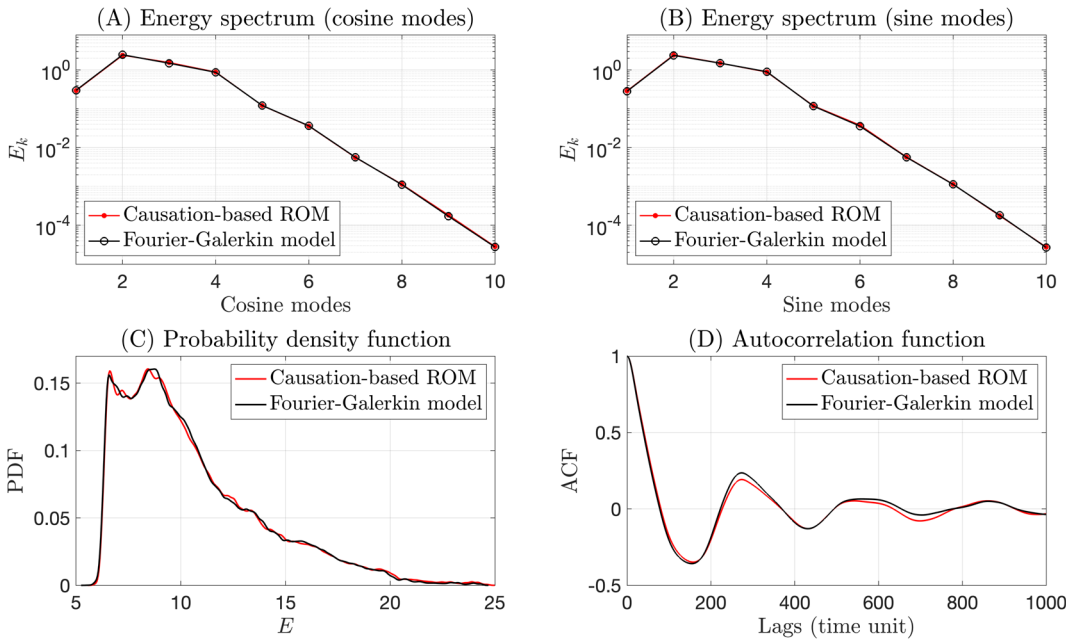

Fig. 6 Comparison of statistics between the 20-dimensional Fourier–Galerkin system and the associated causation-based ROM. The energy spectrum E_k shown here is computed outside of the training window by averaging $(a_k^\ell(t))^2$ over the time window $[5\times10^4, 3\times10^5]$, for all the components $k = 1, \ldots, 10$ and $\ell = 0, 1$. The cosine modes correspond to $\ell = 0$ (Panel A) and sine modes correspond to $\ell = 1$ (Panel B). The PDFs and the ACFs (shown in the bottom panels) for the kinetic energy $E(t) = \sum_{k=1}^{10}\sum_{\ell=0}^{1}(a_k^\ell(t))^2$ are computed over the same time window $[5 \times 10^4, 3 \times 10^5]$

3.3 Data-driven inverse models under the POD basis

We turn now to examine the situation when the underlying orthogonal basis is constructed empirically instead, which is taken to be the POD basis here. For benchmarking purposes, we will compare the performance of the learned model with that of the POD-Galerkin system (17) as well as a thresholded version of the Galerkin system obtained by removing terms whose coefficients in absolute value are below a given threshold to achieve a specified sparsity percentage.

The causation entropy, as computed using the 20-dimensional POD projection of the KSE solution, is shown in Fig. 7. Unlike the case with the Fourier basis shown in Fig. 2, we no longer see a gap that separates a small fraction of larger causation entropy values with the remaining smaller causation entropy values. As mentioned at the end of the previous subsection, a plausible reason is that the POD-Galerkin system (17) itself does not have a sparse structure. Recall that the 20-dimensional Fourier–Galerkin system (13) utilized in the previous subsection has only 295 terms in its vector field, accounting for about 6.41% of the total $20 \times 230 = 4600$ possible monomials in a 20-dimensional quadratic vector field (excluding constant terms). In sharp contrast, almost all the 4600 terms are present in the 20-dimensional POD-Galerkin system (17). As shown in Fig. 8, the absolute value of the coefficients falls in the range $[10^{-5}, 10^{-1}]$ for 96.5% of the terms (namely 4439 terms) in this POD-Galerkin system.

Due to the lack of any obvious cutoff thresholds appearing in the distribution of the causation entropy values, a possible way to proceed is to construct a hierarchy of inverse models that maintain different sparsity percentages by adjusting the cutoff threshold. There are two different ways to carry out this cutoff procedure. One way is to choose a uniform cutoff threshold for all the equations, such as indicated by the red dashed line in Fig. 7. Apparently, this approach only ensures that a given percentage of terms is removed from the learned model but does not guarantee that the percentage of terms removed is the same for each equation in the system. The other way is to choose a

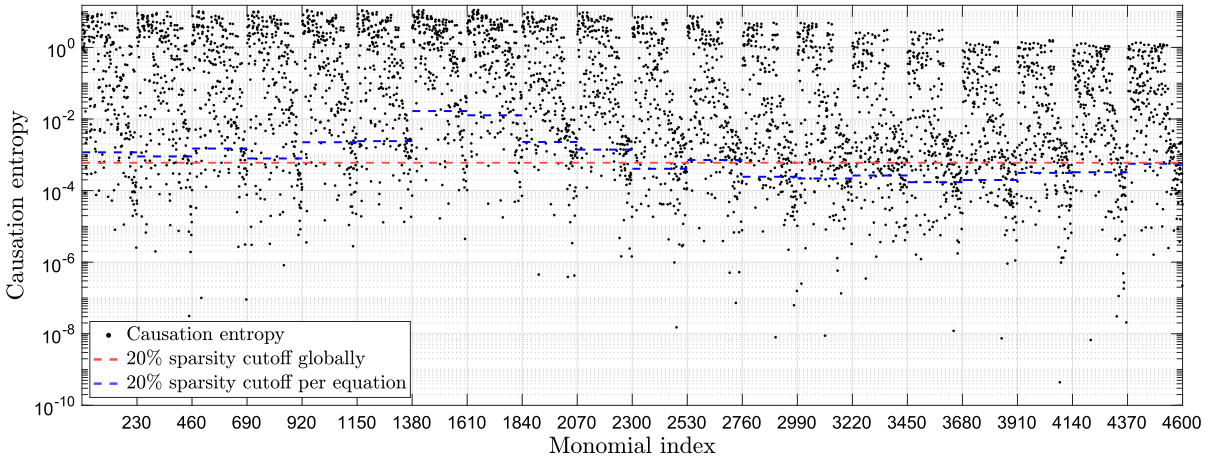

Fig. 7 Causation entropy that ranks the library functions for learning a data-driven quadratic inverse model of the 20-dimensional POD Galerkin system (17). The causation entropy values are grouped by equation in the same way as done in Fig. 2. Also shown are the cutoff thresholds for ensuring 20% sparsity based on two strategies: the red dashed line corresponds to the threshold 6×10^{-4} that separates the lower 20% of all the 4600 causation entropy values from the remaining 80%, while the blue dashed line segments correspond to the thresholds that separate the lower 20% of the 230 causation entropy values for each of the 20 equations. The total number of terms kept in the learned models based on these two strategies are the same, but the constituent terms kept in the corresponding identified models are slightly different

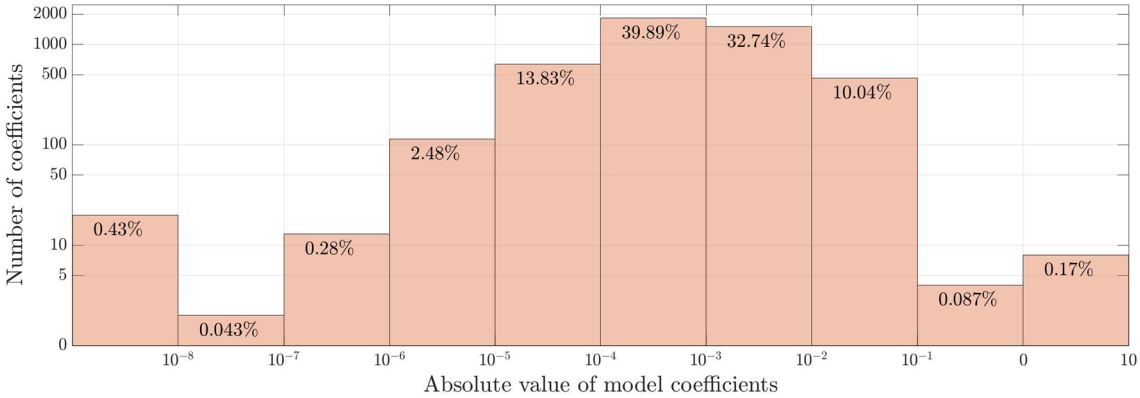

Fig. 8 Distribution of the model coefficients for all the 4600 terms in the 20-dimensional POD-Galerkin system (17) of the KSE (8), for the parameter regime given by Sect. 3.1.3

custom cutoff threshold for each equation, such as indicated by the blue dashed lines in Fig. 7 to achieve the same sparsity percentage for each equation. In principle, the two cutoff procedures can lead to quite different reduced models, especially when the range of the causation entropy values varies significantly from equation to equation. However, for the model considered here, it has been checked that the ROMs obtained by the two approaches for a given sparsity percentage lead to similar modeling performance. For all the numerical results reported below, the causation-based ROMs are constructed using the latter approach to gain the same sparsity percentage for all the equations. Once the constituent terms are determined based on a chosen cutoff threshold strategy for the causation entropy values, we use again the MLE to determine the model coefficients in the causation-based ROMs; see Sect. 2.3.

In Fig. 9, we present the skill of the 20-dimensional causation-based ROM with a 20% sparsity. As can be observed, even though the ROM contains 20% fewer terms than the corresponding POD-Galerkin system, it can faithfully reproduce the essential dynamical fea-

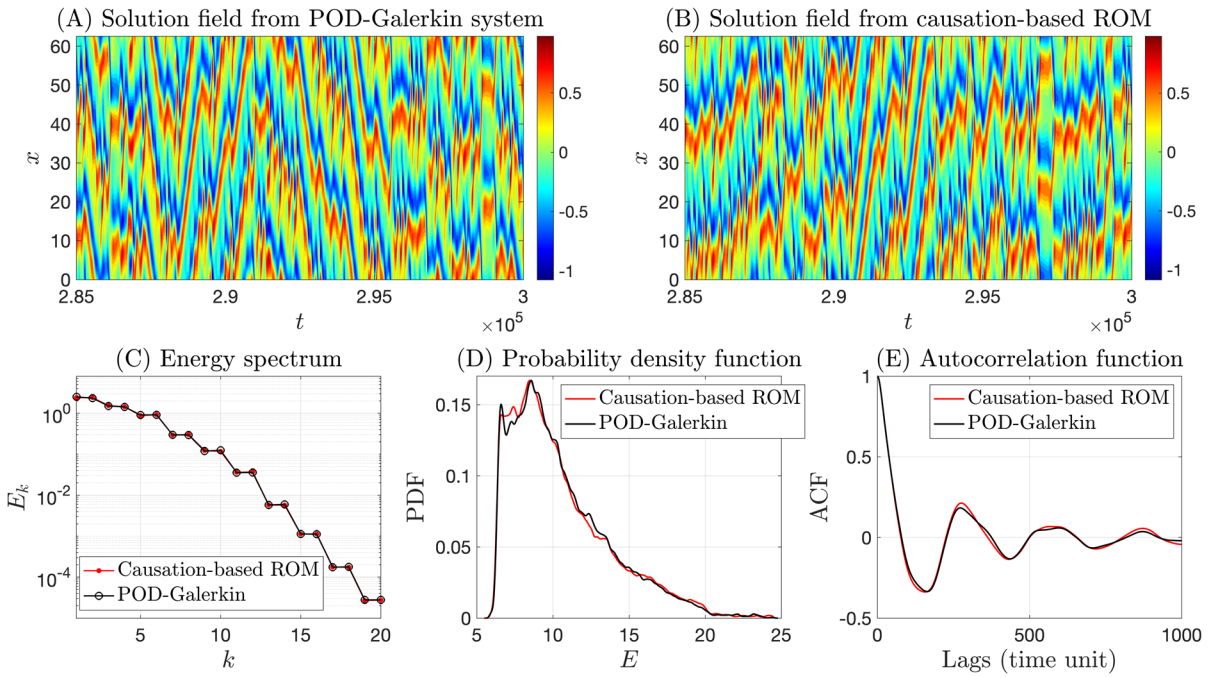

Fig. 9 Performance of the 20-dimensional causation-based ROM under POD basis, with a 20% sparsity cutoff per equation, in comparison with the 20-dimensional POD-Galerkin system. The reconstructed solution fields are shown in Panel A for the POD-Galerkin system and Panel B for the learned model. The energy spectrum E_k is shown in Panel C. The PDFs and the ACFs of the kinetic energy E are shown in Panels D and E, respectively. The energy spectra E_k's and the kinetic energy E are computed in the same way as described in the caption of Fig. 6. Like in Fig. 6, the time window used for computing E_k and E is $[5 \times 10^4, 3 \times 10^5]$, which is outside of the training window $[10^4, 5 \times 10^4]$

tures and the associated statistics appearing in the solution of the POD-Galerkin system. We also checked that even by increasing the sparsity percentage to 50%, the causation-based ROM can still produce chaotic transient dynamics over a long time window with the corresponding solution field resembling that shown in Panels A and B of Fig. 9, although the solution eventually becomes periodic after about 3.6×10^6 time-step iterations. This slow drift to periodic dynamics after long-time integration observed for "severely" truncated causation-based ROMs is not a surprise. The KSE is known to have many periodic dynamics regimes interlaced with chaotic dynamics regimes [53]. In other words, the chaotic attractors observed for the KSE may be prone to instability under perturbations depending on the parameter regimes considered. Since a ROM can be viewed as a perturbation of the original KSE model, it is possible for the dynamics of a highly truncated ROM to be (gradually) pushed towards the basin of attraction of a periodic attractor in a nearby regime.

Although the dynamical features of the KSE dictate that one may not be able to use a too sparse ROM to capture long-term statistics for certain parameter regimes, the fact that such a causation-based ROM can still reproduce accurately short-time features suggests its potential usage for other purposes such as data assimilation and short-term trajectory prediction. In the next subsection, we demonstrate the advantage of such highly sparse causation-based ROMs in the context of data assimilation with partial observations.

3.4 Application to data assimilation with partial observations

We now illustrate the performance of causation-based ROMs in the context of data assimilation to recover unobserved higher-frequency mode dynamics based on observation data for a few low-frequency mode dynamics. As a benchmark, we also compare the results obtained from a thresholded stochastic POD-Galerkin

model described below. For simplicity, we assume that the observation data is available continuously in time, and we perform the data assimilation with the ensemble Kalman–Bucy filter (EnKBF) [7,10] for both reduced systems.

The thresholded Galerkin system is obtained from the true 20-dimensional POD-Galerkin system as follows. We rank the coefficients of the true Galerkin system from large to small in absolute value and then drop the terms with coefficients below a cutoff value that is determined to ensure the number of monomials retained is the same as that of the employed causation-based ROM. After identifying the terms to be kept, we use the MLE to estimate the model coefficients and the covariance matrix of the additive noise term in the final thresholded Galerkin model.

For the sake of clarity, we first provide below some details about the EnKBF applied to a generic n-dimensional SDE system of the form (6):

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \boldsymbol{\Phi}(\boldsymbol{a}) + \boldsymbol{\sigma}\dot{\boldsymbol{W}}(t), \tag{21}$$

in which the noise amplitude matrix $\boldsymbol{\sigma}$ is assumed to be $n \times n$-dimensional and the first r component of $\boldsymbol{a}$ is taken to be observed while the remaining components to be unobserved. We denote

$$\begin{aligned} &\boldsymbol{y} = (a_1, \ldots, a_r)^{\mathrm{T}}, \qquad \boldsymbol{z} = (a_{r+1}, \ldots, a_n)^{\mathrm{T}}, \\ &\boldsymbol{W}_1 = (W_1, \ldots, W_r)^{\mathrm{T}}, \quad \boldsymbol{W}_2 = (W_{r+1}, \ldots, W_n)^{\mathrm{T}}. \end{aligned} \tag{22}$$

We also decompose $\boldsymbol{\sigma}$ into four submatrices

$$\boldsymbol{\sigma} = \begin{pmatrix} \boldsymbol{\sigma}_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\sigma}_{22} \end{pmatrix}, \tag{23}$$

where the dimensions of $\boldsymbol{\sigma}_{11}$ and $\boldsymbol{\sigma}_{22}$ are $r \times r$ and $(n-r) \times (n-r)$, respectively. We then rewrite (21) using $(\boldsymbol{y}, \boldsymbol{z})$ as follows

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} &= \boldsymbol{g}_1(\boldsymbol{y}, \boldsymbol{z}) + \boldsymbol{\sigma}_{11}\dot{\boldsymbol{W}}_1(t), \\ \frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} &= \boldsymbol{g}_2(\boldsymbol{y}, \boldsymbol{z}) + \boldsymbol{\sigma}_{22}\dot{\boldsymbol{W}}_2(t), \end{aligned} \tag{24}$$

where $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ denote respectively the first r and the remaining $n-r$ components of the (nonlinear) function Φ in (21). Note also that compared with the original system (21), we decoupled the noise terms in the $\boldsymbol{y}$- and $\boldsymbol{z}$-subsystems by dropping $\boldsymbol{\sigma}_{12}\dot{\boldsymbol{W}}_2(t)$ in the $\boldsymbol{y}$-subsystem and $\boldsymbol{\sigma}_{21}\dot{\boldsymbol{W}}_1(t)$ in the $\boldsymbol{z}$-subsystem for simplicity. In practice, the noise amplitude matrix $\boldsymbol{\sigma}$ is oftentimes diagonally dominant. This is, in particular, true for the KSE problem considered here. Additionally, the noise in both the causation-based ROM and the thresholded POD-Galerkin system is very weak. Thus, such an approximation has little impact on the accuracy of final data assimilation results.

Assume that a total of p ensemble members are used in the EnKBF. Denote the collection of all the p ensemble members at time t by

$$\boldsymbol{Z}(t) = (z_1(t), z_2(t), \ldots, z_p(t))^{\mathrm{T}}.$$

Denote also the observation data of $\boldsymbol{y}$ at time t by $\boldsymbol{y}_{\mathrm{obs}}(t)$. We define then

$$\begin{aligned} \overline{\boldsymbol{Z}}(t) &= \frac{1}{p}\sum_{\ell=1}^{p} z_\ell(t), \\ \overline{\boldsymbol{g}}_2(\boldsymbol{y}_{\mathrm{obs}}(t), \boldsymbol{Z}(t)) &= \frac{1}{p}\sum_{\ell=1}^{p} \boldsymbol{g}_2(\boldsymbol{y}_{\mathrm{obs}}(t), z_\ell(t)), \end{aligned} \tag{25}$$

and

$$\begin{aligned} \mathcal{N}(\boldsymbol{y}_{\mathrm{obs}}(t), \boldsymbol{Z}(t)) = \frac{1}{(p-1)}\sum_{\ell=1}^{p} &(z_\ell(t) - \overline{\boldsymbol{Z}}(t)) \\ &\Big(\boldsymbol{g}_2(\boldsymbol{y}_{\mathrm{obs}}(t), z_\ell(t)) - \overline{\boldsymbol{g}}_2(\boldsymbol{y}_{\mathrm{obs}}(t), \boldsymbol{Z}(t))\Big)^{\mathrm{T}} C^{-1}, \end{aligned} \tag{26}$$

where $C = \boldsymbol{\sigma}_{22}\boldsymbol{\sigma}_{22}^{\mathrm{T}}$.

Then, each ensemble member z_i, $i = 1, 2, \ldots, p$, of the EnKBF is computed using

$$\begin{aligned} \frac{\mathrm{d}z_i}{\mathrm{d}t} &= \boldsymbol{g}_2(\boldsymbol{y}_{\mathrm{obs}}(t), z_i) + \boldsymbol{\sigma}_{22}\dot{\boldsymbol{W}}_{2,i}(t) \\ &\quad - \mathcal{N}(\boldsymbol{y}_{\mathrm{obs}}(t), \boldsymbol{Z}(t)) \\ &\quad \big[\boldsymbol{g}_1(\boldsymbol{y}_{\mathrm{obs}}(t), z_i) - \dot{\boldsymbol{y}}_{\mathrm{obs}}(t) + \boldsymbol{\sigma}_{11}\dot{\boldsymbol{W}}_{1,i}(t)\big], \end{aligned} \tag{27}$$

where $\boldsymbol{W}_{1,i}$ and $\boldsymbol{W}_{2,i}$ are respectively r-dimensional and $(n-r)$-dimensional Brownian motions for $i = 1, 2, \ldots, p$, with their components to be all mutually independent.

The setup of the data assimilation experiment is as follows. We observe the amplitudes of the first three POD modes of the KSE solutions and aim to recover the amplitudes of the few dominant unobserved modes by applying the EnKBF to either the 20-dimensional causation-based ROM with a large sparsity percentage or the corresponding 20-dimensional thresholded POD-Galerkin system with the same sparsity percentage. We take the size of the EnKBF ensemble to be $p = 500$, and the unobserved variables are initialized to be zero for all the ensemble simulations. The sparsity percentage of the ROMs is taken to be 90%, resulting in 460 terms in the drift part of both the causation-based

ROM and the thresholded POD-Galerkin system. The KSE is simulated over the time window [0, 2000] with the initial data taken to be the solution profile at the last time instant of the training data utilized for constructing the POD basis function as well as the training of the ROMs.

In the first row of Fig. 10, we show the time series of the three observed POD modes. On average, these three modes capture about 63.5% of the kinetic energy in the KSE solution for the considered parameter regime, while above 99% of the kinetic energy is captured by the first 10 POD modes. As shown in Fig. 10 (black curves), modes 4 to 10 still have quite large amplitude oscillations almost comparable with those of the first three modes, and they evolve on different time scales. The fact that the unobserved dynamics still contain, on average, nearly 40% of the kinetic energy and that their projected dynamics reveal multi-scale, highly chaotic oscillatory features present arguably a challenging test ground for the data assimilation experiment.

The red curves in Fig. 10 represent the posterior mean of the unobserved modes a_4 through a_{15} (red curves) obtained from the EnKBF applied to the causation-based ROM, in comparison with the corresponding true POD projections of the KSE solutions (black curves). Despite its highly sparse nature, with 90% sparsity compared with the true POD-Galerkin system of the same dimension, the causation-based ROM is able to recover with high fidelity all the energetic unobserved modes, $a_4, \ldots, a_{12}$. The skill for the remaining small amplitude modes, $a_{13}, \ldots, a_{20}$, deteriorates as the mode index increases, as can be seen in Fig. 10 for modes a_{13}, a_{14}, and a_{15}. However, these 8 modes contain, on average, only approximately 0.13% of the solution energy.

As a comparison, the corresponding results for the 20-dimensional thresholded POD-Galerkin system with 90% sparsity are shown in Fig. 11. The skill is significantly worse than that obtained from the causation-based ROM. When comparing these unobserved dynamics at the spatiotemporal field level, it also reveals that the thresholded POD-Galerkin system with this high truncation ratio suffers particularly severely when there is a relatively abrupt change in the solution dynamics, such as shown at around $t = 1450$ in Fig. 12. Finally, we mention that the time series for all the 500 ensemble members in the data assimilation essentially coincide with each other for both of the two ROMs analyzed due to the fact that the involved noise amplitude matrix $\boldsymbol{\sigma}$ for both ROMs employed has entries all close to zero.

The above results show that causation entropy can indeed be utilized to rank the relative importance of candidate terms from a given function library for the construction of skillful sparse inverse models. The obtained superior data assimilation skills compared with those from the thresholded POD-Galerkin system also illustrate that a naive truncation based on the numerical values of the model coefficients in e.g. a POD-Galerkin system may not be appropriate, especially when a highly truncated ROM is sought.

4 Discussion and conclusions

In this article, we analyzed an efficient approach to identifying data-driven ROMs with a sparse structure using a quantitative indicator called causation entropy. For each potential building-block function f in the vector field of the i-th component a_i, the associated causation entropy measures the difference between the entropy of $\dot{a}_i$ conditioned on the whole set of candidate functions and the one conditioned on the set without f; see (2). Thus, it quantifies statistically the additional contribution of each term to the underlying dynamics beyond the information already captured by all the other terms in the model ansatz.

The ranking of the candidate terms provided by the causation entropy leads to a hierarchy of parsimonious structures for the ROMs controlled by a cutoff threshold parameter. The model coefficients for the corresponding causation-based ROMs can then be learned using standard parameter estimation techniques, such as the MLE; cf. Sect. 2.3. Illustrating on the Kuramoto–Sivashinky equation, we showed in Sect. 3 that the obtained causation-based ROMs are skillful in both recovering long-term statistics and inferring unobserved dynamics via data assimilation when only a small subset of the ROM's state variables is observed.

We conclude by outlining some potential future directions to be explored. For this purpose, we want to emphasize first that, when building up the causation-based ROMs, it is straightforward to add additional physically relevant constraints, such as skew symmetry for certain linear terms and energy conservation for the quadratic nonlinearity. For the results shown in Sect. 2.3, the obtained ROMs turn out to be stable without enforcing energy conservation constraints, even

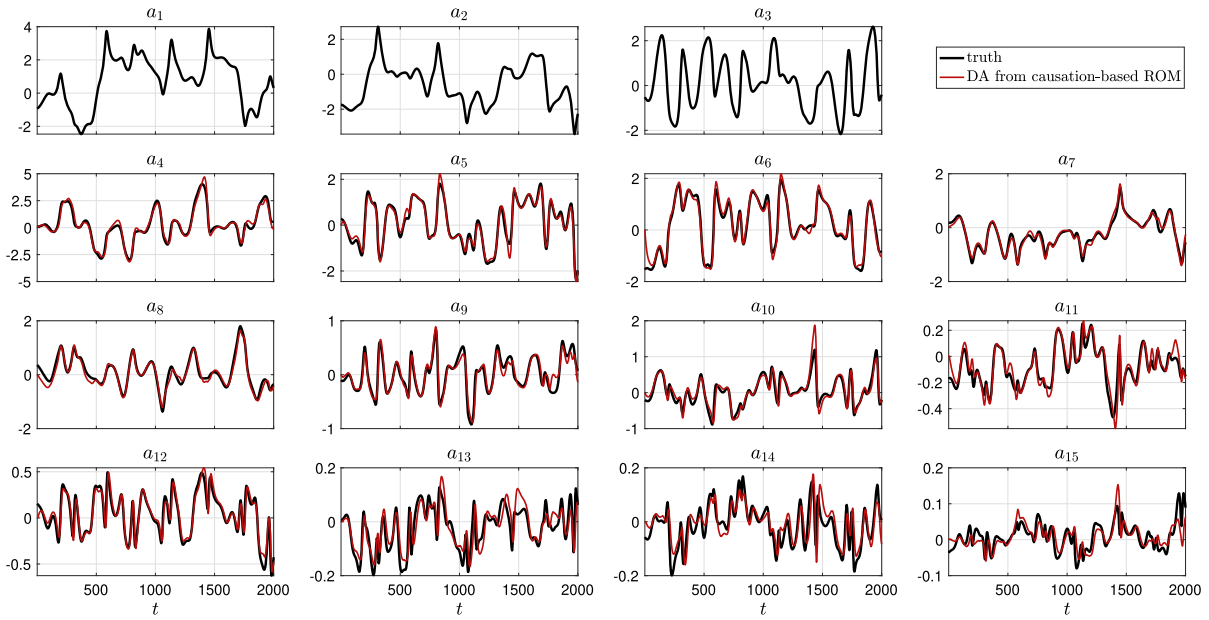

Fig. 10 The black curves show the projections of the true KSE solution onto the first 15 POD modes, with the first three components a_1, a_2, and a_3 taken to be the observed modes in the data assimilation experiments. The red curves show the assimilated ensemble mean dynamics of the unobserved POD modes from the 20-dimensional causation-based ROM, with 90% sparsity

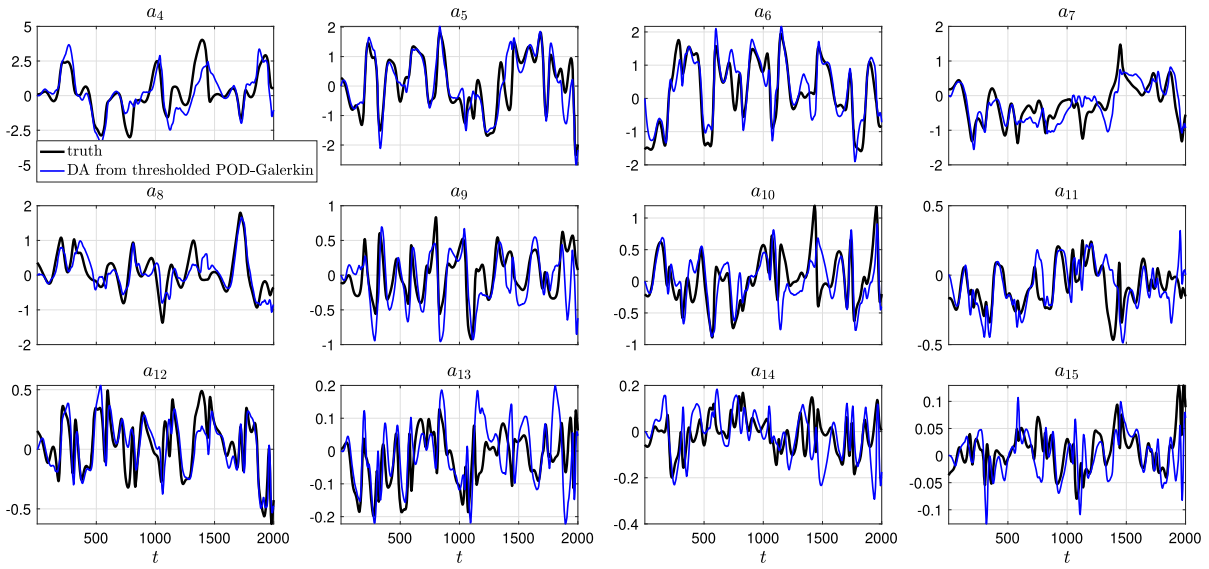

Fig. 11 The assimilated ensemble mean dynamics of the unobserved POD modes from the EnKBF of the 20-dimensional thresholded POD-Galerkin system with 90% sparsity (blue curve), in comparison with the POD projections of the true KSE solution (black curves)

though the quadratic term in the Kuramoto–Sivashinky equation conserves energy. However, such a constraint is expected to be important, e.g., in the reduction of fluid problems in turbulent regimes. To enforce such constraints, we just need to make sure all relevant terms are included in the identified model structure since it can

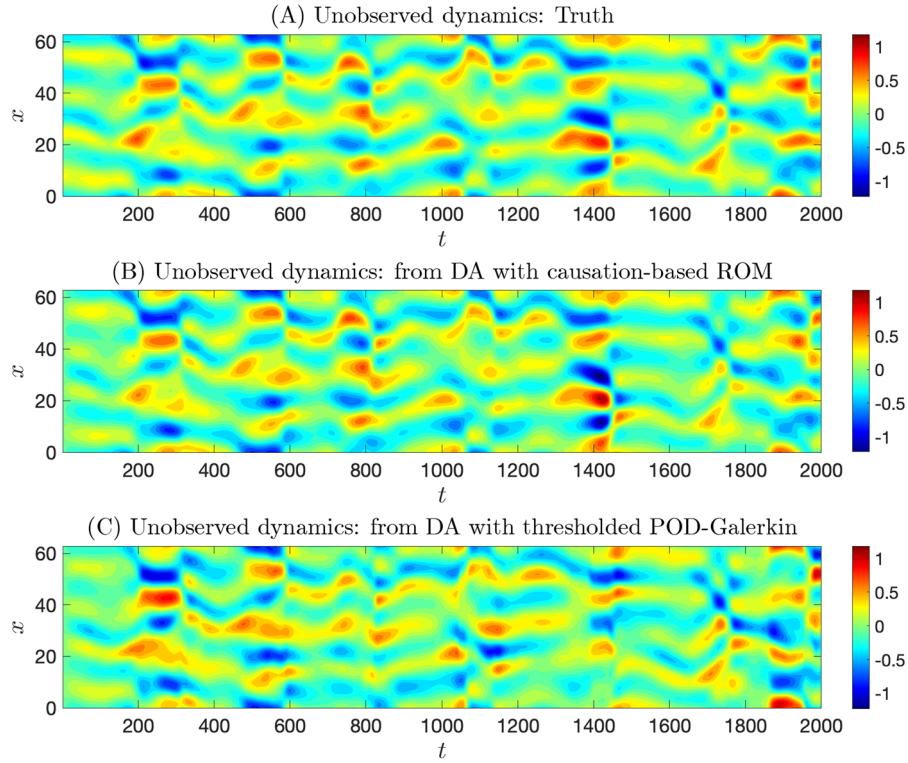

Fig. 12 Comparison of the true unobserved spatiotemporal field (top panel) with those reconstructed based on the assimilated ensemble mean dynamics shown in Fig. 10 for the causation-based ROM (middle panel) and in Fig. 11 for the thresholded POD-Galerkin system (bottom panel)

happen that the causation entropy for some but not all of the terms involved in the constraint is above a given cutoff threshold. Of course, the subsequent parameter estimation is subject to the desired constraints as well, which can be performed using, e.g., the constrained MLE [36, Section 2.5].

Oftentimes, when constructing ROMs for highly chaotic systems, one needs to include closure terms to take into account the impact of the orthogonal dynamics not resolved by the ROMs [2]. Different strategies can be envisioned to extend the current framework for this purpose. For instance, after the drift part of the causation-based ROM is identified (i.e., the $\boldsymbol{\Phi}(\boldsymbol{a})$-term in (6) or (7)), instead of fitting the resulting training residual data by a noise term $\boldsymbol{\sigma}\dot{\boldsymbol{W}}(t)$, we can explore more advanced data-driven techniques such as multilevel approaches and empirical model reduction [60,63,76], nonlinear autoregressive techniques [13,33,37,73], or neural networks. Alternatively, one could first learn a higher-dimensional causation-based ROM, then use parameterization techniques [25,26] to approximate the newly added components by those to be kept.

To what extent one can sparsify a ROM depends apparently on the purposes of the ROMs. However, it can also be tied to the underlying orthogonal basis employed. As already seen in Sect. 3, the causation-based ROMs constructed using the eigenbasis come with a much sparser structure than those built from the POD basis, for the PDE considered. It would be interesting to explore if a coordinate transformation exists that can further enhance the sparsity of the ROMs built on a POD basis. For instance, if we rewrite the POD-ROM under the eigenbasis of the ROM's linear part, we can oftentimes achieve a diagonalization of the linear terms since eigenvalues with multiplicity one are generic. However, whether this transformation can also help aggregate the nonlinearity to form sparser structures (after re-computing the causation entropy matrix under the transformed basis) is up to further investigation.

Another aspect concerns the efficient computation of the causation entropy matrix when the number of functions, M, in the learning library is in the order of several thousand or beyond, which can, for instance, be encountered for ROMs with dimension 100 or higher. The computational cost for determining a causation

entropy lies in the calculation of the log-determinants of the four covariance matrices involved in formula (5), which are of dimension either $M \times M$ or $(M \pm 1) \times (M \pm 1)$. For a ROM of dimension N, there are a total of $N \times M$ causation entropies to determine. Thus, we need to compute the log-determinants of $4 \times N \times M$ covariance matrices, each with dimension about $M \times M$. To gain computational efficiency when M is large, one may benefit from techniques for approximating the log-determinant of a high dimensional symmetric positive definite matrix [16,82], although additional investigation would be needed to see how one can strike a balance between the computational efficiency gained and the approximation error made on each entry of the causation entropy matrix. Alternatively, we can try to reduce the number of functions in the feature library either by an iterative approach using a greedy algorithm [107] or by exploring potential physical/modeling insights for the considered applications. For instance, in [36], a localization strategy is introduced to significantly reduce the size of the feature library when constructing an efficient causation-based ROM for the two-layer Lorenz 1996 model.

Acknowledgements We express our deep gratitude to the reviewers for their valuable feedback, which has substantially enhanced this work.

Funding This research was funded in part by the Army Research Office grant W911NF-23-1-0118 (N.C.), the Office of Naval Research grant N00014-24-1-2244 (N.C.), and the National Science Foundation grants DMS-2108856 and DMS-2407483 (H.L.). We also acknowledge the computational resources provided by Advanced Research Computing at Virginia Tech.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no conflicts to disclose.

References

1. Ahmed, N.A., Gokhale, D.: Entropy expressions and their estimators for multivariate distributions. IEEE Trans. Inf. Theory **35**, 688–692 (1989)
2. Ahmed, S.E., Pawar, S., San, O., Rasheed, A., Iliescu, T., Noack, B.R.: On closures for reduced order models-A spectrum of first-principle to machine-learned avenues. Phys. Fluids **33**(9), 091301 (2021)
3. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proceeding of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest, (1973)
4. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **AC–19**, 716–723 (1974)
5. AlMomani, A.A., Bollt, E.: ERFit: Entropic regression fit MATLAB package, for data-driven system identification of underlying dynamic equations. *arXiv preprint* arXiv:2010.02411, (2020)
6. AlMomani, A.A.R., Sun, J., Bollt, E.: How entropic regression beats the outliers problem in nonlinear system identification. *Chaos: Interdisciplinary J. Nonlinear Sci.*, **30**(1), (2020)
7. Amezcua, J., Ide, K., Kalnay, E., Reich, S.: Ensemble transform Kalman–Bucy filters. Q. J. R. Meteorol. Soc. **140**(680), 995–1004 (2014)
8. Armbruster, D., Heiland, R., Kostelich, E.J., Nicolaenko, B.: Phase-space analysis of bursting behavior in Kolmogorov flow. Physica D **58**, 392–401 (1992)
9. Aubry, N., Lian, W.-Y., Titi, E.S.: Preserving symmetries in the proper orthogonal decomposition. SIAM J. Sci. Comput. **14**, 483–505 (1993)
10. Bergemann, K., Reich, S.: An ensemble Kalman–Bucy filter for continuous data assimilation. Meteorol. Z. **21**, 213–219 (2012)
11. Bertozzi, A.L., Pugh, M.C.: Long-wave instabilities and saturation in thin film equations. Commun. Pure Appl. Math. **51**, 625–661 (1998)
12. Bhola, S., Duraisamy, K.: Estimating global identifiability using conditional mutual information in a Bayesian framework. Sci. Rep. **13**, 18336 (2023)
13. Billings, S.A.: Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-temporal Domains. John Wiley & Sons, New York (2013)
14. Boers, N., Chekroun, M.D., Liu, H., Kondrashov, D., Rousseau, D.-D., Svensson, A., Bigler, M., Ghil, M.: Inverse stochastic-dynamic models for high-resolution Greenland ice-core records. Earth Syst. Dyn. **8**, 1171–1190 (2017)
15. Boninsegna, L., Nüske, F., Clementi, C.: Sparse learning of stochastic dynamical equations. *J. Chem. Pphys.*, 148(24), (2018)
16. Boutsidis, C., Drineas, P., Kambadur, P., Kontopoulou, E.-M., Zouzias, A.: A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. Linear Algebra Appl. **533**, 95–117 (2017)

17. Branicki, M., Majda, A.J.: Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency. Nonlinearity **25**(9), 2543 (2012)
18. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. **113**(15), 3932–3937 (2016)
19. Carlberg, K., Farhat, C., Cortial, J., Amsallem, D.: The GNAT method for nonlinear model reduction: effective implementation and application to computational fluid dynamics and turbulent flows. J. Comput. Phys. **242**, 623–647 (2013)
20. Casella, G., Berger, R.L.: Statistical Inference, 2nd edn. CRC Press, Boca Raton (2024)
21. Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., Kashinath, K.: Towards physically consistent data-driven weather forecasting: integrating data assimilation with equivariance-preserving spatial transformers in a case study with ERA5. In: Geoscientific Model Development Discussions, pp. 1–23, (2021)
22. Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Kashinath, K.: Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. In: Proceedings of the 10th International Conference on Climate Informatics, pp. 106–112, (2020)
23. Chattopadhyay, A., Subel, A., Hassanzadeh, P.: Data-driven super-parameterization using deep learning: experimentation with multiscale Lorenz 96 systems and transfer learning. J. Adv. Model. Earth Syst. **12**(11), e2020MS002084 (2020)
24. Chekroun, M.D., Liu, H., McWilliams, J.C.: The emergence of fast oscillations in a reduced primitive equation model and its implications for closure theories. Comput. Fluids **151**, 3–22 (2017)
25. Chekroun, M.D., Liu, H., McWilliams, J.C.: Variational approach to closure of nonlinear dynamical systems: autonomous case. J. Stat. Phys. **179**, 1073–1160 (2020)
26. Chekroun, M.D., Liu, H., McWilliams, J.C.: Stochastic rectification of fast oscillations on slow manifold closures. Proc. Natl. Acad. Sci. **118**(48), e2113650118 (2021)
27. Chekroun, M.D., Kondrashov, D.: Data-adaptive harmonic spectra and multilayer Stuart-Landau models. Chaos Interdiscip. J. Nonlinear Sci. **27**(9), 093110 (2017)
28. Chekroun, M.D., Kondrashov, D., Ghil, M.: Predicting stochastic systems by noise sampling, and application to the El Niño-southern oscillation. Proc. Natl. Acad. Sci. **108**(29), 11766–11771 (2011)
29. Chen, N.: Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics. J. Comput. Phys. **418**, 109635 (2020)
30. Chen, N., Li, Y.: BAMCAFE: a Bayesian machine learning advanced forecast ensemble method for complex turbulent systems with partial observations. Chaos Interdiscip. J. Nonlinear Sci. **31**(11), 113114 (2021)
31. Chen, N., Li, Y., Liu, H.: Conditional gaussian nonlinear system: a fast preconditioner and a cheap surrogate model for complex nonlinear systems. Chaos Interdiscip. J. Nonlinear Sci. **32**, 053122 (2022)
32. Chen, N., Liu, H.: Minimum reduced-order models via causal inference. *arXiv preprint* arXiv:2407.00271, pp. ges 1–31, (2024)
33. Chen, N., Liu, H., Lu, F.: Shock trace prediction by reduced models for a viscous stochastic burgers equation. Chaos Interdiscip. J. Nonlinear Sci. **32**, 043109 (2022)
34. Chen, N., Majda, A.: Conditional Gaussian systems for multiscale nonlinear stochastic systems: prediction, state estimation and uncertainty quantification. Entropy **20**(7), 509 (2018)
35. Chen, N., Qi, D.: A physics-informed data-driven algorithm for ensemble forecast of complex turbulent systems. Appl. Math. Comput. **466**, 128480 (2024)
36. Chen, N., Zhang, Y.: A causality-based learning approach for discovering the underlying dynamics of complex systems from partial observations with stochastic parameterization. Physica D **449**, 133743 (2023)
37. Chorin, A.J., Lu, F.: Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. Proc. Natl. Acad. Sci. **112**(32), 9804–9809 (2015)
38. Cortiella, A., Park, K.-C., Doostan, A.: Sparse identification of nonlinear dynamical systems via reweighted l1-regularized least squares. Comput. Methods Appl. Mech. Eng. **376**, 113620 (2021)
39. Cover, T., Thomas, J.: Elements of Information Theory, 2nd edn. John Wiley & Sons, New York (2006)
40. Crommelin, D., Majda, A.: Strategies for model reduction: comparing different optimal bases. J. Atmos. Sci. **61**(17), 2206–2217 (2004)
41. Darbellay, G.A., Vajda, I.: Estimation of the information by an adaptive partitioning of the observation space. IEEE Trans. Inf. Theory **45**, 1315–1321 (1999)
42. Elinger, J.: *Information Theoretic Causality Measures For Parameter Estimation and System Identification.* PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, (2021)
43. Elinger, J., Rogers, J.: Causation entropy method for covariate selection in dynamic models. In: 2021 American Control Conference (ACC), pp. 2842–2847. IEEE, (2021)
44. Fish, J., DeWitt, A., AlMomani, A.A.R., Laurienti, P.J., Bollt, E.: Entropic regression with neurologically motivated applications. Chaos Interdiscip. J. Nonlinear Sci., **31**(11), (2021)
45. Ghil, M., Childress, S.: Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics. Springer Science & Business Media, Berlin (2012)
46. Hannachi, A., Jolliffe, I.T., Stephenson, D.B.: Empirical orthogonal functions and related techniques in atmospheric science: a review. Int. J. Climatol. **27**, 1119–1152 (2007)
47. Harlim, J., Mahdi, A., Majda, A.J.: An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. J. Comput. Phys. **257**, 782–812 (2014)
48. Hasselmann, K.: PIPs and POPs: the reduction of complex dynamical systems using principal interaction and oscillation patterns. J. Geophys. Res.: Atmospheres **93**(D9), 11015–11021 (1988)
49. Herawati, N., Nisa, K., Setiawan, E., Nusyirwan, N., Tiryono, T.: Regularized multiple regression methods to deal with severe multicollinearity. Int. J. Stat. Appl. **8**, 167–172 (2018)

50. Hijazi, S., Stabile, G., Mola, A., Rozza, G.: Data-driven pod-galerkin reduced order model for turbulent flows. J. Comput. Phys. **416**, 109513 (2020)
51. Holmes, P., Lumley, J.L., Berkooz, G.: Turbulence, Coherent Structures. Dynamical Systems and Symmetry, Cambridge (1996)
52. Holmes, P., Lumley, J.L., Berkooz, G., Rowley, C.W.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry, 2nd edn. Cambridge University Press, Cambridge (2012)
53. Hyman, J.M., Nicolaenko, B., Zaleski, S.: Order and complexity in the Kuramoto–Sivashinsky model of weakly turbulent interfaces. Physica D **23**, 265–292 (1986)
54. Jardak, M., Navon, I., Zupanski, M.: Comparison of sequential data assimilation methods for the Kuramoto–Sivashinsky equation. Int. J. Numer. Meth. Fluids **62**, 374–402 (2010)
55. Kaiser, J., Reed, W.: Data smoothing using low-pass digital filters. Rev. Sci. Instrum. **48**, 1447–1457 (1977)
56. Kassam, A., Trefethen, L.N.: Fourth-order time-stepping for stiff PDEs. SIAM J. Sci. Comp. **26**(4), 1214–1233 (2005)
57. Kim, P., Rogers, J., Sun, J., Bollt, E.: Causation entropy identifies sparsity structure for parameter estimation of dynamic systems. J. Comput. Nonlinear Dyn. **12**(1), 011008 (2017)
58. Kleeman, R.: Information theory and dynamical system predictability. Entropy **13**(3), 612–649 (2011)
59. Koc, B., Mou, C., Liu, H., Wang, Z., Rozza, G., Iliescu, T.: Verifiability of the data-driven variational multiscale reduced order model. J. Sci. Comput. **93**(54), 1–26 (2022)
60. Kondrashov, D., Chekroun, M.D., Ghil, M.: Data-driven non-Markovian closure models. Physica D **297**, 33–55 (2015)
61. Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. Probl. Inf. Transm. **23**, 95–102 (1987)
62. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**, 066138 (2004)
63. Kravtsov, S., Kondrashov, D., Ghil, M.: Multilevel regression modeling of nonlinear processes: derivation and applications to climatic variability. J. Clim. **18**(21), 4404–4424 (2005)
64. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for parabolic problems. Numer. Math. **90**, 117–148 (2001)
65. Kuramoto, Y., Tsuzuki, T.: Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. Prog. Theor. Phys. **55**(2), 356–369 (1976)
66. Kwasniok, F.: The reduction of complex dynamical systems using principal interaction patterns. Physica D **92**(1–2), 28–60 (1996)
67. Kwasniok, F.: Optimal Galerkin approximations of partial differential equations using principal interaction patterns. Physical Rev. E **55**(5), 5365 (1997)
68. LaQuey, R., Mahajan, S., Rutherford, P., Tang, W.: Nonlinear saturation of the trapped-ion mode. Phys. Rev. Lett. **34**, 391–394 (1975)
69. Larios, A., Pei, Y.: Nonlinear continuous data assimilation. Evolut. Equ. Control Theory **13**, 329–348 (2024)
70. Lee, T.-W.: Independent Component Analysis: Theory and Applications. Springer, Berlin (1998)
71. Lin, K.K., Lu, F.: Data-driven model reduction, wiener projections, and the Koopman–Mori–Zwanzig formalism. J. Comput. Phys. **424**, 109864 (2021)
72. Lozano-Durán, A., Arranz, G.: Information-theoretic formulation of dynamical systems: causality, modeling, and control. Phys. Rev. Res. **4**, 023195 (2022)
73. Lu, F., Lin, K.K., Chorin, A.J.: Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation. Physica D **340**, 46–57 (2017)
74. Lunasin, E., Titi, E.S.: Finite determining parameters feedback control for distributed nonlinear dissipative systems—a computational study. Evolut. Equ. Control Theory **6**, 535–557 (2017)
75. Majda, A.J., Chen, N.: Model error, information barriers, state estimation and prediction in complex multiscale systems. Entropy **20**(9), 644 (2018)
76. Majda, A.J., Harlim, J.: Physics constrained nonlinear regression models for time series. Nonlinearity **26**(1), 201 (2012)
77. Moosavi, A., Stefanescu, R., Sandu, A.: Efficient construction of local parametric reduced order models using machine learning techniques. *arXiv preprint* arXiv:1511.02909, (2015)
78. Mou, C., Koc, B., San, O., Rebholz, L.G., Iliescu, T.: Data-driven variational multiscale reduced order models. Comput. Methods Appl. Mech. Eng. **373**, 113470 (2021)
79. Mou, C., Smith, L.M., Chen, N.: Combining stochastic parameterized reduced-order models with machine learning for data assimilation and uncertainty quantification with partial observations. J. Adv. Modeling Earth Syst. **15**(10), e2022MS003597 (2023)
80. Noack, B.R., Morzynski, M., Tadmor, G.: Reduced-Order Modelling for Flow Control, vol. 528. Springer Science & Business Media, Berlin (2011)
81. Otto, S.E., Rowley, C.W.: Linearly recurrent autoencoder networks for learning dynamics. SIAM J. Appl. Dyn. Syst. **18**, 558–593 (2019)
82. Pace, R.K., LeSage, J.P.: A sampling approach to estimate the log determinant used in spatial likelihood problems. J. Geogr. Syst. **11**(3), 209–225 (2009)
83. Pawar, S., Ahmed, S.E., San, O., Rasheed, A.: Data-driven recovery of hidden physics in reduced order modeling of fluid flows. Phys. Fluids **32**(3), 036602 (2020)
84. Peherstorfer, B., Willcox, K.: Dynamic data-driven reduced-order models. Comput. Methods Appl. Mech. Eng. **291**, 21–41 (2015)
85. Penland, C., Magorian, T.: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. J. Clim. **6**, 1067–1076 (1993)
86. Rigney, D. R., Goldberger, A. L., Ocasio, W. C., Ichimaru, Y., Moody, G. B., Mark, R. G.: Multi-channel physiological data: Description and analysis (Data Set B). In: A. S. Weigend and N. A. Gershenfeld, (eds.) Time Series Prediction: Forecasting the Future and Understanding the Past, pp. 105–130. Routledge, Taylor & Francis Group, New York, London, (1994)
87. Rish, I., Grabarnik, G.Y.: Sparse Modeling: Theory, Algorithms, and Applications. CRC Press, Boca Raton (2014)

88. Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., Henningson, D.S.: Spectral analysis of nonlinear flows. J. Fluid Mech. **641**, 115–127 (2009)
89. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
90. San, O., Maulik, R.: Extreme learning machine for reduced order modeling of turbulent geophysical flows. Phys. Rev. E **97**(4), 042322 (2018)
91. Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. **7**(4), 1307–1330 (1986)
92. Schaeffer, H., Tran, G., Ward, R.: Extracting sparse high-dimensional dynamics from limited data. SIAM J. Appl. Math. **78**(6), 3279–3295 (2018)
93. Schmid, P.J.: Dynamic mode decomposition of numerical and experimental data. J. Fluid Mech. **656**, 5–28 (2010)
94. Schneider, T., Stuart, A.M., Wu, J.-L.: Learning stochastic closures using ensemble Kalman inversion. Trans. Math. Appl. **5**(1), tnab003 (2021)
95. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**, 461–464 (2000)
96. Seber, G.A.F., Lee, A.J.: Linear Regression Analysis, 2nd edn. John Wiley & Sons, Hoboken (2003)
97. Sheard, S.A., Mostashari, A.: Principles of complex systems for systems engineering. Syst. Eng. **12**(4), 295–311 (2009)
98. Sirovich, L.: Turbulence and the dynamics of coherent structures. Parts I-III. Quart. Appl. Math. **45**(3), 561–590 (1987)
99. Sivashinsky, G.: Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations. Acta Astronautica **4**(11–12), 1177–1206 (1977)
100. Sivashinsky, G.I., Michelson, D.M.: On irregular wavy flow of a liquid film down a vertical plane. Progress Theoret. Phys. **63**, 2112–2114 (1980)
101. Smarra, F., Jain, A., De Rubeis, T., Ambrosini, D., D'Innocenzo, A., Mangharam, R.: Data-driven model predictive control using random forests for building energy optimization and climate control. Appl. Energy **226**, 1252–1272 (2018)
102. Snyder, W., Mou, C., Liu, H., San, O., De Vita, R., Iliescu, T.: Reduced order model closures: a brief tutorial. In Recent Advances in Mechanics and Fluid-Structure Interaction with Applications: The Bong Jae Chung Memorial Volume, pp. 167–193. Springer, (2022)
103. Srinivasan, K., Chekroun, M.D., McWilliams, J.C.: Turbulence closure with small, local neural networks: forced two-dimensional and β-plane flows. J. Adv. Modeling Earth Syst. **16**, e2023MS003795 (2024)
104. Stinis, P.: Stochastic optimal prediction for the Kuramoto–Sivashinsky equation. Multiscale Model. Simul. **2**(4), 580–612 (2004)
105. Strogatz, S.H.: Nonlinear Dynamics and Chaos with Student Solutions Manual: With Applications to Physics, Biology, Chemistry, and Engineering. CRC Press, Boca Raton (2018)
106. Sun, J., Bollt, E.M.: Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. Physica D **267**, 49–57 (2014)
107. Sun, J., Taylor, D., Bollt, E.M.: Causal network inference by optimal causation entropy. SIAM J. Appl. Dyn. Syst. **14**, 73–106 (2015)
108. Taira, K., Brunton, S.L., Dawson, S.T.M., Rowley, C.W., Colonius, T., McKeon, B.J., Schmidt, O.T., Gordeyev, S., Theofilis, V., Ukeiley, L.S.: Modal analysis of fluid flows: an overview. AIAA J. **55**, 4013–4041 (2017)
109. Taira, K., Hemati, M.S., Brunton, S.L., Sun, Y., Duraisamy, K., Bagheri, S., Dawson, S.T., Yeh, C.-A.: Modal analysis of fluid flows: applications and outlook. AIAA J. **58**(3), 998–1022 (2020)
110. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Applied Mathematical Sciences, vol. 68, 2nd edn. Springer, New York (1997)
111. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
112. Tippett, M.K., Kleeman, R., Tang, Y.: Measuring the potential utility of seasonal climate predictions. *Geophys. Res. Lett.*, 31(22), (2004)
113. Tu, J.H., Rowley, C.W., Luchtenburg, D.M., Brunton, S.L., Kutz, J.N.: On dynamic mode decomposition: theory and applications. J. Comput. Dyn. **1**, 391–421 (2014)
114. Vallis, G.K.: Atmospheric and Oceanic Fluid Dynamics. Cambridge University Press, Cambridge (2017)
115. Vautard, R., Yiou, P., Ghil, M.: Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. Physica D **58**, 95–126 (1992)
116. Wan, Z.Y., Sapsis, T.P.: Reduced-space Gaussian process regression for data-driven probabilistic forecast of chaotic dynamical systems. Physica D **345**, 40–55 (2017)
117. Wilcox, D.C.: Multiscale model for turbulent flows. AIAA J. **26**(11), 1311–1320 (1988)
118. Williams, M., Kevrekidis, I., Rowley, C.: A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. J. Nonlinear Sci. **25**(6), 1307–1346 (2015)
119. Wolf, A., Swift, J.B., Swinney, H.L., Vastano, J.A.: Determining lyapunov exponents from a time series. Physica D **16**, 285–317 (1985)
120. Wyner, A.D.: A definition of conditional mutual information for arbitrary ensembles. Inf. Control **38**, 51–59 (1978)
121. Xie, X., Mohebujjaman, M., Rebholz, L.G., Iliescu, T.: Data-driven filtered reduced order modeling of fluid flows. SIAM J. Sci. Comput. **40**(3), B834–B857 (2018)