entropy

Article

Can Short and Partial Observations Reduce Model Error and Facilitate Machine Learning Prediction?

Nan Chen

Department of Mathematics, University of Wisconsin-Madison, 480 Lincoln Dr. Madison, Madison, WI 53706, USA; chennan@math.wisc.edu

Received: 23 August 2020; Accepted: 22 September 2020; Published: 24 September 2020

Abstract: Predicting complex nonlinear turbulent dynamical systems is an important and practical topic. However, due to the lack of a complete understanding of nature, the ubiquitous model error may greatly affect the prediction performance. Machine learning algorithms can overcome the model error, but they are often impeded by inadequate and partial observations in predicting nature. In this article, an efficient and dynamically consistent conditional sampling algorithm is developed, which incorporates the conditional path-wise temporal dependence into a two-step forward-backward data assimilation procedure to sample multiple distinct nonlinear time series conditioned on short and partial observations using an imperfect model. The resulting sampled trajectories succeed in reducing the model error and greatly enrich the training data set for machine learning forecasts. For a rich class of nonlinear and non-Gaussian systems, the conditional sampling is carried out by solving a simple stochastic differential equation, which is computationally efficient and accurate. The sampling algorithm is applied to create massive training data of multiscale compressible shallow water flows from highly nonlinear and indirect observations. The resulting machine learning prediction significantly outweighs the imperfect model forecast. The sampling algorithm also facilitates the machine learning forecast of a highly non-Gaussian climate phenomenon using extremely short observations.

Keywords: model error; short and partial observations; data assimilation; conditional sampling; machine learning; non-Gaussian systems

1. Introduction

Predicting complex nonlinear turbulent dynamical systems is an important and practical topic. In many areas including engineering, geophysics, and climate science, dynamical or statistical models are widely used for prediction. However, due to the lack of a complete understanding of nature, the ubiquitous model error may greatly affect the prediction performance. On the other hand, machine learning forecasts based on the properties learned from massive training data possess a large potential to overcome the model error and they have become the predominant forecast methods in quite a few industrial and engineering problems [1–3]. However, the prediction of many climate and geophysical phenomena using machine learning algorithms can still be very challenging. This is because the high-resolution satellites and other refined measurements were not widely developed until recent times. As a result, the available training data are very limited with about only 40 years, which is far from enough to train the machine learning algorithms for predicting interannual or decadal variability. Another fundamental difficulty in utilizing machine learning to predict nature is that only partial observations are available in most applications. In other words, the typical observations involve merely a small subset of the variables and sometimes there is even no direct observations for any variables of interest. Reconstructing the training data of the unobserved variables is quite difficult without the aid of a suitable model.

Since the data in climate science, geophysics, and many other complex nonlinear systems are spatiotemporally correlated and intrinsically chaotic, the traditional data augmentation methods [4–6] for static data are not applicable to expanding the training data set of these problems. On the other hand, thanks to the development of many physics-based dynamical models in describing nature, long correlated time series from these models have been used for training the machine learning algorithms [7,8]. The resulting machine learning prediction can often outperform the simple model forecast. However, such an enhancement of prediction accuracy mainly comes from the improvement of the forecast methodology whereas model error remains in the training data. Thus, incorporating the available short observations into the imperfect model is essential to improving the quality of the training data.

Data assimilation is a well-known technique that combines observations with an imperfect model to improve the statistical state estimation [9–12]. The resulting posterior distribution is typically more informative and accurate than the model's prior distribution. Data assimilation also plays an important role in recovering the states of the unobserved variables via their dynamical coupling with the observed ones in the model.

Despite the idea of combining partial observations with an imperfect model to improve the quality of the training data for machine learning, there remain two key issues to be resolved. First, data assimilation only provides a statistical state estimation at each time instant. The more desired format of training data are the time series that describes the spatiotemporal evolution of the underlying dynamics. In practice, the time series by collecting all the posterior mean estimates is often used as an approximation of the true signal [13]. However, the posterior mean time series can be biased in reproducing even the basic dynamical properties of nature (See Appendix A). Second, since the length of the assimilated states is consistent with that of the short observations, the insufficient training data remain as an unsolved problem.

In this article, an efficient and dynamically consistent conditional sampling algorithm for nonlinear time series is developed that resolves the two issues discussed above. The sampling algorithm starts with a forward and a backward data assimilation procedure to characterize the correct uncertainty of the posterior estimates. Then, the path-wise temporal dependence conditioned on the observations is combined with the point-wise posterior estimates to develop a recursive sampling scheme. For a rich class of nonlinear and non-Gaussian systems, the conditional sampling is carried out by solving a simple stochastic differential equation (SDE), where the short and partial observations serve as the implicit input. This allows an extremely efficient way to sample the nonlinear time series even in high dimensions. One crucial feature of the resulting sampled trajectories is that they outweigh the posterior mean time series in that they succeed in reproducing the exact dynamical and statistical characteristics of nature in the perfect model setup. In the presence of an imperfect model, the model error can be significantly mitigated in the sampled trajectories due to the extra information from observations. What remains is to cope with the short observations. Here, the posterior uncertainties play a vital role in generating multiple distinct sampled trajectories conditioned on the same short observations, which effectively provide a large training set. In addition, this conditional sampling technique allows for sampling the trajectories of the unobserved variables. All these features facilitate the machine learning training and advance the effective prediction.

The rest of the article is organized as follows. The efficient and dynamically consistent conditional sampling algorithm is developed in Section 2. The prediction schemes are described in Section 3. Section 4 aims at improving the prediction of multiscale compressible shallow water flows using indirect observations. Section 5 focuses on advancing the forecast of a highly non-Gaussian climate phenomenon using extremely short observations. The article is concluded in Section 6.

2. The Nonlinear Models and the Conditional Sampling Algorithm

The focus here is on a rich class of high-dimensional nonlinear and non-Gaussian turbulent models, the structure of which allows using closed analytic formulae to develop an efficient and dynamically consistent conditional sampling algorithm. The procedure of deriving such a conditional sampling technique can be extended to general nonlinear systems using particle methods.

2.1. The Nonlinear Modeling Framework

Consider the following class of nonlinear and non-Gaussian systems [14,15],

$$\mathrm{d}\mathbf{X}(t) = \Big[\mathbf{A_0}(\mathbf{X},t) + \mathbf{A_1}(\mathbf{X},t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{B_1}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_1}(t) + \mathbf{B_2}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_2}(t), \tag{1a}$$

$$\mathrm{d}\mathbf{Y}(t) = \Big[\mathbf{a_0}(\mathbf{X},t) + \mathbf{a_1}(\mathbf{X},t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{b_1}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_1}(t) + \mathbf{b_2}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_2}(t), \tag{1b}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are both multi-dimensional state variables. In (1), $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{a_0}$, $\mathbf{a_1}$, $\mathbf{b_1}$, $\mathbf{b_2}$, $\mathbf{B_1}$ and $\mathbf{B_2}$ are vectors and matrices that depend nonlinearly on the state variables $\mathbf{X}$ and time t, while $\mathbf{W_1}$ and $\mathbf{W_2}$ are independent white noise. For nonlinear systems with partial observations, $\mathbf{X}$ can be regarded as the collection of the observed variables while $\mathbf{Y}$ contains the variables that are not directly observed.

The systems in (1) are called conditional Gaussian nonlinear systems due to the fact that conditioned on a given realization of $\mathbf{X}(s)$ for $s \le t$, the distribution of $\mathbf{Y}(t)$ is Gaussian. Despite the conditional Gaussianity, both the joint and marginal distributions can be highly non-Gaussian. Extreme events, intermittency, and complex nonlinear interactions between different variables all appear in such systems. The framework includes many physics-constrained nonlinear stochastic models, large-scale dynamical models in turbulence, fluids, and geophysical flows, as well as stochastically coupled reaction–diffusion models in neuroscience and ecology. A gallery of examples of conditional Gaussian systems in engineering, neuroscience, ecology, fluids, and geophysical flows can be found in [15]. Some well-known dynamical systems that belong to this framework are the noisy versions of various Lorenz systems, a variety of the stochastically coupled FitzHugh–Nagumo model, and the Boussinesq equations with noise.

2.2. The Nonlinear Data Assimilation

Data assimilation or filtering aims at solving the conditional (or posterior) distribution $p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t)$. One advantage of the conditional Gaussian nonlinear systems (1) is that such a conditional distribution can be written down using closed analytic formulae.

Theorem 1 (Nonlinear Optimal Filter [14])**.** *For the conditional Gaussian nonlinear systems* (1)*, given one realization of* $\mathbf{X}(s)$ *for* $s \in [0,t]$*, the conditional distribution* $p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{f}(t), \mathbf{R_f}(t))$ *is Gaussian. The conditional mean* $\boldsymbol{\mu}_\mathbf{f}$ *and the conditional covariance* $\mathbf{R_f}$ *are given by the following explicit formulae:*

$$\mathrm{d}\boldsymbol{\mu}_\mathbf{f} = (\mathbf{a_0} + \mathbf{a_1}\boldsymbol{\mu}_\mathbf{f})\,\mathrm{d}t + (\mathbf{b}\circ\mathbf{B} + \mathbf{R_f}\mathbf{A}_1^*)(\mathbf{B}\circ\mathbf{B}^*)^{-1}(\,\mathrm{d}\mathbf{X} - (\mathbf{A_0} + \mathbf{A_1}\boldsymbol{\mu}_\mathbf{f})\,\mathrm{d}t), \tag{2a}$$

$$\mathrm{d}\mathbf{R_f} = \big(\mathbf{a_1}\mathbf{R_f} + \mathbf{R_f}\mathbf{a}_1^* + \mathbf{b}\circ\mathbf{b} - (\mathbf{b}\circ\mathbf{B} + \mathbf{R_f}\mathbf{A}_1^*)(\mathbf{B}\circ\mathbf{B})^{-1}(\mathbf{B}\circ\mathbf{b} + \mathbf{A_1}\mathbf{R_f})\big)\,\mathrm{d}t, \tag{2b}$$

where $\mathbf{b}\circ\mathbf{b} = \mathbf{b_1}\mathbf{b}_1^* + \mathbf{b_2}\mathbf{b}_2^*$, $\mathbf{B}\circ\mathbf{B} = \mathbf{B_1}\mathbf{B}_1^* + \mathbf{B_2}\mathbf{B}_2^*$ *and* $\mathbf{b}\circ\mathbf{B} = \mathbf{b_1}\mathbf{B}_1^* + \mathbf{b_2}\mathbf{B}_2^*$.

Here, the "optimality" is in the Bayesian sense. From now on, we assume $\mathbf{b}\circ\mathbf{B} = 0$, which is the case in most applications.

2.3. The Optimal Conditional Sampling

The data assimilation exploits the observational information up to the current time instant for improving the initialization of real-time prediction. On the other hand, the optimal offline point-wise statistical state estimation can be carried out by making use of the observational information in the entire training period. This leads to a more accurate state estimation and is achieved by a forward pass of filtering and a backward pass of smoothing.

Theorem 2 (Nonlinear Optimal Smoother [16])**.** *For the conditional Gaussian nonlinear systems* (1)*, given one realization of* $\mathbf{X}(t)$ *for* $t \in [0,T]$*, the optimal smoother estimate* $p(\mathbf{Y}(t)|\mathbf{X}(s), s\in[0,T]) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}(t), \mathbf{R}_\mathbf{s}(t))$ *is conditional Gaussian, where the conditional mean* $\boldsymbol{\mu}_\mathbf{s}(t)$ *and the conditional covariance* $\mathbf{R}_\mathbf{s}(t)$ *satisfy the following equations:*

$$\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}} = \big(-\mathbf{a_0} - \mathbf{a_1}\boldsymbol{\mu}_\mathbf{s} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}(\boldsymbol{\mu}_\mathbf{f} - \boldsymbol{\mu}_\mathbf{s})\big)\,\mathrm{d}t, \tag{3a}$$

$$\overleftarrow{\mathrm{d}\mathbf{R}_\mathbf{s}} = -\big((\mathbf{a_1} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1})\mathbf{R}_\mathbf{s} + \mathbf{R}_\mathbf{s}(\mathbf{a}_1^* + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}) - \mathbf{b}\circ\mathbf{b}\big)\,\mathrm{d}t. \tag{3b}$$

The backward notations on the left-hand side of (3) *are understood as* $\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}} = \lim_{\Delta t\to 0}\boldsymbol{\mu}_\mathbf{s}(t) - \boldsymbol{\mu}_\mathbf{s}(t+\Delta t)$*. The starting value of the nonlinear smoother* $(\boldsymbol{\mu}_\mathbf{s}(T), \mathbf{R}_\mathbf{s}(T))$ *is the same as the filter estimate at the endpoint* $(\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R}_\mathbf{f}(T))$*.*

The nonlinear smoother provides the optimal statistical state estimate at each time instant t. However, merely using the information from the smoother posterior distributions (3) is not sufficient to draw unbiased model trajectories conditioned on the observations. This is because, in addition to the point-wise statistical state estimates (namely at each fixed time instant) from the smoother (3), the conditional temporal dependence at nearby time instants also needs to be taken into account in creating these conditional sampled trajectories. In practice, the posterior mean time series is often used as a surrogate of the recovered model trajectory conditioned on the observations. However, the posterior uncertainty and its temporal correlation are completely ignored in such an approximation. The consequence is that the posterior mean time series fails to capture many key dynamical and statistical features of the underlying dynamics, such as the temporal autocorrelation function (ACF) and the PDF (See Appendix A). Note that a naive way of involving the posterior uncertainty in sampling hidden trajectories conditioned on the observations is to draw independent random numbers from the posterior distributions at different time instants and then connect them together. However, such an approach not only leads to a noisy time series but fails to capture the underlying dynamical features as well due to the lack of incorporating the temporal correlation of the uncertainty.

The result presented below exploits both the conditional path-wise temporal dependence and the point-wise optimal state estimate from the nonlinear smoother (3) to build an optimal conditional sampling algorithm. For the nonlinear systems (1), the conditional sampling can be carried out in an extremely efficient way by solving a simple SDE.

Theorem 3 (Conditional Sampling Formula)**.** *For the conditional Gaussian nonlinear systems* (1)*, conditioned on one realization of* $\mathbf{X}(t)$ *for* $t \in [0,T]$*, the optimal conditional sampling formula of the trajectories of* $\mathbf{Y}(t)$ *in the same time interval satisfies the following explicit formula:*

$$\overleftarrow{\mathrm{d}\mathbf{Y}} = \overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}} - \big(\mathbf{a_1} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}\big)(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{s})\,\mathrm{d}t + \big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_\mathbf{Y}. \tag{4}$$

where $\mathbf{W}_\mathbf{Y}$ *is an independent white noise source and the square root of the positive definite matrix* $\mathbf{b}\circ\mathbf{b}$ *is unique. The initial value of* $\mathbf{Y}$ *is drawn from* $\mathcal{N}(\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R}_\mathbf{f}(T))$*.*

See Appendix D for the proof of this Theorem. The mathematical conciseness of (4) indicates that the sampled trajectories meander around the smoother mean time series. Meanwhile, the second term on the right-hand side indicates a correlated uncertainty in the sampled trajectories. Such a temporal correlated uncertainty plays a crucial role for the sampled trajectories to perfectly reproduce the path-wise and statistical features of nature in the absence of model error, which is however lacked in the posterior mean time series. The uncertainty also facilitates the sampling of multiple distinct trajectories conditioned on the same short observational time series, which effectively provide a large training set for the machine learning algorithms. Notably, in the presence of an imperfect model, the model error can be significantly mitigated in the sampled trajectories due to the extra information from observations. The machine learning prediction based on these sampled trajectories is thus expected to outperform the model forecast using the imperfect model.

Thanks to the explicit formula in (4), the conditional sampling procedure is computationally much cheaper than the traditional particle methods. In Appendix A, an alternative conditional sampling formula is included. It requires only the filter estimate, which further reduces the computational cost. In addition, a block decomposition technique [17] can be adopted here for efficiently sampling many high-dimensional systems.

Finally, Figure 1 shows a schematic illustration of the entire procedure. Here, the nonlinear smoother is crucial for the state estimation. This is because the smoother improves the filter estimates by exploiting the entire observational information and leads to an unbiased conditional state estimation. Utilizing the point-wise state estimates and the temporal dependence to derive the formula of the path-wise conditional sampling is another key step that fundamentally differs from the state estimation using data assimilation.

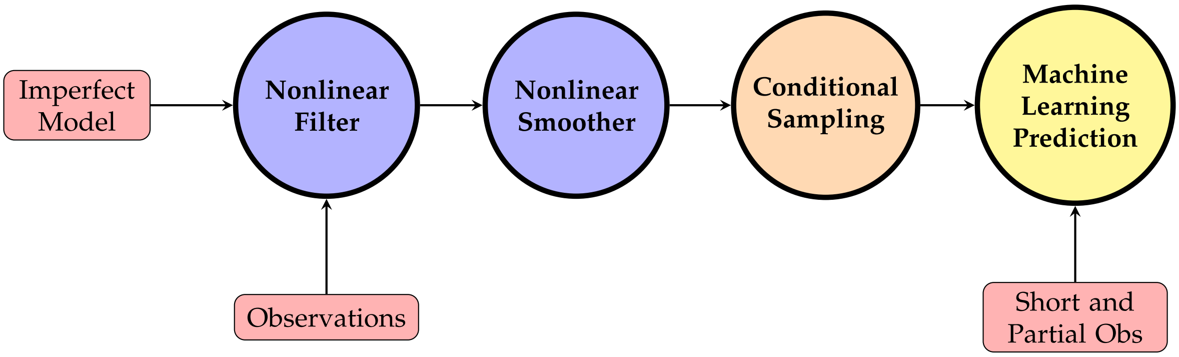

Figure 1. Illustration of the conditional sampling procedure and its application to the augmentation of the training data for machine learning forecasts.

3. The Prediction Schemes

3.1. The Machine Learning Algorithm

The purpose of this article is to elucidate the advantage of using the multiple conditional sampled trajectories to reduce the model error and improve the machine learning prediction. To this end, designing sophisticated machine learning algorithms is not the main focus here. Throughout this article, the long short-term memory (LSTM) network [18] is adopted as the machine learning algorithm. The LSTM network is an artificial recurrent neural network architecture in deep learning. Unlike the standard feed forward neural networks, LSTM has feedback connections. Therefore, it is applicable to predicting sequential data, in particular the time series from complex nonlinear dynamical systems. Only the vanilla LSTM network is used here. It starts with a sequence input layer followed by an LSTM layer. The network ends with a fully connected layer and a regression output layer.

3.2. The Ensemble Forecast

The ensemble forecast is a widely used model-based prediction approach, which involves making a set of forecasts. Due to the intrinsic chaotic behavior of the underlying dynamics, different forecast trajectories are distinct from each other. The ensemble mean forecast time series is often regarded as the path-wise prediction of the true signal and the ensemble spread characterizes the forecast uncertainty. In the following tests, the ensemble forecast is adopted as the prediction scheme for both the perfect and imperfect models. The associated ensemble mean time series will be compared with the LSTM prediction.

4. Predicting Multiscale Compressible Shallow Water Flows Using Lagrangian Observations

In many applications, one fundamental difficulty is that the variables of interest are not directly observed. The goal of this section is to predict multiscale compressible shallow water flows, where the available indirect observations are the Lagrangian tracer trajectories that are transported by the underlying flow field [19–21]. To mimic the real-world scenario, the only information in hand is an imperfect flow model and a short observational trajectory of the Lagrangian tracers.

4.1. The Shallow Water Equation

The linear rotating shallow water equation with double periodic boundary conditions in $[-\pi, \pi)^2$ domain reads [22,23]

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \epsilon^{-1}\mathbf{u}^{\perp} &= -\epsilon^{-1}\nabla\eta, \\ \frac{\partial \eta}{\partial t} + \epsilon^{-1}\delta\nabla\cdot\mathbf{u} &= 0, \end{aligned} \tag{5}$$

where $\mathbf{u} = (u, v)$ is the two-dimensional velocity field and η is the height function. The non-dimensional numbers are $\epsilon = \mathrm{Ro}$, $\delta = \mathrm{Ro}^2\mathrm{Fr}^{-2}$, where Ro is the Rossby number that represents the ratio between the Coriolis term and the advection term, and Fr is the Froude number. For most geophysical problems, ϵ ranges from $O(0.1)$ to $O(1)$, representing fast to moderate rotations while $\delta = 1$ is a typically choice. Applying a plane wave ansatz, the solution of (5) is given by [22]

$$\begin{pmatrix} \mathbf{v}(\mathbf{x},t) \\ \eta(\mathbf{x},t) \end{pmatrix} = \sum_{\mathbf{k}\in\mathbb{Z}^2,\zeta\in\{B,\pm\}} \hat{u}_{\mathbf{k},\zeta}(t)e^{i\mathbf{k}\cdot\mathbf{x}}\mathbf{r}_{\mathbf{k},\zeta}, \tag{6}$$

where $\mathbf{k} = (k_1, k_2)$ is the two-dimensional Fourier wavenumber. The set $\{B, +, -\}$ represents three different modes associated with each Fourier wavenumber. The modes with $\zeta = B$ are the geostrophically balanced (GB) modes, where the GB relation $\mathbf{u}^{\perp} = -\nabla\eta$ always holds. The GB flows are incompressible, which is embodied in the eigenvector $\mathbf{r}_{\mathbf{k},\zeta}$, and the associated phase speed is $\omega_{\mathbf{k},B} = 0$. The modes with $\zeta = \pm$ represent the compressible gravity waves with the phase speeds $\omega_{\mathbf{k},\pm} = \pm\epsilon^{-1}\sqrt{|\mathbf{k}|^2+1}$. See Appendix B for details. Here, the temporal evolution of each random Fourier coefficient $\hat{u}_{\mathbf{k},\zeta}(t)$ is governed by a linear Gaussian process [24],

$$\mathrm{d}\hat{u}_{\mathbf{k},B} = (-d_{\mathbf{k},B}\hat{u}_{\mathbf{k},B} + f_{\mathbf{k},B}(t))\,\mathrm{d}t + \sigma_{\mathbf{k},B}\,\mathrm{d}W_{\mathbf{k},B}(t), \tag{7a}$$

$$\mathrm{d}\hat{u}_{\mathbf{k},\pm} = \Big((-d_{\mathbf{k},\pm} + i\omega_{\mathbf{k},\pm})\hat{u}_{\mathbf{k},\pm} + f_{\mathbf{k},\pm}(t) \Big)\,\mathrm{d}t + \sigma_{\mathbf{k},\pm}\,\mathrm{d}W_{\mathbf{k},\pm}(t). \tag{7b}$$

Such an representation is widely used in practice [25–27], where the damping and random noise allow a simple way to mimic the turbulent features and the interactions between the resolved and unresolved scales. The large-scale deterministic forcings $f_{\mathbf{k},B}(t)$ and $f_{\mathbf{k},\pm}(t)$ represent for example the seasonal cycle. It is clear that, when ϵ is small, the model becomes a multiscale system with slow GB flows and fast gravity waves.

4.2. The Lagrangian Observations

Assume there are L tracers. The equation of each tracer trajectory $\mathbf{x}_l = (x_l, y_l)$ is given by the Newton's law together with some small-scale uncertainties

$$\mathrm{d}\mathbf{x}_l = \mathbf{v}(\mathbf{x}_l)\,\mathrm{d}t + \sigma_{\mathbf{x}}\,\mathrm{d}\mathbf{W}_l = \sum_{\mathbf{k}\in\mathbb{Z}^2,\zeta\in\{B,\pm\}} \hat{u}_{\mathbf{k},\zeta}(t)e^{i\mathbf{k}\cdot\mathbf{x}_l}\tilde{\mathbf{r}}_{\mathbf{k},\zeta} + \sigma_{\mathbf{x}}\,\mathrm{d}\mathbf{W}_l, \tag{8}$$

where $\tilde{\mathbf{r}}_{\mathbf{k},\zeta}$ is a two-dimensional vector, containing the first two components of the eigenvector $\mathbf{r}_{\mathbf{k},\zeta}$. It is important to note that the GB and the gravity modes are coupled in the observations. In addition, the tracer equation is highly nonlinear because the prognostic variable $\mathbf{x}_l$ appears in the exponential function on the right-hand side. These features lead to a tough test in recovering and sampling the underlying velocity fields. Nevertheless, despite the intrinsic nonlinearity, the coupled system (6)–(8) belongs to the modeling framework (1), which facilitates the conditional sampling of the flow trajectories [28].

4.3. Setup of the Numerical Tests

The modes with wavenumbers $-1 \leq k_1, k_2 \leq 1$ are used in the study here, which means the total number of the Fourier modes is $9 \times 3 = 27$. However, the first two components of the $(0, 0)$-th GB mode are zero. This mode often represents the given background flow. Since it has no contribution to the observational process, it is removed here. Thus, the degree of freedom (DoF) of the underlying model is DoF $= 26$. See Appendix B for details. The parameters in (7) are $d_{\mathbf{k},B} = d_{\mathbf{k},g} = 0.5$, $\sigma_{\mathbf{k},B} = \sigma_{\mathbf{k},\pm} = 0.4$ and $f_{\mathbf{k},B} = f_{\mathbf{k},\pm} = 0$, which indicate an equipartition of the energy in different modes. The uncertainty in the tracer Equation (8) is $\sigma_{\mathbf{x}} = 0.1$. The number of tracers is $L = 20$, which is slightly smaller than the DoF of the resolved modes and is the typical case in practice. Appendix B includes a sensitivity analysis of the prediction skill as a function of L, which shows qualitatively similar results as those in the main text here.

Below, a small Rossby number $\epsilon = 0.2$ is used in the perfect model. In practice, the dynamics of the slowly-varying GB modes is often known, but building an accurate model for the fast gravity modes is a challenging task, especially for estimating the phase speeds $\omega_{\mathbf{k},\pm}$ [29]. To mimic such a situation, an imperfect model is used here to sample and predict the gravity modes, where the model error comes from an overestimation of the Rossby number with $\epsilon = 0.5$ in the imperfect model that leads to an inaccurate $\omega_{\mathbf{k},\pm}$. The available observational data has only $T = 10$ units, which is about five decorrelation times of the Fourier modes. The prediction test is taken on an independent period with 50 units, where the true signal generated from the perfect model in this period is used as the reference value to compute the prediction skill scores in the forecast experiments.

4.4. Conditional Sampling

To expand the training data set, $N = 20$ distinct trajectories from the conditional sampling algorithm are generated by exploiting the imperfect model, the short observations, and the formula in (4). Therefore, the effective training data for the LSTM network contains $TN = 200$ units. Figure 2a shows the true signal and the smoother estimate of the gravity mode $(0, -1)$, where the uncertainty in the smoother estimate facilitates the sampling of multiple trajectories. The variability in these trajectories significantly enriches the short-term time evolution patterns that are associated with the underlying flow field but are not fully reflected in the short true time series. It is also important to note that the model error in the phase speed is reduced by a large extent in the conditional sampled trajectories. In fact, (b) includes the ACF of the sampled trajectory, which demonstrates a large improvement compared with that of the imperfect model. Since the ACF characterizes the temporal dependence, the result here suggests a potential enhancement of the prediction accuracy in the LSTM forecast using the conditional sampled trajectories as training data.

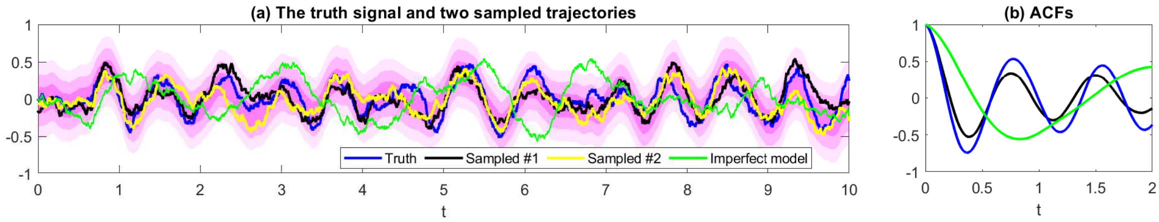

Figure 2. Gravity model $(0, -1)$ of the shallow water equation. (**a**) the true signal and two sampled trajectories conditional on the observations of the tracers' motions. The deep, moderate and light magenta shading areas show the one, two, and three standard deviations (STDs) of the uncertainty in the smoother estimate, respectively; (**b**) the ACFs of the truth, the conditional sampled trajectory and the trajectory from a free run of the imperfect model.

4.5. The LSTM Setup

Recall the total DoF of the underlying flow field is 26, where the DoF of the gravity modes is 18. Since the equations of different Fourier modes are independent from each other in the perfect model, it is reasonable to build nine LSTM networks for predicting these gravity modes, each of which contains two Fourier modes that are the complex conjugate pair. Similar manipulation is used for predicting the GB modes. Since no model error is involved in the GB modes, the LSTM prediction of the GB part of the flow is almost the same as the perfect model ensemble forecast. Note that certain weak coupling between different Fourier modes may exist in the conditional sampled time series because the observations are the mixture of all the Fourier modes. Nevertheless, dividing a complex problem into several independent subproblems facilitates learning the features of the underlying dynamics. The input time series has a length of 0.25 time units while the output is the next 0.005 time units. Each of the LSTM used here contains only one hidden LSTM layer, where the number of the neurons is 100. The maximum epoch is 100. The Adam optimization algorithm is adopted, and a mean squared error loss function is used.

4.6. Prediction

For both the model-based ensemble forecast and the machine learning prediction here, the initial conditions of the flow fields are assumed to be perfectly known in the prediction stage. Though in practice data assimilation is required for the initializations, the idealized setup here rules out the impact from initialization and allows us to study the predictability limit of different methods.

Figure 3 shows the root-mean-square error (RMSE) and the pattern correlation (Corr) of the predicted time series related to the truth as a function of lead time. These skill scores for the gravity modes $(0, -1)$ and $(1, 0)$ with $\zeta = +$ are included in (a,b). The prediction of the other gravity modes has similar behavior. (c) shows the reconstructed velocity field associated with the gravity modes in physical space. Clearly, the skill scores of the LSTM prediction are quite close to those of the perfect model, while the imperfect model has a much larger forecast bias. The improvement in the LSTM prediction is due to the significant reduction of the model error and the extended length of the training time series resulting from the conditional sampling algorithm.

A case study is included in Figure 4, which compares the predicted flow fields at a lead time $t_{lead} = 0.4$. The first row shows the truth and the predicted velocity field associated with only the gravity modes while the second row shows those of total velocity field that includes both the GB and gravity modes. The perfect model succeeds in predicting both the compressible gravity flows and the total velocity field. In contrast, the predicted flow pattern associated with the gravity modes from the imperfect model is completely reversed. As a result, the imperfect model fails to forecast the vortices in the total flow field in that they are overwhelmed by the large error from the forecasted gravity waves. On the other hand, despite a slight overestimation of the flow amplitude, the overall patterns are forecasted accurately by the LSTM network. In particular, both the meandering jets and the predominant vortices in the total velocity field are precisely predicted by the LSTM network.

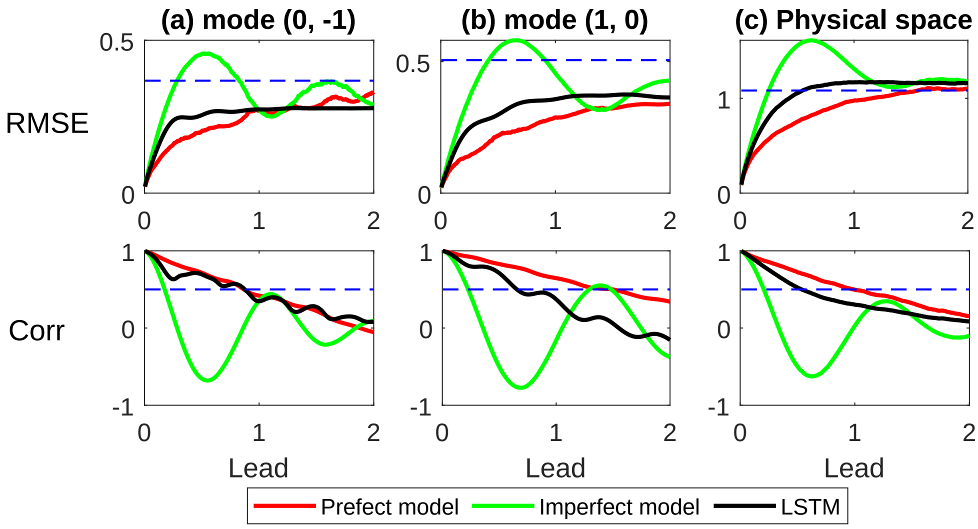

Figure 3. The RMSE and the Corr as a function of lead time. (**a**,**b**) the gravity modes $(0,-1)$ and $(1,0)$ with $\zeta = +$. The RMSE and Corr are computed based only on the real part of the Fourier time series; (**c**) the reconstructed velocity field associated with the gravity modes in physical space, where the skill scores are the averaged values of the u and v components. In reconstructing the velocity field in physical space, a 25×25 mesh grid is used. The red, green, and black curves show the prediction skill scores using the perfect model, the imperfect model, and the LSTM network. The dashed lines in the RMSE panels indicate the one standard deviation of the true signal and those in the Corr panels show the Corr $= 0.5$ threshold.

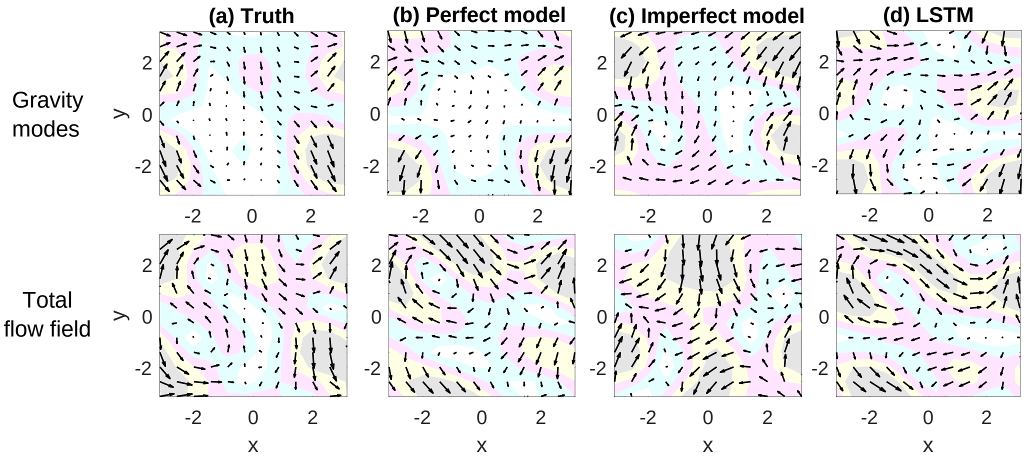

Figure 4. Predicting the flow fields at a lead time $t_{lead} = 0.4$ in physical space. The prediction starting from $t = 12$ in the prediction phase. (**a**–**d**) compare the truth, the perfect model forecast, the imperfect model forecast, and the LSTM forecast. The first row shows the velocity field associated with only the gravity modes while the second row shows the total velocity field. The velocity field in each panel is shown by the quivers and its amplitude is denoted by the shading area.

5. Predicting the Monsoon Intraseasonal Oscillation (MISO)

5.1. The MISO Index and the Low-Order Nonlinear Stochastic Models

Monsoon Intraseasonal Oscillation (MISO) [30–32] is one of the prominent modes of tropical intraseasonal variability. As a slow moving planetary scale envelope of convection propagating northeastward, it strongly interacts with the boreal summer monsoon rainfall over south Asia. The MISO plays a crucial role in determining the onset and demise of the Indian summer monsoon as well as affecting the seasonal amount of rainfall over the Indian subcontinent. Therefore, both real-time monitoring and accurate extended-range forecast of MISO phases have large ecological and societal impacts. Recently, a new MISO index [33], by applying a new nonlinear data analysis technique to the daily Global Precipitation Climatology Project (GPCP) rainfall data [34] over the Asian summer monsoon region (20° S–30° N, 30° E–140° E), was developed. This new index outweighs the traditional ones that are based on the extended empirical orthogonal function (EEOF) in capturing the intermittency and nonlinear features of the MISO. The associated MISO modes also have higher memory and predictability, stronger amplitude, and higher fractional explained variance over the western Pacific, Western Ghats, and adjoining Arabian Sea regions, and more realistic representation of the regional heat sources over the Indian and Pacific Oceans compared with those extracted via EEOF analysis. (a) of Figure 5 shows this MISO index, which is a two-dimensional time series. Both components are intermittent with active phases in summer and quiescent phases in winter. The associated PDFs, as are shown in the blue curves, are highly non-Gaussian.

To develop a model that describes such a MISO index, it is natural to start with a two-dimensional linear stochastic oscillator. However, the linear oscillator model itself is not sufficient to characterize all the observed features of the observed MISO index. In particular, it fails to capture the variability in the amplitude and the oscillation frequency. In fact, as is shown in (d–e) in Figure 5, the dynamical and statistical behavior of the MISO index in different years is quite distinct with each other. Therefore, a simple but effective way to take into account these characteristics is to add two extra processes, representing the randomness in damping and phase [35]. The model reads

$$\begin{aligned}
\mathrm{d}u_1 &= \Big(-d_u u_1 + \gamma(v + v_f(t))u_1 - (a+\omega)u_2\Big)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_{u_1},\\
\mathrm{d}u_2 &= \Big(-d_u u_2 + \gamma(v + v_f(t))u_2 + (a+\omega)u_1\Big)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_{u_2},\\
\mathrm{d}v &= -d_v v\,\mathrm{d}t + \sigma_v\,\mathrm{d}W_v,\\
\mathrm{d}\omega &= -d_\omega \omega\,\mathrm{d}t + \sigma_\omega\,\mathrm{d}W_\omega,
\end{aligned} \tag{9}$$

where the time-periodic function

$$v_f(t) = f_0 + f_1 \sin(\omega_f t + \phi) \tag{10}$$

provides a crude description of the seasonal cycle. In this model, u_1 and u_2 stand for the two components of the MISO index while v and ω are the stochastic damping and the stochastic phase, respectively. The white noise sources W_{u_1}, W_{u_2}, W_v and W_ω are independent from each other. Note that the nonlinear model developed here is slightly different from the physics-constraint model in [35]. However, these two models have almost the same skill in predicting the MISO index. The main reason to adopt the form in (9) and (10) is that the model structure of this coupled nonlinear model facilitates the application of the conditional sampling algorithm.

To illustrate the skill of the coupled model (9) and (10) in capturing the nonlinear and non-Gaussian features of the MISO index, it is shown in (b,c) of Figure 5 that the model can perfectly recover the ACFs, the cross-correlation functions (CCFs), and the non-Gaussian PDFs of the entire MISO index. The model trajectories also highly resemble the true MISO index (see Figure A3 in Appendix C). The associated model parameters are listed in the top row of Table 1. Due to the high skill in describing the MISO index, the model (9) and (10) with these parameters is named as the "nearly perfect model".

Table 1. Two sets of the parameters for the model (9) and (10).

	d_u	d_v	d_ω	γ	a	σ_u	σ_v	σ_ω	f_0	f_1	ω_f	ϕ
I. nearly perfect model	0.9	0.6	0.5	0.2	4.1	0.5	0.5	0.7	1	4.7	$2\pi/12$	-2
II. imperfect model	0.9	0.6	0.5	0.2	5.2	0.5	0.5	0.7	1	4.7	$2\pi/12$	-2

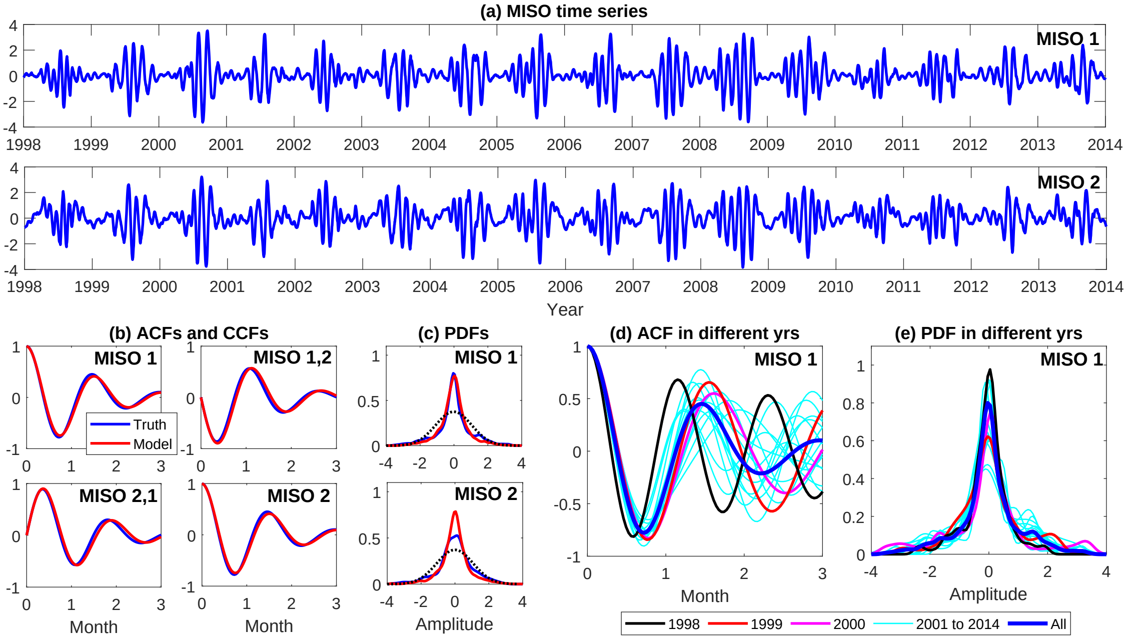

Figure 5. MISO index and the associated statistics. (**a**) the two components of the MISO index; (**b**) the ACFs and the cross-correlation functions (CCFs) associated with the MISO time series in (**a**) and those from the simulation of the model (9) and (10) with the parameters listed in the top row of Table 1; (**c**) similar to (**b**) but for the PDFs; (**d**) the ACF of the MISO 1 component in different years; (**e**) similar to (**d**) but for the PDF.

It is important to note that the nearly perfect model is obtained by exploiting *all* the observational information (1998–2013) for model calibration. Therefore, despite the fact that such a model succeeds in describing nature, it cannot be used for predicting the MISO index during the same period because of the overfitting issue. In the following, only the MISO index in year 1998 is used for model calibration. This mimics most applications in nature where the data for model calibration is very limited. On the other hand, the time series from year 2001 to year 2013 is adopted for testing the prediction skill, which provides a sufficient long validation period to reach a robust conclusion.

Due to the fact that the MISO index has distinct year-to-year behavior ((d,e) in Figure 5), the model parameters estimated by using only the time series in year 1998 are different from the nearly perfect model. These estimated model parameters are shown in the bottom row of Table 1 and the resulting model is named as the “imperfect model” in predicting the MISO time series from year 2001 to year 2013. Below, either a one-year or a three-year time series is combined with this imperfect model to enrich the training data set for the machine learning forecasts by applying the conditional sampling technique. Note that, although the imperfect model can be in principle re-calibrated when the new data comes sequentially in time, the parameters in the imperfect model are unchanged in the tests here for two reasons. First, the online parameter estimation for complex nonlinear systems in real applications can be extremely expensive and thus it is seldom adopted in practice. Second, there often exists an intrinsic model error in real applications. The imperfect parameters here mimic such an intrinsic barrier since otherwise the model is nearly perfect.

5.2. Conditional Sampling

Below, the short observational data in (1) year 1998, (2) year 1999, (3) year 2000, or (4) a three-year period 1998–2000 are utilized together with the imperfect model to enrich the training data set of the machine learning forecast based on the conditional sampling technique. The conditional sampling algorithm here involves two steps:

Step 1. Conditioned on the two components $\mathbf{X} = (u_1, u_2)^T$ of the observed MISO index, sample N trajectories of $\mathbf{Y} = (v, \omega)^T$.

Step 2. Conditioned on each of the sampled trajectories, denoted now by $\mathbf{X} = (v, \omega)^T$, from Step 1, sample a trajectory of $\mathbf{Y} = (u_1, u_2)^T$.

In both the steps, $\mathbf{X}$ denotes the conditional variables while $\mathbf{Y}$ stands for the sampled variables. The structure of the nonlinear model (9) and (10) allows the setups in both the steps belong to the nonlinear modeling family (1) such that the closed analytic Formulae (4) can be applied in sampling both (v, ω) and (u_1, u_2). Note that, despite (u_1, u_2) being not appeared in the equations of (v, ω), the conditional sampling in step 2 is necessary since uncertainty exists in the processes of (v, ω). Simply plugging the sampled trajectories of (v, ω) into the equations of (u_1, u_2) leads to biased results.

5.3. The Setup of the LSTM and the Model Ensemble Forecast

In all the tests here, a 30-year sampled time series of (u_1, u_2) is used as the machine learning training data. This implies $N = 30$ when a one-year observational data are used in the conditional sampling and $N = 10$ when the three-year data from year 1998 to year 2000 is used. The hidden variables (v, ω) are only used to generate the sampled trajectories of (u_1, u_2), but they are not directly utilized in the machine learning training and forecasting steps. The input data (u_1, u_2) in the LSTM has a length of 30 days and the output is the next one day. The number of the neurons in the LSTM layer is 200, and the maximum epoch is 100.

The model ensemble forecast is adopted as a comparison with the machine learning prediction. The model ensemble forecast exploits 50 ensemble members, which have been validated to provide an unbiased ensemble mean prediction time series. Since there is no observational data for the two hidden variables v and ω, the exact data assimilation scheme in (2) is utilized for their initializations.

5.4. Prediction

The MISO index prediction using different methods is shown in Figure 6. The solid curves in both (a) and (b) illustrate the overall prediction skill in the 2001–2013 period using the LSTM. The difference between these two panels is the following: In (a), the training data are the short observational data with either one year (blue, red, and green curves) or three years (black curves). On the other hand, in (b), the 30-year sampled time series is used for the LSTM training. As a comparison, the model ensemble forecast results are also included in these panels, where the dashed pink curve shows the prediction using the nearly perfect model while the dashed cyan curve illustrates this using the imperfect model.

The main conclusions are as follows: first, either a one-year or a three-year time series is not sufficient for training the LSTM. Among all these short training data, using the time series in year 1998 brings about the worst prediction skill. This is because the dynamical and statistical features of the MISO index in year 1998 are more distinguished from those in the entire observational period, compared with year 1999 and year 2000 (see (d,e) in Figure 5). The consequence is that the LSTM prediction using only the time series of year 1998 as the training data performs even significantly worse than the ensemble forecast using the imperfect model. Second, the LSTM prediction using the long sampled trajectories based on the information of the extremely short observations in year 1999 or year 2000 has almost the same skill as the perfect model forecast! In fact, the medium-range forecast (20 to 30 days) using the LSTM even slightly outweighs that using the perfect model. Notably, the LSTM prediction using the long sampling data based on the time series of year 1998 is also greatly improved and outweighs the ensemble forecast using the imperfect model. This seems to be counter-intuitive. Nevertheless, since the model is stochastic, it is able to generate patterns with different oscillation frequencies in the conditional sampled trajectories. The LSTM forecast then exploits the input information to find the most similar patterns in the training data set. In other words, the LSTM gives different weights to all the possible events in the training data set, which can be regarded as different ensemble members. This is, however, not the case in the model based ensemble mean forecast, where the same weight is assigned to all the ensemble members. Thus, despite making use of the same observational information, the LSTM performs better than the model ensemble mean forecast. In Figure A4 of Appendix C, the model generated training data are used for the LSTM prediction, which also indicates such a mechanism.

(e) of Figure 6 shows the maximum lead time of skillful predictions in different years. Here, the skillful prediction is defined as the Corr > 0.5 and RMSE < 1STD (standard deviation) of the true signal. Although the useful predictions vary from year to year, the LSTM prediction using the conditional sampled trajectories based on the one-year observed time series in 1999 is comparable to the nearly perfect model forecast. The associated maximum lead time is longer than that using the three-year short training data and the imperfect model forecast in most of the years. The only exception is year 2010. In fact, the MISO pattern of year 2010 is similar to that of year 1998 with a weak amplitude. Note that the winter of both year 1998 and year 2010 corresponds to a strong positive phase of the Atlantic Zonal Mode (AZM) [36,37]. The results here confirm the negative correlation between the MISO and the AZM as well as the relatively short predictability of the MISO at the positive phase of the AZM. Next, 2002 and year 2004 are recorded as drought years. The MISO index in these two years is more irregular than the other years, which explains its lower overall prediction skill than most of the other years. On the other hand, 2008 has an overall strong and regular MISO activity during the whole monsoon season that results in a long predictability. Finally, Figure 7 shows the predicted time series at the lead time of 30 days, which clearly indicates the LSTM with the conditional sampled training data outperforms the imperfect model ensemble forecast in most of the years.

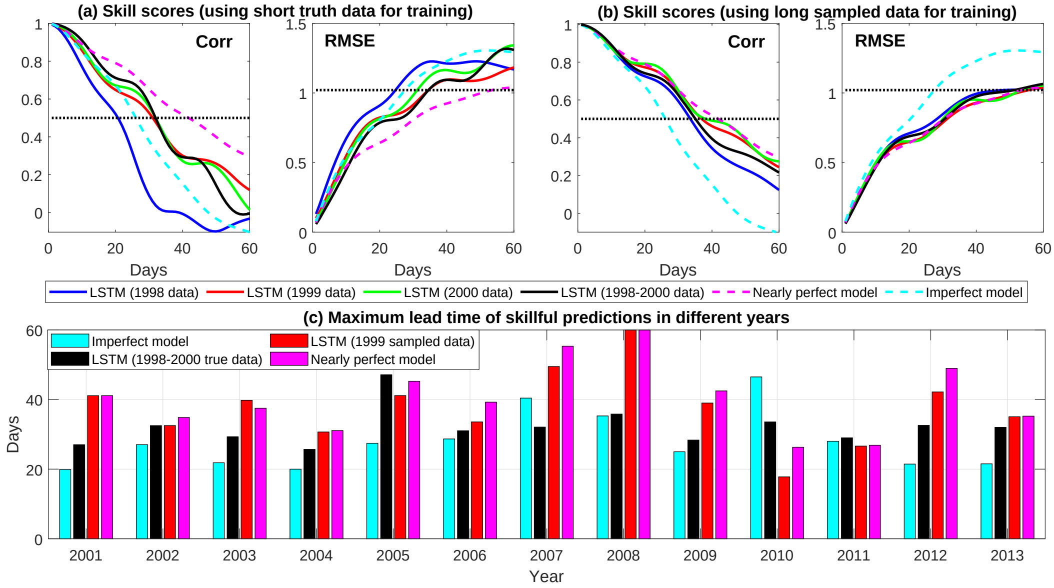

Figure 6. The MISO prediction. (**a**) the RMSE and Corr as a function of lead time (days) in predicting the MISO time series from 2001 to 2013, where the short observational data are used for training the LSTM. Here, the solid blue, red, and green curves are the LSTM prediction based on a one-year training data using the observed time series in year 1998, 1999, and 2000, respectively. The solid black curve is the LSTM prediction based on a three-year training data using the observed time series from 1998 to 2000. The dashed pink and cyan curves show the model predictions using the nearly perfect model and the imperfect model. The black dotted lines show the threshold Corr = 0.5 and one standard deviation of the true time series. (**b**) is similar to (a), but the training data of the LSTM are obtained by applying the conditional sampling algorithm to the observed time series in either one of the three years of 1998 (blue), 1999 (red) and 2000 (green), or all three years (black). The total length of the sampled data are 30 years; (**c**) the maximum lead time of the skillful prediction in different years. Here, the skillful prediction is defined as the Corr > 0.5 and RMSE < 1STD (standard deviation) of the true signal.

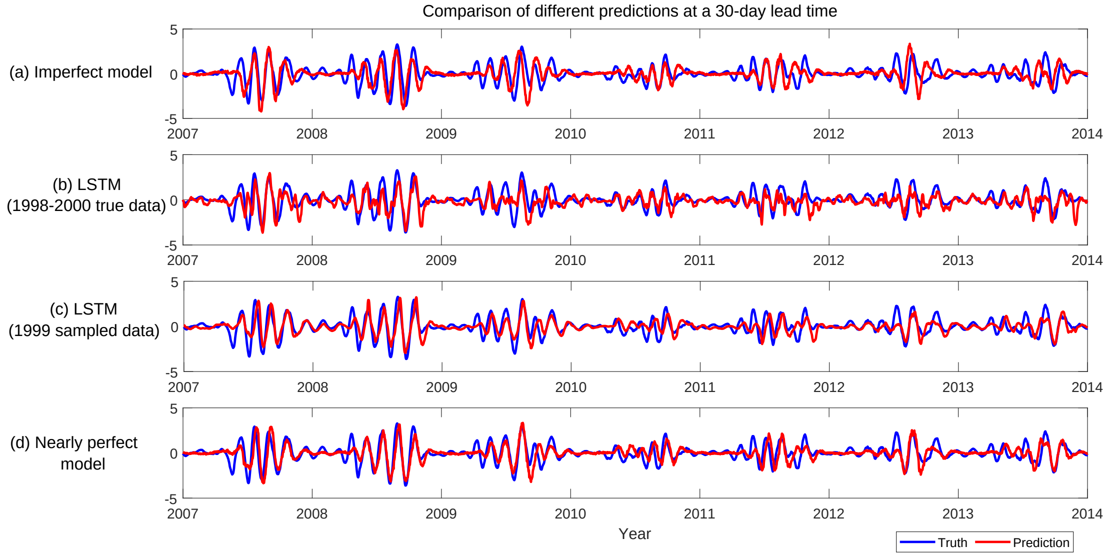

Figure 7. Comparison of the predicted MISO index using different methods. (**a**) prediction using the imperfect model; (**b**) prediction using the LSTM, where the training data are the true observational data from 1998 to 2000; (**c**) prediction using the LSTM, where the training data are obtained by applying the conditional sampling algorithm to the observed MISO time series of year 1999 with in total 30 years sampled data; (**d**) prediction using the nearly perfect model.

6. Conclusions

An efficient and optimal algorithm for sampling nonlinear time series conditioned on the observations is developed in this article. It exploits only short and partial observations to greatly reduce the model error and expand the training data set for the machine learning forecast algorithms. The sampling algorithm succeeds in creating a large training data of multiscale compressible shallow water flows from highly nonlinear and indirect observations. The resulting machine learning prediction significantly outweighs the imperfect model forecast. The sampling algorithm also facilitates the machine learning forecast of the highly non-Gaussian MISO time series using extremely short observations.

Note that the MISO is an intraseasonal variability and therefore the total available data of MISO in the satellite era is actually not that restricted. The reason to adopt only a small portion of the observed MISO data for training is to mimic the real situations of predicting many other nature phenomena. Meanwhile, the remaining relatively long MISO trajectories let us study the improvement of the MISO forecast using the conditional sampling technique. On the other hand, the El Niño-Southern Oscillation (ENSO) is an interannual variability, which has significant impact on the entire earth system. The ENSO has various spatiotemporal patterns, but the observational data of the ENSO is very scarce. In addition, the current operational models have large model error in describing and predicting the ENSO. Thus, an interesting and useful task is to utilize the conditional sampling algorithm for improving the ENSO prediction, which is left as a future work. Another future direction is to generalize the conditional sampling algorithm to general nonlinear and non-Gaussian systems using particle methods.

Funding: The research of N.C. is funded by the Office of VCRGE at UW-Madison and ONR N00014-19-1-2421.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Conditional Sampling

Appendix A.1. An Alternative Conditional Sampling Formula That Requires Only the Filter Estimates

Recall the conditional sampling formula in the main text (Theorem 3), which is mathematically concise as is discussed in the main text. However, it requires the information of both the filter and smoother. An alternative form that makes use of only the filter estimate is given as follows, which is more computationally convenient.

Corollary A1 (An Alternative Form of the Conditional Sampling Formula)**.**

$$\overleftarrow{\mathrm{d}\mathbf{Y}} = \big(- \mathbf{a_0} - \mathbf{a_1}\mathbf{Y}\big)\,\mathrm{d}t + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}\big(\boldsymbol{\mu}_\mathbf{f} - \mathbf{Y}\big)\,\mathrm{d}t + \big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_\mathbf{Y}, \tag{A1}$$

Below is a simple proof of the equivalence of Theorem 3 and Corollary A1.

Proof. Recall the smoother posterior mean equation

$$\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}} = \big(- \mathbf{a_0} - \mathbf{a_1}\boldsymbol{\mu}_\mathbf{s} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}(\boldsymbol{\mu}_\mathbf{f} - \boldsymbol{\mu}_\mathbf{s})\big)\,\mathrm{d}t. \tag{A2}$$

Taking the summation of (4) and (A2) cancels $\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}}$ and $\boldsymbol{\mu}_\mathbf{s}$. The resulting equation yields

$$\begin{aligned}\overleftarrow{\mathrm{d}\mathbf{Y}} - \overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}} &= -\mathbf{a_1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{s})\,\mathrm{d}t + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}\big(\boldsymbol{\mu}_\mathbf{s} - \mathbf{Y}\big)\,\mathrm{d}t + \big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_\mathbf{Y}\\ &= -\big(\mathbf{a_1} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}_\mathbf{f}^{-1}\big)(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{s})\,\mathrm{d}t + \big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_\mathbf{Y},\end{aligned} \tag{A3}$$

which is (A1). □

Appendix A.2. Comparison of the Posterior Mean Time Series and the Trajectories from Conditional Sampling in Reproducing the Dynamical and Statistical Characteristics of Nature

The goal of this subsection is to show that, even in the perfect model setup, the posterior mean time series fails to reproduce the basic dynamical and statistical features of nature. This implies that the machine learning prediction can be biased if the training is based on the posterior mean time series.

Consider a simple example

$$\mathrm{d}u = (-\gamma u + f_u)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u \tag{A4a}$$

$$\mathrm{d}\gamma = (-d_\gamma\gamma + u^2 + f_\gamma)\,\mathrm{d}t + \sigma_\gamma\,\mathrm{d}W_\gamma \tag{A4b}$$

with the following parameters

$$\sigma_u = 1, \qquad \mathrm{d}_\gamma = 0.5, \qquad f_\gamma = 0.8, \qquad \sigma_\gamma = 2. \tag{A5}$$

This is a simple nonlinear dyad model with energy conserving nonlinearity (known also as physics constraints) [38,39]. The variable u is observed while the hidden variable v plays the role of the stochastic damping that triggers intermittent instability and extreme events in u. See the burst events in (a) of Figure A1.

Now conditioned on one observed trajectory of u, the filter and smoother posterior mean time series of v are shown in (b) of Figure A1. At the phases when extreme events occur in u, the smoother posterior mean succeeds in recovering the true signal, but the filter posterior mean fails to predict the timing of the event onset and the duration. However, when the observed signal of u is at its quiescent phases, the small signal-to-noise ratio leads to a large error in the posterior mean time series with a large uncertainty. The dynamical behavior of the posterior mean time series is obviously very different from the truth. It is thus expected that the machine learning prediction can be biased by using these time series as the training data. In fact, as is shown in (d)–(e), both the PDFs and ACFs associated with the posterior mean time series (regardless of filter or smoother) are biased from the truth. The large error in the PDFs from (d) is as expected since the posterior mean time series of γ in (b) essentially follows the model's equilibrium state at the quiescent phases of u and fails to capture the variability of the truth. Its consequence to the machine learning prediction is that the events of γ do not appear in the training period (the posterior mean time series) will never be forecasted and therefore the forecast contains a large error. On the other hand, the ACFs associated with the posterior mean time series are also biased from the truth according to (e). This can be seen intuitively from the time series in (b). Since the ACF represents the averaged path-wise memory of the system, the error in the ACFs affects the short-term prediction. In contrast, the sampled trajectories from the conditional sampling perfectly recover both the PDFs and the ACFs. This indicates that the sampled trajectories are able to fully recover both the dynamical and statistical features of nature. Therefore, training the machine learning algorithms using these sampled trajectories are expected to provide skillful prediction results.

(c) shows two sampled trajectories. It is clear that these sampled trajectories have similar dynamical and statistical features as the truth. In addition, since they are calculated by conditioning on the observations, they are different from a free run of the model. This is an extremely important feature when model error comes, at which time the observations can mitigate the model error and the resulting sampled trajectories will be more accurate than the model free run. On the other hand, the sampled trajectories may have a potential to facilitate the prediction of rare and extreme events. In fact, despite a large variability in the sampled trajectories of γ at the quiescent phases of u, the uncertainty in the sampled trajectories corresponding to the extreme events of u is small. This means the phases corresponding to the extreme events are guaranteed to be sampled in these trajectories. Meanwhile, the sampled trajectories themselves have an extra chance to reach the phases corresponding to extreme events—for example the event at time instants $t = 98$ on the brown curves.

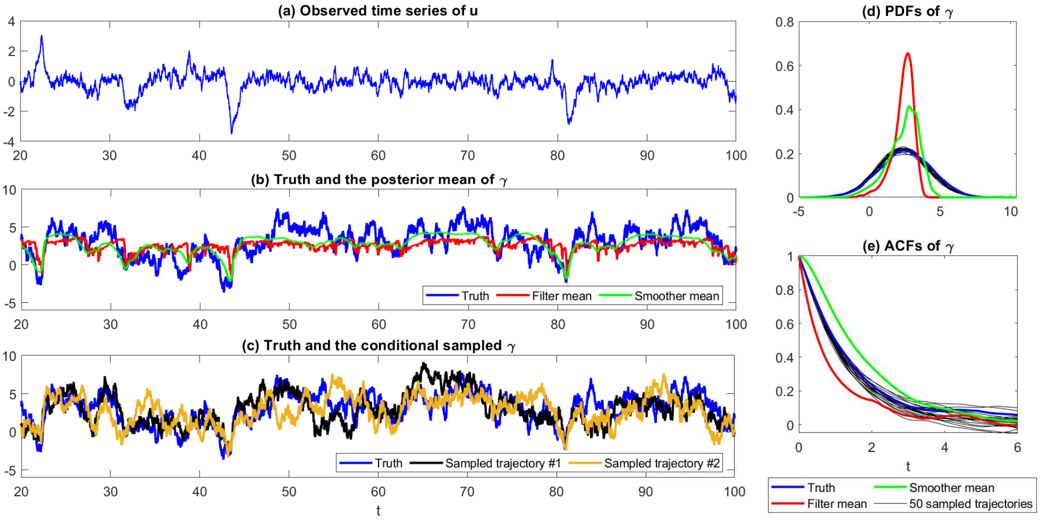

Figure A1. (**a**) the true signal of u; (**b**) the true signal of γ (blue), the filter posterior mean time series (red) and the smoother posterior mean time series (green); (**c**) the true signal of γ (blue) and two sampled trajectories (black and brown); (**d**,**e**) comparison of the PDFs and ACFs of γ. The blue curve is the PDF associated with the true signal. The red and green curves are the PDFs associated with the filter and smoother posterior mean time series, respectively. The black curves are the PDFs associated with 50 trajectories from the conditional sampling. These statistics are computed based on time series with 1000 time units, which are sufficiently long to rule out almost all the undersampling errors.

Appendix B. Details of Predicting Multiscale Compressible Shallow Water Flows Using Lagrangian Observations

Appendix B.1. Details of the Shallow Water Equation

Recall the two-dimensional linear rotating shallow water Equation (5) in the main text, where the solution is given by

$$\begin{pmatrix} \mathbf{v}(\mathbf{x},t) \\ \eta(\mathbf{x},t) \end{pmatrix} = \sum_{\mathbf{k}\in\mathbb{Z}^2,\zeta\in\{B,\pm\}} \hat{u}_{\mathbf{k},\zeta}(t)e^{i\mathbf{k}\cdot\mathbf{x}}\mathbf{r}_{\mathbf{k},\zeta}. \tag{A6}$$

In (A6), the modes with $\zeta = B$ are the geostrophically balanced (GB) modes, where the GB relation $\mathbf{u}^{\perp} = -\nabla\eta$ always holds. The GB flows are incompressible, where the incompressibility $\nabla\cdot\mathbf{u} = 0$ is embodied in the eigenvector $\mathbf{r}_{\mathbf{k},\zeta}$, and the associated phase speed is $\omega_{\mathbf{k},B} = 0$. The modes with $\zeta = \pm$ represent the gravity modes (also known as the Poincaré waves). The associated phase speed is $\omega_{\mathbf{k},\pm} = \pm\epsilon^{-1}\sqrt{|\mathbf{k}|^2+1}$, and they are compressible. The normalized eigenvector of the GB modes $\mathbf{r}_{\mathbf{k},B}$ is given by

$$\mathbf{r}_{\mathbf{k},B} = \frac{1}{\sqrt{|\mathbf{k}|^2+1}}\begin{pmatrix} -ik_2 \\ ik_1 \\ 1 \end{pmatrix}. \tag{A7}$$

The normalized eigenvectors $\mathbf{r}_{\mathbf{k},\pm}$ of the gravity modes are given by

$$\mathbf{r}_{\mathbf{k},\pm} = \frac{1}{|\mathbf{k}|\sqrt{(\delta+\delta^2)|\mathbf{k}|^2+2}}\begin{pmatrix} ik_2 \pm k_1\sqrt{\delta|\mathbf{k}|^2+1} \\ -ik_1 \pm k_2\sqrt{\delta|\mathbf{k}|^2+1} \\ \delta|\mathbf{k}|^2 \end{pmatrix} \tag{A8}$$

For the special case, $\mathbf{k} = \mathbf{0}$, the Poincaré waves have no gravity component and coincide with the inertial waves. The resulting eigenvalues become $\mathbf{r}_{\mathbf{0},\pm} = \pm\epsilon^{-1}$ with the eigenvectors being

$$\mathbf{r}_{\mathbf{0},\pm} = \frac{1}{\sqrt{2}}\begin{pmatrix} \pm i \\ 1 \\ 0 \end{pmatrix}. \tag{A9}$$

Note from the eigenvector of the GB modes (A7) that the first two components of the $(0,0)$-th mode $\mathbf{r}_{\mathbf{0},B}$ are zero. This mode often represents the given background flow. Since it has no contribution to the observational process, it is removed here. Recall that the modes with Fourier wavenumbers $-1 \leq k_1, k_2 \leq 1$ are used in the main text. Therefore, the total degree of freedom (DoF) of the underlying flow model is DoF = 26, which includes 8 GB modes and 2×9 gravity modes.

Appendix B.2. Sensitivity Analysis

Sensitivity analysis is carried out in this subsection to study the dependence of the prediction skill on several key parameters.

One of the important parameters for the sensitivity study is the number of the Lagrangian tracers L. To begin with, it is expected that the LSTM network prediction can become less accurate in the two extreme situations with either a very large or a very small L. In fact, when $L = 1$, the information provided by the observation is very limited. As a result, the LSTM prediction is only slightly better than the forecast based on the imperfect model. On the other hand, when L becomes infinity, the posterior estimate converges to the true short signal with no uncertainty. However, a short training period is not sufficient for training the LSTM network, which therefore leads to the deterioration of the prediction skill. In such a situation, the LSTM network prediction is even worse than the imperfect model forecast. Figure A2a shows the maximum lead time of the useful prediction of the LSTM network as a function of L. Here, the useful prediction is defined as the lead time, within which the RMSE is less than the one standard deviation of the true signal and the Corr is above the threshold Corr $= 0.5$. It is seen that, within a relatively wide range $L \in [17, 26]$, the LSTM network prediction is significantly more skillful than the imperfect model forecast. The LSTM network prediction remains more skillful than the imperfect model prediction by a large amount even with a further increasing or decreasing of L. Note that, despite the useful prediction based on the LSTM network having an obvious gap compared with that from the perfect model prediction, the overall difference in the prediction skill between these two methods is not as significant as that between the LSTM network prediction with the imperfect model forecast. In fact, as shown in (b), the skill score curves associated with both the LSTM network and the perfect model forecast flatten after a short time, which lead to the large gap in (a), but the two curves are not too far from each other.

Another key parameter for the sensitivity analysis is the length of the observational period T. Recall that the length of the observation is $T = 10$ and $N = 20$ repeated sampled trajectories are used to form the training data set of the LSTM network in the main text. Two extra experiments are carried out here with $T = 50$ and $N = 4$ (yellow circle/curve) and $T = 200$ and $N = 1$ (pink circle/curve). It turns out that, within $t_{lead} = 0.3$, the prediction skill with $T = 10$ is nearly the same as those with $T = 50$ and $T = 200$. As the lead time increases, the Corr with $T = 50$ and $T = 200$ becomes slightly higher than the short observational case with $T = 10$, as is expected. Nevertheless, comparing with the enhanced prediction skill from using the imperfect model (green curve) to the LSTM forecast with $T = 10$, the above improvement is relatively marginal. Therefore, these tests indicate that a short observation with $T = 10$ is sufficient to allow a skillful prediction using the LSTM network since the multiple sampled trajectories include a large variability of the underlying systems.

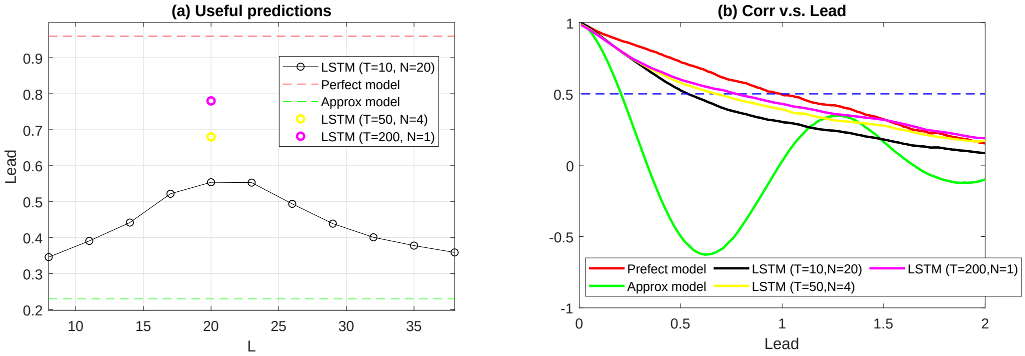

Figure A2. Sensitivity analysis. (**a**) the black curve shows the maximum lead time of the useful prediction using the LSTM network as a function of the number of tracers L with $T = 10$ and $N = 20$. Here, the useful prediction is defined as the lead time, within which the RMSE is less than the one standard deviation of the true signal and the Corr is above the threshold Corr $= 0.5$. The yellow and pink circles show the lead time of the useful prediction using the LSTM network with $T = 50, N = 4$ and $T = 200, L = 1$, respectively. The red and green dashed lines show the useful predictions using the perfect and the imperfect models, which do not depend on L when the initial values are perfectly known; (**b**) the Corr as a function of the lead time using different methods.

Appendix C. Additional Results for Predicting the MISO Index

(c,d) of Figure A3 show a simulation from the nearly perfect model (9) and (10). The path-wise behavior of the model resembles the observed MISO index, which is shown in (a,b).

Figure A4 shows the skill scores of the MISO prediction. the training data of the LSTM in this figure is provided by the model simulation. Here, the model simulation means simply running the model (9) and (10) forward for 30 years. There is no observational data incorporated here and there is no conditional sampling either. The following conclusions can be drawn from this figure. If the nearly perfect model is used to provide a massive training data, then the skill scores of the LSTM prediction are comparable to those of the model ensemble forecast using the nearly perfect model. On the other hand, if the imperfect model is used to generate the training data, then the pattern correlation of the LSTM prediction outweighs the imperfect model ensemble forecast. The underlying reason is the following. Since the model is a stochastic model, it is able to generate patterns with different oscillation frequencies. The LSTM forecast then exploits the input information to find the most similar patterns in the training data set. In other words, the LSTM gives different weights to all the possible events in the training data set, which can be regarded as different ensemble members. This is, however, not the case in the ensemble mean based model forecast, where the same weight is assigned to different ensemble members.

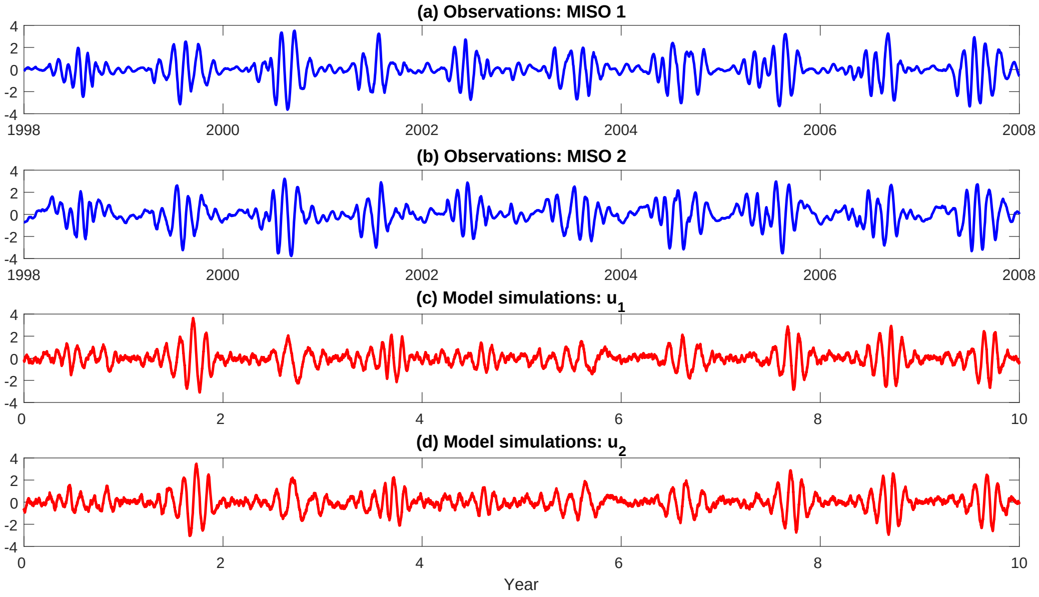

Figure A3. (**a**,**b**) the two components of the observed MISO index (the same as those in Figure 5); (**c**,**d**) a simulation from the nearly perfect model (9) and (10), where the parameters are listed in the top row of Table 1.

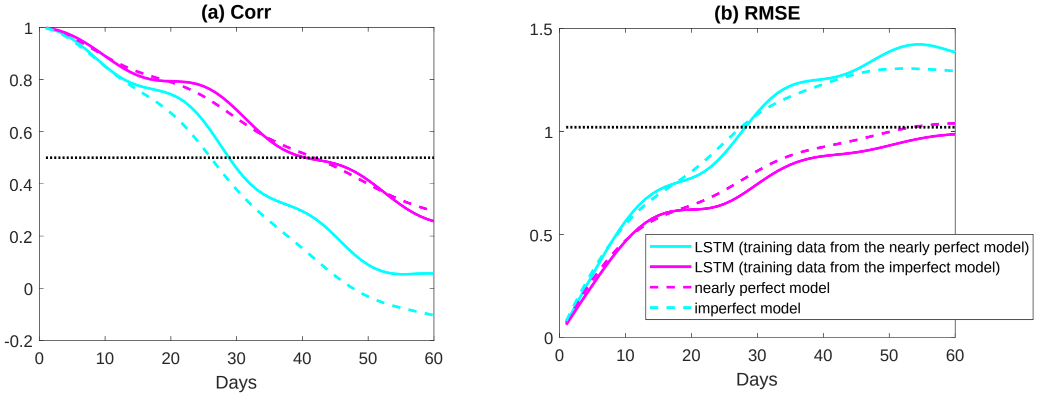

Figure A4. Skill scores of predicting the MISO index. The two panels here are similar to those in Figure 6, but the solid curves here show the training data of the LSTM is from the model simulation with a 30-year length. The dashed curves show the model ensemble forecasts.

Appendix D. The Derivation of the Conditional Sampling Formula

For the convenience of discussion, the statement below starts with a discrete approximation of the original nonlinear continuous system in time (1) by adopting a Euler–Maruyama scheme,

$$\begin{aligned}\widetilde{\mathbf{X}}^{j+1} &= \widetilde{\mathbf{X}}^{j} + \left(\widetilde{\mathbf{A}}_{\mathbf{0}}^{j} + \widetilde{\mathbf{A}}_{\mathbf{1}}^{j}\widetilde{\mathbf{Y}}^{j}\right)\Delta t + \widetilde{\mathbf{B}}_{\mathbf{1}}^{j}\Delta\widetilde{\mathbf{W}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{B}}_{\mathbf{2}}^{j}\Delta\widetilde{\mathbf{W}}_{\mathbf{2}}^{j},\\ \widetilde{\mathbf{Y}}^{j+1} &= \widetilde{\mathbf{Y}}^{j} + \left(\widetilde{\mathbf{a}}_{\mathbf{0}}^{j} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{Y}}^{j}\right)\Delta t + \widetilde{\mathbf{b}}_{\mathbf{1}}^{j}\Delta\widetilde{\mathbf{W}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}_{\mathbf{2}}^{j}\Delta\widetilde{\mathbf{W}}_{\mathbf{2}}^{j},\end{aligned} \tag{A10}$$

Thus, the values of $\mathbf{X}$ and $\mathbf{Y}$ are taken at discrete points in time $\{\widetilde{\mathbf{X}}^0, \ldots, \widetilde{\mathbf{X}}^j, \ldots, \widetilde{\mathbf{X}}^J\}$ and $\{\widetilde{\mathbf{Y}}^0, \ldots, \widetilde{\mathbf{Y}}^j, \ldots, \widetilde{\mathbf{Y}}^J\}$, where $\widetilde{\mathbf{X}}^j := \mathbf{X}(t_j)$ and $\widetilde{\mathbf{Y}}^j = \mathbf{Y}(t_j)$. Here, the variable with tilde and superscript j, namely $\tilde{\cdot}^j$, denotes the discrete approximation of its continuous form at time t_j, where the entire time interval $[0, T]$ is divided into J equipartition subintervals with $0 = t_0, t_1, t_2, \ldots, t_J = T$. Denote $\Delta t = t_{j+1} - t_j$ and therefore $J\Delta t = T$. In the analysis of the system with the discrete approximation, Δt is assumed to be a small value. At the end, the limit $\Delta t \to 0$ will be taken for the discrete approximation to retrieve the original continuous dynamics.

Different from the point-wise state estimation (i.e., filtering or smoothing), which is only conditioned on the observations, calculating the current value $\widetilde{\mathbf{Y}}^j$ on the conditional sampled trajectory requires the given conditions on both the observations $\widetilde{\mathbf{X}}^s, s \leq j$ and the previous sampled point $\widetilde{\mathbf{Y}}^{j+1}$. Therefore, the following Lemma is useful for deriving the conditional sampling formula.

Lemma A1. *The conditional distribution*

$$p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \leq j) \sim \mathcal{N}(\widetilde{\mathbf{m}}^j, \widetilde{\mathbf{P}}^j) \tag{A11}$$

is Gaussian, where the conditional mean $\widetilde{\mathbf{m}}^j$ and conditional covariance $\widetilde{\mathbf{P}}^j$ satisfies the following equations:

$$\widetilde{\mathbf{m}}^j = \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big), \tag{A12a}$$

$$\widetilde{\mathbf{P}}^j = \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j\Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big)(\mathbf{C}^j)^*, \tag{A12b}$$

and the auxiliary matrix $\mathbf{C}$ is given by

$$\widetilde{\mathbf{C}}^j = \widetilde{\mathbf{R}}^j(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j\Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j)^*\big)^{-1}. \tag{A13}$$

Proof. The conditional distribution $p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \leq j)$ can be solved by making use of the joint distribution $p(\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \leq j)$. In light of the second equation in (A10), the marginal distribution $p(\widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \leq j)$ is given by

$$p(\widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \leq j) \sim \mathcal{N}\Big(\widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\tilde{\mu}, \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j\Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\Big). \tag{A14}$$

On the other hand, the marginal distribution $p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{X}}^s, s \leq j)$ is simply given by the filtering formula,

$$p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{X}}^s, s \leq j) \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}^j, \widetilde{\mathbf{R}}^j). \tag{A15}$$

The cross covariance term conditioned on the observations is given by

$$\langle\widetilde{\mathbf{Y}}'^{,j+1}(\widetilde{\mathbf{Y}}'^{,j})^*\rangle|\widetilde{\mathbf{X}}^s, s \leq j = (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j, \tag{A16}$$

where $\widetilde{\mathbf{Y}}'^{,j+1}$ and $\widetilde{\mathbf{Y}}'^{,j}$ are $\widetilde{\mathbf{Y}}^{j+1}$ and $\widetilde{\mathbf{Y}}^{j}$ by removing their mean values. Therefore, collecting (A14)–(A16) leads to

$$\begin{aligned}&p(\widetilde{\mathbf{Y}}^j,\widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s,s\le j)\\ &\sim\mathcal{N}\left(\begin{pmatrix}\widetilde{\boldsymbol{\mu}}^j\\ \widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\end{pmatrix},\begin{pmatrix}\widetilde{\mathbf{R}}^j & \widetilde{\mathbf{R}}^j(1+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\\ (\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j & \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\end{pmatrix}\right).\end{aligned}\tag{A17}$$

In light of the rule of the multivariate Gaussian distribution, the result in (A17) yields the conditional distribution,

$$p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1},\widetilde{\mathbf{X}}^s,s\le J)=p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1},\widetilde{\mathbf{X}}^s,s\le j)=\mathcal{N}(\widetilde{\mathbf{m}}^j,\widetilde{\mathbf{P}}^j),\tag{A18}$$

where

$$\widetilde{\mathbf{m}}^j=\widetilde{\boldsymbol{\mu}}^j+\widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t-(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big),\tag{A19a}$$

$$\widetilde{\mathbf{P}}^j=\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\mathbf{a}_{\mathbf{1}}^j\Delta t)^*\big)(\mathbf{C}^j)^*,\tag{A19b}$$

and the auxiliary matrix $\mathbf{C}$ is given by

$$\widetilde{\mathbf{C}}^j=\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j)^*\big)^{-1}.\tag{A20}$$

Note that the first equality in (A18) is due to the Markovian property of the underlying system [40]. In fact, if $\widetilde{\mathbf{Y}}^{j+1}$ is known, then the conditional distribution of $\widetilde{\mathbf{Y}}^j$ has no dependence on $\widetilde{\mathbf{X}}^s, s\ge j+1$. More specifically, the information of $\widetilde{\mathbf{X}}^s, s>j$ has been included in $\widetilde{\mathbf{Y}}^{j+1}$. The conditional terms $\widetilde{\mathbf{Y}}^{j+1}$ and $\widetilde{\mathbf{X}}^s, s\le j$ represent the information in the future and past, respectively, which can be seen in (A19). This finishes the proof of Lemma A1. □

The result in Lemma A1 allows the derivation of the conditional sampling formula. Each sampled point $\widetilde{\mathbf{Y}}^j$ is conditioned on the previous one $\widetilde{\mathbf{Y}}^{j+1}$ and the smoother estimate. The key step is to make use of the conditional probability in Lemma A1 to draw the sample and finally write down a SDE for an efficient calculation of the conditional sampled path. The proof is shown below.

Proof. Recall from (A18) that $\widetilde{\mathbf{Y}}^{j+1}$ is calculated by generating a multivariate Gaussian random variable with mean $\widetilde{\mathbf{m}}^j$ and covariance $\widetilde{\mathbf{P}}^j$. The backward equation of sampling $\mathbf{Y}$ can be derived by making use of (A19) and (A20).

Plugging (A20) into (A19b) yields

$$\begin{aligned}\widetilde{\mathbf{P}}^j&=\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j\\ &=\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{R}}^j+(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j+O(\Delta t^2)\\ &=\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{R}}^j(\mathbf{I}+(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t)\big)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j+O(\Delta t^2)\\ &=\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\mathbf{I}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j+O(\Delta t^2)\end{aligned}\tag{A21}$$

For notation simplicity, we define

$$\mathbf{F}=(\widetilde{\mathbf{R}}^j)^{-1}\big(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\big),\tag{A22}$$

and thus

$$\begin{aligned}
&\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\mathbf{I}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j\\
=&\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\mathbf{I}-\mathbf{F}\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j\\
=&(\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)((\widetilde{\mathbf{R}}^j)^{-1}-\mathbf{F}(\widetilde{\mathbf{R}}^j)^{-1}\Delta t)(\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{a}}_{\mathbf{j}}^j\widetilde{\mathbf{R}}^j\Delta t)\\
=&(\mathbf{I}-\widetilde{\mathbf{R}}^j\mathbf{F}(\widetilde{\mathbf{R}}^j)^{-1}+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j(\widetilde{\mathbf{R}}^j)^{-1}\Delta t)(\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j\Delta t)\\
=&\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j\mathbf{F}\Delta t+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j\Delta t+O(\Delta t^2).
\end{aligned}\tag{A23}$$

Plugging (A23) back to (A21) yields

$$\begin{aligned}
\widetilde{\mathbf{P}}^j&=\widetilde{\mathbf{R}}^j-\Big(\widetilde{\mathbf{R}}^j-\widetilde{\mathbf{R}}^j\mathbf{F}\Delta t+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j\Delta t\Big)+O(\Delta t^2)\\
&=\widetilde{\mathbf{R}}^j\mathbf{F}\Delta t-\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t-\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j\Delta t+O(\Delta t^2)\\
&=\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+O(\Delta t^2)
\end{aligned}\tag{A24}$$

Applying a similar argument, plugging (A20) into (A19a) yields and subtracting $\widetilde{\mathbf{Y}}^{j+1}$ on both sides of (A19a) yields

$$\begin{aligned}
\widetilde{\mathbf{m}}^j-\widetilde{\mathbf{Y}}^{j+1}&=\widetilde{\boldsymbol{\mu}}^j+\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t+(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j)^*\big)^{-1}\\
&\qquad\qquad\times\big(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t-(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big)-\widetilde{\mathbf{Y}}^{j+1}\\
&=\widetilde{\boldsymbol{\mu}}^j+\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)^*\big(\mathbf{I}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}\\
&\qquad\qquad\times\big(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t-(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big)-\widetilde{\mathbf{Y}}^{j+1}+O(\Delta t^2)\\
&=\widetilde{\boldsymbol{\mu}}^j+(\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\big((\widetilde{\mathbf{R}}^j)^{-1}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{R}}^j+\widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)(\widetilde{\mathbf{R}}^j)^{-1}\Delta t\big)\\
&\qquad\qquad\times\big(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\mathbf{a}}_{\mathbf{0}}^j\Delta t-(\mathbf{I}+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big)-\widetilde{\mathbf{Y}}^{j+1}+O(\Delta t^2)\\
&=\widetilde{\boldsymbol{\mu}}^j+\big(\mathbf{I}-(\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)\big(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\boldsymbol{\mu}}^j-(\widetilde{\mathbf{a}}_{\mathbf{0}}^j+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\boldsymbol{\mu}}^j)\Delta t\big)-\widetilde{\mathbf{Y}}^{j+1}+O(\Delta t^2)\\
&=-(\widetilde{\mathbf{a}}_{\mathbf{0}}^j+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\boldsymbol{\mu}}^j)\Delta t-(\widetilde{\mathbf{a}}_{\mathbf{1}}^j+\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j(\widetilde{\mathbf{R}}^j)^{-1})(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\boldsymbol{\mu}}^j)\Delta t+O(\Delta t^2)\\
&=-(\widetilde{\mathbf{a}}_{\mathbf{0}}^j+\widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{Y}}^{j+1})\Delta t-\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{Y}}^{j+1}-\widetilde{\boldsymbol{\mu}}^j)\Delta t+O(\Delta t^2)
\end{aligned}\tag{A25}$$

Combining (A23) and (A25) and taking the limit $\Delta t \to 0$ yields an explicit formula of sampling the unobserved processes $Z(t)$,

$$\overleftarrow{\mathrm{d}\mathbf{Y}}=\big(-\mathbf{a_0}-\mathbf{a_1}\mathbf{Y}\big)\,\mathrm{d}t+(\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}\big(\boldsymbol{\mu}-\mathbf{Y}\big)dt+\big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_{\mathbf{Y}},\tag{A26}$$

which is Corollary A1. Subtracting the smoother posterior mean from (A26) yields

$$\overleftarrow{\mathrm{d}\mathbf{Y}}=\overleftarrow{\mathrm{d}\boldsymbol{\mu}_{\mathbf{s}}}-\big(\mathbf{a_1}+(\mathbf{b}\circ\mathbf{b})\mathbf{R}_{\mathbf{f}}^{-1}\big)(\mathbf{Y}-\boldsymbol{\mu}_{\mathbf{s}})\,\mathrm{d}t+\big(\mathbf{b}\circ\mathbf{b}\big)^{1/2}\,\mathrm{d}\mathbf{W}_{\mathbf{Y}},\tag{A27}$$

which is Theorem 3. □

References

1. Perretti, C.T.; Munch, S.B.; Sugihara, G. Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5253–5257. [CrossRef] [PubMed]
2. Brunton, S.L.; Kutz, J.N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*; Cambridge University Press: Cambridge, MA, USA, 2019.
3. Wan, Z.Y.; Sapsis, T.P. Machine learning the kinematics of spherical particles in fluid flows. *J. Fluid Mech.* **2018**, *857*, R2-1–R2-11. [CrossRef]

4. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Not Enough Data? Deep Learning to the Rescue! *arXiv* **2019**, arXiv:1911.03118.
5. Wong, S.C.; Gatt, A.; Stamatescu, V.; McDonnell, M.D. Understanding data augmentation for classification: when to warp? In Proceedings of the 2016 IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, Australia, 30 November–2 December 2016; pp. 1–6.
6. Wei, J.W.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
7. Ham, Y.G.; Kim, J.H.; Luo, J.J. Deep learning for multi-year ENSO forecasts. *Nature* **2019**, *573*, 568–572. [CrossRef]
8. O'Gorman, P.A.; Dwyer, J.G. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.* **2018**, *10*, 2548–2563. [CrossRef]
9. Kalnay, E. *Atmospheric Modeling, Data Assimilation and Predictability*; Cambridge University Press: Cambridge, MA, USA, 2003.
10. Evensen, G. *Data Assimilation: The Ensemble Kalman Filter*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
11. Majda, A.J.; Harlim, J. *Filtering Complex Turbulent Systems*; Cambridge University Press: Cambridge, MA, USA, 2012.
12. Law, K.; Stuart, A.; Zygalakis, K. *Data Assimilation: A Mathematical Introduction*; Springer: New York, NY, USA, 2015.
13. Brajard, J.; Carassi, A.; Bocquet, M.; Bertino, L. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *arXiv* **2020**, arXiv:2001.01520.
14. Liptser, R.S.; Shiryaev, A.N. *Statistics of Random Processes II: Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
15. Chen, N.; Majda, A. Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy* **2018**, *20*, 509. [CrossRef]
16. Chen, N. Improving the Prediction of Complex Nonlinear Turbulent Dynamical Systems Using Nonlinear Filter, Smoother and Backward Sampling Techniques. *Res. Math. Sci.* **2020**, *7*, 1–39.
17. Chen, N.; Majda, A.J. Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 12864–12869. [CrossRef]
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
19. Gould, J.; Roemmich, D.; Wijffels, S.; Freeland, H.; Ignaszewsky, M.; Xu, J.; Pouliquen, S.; Desaubies, Y.; Send, U.; Radhakrishnan, K.; et al. Argo profiling floats bring new era of in situ ocean observations. *Eos Trans. Am. Geophys. Union* **2004**, *85*, 185–191. [CrossRef]
20. LaCasce, J. Statistics from Lagrangian observations. *Prog. Oceanogr.* **2008**, *77*, 1–29. [CrossRef]
21. Apte, A.; Jones, C.K.; Stuart, A. A Bayesian approach to Lagrangian data assimilation. *Tellus Dyn. Meteorol. Oceanogr.* **2008**, *60*, 336–347. [CrossRef]
22. Majda, A. *Introduction to PDEs and Waves for the Atmosphere and Ocean*; American Mathematical Soc.: Providence, RI, USA, 2003.
23. Vallis, G.K. *Atmospheric and Oceanic Fluid Dynamics*; Cambridge University Press: Cambridge, MA, USA, 2017.
24. Chen, N.; Majda, A.J.; Tong, X.T. Noisy Lagrangian tracers for filtering random rotating compressible flows. *J. Nonlinear Sci.* **2015**, *25*, 451–488. [CrossRef]
25. Majda, A.J. *Introduction to Turbulent Dynamical Systems in Complex Systems*; Springer: New York, NY, USA, 2016.
26. Farrell, B.F.; Ioannou, P.J. Stochastic forcing of the linearized Navier–Stokes equations. *Phys. Fluids Fluid Dyn.* **1993**, *5*, 2600–2609. [CrossRef]
27. Geurts, B.J.; Kuerten, J.G. Ideal stochastic forcing for the motion of particles in large-eddy simulation extracted from direct numerical simulation of turbulent channel flow. *Phys. Fluids* **2012**, *24*, 081702. [CrossRef]
28. Chen, N.; Majda, A.J.; Tong, X.T. Information barriers for noisy Lagrangian tracers in filtering random incompressible flows. *Nonlinearity* **2014**, *27*, 2133. [CrossRef]
29. Fritts, D.C.; Vanzandt, T.E. Spectral estimates of gravity wave energy and momentum fluxes. Part I: Energy dissipation, acceleration, and constraints. *J. Atmos. Sci.* **1993**, *50*, 3685–3694. [CrossRef]

30. Sikka, D.; Gadgil, S. On the maximum cloud zone and the ITCZ over Indian, longitudes during the southwest monsoon. *Mon. Weather Rev.* **1980**, *108*, 1840–1853. [CrossRef]
31. Goswami, B.; Mohan, R.A. Intraseasonal oscillations and interannual variability of the Indian summer monsoon. *J. Clim.* **2001**, *14*, 1180–1198. [CrossRef]
32. Lau, W.K.M.; Waliser, D.E. *Intraseasonal Variability in the Atmosphere-Ocean Climate System*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
33. Sabeerali, C.; Ajayamohan, R.; Giannakis, D.; Majda, A.J. Extraction and prediction of indices for monsoon intraseasonal oscillations: An approach based on nonlinear Laplacian spectral analysis. *Clim. Dyn.* **2017**, *49*, 3031–3050. [CrossRef]
34. Huffman, G.J.; Adler, R.F.; Morrissey, M.M.; Bolvin, D.T.; Curtis, S.; Joyce, R.; McGavock, B.; Susskind, J. Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorol.* **2001**, *2*, 36–50. [CrossRef]
35. Chen, N.; Majda, A.J.; Sabeerali, C.; Ajayamohan, R. Predicting monsoon intraseasonal precipitation using a low-order nonlinear stochastic model. *J. Clim.* **2018**, *31*, 4403–4427. [CrossRef]
36. Cabos, W.; de la Vara, A.; Koseki, S. Tropical Atlantic variability: Observations and modeling. *Atmosphere* **2019**, *10*, 502. [CrossRef]
37. Sabeerali, C.; Ajayamohan, R.; Bangalath, H.K.; Chen, N. Atlantic Zonal Mode: An Emerging Source of Indian Summer Monsoon Variability in a Warming World. *Geophys. Res. Lett.* **2019**, *46*, 4460–4467. [CrossRef]
38. Majda, A.J.; Harlim, J. Physics constrained nonlinear regression models for time series. *Nonlinearity* **2012**, *26*, 201. [CrossRef]
39. Harlim, J.; Mahdi, A.; Majda, A.J. An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *J. Comput. Phys.* **2014**, *257*, 782–812. [CrossRef]
40. Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press: Cambridge, MA, USA, 2013.