

# A Physics-Informed Data-Driven Algorithm for Ensemble Forecast of Complex Turbulent Systems

Nan Chen<sup>a</sup> and Di Qi<sup>b</sup>

<sup>a</sup> Department of Mathematics, University of Wisconsin-Madison, 480 Lincoln Drive, Madison, 53706, WI, USA

<sup>b</sup> Department of Mathematics, Purdue University, 150 North University Street, West Lafayette, 47907, IN, USA

## Abstract

A new ensemble forecast algorithm, named as the physics-informed data-driven algorithm with conditional Gaussian statistics (PIDD-CG), is developed to predict the time evolution of the probability density functions (PDFs) of complex turbulent systems with partial observations. The PIDD-CG algorithm integrates a unique multiscale statistical closure model with an extremely efficient nonlinear data assimilation scheme to represent the PDF as a mixture of conditional statistics, which overcomes the curse of dimensionality for high-dimensional systems. The multiscale features in the time evolution of each conditional statistics ensemble member effectively captured by an appropriate combination of physics-informed analytic formulae and recurrent neural networks. An information metric is adopted as the loss function for the latter to more accurately calibrate the key turbulent signals with strong fluctuations. The proposed algorithm succeeds in forecasting both the transient and statistical equilibrium non-Gaussian PDFs of strongly turbulent systems with intermittency, regime switching and extreme events.

**Keywords**— turbulent systems, multiscale statistical closure model, conditional Gaussian mixture, recurrent neural network, information metric

## 1 Introduction

Complex turbulent systems are ubiquitous in many fields, such as geophysics, climate science, neural science, engineering, and plasma physics [44, 43, 17]. These systems contain rich nonlinear dynamics and statistical features, including multiscale structures, intermittency, extreme events, regime switching, and strong non-Gaussian probability density functions (PDFs) [26, 16, 49, 28, 45]. Predicting the future states of complex turbulent systems is a central challenge in contemporary science with large societal impacts. Due to the turbulent nature of such systems, the *trajectory forecast* based on a single realization of the model state quickly loses track of the truth. Alternatively, the *ensemble forecast*, which adopts a probabilistic characterization of the model states utilizing a Monte Carlo (MC) type approach, is the predominant strategy in predicting complex turbulent systems in practice [34, 46, 24]. In the ensemble forecast, different ensemble members are sampled from an initial conditions and are subject to different random forcing, accounting for the uncertainty in the initialization and model errors, respectively. The ensemble forecast aims at providing an indication of the PDF of possible future states by tracking the evolution of the group of ensemble members.

Despite the simplicity of the general framework, there exists a major computational challenge in applying the traditional ensemble forecast method to realistic scenarios. In fact, as the dimension of the system becomes large, an exponential increase of the ensemble size is needed to maintain the accuracy of the forecast PDF, which is known as the curse of dimensionality [11]. However, since the computational cost of each single model realization also shoots up significantly as the dimension of the system increases, only a small ensemble size is affordable in practical situations, such as climate and weather forecast [18, 13]. As a result, although the traditional ensemble forecast method can ideally provide a reasonably accurate characterization of the mean state, the lack of a sufficient number of samples makes it extremely difficult to accurately forecast the intrinsic uncertainty of the system. Especially, the direct ensemble method often fails to capture the non-Gaussian joint PDF in high dimensions, and thus leads to large errors in characterizing many key turbulent phenomena, such as regime switching, intermittency and extreme events [28, 15]. A similar issue occurs at the initialization stage of the forecast. Since only the time series of a subset of the state variables can be observed in most of the practical problems (known as partial observations), ensemble data assimilation is often required for the state estimation of the unobserved variables [14, 2]. Yet, the error in

quantifying the initial uncertainty due to the lack of a sufficient number of samples can be rapidly amplified in the subsequent forecast, leading to large biases in predicting even short-term transient features.

In this paper, a new physics-informed data-driven conditional Gaussian (PIDD-CG) algorithm is developed that aims at efficiently and accurately forecasting the key non-Gaussian PDF for a wide class of high-dimensional complex systems. The PIDD-CG algorithm starts with a phase space decomposition by projecting the model states into a low-dimensional subspace containing the observed state variables and a remaining multiscale high-dimensional subspace. A systematic multiscale data-driven closure approximation is developed in the low-dimensional subspace, with which a small number of samples is sufficient to characterize the associated uncertainty propagations. The PIDD-CG algorithm then exploits an effective physics-based decomposition of the PDF in the high-dimensional subspace into a conditional Gaussian mixture and integrates the evolution equations of the conditional statistics associated with each mixture component to obtain the forecast PDF [8]. A continuous data assimilation scheme is used to determine the characteristics in each conditional Gaussian component that is associated with one realization in the low-dimensional observed state [6]. This captures the correlation between the two nonlinearly coupled subspaces. There are several remarkable advantages of the PIDD-CG algorithm. First, by creating the mixture distribution the algorithm does not suffer from the curse of dimensionality with respect to the number of mixture components [10]. In fact, fundamentally different from the purely data-driven approaches that often require a large number of samples, the development of the mixture distribution in the PIDD-CG algorithm uses only the same small number of samples in characterizing the associated low-dimensional subspace and is sufficient to represent the high-dimensional full PDF thanks to the conditional mixture. Second, the governing equations of the time-evolution of the conditional statistics have closed analytic formulae [25], which further reduce the computational cost and avoid the direct ensemble approximation of obtaining such statistical moments that are commonly required in applying ensemble simulations.

Yet, despite the analytically solvable properties, the computational cost of solving the governing equations of the full conditional statistics can still be demanding due to the existence of a large number of complicated nonlinear terms involving a wide spectrum of multiscale fluctuating variables, especially those associated with unresolved scales. Therefore, the PIDD-CG algorithm proposes a balanced physics-informed and data-driven construction of these governing equations, aiming at explicitly preserving the crucial dynamical structure while using data-driven approaches to effectively forecast the complicated unresolved details with a much lower computational cost. Specifically, the PIDD-CG algorithm approximates the complicated nonlinear feedbacks in these governing equations by a recurrent neural network (RNN) [50]. Since the neural network aims at predicting the statistics, a simple but effective information metric is adopted as the loss function to train the RNN [39, 21], which significantly outweighs the traditional loss functions that are based on minimizing the path-wise errors. Finally, in light of the evolution equations of the conditional statistics, the PIDD-CG algorithm naturally provides a systematic way of developing statistical reduced order models [29, 36], in which the feedback from the unresolved-scale variables is approximated by the RNNs. This further enhances the forecast efficiency for complex systems with very large dimensions in practice when the primary interest lies in the statistical forecast of certain large-scale modes.

In the rest of the paper, the PIDD-CG algorithm is illustrated based on a prototype model in geophysical turbulence: the topographic barotropic model, which displays many representative turbulent features, including extreme events and switching regimes [27]. The method for a general group of nonlinear systems is described in *Methods* Section 4. More detailed results including a coupled dyad model as a proof-of-concept and a complete analysis of the computational performance are listed in the Supplementary Information (SI).

## 2 Results

### 2.1 The PIDD-CG algorithm

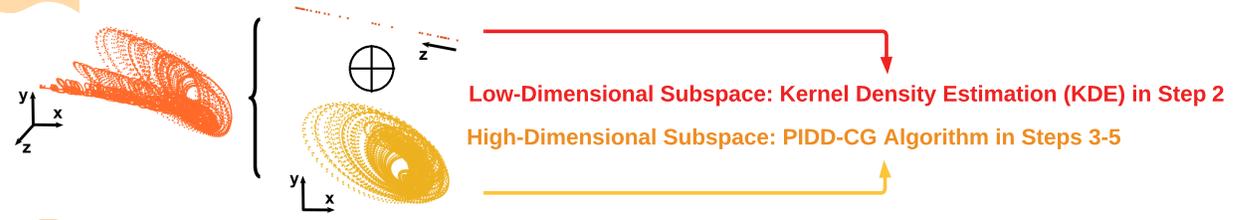
The general framework of the PIDD-CG algorithm is schematically illustrated in Figure 1, which consists of five key steps.

**Step 1. Phase space decomposition.** A phase space decomposition is carried out to project the state variables into two subspaces admitting conditional Gaussian structures [6]. A collection of the leading state variables  $\mathbf{X}$  (such as the large-scale, resolved or observed states) belongs to this relatively low-dimensional subspace. The rest of the variables  $\mathbf{Y}$ , which are multiscale, unresolved or unobserved, are contained in the remaining high-dimensional subspace.

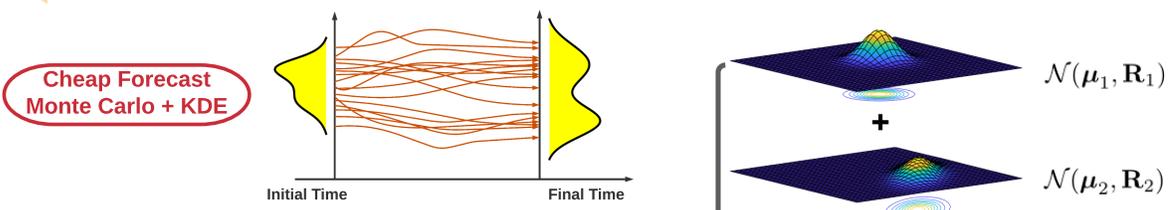
**Step 2. Systematic multiscale statistical closure of the large-scale dynamics.** A systematic multiscale statistical closure approximation is proposed to avoid running the original full-dimensional system when predicting the statistics of the low-dimensional variable  $\mathbf{X}$  that is fully coupled with  $\mathbf{Y}$ . The statistical closure of the equations of  $\mathbf{X}$  depends explicitly on the conditional mean of only a few modes of  $\mathbf{Y}$  conditioned on realizations of  $\mathbf{X}$ , which can be solved via closed analytic formulae (see *Step 4*), while the residual part is effectively approximated by a RNN (see *Step 5*). Such a unique way of building the closure model allows to forecast the statistics of  $\mathbf{X}$  from an intrinsically low-dimensional subsystem, which requires only a small number of samples. Denote such a number by  $J$ . The forecast PDF of the low-dimensional variable  $\mathbf{X}$  is then approximated by a kernel density estimation [4] using Gaussian kernels.

# Overview of the PIDD-CG Ensemble Forecast Algorithm for Turbulent Systems with Partial Observations

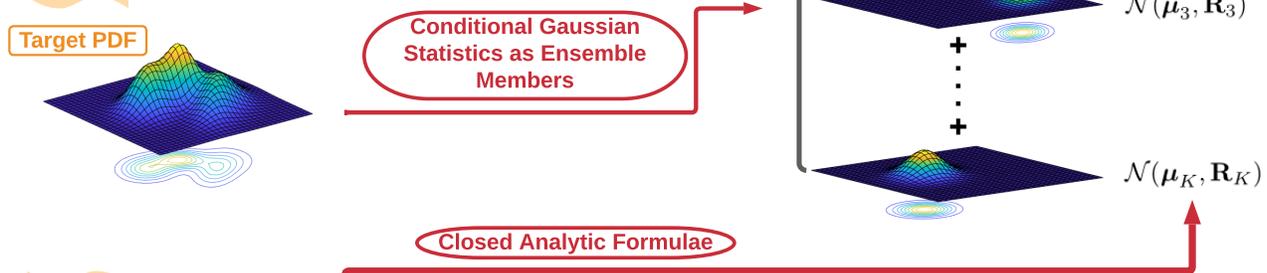
## Step 1. Phase Space Decomposition



## Step 2. Systematic Multiscale Statistical Closure Approximation of the Dynamics in the Low-Dimensional Subspace



## Step 3. Effective Physics-Informed Conditional Gaussian Mixture via Data Assimilation



## Step 4. Time Evolution of Conditional Statistics

$$\left. \begin{aligned} \frac{d\mu_k}{dt} \\ \frac{d\mathbf{R}_k}{dt} \end{aligned} \right\} = \text{Basic Dynamics} + \boxed{\text{Complicated Nonlinear Interactions}} \quad (\text{for } k = 1, \dots, K)$$

## Step 5. Effective Approximation Utilizing Recurrent Neural Network

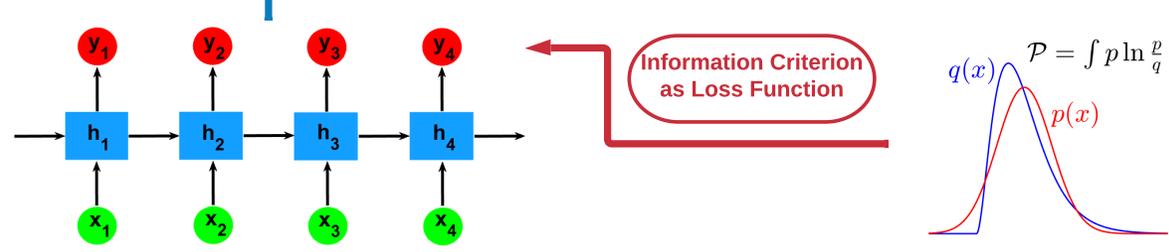


Figure 1: Schematic diagram of the PIDD-CG algorithm for predicting the PDF of the high-dimensional complex turbulent systems.

**Step 3. Effective physics-informed conditional Gaussian mixture via data assimilation.** Given the physical model formulation conditioned on each of the  $J$  forecast trajectories of  $\mathbf{X}$  from *Step 2*, the conditional distribution of  $\mathbf{Y}$  can be computed via a nonlinear data assimilation method [6]. These  $J$  conditional distributions are the  $J$  conditional statistics ensemble members. Notably, the center and the bandwidth associated with each conditional statistics ensemble member are automatically optimized by the nonlinear data assimilation that takes into account the model physics. Because of this, the resulting mixture distribution consisting of the conditional statistics ensemble avoids the curse of dimensionality with respect to the number of mixture components [10]. In addition, since each sample of  $\mathbf{Y}$  is conditioned on one realization of  $\mathbf{X}$ , the cross-correlation between  $\mathbf{X}$  and  $\mathbf{Y}$  is also captured when forming their joint distribution.

**Step 4. Analytic formulae for the time evolution of the conditional statistics in smaller-scale dynamics.** One important feature of the PIDD-CG algorithm is that the time evolution of each conditional statistics ensemble member from *Step 3* can be solved via closed analytic formulae [25]. Thus, it avoids using the expensive MC methods for finding such data assimilation solutions and prevents the sampling errors when handling high-dimensional systems.

**Step 5. Data-driven modeling of the nonlinear feedbacks in conditional statistics with information theory.** Despite the closed analytic formulae, the computational cost of running the evolution equations of the conditional statistics, especially the time evolution of the conditional covariance, can still be demanding for high-dimensional systems. To improve computational efficiency, a model reduction strategy is applied to approximate certain complicated nonlinear components in the unresolved fluctuation modes feedbacks using a recurrent neural network (RNN). Since the output variables are associated with the conditional distribution, a simple but effective information loss function [39] is adopted as a natural metric to train the RNN.

The PIDD-CG algorithm also allows an efficient and accurate data assimilation scheme to obtain the conditional statistics ensemble at the forecast initialization stage (see the *Methods* Section 4), which facilitates the application of the algorithm to the more realistic situations with only partial observations.

## 2.2 The topographic barotropic model

The topographic barotropic flow is a prototype model in geophysics [27], which involves multiscale interactions and transport among the zonal mean flow and the fluctuations. It also contains many key features of interests in turbulence, such as emerging non-Gaussian PDFs, regime switching and extreme events.

The spectral formulation of the model with layered topography reads (see SI for the derivations):

$$\frac{dU}{dt} = \sum_k \hat{h}_k^* \hat{v}_k - d_0 U + \sigma_0 \dot{W}_0, \quad (1a)$$

$$\frac{d\hat{v}_k}{dt} = [-\gamma_{v,k}(U) + i\omega_{v,k}(U)] \hat{v}_k - l_x^2 \hat{h}_k U - d_{v,k} \hat{v}_k + \sigma_{v,k} \dot{W}_k, \quad (1b)$$

$$\frac{d\hat{T}_k}{dt} = [-\gamma_{T,k}(U) + i\omega_{T,k}(U)] \hat{T}_k - d_{T,k} \hat{T}_k - \alpha \hat{v}_k. \quad (1c)$$

In (1), the wavenumbers are given by  $k\mathbf{l}$ ,  $k = \pm 1, \dots, \pm K$ , expanded along one characteristic direction  $\mathbf{l} = (l_x, l_y)$  with  $|\mathbf{l}| = 1$ . The state variable  $U$  represents the large-scale zonal mean flow velocity while  $\hat{v}_k$ ,  $\hat{T}_k$  and  $h_k$  are the coefficients of the  $k$ -th Fourier modes corresponding to the fluctuation components of the flow velocity  $v$ , the turbulent transport of passive tracer field  $T$ , and the topography  $h$ , respectively. The notation  $\cdot^*$  stands for the complex conjugate while  $\dot{W}_0$  and  $\dot{W}_k$  are independent white noise sources with strengths  $\sigma_0$  and  $\sigma_k$ . The model parameters  $\gamma_{v,k}$ ,  $\gamma_{T,k}$  and  $\omega_{v,k}$ ,  $\omega_{T,k}$  represent the dispersion and dissipation effects. The details of the parameter values are included in the SI.

The mean flow  $U$  is driven by the topographic stress from  $h$  combining all the feedbacks from the fluctuations  $v$ , while  $U$  also inversely contributes to each spectral mode through the nonlinear advection and topographic effect. The coupling between the mean flow and the fluctuations is through the topographic stress. The PIDD-CG algorithm is applied to two regimes with distinct dynamical and statistical features. Depending on the statistical equilibrium distributions of  $U$ ,  $\hat{v}_k$  and  $\hat{T}_k$  [38], these two regimes are named as:

- *Strongly non-Gaussian regime:* The zonal mean flow  $U$  is driven by strong white noise forcing while only small noises are added to the fluctuation modes  $\hat{v}_k$ .
- *Near-Gaussian regime:* The fluctuation modes  $\hat{v}_k$  are subject to strong white noise forcing while the noise strength in the zonal mean flow  $U$  is relatively weak.

Figure 2 illustrates dynamical features of both regimes. In the non-Gaussian regime, the large white noise forcing in  $U$  excites a strong competition between two alternating states: a highly intermittent flow field  $v$  when the eastward jet appears ( $U > 0$ ) and a nearly steady flow when the westward jet occurs ( $U < 0$ ). The intermittent nature of the flow field triggers

non-Gaussian fat-tailed distributions of  $v$  and  $T$  (see e.g., the blue curves in Figure 4). On the other hand, in the near-Gaussian regime, strongly multiscale features emerge in the time series of all the variables, which include multiple fast scales with rapid oscillations and a slowly varying long-term tendency. In particular, there are two distinct dominant frequencies in the fast oscillations. The extremely fast oscillation appears when the zonal flow goes steadily towards the west ( $U < 0$ ) while the moderately fast one occurs when the zonal flow becomes intermittent with an average eastward velocity ( $U > 0$ ). In contrast to the non-Gaussian regime, the near-Gaussian statistics in this regime is due to the comparable amplitude of  $\hat{v}_k$  and  $\hat{T}_k$  at different states.

In the following forecast tests, the zonal mean flow  $U$  is assumed to be the only observed variable.

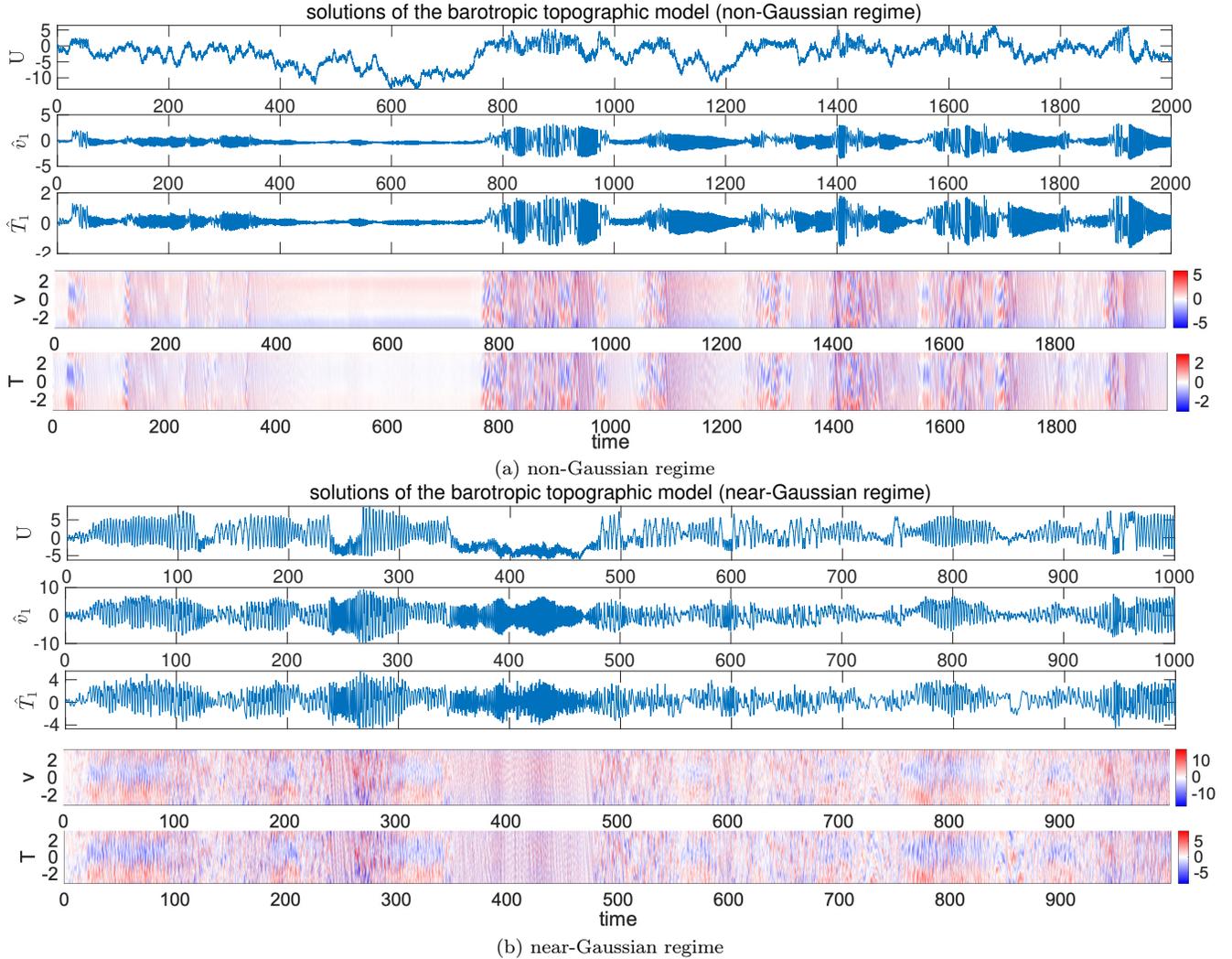


Figure 2: Solutions of the barotropic topographic model with in total 41 modes (i.e.,  $K = 10$ ). Panel (a) shows the solution in the non-Gaussian regime while Panel (b) shows that in the near-Gaussian regime. In each panel, the zonal velocity  $U$ , the real parts of the first fluctuation mode  $\hat{v}_1$  and  $\hat{T}_1$ , and the entire fields of  $v$  and  $T$  in physical space as a function of time are presented.

### 2.3 Predicting key PDFs in the barotropic topographic model using the PIDD-CG algorithm

In *Step 1* of the PIDD-CG algorithm described in Section 2.1, the entire phase space is decomposed into a low-dimensional subspace that contains only the observed variable, namely the zonal mean flow  $U$ , and a high-dimensional subspace that includes all the fluctuation modes  $\hat{v}_k$  and  $\hat{T}_k$  for  $k = \pm 1, \dots, \pm K$ . In *Step 2*, only the explicit equations of the conditional

	non-Gaussian regime				near-Gaussian regime		
	$t = 0.5$	$t = 1$	$t = 1.5$	$t = 2$	$t = 0.2$	$t = 0.5$	$t = 1$
$U$	$6.80 \times 10^{-3}$	$4.74 \times 10^{-3}$	$7.55 \times 10^{-3}$	$1.76 \times 10^{-2}$	$1.58 \times 10^{-2}$	$1.68 \times 10^{-2}$	$1.85 \times 10^{-2}$
$\hat{v}_1$	$4.56 \times 10^{-4}$	$1.68 \times 10^{-3}$	$4.09 \times 10^{-3}$	$4.93 \times 10^{-3}$	$7.13 \times 10^{-4}$	$4.76 \times 10^{-3}$	$4.06 \times 10^{-3}$
$\hat{v}_2$	$2.89 \times 10^{-4}$	$4.19 \times 10^{-4}$	$9.74 \times 10^{-4}$	$4.48 \times 10^{-3}$	$1.44 \times 10^{-3}$	$1.27 \times 10^{-2}$	$4.14 \times 10^{-2}$
$\hat{T}_1$	$9.21 \times 10^{-4}$	$4.52 \times 10^{-3}$	$5.11 \times 10^{-3}$	$9.58 \times 10^{-3}$	$4.03 \times 10^{-4}$	$3.28 \times 10^{-3}$	$2.55 \times 10^{-3}$
$\hat{T}_2$	$3.40 \times 10^{-4}$	$4.05 \times 10^{-4}$	$4.08 \times 10^{-4}$	$2.91 \times 10^{-3}$	$4.74 \times 10^{-4}$	$1.41 \times 10^{-3}$	$2.21 \times 10^{-3}$

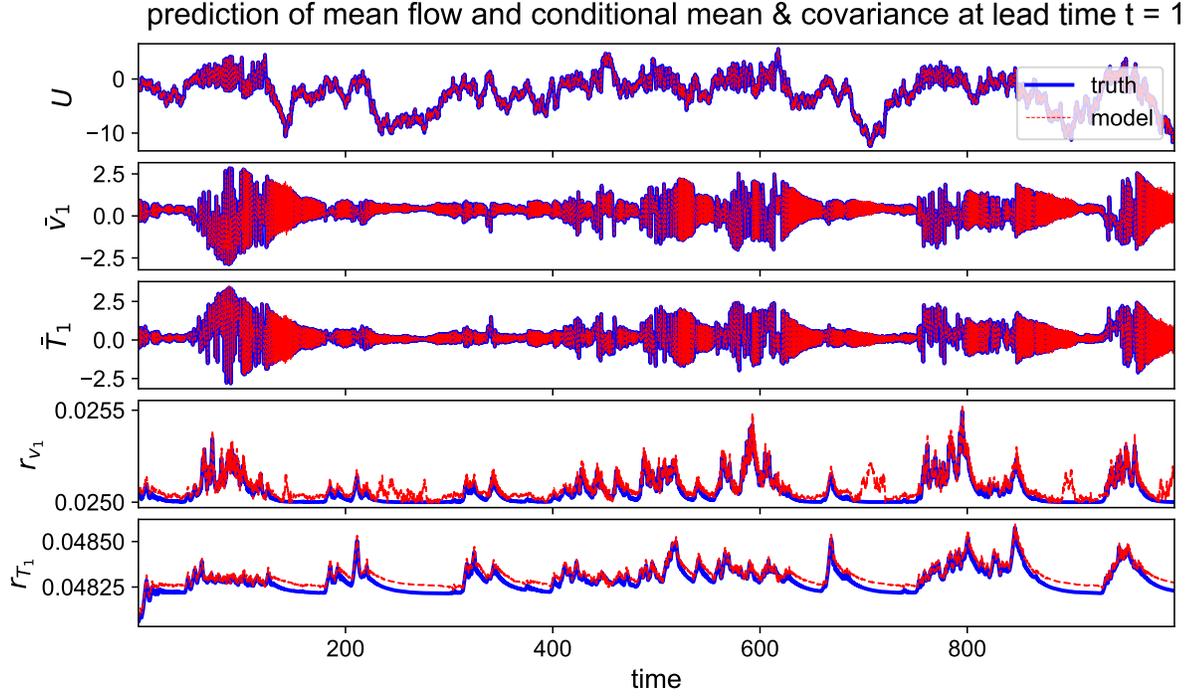
Table 1: Information error (relative entropy) between the truth and forecast PDF in  $U$ ,  $\hat{v}_1$ ,  $\hat{v}_2$ ,  $\hat{T}_1$  and  $\hat{T}_2$  of the barotropic topographic model at different lead time before arriving at the statistical equilibrium state.

means of the leading two complex modes of  $\hat{v}_k$  (e.g.,  $k = \pm 1$  and  $\pm 2$ ) are utilized to build the closure model of  $U$  while all the remaining small-scale feedbacks are automatically learned by the RNN. This accounts for a strongly reduced model of 2 resolved modes compared with the full model with  $K = 10$ . *Steps 3-5* follow directly the description in Section 2.1 with the detailed step-by-step explanations being included in the Method Section and the SI.

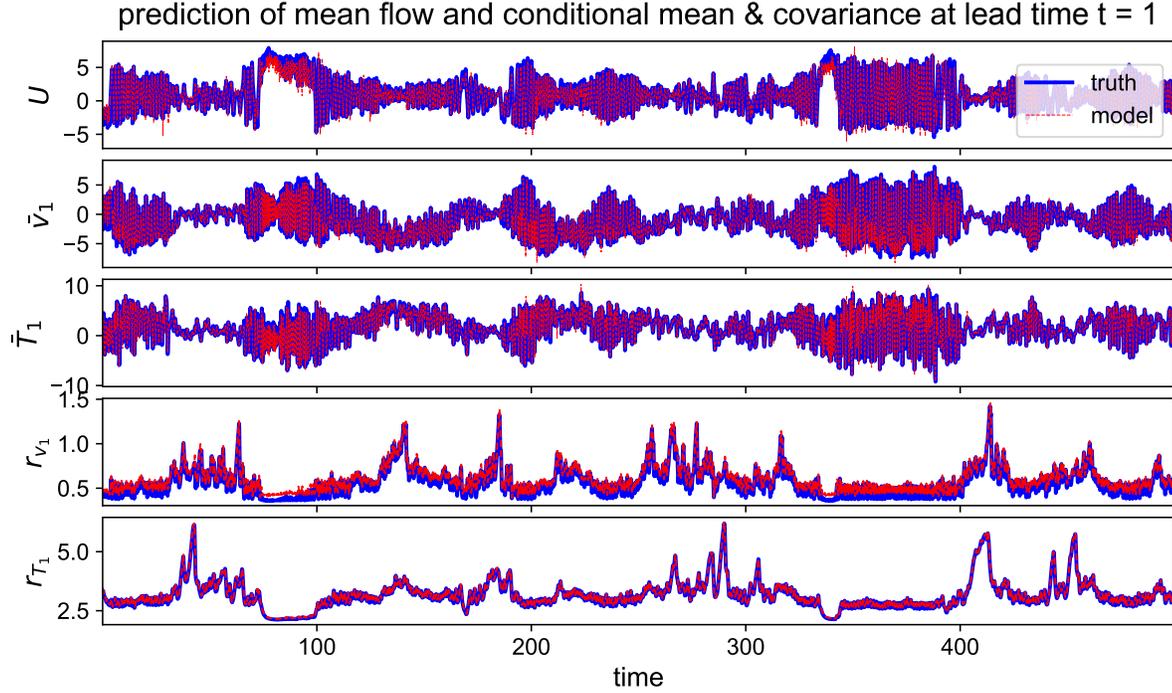
Since the joint PDF from the PIDD-CG algorithm is given by a mixture distribution, where each mixture component is uniquely determined by  $U$  and the conditional statistics of  $\hat{v}_k$  and  $\hat{T}_k$ , it is natural to start with the study of the forecast of these quantities. Figure 3 compares the forecast trajectories at the lead time  $t = 1$  with the truth. The lead time forecast here means each point in the forecast trajectory is the forecast value starting from 1 unit prior to it. Such a lead time is around the decorrelation time of the first a few modes of  $\hat{v}_k$  and  $\hat{T}_k$  and is comparable with the time scale of the fast component of  $U$  (see the SI). Note that the forecast at a lead around the decorrelation time of a turbulent system is extremely challenging if the entire dynamics is fully approximated by the neural network due to the quick accumulation of errors [39, 36]. Nevertheless, basic dynamics are retained in the PIDD-CG algorithm while the neural network plays the supplementary role of only modeling the unresolved residual part. Therefore, good agreements between the truth and the forecast are achieved in the zonal mean flow  $U$  as well as the conditional statistics. In particular, the multiscale structures are accurately predicted in both regimes despite the small errors in forecasting the highly oscillating small scales, which are the unpredictable part due to the turbulent nature. More importantly, both the locations and the amplitudes of the dominant intermittent features are successfully predicted by the PIDD-CG algorithm, which is a central prerequisite for accurately predicting the joint PDFs.

Figure 4 shows the main prediction results for the true PDFs and the forecast ones. The forecast starts from one specific observed initial value of  $U$  while the initial values of  $\hat{v}_k$  and  $\hat{T}_k$  are recovered from the efficient and accurate data assimilation algorithm, which is described in the Method Section 4. In each panel, the one-dimensional PDFs of the zonal mean flow  $U$  and the leading fluctuation modes  $\hat{v}_1$  and  $\hat{T}_1$  as well as the two-dimensional joint PDFs between these variables are presented. A complete comparison of the second fluctuation modes and other joint distributions at several different lead times can be found in the SI. Here, the true PDFs are generated from a direct MC simulation of the topographic barotropic flow model (1). The MC simulation contains  $J_{MC} = 50000$  samples to fully characterize the non-Gaussian statistics, which is computationally very expensive. In contrast, the PIDD-CG algorithm exploits a much smaller ensemble size with only  $J = 100$  samples.

For the purpose of illustrating the skill of the PIDD-CG algorithm for predicting the entire time evolution of the system, Figure 4 includes the predicted PDFs at both a short-term transient phase and the nearly final saturated statistical equilibrium state. The non-Gaussian regime has a longer mixing time with the statistical equilibrium being reached at around the lead time  $t = 2$ , while the solution in the near-Gaussian regime mixes faster and reaches the statistical equilibrium state within  $t = 1$ . The PIDD-CG algorithm succeeds in capturing the transient PDFs as well as the final equilibrium state in both regimes containing distinct statistics. Particularly, the highly skewed and fat-tailed PDFs in the non-Gaussian regime are accurately reproduced by the PIDD-CG algorithm. Table 1 includes a quantitative assessment of predicting the one-dimensional PDFs. It shows the relative error, quantified by an information measurement (the relative entropy) [21, 27], at different forecast lead times before the system reaches the statistical equilibrium. The forecast error remains in a negligible level at the order of  $O(10^{-3})$  in most cases, and the biggest error occurs at the longest lead time, which is nevertheless at most of order  $O(10^{-2})$ . Among different variables, the predicted zonal mean flow  $U$  shows a slightly larger error than the fluctuation modes, which is due to the relatively more severe approximation in  $U$ . In fact, only the conditional means of the leading two fluctuation modes are explicitly included in the development of the closure model of  $U$  while the combined feedback from the remaining multiple fluctuation modes is completely approximated by the RNN. Nevertheless, the error in predicting  $U$  lies in an overall low level, which justifies the strategy in the PIDD-CG algorithm that combines the conditional mean time series with the RNN in facilitating the statistical closure of  $U$ .



(a) non-Gaussian regime



(b) near-Gaussian regime

Figure 3: Forecast at lead time  $t = 1$  of the trajectories for the zonal mean flow  $U$  as well as the conditional mean and conditional variance of the first fluctuation modes  $\bar{v}_1, \bar{T}_1$  and  $r_{v_1}, r_{T_1}$ . Panel (a) shows the solutions in the non-Gaussian regime while Panel (b) shows those in the near-Gaussian regime.

### 3 Discussion

Different from the traditional ensemble forecast methods, the PIDD-CG algorithm incorporates the evolution equations of the conditional statistics as part of the forecast scheme. A data-driven component using RNNs is incorporated into the method

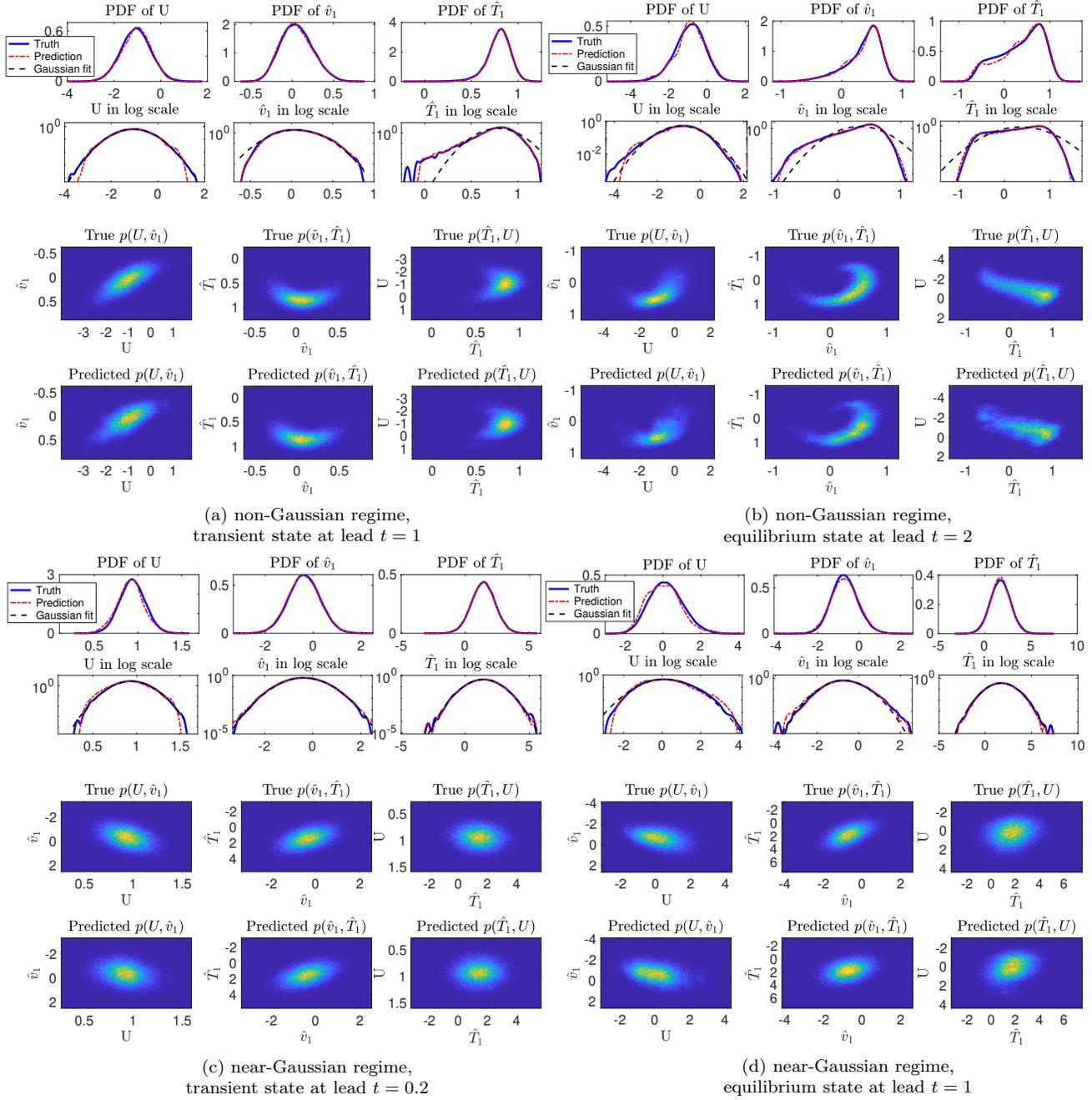


Figure 4: Comparison of the truth and the forecast PDFs of the zonal mean flow  $U$  and the first fluctuation modes  $\hat{v}_1, \hat{T}_1$ . The truth is generated from a direct Monte-Carlo simulation with  $J_{MC} = 50000$  ensembles, while the prediction from the PIDD-CG algorithm only uses  $J = 100$  ensembles. The prediction at both a transient state and a nearly statistical equilibrium state are compared in both regimes. To better demonstrate the non-Gaussianity of each variable, the comparison of the one-dimensional PDFs in the logarithm scale is also included. Note that the ranges of  $x$ -axes for the same variable are different with each other in two panels, representing distinct features in the transient and the nearly statistical equilibrium states.

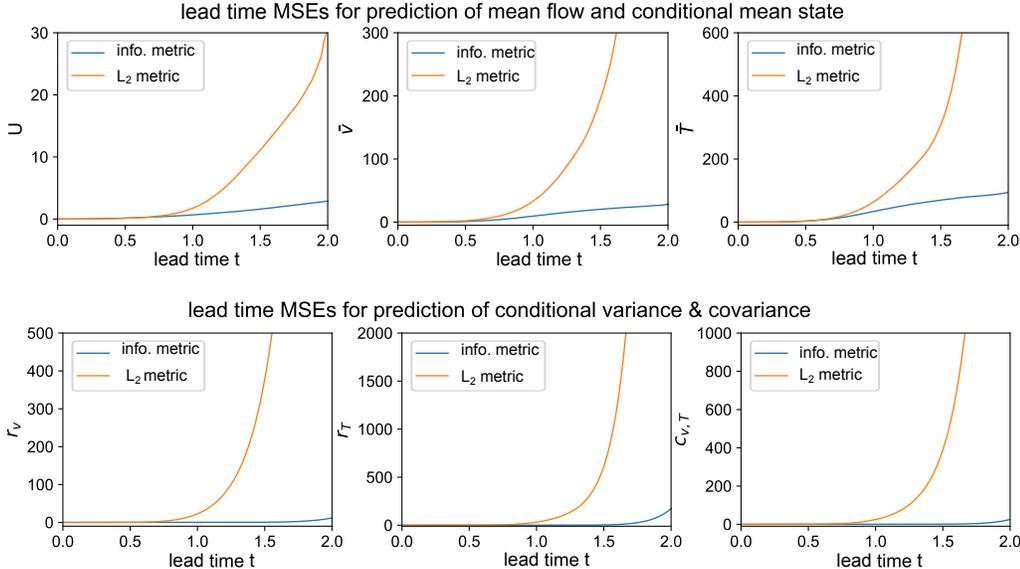


Figure 5: Prediction errors using the RNN optimized under different loss functions in the non-Gaussian regime. The mean square errors (MSEs) in forecasting  $U$ , the conditional mean  $\bar{v}_1, \bar{T}_1$ , the conditional variance  $r_{v_1}, r_{T_1}$  and the conditional cross-covariance  $C_{v_1, T_1}$  under the  $L_2$  loss and the information loss (relative entropy) are compared as a function of the lead time  $t$ .

to facilitate an efficient estimation of the combined fluctuation feedback in the observed state and the multiscale fluctuation modes in the unobserved high-dimensional subspace. Since the learning target has become a conditional distribution, an information metric is utilized as the loss function to train the RNN, which is a natural choice to avoid unnecessary fluctuating errors in turbulent signals. Yet, a quantitative assessment of the advantage of using such an information metric compared with the traditional  $L^2$  loss function is still needed. To this end, Figure 5 illustrates the forecast skill of the RNN in the non-Gaussian regime trained by these two different losses during the optimization process. Due to the strong turbulent nature of the system, the errors accumulate in time as the model is updated recurrently. Nevertheless, the forecast error using the RNN trained with the information loss grows much slower than that using the RNN optimized via the  $L^2$  loss. In fact, there is already an intrinsic barrier in the training phase if the  $L^2$  loss is utilized, which indicates that the path-wise measurement is not the most appropriate choice in minimizing the statistical error. In particular, it is noticeable from Figure 5 that the prediction of the conditional mean using the information metric remains accurate even at a very long lead time. This is crucial for accurately recovering of the joint PDFs, as the conditional mean explicitly appears in the closure term in the  $U$  equation.

Finally, it is worthwhile to point out that predicting the full spectrum of the state variables, especially in high-resolution systems, is not computational feasible. The primary interest often lies in predicting the large-scale coherent structures. Therefore, developing computationally efficient reduced order models with suitable parameterizations and forecasting the statistics of the leading a few resolved modes, which still correspond to a relatively high-dimensional PDF for direct numerical simulations, have more practical significance in more realistic applications [29, 1, 20, 41, 5, 3, 22, 42, 12]. The PIDD-CG algorithm can facilitate the development of efficient statistical reduced order models as well as accelerating the associated statistical forecast. Specifically, a large portion of the state variables in  $\mathbf{Y}$  can be truncated and only those that are utilized explicitly in the closure of  $\mathbf{X}$  in *Step 2* are preserved in the reduced order model. The contribution from the truncated modes in  $\mathbf{Y}$  can be effectively approximated by RNNs, which are then added to the equations of both  $\mathbf{X}$  and the resolved components of  $\mathbf{Y}$  as the additional closure terms. Note that, in applying the PIDD-CG algorithm to the topographic barotropic flow model (1), the statistical evolutions of the pairs  $(\hat{v}_k, \hat{T}_k)$  and  $(\hat{v}_{k'}, \hat{T}_{k'})$  with  $|k| \neq |k'|$  do not explicitly influence each other but they are coupled through the zonal mean flow  $U$  (see the SI). Therefore, the resolved subsystem consisting of these variables serves as a natural reduced order formulation for predicting the joint PDF of  $U$  and  $(\hat{v}_k, \hat{T}_k)$  with  $k = \pm 1$  and  $\pm 2$ . In fact, the forecast PDF from such a 9-dimensional reduced order model is exactly the same as the associated 9-dimensional marginal distribution from the 41-dimensional model when applying the PIDD-CG algorithm to the topographic barotropic model (1), due to the lack of explicit dependence between different  $(\hat{v}_k, \hat{T}_k)$  pairs.

## 4 Methods

### 4.1 General mathematical framework of the PIDD-CG algorithm

In this section, we present the general strategy of the PIDD-CG algorithm step-by-step, following the schematic illustration in Figure 1. In the PIDD-CG algorithm, high computational efficiency is achieved by avoiding the direct MC sampling in the whole high-dimensional phase space of the entire system. After a proper phase space decomposition (*Step 1*), the ensemble sampling is performed only inside a low-dimensional subspace, where a systematic statistical closure approximation has been applied in the leading-order states (*Step 2*). On the other hand, an effective physics-informed data-driven estimation is adopted to advance the conditional statistics forecast in the remaining high-dimensional subspace (*Step 3*), which allows an accurate recovery of the conditional statistics only relying on the small number of samples. In particular, the ensemble size is independent of the full dimension of the phase space [10], and thus the curse of dimensionality is avoided. The ensemble members in the PIDD-CG algorithm consist of a conditional Gaussian mixture, which can be tracked by closed analytic formulae for the conditional mean and covariance (*Step 4*). Finally, a RNN is introduced to approximate the nonlinear unresolved-scale feedbacks in the analytic expressions of the conditional statistics as well as the statistical feedback in model closure in *Step 2* and *3*. In particular, since the model outputs are associated with the conditional distribution, a simple but effective information loss function is adopted to train the RNN (*Step 5*). The combination of the physics-informed analytic time evolution of the statistics and the data-driven RNN closure of unresolved multiscale feedbacks accelerates the overall computational efficiency and model accuracy by a significant amount.

We start with the general formulation of turbulent dynamical systems [47, 40, 19, 28],

$$\frac{d\mathbf{u}}{dt} = (L + D)\mathbf{u} + B(\mathbf{u}, \mathbf{u}) + \mathbf{F}(t) + \boldsymbol{\sigma}(\mathbf{u}, t)\dot{\mathbf{W}}(t), \quad (2)$$

where the state variable  $\mathbf{u} \in \mathbb{C}^N$  lies in a high-dimensional phase space. In (2), the first two components,  $(L + D)\mathbf{u}$ , represent linear dispersion and dissipation effects, where  $L^* = -L$  is a skew-symmetric operator; and  $D$  is negative definite. The nonlinear effect is introduced through a quadratic form,  $B(\mathbf{u}, \mathbf{u})$ , which conserves the total energy with  $\mathbf{u} \cdot B(\mathbf{u}, \mathbf{u}) = 0$ . Besides, the system is subject to external forcing effects that are decomposed into a deterministic component,  $\mathbf{F}(t)$ , and a stochastic one represented by a Gaussian random process,  $\boldsymbol{\sigma}(\mathbf{u}, t)\dot{\mathbf{W}}(t)$ , where  $\boldsymbol{\sigma}$  measures the noise amplitude and  $\dot{\mathbf{W}}$  is the white noise.

#### *Step 1. Phase space decomposition*

To facilitate the analytically solvable properties in the PIDD-CG algorithm, we introduce a proper phase space decomposition of general system (2). Decompose the original model state  $\mathbf{u}$  into two multi-dimensional state variables,  $\mathbf{X} \in \mathbb{C}^{N_1}$  and  $\mathbf{Y} \in \mathbb{C}^{N_2}$ , with  $N_1 + N_2 = N$ . Usually,  $\mathbf{X}$  is a collection of the large-scale, resolved leading modes or observed state variables while  $\mathbf{Y}$  contains the remaining relatively smaller-scale modes, including unresolved and unobserved ones. Therefore,  $\mathbf{X}$  belongs to a relatively low-dimensional subspace while  $\mathbf{Y}$  remains high-dimensional. Since  $\mathbf{Y}$  by design denotes the faster and smaller scale components of the system, the terms corresponding to the nonlinear self-interaction inside  $\mathbf{Y}$ , i.e.  $B(\mathbf{Y}, \mathbf{Y})$ , mostly involve high frequencies and homogeneous dynamics [31]. Thus, these terms can often be effectively parameterized either by simple stochastic noise [30, 35, 32, 3] or suitable closures that are nonlinear functions of  $\mathbf{X}$  and conditionally linear functions of  $\mathbf{Y}$  [33]. The resulting approximate system can successfully capture the dominant dynamical and statistics features of the original one as well as reproducing very similar ensemble forecast solutions.

With such a decomposition of the model states, the following nonlinearly coupled multiscale stochastic model is reached as the approximation of the original system (2):

$$\frac{d\mathbf{X}}{dt} = \left[ \mathbf{A}_0(\mathbf{X}, t) + \mathbf{A}_1(\mathbf{X}, t)\mathbf{Y}(t) \right] + \mathbf{B}_1(\mathbf{X}, t)\dot{\mathbf{W}}_1(t), \quad (3a)$$

$$\frac{d\mathbf{Y}}{dt} = \left[ \mathbf{a}_0(\mathbf{X}, t) + \mathbf{a}_1(\mathbf{X}, t)\mathbf{Y}(t) \right] + \mathbf{b}_2(\mathbf{X}, t)\dot{\mathbf{W}}_2(t), \quad (3b)$$

where  $\mathbf{A}_0, \mathbf{a}_0, \mathbf{A}_1, \mathbf{a}_1, \mathbf{B}_1$  and  $\mathbf{b}_2$  are vectors or matrices that can depend nonlinearly on the state variables  $\mathbf{X}$  and time  $t$  while  $\dot{\mathbf{W}}_1$  and  $\dot{\mathbf{W}}_2$  are independent white noise sources. One desirable feature of (3) is that, given one realization of the time series  $\mathbf{X}(s)$  for  $s \in [0, t]$ , the conditional distribution

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \leq t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \mathbf{R}(t)) \quad (4)$$

becomes Gaussian, where the conditional mean  $\boldsymbol{\mu}$  and the conditional covariance  $\mathbf{R}$  can be solved via closed analytic formulae [25] (details will be shown next in *Step 4*).

It is worthwhile to highlight that many complex nonlinear systems already fit into the framework of (3) [6, 7]:

- Physics-constrained nonlinear stochastic models. Examples include the noisy versions of Lorenz models, Charney-DeVore flows, and the paradigm model for topographic mean flow interactions.

- Stochastically coupled reaction-diffusion models in neuroscience and ecology. Examples include FitzHugh-Nagumo models and SIR epidemic models.
- Multi-scale models in turbulence and geophysical flows. Example include the Boussinesq equations and rotating shallow water equation.

These examples further justify that the modeling framework (3) is appropriate to characterize or approximate many nonlinear and non-Gaussian systems in various disciplines.

### Step 2. Systematic multiscale statistical closure of the large-scale dynamics

Since the observed large-scale state variable  $\mathbf{X}$  lies in a low-dimensional subspace, a small number of suitable random sample points is sufficient to effectively characterize such a low-dimensional PDF. However, as the dynamics of  $\mathbf{X}$  is nonlinearly coupled with  $\mathbf{Y}$ , the high-dimensional system (3) has to be integrated all together requiring solutions of both  $\mathbf{X}$  and  $\mathbf{Y}$  to obtain the ensemble forecast of  $\mathbf{X}$ . This is not only computationally challenging for the simulation of each single realization, but also requires a large ensemble size to accurately reconstruct the statistics of the entire system. To reduce the high computational cost, a systematic multiscale statistical closure model of  $\mathbf{X}$  is thus developed, the purpose of which is to avoid running the full set of the equations of  $\mathbf{Y}$  during predicting the marginal PDF of only  $\mathbf{X}$ .

The multiscale statistical closure here depends on the crucial feature that the dynamics of  $\mathbf{Y}$  in (3a) becomes linear given one realization of the trajectory of  $\mathbf{X}$ . Thus the conditional mean of  $\mathbf{Y}(t)$  relying on the history observation of  $\mathbf{X}(s)$ ,  $s < t$  can be solved via closed analytic formulae (shown in Step 4). Next, decompose  $\mathbf{Y}$  as  $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ , where  $\mathbf{Y}_1$  is the resolved subscale processes and  $\mathbf{Y}_2$  represent the rest unresolved ones. Correspondingly, the conditional mean of  $\mathbf{Y}$  can be decomposed as  $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ . Accordingly, rewrite  $\mathbf{A}_1$  in (3a) as  $\mathbf{A}_1 = [\mathbf{A}_{1,1}, \mathbf{A}_{1,2}]$ . The statistical closure model of  $\mathbf{X}$  in (3a) reads:

$$\begin{aligned}
\frac{d\mathbf{X}}{dt} &= [\mathbf{A}_0 + \mathbf{A}_1 \mathbf{Y}] + \mathbf{B}_1 \dot{\mathbf{W}}_1 \\
&= [\mathbf{A}_0 + \mathbf{A}_{1,1} \mathbf{Y}_1 + \mathbf{A}_{1,2} \mathbf{Y}_2] + \mathbf{B}_1 \dot{\mathbf{W}}_1 \\
&= [\mathbf{A}_0 + \mathbf{A}_{1,1} \boldsymbol{\mu}_1] + \mathbf{B}_1 \dot{\mathbf{W}}_1 + [\mathbf{A}_{1,1} (\mathbf{Y}_1 - \boldsymbol{\mu}_1) + \mathbf{A}_{1,2} \mathbf{Y}_2] \\
&:= [\mathbf{A}_0 + \mathbf{A}_{1,1} \boldsymbol{\mu}_1] + \mathbf{B}_1 \dot{\mathbf{W}}_1 + \mathcal{F}_{\mathbf{X}}.
\end{aligned} \tag{5}$$

The contributions from the resolved and unresolved components of  $\mathbf{Y}$  are separated in the above closure model (5). First, the contribution from the resolved leading modes of  $\mathbf{Y}$  is modeled by the conditional mean and its stochastic deviation from the mean, namely  $\mathbf{A}_{1,1} \mathbf{Y}_1 = \mathbf{A}_{1,1} \boldsymbol{\mu}_1 + \mathbf{A}_{1,1} (\mathbf{Y}_1 - \boldsymbol{\mu}_1)$ . Second, the combined contribution of the unresolved fast processes  $\mathbf{A}_{1,2} \mathbf{Y}_2$  is modeled together as the coupled feedback from the various multiscale fluctuations. The conditional mean part with  $\boldsymbol{\mu}_1$  is explicitly modeled through the conditional dynamics of reduced order model in Step 4 and Step 5, while the remaining components of the deviation from the mean and the fast fluctuations in a high-dimensional space are both unresolved. In the closure approximation (5), we denote the unresolved ‘residual’ part as  $\mathcal{F}_{\mathbf{X}}$  by combining contributions from both the mean deviation  $\mathbf{Y}_1 - \boldsymbol{\mu}_1$  and the remaining high-dimensional unresolved fluctuations  $\mathbf{Y}_2$  together. Usually, these terms will include complex nonlinear coupling between multiple scales. Nevertheless, we only need their combined feedback for the prediction of the large-scale state  $\mathbf{X}$ .

Here, the unresolved mean feedback  $\mathcal{F}_{\mathbf{X}}$  is effectively approximated from data by a RNN:

$$\mathcal{F}_{\mathbf{X}}(t+1) = \text{RNN}(\mathbf{X}(t-m:t), \boldsymbol{\mu}_1(t-m:t), \mathcal{F}_{\mathbf{X}}(t-m:t)), \tag{6}$$

where the input of the RNN depends on the discrete time series from a past time instant  $t-m$  to the current time instant  $t$  of the state variable  $\mathbf{X}$ , the conditional mean of the resolved leading modes  $\boldsymbol{\mu}_1$  and  $\mathcal{F}_{\mathbf{X}}$  itself (see more details of the neural network architecture in the SI). Therefore, the closure model (5) together with the governing equations of the conditional mean  $\boldsymbol{\mu}_1$  leads to a closed system. An ensemble simulation of this set of equations with a small number of samples can be carried out to sufficiently characterize the low-dimensional PDF of  $\mathbf{X}$  at future time instants via running the RNN (6) iteratively forward starting from an initial time instant.

Finally, denote the total number of ensemble members by  $J$ , and the forecast value at time  $t$  associated with the  $j$ -th ensemble member by  $\mathbf{X}^{\{j\}}(t)$ . Then a smoothed PDF of  $\mathbf{X}(t)$  can be reached by a kernel density estimation (KDE) [4],

$$p(\mathbf{X}(t)) = \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J \tilde{p}(\mathbf{X}^{\{j\}}(t)). \tag{7}$$

In (7),  $\tilde{p}(\mathbf{X}^{\{j\}}(t))$  is the  $j$ -th member from the KDE that is associated with  $\mathbf{X}^{\{j\}}(t)$  using the ‘solve-the-equation plug-in’ algorithm [4], which is an appropriate KDE method for approximating non-Gaussian distributions. Note that the asymptotic

expression as  $J \rightarrow \infty$  is used in (7) for the mathematical rigor of the formula. In practice, only a finite value of  $J$  is adopted as the approximation. To facilitate the computation in *Step 3*, a Gaussian kernel is used in (7) such that  $\tilde{p}(\mathbf{X}^{\{j\}}(t))$  is a Gaussian distribution centered at  $\mathbf{X}^{\{j\}}(t)$ .

### Step 3. Effective physics-informed conditional Gaussian mixture via data assimilation

In the ensemble simulation of the large-scale dynamics in *Step 2*, each ensemble member provides one trajectory of  $\mathbf{X}$  up to time  $t$ , denoted by  $\mathbf{X}^{\{j\}}(s \leq t)$ , where  $j = 1, \dots, J$ . Given the coupled model formulation in (3) and conditioned on the realization  $\mathbf{X}^{\{j\}}(s \leq t)$ , there is one corresponding distribution of  $\mathbf{Y}$  at time instant  $t$ , namely  $p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t))$ . Such a conditional distribution can be viewed as the posterior distribution of the analysis state  $\mathbf{Y}(t)$  from data assimilation, where  $\mathbf{X}^{\{j\}}(s \leq t)$  plays the role of the observed time series. One desirable feature in the model (3) is that it becomes a conditional Gaussian distribution (4), due to the linear dynamics of  $\mathbf{Y}$  conditioned on  $\mathbf{X}(s \leq t)$  with Gaussian white noise [6]. In light of these conditional Gaussian distributions, the marginal distribution of  $\mathbf{Y}(t)$  is provided by a conditional Gaussian mixture,

$$p(\mathbf{Y}(t)) = \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t)). \quad (8)$$

Combining (8) with (7) yields the formula for the joint distribution,

$$p(\mathbf{X}(t), \mathbf{Y}(t)) = \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J \tilde{p}(\mathbf{X}^{\{j\}}(t)) p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t)). \quad (9)$$

Since the component  $p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(\cdot))$  with each sample  $j$  is a Gaussian distribution, the overall joint distribution in (9) is given by a Gaussian mixture. It has been shown in rigorous analysis [10] that the error bound of the joint distribution in (9) does not depend on the dimension of  $\mathbf{Y}$ . Therefore, as long as  $\mathbf{X}$  stays in a relatively low dimensional subspace, it suffices to adopt a small number of samples (i.e., a small  $J$ ) to accurately approximate the joint PDF based on the formula in (9).

Here, the physics-informed ingredient is embodied in the data assimilation that takes into account the large-scale model information to improve the conditional distribution. In fact, the center of each  $p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t))$  can be very different from the actual value of  $\mathbf{Y}^{\{j\}}(t)$  when it is explicitly simulated with  $\mathbf{X}^{\{j\}}(t)$  from the original coupled system. This distinguishes from the KDE, where the mixture components are often centered at the simulated data points. In addition, the covariance of  $p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t))$ , which is the correspondence to the fixed bandwidth in the standard KDE, is determined utilizing the dynamical properties via data assimilation and can be adaptive for different  $j$ 's. The optimization procedure of automatically determining the center and the bandwidth of each Gaussian mixture component in such a physics-driven method facilitates the use of only a small number of samples to accurately approximate the high-dimensional PDF.

### Step 4. Analytic formulae for time evolution of conditional statistics in smaller-scale dynamics

To forecast the joint PDF using (9), what remains to compute is the conditional Gaussian distribution  $p(\mathbf{Y}(t)|\mathbf{X}^{\{j\}}(s \leq t))$  for the sample realizations  $j = 1, \dots, J$ . However, since  $\mathbf{Y}$  contains all the smaller scale dynamics in a high-dimensional subspace, applying a direct ensemble method to forecast its statistics has the same the curse of dimensionality issue as simulating the original physical system (2) or (3). From a different approach exploiting the important structural feature of the system (3), the dynamical equations for the conditional mean  $\boldsymbol{\mu}$  and the conditional covariance  $\mathbf{R}$  in (4) are available via the following explicit analytic formulae [25]

$$\frac{d\boldsymbol{\mu}}{dt} = (\mathbf{a}_0 + \mathbf{a}_1\boldsymbol{\mu}) + (\mathbf{R}\mathbf{A}_1^*)(\mathbf{B}_1\mathbf{B}_1^*)^{-1} \left( \frac{d\mathbf{X}}{dt} - (\mathbf{A}_0 + \mathbf{A}_1\boldsymbol{\mu}) \right), \quad (10a)$$

$$\frac{d\mathbf{R}}{dt} = \mathbf{a}_1\mathbf{R} + \mathbf{R}\mathbf{a}_1^* + \mathbf{b}_2\mathbf{b}_2^* - (\mathbf{R}\mathbf{A}_1^*)(\mathbf{B}_1\mathbf{B}_1^*)^{-1}(\mathbf{A}_1\mathbf{R}), \quad (10b)$$

with  $*$  being the complex conjugate transpose. Once a single trajectory of  $\mathbf{X}$  is given, the system (10) can be solved using the standard ODE solvers, such as the Euler or the Runge-Kutta schemes. In this way, the extensive MC simulation of the entire high-dimensional system can be effectively avoided. In addition, the analytic formulae of the moment equations in (10) avoid the potential computational issues of random sampling errors and recover the true conditional statistics characterized by the deterministic solutions for the mean  $\boldsymbol{\mu}$  and covariance  $\mathbf{R}$ .

### Step 5. Data-driven modeling of the nonlinear feedbacks in conditional statistics with information theory

Despite the closed analytic formulae for solving the conditional statistics, the cost of running (10), especially the full spectrum of the conditional covariance  $\mathbf{R} \in \mathbb{C}^{N_2 \times N_2}$  in (10b), remains to be computational demanding for high-dimensional systems.

Therefore, an additional model reduction strategy is required to further reduce the computational cost and focus on the conditional statistics of the resolved states  $\mathbf{Y}_1$ . The contributions from the remaining unresolved fluctuations are modeled through a closure scheme as described in (5) of *Step 2*. Usually,  $\mathbf{Y}_1$  can be separated by taking the most energetic leading modes of the high-dimensional state  $\mathbf{Y}$ . In addition, certain localized structures [48, 2] can be exploited to approximate the covariance matrix with a diagonal or block diagonal structure. Thus only the entries near the diagonal line need to be computed in the algorithm. For example in our illustrative example (1) in Section 2.2, the coupling coefficient  $\mathbf{a}_1$  is automatically diagonalized since it represents linear dispersion and dissipation effects in small scales as well as the noise coefficients  $\mathbf{B}_1$  and  $\mathbf{b}_2$ .

To accelerate the computational efficiency, the most time consuming parts including the complicated nonlinear and possibly unresolved information in solving (10) are directly learned from data. To this end, define

$$\begin{aligned}\mathcal{F}_{\mathbf{Y}} &= \dot{\mathbf{X}} - (\mathbf{A}_0 + \mathbf{A}_1\boldsymbol{\mu}), \\ \mathcal{G}_{\mathbf{Y}} &= \mathbf{R}\mathbf{A}_1^*,\end{aligned}\tag{11}$$

where  $\mathcal{F}_{\mathbf{Y}}$  and  $\mathcal{G}_{\mathbf{Y}}$  are the ‘embedded’ feedbacks to the resolved mean and covariance dynamics. The explicit expressions on the right hand sides of (11) contain the full information in  $\boldsymbol{\mu}$  and  $\mathbf{R}$ , while the RNNs can help us learn the resolved state information without running the full spectrum of unresolved fluctuation modes. More importantly, note that the mean feedback will be the same as the unresolved term in (6), i.e.,  $\mathcal{F}_{\mathbf{Y}} = \mathcal{F}_{\mathbf{X}}$ , when we take  $\mathbf{Y} = \mathbf{Y}_1$ . Thus the computational cost is further reduced. As an analog to  $\mathcal{F}_{\mathbf{Y}}$  and  $\mathcal{G}_{\mathbf{Y}}$  in (11), define  $\mathcal{F}_{\mathbf{Y}_1}$  and  $\mathcal{G}_{\mathbf{Y}_1}$  as the functions constrained on the resolved subspace of  $\mathbf{Y}_1$ . Then the associated low-order dynamics from the full formulae in (10) can be rewritten as

$$\frac{d\boldsymbol{\mu}_1}{dt} = (\mathbf{a}_0 + \mathbf{a}_1\boldsymbol{\mu}_1) + \mathcal{G}_{\mathbf{Y}_1}(\mathbf{B}_1\mathbf{B}_1^*)^{-1}\mathcal{F}_{\mathbf{Y}_1},\tag{12a}$$

$$\frac{d\mathbf{R}_1}{dt} = \mathbf{a}_1\mathbf{R}_1 + \mathbf{R}_1\mathbf{a}_1^* + \mathbf{b}_2\mathbf{b}_2^* - \mathcal{G}_{\mathbf{Y}_1}(\mathbf{B}_1\mathbf{B}_1^*)^{-1}\mathcal{G}_{\mathbf{Y}_1}^*.\tag{12b}$$

Then  $\mathcal{F}_{\mathbf{Y}_1}$  and  $\mathcal{G}_{\mathbf{Y}_1}$  are approximated by the following RNNs,

$$\begin{aligned}\mathcal{F}_{\mathbf{Y}_1}(t+1) &= \text{RNN}(\mathbf{X}(t-m:t), \boldsymbol{\mu}_1(t-m:t), \mathcal{F}_{\mathbf{Y}_1}(t-m:t)), \\ \mathcal{G}_{\mathbf{Y}_1}(t+1) &= \text{RNN}(\mathbf{X}(t-m:t), \mathbf{R}_1(t-m:t), \mathcal{G}_{\mathbf{Y}_1}(t-m:t)).\end{aligned}\tag{13}$$

What remains is to train the RNNs. It is worthwhile to highlight that these RNNs are used to approximate certain components in the moment equations. Therefore, it is important to develop a suitable criterion as the loss function of the RNNs such that the resulting moments  $\boldsymbol{\mu}$  and  $\mathbf{R}$  or the associated PDF are forecasted as accurate as possible. Since the path-wise error is not necessarily related to the calibration of the forecast statistics, minimizing the path-wise errors in forecasting the conditional mean and conditional covariance is not the most appropriate choice of the loss function for training the RNNs. Instead, an information loss function is adopted here that specifically emphasizes the minimization of the forecast error in terms of the PDF. The information criterion used here is the so-called relative entropy or the Kullback-Leibler divergence (KL divergence) [23, 21],

$$\mathcal{P}(p^{\text{ref}}(\mathbf{z}), p^f(\mathbf{z})) = \int p^{\text{ref}}(\mathbf{z}) \ln \frac{p^{\text{ref}}(\mathbf{z})}{p^f(\mathbf{z})} d\mathbf{z},\tag{14}$$

where  $p^{\text{ref}}(\mathbf{z})$  is the true PDF while  $p^f(\mathbf{z})$  is the predicted one.

## 4.2 Implementation details of the PIDD-CG algorithm

### 4.2.1 ML training and the use of the relative entropy as cost function

In the training process, the neural network parameters are achieved through the optimization using a proper loss function. A straightforward choice of the loss function is the standard  $L^2$  loss, which measures the mean square error (MSE) between the truth and the predicted conditional mean or the predicted conditional covariance. That is,

$$\mathcal{L}_{\text{MSE}}(t) = \left\| \bar{u}^{\text{NN}}(t) - \bar{u}^{\text{ref}}(t) \right\|_{L^2}^2 + \sum_k \alpha_k \left\| r_k^{\text{NN}}(t) - r_k^{\text{ref}}(t) \right\|_{L^2}^2.\tag{15}$$

In (15),  $\bar{u}^{\text{NN}}$  and  $\bar{u}^{\text{ref}}$  are the conditional mean of neural network output and the truth, respectively; and  $r_k^{\text{NN}}$  and  $r_k^{\text{ref}}$  are the conditional covariance entries. However, as is shown in the SI, this  $L^2$ -loss becomes insufficient for guiding the training convergence to the optimal critical point, especially when highly turbulent fluctuations become dominant in the solution fields.

In the regimes with stronger extreme events and many noisy small-scale fluctuations, it becomes essential to focus on the dominant solution structures and ‘filter out’ the noises in small amplitudes in the training phase. To this end, a more balanced

measurement of the training error, is introduced here. It is named as the information loss as it exploits an information metric — the relative entropy — as the loss function:

$$\mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{j=1}^M L_j, \quad L_j(\mathbf{x}^j, \mathbf{y}^j) = \sum_i \tilde{y}_i^{(j)} \log \frac{\tilde{y}_i^{(j)}}{\tilde{x}_i^{(j)}}, \quad (16)$$

where  $\mathbf{x}$  is the PDF reconstructed by the predicted conditional statistics while  $\mathbf{y}$  is the reconstruction of the true PDF made by the true conditional statistics. The superscript  $j$  in (16) represents the mini-batch members and the subscript  $i$  goes through the dimensions of the variable. In addition, to highlight more towards the extreme events, the following two sets of positive and negative temperatures are added to rescale the data from the partition functions

$$\tilde{x}_i^+ = \frac{\exp(x_i/T_+)}{\sum_i \exp(x_i/T_+)}, \quad \tilde{x}_i^- = \frac{\exp(-x_i/T_-)}{\sum_i \exp(-x_i/T_-)}, \quad (17)$$

where  $T_+ > 0, T_- > 0$  are the positive and negative temperatures weighting the importance of extreme events in the scaled measure.

#### 4.2.2 Initialization of the conditional statistics ensembles

Assume the initial time instant for the forecast is at  $t = T$ . In the traditional ensemble forecast, the initialization of the ensembles is provided by data assimilation. Consider the modeling framework in (3). Assume the time series of the large-scale variable  $\mathbf{X}$  is fully observed with no additional observational error while the observation of  $\mathbf{Y}$  is not available. The situation in which the observation of  $\mathbf{X}$  contains observational error can be easily handled by imposing another data assimilation procedure for  $\mathbf{X}$ , which is however not the main scope of the current framework. Therefore, the precise observational value of  $\mathbf{X}$  at time  $T$  is naturally served as the initial condition of  $\mathbf{X}(T)$ . This also means all the initial ensembles of  $\mathbf{X}(T)$  are the same as the observed value. On the other hand, the initial ensemble of  $\mathbf{Y}(T)$  is provided by sampling  $J$  different samples from the conditional (or the so-called posterior) distribution (4), which is given by the data assimilation formulae in (10).

The initialization of the PIDD-CG algorithm has two major differences compared with the traditional ensemble forecast initialization. First, each ensemble member in the PIDD-CG algorithm is no longer a single point but instead a conditional Gaussian distribution. Second, due to the use of neural network approximations in (6) and (13), each initial ensemble of  $\mathcal{F}_{\mathbf{X}}$ ,  $\mathcal{F}_{\mathbf{Y}}$  and  $\mathcal{G}_{\mathbf{Y}}$  contains a time series that requires the past information. This is different from the traditional ensemble forecast that only exploits the state estimation at the initial time instant. The details of the initialization in the PIDD-CG algorithm are as follows.

(a). *The initialization of the state variable  $\mathbf{Y}$  in (3).*

Since the initial ensembles of  $\mathbf{X}$  are all the same, it is natural to use the same conditional statistics ensembles for the initialization of  $\mathbf{Y}$  as well. In particular, all the initial ensembles of  $\mathbf{Y}$  take the values where  $\boldsymbol{\mu}$  and  $\mathbf{R}$  are the posterior mean and posterior covariance computed from the direct data assimilation (10). This also makes the entire initial distribution of  $\mathbf{X}$  and  $\mathbf{Y}$  to be consistent as the traditional ensemble forecast method.

(b). *The initialization of the functions  $\mathcal{F}_{\mathbf{Y}}$  and  $\mathcal{G}_{\mathbf{Y}}$  in (13).*

The functions  $\mathbf{F}_{\mathbf{Y}}$  and  $\mathcal{G}_{\mathbf{Y}}$  depend on the past information of  $\boldsymbol{\mu}$ ,  $\mathbf{R}$ ,  $\mathbf{X}$  and themselves. Therefore, the data assimilation scheme (10) starts from a certain time instant in the past, and then results in the time series of  $\boldsymbol{\mu}$  and  $\mathbf{R}$  from  $T - m$  to the current time instant  $T$ . The time series of  $\mathbf{X}$  from  $T - m$  to  $T$  is available from observations.

(c). *The initialization of the functions  $\mathcal{F}_{\mathbf{X}}$  in (6).*

The input of the neural network in (6) requires additional path-wise information of the unobserved trajectory  $\mathbf{Y}$ , which are related but are not directly available using the point-wise posterior mean and posterior covariance. The trajectory of  $\mathbf{Y}$  can be sampled from an infinite dimensional (or high-dimensional with temporal the discretization) joint posterior distribution, where the infinity (or high) dimensionality comes not only from the number of the state variables in  $\mathbf{Y}$  but also the temporal direction. Nevertheless, the modeling framework (3) allows such a high-dimensional sampling problem to be solved by integrating a backward stochastic differential equation [9]. It is given by

$$\overleftarrow{\frac{d\mathbf{Y}}{dt}} = \overleftarrow{\frac{d\boldsymbol{\mu}_s}{dt}} - (\mathbf{a}_1 + (\mathbf{b}_2 \mathbf{b}_2^* \mathbf{R}^{-1})(\mathbf{Y} - \boldsymbol{\mu}_s) + \mathbf{b}_2 \dot{\mathbf{W}}_{\mathbf{Y}}(t)), \quad (18)$$

where the notation  $\overleftarrow{d}/dt$  corresponds to the negative of the usual derivative, which means that the system (20) is solved backward over  $[0, T]$ . Here  $\boldsymbol{\mu}_s(t)$  and  $\mathbf{R}_s(t)$  are the so-called smoother mean and smoother covariance

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0, T]) \sim \mathcal{N}(\boldsymbol{\mu}_s(t), \mathbf{R}_s(t)), \quad (19)$$

which are also provided with closed analytic formulae

$$\frac{\overleftarrow{d}\boldsymbol{\mu}_s}{dt} = -\mathbf{a}_0 - \mathbf{a}_1\boldsymbol{\mu}_s + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_s), \quad (20a)$$

$$\frac{\overleftarrow{d}\mathbf{R}_s}{dt} = -(\mathbf{a}_1 + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}^{-1})\mathbf{R}_s - \mathbf{R}_s(\mathbf{a}_1^* + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}) + \mathbf{b}_2\mathbf{b}_2^*, \quad (20b)$$

The initial condition of solving (18) is  $(\boldsymbol{\mu}_s(T), \mathbf{R}_s(T)) = (\boldsymbol{\mu}(T), \mathbf{R}(T))$ , which is the same as the data assimilation (filtering) estimate  $(\boldsymbol{\mu}(T), \mathbf{R}(T))$ .

## Acknowledgement

The research of N.C. is partially funded by the Office of VCRGE at UW-Madison and ONR N00014-21-1-2904. The research of D. Q. is partially supported by the start-up funds provided by Purdue University.

## References

- [1] Shady E Ahmed, Suraj Pawar, Omer San, Adil Rasheed, Traian Iliescu, and Bernd R Noack. On closures for reduced order models—a spectrum of first-principle to machine-learned avenues. *Physics of Fluids*, 33(9):091301, 2021.
- [2] Jeffrey L Anderson. Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, 140(7):2359–2371, 2012.
- [3] Judith Berner, Ulrich Achatz, Lauriane Batte, Lisa Bengtsson, Alvaro de la Cámara, Hannah M Christensen, Matteo Colangeli, Danielle RB Coleman, Daan Crommelin, Stamen I Dolaptchiev, et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98(3):565–588, 2017.
- [4] Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. Kernel density estimation via diffusion. *The annals of Statistics*, 38(5):2916–2957, 2010.
- [5] Mickaël D Chekroun, Honghu Liu, and James C McWilliams. Variational approach to closure of nonlinear dynamical systems: Autonomous case. *Journal of Statistical Physics*, 179(5):1073–1160, 2020.
- [6] Nan Chen and Andrew Majda. Conditional gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy*, 20(7):509, 2018.
- [7] Nan Chen and Andrew J Majda. Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics. *Monthly Weather Review*, 144(12):4885–4917, 2016.
- [8] Nan Chen and Andrew J Majda. Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems. *Proceedings of the National Academy of Sciences*, 114(49):12864–12869, 2017.
- [9] Nan Chen and Andrew J Majda. Efficient nonlinear optimal smoothing and sampling algorithms for complex turbulent nonlinear dynamical systems with partial observations. *Journal of Computational Physics*, page 109381, 2020.
- [10] Nan Chen, Andrew J Majda, and Xin T Tong. Rigorous analysis for efficient statistically accurate algorithms for solving Fokker–Planck equations in large dimensions. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1198–1223, 2018.
- [11] Vladimir Cherkassky and Filip M Mulier. *Learning from data: concepts, theory, and methods*. John Wiley & Sons, Hoboken, New Jersey, USA, 2007.
- [12] Wouter Edeling and Daan Crommelin. Reducing data-driven dynamical subgrid scale models by physical constraints. *Computers & Fluids*, 201:104470, 2020.
- [13] Jason P Evans, Fei Ji, Gab Abramowitz, and Marie Ekström. Optimally choosing small ensemble members to produce robust climate simulations. *Environmental Research Letters*, 8(4):044050, 2013.
- [14] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, Berlin Heidelberg, Germany, 2009.
- [15] Mohammad Farazmand and Themistoklis Sapsis. Extreme events: Mechanisms and prediction. *Applied Mechanics Reviews*, 2018.
- [16] Christian LE Franzke, Terence J O’Kane, Judith Berner, Paul D Williams, and Valerio Lucarini. Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 6(1):63–78, 2015.
- [17] Michael Ghil and Stephen Childress. *Topics in geophysical fluid dynamics: atmospheric dynamics, dynamo theory, and climate dynamics*. Springer Science & Business Media, New York, USA, 2012.

- [18] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.
- [19] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, Cambridge, England, 2003.
- [20] M Amin Khodkar and Pedram Hassanzadeh. Data-driven reduced modelling of turbulent Rayleigh–Bénard convection using DMD-enhanced fluctuation–dissipation theorem. *Journal of Fluid Mechanics*, 852, 2018.
- [21] Richard Kleeman. Information theory and dynamical system predictability. *Entropy*, 13(3):612–649, 2011.
- [22] Dmitri Kondrashov, Mickaël D Chekroun, and Michael Ghil. Data-driven non-markovian closure models. *Physica D: Nonlinear Phenomena*, 297:33–55, 2015.
- [23] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [24] Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of computational physics*, 227(7):3515–3539, 2008.
- [25] Robert S Liptser and Albert N Shiryaev. *Statistics of random processes II: Applications*, volume 6. Springer Science & Business Media, Berlin Heidelberg, Germany, 2013.
- [26] Valerio Lucarini, Davide Faranda, Jorge Miguel Milhazes de Freitas, Mark Holland, Tobias Kuna, Matthew Nicol, Mike Todd, Sandro Vaienti, et al. *Extremes and recurrence in dynamical systems*. John Wiley & Sons, Hoboken, New Jersey, USA, 2016.
- [27] Andrew Majda and Xiaoming Wang. *Nonlinear dynamics and statistical theories for basic geophysical flows*. Cambridge University Press, Cambridge, England, 2006.
- [28] Andrew J Majda. *Introduction to turbulent dynamical systems in complex systems*. Springer, Switzerland, 2016.
- [29] Andrew J Majda and Di Qi. Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Review*, 60(3):491–549, 2018.
- [30] Andrew J Majda, Ilya Timofeyev, and Eric Vanden Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics*, 54(8):891–974, 2001.
- [31] Andrew J Majda, Ilya Timofeyev, and Eric Vanden-Eijnden. Systematic strategies for stochastic mode reduction in climate. *Journal of the Atmospheric Sciences*, 60(14):1705–1722, 2003.
- [32] PierGianLuca Porta Mana and Laure Zanna. Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling*, 79:1–20, 2014.
- [33] Changhong Mou, Nan Chen, and Traian Iliescu. An efficient data-driven multiscale stochastic reduced order modeling framework for complex turbulent systems. *arXiv preprint arXiv:2203.11438*, 2022.
- [34] Tim Palmer. The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.
- [35] RS Plant and George C Craig. A stochastic parameterization for deep convection based on equilibrium statistics. *Journal of the Atmospheric Sciences*, 65(1):87–105, 2008.
- [36] Di Qi and John Harlim. Machine learning-based statistical closure models for turbulent dynamical systems. *arXiv preprint arXiv:2108.13220*, 2021.
- [37] Di Qi and Andrew J Majda. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Two-layer baroclinic turbulence. *Journal of the Atmospheric Sciences*, 73(12):4609–4639, 2016.
- [38] Di Qi and Andrew J Majda. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Barotropic turbulence with topography. *Physica D: Nonlinear Phenomena*, 343:7–27, 2017.
- [39] Di Qi and Andrew J Majda. Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, 117(1):52–59, 2020.
- [40] Rick Salmon. *Lectures on geophysical fluid dynamics*. Oxford University Press, Oxford, England, 1998.
- [41] Manuel Santos Gutiérrez, Valerio Lucarini, Mickaël D Chekroun, and Michael Ghil. Reduced-order models for coupled dynamical systems: Data-driven methods and the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(5):053116, 2021.
- [42] Tapio Schneider, Andrew M Stuart, and Jin-Long Wu. Learning stochastic closures using ensemble Kalman inversion. *Transactions of Mathematics and Its Applications*, 5(1):tnab003, 2021.
- [43] Sarah A Sheard and Ali Mostashari. Principles of complex systems for systems engineering. *Systems Engineering*, 12(4):295–311, 2009.

- [44] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, Boca Raton, Florida, 2018.
- [45] Wei-Kuo Tao, Jiun-Dar Chern, Robert Atlas, David Randall, Marat Khairoutdinov, Jui-Lin Li, Duane E Waliser, Arthur Hou, Xin Lin, Christa Peters-Lidard, et al. A multiscale modeling system: Developments, applications, and critical issues. *Bulletin of the American Meteorological Society*, 90(4):515–534, 2009.
- [46] Zoltan Toth and Eugenia Kalnay. Ensemble forecasting at ncep and the breeding method. *Monthly Weather Review*, 125(12):3297–3319, 1997.
- [47] Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, Cambridge, England, 2017.
- [48] Jeffrey S Whitaker, Gilbert P Compo, Xue Wei, and Thomas M Hamill. Reanalysis without radiosondes using ensemble data assimilation. *Monthly Weather Review*, 132(5):1190–1200, 2004.
- [49] David C Wilcox. Multiscale model for turbulent flows. *AIAA journal*, 26(11):1311–1320, 1988.
- [50] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

## Supplementary Information:

This Supplementary Information (SI) contains the numerical construction and experiments for

- A) a coupled dyad model as a proof-of-concept of the PIDD-CG algorithm; and
- B) a complete analysis of the computational performance of the barotropic topographic model with multiscale coupling.

## A The coupled dyad model

First, we verify the effectiveness of the PIDD-CG algorithm on a prototype two-mode dyad model as a standard proof of concept. The purpose of using this model with the simplest possible turbulent structure is to illustrate the basic ideas in the algorithm construction and the key features in the method to predict crucial statistics in a simple and clean setup.

### A.1 Model description

The dyad model is described by the two nonlinearly coupled states  $(u, v)$  following the dynamics

$$\begin{aligned}\frac{du}{dt} &= -d_u u + cuv + F_u + \sigma_u \dot{W}_u, \\ \frac{dv}{dt} &= -d_v v - cu^2 + F_v + \sigma_v \dot{W}_v.\end{aligned}\tag{S1}$$

Naturally, we can view  $u$  as the ‘observed state’ (as  $\mathbf{X}$  in the main text) and  $v$  the ‘unresolved process’ ( $\mathbf{Y}$  in the main text) satisfying the general conditional Gaussian framework (Eqn. (3) in the main text). Different statistics can be generated by varying the model parameters  $(d_u, d_v, c, F_u, F_v, \sigma_u, \sigma_v)$ . The common parameters are taken as  $(d_u, d_v, c, F_u, F_v) = (0.8, 0.8, 1.2, 1, 1)$ . In particular, we consider two typical statistical regimes by changing the noise forcing strength for: i) near-Gaussian regime in  $u$  with  $(\sigma_u, \sigma_v) = (3, 0.2)$ ; and ii) non-Gaussian regime in  $u$  with  $(\sigma_u, \sigma_v) = (0.5, 2)$ . For conciseness, these two regimes are referred to as *near-Gaussian*  $u$  and *non-Gaussian*  $u$  in the following. The typical trajectories of the dyad model are illustrated in Figure S1. It can be seen that different strongly turbulent features including significant skewness and kurtosis appear in the solutions of  $u$  and  $v$ . Thus the dyad model becomes a desirable first test for confirming the basic features and prediction skill of the general PIDD-CG algorithm.

The conditional Gaussian dynamics can be proposed for the dyad model (S1) based on the general framework (3) in the main text. Given a realization of the observed state  $u$ , the unresolved state  $v$  has conditional Gaussian statistics  $v \sim \mathcal{N}(\bar{v}, r_v)$  with  $\bar{v} \equiv \bar{v}(t; u(\cdot))$  and  $r_v \equiv r_v(t; u(\cdot))$  being the conditional mean and conditional variance dependent on the history trajectory of  $u(s)$ ,  $s \leq t$ . Therefore, we can find the explicit dynamical equations for the conditional statistics as

$$\begin{aligned}\frac{d\bar{v}}{dt} &= -d_v \bar{v} - cu^2 + F_u + \sigma_u^{-2} \mathcal{F}_u \cdot \mathcal{G}_v, \\ \frac{dr_v}{dt} &= -2d_v r_v + \sigma_v^2 - \sigma_u^{-2} \mathcal{G}_v^2.\end{aligned}\tag{S2}$$

In the above formulation, we have the unresolved nonlinear coupling terms  $\mathcal{F}_u$  and  $\mathcal{G}_v$  as follows

$$\mathcal{F}_u = \dot{u} + d_u u - f_u - cu\bar{v}, \quad \mathcal{G}_v = cur_v.\tag{S3}$$

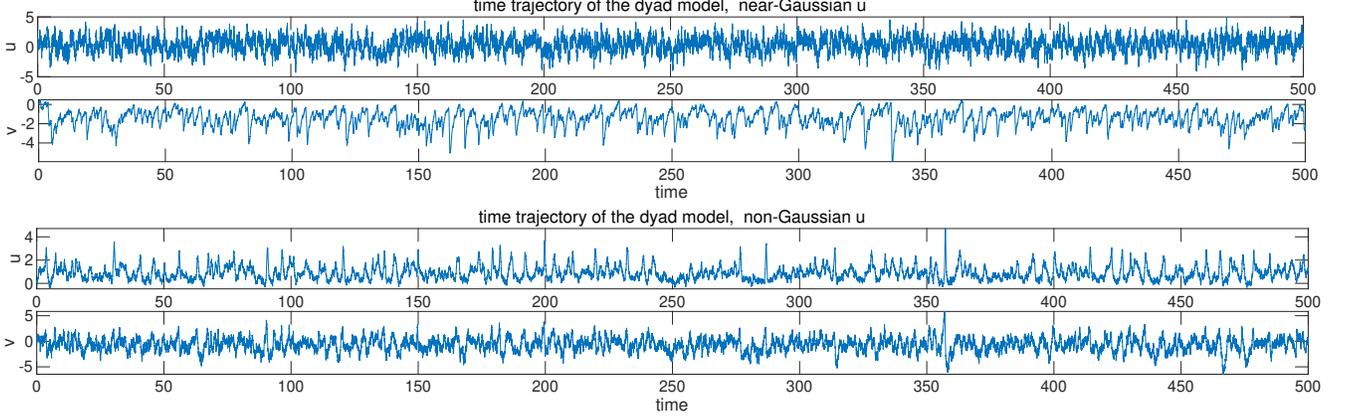


Figure S1: Trajectories of the dyad model in different statistical regimes with near-Gaussian observed state  $u$  (upper row) and non-Gaussian observed state  $u$  (lower row).

In the effective approximation of the nonlinear terms, we propose to use RNNs to replace the crucial nonlinear feedback terms in the PIDD-CG algorithm as

$$\begin{aligned}\mathcal{F}_u(t+1) &= \text{RNN}(u(t-m:t), \bar{v}(t-m:t), \mathcal{F}_u(t-m:t)), \\ \mathcal{G}_v(t+1) &= \text{RNN}(u(t-m:t), r_v(t-m:t), \mathcal{G}_v(t-m:t)).\end{aligned}\tag{S4}$$

Notice that we need to include the history of the nonlinear terms  $\mathcal{F}_u, \mathcal{G}_v$  in the inputs of the RNNs.

## A.2 Forecast results

### A.2.1 Training and lead time prediction

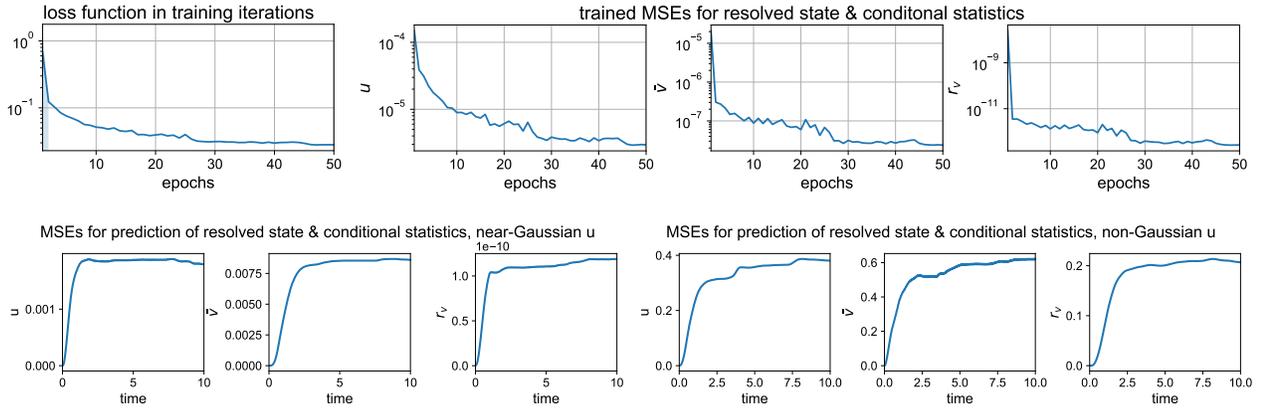


Figure S2: Training and lead-time prediction errors in the dyad model. The first row shows the iterations of training loss and training errors in the resolved state  $u$  and conditional mean and variance in the unresolved state  $v$ . The second row shows the development of lead time errors using the trained model in the two parameter regimes.

First, we train the RNNs proposed in (S4) by using the standard LSTM network. The loss function is from the information metric (16) in the main text and the model is trained in 50 epochs. The error evolution of the loss and the mean square errors (MSEs) in the target states are compared in the first row of Figure S2. Rapid convergence and accurate training are achieved. Then we confirm the trajectory prediction performance in the two tested statistical regimes. The second row of Figure S2 displays the prediction errors in the states with different lead times. The errors all saturate in small amplitudes, inferring accurate prediction for both the resolved state  $u$  and the conditional statistics for the unresolved state  $v$ . The truth and model prediction of the trajectory realizations are also compared in Figure S3. It confirms the good performance by using the PIDD-CG algorithm to recover the true trajectory by directly learning the nonlinear dynamics from data. Notice

that the near-Gaussian regime is usually easier to predict and can stay accurate with smaller errors for longer lead time prediction. In the trajectory prediction of the resolved state  $u$  in its non-Gaussian regime, larger errors will emerge as the model is iterated for longer time forecast (right panel of Figure S3).

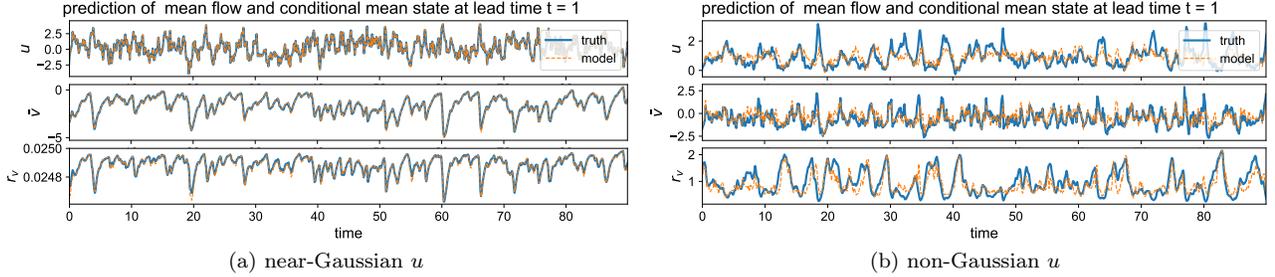


Figure S3: Lead time prediction of the resolved state  $u$  and conditional mean and variance in the unresolved state  $v$  in the two test regimes.

## A.2.2 Prediction of PDFs

Next, we use the PIDD-CG algorithm to predict the evolution of state PDFs from a given initial distribution. The initial distribution is taken as Gaussian and is very different from the final highly non-Gaussian equilibrium state. To show the true statistics as the reference solution, we carry out direct Monte-Carlo simulation with a large ensemble size  $N = 50000$ . In the PIDD-CG algorithm, we only take  $M = 100$  samples to recover the resolved space of  $u$ . The first row of Figure S4 shows the prediction of the final equilibrium PDFs in the two tested regimes. As is implied from the trajectory prediction in Figure S3, the near-Gaussian regime of  $u$  gives accurate recovery of the PDFs. On the other hand, the non-Gaussian regime of  $u$  is more challenging to capture the entire non-Gaussian statistics with a very small sample size. Nevertheless, the major statistical structures are still captured using the PIDD-CG algorithm. The second and third rows of Figure S4 show the predicted PDFs at several different lead time instants before the final statistical equilibrium. We observe the development of skewed PDFs in time in the transient states. Again, the efficient PIDD-CG algorithm is able to capture the statistical development of the transient PDFs. Especially in the non-Gaussian regime of  $u$ , the forecast is accurate at the starting time then errors are gradually developed as we try to predict the PDFs at a longer lead time due to the strong nonlinearity in the dynamics.

# B The barotropic topographic model with multiscale coupling

Next, we display detailed results about applying the PIDD-CG algorithm on the multiscale barotropic topographic model discussed as the major test model in the main text.

## B.1 Model description

### B.1.1 The starting model

The topographic barotropic flow is a prototype model in geophysics. It is expressed as follows [27],

$$\begin{aligned} \frac{\partial q}{\partial t} + \nabla^\perp \psi \cdot \nabla q &= \mathcal{D}(\Delta) \psi + F_q, \\ \frac{dU}{dt} + \int \frac{\partial h}{\partial x} \psi' &= -d_U U + F_U, \end{aligned} \quad (\text{S5})$$

which is defined in a two-dimensional domain  $D : \mathbf{x} = (x, y) \in [-\pi, \pi]^2$  with double periodic boundary conditions. In (S5),  $\mathcal{D}(\Delta)$  is the dissipation operator,  $h$  is the topographic effect, and  $F_q$  and  $F_U$  are external forcings. The state variable  $U$  represents the large-scale zonal flow velocity while  $q$  and  $\psi$  are the potential vorticity and the stream function, respectively. They are related by

$$q = q' + f = \nabla^2 \psi' + h + \beta y, \quad \psi = -U(t)y + \psi', \quad (\text{S6})$$

where the prime terms denote the fluctuations subject to the large-scale mean flow. The averaged integration in the mean dynamics  $U$  in (S5) is defined as  $f f d\mathbf{x} = \frac{1}{|D|} \int_D f d\mathbf{x}$ , where  $|D|$  is the total area of the domain. The topographic barotropic

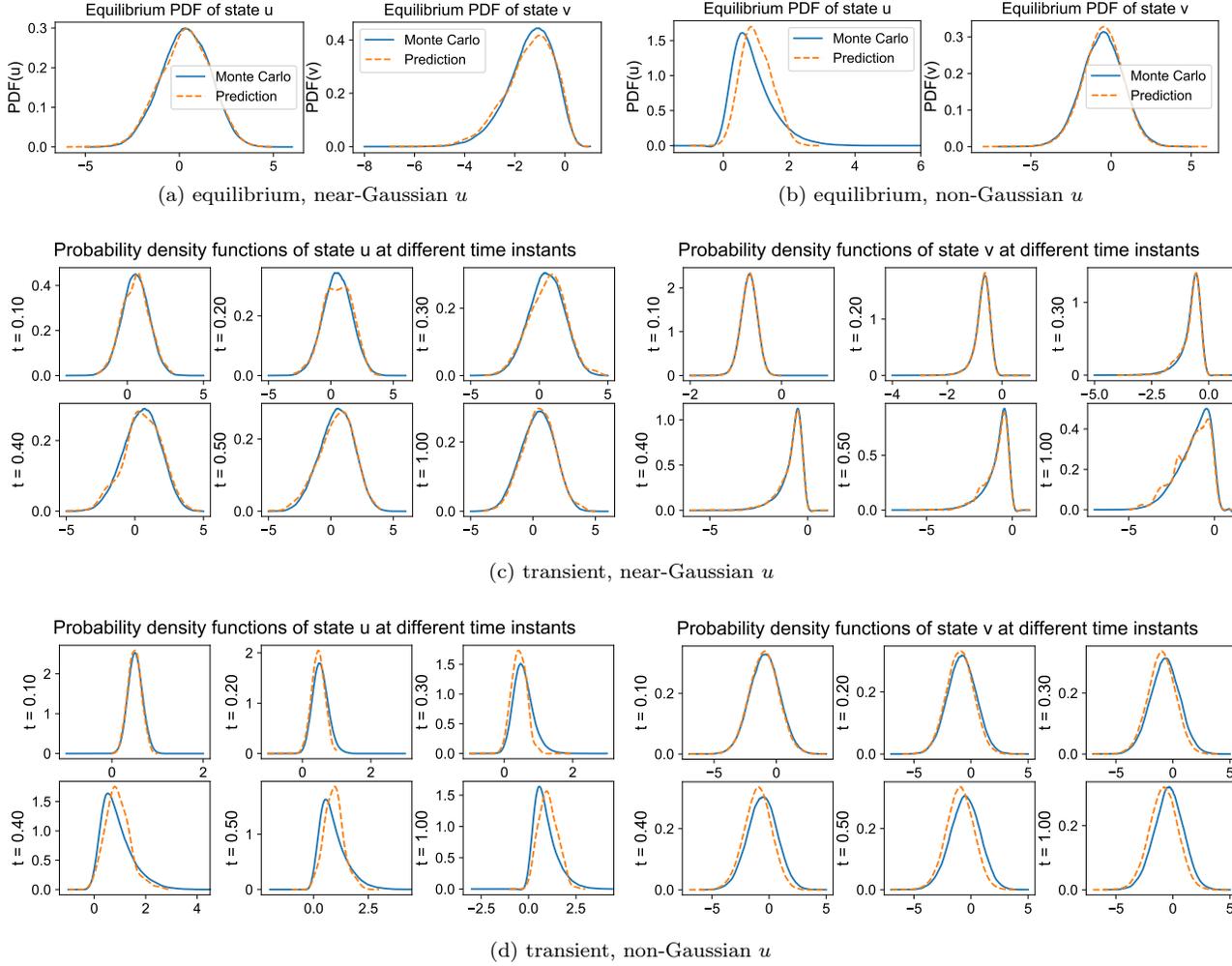


Figure S4: Prediction of equilibrium (first row) and transient (second and third rows) PDFs of the dyad states at different time instants before the equilibrium. The truth from Monte-Carlo samples is shown in solid blue line, and the reduced model prediction in dashed orange line. The two tested regimes with different statistics are compared.

flow model in (S5) is supplemented by a passive tracer model that characterizes the advection-diffusions of the transport of a tracer density field  $T(\mathbf{x}, t)$ , such that

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = -d_T T + \kappa \Delta T, \quad (\text{S7})$$

where  $\mathbf{u} = \nabla^\perp \psi$ ,  $d_T$  is the drag term and  $\kappa$  is the diffusion coefficient. The model (S5)–(S7) exhibits very rich dynamical and statistical features, such as the switching behavior between blocked and unblocked zonal flow regimes, non-Gaussian distributions and extreme events.

### B.1.2 The barotropic model with layered topography

A particularly interesting case of the above barotropic model (S5) is the one with layered topography, where the topography and stream function have the following expansion form

$$h(x, y) = \sum_k \hat{h}_k e^{ik\mathbf{l} \cdot \mathbf{x}}, \quad \psi(x, y, t) = \sum_k \hat{\psi}_k(t) e^{ik\mathbf{l} \cdot \mathbf{x}}. \quad (\text{S8})$$

The expansion is along one characteristic wavenumber direction  $\mathbf{l} = (l_x, l_y)$  with  $|\mathbf{l}| = 1$ . The corresponding full velocity field can be found combining the zonal mean flow and the fluctuations

$$\mathbf{u} = (U + u', v') = \left( U - il_y \sum_k k \hat{\psi}_k e^{ik\mathbf{l} \cdot \mathbf{x}}, il_x \sum_k k \hat{\psi}_k e^{ik\mathbf{l} \cdot \mathbf{x}} \right). \quad (\text{S9})$$

Then nonlinear coupling term,  $\nabla^\perp \psi \cdot \nabla q$ , vanishes under the above layered topography expansion. As a further simplification of the passive tracer model, we introduce a background mean gradient  $\boldsymbol{\alpha} = (\alpha_x, \alpha_y)$  on top of the tracer field fluctuations  $T'$  and a stochastic velocity field  $\mathbf{u}$  is assumed such that

$$\begin{aligned} T(\mathbf{x}, t) &= \boldsymbol{\alpha} \cdot \mathbf{x} + T'(\mathbf{x}, t), \\ \mathbf{u}(\mathbf{x}, t) &= (U(t), v(x, t)). \end{aligned} \quad (\text{S10})$$

Here the zonal cross-sweep  $U$  and the fluctuations  $v$  can be adopted from the topographic barotropic model solution. The same layered structure can be also assumed for the tracer fluctuation state so that we consider the tracer mean gradient  $\alpha_x \equiv 0, \alpha_y = \alpha$  with

$$T'(x, y, t) = \sum_k \hat{T}_k(t) e^{ik\mathbf{l} \cdot \mathbf{x}}. \quad (\text{S11})$$

The resulting equation gives a simplified formulation for the turbulent transport of passive tracer field

$$\frac{\partial T'}{\partial t} + U \frac{\partial T'}{\partial x} = -d_T T' + \kappa \frac{\partial^2 T'}{\partial x^2} - \alpha v(x, t). \quad (\text{S12})$$

The above equation for the tracer fluctuation field provides a judicious simplified formulation in modeling the tracer passive transport with many interesting statistical features such as intermittency and skewed statistics.

### B.1.3 The spectral formulation of the barotropic model with layered topography

Based on the above justifications, the original equations (S5) can be reformulated in the following form for each Fourier spectral mode

$$\frac{dU}{dt} = \sum_k h_k^* \hat{v}_k - d_0 U + \sigma_0 \dot{W}_0, \quad (\text{S13a})$$

$$\frac{d\hat{v}_k}{dt} = [-\gamma_{v,k}(U) + i\omega_{v,k}(U)] \hat{v}_k - l_x^2 \hat{h}_k U - d_{v,k} \hat{v}_k + \sigma_{v,k} \dot{W}_k, \quad (\text{S13b})$$

$$\frac{d\hat{T}_k}{dt} = [-\gamma_{T,k}(U) + i\omega_{T,k}(U)] \hat{T}_k - d_{T,k} \hat{T}_k - \alpha \hat{v}_k, \quad (\text{S13c})$$

where  $k$  is the wavenumber with  $|k| = 1, \dots, K$  and model parameters

$$\begin{aligned} \gamma_{T,k} &= d_T + \kappa k^2, & \omega_{T,k} &= -k(U + u) \\ \gamma_{v,k} &= 0, & \omega_{v,k} &= l_x(k^{-1}\beta - kU), & \sigma_{v,k} &= -ik^{-1}\sigma_k. \end{aligned} \quad (\text{S14})$$

Note that additional dependent parameters  $\gamma_{v,k}, \gamma_{T,k}$  are introduced in (S13) as extra parameterization for the unresolved multiscale interactions between fluctuation modes.

Notably, given one realization of  $U$ , the processes  $\hat{v}_k$  and  $\hat{T}_k$  in (S13) become conditionally linear and Gaussian. Thus, the system automatically fits into the general modeling framework (3) proposed in the main text.

### B.1.4 Step-by-step illustration of predicting the barotropic topographic model using the PIDD-CG algorithm

**Step 1. Phase space decomposition.** The fact that the large-scale zonal velocity  $U$  is observed offers a natural way for the phase space decomposition: the low-dimensional subspace contains only  $U$  while the remaining high-dimensional subspace involves all the Fourier modes for  $\hat{v}_k$  and  $\hat{T}_k$ . Putting into the general framework in (3), this means:

$$\mathbf{X} := U \quad \text{and} \quad \mathbf{Y} := (\mathbf{v}, \mathbf{T}) = (\hat{v}_1, \hat{v}_2, \dots, \hat{T}_1, \hat{T}_2, \dots).$$

**Step 2. Systematic multiscale statistical closure approximation of the large-scale dynamics in the low-dimensional subspace.** Since  $U$  is coupled with  $\hat{v}_k$  and  $\hat{T}_k$ , a suitable closure equation of  $U$  needs to be developed before applying the traditional MC method to forecast  $U$  in a closed intrinsic low-dimensional subspace.

Following the general framework in (5) in the main text, the fluctuation modes are decomposed into two subsets of the *resolved scales* for wavenumbers in  $\mathcal{I} = \{k : |k| \leq M\}$  and the long spectrum of the *unresolved scales* in  $\mathcal{I}^c = \{k : M < |k| \leq N\}$ , where the modes of  $\hat{v}_k$  and  $\hat{T}_k$  belonging to  $\mathcal{I}$  correspond to the state variable  $\mathbf{Y}_1$  in (5) while the remaining modes are  $\mathbf{Y}_2$ . Therefore, the mean flow equation (S13) in the topographic model can be rewritten as

$$\begin{aligned} \frac{dU}{dt} &= \sum_{k \in \mathcal{I}} h_k^* \bar{v}_k - d_0 U + \mathcal{M}_U + \mathcal{N}_U + \sigma_0 \dot{W}_0 \\ &:= \sum_{k \in \mathcal{I}} h_k^* \bar{v}_k - d_0 U + \mathcal{H}_U + \sigma_0 \dot{W}_0 \\ \mathcal{M}_U &= \sum_{k \in \mathcal{I} \cup \mathcal{I}^c} h_k^* (\hat{v}_k - \bar{v}_k), \quad \mathcal{N}_U = \sum_{k \in \mathcal{I}^c} h_k^* \hat{v}_k, \quad \mathcal{H}_U = \mathcal{M}_U + \mathcal{N}_U, \end{aligned} \tag{S15}$$

where  $\bar{v}_k$  is the conditional mean of  $v_k$  given the past trajectory of  $U$  that will be introduced in the next two steps. Clearly,  $\mathcal{H}_U = \mathcal{M}_U + \mathcal{N}_U$  corresponds to  $\mathcal{F}_{\mathbf{X}}$  in the general framework (5). Following (6),  $\mathcal{H}_U$  is approximated by a RNN that approximates the contribution from both the fluctuation part of the resolved modes and the entire unresolved modes,

$$\mathcal{H}_U(t+1) = \text{RNN} \left( U(t-m:t), \{\bar{v}_k(t-m:t)\}_{k \in \mathcal{I}}, \mathcal{H}_U(t-m:t) \right).$$

With the governing equation of  $\hat{v}_k$  being provided, the intrinsic dimension of the approximated governing equation of  $U$  is low. It therefore allows to use a MC simulation with a small number of ensembles  $N$  to forecast its PDF up to a given time instant, which is then smoothed using a kernel density estimation,

$$p(U) = \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J \tilde{p}(U^{\{j\}})$$

with  $\tilde{p}(U^{\{j\}})$  the  $j$ -th member from kernel density estimation that is associated with  $U^{\{j\}}$ .

**Step 3. Effective physics-informed conditional Gaussian mixture via data assimilation.** Each of the ensemble member from the MC simulation in Step 1 provides one trajectory of  $U$ , denoted by  $U^{\{j\}}$ . Conditioned on such a trajectory, there is one corresponding distribution of  $\mathbf{v}$  and  $\mathbf{T}$ , namely  $p(\mathbf{v}, \mathbf{T} | U^{\{j\}})$ . Note that  $p(\mathbf{v}, \mathbf{T} | U^{\{j\}})$  is a conditional Gaussian distribution for the layered topographic model (S13) since conditioned on  $U$  the processes of  $v_k$  and  $T_k$  are conditional linear. The joint distribution of  $U, \mathbf{v}$  and  $\mathbf{T}$  is thus given by

$$p(U, \mathbf{v}, \mathbf{T}) = \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J \tilde{p}(U^{\{j\}}) p(\mathbf{v}, \mathbf{T} | U^{\{j\}}). \tag{S16}$$

This corresponds to (9) in the main text of the general PIDD-CG forecast framework.

**Step 4. Time evolution of the conditional statistics in smaller-scale dynamics.** Following the general framework (10) or (12) in the main text, the time evolutions of the conditional mean and the conditional covariance for the

barotropic model with layered topography are given as follows,

$$\frac{d\bar{v}_k}{dt} = -l_x^2 \hat{h}_k U + [i\omega_{v,k}(U) - d_{v,k}] \bar{v}_k + \sigma_0^{-2} \mathcal{F}_U \cdot \mathcal{G}_{v,k}, \quad (\text{S17a})$$

$$\frac{d\bar{T}_k}{dt} = [i\omega_{T,k}(U) - d_{T,k}] \bar{T}_k - \alpha \bar{v}_k + \sigma_0^{-2} \mathcal{F}_U \cdot \mathcal{G}_{c,k}, \quad (\text{S17b})$$

$$\frac{dr_{v,k}}{dt} = -2d_{v,k} r_{v,k} + \sigma_{v,k}^2 - \sigma_0^{-2} |\mathcal{G}_{v,k}|^2, \quad (\text{S17c})$$

$$\frac{dr_{T,k}}{dt} = -2d_{T,k} r_{T,k} - \alpha (c_k + c_k^*) - \sigma_0^{-2} |\mathcal{G}_{c,k}|^2, \quad (\text{S17d})$$

$$\begin{aligned} \frac{dc_k}{dt} = & - (d_{v,k} + d_{T,k}) c_k + i [\omega_{T,k}(U) - \omega_{v,k}(U)] c_k \\ & - \alpha r_{v,k} - \sigma_0^{-2} \mathcal{G}_{c,k} \cdot \mathcal{G}_{v,k}^*, \end{aligned} \quad (\text{S17e})$$

where  $\bar{v}_k$  and  $\bar{T}_k$  are the conditional mean of  $\hat{v}_k$  and  $\hat{T}_k$  while  $r_{v,k}$ ,  $r_{T,k}$  and  $c_k$  are the conditional variance of  $\hat{v}_k$ ,  $\hat{T}_k$  and the cross covariance between  $\hat{v}_k$  and  $\hat{T}_k$ , respectively. In (S17), only the leading modes  $|k| \leq M$  are resolved explicitly. In addition, the covariance equations have been further simplified by applying a block diagonal approximation, where each block has the size  $2 \times 2$  including the cross-correlation of the mode with the same wavenumber  $c_k$ . The central (quasi) linear dynamics are explicitly expressed while the complicated nonlinear functions are denoted by  $\mathcal{F}_U$ ,  $\mathcal{G}_{v,k}$  and  $\mathcal{G}_{c,k}$  with

$$\begin{aligned} \mathcal{F}_U &= \dot{U} - \sum_m \hat{h}_m^* \bar{v}_m + d_0 U, \\ \mathcal{G}_{v,k} &= \sum_m \hat{h}_m r_{v,km}, \quad \mathcal{G}_{c,k} = \sum_m \hat{h}_m c_{km}, \end{aligned}$$

where  $\mathcal{F}_U$  corresponds to  $\mathcal{F}_\mathbf{X}$  while  $\mathcal{G} = (\mathcal{G}_{v,k}, \mathcal{G}_{c,k})_{k \in \mathcal{I}}$  correspond to  $\mathcal{G}_\mathbf{Y}$  in (11) of the main text.

**Step 5. Data-driven modeling of the nonlinear feedbacks in conditional statistics via recurrent neural networks.** Finally, corresponding to (13), the complicated nonlinear functions are approximated by RNNs,

$$\begin{aligned} \mathcal{F}_U(t+1) &= \text{RNN}(U(t-m:t), \{\bar{v}_k(t-m:t)\}_k, \mathcal{F}_U(t-m:t)), \\ \mathcal{G}(t+1) &= \text{RNN}(U(t-m:t), \{r_k(t-m:t), c_k(t-m:t), \mathcal{G}_k(t-m:t)\}_{k \in \mathcal{I}}). \end{aligned} \quad (\text{S18})$$

## B.2 Model parameters

This section includes the basic numerical setup and model parameters for the barotropic topographic model (S5) as well as the corresponding neural network architecture for the unresolved subscale processes in (S18).

### B.2.1 Parameters in the topographic barotropic model

The barotropic topographic model displays distinct dynamical and statistical features with different values of the white noise forcing amplitudes  $\sigma_U$  and  $\sigma_{v,k}$ . In particular, the two representative regimes in the main text corresponding to the highly non-Gaussian and the nearly Gaussian regimes, and the PDFs are equipped with the following parameters:

- *Strongly non-Gaussian regime:* The zonal mean flow is strongly forced with white noise strength  $\sigma_U = \frac{1}{\sqrt{2}}$  while only small noises  $\sigma_{v,k} = \frac{k^{-1}}{20\sqrt{2}}$  are added to the fluctuation modes.
- *Near-Gaussian regime:* The fluctuation modes are subject to relatively stronger noise forcing with strength  $\sigma_{v,k} = \frac{k^{-1}}{\sqrt{2}}$  compared with the noise strength in the zonal mean flow  $\sigma_U = \frac{1}{2\sqrt{2}}$ .

Notice that even in the near-Gaussian regime, nonlinear dynamics takes a dominant role in the multiscale interactions. Therefore, the feedbacks from the fluctuations cannot be neglected.

Next, the topography structure is given by

$$h = H_1 (\cos x + \sin x) + H_2 (\cos 2x + \sin 2x) + \sum_{k=3}^K k^{-2} e^{i\theta_0} + c.c.,$$

with the first two dominant leading modes  $H_1 = 1, H_2 = \frac{1}{2}$ , and initial phase parameter  $\theta_0 = -\frac{\pi}{4}$ . The ‘c.c.’ denotes the complex conjugate. This topography can be viewed as an analog to a long north-south ridge [38]. For the rest parts of the model parameters, a uniform damping is adopted in both the mean and the fluctuation modes  $d_U = d_{v,k} \equiv 0.0125$ . The damping, diffusion, and mean cross-sweep for tracer field are  $d_T = 0.1, \kappa_T = 0.001, \alpha = 1$ , respectively, and the rotation

parameter is  $\beta = 1$ . These parameter values are derived from non-dimensionalization of the real physics measurements of the characteristic scales [27].

For the numerical integration in the true model to generate the simulated data, the standard 4th-order Runge-Kutta scheme with time step size  $dt = 1 \times 10^{-3}$  is adopted, which is essential to maintain stability due to the stiffness in the small-scale flow and tracer dynamics, as well as the full conditional statistical equations. In contrast, only a forward Euler scheme with a much larger time step size  $\Delta t = 0.01$  is utilized in the RNN and thus the numerical cost is further reduced. Notice that this large time step cannot guarantee the numerical stability in the original model.

## B.2.2 Parameters in the neural network

Recurrent neural networks (RNNs) offer the desirable structure to incorporate temporal processes of sequential data, and keep tracking of hidden processes. The *long short-time memory* (LSTM) network is a special RNN that is useful to recover the time-series including very long time correlations. The LSTM designed to learn the multi-scale temporal structures overcoming the problem of vanishing gradients. In the computational cell of the LSTM network, it consists of the basic building cell as

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f), \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i), \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes \tanh(W_c x_t + U_c h_{t-1} + b_c), \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + V_o c_t + b_o), \\ h_t &= o_t \otimes \tanh(c_t). \end{aligned} \tag{S19}$$

Above, the  $\sigma_g = \frac{1}{1+e^{-x}}$  is the sigmoid activation function, and  $\otimes$  represents the element-wise product. The model cell includes forget, input, and output gates  $f_t, i_t, o_t$ , and the cell state  $c_t$ . The hidden process  $\{h_{t-m}, \dots, h_{t-1}, h_t\}$  represents the time-series of the unresolved process. The final output data is given by a final linear layer  $y_t = W_x h_t$  applying on the final state of the hidden process.

The LSTM net is constructed from  $m$  LSTM cells  $\mathbf{h}_{i+1} = \text{Lc}(\mathbf{x}_i, \mathbf{h}_i; \mathbf{W})$  with the same structure and parameters  $\mathbf{W}$ . The cells are connected by the intermediate hidden state  $\mathbf{h}_i \in \mathbb{R}^h$ . Every LSTM cell takes in the input data  $\mathbf{x}_i$  at the  $i$ -th step and the output  $\mathbf{h}_i$  from the previous adjacent cell, and gives out the inner hidden state  $\mathbf{h}_{i+1}$  to be used for prediction of the next state. The full LSTM chain is connected by  $m$  sequential cell structures, that is,

$$\mathbf{h}_m = \text{Lc}^{(m)}\{\mathbf{h}_0; \mathbf{x}_{t-m}, \dots, \mathbf{x}_{t-i}, \dots, \mathbf{x}_{t-1}\} \equiv \text{Lc}(\mathbf{x}_{t-1}) \circ \dots \circ \text{Lc}(\mathbf{x}_{t-i}) \circ \dots \circ \text{Lc}(\mathbf{x}_{t-m})(\mathbf{h}_0). \tag{S20}$$

Above, the data at different time instants  $\mathbf{x}_i$  is fed into the corresponding LSTM cell, and  $\mathbf{h}_i$  is the hidden state as the output of the previous cell and input for the next cell. For simplicity, the initial value of the hidden state is often set as zero,  $\mathbf{h}_0 = 0$ . The final output  $\mathbf{h}_m$  from the last step of the LSTM chain goes through another single layer fully connected network to give the model approximation of the dynamical increment for  $f$

$$f_m^M = \sigma(\mathbf{W}^f \mathbf{h}_m + \mathbf{b}^f), \tag{S21}$$

where  $\mathbf{W}^f, \mathbf{b}^f$  are the model coefficients in the final layer, and  $\sigma$  is a nonlinear activation function adopting the rectified linear unit (ReLU).

The above standard architecture of the LSTM is applied to estimate the feedback  $\mathcal{F}_U, \mathcal{G}$  to the leading resolved scales in (S15) and (S17) in (S18) to learn the embedded processes. In addition, a residual structure is adopted in the LSTM neural network to update the correlated time sequence

$$\theta_{i+1} = \theta_i + \mathcal{F},$$

where  $\mathcal{F}$  is the LSTM output for the increment.

The major hyperparameters of the neural network used in the tests are as follows. The LSTM chain consists of  $m = 100$  repeating cells with the same structure, taking a time sequence of time length  $t = 1$  which is about the decorrelation time of the system state. The dimension of the hidden state is taken as  $h_v = 100$  for the conditional variance equations and  $h_m = 20$  for the conditional mean feedback. The LSTM output is reiterated forward for  $n = 10$  steps (only a short forward time of  $t = 0.1$ ) to account for the integrated error along the time integration steps. The optimization is carried out by stochastic gradient descent using the ADAM scheme for a training batch of size 100 samples. During the training process, a total of 100 epochs are repeated starting from the learning rate  $\text{lr} = 1 \times 10^{-3}$  which is reduced three times to half of its original values at epoch number 25, 50 and 75. In the training data using the neural network, we pick a larger integration time step  $\Delta t = 10dt = 0.01$ . The reduced-order model can be integrated using the simple forward Euler scheme with the large time step  $\Delta t$  thanks to the autocorrections directly learned during the training process of the LSTM neural network.

## B.3 Forecast results

Finally, we display detailed test results for the training and prediction using the PIDD-CG algorithm of the barotropic topographic model with strong non-Gaussian and intermittent dynamics.

layered topo.	$N$	$\beta$	$H_1$	$H_2$	$(l_x, l_y)$	$d_T$	$\kappa_T$	$\alpha$	$d_0$	$\sigma_0$
$H_1(\cos x + \sin x) +$ $H_2(\cos 2x + \sin 2x)$	10	1	1	0.5	(1, 0)	0.1	0.001	1	0.0125	$\frac{1}{2\sqrt{2}}$

Table S1: Standard model parameters for the barotropic topographic model simulations.

### B.3.1 Transition in true model statistics

In the numerical test, we consider different statistical regimes with the inclusion of damping and stochastic forcing effects. Especially, we would like to check the transition in statistics in the flow and tracer field solutions with varying model parameters. The standard model parameters used in the tests are listed in Table S1 according to the reference values proposed in [37]. The same damping amplitude  $d_U = d_k = d_0$  is applied to all the spectral modes and two sets stochastic forcing strength are considered for i) the near-Gaussian regime  $\sigma_U = \sigma_0, \sigma_k = 2\sigma_0$ ; and ii) the non-Gaussian regime  $\sigma_U = 2\sigma_0, \sigma_k = 0.1\sigma_0$ . In the near-Gaussian regime, the zonal flow  $U$  is subject to a smaller white noise forcing compared with the small scale modes forcing  $\sigma_U < \sigma_k$ . Both flow and tracer modes show PDFs close to Gaussian. In the non-Gaussian regime, a stronger forcing in the zonal flow field  $\sigma_k < \sigma_U$  gives a negatively skewed zonal state  $U$ , and the small scale modes become strongly fat tailed. The corresponding passive tracer field in this case also displays fat tails and also a large skewness in the leading modes.

We first display the key features in the two test regimes with distinct statistics. The equilibrium energy spectra in the flow and tracer modes as well as the autocorrelation functions are shown in Figure S5. The time development of the marginal PDFs of the zonal flow  $U$  and the leading fluctuation and tracer modes  $\hat{v}_1, \hat{v}_2$  and  $\hat{T}_1, \hat{T}_2$  at different time instants until equilibrium are compared in Figure S6 for the near-Gaussian and non-Gaussian regime respectively. First, in the equilibrium energy spectra it is observed that even though the first two leading modes are the most energetic, the other smaller scale modes still contain large amount of energy with strong feedback to the mean flow thus cannot be directly neglected in the reduced model approximation. This confirms the crucial role to consider the conditional Gaussian structure in the PIDD-CG algorithm to achieve accurate statistical forecast. In addition to the multiscales in spatial modes, the time series also display multiscale structures as shown in the autocorrelation functions. The zonal flow  $U$  has a much slower decay in time correlation compared with the rapid decay in the fluctuation modes. From the comparison of the time-series in the two cases, the multiscale structure and competition between the blocked and unblocked regimes is also observed. Especially, we see the generation of high skewness in the zonal mean flow  $U$  is from the strong zonal transport with suppressed fluctuations. These multiscale features are consistent with the time series in Figure 2 of the main text.

To see the distinct statistical features in the two test regimes, Figure S6 illustrates the time evolution of the marginal PDFs in time up to the final equilibrium state. Both regimes start with a Gaussian initial distribution with accurate observation in  $U$  (thus with zero variance) and the fluctuation modes are sampled from the conditional Gaussian distribution. In the near-Gaussian regime, the marginal PDFs of the states rapidly develop into the equilibrium steady state with all near-Gaussian structures. In contrast, in the non-Gaussian regime, strongly skewed PDFs are gradually developed in time with highly non-Gaussian features.

### B.3.2 Training and lead time prediction

The PIDD-CG algorithm includes the data-driven component to learn the unresolved feedbacks from data. The standard LSTM network is used and thus a training stage is required. In the training process using the neural network, we pick a larger integration time step  $\Delta t = 10dt = 0.01$ . The training output is reiterated recurrently for  $n = 10$  times to improve the stability of the scheme.

First, we show the convergence in the training stage. In Figure S7, the training loss and mean square errors are displayed during the training iterations with 100 epochs. It can be seen that the training error drops rapidly during the first few iterations and is quickly saturated at a low level. The corresponding errors in the conditional mean and covariance can also be minimized very quickly. Furthermore, we compare the improvement with the reiterated multistep forecast  $n = 10$  compared with single step update  $n = 1$  in the training calibration. The multi-step model achieves a higher accuracy during training and is faster to converge with a fewer number of iterations.

Next, we check the lead time prediction using the trained model. The neural network model prediction is in general challenging with growing imperfect errors in time due to the high model uncertainty and strong internal instability. The prediction accuracy gradually grows larger as the errors accumulate for the prediction in longer lead time shown in Figure S8 (also in the trajectory predictions in Figure 5 of the main text). Again, the multi-step time integration shows much higher stability during longer time iterations and maintains high accuracy beyond the decorrelation time. In comparison, if only the one-step updating scheme is used in the training stage, the prediction keeps accurate for short leading time (around  $T < 0.5$ ) while it quickly diverges to much larger errors when the leading time becomes larger. This is related with the inherent difficulty in the unstable numerical integration with this large time step  $\Delta t$ .

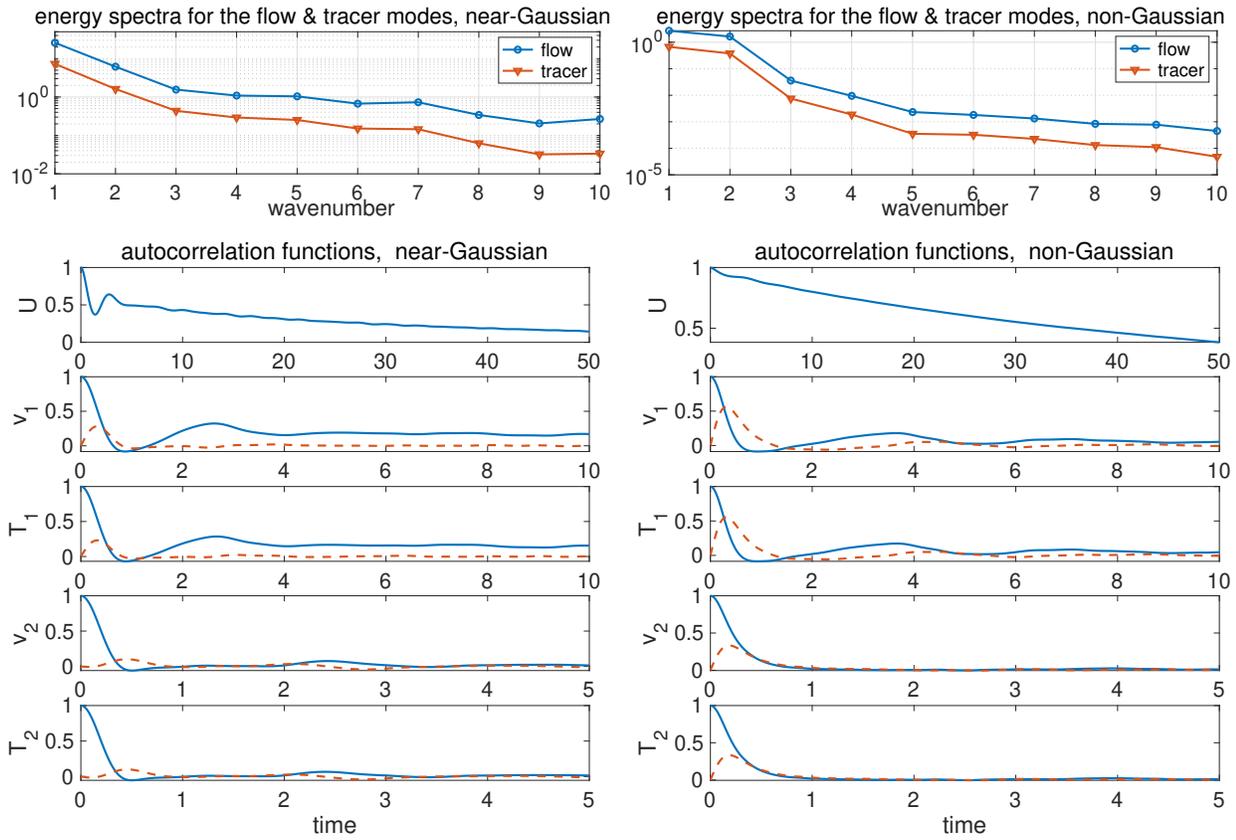


Figure S5: Equilibrium energy spectrum and autocorrelation functions of the barotropic topographic model in the near-Gaussian (left) and non-Gaussian (right) regimes.

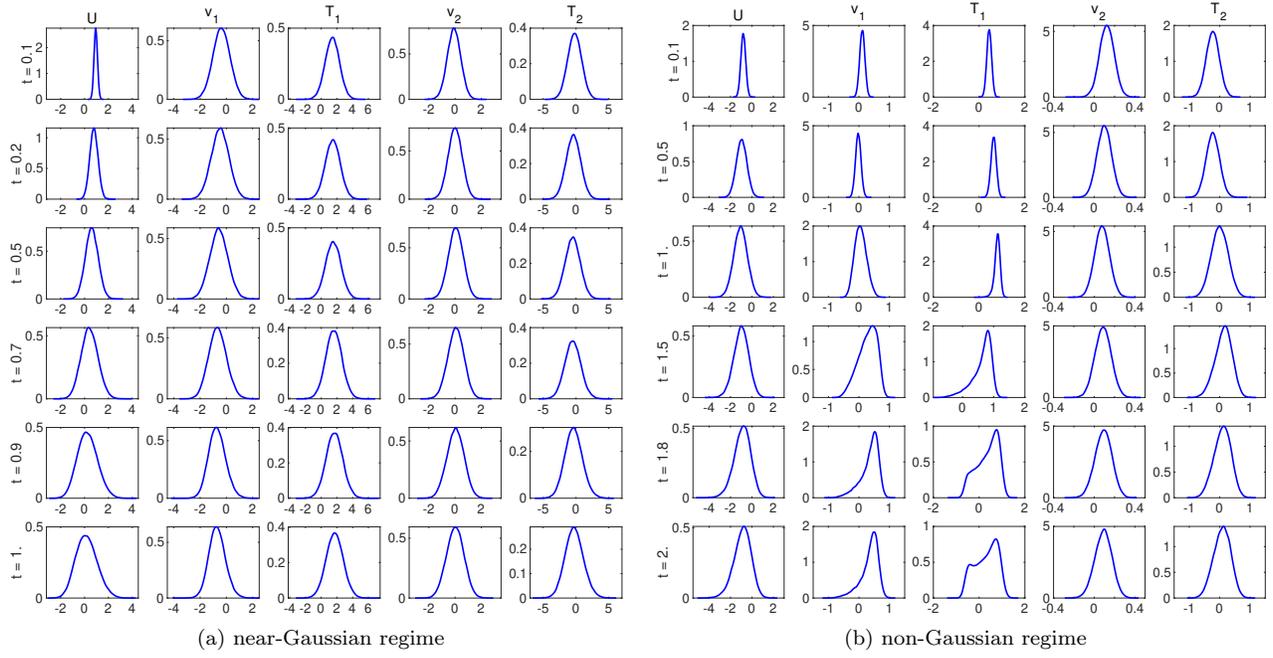


Figure S6: Marginal PDFs for the zonal flow  $U$  and the leading flow and tracer modes  $\hat{v}_1, \hat{v}_2$  and  $\hat{T}_1, \hat{T}_2$  at several time instants until equilibrium state is reached in the near-Gaussian (left) and non-Gaussian (right) regime.

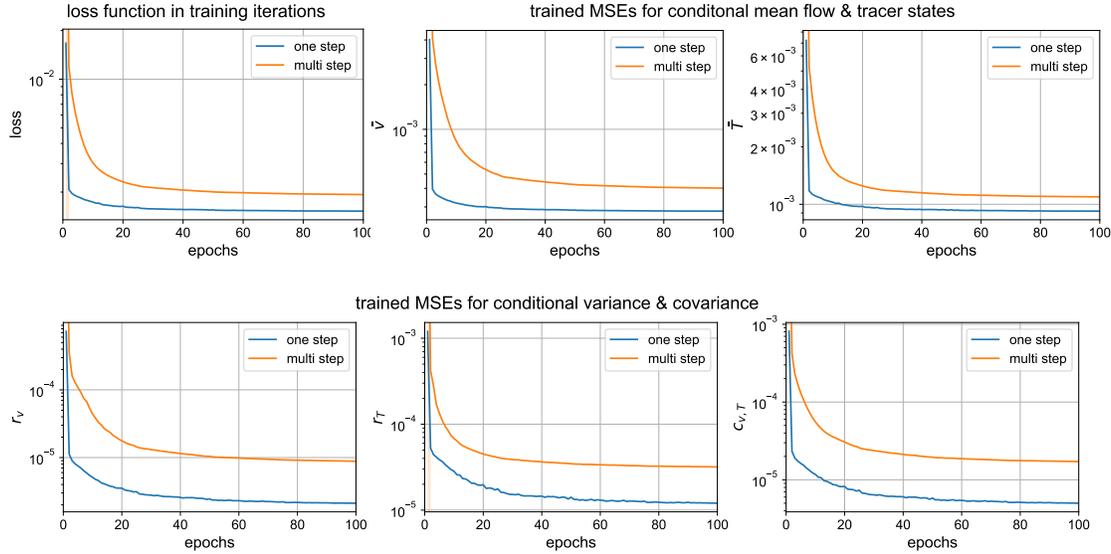


Figure S7: Errors in the loss function and the MSEs in conditional mean, variance, and covariance during training iterations.

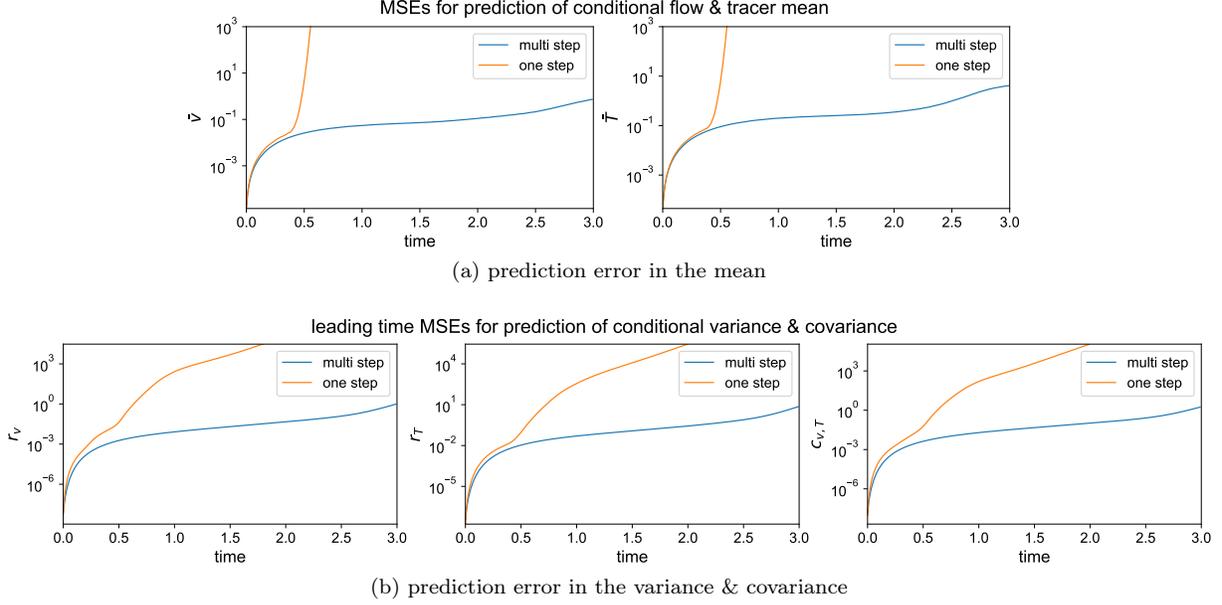


Figure S8: MSEs with different lead time predictions for errors in the conditional mean and variance. The two trained models with different iterating forward steps are compared.

### B.3.3 Prediction of the transient PDFs

Finally, we display more detailed prediction results of using the PIDD-CG algorithm to efficiently capture the transient PDFs of key model states. Using the same initial distribution as in the main text, Figure S9-S12 display the predicted marginal PDFs of the resolved states and the joint distributions between the zonal flow  $U$  and the first two leading flow and tracer modes  $\hat{v}_1, \hat{v}_2$  and  $\hat{T}_1, \hat{T}_2$  in the transient states before equilibrium is reached. Especially, we observe the development of non-Gaussian features from the initial mixed Gaussian state. Notice the difference scales in the values of  $U$  and leading modes at different time instants. The direct Monte-Carlo simulation requires a sample size of 50000 particles to capture the PDFs in accuracy, while in contrast the PIDD-CG algorithm only needs  $N = 100$  samples to achieve comparable accuracy with the truth. The PIDD-CG algorithm maintains high accuracy in capturing the highly non-Gaussian statistics regardless of the relatively high full dimension of the system. On the other hand, in the near-Gaussian regime, the convergence to equilibrium is faster and is also accurately captured with accuracy in the much efficient PIDD-CG algorithm.

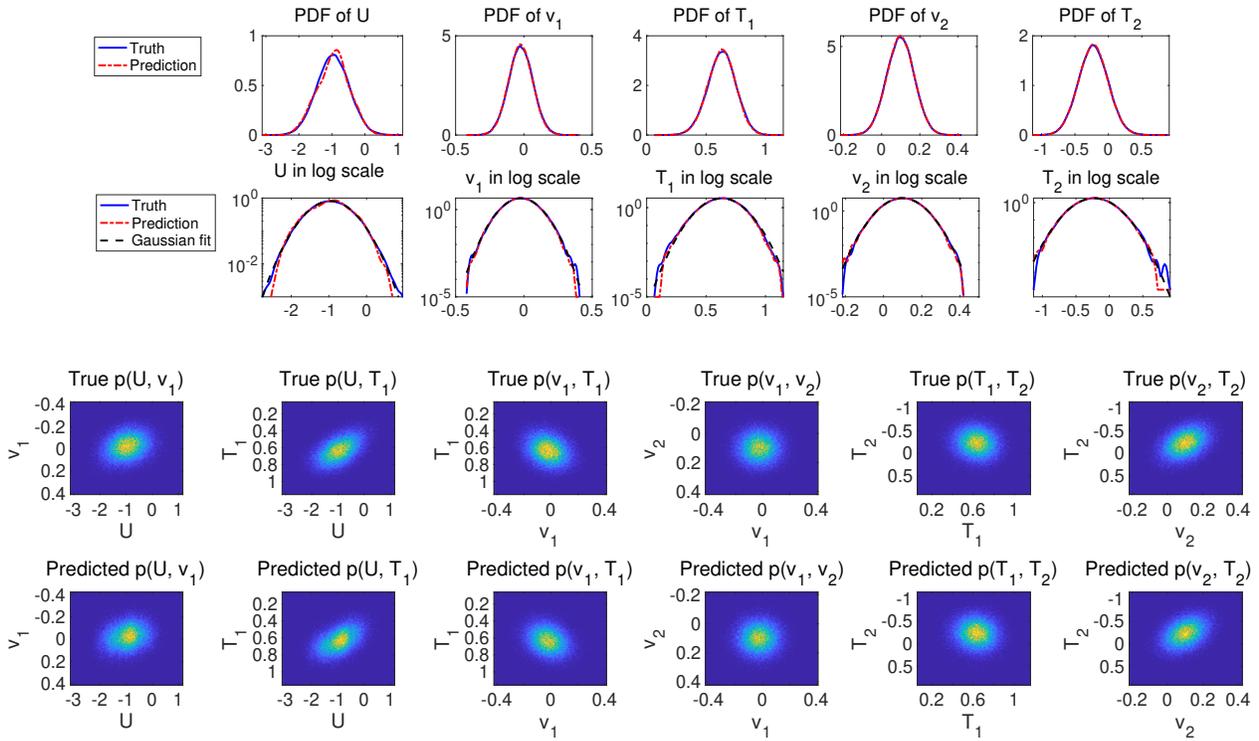


Figure S9: Prediction of the marginal PDFs and joint PDFs in the non-Gaussian regime at lead time  $t = 0.5$ .

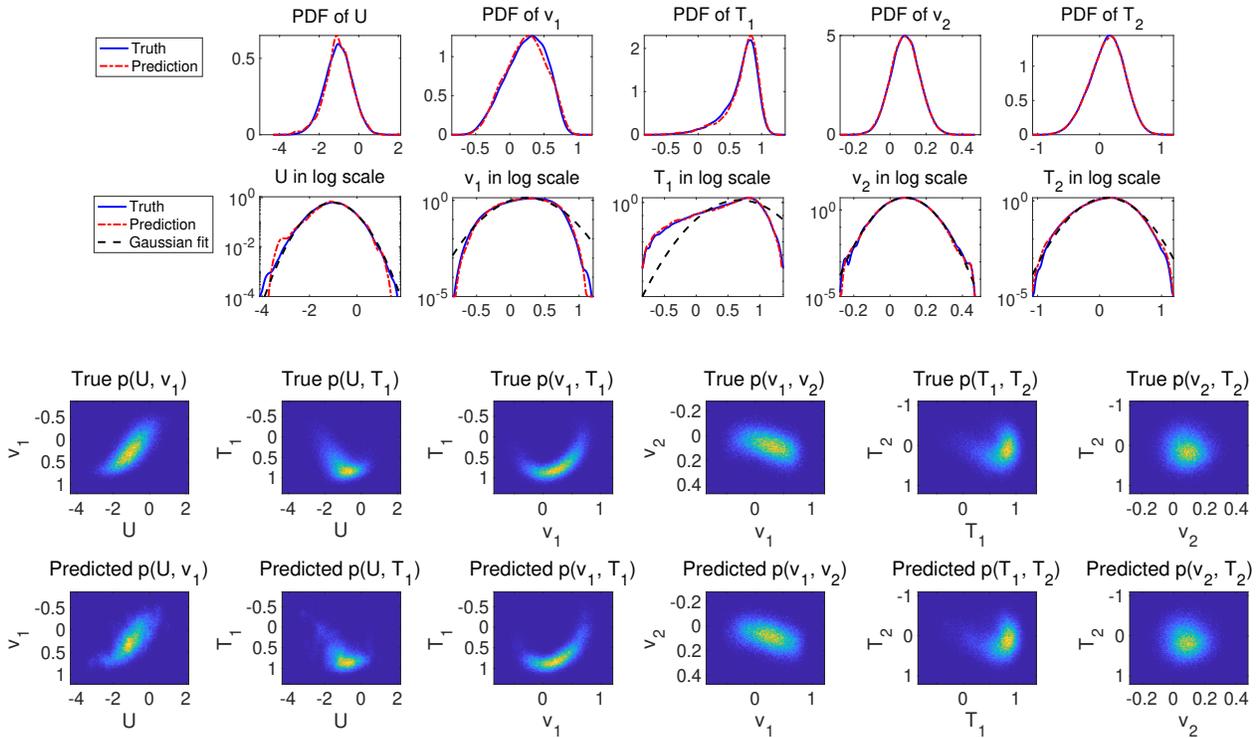


Figure S10: Prediction of the marginal PDFs and joint PDFs in the non-Gaussian regime at lead time  $t = 1.5$ .

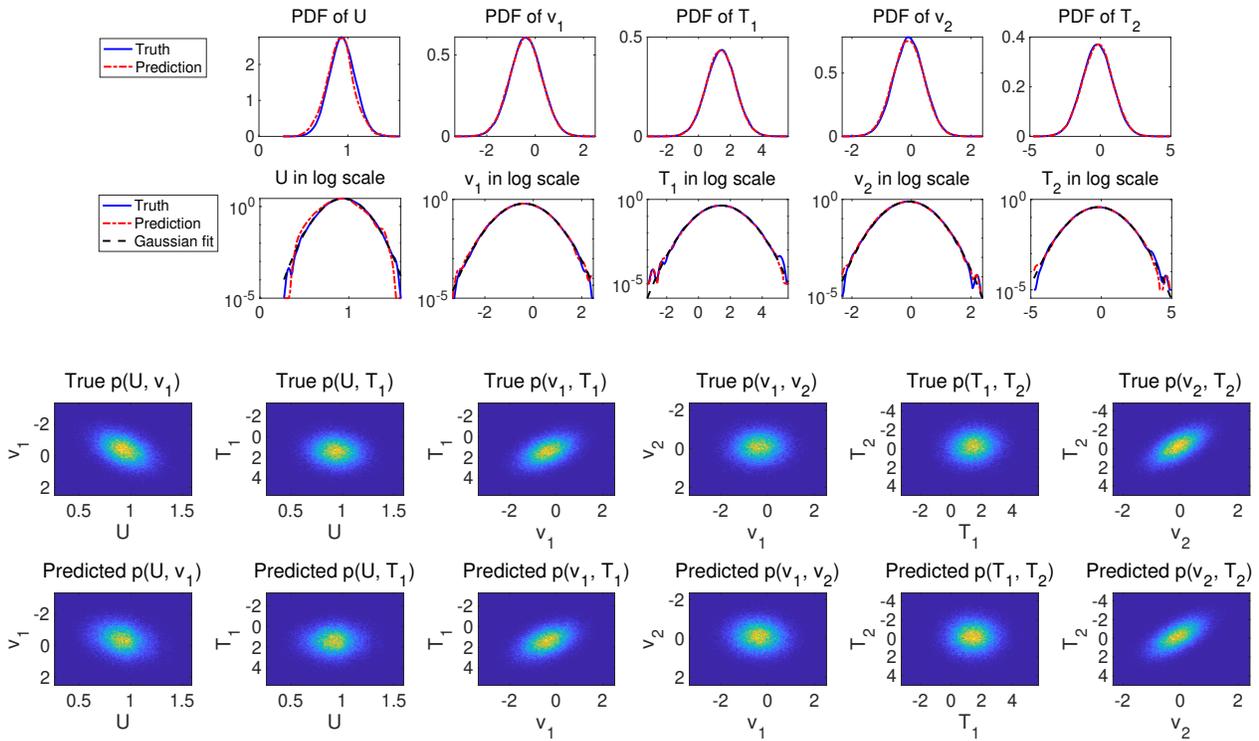


Figure S11: Prediction of the marginal PDFs and joint PDFs in the near-Gaussian regime at lead time  $t = 0.1$ .

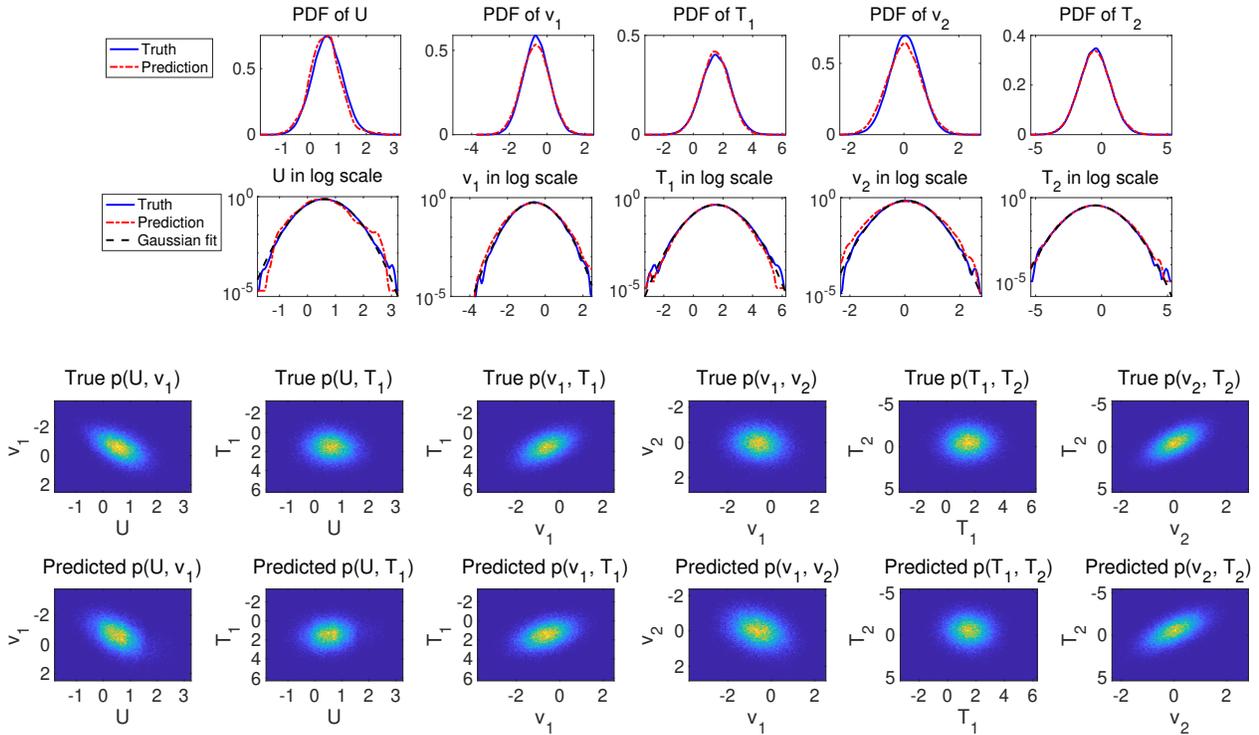


Figure S12: Prediction of the marginal PDFs and joint PDFs in the near-Gaussian regime at lead time  $t = 0.5$ .