

Ensemble Kalman Inversion For Sparse Learning Of Dynamical Systems From Time-Averaged Data

Tapio Schneider^a, Andrew M. Stuart^a, Jin-Long Wu^a

^a*California Institute of Technology, Pasadena, CA 91125.*

Abstract

Enforcing sparse structure within learning has led to significant advances in the field of data-driven discovery of dynamical systems. However, such methods require access not only to time-series of the state of the dynamical system, but also to the time derivative. In many applications, the data are available only in the form of time-averages such as moments and autocorrelation functions. We propose a sparse learning methodology to discover the vector fields defining a (possibly stochastic or partial) differential equation, using only time-averaged statistics. Such a formulation of sparse learning naturally leads to a nonlinear inverse problem to which we apply the methodology of ensemble Kalman inversion (EKI). EKI is chosen because it may be formulated in terms of the iterative solution of quadratic optimization problems; sparsity is then easily imposed. We then apply the EKI-based sparse learning methodology to various examples governed by stochastic differential equations (a noisy Lorenz 63 system), ordinary differential equations (Lorenz 96 system and coalescence equations), and a partial differential equation (the Kuramoto-Sivashinsky equation). The results demonstrate that time-averaged statistics can be used for data-driven discovery of differential equations using sparse EKI. The proposed sparse learning methodology extends the scope of data-driven discovery of differential equations to previously challenging applications and data-acquisition scenarios.

Keywords: Ensemble Kalman inversion, sparse learning, dynamical systems, time-averaged data

Email addresses: tapio@caltech.edu (Tapio Schneider), astuart@caltech.edu (Andrew M. Stuart), jinlong@caltech.edu (Jin-Long Wu)

1. Introduction

1.1. Overview and Literature Review

The goal of this paper is to describe a sparse learning methodology to discover the vector fields defining a (possibly stochastic or partial) differential equation, using time-series data. The approach is to use ensemble Kalman inversion (EKI) to learn the unknown vector fields by matching the model to time-averaged statistics derived from the time-series data. Sparse learning allows for the discovery of dynamical models from within a large dictionary of models; learning from time-averaged statistics is often necessary either because data are available only in that form, or because use of time-averages avoids incompatibility issues between model and data at small time increments and the lack of differentiability of sample paths of stochastic differential equations (SDEs). The EKI methodology is an approach to parameter identification, which lends itself naturally to the imposition of constraints such as sparsity, and which is known empirically to be robust and flexible. The work presented here leads to a new approach which widens the scope of sparse learning problems in dynamical systems and that is demonstrably both flexible (applying to a range of examples) and robust (works very well in practice). Our work suggests wider deployment of the proposed methodology, and the need for development of an underpinning theory.

Seeking sparse structure in learning has played a significant role in recent decades. Sparse dictionary learning techniques are well known as compressed sensing [1–3] and have already been extensively studied in application domains such as image and signal processing [4]. The general concept of incorporating sparsity into optimization has also been studied in a variety of applied disciplines for several decades, for example in applications in geophysics [5]. Since then it has been formulated as a theoretical framework known as LASSO [6, 7]. In addition, sparsity-promoting techniques have been found useful in emerging areas in artificial intelligence, such as deep learning [8].

Exploitation of sparsity in the data-driven discovery of differential equations was pioneered in a recent series of papers [9–12], all of which make the assumption that nature favors simplicity and that the vector fields to be discovered are sparse within a high-dimensional dictionary. More recently, a sparsity-promoting joint outlier detection and model discovery method was proposed in [13], and a sparsity-promoting method was proposed in [14] for learning governing equations of dynamical systems from undersampled data. The data-driven discovery of differential equations with sparsity has also been investigated for the learning of stochastic differential equations [15, 16]. These methods need to be provided with, or need to numerically evaluate, time-series of the time derivative of the state variables, as well as the time-series of state variables themselves. In this setting, the learning problem may be phrased as an over-determined system of linear equations, and the solution may be sought through a regularized least squares approach in which the regularization imposes sparsity.

When numerical differentiation is required, it is susceptible to noise in the time-series of the state variables. Techniques such as total variation regularization [17] have been adopted to alleviate this issue by denoising the time derivative [18]. Nonetheless, the presence of noise in time-series presents a significant issue for these approaches.

One way of circumventing the need to numerically differentiate time-series is to fit dynamical models to statistics derived from time-series, such as moments or autocorrelation functions. This idea is widely used in the study of autoregressive (AR) processes [19–24] and in the study of SDEs [25–29]. There are also a plethora of applied papers that take this approach, in both discrete and continuous time, such as [30, 31]. In addition to avoiding numerical differentiation, such methods also have the potential to learn models when only a subset of the state variables are observed. Furthermore, there are settings in which only time-averaged data are available. It is important to note that the parameter-to-data map for such problems is nonlinear.

EKI is a general methodology for nonlinear inverse problems described in [32], building on algorithms designed for the solution of inverse problems arising in oil reservoir simulation [33, 34]. The incorporation of regularization into EKI is discussed in [35, 36], and the incorporation of constraints in [37–41]. In this work, we propose an EKI-based sparsity-promoting methodology for parameter learning. This sparse EKI method combines ideas from [35, 37] to create a derivative-free optimization approach to parameter learning that enforces sparsity. We apply the method to the learning of vector fields in (possibly stochastic or partial) differential equations, building on the ideas in [9], but using nonlinear indirect measurements defined by time-averaging, rather than linear direct observations. It is a remarkable property of EKI that, despite the nonlinearity of the observation operator, the core computational task is the minimization of a quadratic objective functional to which a sparsity constraint maybe easily added, just as it is in the original work on sparse learning of dynamical systems in [9]. This fact allows transfer of the learning framework introduced in [9] to more complex indirect, nonlinearly and partially observed dynamics, and indeed to a wide range of nonlinear inverse and parameter identification problems.

1.2. Our Contribution

Our contributions in this paper are as follows:

- We demonstrate how to impose a sparsity constraint within the EKI algorithm, by formulating the update step as an ℓ_1 and/or ℓ_0 regularized least squares problem.
- We demonstrate the use of sparsity-promoting EKI to discover the governing equations of (possibly stochastic) dynamical systems based on statistics derived from averaging time-series. The results are compared with those obtained using standard EKI to illustrate the merits of imposing sparsity.

- We illustrate the methodology in two simulation studies, discovering the stochastic Lorenz '63 and the deterministic Lorenz '96 systems from data, and we also illustrate the methodology to find a closure model for the slow variables within a multiscale Lorenz '96 system.
- We illustrate the methodology by discovering coalescence equations for collisional dynamics, using both simulation studies and closure models.
- We illustrate the methodology in the context of discovering the Kuramoto-Sivashinsky equation from a larger family of linear dissipative and dispersive linear systems subject to energy conserving quadratic nonlinearities.
- We demonstrate how to impose constraints on the parameter learning, which ensure that the subset of parameters that are queried during the algorithm all lead to well-posed dynamical systems.

Furthermore, although we apply the method in the context of learning dynamical systems, the EKI-based sparse methodology may be more widely applied within nonlinear inverse problems generally.

In Section 2, we formulate the inverse problem of interest and introduce the four problem classes to which we will apply our methodology. Section 3 describes the ensemble Kalman-based methodology which we employ to solve the inverse problem. It also discusses the quadratic programming approach we employ to incorporate the ℓ_1 -penalty into the ensemble Kalman-based methodology, and the proximal gradient methodology used to incorporate the ℓ_0 -penalty. In Section 4, we describe numerical results relating to each of the four example problems. We conclude in Section 5.

1.3. Notation

Throughout we use $|\cdot|_{\ell_p}$ to denote the p -norm on Euclidean space, extended to include the case $p = 0$, which counts the number of non-zero entries of the vector. The commonly occurring case $p = 2$ is simply denoted by $|\cdot|$, and the notation $|\cdot|_A := |A^{-\frac{1}{2}} \cdot|$ is used for symmetric positive-definite A .

2. Problem Formulation

The aim of this work is to use time-series data to learn the right hand side of a differential equation

$$\frac{dx}{dt} = f(x), \tag{2.1}$$

where $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \mapsto \mathbb{R}^n$, from time-averaged information about x . To this end, we first approximate f with a set of basis functions $\phi = \{\phi_i\}$, $i \in \{1, \dots, p\}$, leading to a modeled differential equation

$$\frac{dX_k}{dt} = \sum_{i=1}^p \theta_{ki} \phi_i(X), \quad k = 1, \dots, n, \quad (2.2)$$

where $X \in \mathbb{R}^n$ with components X_k , $\phi_i : \mathbb{R}^n \mapsto \mathbb{R}$, and the parameter matrix $\theta \in \Theta \subseteq \mathbb{R}^{n \times p}$. We assume that, with appropriate choice of the basis functions $\{\phi_i\}$, the function $X(t)$ provides a good approximation of $x(t)$ for some choice of parameter matrix θ ; furthermore, we assume that this choice of θ is sparse in the sense that $|\theta|_{\ell_0} \ll np$. We will also consider generalizations to SDEs and to partial differential equations (PDEs).

We assume that the data y available to us is in the form of time-averages of quantities derived from $x(t)$, or linear transformations of such quantities. This includes moments, autocorrelation functions, and the power spectral density. If $x(\cdot), X(\cdot; \theta) \in \mathcal{X} := C(\mathbb{R}^+, \mathbb{R}^n)$ denote solutions of the true and modeled systems, Θ denotes the subset of parameter space over which we seek modeled solutions, and $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}^J$ is a function on the space of solution trajectories, then define $\mathcal{G}(\theta) := \mathcal{F}(X(\cdot; \theta)) : \Theta \mapsto \mathbb{R}^J$. In this work, \mathcal{F} corresponds to time-averaged functions of solution trajectories $x(\cdot)$. We focus on solution of the following inverse problem to determine θ from y :

$$y = \mathcal{G}(\theta). \quad (2.3)$$

For simplicity we have assumed independence of \mathcal{G} on the initial condition (and the driving Brownian noise in the SDE case), noting that for ergodic problems this dependence indeed disappears when time-averages over the infinite time horizon are used. In practice, a noisy finite-time average is used to generate the data, and the resulting fluctuations may be viewed as small noise around the infinite time average, and we will account for this in our algorithms. The ergodic setting will obtain for most of the examples considered in this paper. However, one of the examples we study is not ergodic (the coalescence equations), and in that setting we study the dependence of our learned parameters on the initial condition.

The formulation in Eq. (2.3) has the advantage that it does not involve the matching of trajectories $x(t)$ and $X(t)$, a problem that can be difficult when noise is present in the data (for example from using finite-time rather than ergodic averages) or when the trajectory is not differentiable (as arises in SDEs). However the approach we adopt has the apparent disadvantage that the data available may be of small volume. Indeed, it may be the case that $J \ll np$ — that is, the amount of data is far less than the number of unknowns. Nonetheless, nature favors simplicity in many cases, and then a sparse solution for θ provides a better modeled system than a dense one and can still be identifiable with limited data. Therefore, we aim to solve the inverse problem formulated in Eq. (2.3) by using a modified version of ensemble Kalman inversion (EKI) that promotes sparsity in θ .

We now describe four examples that will be used to illustrate the methodology. In all four cases, we demonstrate how to ensure that parameter learning takes place within a subset of parameter models that lead to well-posed dynamics. The issue of ensuring this does not arise in the approach of [9] because the dynamical system is not simulated as part of the algorithm. For the EKI approach adopted here, it is integral to the method that the candidate model problems are simulated for a variety of parameter values during the learning algorithm, and the resulting outputs compared with the data available. This ensures that the candidate parameter values lead to well-posed dynamics. The following four examples will be used in the numerical illustrations in Section 4. The reader primarily interested in the form of the algorithm can skip straight to Section 3 and return to these examples in conjunction with reading Section 4.

Example 2.1 (Lorenz 63 System [42]). *The noisy Lorenz equations are a system of three ordinary differential equations taking the form*

$$\dot{x} = f(x) + \sqrt{\sigma^\dagger} \dot{W}, \quad (2.4)$$

where W is an \mathbb{R}^3 -valued Brownian motion, $x = [x_1, x_2, x_3]^\top$, and $f : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is given by

$$\begin{aligned} f_1(x) &= \alpha(x_2 - x_1), \\ f_2(x) &= x_1(\rho - x_3) - x_2, \\ f_3(x) &= x_1x_2 - \beta x_3. \end{aligned} \quad (2.5)$$

We will seek a modeled system of the form

$$\dot{X}_k = \sum_{i=1}^9 \theta_{ki} \phi_i(X) + \sqrt{\sigma} \dot{W}_k, \quad k = 1, 2, 3, \quad (2.6)$$

where $\phi = \{\phi_i\}$, $i \in \{1, \dots, 9\}$ contains all the first ($i \in \{1, 2, 3\}$) and second-order ($i \in \{6, \dots, 9\}$) polynomial basis functions and the W_k are independent \mathbb{R} -valued Brownian motions.

In this setting, the modeled system in Eq. (2.6) coincides with the true system in Eq. (2.4) with a proper choice of parameters θ and noise level σ . This example thus serves as a simulation study, while also illustrating the applicability of our sparse discovery method to SDEs, hence going beyond [9].

The parameter vector θ contains 27 unknowns. To ensure well-posedness of the explored model-class we further impose that the quadratic terms are energy conserving; specifically, we enforce that the inner-product of the quadratic terms with X is identically zero:

$$\sum_{k=1}^3 \sum_{i=4}^9 X_k \theta_{ki} \phi_i(X) \equiv 0. \quad (2.7)$$

This ensures that the quadratic term contributes zero energy to the system and is natural from the viewpoint of the geophysical modelling considerations that underpin the model. Mathematically, imposition of (2.7) ensures that the stochastic differential equation does not explode in finite time as it implies boundedness of the second moment at any fixed positive time [43, 44]. The constraints in (2.7) number 10, corresponding to removal of X_1^3 , X_2^3 , X_3^3 , $X_1^2X_2$, $X_1^2X_3$, $X_2^2X_1$, $X_2^2X_3$, $X_3^2X_1$, $X_3^2X_2$, and $X_1X_2X_3$ from the energy. Consequently, there remain 17 independent unknown coefficients in $\theta = \{\theta_{ki}\}$, $k \in \{1, 2, 3\}$, $i \in \{1, \dots, 9\}$ after incorporating the energy constraint. Our goal is to learn a sparse solution θ which has less than 17 non-zero elements, as well as the noise level σ . Ideally, in this simulation study, the learnt solution for θ will have 7 non-zero elements that agree with the true system in Eq. (2.4), and a value of σ which agrees with the true value σ^\dagger .

Proposition 1. Assume that the constraints on parameters $\{\theta_{ki}\}$ are chosen as detailed above, so as to ensure (2.7) holds. Then for any $T > 0$, there are constants $c_1, c_2 > 0$ such that equation (2.6) has, almost surely, a unique solution satisfying $u = (X_1, X_2, X_3) \in C([0, T]; \mathbb{R}^3)$, and

$$\sup_{t \in [0, T]} \mathbb{E}|u|^2 \leq (|u_0|^2 + c_1)e^{c_2 T}. \quad \diamond$$

Proof. Define the Lyapunov function $V(u) = \frac{1}{2}|u|^2$. Applying the Itô formula to u solving (2.6) gives

$$\frac{d}{dt} \{\mathbb{E}V(u)\} = \mathbb{E} \sum_{k=1}^3 \sum_{i=1}^9 X_k \theta_{ki} \phi_i(X) + \sigma.$$

(The precise interpretation of this inequality is in time-integrated form). Applying (2.7) and noting that the ϕ_i are linear in u for $i = 1, 2, 3$ leads, after application of the Cauchy-Schwarz inequality, to the bound

$$\frac{d}{dt} \{\mathbb{E}V(u)\} \leq \alpha \mathbb{E}V(u) + \sigma$$

for some $\alpha > 0$ (again to be interpreted in integrated form). Integration of the inequality yields the conclusion of the proposition, by application of the moment bound theory of Itô SDEs explained in [43]. \square

Example 2.2 (Lorenz 96 System [45]). The Lorenz 96 single scale system describes the time evolution of a set of variables $\{x_k\}_{k=1}^K$ according to the equations

$$\begin{aligned} \dot{x}_k &= -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F, \quad k \in \{1, \dots, K\}, \\ x_{k+K} &= x_k. \end{aligned} \tag{2.8}$$

We choose $K = 36$ and use the system in Eq. (2.8) as the true system for a simulation study. We aim at modeling the unknown tendency with first and second-order polynomial

basis functions:

$$\dot{X}_k = \sum_{i=1}^{702} \theta_{ki} \phi_i(X), \quad k = 1, \dots, 36, \quad (2.9)$$

where $\phi = \{\phi_i\}$, $| i \in \{1, \dots, 702\}$ contains all the first ($\phi = \{\phi_i\}$, $| i \in \{1, \dots, 36\}$) and second-order polynomial basis functions.

As in Example 2.1, imposition of energy conserving quadratic nonlinearities is important from a modeling point of view and as a means to ensure well-posedness, i.e., existence of solutions to the equation for all time. To this end, we work with a simpler modeled system, from a subclass of the models (2.9), taking the form

$$\begin{aligned} \dot{X}_k = & -X_{k-1}(\beta_k^{(1)} X_{k-2} - \beta_{k+1}^{(1)} X_{k+1}) - (\beta_k^{(2)} X_{k-1} X_k - \beta_{k+1}^{(2)} X_{k+1}^2) \\ & - (\beta_k^{(3)} X_k X_{k+1} - \beta_{k-1}^{(3)} X_{k-1}^2) - (\beta_k^{(4)} X_{k-1} X_{k+1} - \beta_{k+1}^{(4)} X_{k+1} X_{k+2}) \\ & - \alpha_k X_k + F, \quad k \in \{1, \dots, K\}, \end{aligned} \quad (2.10)$$

$$X_{k+K} = X_k.$$

Thus we only introduce the second-order polynomial basis functions that are constructed by a single variable and its two nearest neighbors, together with a linear diagonal term. This incorporates the energy conservation constraint — the inner-product of the quadratic terms with X is identically zero. The boundary conditions for the unknown parameters in Eq. (2.10) are $\beta_{k+K}^{(i)} = \beta_k^{(i)}$ and $\alpha_{k+K} = \alpha_k$ for $i \in \{1, 2, 3, 4\}$ and $k \in \{1, \dots, K\}$. Therefore, we have 144 unknowns in β and 36 unknowns in α . Our goal is to learn a sparse solution $\{\{\beta_k^{(i)}\}_{i=1}^4, \alpha_k\}_{k=1}^{36}$ which has considerably fewer than 180 non-zero elements. Ideally, of course, in this simulation study setting, the sparse solution will have 72 non-zero elements that agree with the true system in Eq. (2.8).

Proposition 2. Equation (2.10) has unique solution $u = (X_1, \dots, X_K) \in C([0, \infty); \mathbb{R}^K)$. \diamond

Proof. Define the Lyapunov function $V(u) = \frac{1}{2}|u|^2$. Straightforward computation using (2.10) gives

$$\frac{d}{dt} V(u) \leq \alpha V(u) + \beta$$

for some $\alpha, \beta > 0$ after using the fact that (2.9) holds and using Cauchy-Schwarz. Integration of the inequality yields the conclusion of the proposition since, for finite dimensional systems, blow-up in finite time is the only way the solution can cease to exist, and the bound precludes this. \square

In addition, we also consider a situation in which data are generated by the multiscale

Lorenz 96 system:

$$\begin{aligned}
\dot{x}_k &= -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - \frac{hc}{J} \sum_{j=1}^J y_{j,k}, \quad k \in \{1, \dots, K\}, \\
\frac{1}{c} \dot{y}_{j,k} &= -by_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - y_{j,k} + \frac{h}{J} x_k, \quad (j, k) \in \{1, \dots, J\} \times \{1, \dots, K\} \\
x_{k+K} &= x_k, \quad y_{j,k+K} = y_{j,k}, \quad y_{j+J,k} = y_{j,k+1}.
\end{aligned} \tag{2.11}$$

We will take the values $K = 36$ and $J = 10$ and work with parameter values in which the X and Y variables are scale-separated. In this setting, the averaging principle enables elimination of the Y variables from the X equation because they are a function of X . Thus, it is natural to try and fit data from the X variable in the multiscale system Eq. (2.11) to a closed equation in X alone, of the form

$$\begin{aligned}
\dot{X}_k &= -X_{k-1}(\beta_k^{(1)} X_{k-2} - \beta_{k+1}^{(1)} X_{k+1}) - (\beta_k^{(2)} X_{k-1} X_k - \beta_{k+1}^{(2)} X_{k+1}^2) \\
&\quad - (\beta_k^{(3)} X_k X_{k+1} - \beta_{k-1}^{(3)} X_{k-1}^2) - (\beta_k^{(4)} X_{k-1} X_{k+1} - \beta_{k+1}^{(4)} X_{k+1} X_{k+2}) \\
&\quad - \alpha_k X_k + F + g(X_k), \quad k \in \{1, \dots, K\}, \\
X_{k+K} &= X_k.
\end{aligned} \tag{2.12}$$

Comparison with Eq. (2.10) shows that the only difference is an additional function $g(X_k)$. The averaging principle alone does not justify the diagonal and universal form of the closure $g(\cdot)$ but empirical evidence, and arguments based on $J \gg 1$, show that it is a reasonable closure model to employ, an idea developed in [46] and studied further in [29, 47]. We use a hierarchical Gaussian process (GP) with 10 unknowns to parameterize the function $g(X_k)$, as introduced in [29], and learn the GP together with unknown parameters $\{\{\beta_k^{(i)}\}_{i=1}^4, \alpha_k\}_{k=1}^{36}$, based on the data from multiscale Lorenz 96 system in Eq. (2.11). The sparsity constraint is not put on the GP parameters but only on the $\{\{\beta_k^{(i)}\}_{i=1}^4, \alpha_k\}_{k=1}^{36}$.

Example 2.3 (Coalescence Equations). Coagulation and fragmentation equations [48, 49] for systems of particles or droplets may be found in the modeling of a wide range of phenomena arising in science and engineering, for example in cloud microphysics [50, 51], or 3D printing [52]. We consider models in which fragmentation does not occur and refer to the resulting process as one of coalescence. The transport equations in Eq. (2.13) below describe the evolution of the coalescence of particles or droplets, by tracking the evolution of the moments x_k of the mass distribution:

$$\frac{dx_k}{dt} = \frac{1}{2} \int_0^\infty \int_0^\infty \left((m + m')^k - m^k - m'^k \right) \mathcal{C}(m, m') f(m) f(m') dm dm'. \tag{2.13}$$

Here $f(\cdot)$ denotes the mass distribution, and the kernel \mathcal{C} describes the probability of coalescence of two particles or droplets with masses m and m' . We employ the polynomial kernel

\mathcal{C} defined via non-negative weights $\{c_{ab}\}$ and the non-negative integer r :

$$\mathcal{C}(m, m') = \sum_{a,b=0}^r c_{ab} m^a m'^b. \quad (2.14)$$

By substituting Eq. (2.14) into the governing equations (2.13), and truncating to consider only moments $k = 1, \dots, K$, we derive a modeled system to describe the evolution of moments:

$$\begin{aligned} \frac{dX_k}{dt} &= \frac{1}{2} \sum_{a,b=0}^r \sum_{j=1}^{k-1} c_{ab} \binom{k}{j} X_{a+j} X_{b+k-j}, \quad k = 2, 3, \dots, K, \\ \frac{dX_k}{dt} &= \begin{cases} -\frac{1}{2} \sum_{a,b=0}^r c_{ab} X_a X_b & k = 0, \\ 0 & k = 1. \end{cases} \end{aligned} \quad (2.15)$$

It should be noted that Eq. (2.15) is not a closed system: X_ℓ for $\ell \in \{K+1, \dots, K+r-1\}$ are needed in the modeled system. We base the closure model for these higher-order moments on the fitting of a Gamma distribution for $f(\cdot)$. The resulting moment-based coalescence equation with polynomial kernel and Gamma distribution closure is proposed in [53].

Since the mass of the system is X_0 , all integrals should be normalized by this number to have the standard probabilistic interpretation. Then, the mean of this probabilistic distribution is X_1/X_0 and the variance is $X_2/X_0 - (X_1/X_0)^2$. If κ and η are the shape and scale parameters of the Gamma distribution, $\kappa\eta$ is the mean and $\kappa\eta^2$ is the variance. This leads to the following Gamma distribution closure, noting that $\Gamma(n) = (n-1)!$:

$$\begin{aligned} X_k &= X_0 \eta^k \frac{\Gamma(\kappa + k)}{\Gamma(\kappa)}, \quad k > K, \\ \kappa &= \frac{X_1^2}{X_0 X_2 - X_1^2}, \quad \eta = \frac{X_2}{X_1} - \frac{X_1}{X_0}. \end{aligned} \quad (2.16)$$

We study the modeled system in Eq. (2.15), (2.16) with $K = 2$ and $r = 3$; thus, we use the closure to determine the variables X_3, X_4 in terms of the primary moments X_0, X_1 and X_2 . Our goal in this example is to learn a sparse solution of coefficients c_{ab} in Eq. (2.15) based on the data in the following different settings:

- a simulation study where data are generated by and fitted to the model in Eq. (2.15) with $K = 2$ and $r = 3$, and with the Gamma distribution closure in Eq. (2.16);
- data are generated by the model in Eq. (2.15) with $K > 2$ and $r = 3$, and we fit a model for $K = 2$ and $r = 3$, both using the Gamma distribution closure in Eq. (2.16);
- data are generated by the model in Eq. (2.15) with $K = 2$ and $r = 3$ and an exponential closure distribution, and we fit a model for $K = 2$ and $r = 3$, with the Gamma distribution closure in Eq. (2.16).

For the last bullet we note that the exponential distribution closure has the following form:

$$\begin{aligned} X_k &= X_0 \frac{k!}{\mu^k}, \quad k > 2, \\ \mu &= \frac{X_0}{X_1}, \end{aligned} \tag{2.17}$$

where μ denotes the rate parameter of an exponential distribution, chosen to agree with information present in X_0 and X_1 .

To prevent unphysical responses, we constrain the parameters κ and η of the Gamma distribution to prescribed intervals. In so doing, we obtain a closed pair of equations for (X_0, X_2) of the form

$$\begin{aligned} \dot{X}_0 &= -\frac{1}{2} \sum_{a,b=0}^3 c_{ab} X_a X_b, \\ \dot{X}_2 &= \sum_{a,b=0}^3 c_{ab} X_{a+1} X_{b+1}, \end{aligned} \tag{2.18}$$

with $X_1(t) \equiv X_1(0)$ and

$$\begin{aligned} X_k &= X_0 \eta^k \frac{\Gamma(\kappa + k)}{\Gamma(\kappa)}, \quad k = 3, 4, \\ \kappa' &= \frac{X_1^2}{X_0 X_2 - X_1^2}, \\ \eta' &= \frac{X_2}{X_1} - \frac{X_1}{X_0}; \\ \kappa &= \max(\min(\kappa', \kappa_{\max}), \kappa_{\min}), \\ \eta &= \max(\min(\eta', \eta_{\max}), \eta_{\min}). \end{aligned} \tag{2.19}$$

The moment X_1 is a constant which, throughout the simulations in this paper, is set to be 2. Furthermore we take κ_{\min} and κ_{\max} to be 10^{-3} and 10, and η_{\min} and η_{\max} are set to 10^{-3} and 1.

Proposition 3. *Let X_0, X_1, X_2 be non-negative at $t = 0$, assume that $c_{11} = 0$ and that $c_{ab} \geq 0$ for $0 \leq a, b \leq 4$ and that $0 < \kappa_{\min} < \kappa_{\max} < \infty$, $0 < \eta_{\min} < \eta_{\max} < \infty$. Let $T \in (0, \infty]$ be the first time at which X_0 or X_2 becomes zero. Then, the equations (2.18) and (2.19) for $u = (X_0, X_2)$ have a unique solution $u \in C^1([0, T]; \mathbb{R}^2)$. \diamond*

Proof. Recall that X_1 is constant in time. We consider solutions only in the time-interval $[0, T]$. The imposed upper and lower bounds on κ and η ensure that the closure model has the property that there is a universal constant $c \in (0, 1)$ such that, whilst a solution to the equations for X_0, X_2 exists (and necessarily remains non-negative) in $[0, T]$,

$$cX_0 \leq X_3 \leq c^{-1}X_0, \quad cX_0 \leq X_4 \leq c^{-1}X_0.$$

It follows from the equation for X_0 that, for $t \in [0, T]$,

$$0 \leq X_0(t) \leq X_0(0)$$

since all quantities on the right hand-side of the equation for \dot{X}_0 in (2.18) are negative. Now note that the right hand side of the equation for \dot{X}_2 in (2.18) contains no quadratic terms in X_2 , since $c_{11} = 0$, and that the constant term and linear coefficient (with respect to X_2) on the right hand side are, for $t \in [0, T]$, bounded. Multiplying the equation by X_2 shows that $V(X_2) = |X_2|^2$ satisfies

$$\frac{d}{dt}V(X_2) \leq \alpha + \beta V(X_2)$$

with α, β determined by the initial conditions and \mathbf{c} . Hence X_2 cannot blow-up in $[0, T]$ and the result is proved. \square

The total number of unknowns to be learned is thus 9, after imposing symmetry on c_{ab} and setting c_{11} to zero. We also have a positivity constraint on all of the unknowns. Adding a further constraint $c_{22} = 0$ can be used to prevent X_0 from becoming negative, thereby extending the preceding proposition to hold for all $t \geq 0$. In practice, however, we find that the sparse solution always enforces $c_{22} = 0$ and that we do not need to impose it.

Example 2.4 (Kuramoto-Sivashinsky Equation). Let \mathbb{T}^L denote the torus $[0, L]$ and consider the equation

$$\begin{aligned} \partial_t u &= -\partial_x^4 u - \partial_x^2 u - u \partial_x u, \quad x \in \mathbb{T}^L, \\ u|_{t=0} &= u_0. \end{aligned} \tag{2.20}$$

We are interested in learning this model from a library of equations of the form

$$\begin{aligned} \partial_t u &= -\sum_{j=1}^5 \left(\alpha_j \partial_x^j u + \beta_j u^j \partial_x u \right), \quad x \in \mathbb{T}^L, \\ u|_{t=0} &= u_0. \end{aligned} \tag{2.21}$$

Thus, we have 10 unknowns. We wish to constrain the model so that solutions do not blow up in finite time. To do this, we ensure that $\|u\|_{L^2(\mathbb{T}; \mathbb{R})}$ remains bounded. Note that the nonlinear terms disappear when multiplied by u and integrated over \mathbb{T}^L , because of periodicity. On the other hand, the linear term will damp all high spatial frequencies, uniformly in wave-number sufficiently large, provided that $\alpha_4 > 0$. Thus we have:

Proposition 4. *Let $\alpha_4 > 0$. Then there is constant $\mathbf{c} > 0$ such that the solution of equation (2.21), if it exists as a function $u \in C([0, T]; L^2(\mathbb{T}; \mathbb{R}))$, satisfies*

$$\sup_{t \in [0, T]} \|u\|_{L^2(\mathbb{T}; \mathbb{R})}^2 \leq \|u_0\|_{L^2(\mathbb{T}; \mathbb{R})}^2 e^{\mathbf{c}T}. \quad \diamond$$

Proof. Let

$$V(u) = \frac{1}{2} \|u\|_{L^2(\mathbb{T}; \mathbb{R})}^2$$

Using periodicity we note that

$$\int_{\mathbb{T}} u \times \beta_j u^j \partial_x u \, dx = 0$$

and that, for j odd,

$$\int_{\mathbb{T}} u \times \alpha_j \partial_x^j u \, dx = 0.$$

Thus we obtain, provided a solution exists,

$$\frac{d}{dt} V(u) \leq -\alpha_2 \int_{\mathbb{T}} u \times \partial_x^2 u \, dx - \alpha_4 \int_{\mathbb{T}} u \times \partial_x^4 u \, dx.$$

A straightforward calculation based on the spectrum of the operator L defined by $Lu := \alpha_2 \partial_x^2 u + \alpha_4 \partial_x^4 u$ on \mathbb{T} shows that there is constant $c > 0$ such that

$$\frac{d}{dt} V(u) = - \int_{\mathbb{T}} u \times Lu \, dx \leq cV(u),$$

and the proof is complete. \square

In summary we have 10 unknowns, and a positivity constraint on one of the unknowns. Although we ensure that the ℓ_2 -norm of the simulated state remains bounded, it is still possible that the simulated state may blow-up due to numerical discretization. Therefore, we also implement numerical clipping to bound the simulated state at every time step. Details concerning the numerical solution of the K-S equation, including the clipping used, are presented in [Appendix A](#).

3. Algorithms

Recall the inverse problem of interest, encapsulated in (2.3). In practice the data we are given, $y \in \mathbb{R}^J$, is a noisy evaluation of the function $\mathcal{G}(\theta)$. We use the notation $G(\theta)$ to denote this noisy evaluation, and we typically envision the noisy evaluation coming from finite-time averaging. Appealing to central-limit theorem type results, which quantify rates of convergence towards ergodic averages, we assume that $\mathcal{G}(\theta) - G(\theta)$ is Gaussian. Then the inverse problem can be formulated as follows: given $y \in \mathbb{R}^J$, find $\theta \in \Theta$ so that

$$y = G(\theta) + \eta, \quad \eta \sim N(0, \Gamma). \tag{3.1}$$

The natural objective function associated with the inverse problem (3.1) is

$$\frac{1}{2} |\Gamma^{-\frac{1}{2}}(y - G(\theta))|^2. \tag{3.2}$$

We study the use of EKI based methods for solving this inverse problem. The four primary reasons for using EKI to solve the inverse problem are: (a) EKI does not require derivatives, which are difficult to compute in this setting of SDEs; nonetheless, EKI provably, in the linear case [54], and approximately, in the nonlinear case [55] behaves like a gradient descent with respect to objective (3.2), projected into a finite dimensional space defined by the ensemble; (b) EKI is robust to noisy evaluations of the forward map as shown in [56]. (c) EKI is inherently parallelizable and scales well to high-dimensional unknowns [57]. (d) EKI lends itself naturally to the imposition of constraints on the unknown parameter [37]. The primary novelty of the approach proposed in this paper is the demonstration that imposition of sparsity within EKI is something that can be achieved easily and that enables generalization of the approach pioneered in [9] to settings in which the observations are nonlinear and partial and in which the dynamical system comes from an SDE. Such problems lead to the parameter learning of θ from y related by (3.1).

In subsection 3.1, we recap the basic EKI algorithm to fit unknown parameters. In particular, we demonstrate how the iterative algorithm, which is nonlinear, has at its core a quadratic optimization problem. In subsection 3.2, we describe a method of inducing sparsity within the EKI algorithm, by introducing an ℓ_1 constraint and/or an ℓ_0 -type penalty on top of the core optimization task solved by the basic EKI algorithm. In subsection 3.3, we describe how we reformulate the sparsity inducing step of the algorithm as a standard quadratic programming problem.

There are many variants on the precise manner in which sparsity constraints are imposed, and the algorithms used to solve the resulting optimization problems. Our purpose is not to determine the best way to impose sparsity or to the best way to solve the optimization problems: these are well-studied problems and we simply employ some successful approaches to their resolution. Rather, our purpose is to demonstrate that the EKI approach to parameter learning is easily extended to incorporate sparsity constraints because its core is solution of a quadratic optimization problem.

3.1. Ensemble Kalman Inversion

The ensemble Kalman inversion (EKI) algorithm that we employ to solve the inverse problem in (3.1) is described in [32, 37]. First, we introduce a new variable $w = G(\theta)$ and variables v and $\Psi(v)$:

$$\begin{aligned} v &= (\theta, w)^\top, \\ \Psi(v) &= (\theta, G(\theta))^\top. \end{aligned} \tag{3.3}$$

Using these variables we formulate the following noisily observed dynamical system:

$$\begin{aligned} v_{m+1} &= \Psi(v_m) \\ y_{m+1} &= H v_{m+1} + \eta_{m+1}. \end{aligned} \tag{3.4}$$

Here $H = [0, I]$, $H^\perp = [I, 0]$, and hence $Hv = w$, $H^\perp v = \theta$. In this setting, $\{v_m\}$ is the state and $\{y_m\}$ are the data. The objective is to estimate $H^\perp v_m = \theta_m$ from $\{y_\ell\}_{\ell=1}^m$ and to do so iteratively with respect to m . In practice we only have one data point y and not a sequence y_m ; we address this issue in what follows below.

The EKI methodology creates an ensemble $\{v_m^{(j)}\}_{j=1}^J$ defined iteratively in m as follows:

$$\begin{aligned} J_m^{(j)}(v) &:= \frac{1}{2} |y_{m+1}^{(j)} - Hv|_\Gamma^2 + \frac{1}{2} |v - \Psi(v_m^{(j)})|_{C_m^{\Psi\Psi}}^2, \\ v_{m+1}^{(j)} &= \arg \min_v J_m^{(j)}(v). \end{aligned} \quad (3.5)$$

The matrix $C^{\Psi\Psi}$ is the empirical covariance of $\{\Psi(v_m^{(j)})\}_{j=1}^J$. The data $y_{m+1}^{(j)}$ is either fixed so that $y_{m+1}^{(j)} \equiv y$ or created by adding random draws to y from the distribution of the η , independently for all m and j . At each step, m ensemble parameter estimates indexed by $j = 1, \dots, J$ are found from $\theta_m^{(j)} = H^\perp v_m^{(j)}$.

Using the fact that $v = (\theta, w)^T$, the minimizer $v_{m+1}^{(j)}$ in (3.5) decouples to give the update formulae

$$\theta_{m+1}^{(j)} = \theta_m^{(j)} + C_m^{\theta G} (C_m^{GG} + \Gamma)^{-1} (y_{m+1}^{(j)} - G(\theta_m^{(j)})), \quad (3.6)$$

the matrix C_m^{GG} is the empirical covariance of $\{G(\theta_m^{(j)})\}_{j=1}^J$, while matrix $C_m^{\theta G}$ is the empirical cross-covariance of $\{\theta_m^{(j)}\}_{j=1}^J$ with $\{G(\theta_m^{(j)})\}_{j=1}^J$. Details of the derivation may be found in [32, 37]. The algorithm preserves the linear span of the initial ensemble $\{\theta_0^{(j)}\}_{j=1}^J$ for each m and thus operates in a finite dimensional vector space, even if Θ is an infinite dimensional vector space.

3.2. Sparse Ensemble Kalman Inversion (EKI)

We are interested in finding a sparse solution θ of the inverse problem in (3.1), building on the key features (a)–(d) possessed by EKI and outlined in the preamble to this section. To impose sparsity on the solution of θ from EKI, we replace the step (3.5) with the step

$$\begin{aligned} J_m^{(j)}(v, \lambda) &:= \frac{1}{2} |y_{m+1}^{(j)} - Hv|_\Gamma^2 + \frac{1}{2} |v - \Psi(v_m^{(j)})|_{C_m^{\Psi\Psi}}^2 + \lambda |H^\perp v|_{\ell_0}, \\ v_{m+1}^{(j)} &= \arg \min_{v \in \mathfrak{V}} J_m^{(j)}(v), \end{aligned} \quad (3.7)$$

where

$$\mathfrak{V} = \{v : |H^\perp v|_{\ell_1} \leq \gamma\}. \quad (3.8)$$

On occasion we will also impose positivity constraints on some of the parameters and will then choose, for some matrix A ,

$$\mathfrak{V} = \{v : |H^\perp v|_{\ell_1} \leq \gamma, AH^\perp v \geq 0\}. \quad (3.9)$$

The parameters γ and λ may be adjusted and indeed could be learned via cross-validation. To solve the resulting optimization problem, we alternate minimization of (3.7) for $\lambda = 0$, which approximates a gradient descent step for (3.2) subject to an ℓ_1 constraint, with a proximal gradient step on the $|\cdot|_{\ell_0}$ norm, projected into the ℓ_1 constraint set; however, the latter cannot leave either of the constraint sets (3.8) or (3.9), and so reduces to a simple thresholding. To this end, we introduce the function \mathcal{T} on vectors defined by

$$\mathcal{T}(\theta_i) = \begin{cases} 0, & \text{if } |\theta_i| < \sqrt{2\lambda} \\ \theta_i, & \text{otherwise} \end{cases} \quad (3.10)$$

With this definition, we arrive Algorithm 1 below.

Algorithm 1 Sparse EKI algorithm

- 1: Choose $\{\theta_m^{(j)}\}_{j=1}^J$ for $m = 0$
 - 2: $\{w_m^{(j)} = G(\theta_m^{(j)})\}_{j=1}^J$
 - 3: **for** $j = 1, 2, \dots, J$ **do**
 - 4: $v_{m+1}^{(j)} \leftarrow \text{argmin of (3.7) with } \lambda = 0$
 - 5: Extract $\theta_{m+1}^{(j)} = H^\perp v_{m+1}^{(j)}$
 - 6: $\theta_{m+1}^{(j)} = \mathcal{T}(\theta_{m+1}^{(j)})$
 - 7: **end for**
 - 8: $m \leftarrow m + 1$, go to 2
-

We note that taking $\gamma = \infty$ results simply in an ℓ_0 penalty and alternation of standard EKI (which, recall, behaves like a step of gradient descent) with a hard-thresholding algorithm. On the other hand, taking $\lambda = 0$ results in a modification of EKI that promotes smaller ℓ_1 -norm solutions. In the next subsection, we give details about how to formulate the optimization problem Eq. (3.7) with $\lambda = 0$ as a standard quadratic programming problem, rendering the preceding algorithm not only implementable, but efficient.

In practice, the coefficients $\theta^{(j)}$ identified by a single sparsity-promoting optimization, such as Algorithm 1, will exhibit bias. To enhance the performance of identifying the coefficients $\theta^{(j)}$, it is sometimes useful to run the sparse EKI Algorithm 1 in multiple batches, removing unnecessary basis functions in each batch, until the number of basis functions cannot be further reduced. Similar concepts, employing multiple optimizations sequentially, are also advocated in [9, 12]. More specifically, iteratively thresholded least squares optimization is recommended in [9], that is, iteratively solving the least squares optimization on reduced basis functions identified by the optimization in the previous step. On the other hand, a second least squares optimization restricted to the features identified from the original ℓ_1 penalized least squares optimization is recommended in [12]. There is also theoretical work

related to debiasing the output of sparse solution algorithms; see [58]. However our approach is more closely linked to the ad hoc approaches advocated in [9, 12].

3.3. Quadratic Programming with ℓ_1 Penalty

The objective function in (3.7) with $\lambda = 0$ can be rewritten (neglecting constants in v) as, for $C_m = C_m^{\Psi\Psi}$,

$$\frac{1}{2}v^\top (H^\top \Gamma^{-1} H + C_m^{-1}) v - (C_m^{-1} \Psi(v_m^{(j)}) + H^\top \Gamma^{-1} g^{(j)})^\top v. \quad (3.11)$$

We wish to minimize over \mathfrak{V} defined in (3.9) (the case (3.8) may be extracted from what follows simply by setting $A = 0$). By appropriate definition of Q and q , we may write the resulting minimization problem as

$$\begin{aligned} \min_v \quad & \frac{1}{2}v^\top Qv + q^\top v \\ \text{s.t.} \quad & AH^\perp v \geq 0, |H^\perp v|_{\ell_1} \leq \gamma. \end{aligned} \quad (3.12)$$

The following decomposition as described in [6] can be employed to convert (3.12) into the standard form of quadratic programming: introduce variables

$$\begin{aligned} v_i &= v_i^+ - v_i^-, \\ |v_i| &= v_i^+ + v_i^-, \end{aligned} \quad (3.13)$$

where $v_i^+ \geq 0$ and $v_i^- \geq 0$ denote the positive and negative part of the i^{th} element of v , respectively. This decomposition leads to the following minimization problem:

$$\begin{aligned} \min_{v^+, v^-} \quad & \frac{1}{2} (v^+ - v^-)^\top Q (v^+ - v^-) + q^\top (v^+ - v^-) \\ \text{s.t.} \quad & AH^\perp (v^+ - v^-) \geq a, H^\perp (v^+ + v^-) \leq \gamma, v^+ \geq 0, v^- \geq 0. \end{aligned} \quad (3.14)$$

If we define the augmented vector $u^\top = [v^+, v^-] \in \mathbb{R}^{2z}$, we see that the problem takes the form of a standard quadratic programming problem; alternatively one may work with the variable $u^\top = [(v^+ - v^-)^\top, (v^+ + v^-)^\top] \in \mathbb{R}^{2z}$.

Remark 3.1. *It should be noted that the quadratic programming problem arising here by application of the approach pioneered in [6] is a classic well-studied problem; we solve it using a state-of-the-art package [59]. In addition to the approach we adopt to imposing sparsity, there are other methods that solve convex optimization problems subject to ℓ_1 regularization or penalty, such as in (3.7). These include the alternating direction method of multipliers (ADMM) [60] and split Bregman method [61]. Our formulation of the problem in (3.7) as a standard quadratic programming problem has the benefit that other equality and/or inequality constraints are readily imposed on the EKI methodology, along with the ℓ_1 regularization of penalization, as explained in [37].*

Remark 3.2. *In some applications, it may be of interest to impose sparsity only on a subset of the parameters θ . The modification required to do this is straightforward and so we do not detail it here. Such a modification is employed in the next section when we learn a closure model for the Lorenz 96 multiscale equations.*

4. Numerical Results

We demonstrate the capability of the proposed methodology by studying the four examples introduced in Section 2. In all cases, the unknown parameters are detailed in Section 2 and the data used to learn them are detailed in what follows. For the noisy Lorenz 63 system, we use the Euler-Maruyama method to solve the Itô SDEs. For the Lorenz 96 systems, we use an adaptive numerical integrator [62, 63] that automatically chooses between the nonstiff Adams method and the stiff BDF method. For the coalescence equation, we use the fourth-order Runge–Kutta method. The numerical integrator for Kuramoto–Sivashinsky equation is presented in Appendix A. In all cases, the results are initially presented in two figures, one showing the ability of the sparse EKI method to fit the data, and a second showing that the proposed methodology indeed provides a sparse solution in terms of the ℓ_1 norm of redundant coefficients. For the canonical chaotic systems, we then show how well the fitted dynamical system performs in terms of reproducing the invariant measure and time correlation. For the coalescence equation, we show how well the fitted dynamical system performs in terms of reproducing the time trajectories of states with a different initial condition. For the Kuramoto–Sivashinsky equation, we present the results of an additional sparse EKI with reduced basis functions identified by the first sparse EKI. All these numerical studies confirm that the sparsity-promoting EKI is able to discover the governing equations of dynamical systems based on statistics derived from averaging time series.

4.1. Lorenz 63 System

We first study a noisy Lorenz 63 system for which the data are obtained by simulating (2.4), with a given set of parameters $\alpha = 10$, $\rho = 28$, $\beta = 8/3$, and $\sigma = 10$. The goal is to fit a modeled system (2.6) by learning unknown coefficients θ_{ki} and σ . In this study, $\phi = \{\phi_i \mid i \in \{1, \dots, 9\}\}$ contains all the first ($\phi = \{\phi_i \mid i \in \{1, 2, 3\}\}$) and second order polynomial basis functions. It is well known that the existing sparsity-promoting model discovery frameworks such as SINDy [9] would encounter some difficulties for such a system due to noise in the time trajectories. We show that the sparse EKI is able to learn a noisy chaotic system based on statistics derived from averaging time series. Results are presented in Figs. 1 to 4.

The data in this case are finite-time averaged approximations of $\{\mathcal{G}_1(X), \mathcal{G}_2(X)\}$, i.e., first and second moments of simulated states. The time-interval used to gather time-averaged

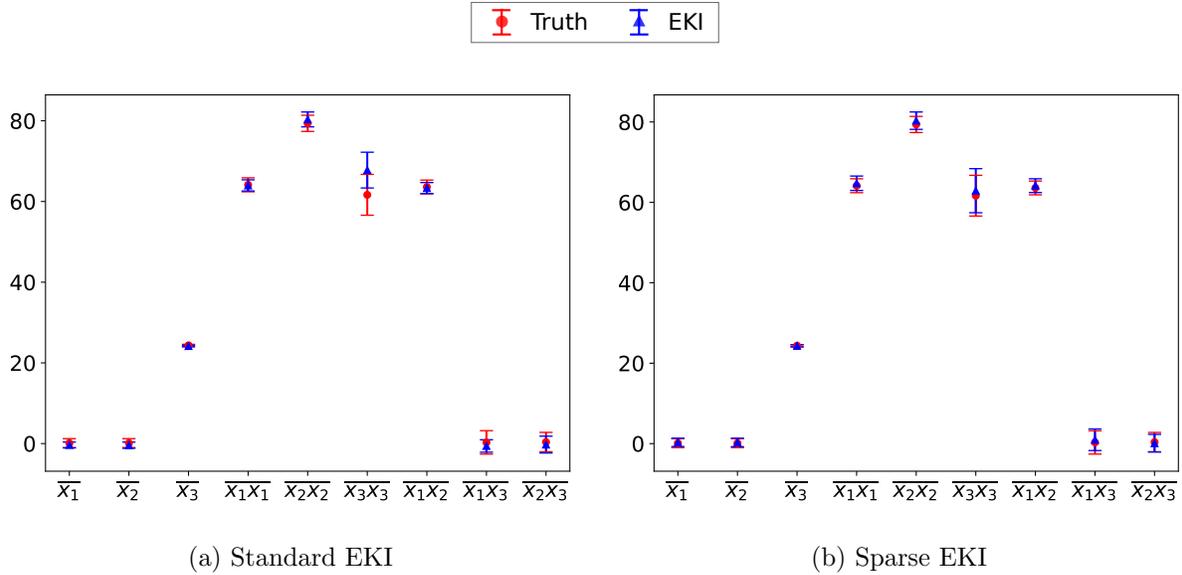


Figure 1: First two moments of state X for noisy Lorenz 63 system found by using (a) standard EKI and (b) sparse EKI.

statistics is $T = 100$. Therefore, we are learning 18 unknown coefficients (17 independent coefficients in $\{\theta_{ki}\}$ and a constant σ), using a data vector y of dimension 9, as shown in Fig. 1. The comparison between the true data and the results of estimated systems in Fig. 1 shows that the sparse EKI has slightly better agreement with the true data than does standard EKI.

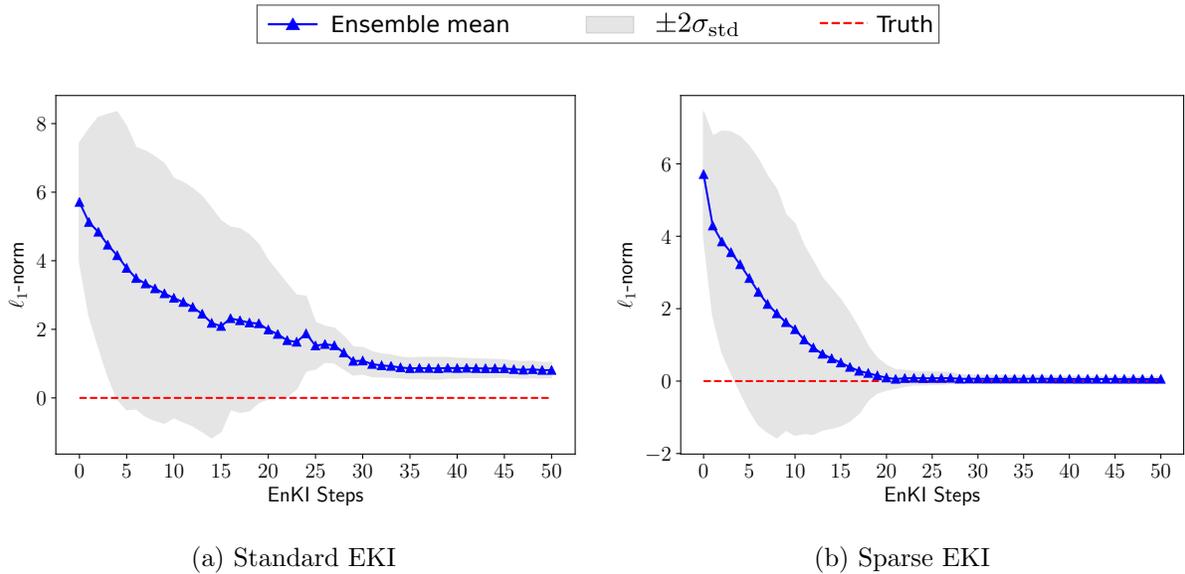


Figure 2: ℓ_1 -norm of redundant coefficients for noisy Lorenz 63 system found by using (a) standard EKI and (b) sparse EKI.

The ℓ_1 -norm of all redundant coefficients also demonstrates the improved performance using sparse EKI, as presented in Fig. 2. Compared to the results of standard EKI, the ℓ_1 -norm of all redundant coefficients is driven much closer to zero using sparse EKI. The coefficients of redundant terms estimated by standard EKI are presented in Table 1, where we can see that there are a few terms being identified with coefficients noticeably larger than zero, such as the linear term X_3 in the equation for X_1 . On the other hand, sparse EKI drives all coefficients of the redundant terms close to zero as shown by Fig. 2b, and the detailed results are omitted here for simplicity.

Table 1: Mean value of coefficients estimated by standard EKI for all redundant terms.

	Redundant terms					
Equation X_1	X_3	X_2X_2	X_3X_3	X_1X_2	X_1X_3	X_2X_3
Coefficient	-0.229	0.026	0.011	-0.062	-0.093	0.094
Equation X_2	X_3	X_1X_1	X_3X_3	X_1X_2	X_2X_3	
Coefficient	0.099	0.062	-0.008	-0.026	-0.001	
Equation X_3	X_1	X_2	X_1X_1	X_2X_2	X_1X_3	X_2X_3
Coefficient	-0.107	-0.066	0.093	0.001	-0.011	0.008

We further investigate the performances of the estimated systems by evaluating the invariant measure. As presented in Fig. 3, the results of sparse EKI show better agreement with the true invariant measure for all three states, confirming the improved performance of the sparse EKI-estimated system over that found from standard EKI, in the long time limit. The comparison of autocorrelation functions is presented in Fig. 4, demonstrating a good agreement of both EKI estimated systems with the true system in terms of time correlation. When comparing results in Fig. 4, the time correlation results of sparse EKI show slightly better performance than the ones of standard EKI, especially for the simulated states X_1 .

Remark 4.1. *For this example we studied in detail the choice of the ℓ_1 penalty parameter γ arising in (3.12) and the thresholding parameter λ in (3.10). As presented in Fig. 5a, smaller γ tends to provide smaller $|\theta|_{\ell_1}$, which indicates a simpler model. The data mismatch dramatically increases when γ is less than around $\gamma = 50$, demonstrating underfitting. On the other hand, both data mismatch and model complexity increase slowly for $\gamma \geq 60$, showing that the performance of the proposed method is not sensitive to the choice of γ provided it is sufficiently large. The comparison of results using different thresholding parameter λ in Fig. 5b shows that the performance of the proposed method is not sensitive to λ . In the example of the noisy Lorenz 63 system we choose $\gamma = 60$ and $\lambda = 0.1$, based on these observations. Since the proposed method is not sensitive to γ outside the underfitting regime and λ , the parametric study is not presented in other examples for simplicity. It would, of*

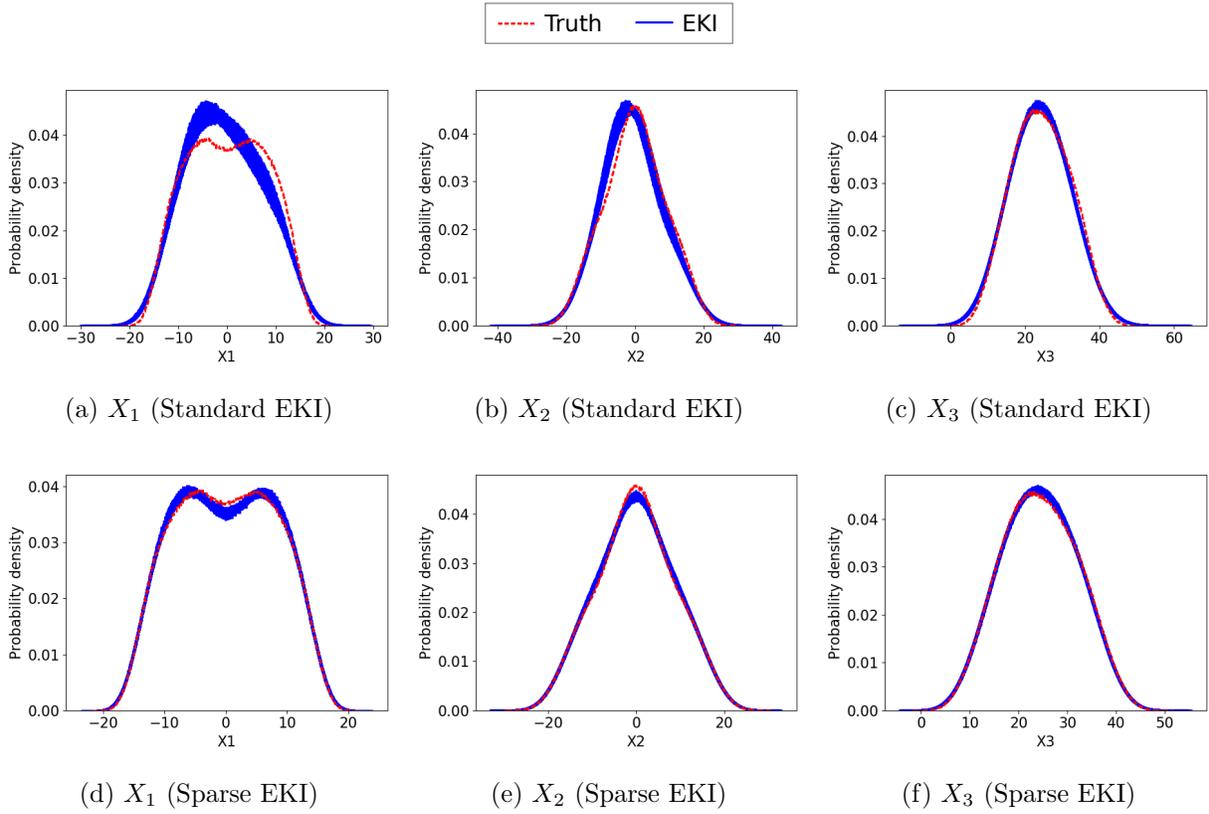


Figure 3: Invariant measure for noisy Lorenz 63 system found by using (a-c) standard EKI and (d-f) sparse EKI.

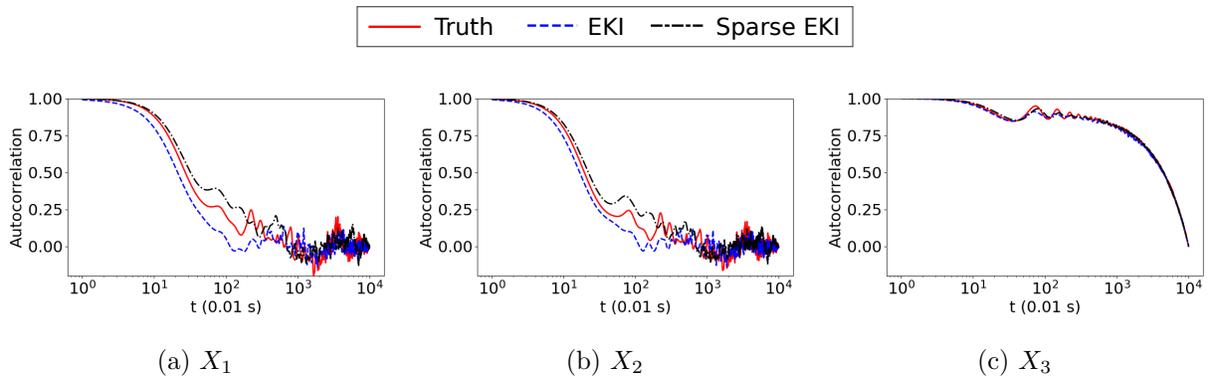


Figure 4: Autocorrelation for noisy Lorenz 63 system found by using standard EKI and sparse EKI.

course, be useful to automate the choice of γ and λ via cross-validation.

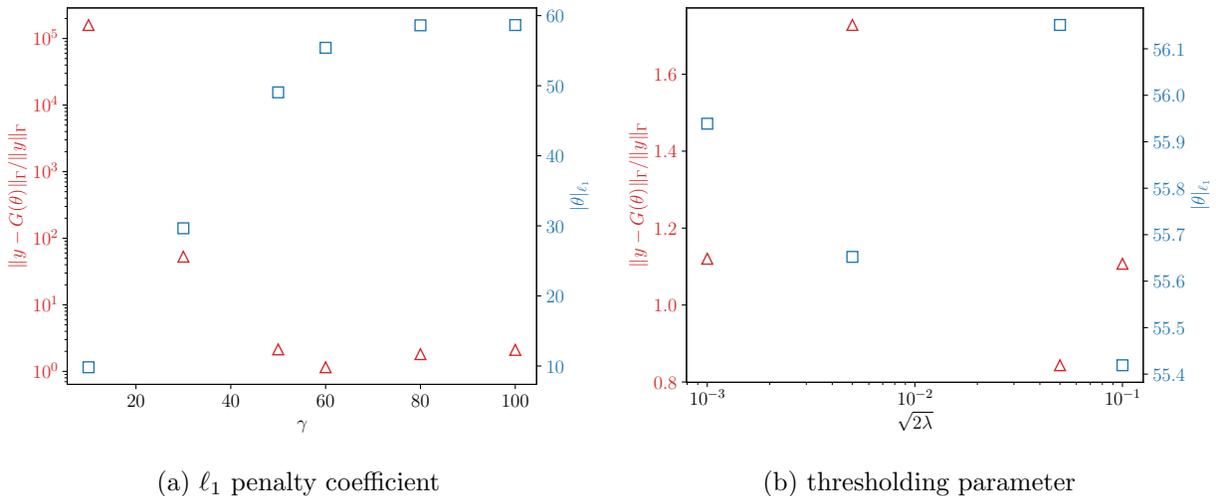


Figure 5: Parametric study of (a) ℓ_1 penalty coefficient γ and (b) thresholding parameter λ . The data mismatch (left axis) is denoted by triangles (Δ), and the model complexity (right axis) is denoted by squares (\square).

4.2. Lorenz 96 System

In this subsection, we study two examples of the Lorenz 96 system: (1) a simulation study for which the true and modeled systems are both single-scale Lorenz 96 systems; (2) a more realistic study for which the true system is a multi-scale Lorenz 96 system and the modeled system only resolves the slow variables.

4.2.1. Simulation Study

For this simulation study, the data are generated from the single-scale Lorenz 96 system in (2.8). The goal is to fit a model as shown in (2.10) by using sparse EKI. Therefore, we are fitting 180 unknown coefficients in total as denoted by $\{\{\beta_k^{(i)}\}_{i=1}^4, \alpha_k\}_{k=1}^{36}$, using a data vector y of dimension 44 (only observing the finite-time average approximation of first and second moments $\{\{\mathcal{G}_1(X), \mathcal{G}_2(X)\}$ for the first 8 state variables). The duration used for time-averaging is $T = 100$. Results are presented in Figs. 6 to 9.

The comparison of EKI results with data from the true system is presented in Fig. 6. This shows that the results of both standard EKI and sparse EKI have good agreement with the true system in data space. However, the ℓ_1 -norm of redundant coefficients presented in Fig. 7 indicates that some redundant coefficients are not close to zero for the system identified by the standard EKI, while most redundant coefficients are driven to zero using sparse EKI.

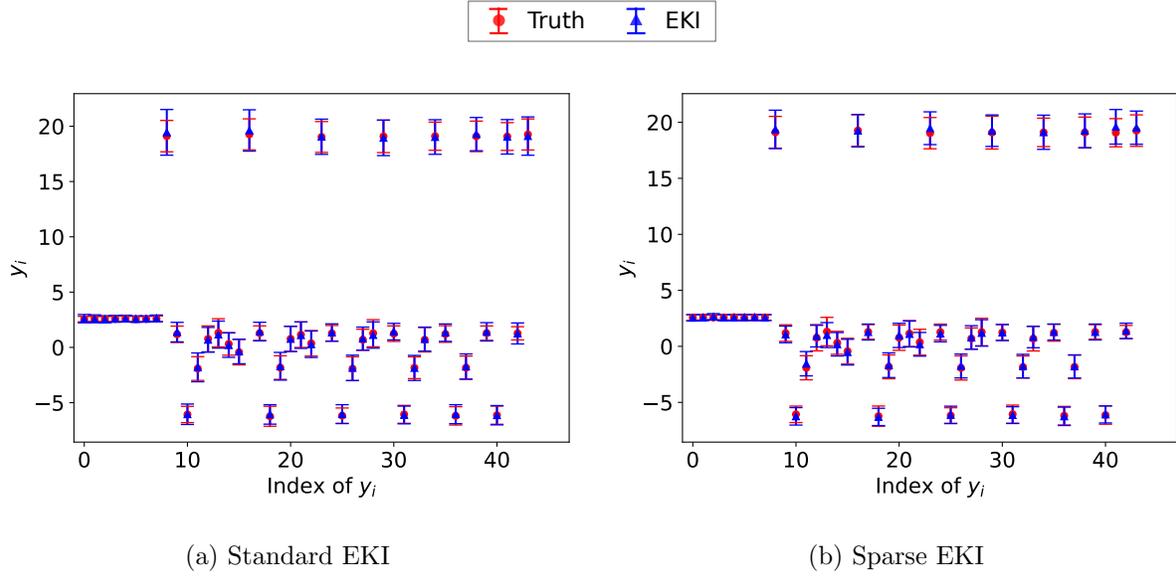


Figure 6: First two moments of state X for single-scale Lorenz 96 system found by using (a) standard EKI and (b) sparse EKI.

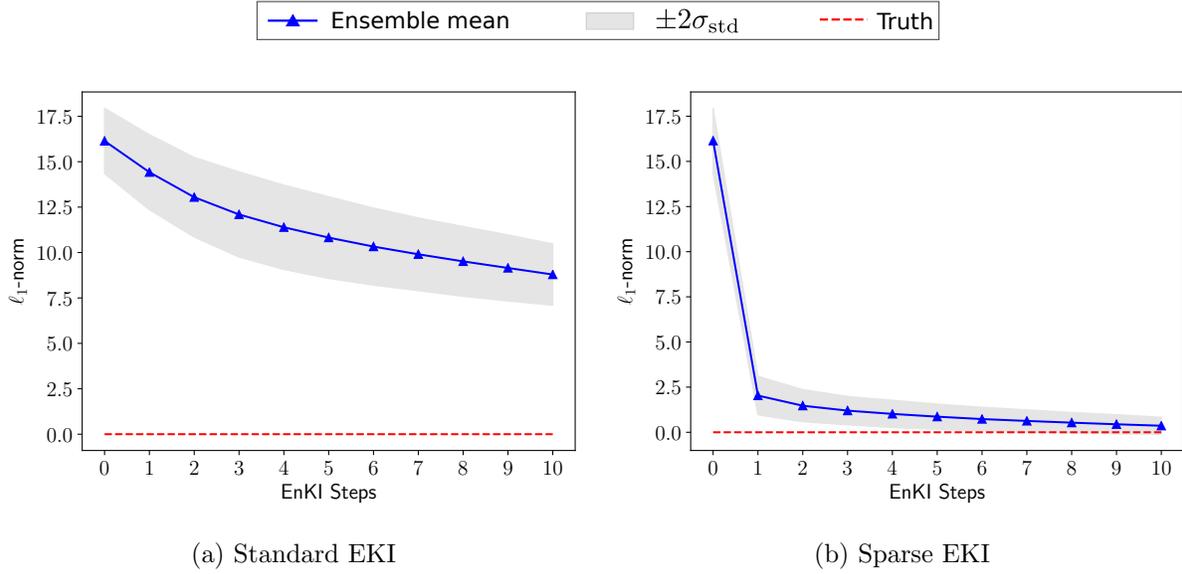


Figure 7: ℓ_1 -norm of redundant coefficients for single-scale Lorenz 96 system found by using (a) standard EKI and (b) sparse EKI.

The long-time limit performance is investigated by evaluating the invariant measure, as presented in Fig. 8. We can see that both the systems identified by standard EKI and sparse EKI show a good agreement with the invariant measure of the true system, while there is slightly greater uncertainty in the invariant measures of the ensemble simulations from standard EKI. As for the invariant measures, the comparison of autocorrelation functions presented in Fig. 9 shows similar performance of the systems identified by standard EKI and sparse EKI, and both systems have a good agreement with the autocorrelation of the true system.

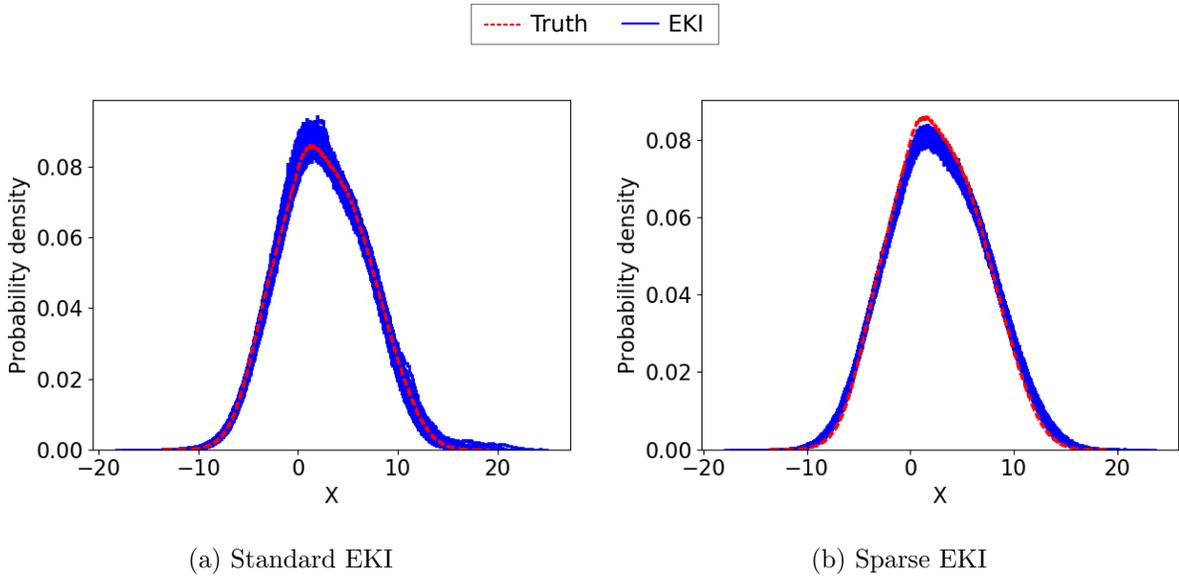


Figure 8: Invariant measure for single-scale Lorenz 96 system found by using (a) standard EKI and (b) sparse EKI.

4.2.2. Multi-scale Data

We now study a more realistic problem for which the data are generated from the multi-scale Lorenz 96 system in (2.11), and the goal is to fit a reduced-order model as shown in (2.12) by using sparse EKI. Therefore, we are fitting 190 unknown coefficients (180 coefficients as denoted by $\{\{\beta_k^{(i)}\}_{i=1}^4, \alpha_k\}_{k=1}^{36}$ and 10 coefficients of GP), using a data vector y of dimension 44 (only observing the finite-time average approximation of first and second moments $\{\{\mathcal{G}_1(X), \mathcal{G}_2(X)\}$ for the first 8 state variables). The time for gathering averaged statistics is $T = 100$. Results are presented in Figs. 10 to 13.

We first present the comparison between EKI results and observation data from the true system in Fig. 10. Although the results of standard EKI show relatively good agreement with the true observation data in Fig. 10a, the results of sparse EKI demonstrate a better agreement with true data in Fig. 10b. The better performance of sparse EKI is also confirmed by the comparison of the ℓ_1 -norm of all redundant coefficients as presented in Fig. 11. The

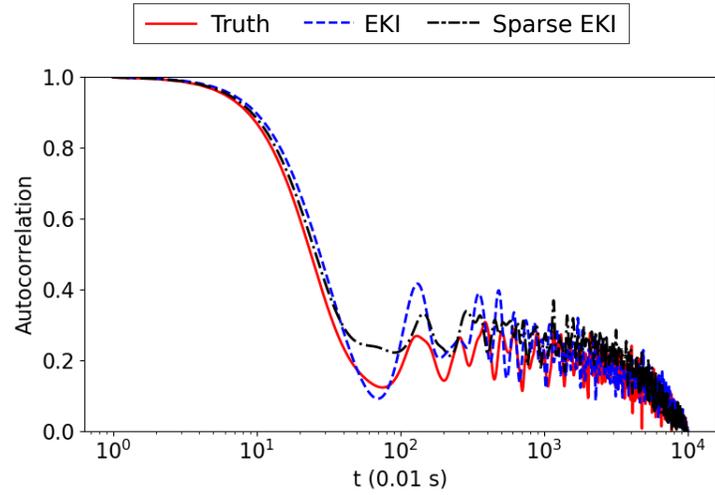
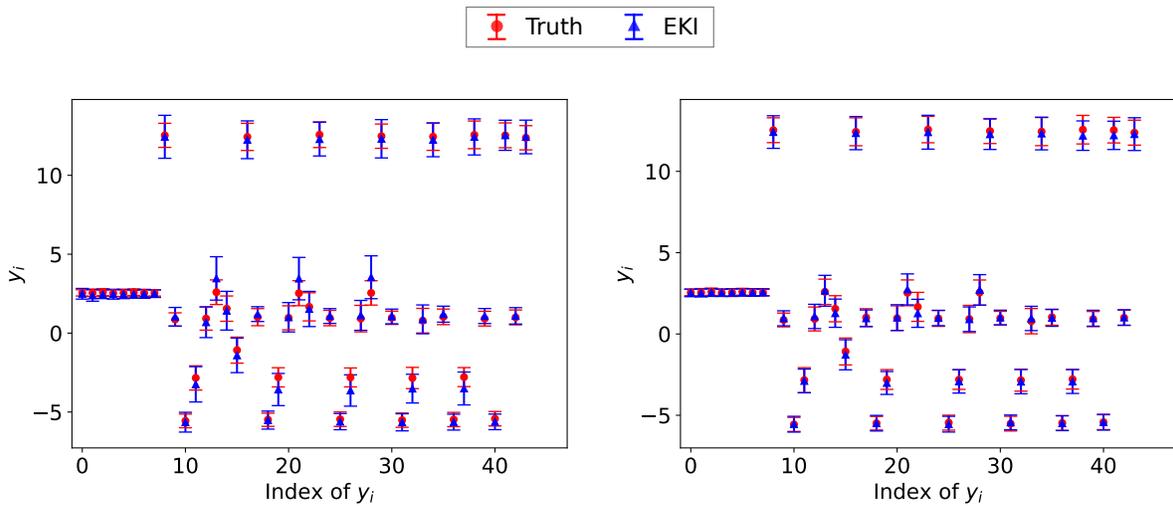


Figure 9: Autocorrelation for single-scale Lorenz 96 system found by using standard EKI and sparse EKI.



(a) Standard EKI

(b) Sparse EKI

Figure 10: First two moments of state X for multi-scale Lorenz 96 system found by using (a) standard EKI and (b) sparse EKI.

comparison in Fig. 11 indicates that most redundant coefficients are successfully driven to zero using sparse EKI, while there are still some non-zero redundant coefficients in the system identified by standard EKI.

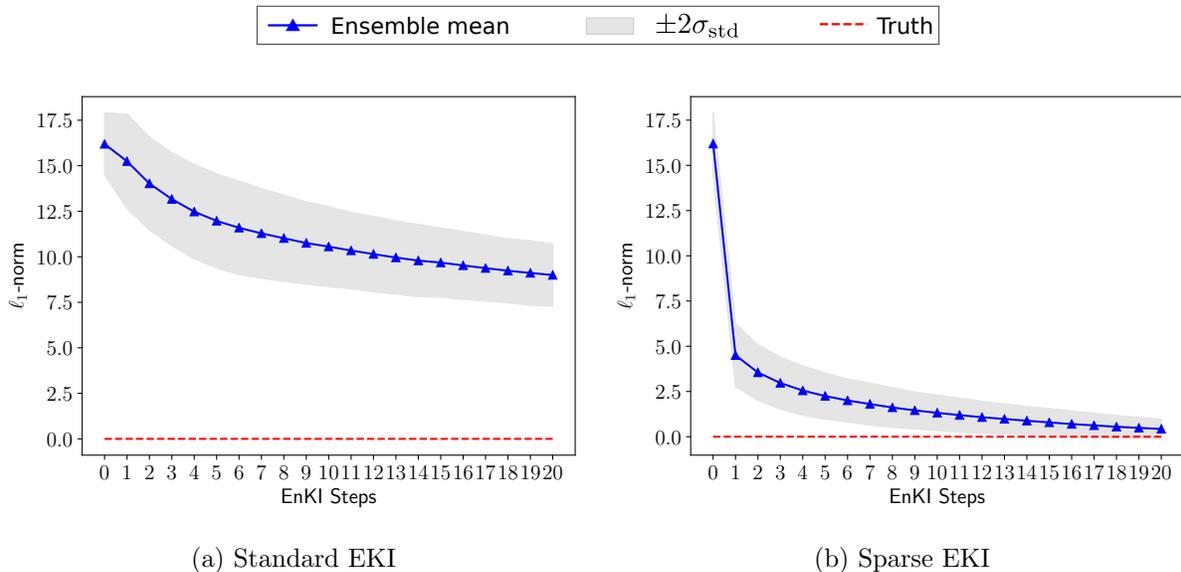


Figure 11: ℓ_1 -norm of redundant coefficients for multi-scale Lorenz 96 system found by using (a) standard EKI and (b) sparse EKI.

The generalization capability of the identified systems is investigated by evaluating the invariant measure. As presented in Fig. 12, the invariant measure of the system identified by sparse EKI shows a much better agreement with the true system, indicating a better performance in the long-time limit. The comparison of the autocorrelation for a chosen ensemble is also studied in Fig. 13. It demonstrates a better agreement with the true system for the system identified by sparse EKI, in terms of capturing the autocorrelation information.

4.3. Coalescence Equations

We further apply the sparse EKI to fit coalescence equations based on statistics derived from time averaging. Specifically, we study three examples: (i) a simulation study where the true system and modeled system share the same closure (Gamma distribution closure) and the same number of resolved states ($K = 2$); (ii) an example where true system and modeled system share the same closure (Gamma distribution closure), while the true system resolves more states ($K = 3$); (iii) an example where true system and modeled system share the same number of resolved states ($K = 2$), while the true system has a different closure (exponential distribution closure). For all three tests of coalescence equations, we impose the symmetry $c_{ab} = c_{ba}$ and thus fit 9 unknown coefficients (recall that we always set $r = 3$, and $c_{11} = 0$), using a data vector of dimension 5 (observing the finite-time average approximation of first

and second moments $\{\{\mathcal{G}_1(X), \mathcal{G}_2(X)\}\}$. The time used for gathering averaged statistics is $T = 50$. Furthermore we impose positivity on all the learned parameters c_{ab} to ensure searching in the space of well-posed models.

4.3.1. Simulation Study

In this simulation study, the data are generated by simulating the coalescence equations in (2.15) with $K = 2$, $r = 3$ and the Gamma distribution closure in (2.16). The goal is to fit a model with the same K , r , and closure by using EKI to estimate unknown coefficients c_{ab} . Results are presented in Figs. 14 to 17.

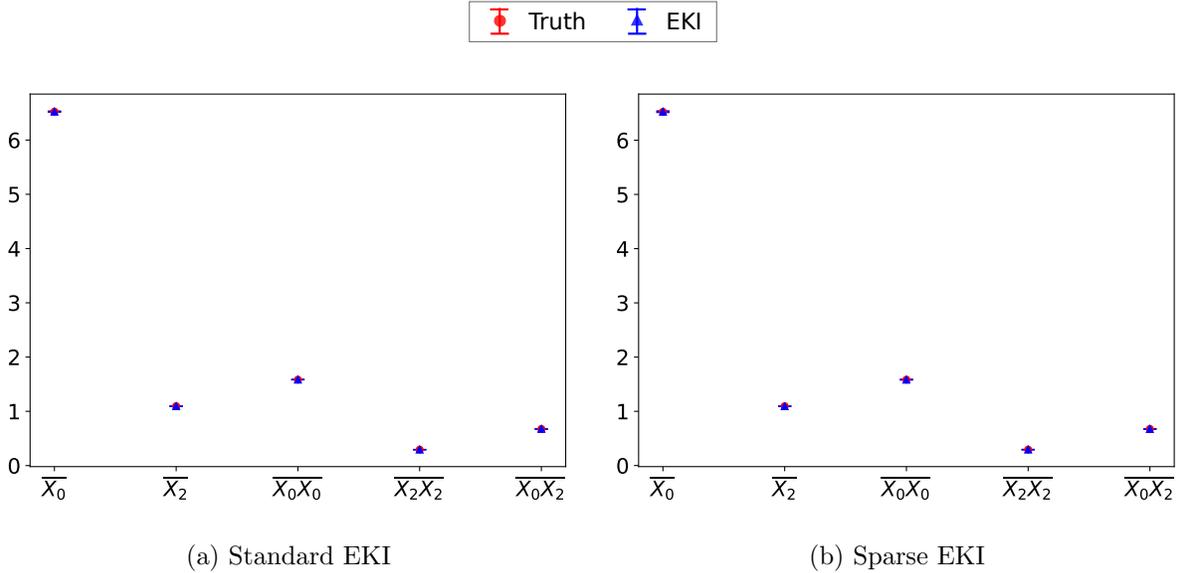


Figure 14: First two moments of state X for coalescence equations found by using (a) standard EKI and (b) sparse EKI.

The comparison of moments data are presented in Fig. 14, which shows that the system identified by using either standard EKI or sparse EKI can have a very good agreement with the true system in terms of matching the first two moments of simulated states. However, it is clear in Fig. 15 that the sets of parameters c_{ab} identified by standard EKI and sparse EKI are quite different. The ℓ_1 -norm of redundant coefficients in Fig. 15a indicates that some of the redundant coefficients are still non-zero for the system identified by standard EKI, while all redundant coefficients are driven to zero as presented in Fig. 15b by using sparse EKI.

The comparison of the non-zero coefficient c_{00} in the true system is presented in Fig. 16. The ensemble mean of the estimated parameter matches with its true value using either standard EKI or sparse EKI, while the result of sparse EKI demonstrates better convergence of the ensemble to the true value.

We further investigate the generalization capability of EKI identified systems by comparing the simulated trajectories of states with an initial condition different from the training

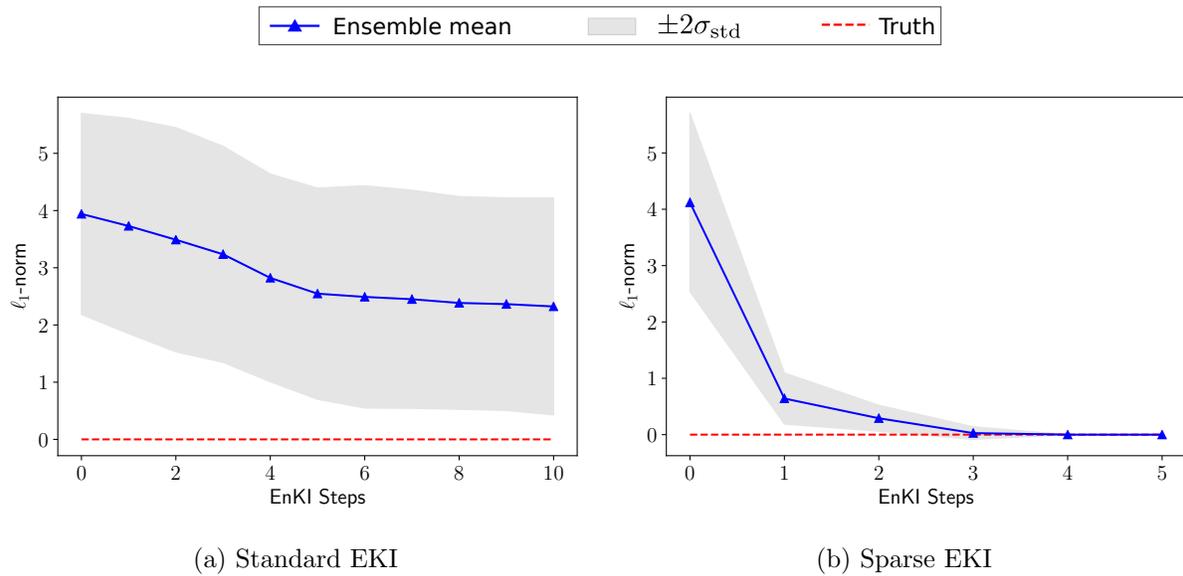


Figure 15: ℓ_1 -norm of redundant coefficients for coalescence equations found by using (a) standard EKI and (b) sparse EKI.

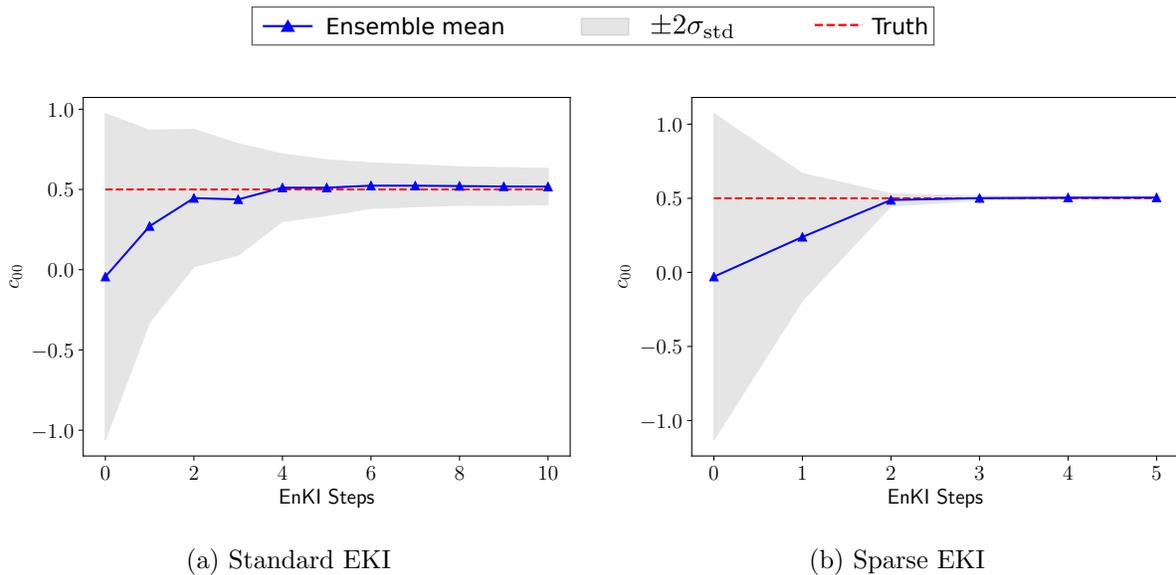


Figure 16: Non-zero parameter for coalescence equations found by using (a) standard EKI and (b) sparse EKI.

set. It can be seen in Fig. 17 that non-zero redundant coefficients and the disagreement of c_{00} among ensemble members do have a negative effect upon the generalization capability of the identified system, as the ensemble of simulated trajectories start to diverge after some time. On the other hand, the system identified by sparse EKI shows a much better agreement with the true trajectories in Fig. 17b, even though the initial condition in this test is different from the one used in the training of the sparse model.

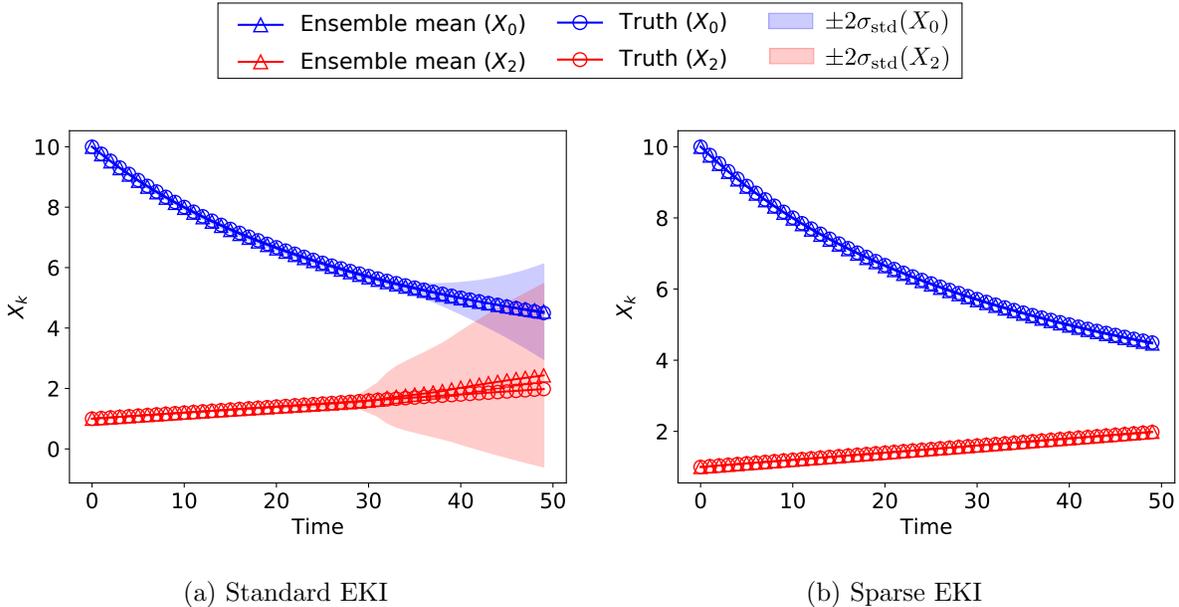


Figure 17: Simulated states for coalescence equations with coefficients found by using (a) standard EKI and (b) sparse EKI. The initial condition of the training dataset is $(X_0, X_1, X_2) = (10, 2, 0.6)$, and the initial condition of the simulations here is $(X_0, X_1, X_2) = (10, 2, 1)$.

4.3.2. Higher-Order Closure Data

We now perform a more realistic study, in which the data are generated by simulating the coalescence equations in (2.15) with $K = 3$, $r = 3$ and the Gamma distribution closure in (2.16). The goal is to fit a model with the same r and closure but different K , namely $K = 2$, using EKI to estimate unknown coefficients c_{ab} . Results are presented in Figs. 18 to 21.

The comparison of data in Fig. 18 shows a comparable performance of the identified system by using either standard EKI or sparse EKI. However, the ℓ_1 -norm of all coefficients is significantly different (as presented in (Fig. 19)). It shows that sparse EKI leads to a set of parameters with smaller ℓ_1 -norm. However, we cannot directly tell whether such a set of parameters is better, since the identified system has a closure distribution different from the one of the true system.

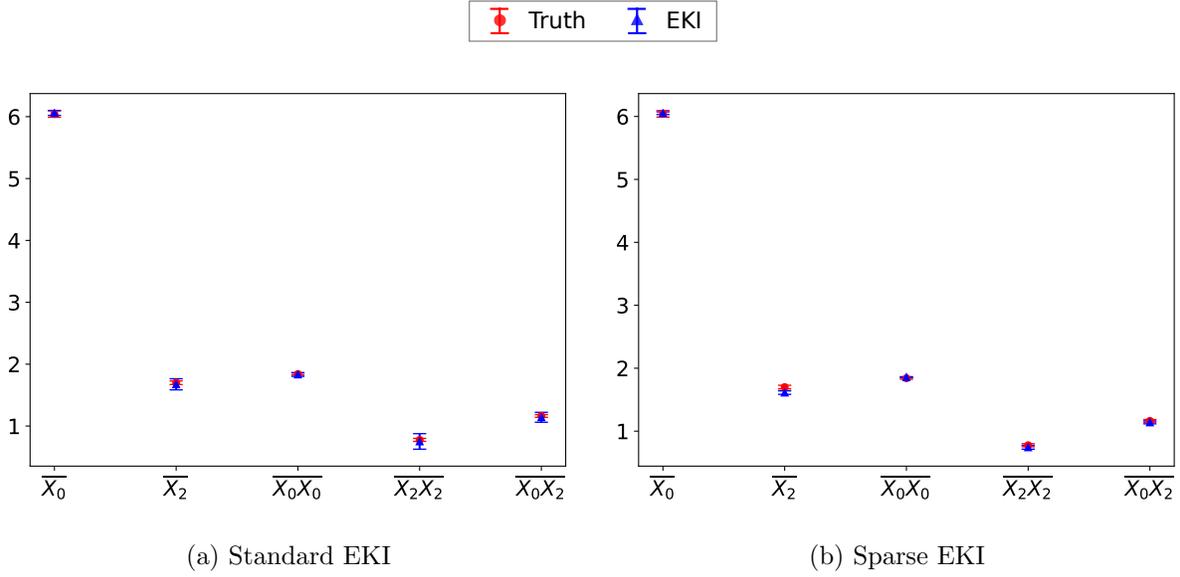


Figure 18: First two moments of state X for coalescence equations found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with a higher-order closure ($K = 3$).

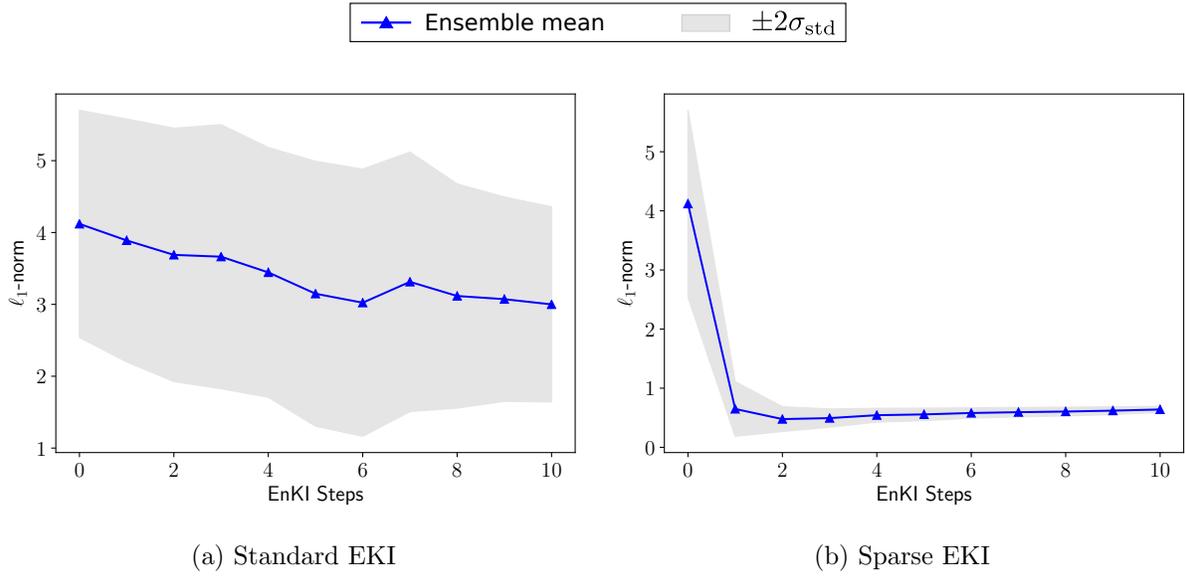


Figure 19: ℓ_1 -norm of all coefficients for coalescence equations found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with a higher order closure ($K = 3$).

Therefore, we investigate the performances of the EKI identified systems by studying the generalization capability in Fig. 20, i.e., simulating identified systems with an initial condition different from the training data. It is clear in Fig. 20 that the simulated trajectories of the system identified by sparse EKI generally have better agreement with the trajectories of the true system.

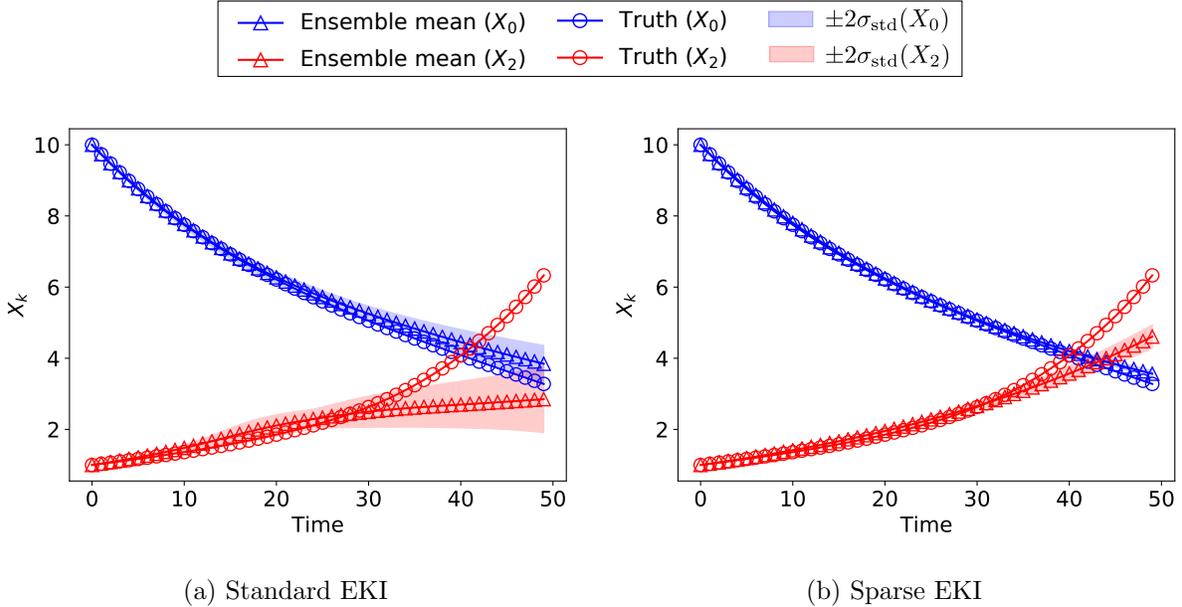


Figure 20: Simulated states for coalescence equations with coefficients found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with a higher-order closure ($K = 3$). The initial condition of the training dataset is $(X_0, X_1, X_2) = (10, 2, 0.6)$, and the initial condition of the simulations here is $(X_0, X_1, X_2) = (10, 2, 1)$.

We further try to improve the performance of the identified system by using training sets with different initial conditions, noting that the coalescence equations are not ergodic, and the initial condition does have an effect on the system prediction. Specifically, we use two training sets with initial conditions $(X_0, X_1, X_2) = (10, 2, 0.6)$ and $(X_0, X_1, X_2) = (10, 2, 2)$, and we then test the performance of the identified systems with a different initial condition $(X_0, X_1, X_2) = (10, 2, 1)$. The results in Fig. 21 show that the identified systems with multiple training sets provide better agreement of simulated trajectories with the true system, and the improvement of performance is more significant for the system identified by standard EKI. This is not surprising since we still fit 9 unknown coefficients here but with twice the data (10 elements in total).

4.3.3. Gamma Versus Exponential Closure Data

We perform a further study where, now, the true system and the modeled system have different closures. Specifically, the data are generated by simulating the coalescence equations

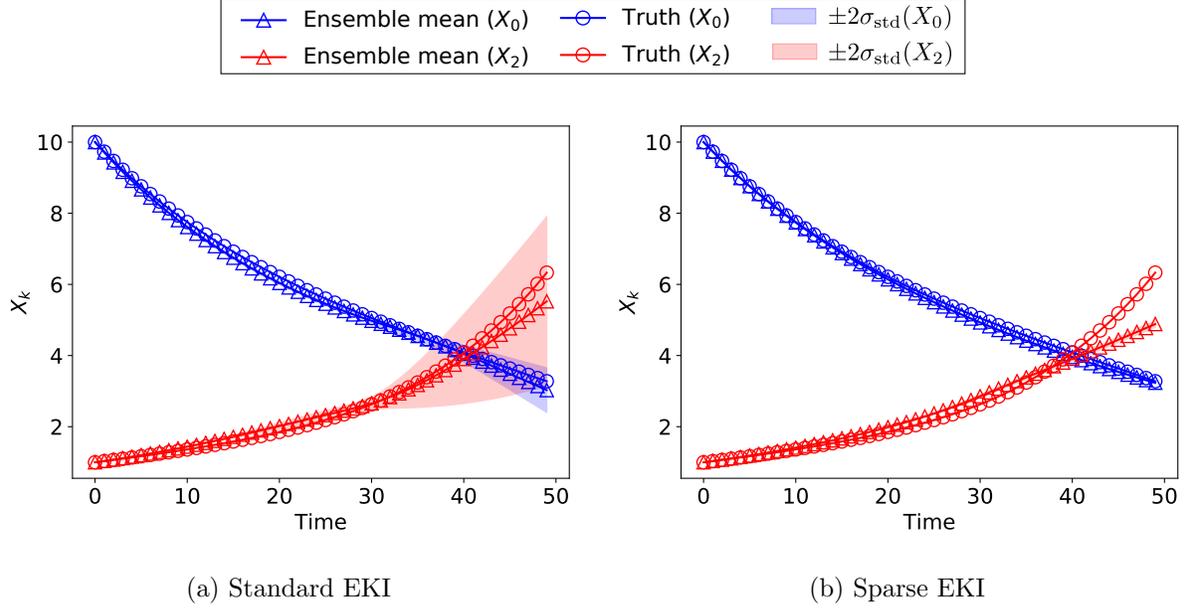


Figure 21: Simulated states for coalescence equations with coefficients found by using (a) standard EKI and (b) sparse EKI. Two datasets are generated by coalescence equations with a higher order closure ($K = 3$). The initial conditions of the training datasets are $(X_0, X_1, X_2) = (10, 2, 0.6)$ and $(X_0, X_1, X_2) = (10, 2, 2)$, and the initial condition of the simulations here is $(X_0, X_1, X_2) = (10, 2, 1)$.

in (2.15) with $K = 2$, $r = 3$, and with the exponential distribution closure in (2.17). The goal is to fit a model with the same K and r but with a Gamma distribution closure, using EKI to estimate unknown coefficients c_{ab} . Results are presented in Figs. 22 to 25.

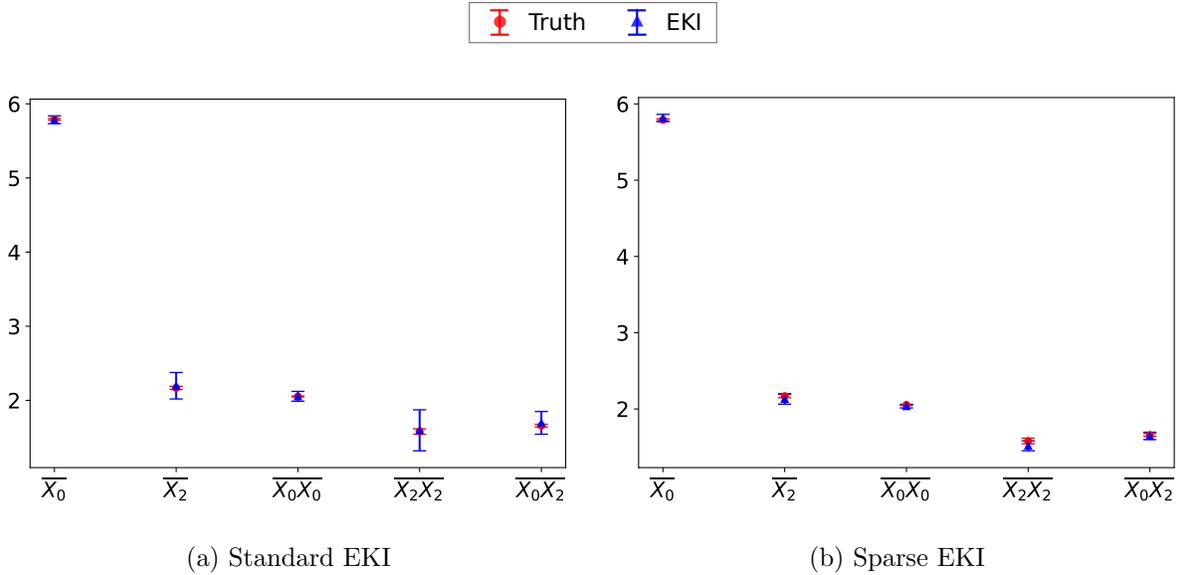


Figure 22: First two moments of state X for coalescence equations found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with an exponential closure.

The comparison of data in Fig. 22 shows larger uncertainties for the results of standard EKI, while the ensemble mean agrees relatively well with the data of the true system. The larger uncertainties can also be seen in Fig. 23: the ℓ_1 -norm of all coefficients estimated using standard EKI remains relatively large, while the sparse EKI identifies another set of parameters with smaller ℓ_1 -norm.

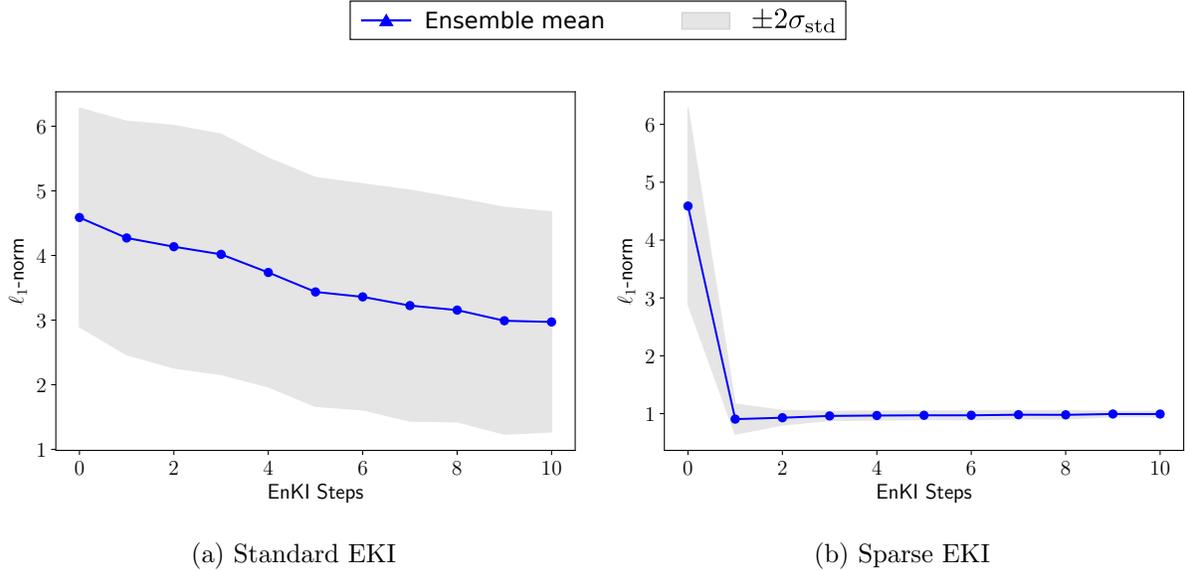


Figure 23: ℓ_1 -norm of all coefficients for coalescence equations found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with an exponential closure.

We then investigate the performance of EKI identified systems by simulating state trajectories with an initial condition different from the training data. The comparison of simulated trajectories is presented in Fig. 24. Although the ensemble mean of either standard EKI or sparse EKI has similar agreement with the trajectories of the true system, there are also larger uncertainties in the ensemble of simulated trajectories for the results of standard EKI.

We further demonstrate that the performance of EKI identified systems can be improved by using multiple training sets. Specifically, two training sets are used with initial conditions $(X_0, X_1, X_2) = (10, 2, 0.6)$ and $(X_0, X_1, X_2) = (20, 2, 0.6)$, and the initial condition of the test presented in Fig. 25 is $(X_0, X_1, X_2) = (15, 2, 0.6)$. Compared to the trajectories of EKI identified systems with a single training set in Fig. 24, the agreement of simulated trajectories with true ones is generally better in Fig. 25 when multiple training set being used.

4.4. Kuramoto-Sivashinsky Equation

We conclude the numerical study by applying the sparse EKI to fit the Kuramoto-Sivashinsky Equation (2.20). We first observe that applying the standard EKI approach to learn the equation, from within the class represented in (2.21) and using the same data

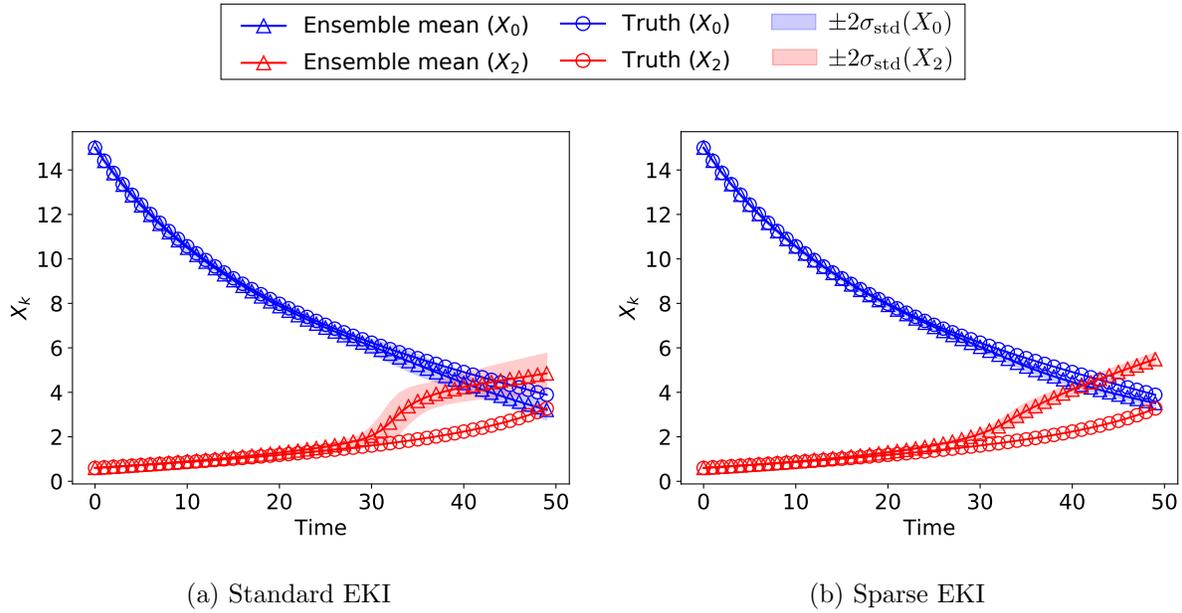


Figure 24: Simulated states for coalescence equations with coefficients found by using (a) standard EKI and (b) sparse EKI. The data are generated by coalescence equations with an exponential closure. The initial condition of the training dataset is $(X_0, X_1, X_2) = (10, 2, 0.6)$, and the initial condition of the simulations here is $(X_0, X_1, X_2) = (15, 2, 0.6)$.

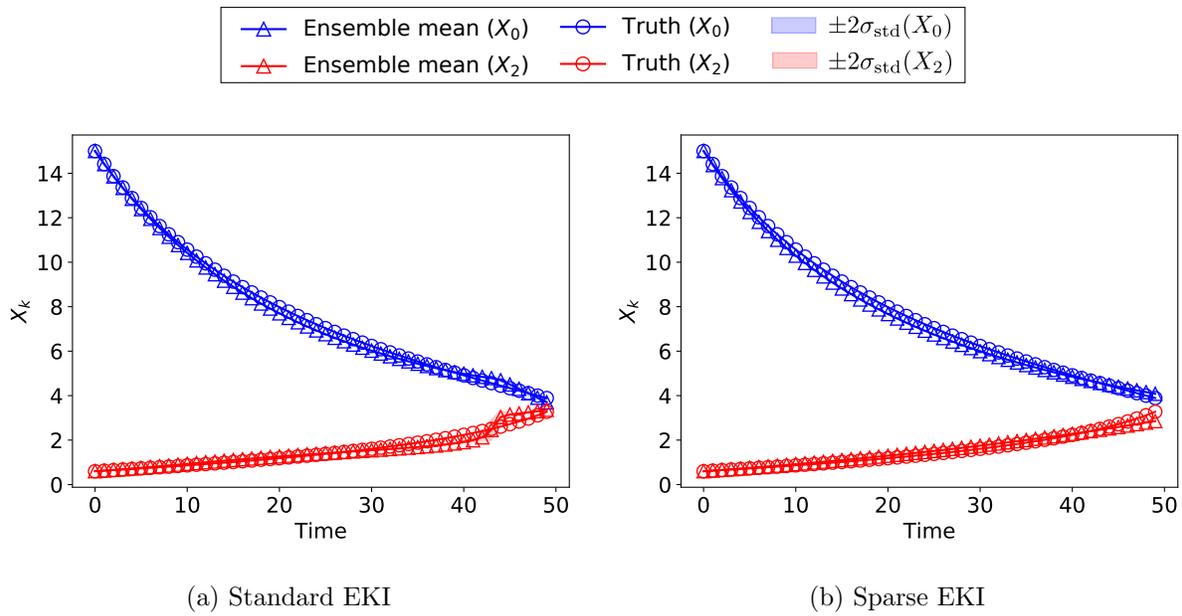


Figure 25: Simulated states for coalescence equations with coefficients found by using (a) standard EKI and (b) sparse EKI. Two datasets are generated by coalescence equations with an exponential closure. The initial conditions of the training datasets are $(X_0, X_1, X_2) = (10, 2, 0.6)$ and $(X_0, X_1, X_2) = (20, 2, 0.6)$, and the initial condition of the simulations here is $(X_0, X_1, X_2) = (15, 2, 0.6)$.

detailed below for application of sparse EKI, leads to a solution as presented in Table 2. It fails to find a solution from within the class (2.21) that is close to the data-generating equation (2.20), clearly motivating the need for the sparse EKI method, results from which are also shown in the same table.

Table 2: Mean value of coefficients estimated by standard EKI.

Linear terms	$\partial_x^1 u$	$\partial_x^2 u$	$\partial_x^3 u$	$\partial_x^4 u$	$\partial_x^5 u$
Coefficient (Standard EKI)	-0.330	1.385	0.659	1.262	-0.130
Coefficient (Sparse EKI)	0	1.020	0	1.020	0
Coefficient (Truth)	0	1	0	1	0
Non-linear terms	$u\partial_x u$	$u^2\partial_x u$	$u^3\partial_x u$	$u^4\partial_x u$	$u^5\partial_x u$
Coefficient (Standard EKI)	1.420	0.224	-0.455	0.104	0.149
Coefficient (Sparse EKI)	1.024	0	0	0	0
Coefficient (Truth)	1	0	0	0	0

We now turn to the sparse setting. Working within the class of models (2.21) requires 10 unknown coefficients $\{\alpha_j, \beta_j\}_{j=1}^5$ to be learnt. To do this we will use a data vector y of dimension 114. Specifically, the data vector consists of: (i) the first to fourth moments at eight locations $\{x_j\}_{j=1}^8$ that are evenly distributed across the range of x , namely $\{\bar{u}_j, \{\bar{u}_j \bar{u}_k\}_{k=1}^8, \bar{u}_j \bar{u}_j \bar{u}_j, \bar{u}_j \bar{u}_j \bar{u}_j \bar{u}_j\}_{j=1}^8$, giving a total moment-data vector of size $8+36+8+8=60$; (ii) temporal autocorrelation of $u(x_j, t)$ at the same eight locations of x and using five points in time, giving a total autocorrelation-data vector of size 40; and (iii) the time-averaged spatial correlation function at 14 locations in space x . The time used for averaging is $T=1000$ and all simulations are performed on the torus $[0, L]$, with $L=128$. Details of the methods employed to solve the extended K-S equation (2.21) are detailed in Appendix A, including the Fourier-based approach to finding the spatial correlation function.

Results of the first sparse EKI (with all ten basis functions) are presented in Figs. 26 and 27, and results of the second sparse EKI (using a reduced number (four) of basis functions, informed by the first phase of the algorithm, using the approach discussed in subsection 3.2), are presented in Figs. 28 and 29.

The comparison of data is presented in Fig. 26 for the first sparse EKI. The comparison of the autocorrelation results at the eight locations is similar, and thus we only present the autocorrelation results at $x=0$. In terms of all three types of data, there are some mismatches between the results of sparse EKI and the true data. In order to evaluate the performance of sparse EKI more precisely, we present the learning of three necessary coefficients (α_2 , α_4 , and β_1) and the ℓ_1 -norm of redundant coefficients in Fig. 27. It is clear that there are some biases in the estimated parameters α_2 and α_4 , while the sparse EKI successfully drives the ℓ_1 -norm of redundant coefficients close to zero.

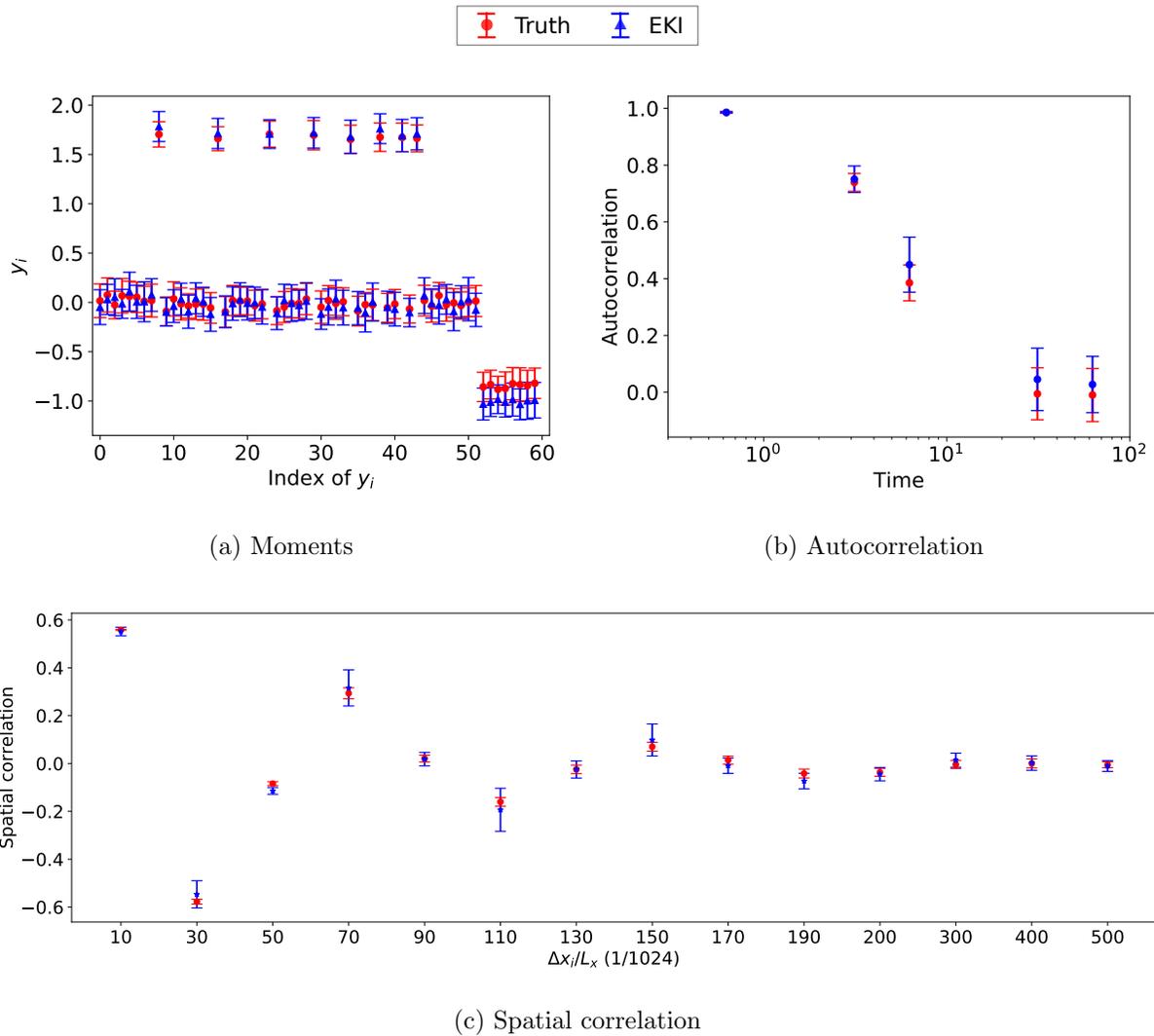


Figure 26: Comparison between the data from the true system and results of the first sparse EKI, including (a) first four moments, (b) autocorrelation at $x = 0$, and (c) time-averaged spatial correlation.

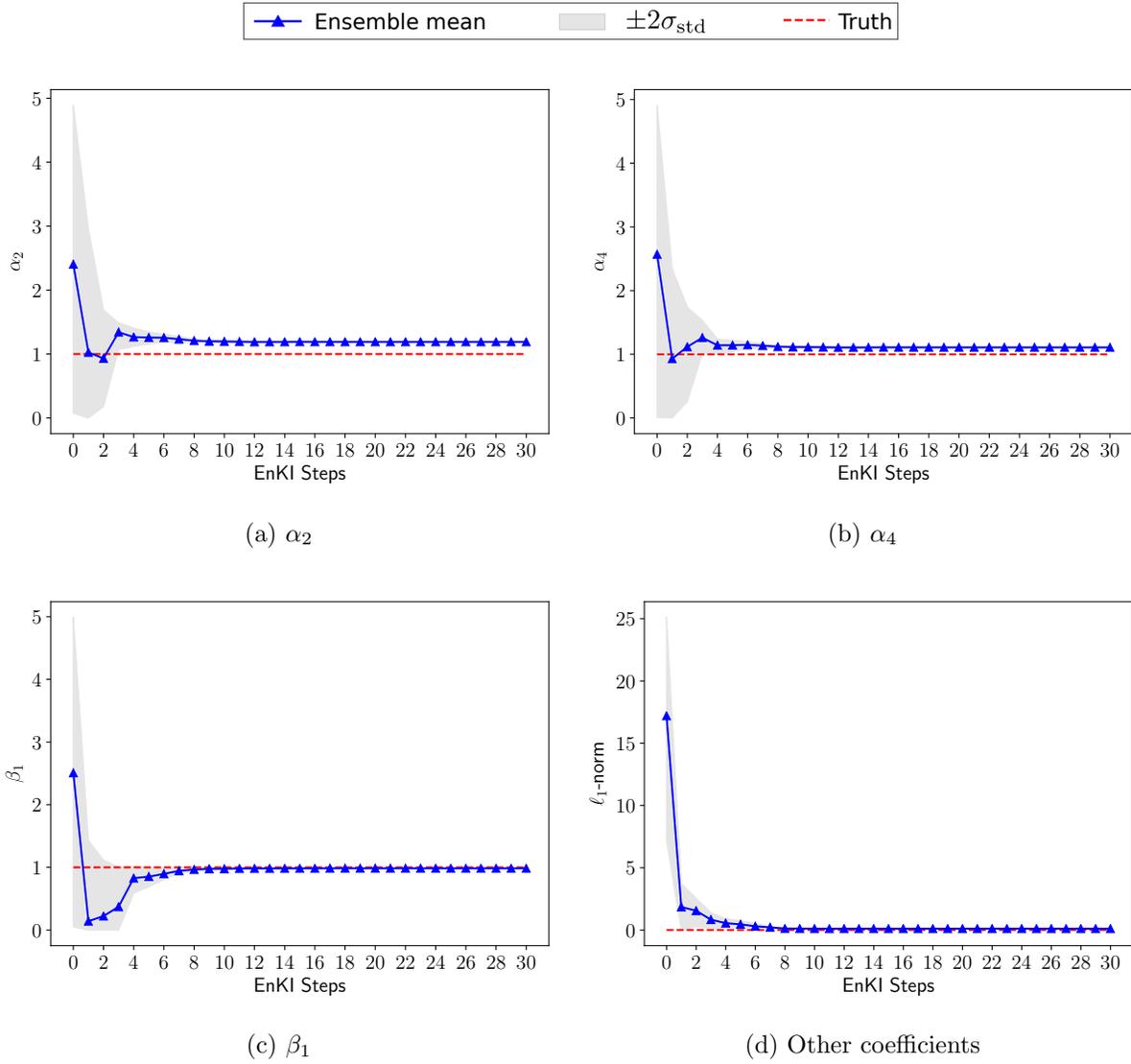


Figure 27: Estimated coefficients from the first EKI, including (a)–(c) necessary coefficients of Kuramoto-Sivashinsky Equation (α_2 , α_4 , and β_1), and (d) the ℓ_1 -norm of other coefficients. There are four nonzero coefficients (α_2 , α_4 , β_1 , and β_3) in the results of first EKI.

In order to improve the accuracy of the estimated parameters in Fig. 27, we perform a second sparse EKI with reduced basis functions, i.e., those with non-zero coefficients (α_2 , α_4 , β_1 , and β_3) in the results obtained in the first application of sparse EKI. The comparison of data is presented in Fig. 28 for the second sparse EKI, which shows a much better agreement with all three types of data. The estimated parameters from the second sparse EKI are presented in Fig. 29, confirming that the Kuramoto-Sivashinsky equation can be accurately identified by using sparse EKI. The results are summarized in Table 2 demonstrating that the sparse EKI method correctly recovers the three non-zero coefficients to an accuracy of less than 2.5% and correctly zeros out all other coefficients; in contrast, the standard EKI finds a non-sparse fit to the data in which all 10 basis coefficients are active. This concluding example demonstrates both the power of sparsity promoting learning of dynamical systems, and the ability of the sparse EKI method to learn dynamical systems from indirect, partial and nonlinear observations.

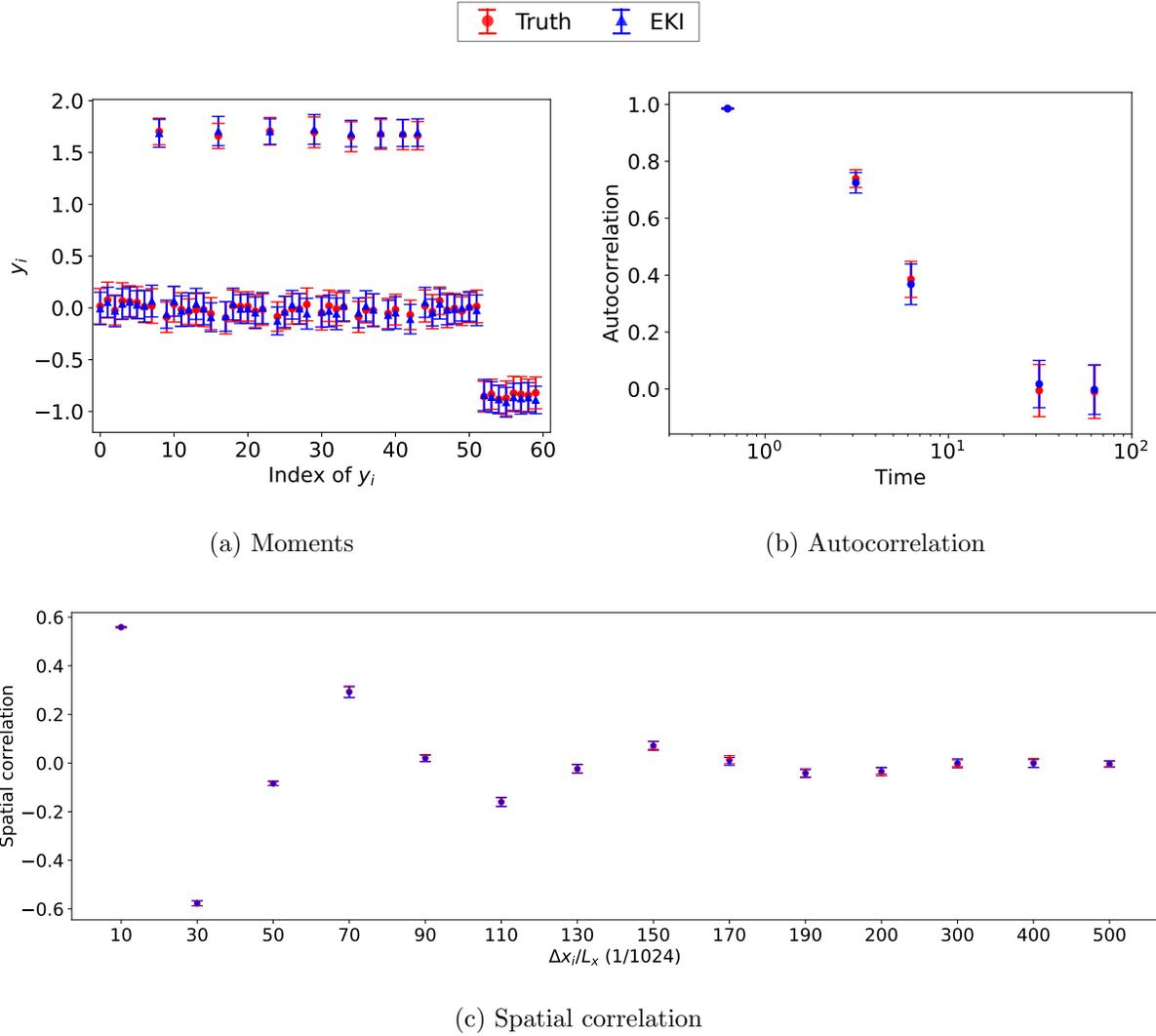


Figure 28: Comparison between the data from the true system and results of the second sparse EKI, including (a) first four moments, (b) autocorrelation at $x = 0$, and (c) time-averaged spatial correlation.

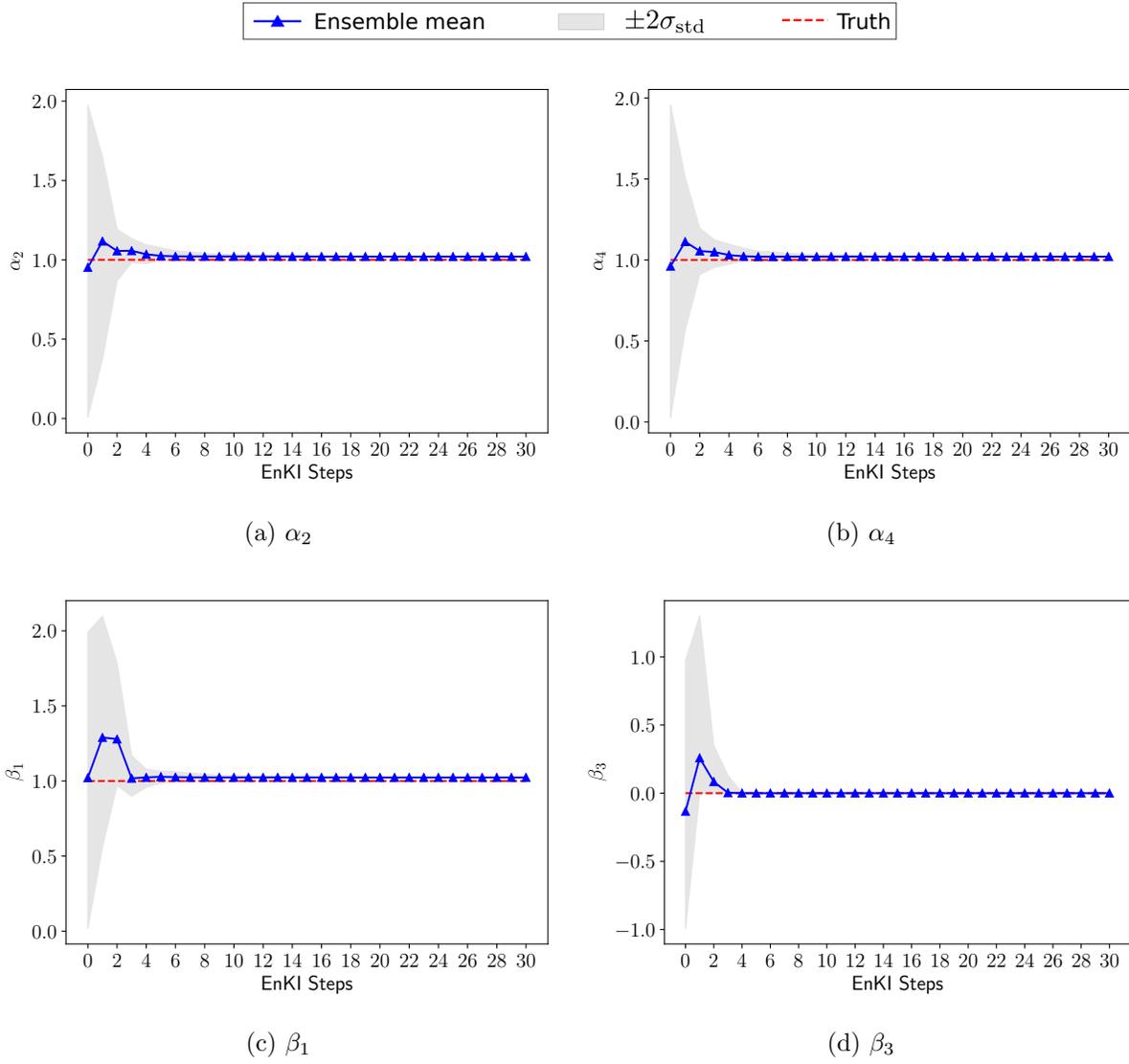


Figure 29: Estimated coefficients from the second sparse EKI, including (a)–(c) necessary coefficients of Kuramoto-Sivashinsky Equation (α_2 , α_4 , and β_1), and (d) the redundant coefficient β_3 .

5. Conclusions

We have demonstrated that sparsity may be naturally incorporated into ensemble Kalman-based inversion methods, leading to the sparse EKI algorithm. The focus of the paper is on learning dynamical models from indirect, partial and nonlinear observations, because the solution of such inverse problems in the sparse setting has been an outstanding challenge in the field. We focus on time-averaged data as a canonical example of such data. The numerical results presented showcase the success of discovering dynamical models in a variety of different examples, demonstrating that sparse learning for model discovery from time-averaged functions of states can be effectively achieved. The proposed sparse learning methodology extends the scope of data-driven discovery of dynamical models from linear observation operators to previously challenging applications where the observation operator is nonlinear. The methodology may in principle be used for the solution of a wide class of nonlinear inverse problems in which sparse solutions are sought. As with existing methods to find sparse solutions of *linear* inverse problems, the core computational task is a quadratic programming problem, subject to linear inequality constraints, and therefore easily implemented. Remarkably, this same computational task allows for solution of *nonlinear* inverse problems when using the proposed sparse EKI method.

Directions for future research stemming from our work include:

- application of the method to other nonlinear inverse problems where sparse learning from a dictionary of functions is valuable;
- development of theory to support the use of the algorithm, noting however that, even in the absence of constraints and imposition of sparsity, the theoretical underpinnings of ensemble inversion methods are only starting to be understood [54, 55];
- detailed study of the use of different algorithms for the imposition of sparsity on the basic quadratic programming task, which is undertaken iteratively during ensemble Kalman inversion;
- careful comparison of ensemble Kalman inversion with other sparsity imposing methods for solving the nonlinear inverse problem of learning dynamical systems from data in time-averaged form.

Acknowledgements We thank Melanie Bieli, Tobias Bischoff and Anna Jaruga for sharing their formulation of the moment-based coalescence equation, and for discussions about it. All authors are supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, by Earthrise Alliance, Mountain Philanthropies, the Paul G.

Allen Family Foundation, and the National Science Foundation (NSF, award AGS1835860). A.M.S. is also supported by NSF (award DMS-1818977) and by the Office of Naval Research (award N00014-17-1-2079).

References

- [1] D. L. Donoho, Compressed sensing, *IEEE Transactions on Information Theory* 52 (4) (2006) 1289–1306.
- [2] E. J. Candes, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59 (8) (2006) 1207–1223.
- [3] E. J. Candes, M. B. Wakin, S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *Journal of Fourier Analysis and Applications* 14 (5-6) (2008) 877–905.
- [4] A. M. Bruckstein, D. L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Review* 51 (1) (2009) 34–81.
- [5] F. Santosa, W. W. Symes, Linear inversion of band-limited reflection seismograms, *SIAM Journal on Scientific and Statistical Computing* 7 (4) (1986) 1307–1330.
- [6] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1) (1996) 267–288.
- [7] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (3) (2011) 273–282.
- [8] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press, 2016.
- [9] S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proceedings of the National Academy of Sciences* 113 (15) (2016) 3932–3937.
- [10] S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations, *Science Advances* 3 (4) (2017) e1602614.
- [11] H. Schaeffer, R. Caflisch, C. D. Hauck, S. Osher, Sparse dynamics for partial differential equations, *Proceedings of the National Academy of Sciences* 110 (17) (2013) 6634–6639.
- [12] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473 (2197) (2017) 20160446.
- [13] G. Tran, R. Ward, Exact recovery of chaotic systems from highly corrupted data, *Multiscale Modeling & Simulation* 15 (3) (2017) 1108–1129.

- [14] H. Schaeffer, G. Tran, R. Ward, Extracting sparse high-dimensional dynamics from limited data, *SIAM Journal on Applied Mathematics* 78 (6) (2018) 3279–3295.
- [15] J. Pereira, M. Ibrahimi, A. Montanari, Learning networks of stochastic differential equations, in: *Advances in Neural Information Processing Systems*, 2010, pp. 172–180.
- [16] L. Boninsegna, F. Nüske, C. Clementi, Sparse learning of stochastic dynamical equations, *The Journal of Chemical Physics* 148 (24) (2018) 241723.
- [17] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D: Nonlinear Phenomena* 60 (1-4) (1992) 259–268.
- [18] R. Chartrand, Numerical differentiation of noisy, nonsmooth data, *ISRN Applied Mathematics* 2011.
- [19] P. Shaman, R. A. Stine, The bias of autoregressive coefficient estimators, *Journal of the American Statistical Association* 83 (403) (1988) 842–848.
- [20] R. Martin, C. Openshaw, Autoregressive modelling in vector spaces: An application to narrow-bandwidth spectral estimation, *Signal processing* 50 (3) (1996) 189–194.
- [21] A. Neumaier, T. Schneider, Estimation of parameters and eigenmodes of multivariate autoregressive models, *ACM Trans. Math. Softw.* 27 (2001) 27–57.
- [22] T. Schneider, A. Neumaier, Algorithm 808: ARfit — A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models, *ACM Trans. Math. Softw.* 27 (2001) 58–65.
- [23] H. Lütkepohl, *Introduction to multiple time series analysis*, Springer Science & Business Media, 2013.
- [24] P. J. Brockwell, R. A. Davis, S. E. Fienberg, *Time series: theory and methods: theory and methods*, Springer Science & Business Media, 1991.
- [25] S. Krumscheid, G. A. Pavliotis, S. Kalliadasis, Semiparametric drift and diffusion estimation for multiscale diffusions, *Multiscale Modeling & Simulation* 11 (2) (2013) 442–473.
- [26] S. Krumscheid, M. Pradas, G. Pavliotis, S. Kalliadasis, Data-driven coarse graining in action: Modeling and prediction of complex systems, *Physical Review E* 92 (4) (2015) 042139.

- [27] S. Kalliadasis, S. Krumscheid, G. A. Pavliotis, A new framework for extracting coarse-grained models from time series with multiscale structure, *Journal of Computational Physics* 296 (2015) 314–328.
- [28] S. Klus, F. Nüske, S. Peitz, J.-H. Niemann, C. Clementi, C. Schütte, Data-driven approximation of the Koopman generator: Model reduction, system identification, and control, *Physica D: Nonlinear Phenomena* 406 (2020) 132416.
- [29] T. Schneider, A. M. Stuart, J.-L. Wu, Learning stochastic closures using ensemble Kalman inversion, arXiv preprint arXiv:2004.08376.
- [30] B. G. Brown, R. W. Katz, A. H. Murphy, Time series models to simulate and forecast wind speed and wind power, *Journal of Climate and Applied Meteorology* 23 (8) (1984) 1184–1195.
- [31] F. Kwasniok, G. Lohmann, Deriving dynamical models from paleoclimatic records: Application to glacial millennial-scale climate variability, *Physical Review E* 80 (6) (2009) 066104.
- [32] M. A. Iglesias, K. J. Law, A. M. Stuart, Ensemble Kalman methods for inverse problems, *Inverse Problems* 29 (4) (2013) 045001.
- [33] Y. Chen, D. Oliver, Ensemble randomized maximum likelihood method as an iterative ensemble smoother, *Mathematical Geosciences* 44 (1) (2002) 1–26.
- [34] A. Emerick, A. Reynolds, Investigation of the sampling performance of ensemble-based methods with a simple reservoir model, *Computational Geosciences* 17 (2) (2013) 325–350.
- [35] M. A. Iglesias, A regularizing iterative ensemble Kalman method for pde-constrained inverse problems, *Inverse Problems* 32 (2) (2016) 025002.
- [36] N. K. Chada, A. M. Stuart, X. T. Tong, Tikhonov regularization within ensemble Kalman inversion, arXiv preprint arXiv:1901.10382.
- [37] D. J. Albers, P.-A. Blancquart, M. E. Levine, E. E. Seylabi, A. M. Stuart, Ensemble Kalman methods with constraints, *Inverse Problems*.
- [38] N. K. Chada, C. Schillings, S. Weissmann, On the incorporation of box-constraints for ensemble Kalman inversion, arXiv preprint arXiv:1908.00696.

- [39] J. Wu, J.-X. Wang, S. C. Shadden, Adding constraints to Bayesian Inverse Problems, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 1666–1673.
- [40] X. Zhang, C. Michelén-Ströfer, H. Xiao, Regularized ensemble kalman methods for inverse problems, *Journal of Computational Physics* (2020) 109517.
- [41] C. A. M. Ströfer, X.-L. Zhang, H. Xiao, O. Coutier-Delgosha, Enforcing boundary conditions on physical fields in bayesian inversion, *Computer Methods in Applied Mechanics and Engineering* 367 (2020) 113097.
- [42] E. N. Lorenz, Deterministic nonperiodic flow, *Journal of the Atmospheric Sciences* 20 (2) (1963) 130–141.
- [43] X. Mao, *Stochastic Differential Equations and Applications*, Elsevier, 2007.
- [44] J. C. Mattingly, A. M. Stuart, D. J. Higham, Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise, *Stochastic processes and their applications* 101 (2) (2002) 185–232.
- [45] E. N. Lorenz, Predictability: A problem partly solved, in: *Proc. Seminar on Predictability*, Vol. 1, 1996.
- [46] I. Fatkullin, E. Vanden-Eijnden, A computational strategy for multiscale systems with applications to lorenz 96 model, *Journal of Computational Physics* 200 (2) (2004) 605–638.
- [47] D. Burov, D. Giannakis, K. Manohar, A. Stuart, Kernel analog forecasting: Multiscale test problems, arXiv preprint arXiv:2005.06623.
- [48] J. M. Ball, J. Carr, O. Penrose, The becker-döring cluster equations: basic properties and asymptotic behaviour of solutions, *Communications in Mathematical Physics* 104 (4) (1986) 657–692.
- [49] J. M. Ball, J. Carr, The discrete coagulation-fragmentation equations: existence, uniqueness, and density conservation, *Journal of Statistical Physics* 61 (1-2) (1990) 203–234.
- [50] H. R. Pruppacher, J. D. Klett, *Microphysics of Clouds and Precipitation*, Springer, 2010.
- [51] A. Seifert, K. D. Beheng, A two-moment cloud microphysics parameterization for mixed-phase clouds. part 1: Model description, *Meteorol. Atmos. Phys.* 92 (2006) 45–66.

- [52] I. Gibson, D. W. Rosen, B. Stucker, et al., Additive manufacturing technologies, Vol. 17, Springer, 2014.
- [53] M. Bieli, T. Bischoff, A. Jaruga, T. Schneider, Cloudy – a new flexible n-moment scheme for warm rain microphysics: Model description and experiments, arXiv preprint arXiv:.
- [54] C. Schillings, A. M. Stuart, Analysis of the ensemble Kalman filter for inverse problems, *SIAM Journal on Numerical Analysis* 55 (3) (2017) 1264–1290.
- [55] A. Garbuno-Inigo, F. Hoffmann, W. Li, A. M. Stuart, Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler, *SIAM Journal on Applied Dynamical Systems* 19 (1) (2020) 412–441.
- [56] A. Duncan, A. M. Stuart, M.-T. Wolfram, In preparation, arXiv preprint arXiv:.
- [57] E. Kalnay, Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, 2003.
- [58] A. Javanmard, A. Montanari, et al., Debiasing the lasso: Optimal sample size for gaussian designs, *The Annals of Statistics* 46 (6A) (2018) 2593–2622.
- [59] M. Andersen, J. Dahl, Z. Liu, L. Vandenberghe, S. Sra, S. Nowozin, S. Wright, Interior-point methods for large-scale cone programming, *Optimization for machine learning* 5583.
- [60] S. Boyd, N. Parikh, E. Chu, Distributed optimization and statistical learning via the alternating direction method of multipliers, Now Publishers Inc, 2011.
- [61] T. Goldstein, S. Osher, The split bregman method for l1-regularized problems, *SIAM journal on imaging sciences* 2 (2) (2009) 323–343.
- [62] L. Petzold, Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations, *SIAM Journal on Scientific and Statistical Computing* 4 (1) (1983) 136–148.
- [63] A. C. Hindmarsh, Odepack, a systematized collection of ode solvers, *Scientific Computing* (1983) 55–64.
- [64] S. M. Cox, P. C. Matthews, Exponential time differencing for stiff systems, *Journal of Computational Physics* 176 (2) (2002) 430–455.

Appendix A. Numerical Solution Of The Extended K-S Equation

We consider the Extended Kuramoto–Sivashinsky (E-K-S) equation on a periodic domain in one dimension:

$$\partial_t u = - \sum_{j=1}^5 \left(\alpha_j \partial_x^j u + \beta_j u^j \partial_x u \right), \quad x \in \mathbb{T}^L, \quad (\text{A.1})$$

$$u|_{t=0} = u_0;$$

here \mathbb{T}^L denotes the torus $[0, L]$. We write the E-K-S equation as

$$\partial_t u = \mathcal{L}u + \mathcal{N}(u), \quad (\text{A.2})$$

where

$$\begin{aligned} \mathcal{L}u &= - \sum_{j=1}^5 \alpha_j \partial_x^j u, \\ \mathcal{N}(u) &= - \sum_{j=1}^5 \frac{\beta_j}{(j+1)} \partial_x (u^{j+1}). \end{aligned} \quad (\text{A.3})$$

Using the Crank-Nicolson/Adams-Bashforth scheme, the above equation can be discretized as

$$\frac{u_{(n+1)} - u_{(n)}}{\Delta t} = \mathcal{L} \frac{u_{(n+1)} + u_{(n)}}{2} + \frac{3}{2} \mathcal{N}(u_{(n)}) - \frac{1}{2} \mathcal{N}(u_{(n-1)}), \quad (\text{A.4})$$

where $u_{(n)} = u(x, n\Delta t)$. We introduce the Fourier transform in the spatial domain:

$$\hat{u}(\xi) = \mathcal{F}(u) = \int_0^L u(x) e^{-2\pi i x \xi} dx. \quad (\text{A.5})$$

Using this notation we obtain the discretization:

$$\frac{\hat{u}_{(n+1)} - \hat{u}_{(n)}}{\Delta t} = \hat{\mathcal{L}} \left(\frac{\hat{u}_{(n+1)} + \hat{u}_{(n)}}{2} \right) + \frac{3}{2} \hat{\mathcal{N}}(\hat{u}_{(n)}) - \frac{1}{2} \hat{\mathcal{N}}(\hat{u}_{(n-1)}), \quad (\text{A.6})$$

where

$$\begin{aligned} \hat{\mathcal{L}}\hat{u} &= - \sum_{j=1}^5 \alpha_j (2\pi i \xi)^j \hat{u}, \\ \hat{\mathcal{N}}(\hat{u}) &= - \sum_{j=1}^5 \frac{2\pi i \xi \beta_j}{(j+1)} \mathcal{F} \left((\mathcal{F}^{-1}(\hat{u}))^{j+1} \right), \end{aligned} \quad (\text{A.7})$$

and where \mathcal{F}^{-1} denotes the inverse Fourier transform. Note that if $\alpha_4 > 0$ then $\lim_{|\xi| \rightarrow \infty} \text{Re}(\hat{\mathcal{L}}) = -\infty$ which makes the equation well-posed. The discretization in (A.6) can be further formulated as

$$\left(I - \frac{\Delta t}{2} \hat{\mathcal{L}} \right) \hat{u}_{(n+1)} = \left(I + \frac{\Delta t}{2} \hat{\mathcal{L}} \right) \hat{u}_{(n)} + \frac{3\Delta t}{2} \hat{\mathcal{N}}(\hat{u}_{(n)}) - \frac{\Delta t}{2} \hat{\mathcal{N}}(\hat{u}_{(n-1)}). \quad (\text{A.8})$$

Alternatively, a standard integrating factor method can be obtained by introducing the Fourier transform in the spatial domain and rewriting the Fourier transform of (A.2) as

$$\partial_t \left(e^{-\hat{\mathcal{L}}t} \hat{u} \right) = e^{-\hat{\mathcal{L}}t} \hat{\mathcal{N}}(\hat{u}). \quad (\text{A.9})$$

This equation can be solved numerically by using the second-order Adams–Bashforth scheme to obtain

$$\hat{u}_{(n+1)} = e^{\hat{\mathcal{L}}\Delta t} \hat{u}_{(n)} + \frac{3\Delta t}{2} e^{\hat{\mathcal{L}}\Delta t} \hat{\mathcal{N}}(\hat{u}_{(n)}) - \frac{\Delta t}{2} e^{2\hat{\mathcal{L}}\Delta t} \hat{\mathcal{N}}(\hat{u}_{(n-1)}). \quad (\text{A.10})$$

The two algorithms (A.8), (A.10) may be found in [64].

In this work, numerical clipping is implemented at every time step to avoid possible blow-up induced by the numerical discretization (A.8) or (A.10):

$$[\hat{u}_{(n+1)}] = \mathcal{F} \left(\max \left(\min \left(\mathcal{F}^{-1}(\hat{u}_{(n+1)}), u_{(n+1)}^{\max} \right), u_{(n+1)}^{\min} \right) \right), \quad (\text{A.11})$$

where $u_{(n+1)}^{\max}$ and $u_{(n+1)}^{\min}$ are upper and lower bounds imposed on the simulated state in the spatial domain. Both algorithms (A.8), (A.10), with clipping, were used, initially, to test the robustness of results to choice of time-stepper; having verified this robustness, all results presented in the paper use algorithm (A.8).

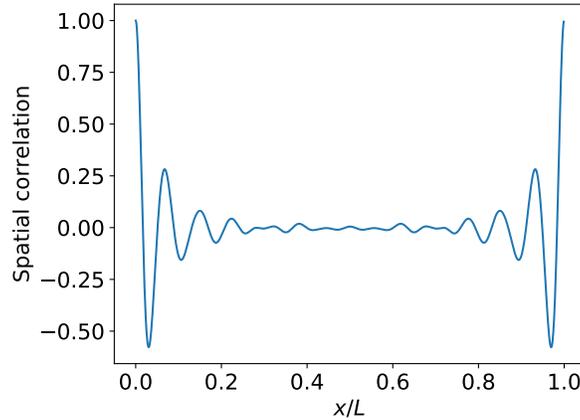


Figure A.30: The time-averaged spatial correlation of the simulated states of K-S equation. Normalization has been performed to set the largest value to 1.

Using algorithm (A.8) we compute the time-averaged spatial correlation function defined by

$$C(x) = \frac{1}{T} \int_0^T \int_0^L u(z, t) u(z + x, t) dz dt. \quad (\text{A.12})$$

Notice that

$$(\mathcal{FC})(\xi) = \frac{1}{T} \int_0^T |(\mathcal{F}u)(\xi, t)|^2 dt$$

facilitating straightforward computation in Fourier space. The function $C(x)$ is shown in Fig. A.30. It is used as part of the definition of G , along with moments and autocorrelation information, from which we learn parameters.