Model Error, Information Barriers, State Estimation and Prediction in Complex Multiscale Systems

Andrew J. Majda [1,2] **and Nan Chen** [1,]*

[1]   Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA; jonjon@cims.nyu.edu

[2]   Center for Prototype Climate Modeling, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi 129188, UAE

*   Correspondence: nan.chen@nyu.edu or chennan@cims.nyu.edu

Abstract: Complex multiscale systems are ubiquitous in many areas. This research expository article discusses the development and applications of a recent information-theoretic framework as well as novel reduced-order nonlinear modeling strategies for understanding and predicting complex multiscale systems. The topics include the basic mathematical properties and qualitative features of complex multiscale systems, statistical prediction and uncertainty quantification, state estimation or data assimilation, and coping with the inevitable model errors in approximating such complex systems. Here, the information-theoretic framework is applied to rigorously quantify the model fidelity, model sensitivity and information barriers arising from different approximation strategies. It also succeeds in assessing the skill of filtering and predicting complex dynamical systems and overcomes the shortcomings in traditional path-wise measurements such as the failure in measuring extreme events. In addition, information theory is incorporated into a systematic data-driven nonlinear stochastic modeling framework that allows effective predictions of nonlinear intermittent time series. Finally, new efficient reduced-order nonlinear modeling strategies combined with information theory for model calibration provide skillful predictions of intermittent extreme events in spatially-extended complex dynamical systems. The contents here include the general mathematical theories, effective numerical procedures, instructive qualitative models, and concrete models from climate, atmosphere and ocean science.

Contents

1. Introduction

Complex multiscale turbulent dynamical systems are ubiquitous in geoscience, engineering, neural science and material science [1–7]. They are characterized by a large dimensional state space and a large dimension of strong instabilities which transfer energy throughout the system. Key mathematical issues are their basic mathematical structural properties and qualitative features [2,3,8,9], statistical prediction and uncertainty quantification (UQ) [10–12], state estimation or data assimilation [13–17], and coping with the inevitable model errors that arise in approximating such complex systems [10,18–21]. Understanding and predicting complex multiscale turbulent dynamical systems involve the blending of rigorous mathematical theory, qualitative and quantitative modelling, and novel numerical procedures [2,22].

One of the central difficulties in studying these complex multiscale turbulent dynamical systems is that either the dynamical equations for the truth are unknown due to the lack of physical understanding or the resolution in the models is inadequate due to the current computing power [1,13,18,23–25]. Therefore, understanding the model error from the imperfect dynamics as well as the coarse-grained processes becomes important. From both the theoretical and practical point of view, the following issues are of great interest.

1. How to measure the skill (i.e., the statistical accuracy) of a given imperfect model in reproducing the present states and predicting the future states in an unbiased fashion?
2. How to make the best possible estimate of model sensitivity to changes in external or internal parameters by utilizing the imperfect knowledge available of the present state? What are the most sensitive parameters for the change of the model status given uncertain knowledge of the present state?
3. How to design cheap and practical reduced models that are nevertheless able to capture both the main statistical features of nature and the correct response to external/internal perturbations?
4. How to develop a systematic data-driven nonlinear modeling and prediction framework that provides skillful forecasts and allows accurate quantifications of the forecast uncertainty?
5. How to build effective models, efficient algorithms and unbiased quantification criteria for online data assimilation (state estimation or filtering) and prediction especially in the presence of model error?

Recently, an information-theoretic framework has been developed and is applied together with other mathematical tools to address all the issues mentioned above [10,26–34]. This information-theoretic framework provides an unbiased way to quantify the model error and model fidelity [18,35–37] in complex nonlinear dynamical systems, which in turn offers a systematic procedure for model selection and parameter estimation within a given class of imperfect models [1,26–28]. The information-theoretic framework is also capable of estimating the model sensitivity in response to the changes in both internal and external parameters [27,28]. Practically, by incorporating the so-called fluctuation–dissipation theorem for the linear statistical response [38–40], the information-theoretic framework allows an extremely efficient approach to assess the model sensitivity [27,28,41]. Such a sensitivity analysis becomes particularly useful in for example detecting the climate change and preventing the occurrence of undesirable extreme events [42,43]. The combination of the model fidelity and the model sensitivity then provides important guidelines for developing reduced-order models [11,44,45] and data-driven prediction strategies using

physics-constrained nonlinear stochastic models [46–48]. Applying the information-theoretic framework for model calibration, the reduced-order models with suitable model structures are able to capture both the key dynamical and statistical features as well as the crucial nonlinear and non-Gaussian characteristics such as intermittency and extreme/rare events as observed in nature.

Nevertheless, the choice of the reduced or simplified models plays a crucial role in approximating nature. Within an improper model family, even the best model with the most elaborate calibration will result in a large model error. This is known as the information barrier and can be quantified by the information-theoretic framework [27,49–52]. In fact, the information-theoretic framework allows a rigorous decomposition of the total model error into an intrinsic barrier and an actual model error. The latter can be eliminated or at least be minimized to a negligible level by the information-optimization criterion [27,28]. Quantifying such information barriers have both theoretic and practical importance. It indicates the futility of model calibration if the information barrier is significant. It can also be used as a guidance to expand the model family of reduced models for the improvement of imperfect models. Note that information barriers appear in both the model fidelity and model sensitivity. A model with perfect model fidelity can still have a significant information barrier in response to the internal/external perturbation and in short term predictions [27].

Another important application of the information-theoretic framework is that it provides a novel and unbiased approach to assess the online data assimilation/filtering and prediction skill in complex multiscale dynamical systems [31,53–58]. The traditional path-wise measurements such as the root-mean-square error and pattern correlation [59,60] are misleading in assessing the model error in both filtering and prediction [31,61]. In fact, these traditional measurements completely fail to quantify the ability of the imperfect models in reproducing the extreme events in nature even in the linear and Gaussian setup. In addiction, these traditional path-wise measurements take into account the information only up to the second order statistics and therefore they have no skill in quantifying the features of intermittency and non-Gaussian probability density functions (PDFs) as well as other salient characteristics in nonlinear multiscale turbulent dynamical systems. On the other hand, the information-theoretic framework combining different information measurements is able to quantify the model error in an unbiased fashion and succeeds in assessing the ability of imperfect models in reproducing both the Gaussian and non-Gaussian features in filtering and forecasting complex nonlinear dynamical systems [31,61].

In practice, due to the incomplete knowledge and the limited computing power for dealing with the complex nonlinear turbulent dynamical systems of nature, reduced-order models are often designed for the state estimation and prediction [62–68]. Parameterizations of unresolved variables and coarse-grained processes are typically involved in the reduced-order models [69–72], which result in large uncertainties. Therefore, the reduced-order models aim at capturing the statistical features rather than a single realization of random trajectories of the complex nonlinear turbulent dynamical systems. Among all types of the reduced-order models, linear tangential approximations and Gaussian closure models are widely used in approximating the time evolutions of the statistics of nature [71,73–75]. Despite their simplicity and skillful behavior in some scenarios, these crude approximations fail to capture many crucial features in nature that result from the nonlinear interactions between different variables or nonlinear feedback within different scales. Therefore, nonlinear and non-Gaussian closure becomes important in describing the turbulence [76–79]. Recently, a new framework of statistical closure models has been developed for improving the skill of the reduced-order models. The new reduced-order models take into account higher-order moments but nevertheless remain computationally efficient [1,45,50]. With the model calibration based on effective information criteria, these new reduced-order models succeed in capturing the non-Gaussian statistical characteristics including intermittency and extreme events as well as memory effect and temporal correlation. The new reduced-order models have also been used to provide accurate state estimation and prediction of many high-dimensional complex nonlinear turbulent systems [80–82].

This research expository article blends new viewpoints and results with a current research summary of a specific perspective. It focuses on both the development and applications of the information-theoretic framework as well as the new reduced-order nonlinear modeling strategies for dealing with model error, information barriers, state estimation and prediction in complex multiscale systems. The contents include the general mathematical framework and theory, effective numerical procedures, instructive qualitative models, and concrete models from climate, atmosphere and ocean science. The remaining of the article is organized as follows. The information-theoretic framework is developed in Section 2. In the same section, various information barriers in the presence of model error are shown via simple but instructive stochastic models. In Section 3, the information-theoretic framework is applied to assess model error in state estimation and prediction with examples coming from both complex scalar models and spatially-extended multiscale turbulent systems. The advantage of the information-theoretic framework over the traditional path-wise measurements are illustrated. Section 4 deals with sensitivity and linear statistical response using the fluctuation–dissipation theorem. An efficient and effective algorithm in finding the most sensitive change directions using information theory is also included in this section. Then, in Section 5, a novel framework of data-driven physics-constrained nonlinear stochastic models and predictions is developed and is applied to predicting the time series of an important atmospheric phenomenon with strong intermittent instabilities and extreme events. Section 6 includes the development of the new effective reduced-order models that involve higher order statistical features. These new models together with the information-optimization model calibration strategy are applied to predicting passive tracer extreme events driven by spatially-extended complex dynamical systems. The article is concluded in Section 7.

2. Information Theory and Information Barriers with Model Error and Some Instructive Stochastic Models

2.1. An Information-Theoretic Framework of Quantifying Model Error and Model Sensitivity

An information-theoretic framework has recently been developed and applied to quantify model error, model sensitivity and prediction skill [10,26–34]. The natural way to measure the lack of information in one probability density $q(\mathbf{u})$ compared with the true probability density $p(\mathbf{u})$ is through the relative entropy $\mathcal{P}(p,q)$ [26,32,40],

$$\mathcal{P}(p,q) = \int p \log\left(\frac{p}{q}\right),\tag{1}$$

which is also known as Kullback–Leibler divergence or information divergence [83–85]. Despite the lack of symmetry, the relative entropy has two attractive features. First, $\mathcal{P}(p,q) \geq 0$ with equality if and only if $p = q$. Second, $\mathcal{P}(p,q)$ is invariant under general nonlinear changes of variables. These provide an attract framework for assessing model errors in many applications [23,26,33,34,86–89].

To quantify the model error and information barriers, let us denote π the probability density of the perfect model, which is actually unknown in practice. Nevertheless, the least biased estimate π_L based on L measurements of the perfect model during the training phase is typically available. Therefore, $\mathcal{P}(\pi, \pi_L)$ precisely quantifies the intrinsic error that is due to the ignorance of the information beyond the L measurements in the perfect model. On the other hand, denote π^M the probability density associated with an imperfect model. Then, the model error in the imperfect model compared with the truth is given by the difference between π^M and π, which is quantified by $\mathcal{P}(\pi, \pi^M)$. Note that $\mathcal{P}(\pi, \pi^M)$ quantifies the overall model error. Signal-dispersion (e.g., in Equation (6)) and other decomposition methods are often used to access different components of the model error. In addition, a general description of the model error in complex turbulent systems includes both the statistical information in terms of the PDFs and the dynamical information such as the temporal autocorrelation function. The latter will be emphasized in Sections 2.3–2.5.

In practice, π^M is determined by no more information than the available in the prefect model. In addition, the imperfect model is typically defined on a subspace of the coarse-grained, resolved variables of the perfect model. Therefore, throughout the following analysis, we focus on characterizing the statistical departures of the imperfect model dynamics relative to the perfect model on these coarse-grained variables **u**.

Now, consider a class of imperfect models $\mathcal{M}$. The best imperfect model for the coarse-grained variables **u** is the $M_* \in \mathcal{M}$ so that the perfect model has the smallest additional information beyond the imperfect model distribution π^{M_*}, namely

$$\mathcal{P}(\pi, \pi^{M_*}) = \min_{M \in \mathcal{M}} \mathcal{P}(\pi, \pi^M). \tag{2}$$

The following general principle [26] facilitates the practical calculation of (2),

$$\begin{aligned}
\mathcal{P}(\pi, \pi_{L'}^M) &= \mathcal{P}(\pi, \pi_L) + \mathcal{P}(\pi_L, \pi_{L'}^M) \\
&= (\mathcal{S}(\pi_L) - \mathcal{S}(\pi)) + \mathcal{P}(\pi_L, \pi_{L'}^M).
\end{aligned} \tag{3}$$

In (3), we have assumed in practice that only L measurements are available in the perfect system and the imperfect model takes into accounts L' measurements with $L' \le L$. In (3),

$$\mathcal{S}(\pi_L) - \mathcal{S}(\pi) = -\int \pi_L \log \pi_L + \int \pi \log \pi \tag{4}$$

is the entropy difference, which precisely measures an intrinsic error from the L measurements of the prefect system. Consequently, the optimization in (2) can be computed by replacing the unknown π by the hypothetically known π_L so that the optimal model within the given class satisfies

$$\mathcal{P}(\pi_L, \pi_{L'}^{M_*}) = \min_{M \in \mathcal{M}} \mathcal{P}(\pi_L, \pi_{L'}^M). \tag{5}$$

The most practical setup for utilizing the framework of empirical information theory in many applications arises when both the measurements of the perfect system and its imperfect model involve only the mean and covariance of the resolved variables **u** so that $\pi_L = \pi_G \sim \mathcal{N}(\bar{\mathbf{u}}, R)$ and $\pi^M := \pi_G^M \sim \mathcal{N}(\bar{\mathbf{u}}_M, R_M)$ are both Gaussian. In this case, $\mathcal{P}(\pi_G, \pi_G^M)$ has the explicit formula [2,26]

$$\mathcal{P}(\pi_G, \pi_G^M) = \left[\frac{1}{2}(\bar{\mathbf{u}} - \bar{\mathbf{u}}_M)^*(R_M)^{-1}(\bar{\mathbf{u}} - \bar{\mathbf{u}}_M) \right] + \left[-\frac{1}{2}\log\det(RR_M^{-1}) + \frac{1}{2}(\mathrm{tr}(RR_M^{-1}) - N) \right]. \tag{6}$$

The first term in brackets in (6) is called 'signal', reflecting the model error in the mean but weighted by the inverse of the model variance, R_M, whereas the second term in brackets, called 'dispersion', involves only the model error covariance ratio, RR_M^{-1}. The signal and dispersion terms in (9) are individually invariant under any (linear) change of variables which maps Gaussian distributions to Gaussians. This property is very important for unbiased model calibration.

Next, we introduce the framework for improving model fidelity and model sensitivity [26,27]. Assume both the perfect and imperfect models are perturbed and both distributions vary smoothly with parameter δ, namely,

$$\begin{aligned}
\pi_{L,\delta}(\mathbf{u}) &= \pi_L(\mathbf{u}) + \delta\pi_L(\mathbf{u}), &\int \delta\pi_L(\mathbf{u})d\mathbf{u} = 0, \\
\pi_\delta^M(\mathbf{u}) &= \pi^M(\mathbf{u}) + \delta\pi^M(\mathbf{u}), &\int \delta\pi^M(\mathbf{u})d\mathbf{u} = 0.
\end{aligned}$$

Rigorous theorems guarantee this smooth dependence under minimal hypothesis for stochastic systems [90]. By assuming the parameter δ is small enough and doing leading order Taylor expansion of (3), we reach the following result:

$$
\begin{aligned}
\mathcal{P}(\pi_\delta, \pi_\delta^M) =& \mathcal{S}(\pi_{L,\delta}) - \mathcal{S}(\pi_\delta) + \mathcal{P}(\pi_L, \pi^M) + \int \log\left(\frac{\pi_L}{\pi^M}\right) \delta\pi_L - \frac{\pi_L}{\pi^M}\delta\pi^M \\
&+ \frac{1}{2}\int \left[\pi_L^{-1}(\delta\pi_L)^2 + \frac{\pi_L}{(\pi^M)^2}(\delta\pi^M)^2 - 2\frac{\delta\pi_L\delta\pi^M}{\pi^M}\right] + O(\delta^3).
\end{aligned}
\tag{7}
$$

In the case with perfect model fidelity in terms of the L measurements, namely $\mathcal{P}(\pi_L, \pi^M) = 0$ or $\pi_L(\mathbf{u}) = \pi^M(\mathbf{u})$, the expansion in (7) becomes

$$
\mathcal{P}(\pi_\delta, \pi_\delta^M) = \mathcal{S}(\pi_{L,\delta}) - \mathcal{S}(\pi_\delta) + \frac{1}{2}\int \pi_L^{-1}(\delta\pi_L - \delta\pi^M)^2 + O(\delta^3),
\tag{8}
$$

where the quadratic discrepancy is measured in the Fisher information metric [91–93].

One important scenario in practice involves measuring only the mean and covariance for an imperfect model. We denote $\pi_{2,\delta} = \pi_{G,\delta}$ the unbiased Gaussian estimate of the perfect model. For the simplicity of statement, we further assume both the covariance of the perfect and imperfect models, R_δ and $R_{M,\delta}$, are diagonal such that $R_\delta = (R_k) + (\delta R_k)$ and $R_{M,\delta} = (R_{M,k}) + (\delta R_{M,k})$, where $|k| \leq N$ and (δR_k) and $(\delta R_{M,k})$ are the covariance response to the external perturbation which are all scalar variances. In such a Gaussian setup, Equation (7) becomes

$$
\begin{aligned}
\mathcal{P}(\pi_\delta, \pi_\delta^M) =& \mathcal{S}(\pi_{G,\delta}) - \mathcal{S}(\pi_\delta) + \mathcal{P}(\pi_G, \pi^M) \\
&+ \sum_{|k|\leq N}(\delta\mathbf{u}_k - \delta\mathbf{u}_{M,k})^* R_{M,k}^{-1}(\mathbf{u}_k - \mathbf{u}_{M,k}) - \frac{1}{2}(\mathbf{u}_k - \mathbf{u}_{M,k})^* \frac{\delta R_{M,k}}{R_{M,k}^2}(\mathbf{u}_k - \mathbf{u}_{M,k}) \\
&+ \frac{1}{2}\sum_{|k|\leq N}\left[\frac{R_k}{R_{M,k}} - 1\right]\left[\frac{\delta R_k}{R_k} - \frac{\delta R_{M,k}}{R_{M,k}}\right] + O(\delta^2).
\end{aligned}
\tag{9}
$$

Under the same Gaussian assumptions and perfect model fidelity, the formula in (8) becomes

$$
\begin{aligned}
\mathcal{P}(\pi_\delta, \pi_\delta^M) =& \mathcal{S}(\pi_{G,\delta}) - \mathcal{S}(\pi_\delta) + \frac{1}{2}\sum_{|k|\leq N}(\delta\bar{\mathbf{u}}_k - \delta\bar{\mathbf{u}}_{M,k})^* R_k^{-1}(\delta\bar{\mathbf{u}}_k - \delta\bar{\mathbf{u}}_{M,k}) \\
&+ \frac{1}{4}\sum_{|k|\leq N} R_k^{-2}(\delta R_k - \delta R_{M,k})^2 + O(\delta^3),
\end{aligned}
\tag{10}
$$

where the first summation represents the signal contribution whereas the second summation represents the dispersion contribution. The formula in (9) or (10) can be applied to quantify the information barrier in the model sensitivity using imperfect models (see, for example, Sections 2.3 and 4).

The information theory developed here plays important roles in quantifying model error, model sensitivity and information barrier, assessing data assimilation and prediction skill as well as developing new reduced-order nonlinear modeling strategies. These topics will all be addressed with instructive examples in the following sections.

2.2. Information Barriers in Capturing Model Fidelity

Information barriers are defined broadly as the gap of information obtained from the imperfect model related to that from the perfect one that can never be overcome. In other words, the information barriers imply the impossibility of generating stochastic models in a given model family that capture the missing physics. For example, if the minimization of the information model error in (5) remains significant, then there always exists a portion of information in the perfect model that cannot be recovered by the reduced imperfect models. Such information barriers play important roles in both

model fidelity and sensitivity. Quantifying the information barriers have both theoretic and practical importance. It indicates the futility of model calibration if the information barrier is significant. It can also be used as a guidance to expand the model family of reduced models for the improvement of imperfect models.

Below, two simple but illustrative examples will be shown for the information barriers in capturing model fidelity. The study of these information barriers to more sophisticated turbulent dynamical models can be found in [30,50]. The information barrier in the model sensitivity will be discussed in Section 2.3.

2.2.1. First Information Barrier: Using Gaussian Approximation in Non-Gaussian Models

The first piece of information involves using linear Gaussian models to approximate non-Gaussian nature, which is a typical (crude) strategy in many real-world applications [30,94,95]. In addition to the intrinsic barrier in capturing the higher-order statistics (namely the non-Gaussian features), the goal here is to show that there exists an information barrier using the linear Gaussian models in capturing even the second order statistics of the truth in the presence of a time-periodic forcing.

As a simple but illustrative example, consider the following nonlinear dynamical system:

$$
\begin{aligned}
\frac{du}{dt} &= -\gamma u + F(t) + \sigma_u \dot{W}_u, \\
\frac{d\gamma}{dt} &= -d_\gamma(\gamma - \hat{\gamma}) + \sigma_\gamma \dot{W}_\gamma.
\end{aligned}
\tag{11}
$$

This is a simplified version of the model named as "stochastic parameterized extended Kalman filter (SPEKF) model" that is widely used in nonlinear data assimilation and prediction [96–99]. Here, u can be regarded as a resolved variable and γ is an unresolved process which interacts with u in a nonlinear way. The external forcing $F(t)$ is usually a periodic function that mimics the seasonal/annual cycle or any deterministic cycle that contributes to the system. In (11), the unresolved process γ plays the role of stochastic damping and therefore the statistics of u can be highly non-Gaussian with intermittent instabilities. One nice property of the model in (11) is that the time evolution of all the moments can be written down in closed analytical forms [15,97].

A natural way to approximate u without knowing the detailed structure of the unresolved process γ is the following mean stochastic model (MSm) [15,41],

$$
\frac{du^M}{dt} = -\hat{\gamma} u^M + F^M(t) + \sigma_u^M \dot{W}_u^M,
\tag{12}
$$

where the mean value of the hidden process γ is used in the dynamics of the resolved variable u. The MSm in (12) is a linear model with additive noise and therefore it has Gaussian statistics. To understand the information barrier in using the linear and Gaussian MSm to approximate the nonlinear and non-Gaussian SPEKF-type model in (11), we study the following two dynamical regimes:

$$
\begin{array}{lcccc}
\text{Highly Intermittent Regime:} & \sigma_u = 0.5, & d_\gamma = 1.2, & \sigma_\gamma = 1, & \hat{\gamma} = 1.5, \\
\text{Nearly Gaussian Regime:} & \sigma_u = 0.5, & d_\gamma = 1.2, & \sigma_\gamma = 1, & \hat{\gamma} = 5.0.
\end{array}
\tag{13}
$$

In both regimes, the periodic forcing is given by $F(t) = 5\sin(t)$. Figure 1 shows sample trajectories of u and γ in the two regimes, respectively. It is clear that in the highly intermittent regime, γ has a frequent transition to values below zero, which triggers large bursts in the signal of u, namely the intermittent instability. On the other hand, in the nearly Gaussian regime, γ stays positive and therefore the signal of u has no intermittent instability. In panels (a)–(d) of Figure 2, the time evolution of the first four moments of u is shown. Despite the strong nonlinear interactions between γ and u, the external forcing $F(t)$ results in periodic behavior in all these statistics. In panels (e)–(g), the PDFs of u at three different time instants within one period $t = 13.5, 15$ and 16.5 are demonstrated, which are all highly

non-Gaussian with significant skewness and kurtosis. As comparison, the skewness and kurtosis in the nearly Gaussian regime (Figure 3) are tiny and the amplitudes of the periodic oscillation in the variance, skewness and kurtosis become much weaker.
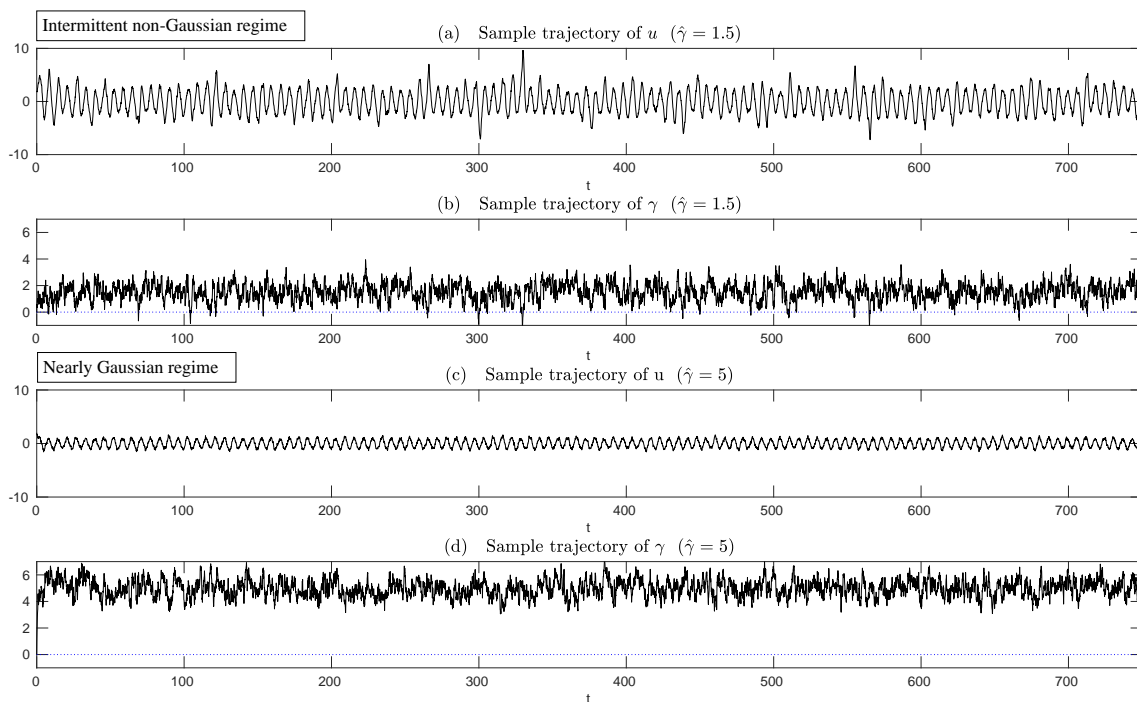
Figure 1. Sample trajectories of u and γ in the highly intermittent regime (**a**,**b**) and nearly Gaussian regime (**c**,**d**), respectively. The parameters are given in (13). In (**b**,**d**), the dotted line $\gamma = 0$ indicates the instability threshold, where γ below zero corresponds to the unstable phases of u.
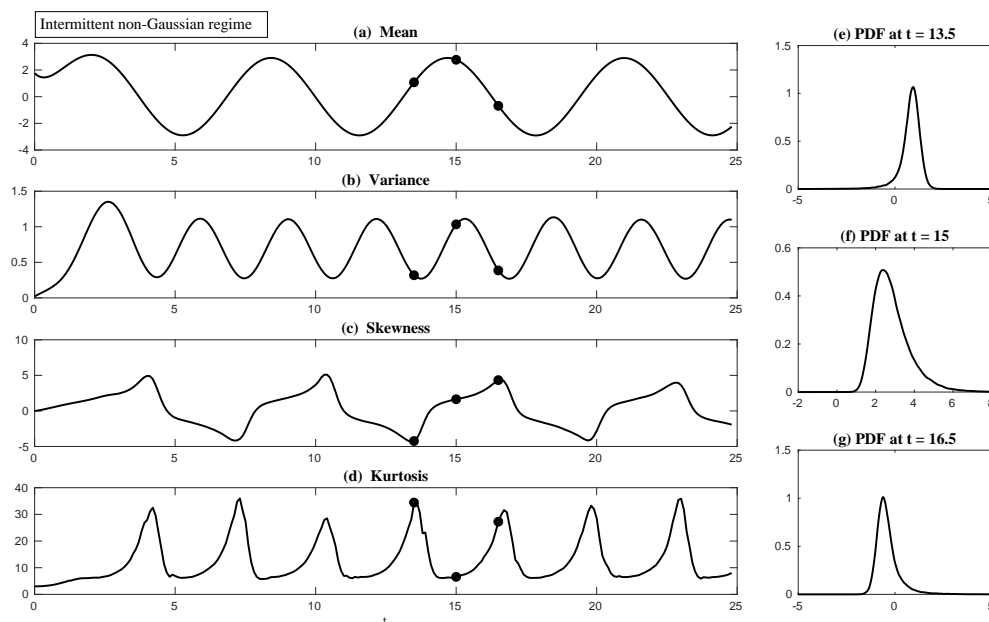
Figure 2. Highly intermittent regime. (**a**–**d**): the first four moments (mean, variance, skewness and kurtosis) of u. (**e**–**g**): Probability density functions (PDFs) of u at $t = 13.5, 15$ and 16.5. These simulations are based on Monte Carlo with 100,000 samples. Note that, due to the intermittent unstable events, the calculation of the high order moments of u becomes sensitive to the samples.
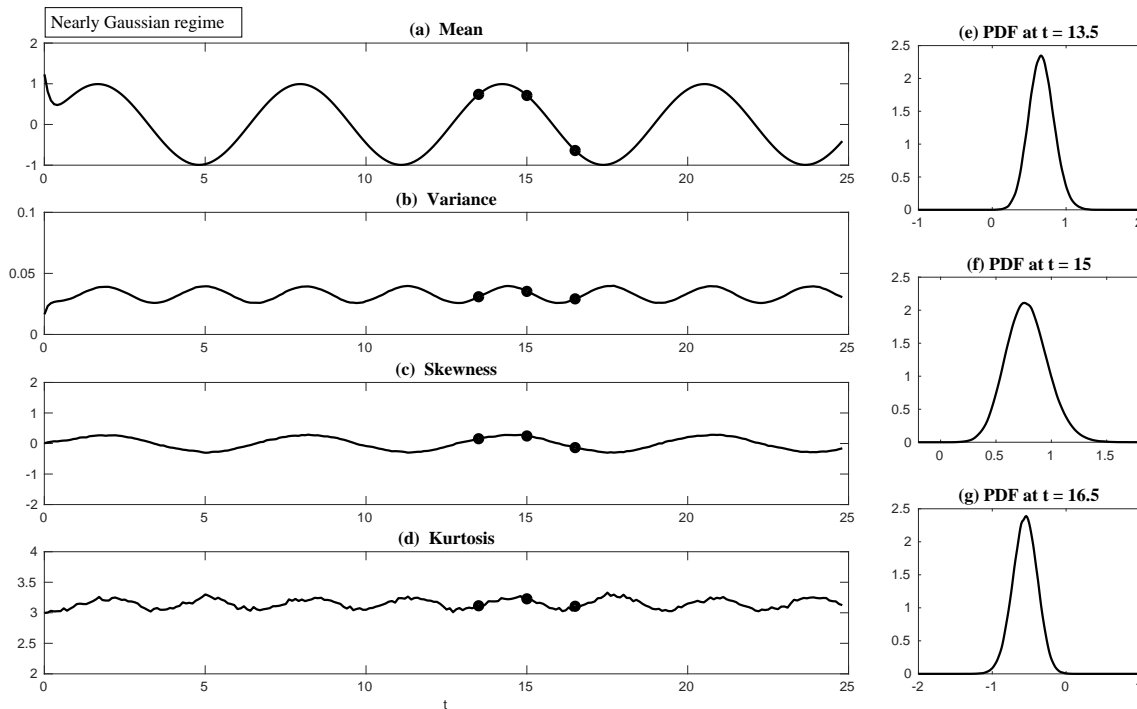
Figure 3. Nearly Gaussian regime. Same captions as in Figure 2.

Let's denote π the PDF of the nonlinear dynamics in (11) and $\pi_G^M = \pi_2^M$ that of the MSm in (12). It is clear according to (3) that there is an intrinsic barrier $\mathcal{S}(\pi) - \mathcal{S}(\pi_G)$ due to the non-Gaussian nature of π. Next, we show that there exists an information barrier in the statistics of MSm π_G^M even compared with the Gaussian approximation of π. Below, the mean and variance of u from the perfect model are both computed using their closed analytical forms [97].

First, we take the same parameters $F^M(t) = F(t)$ and $\sigma_u^M = \sigma_u$ in the MSm (12) as those in the perfect model (11). The time evolutions of the mean $\langle u(t) \rangle$ and the variance $\mathrm{Var}(u(t))$ within one period at the statistical equilibrium from the MSm (blue) are shown in panels (a) and (b) of Figure 4. Although the evolution of the mean using MSm is quite close to the truth (black), the variance is strongly underestimated. This is as expected since a large portion of the variance comes from the intermittent events while the MSm stabilizes the system and does not allow such large bursts. As a consequence, the dispersion part of the model error, as defined in (6), becomes huge. Interestingly, despite the small error in the time evolution of the mean (panel (a)), the signal part of the model error remains significant. In fact, according to (6), the signal part of the model error is weighted by the inverse of the model variance, the severe underestimation of which results in such a large error. To overcome the issue of underestimating the variance, a common strategy in improving imperfect model is to inflate the stochastic forcing coefficient σ_u^M [29,100,101]. Here, the optimization is based on the minimization of the averaged information content $\mathcal{P}(\pi, \pi^M)$ between the perfect and imperfect models within one period at the statistical equilibrium. In panel (f), we show $\overline{\mathcal{P}(\pi, \pi^M)}$ as a function of σ_u^M. With an inflation of σ_u^M, the model error does decrease significantly. However, even at the minimum of the curve where $\sigma_u^{M*} = 1.5$, the model error is still far from zero. In fact, due to the linear nature of the MSm, the forcing $F^M(t)$ in the MSm only affects the evolution of the mean. The periodic behavior in the variance can never be captured by the MSm (see panel (h)), which leads to an information barrier. As comparison, the nonlinearity in the nearly Gaussian regime as shown in Figure 5 is much weaker and therefore the information barrier becomes insignificant. In Figure 6, the optimal stochastic forcing σ_u^{M*} as well as the information barrier as a function of $\hat{\gamma}$ are shown. It is clear that the information barrier decreases when the dynamical regime goes more towards Gaussian (with an increase of $\hat{\gamma}$).
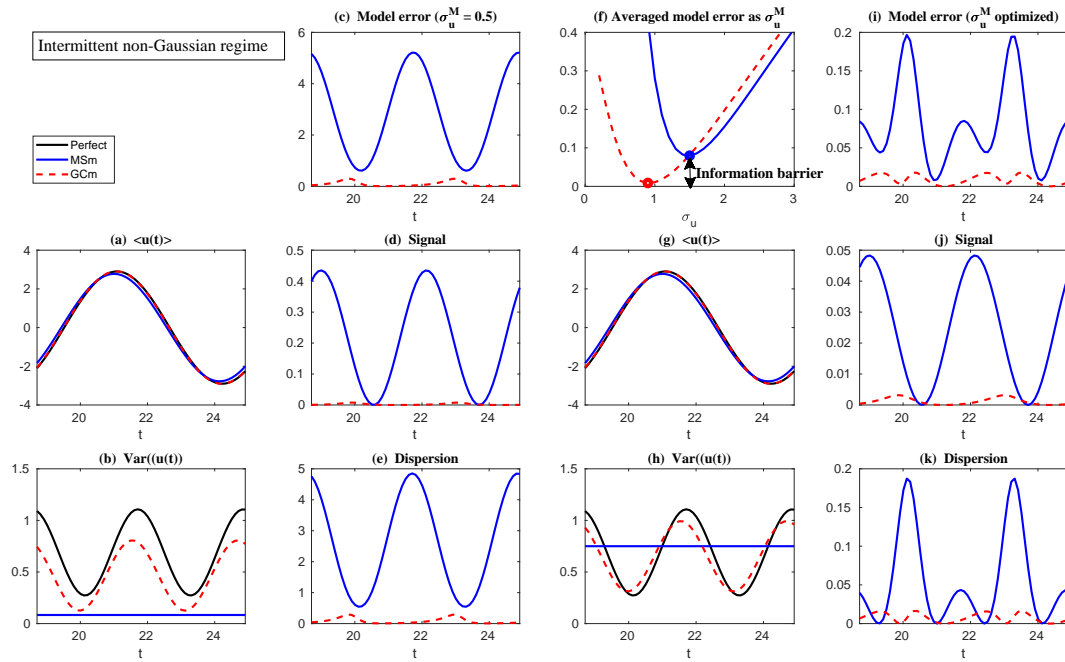
Figure 4. Highly intermittent regime. (**a**,**b**): time evolutions of the mean $\langle u(t) \rangle$ and variance $\mathrm{Var}(u(t))$ within one period in the statistical equilibrium; (**c**–**e**): total model error, model error in the signal part and model error in the dispersion part. The results shown from both mean stochastic model (MSm) and Gaussian closure model (GCm) in (**a**–**e**) are equipped with the same parameters as those in the perfect model; (**f**): averaged model error $\overline{\mathcal{P}(\pi, \pi^M)}$ as a function of σ_u^M; (**g**–**k**) are similar to those in (**a**–**e**) except that the stochastic forcing coefficient σ_u^M is optimized.
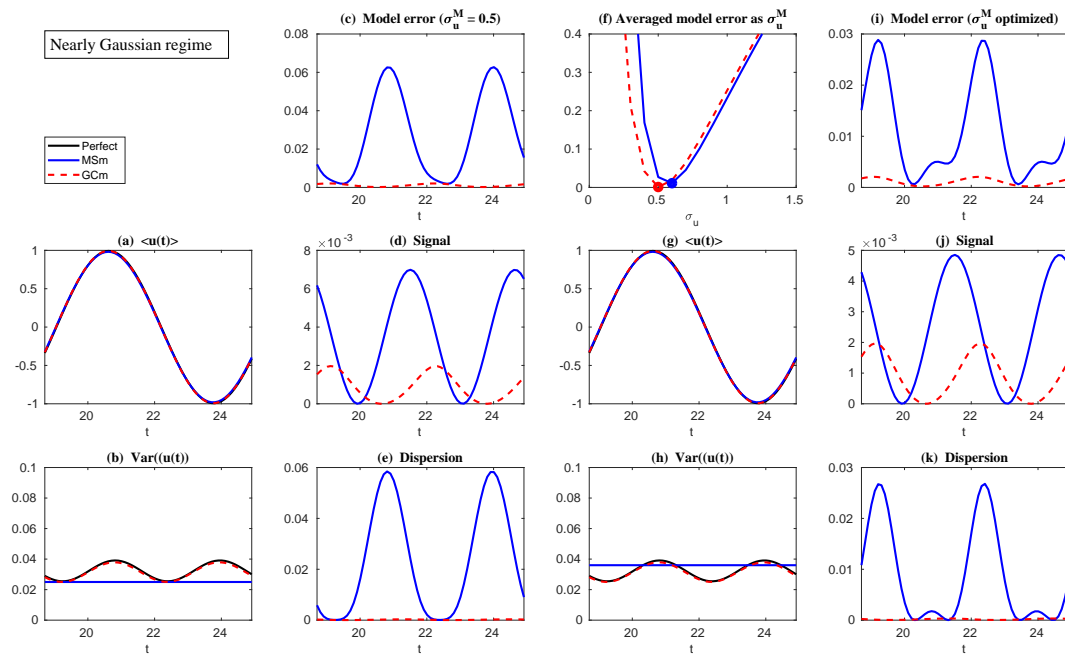
Figure 5. Nearly Gaussian regime. Same captions as those in Figure 4.
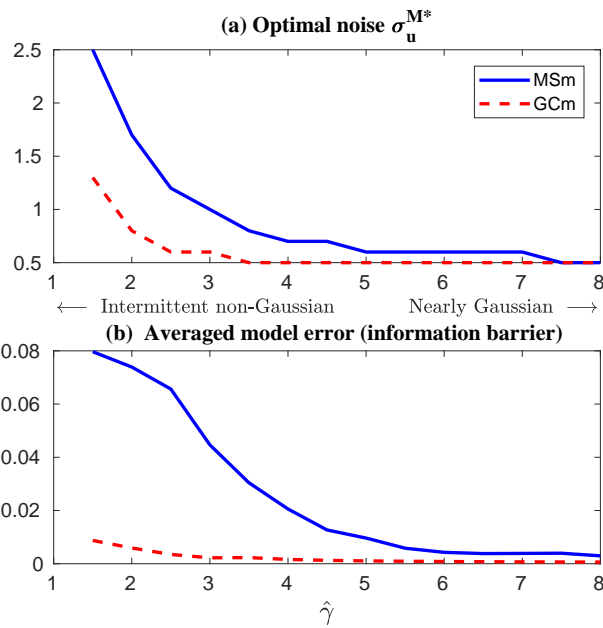
Figure 6. (**a**): optimal noise coefficient σ_u^M in MSm and GCm as a function of $\hat{\gamma}$. The larger the $\hat{\gamma}$ is, the corresponding dynamical regime is more Gaussian; (**b**): the corresponding minimal information model error (information barrier) averaged over a period in the statistical equilibrium.

The above analysis indicates an important fact. That is, even if only the first two moments (mean and variance) are taken into account in the perfect model, the linear MSm still fails to capture the evolution of these Gaussian statistics, which evolve in a strongly nonlinear way driven by the underlying nonlinear perfect dynamics. Such an information barrier cannot be overcome unless the imperfect model contains nonlinear information. In practice, various closure models are used to approximate the nonlinear behavior in the underlying perfect model. Below, we briefly report the results using a simple Gaussian closure model (GCm) [73,102,103]. Recall the Reynolds' decomposition

$$u = \bar{u} + u', \qquad \gamma = \bar{\gamma} + \gamma',$$

where $\bar{\cdot}$ is the ensemble mean and $\cdot'$ is the fluctuation with $\overline{\cdot'} = 0$. With the Reynolds' decomposition, the evolution equations of the mean $\langle u \rangle = \bar{u}, \langle \gamma \rangle = \bar{\gamma}$, the variance $\mathrm{Var}(u) = \overline{u'^2}, \mathrm{Var}(\gamma) = \overline{\gamma'^2}$ and the covariance $\mathrm{Cov}(u', \gamma') = \overline{u'\gamma'}$ are given by

$$
\begin{aligned}
d\bar{u} &= (-\bar{\gamma}\bar{u} - \overline{u'\gamma'} + F^M(t))dt, \\
d\bar{\gamma} &= -d_\gamma^M(\bar{\gamma} - \hat{\gamma}^M)dt, \\
d\overline{u'^2} &= (-2\bar{u}\,\overline{u'\gamma'} - 2\overline{u'^2}\bar{\gamma} + (\sigma_u^M)^2 - 2\overline{u'^2\gamma'})dt, \\
d\overline{\gamma'^2} &= (-2d_\gamma^M\overline{\gamma'^2} + (\sigma_\gamma^M)^2)dt, \\
d\overline{u'\gamma'} &= [-(\bar{\gamma} + d_\gamma^M)\overline{u'\gamma'} - \bar{u}\,\overline{u'^2} - \overline{u'\gamma'^2}]dt.
\end{aligned}
\tag{14}
$$

Note that the third and the fifth equations of (14) involve triad interactions $\overline{u'\gamma'^2}$ and $-2\overline{u'^2\gamma'}$, which represent the third order moments. These triad terms come from the nonlinearity of the underlying perfect system. In fact, the perfect model involves quadratic nonlinearity, and therefore the evolution of the k-th order moments always depend on the $k+1$-th order ones. To close the system, the Gaussian closure model assumes $\overline{u'\gamma'^2} = -2\overline{u'^2\gamma'} = 0$. The resulting system then involves only the interactions between the mean and covariance,

$$
\begin{aligned}
d\bar{u} &= (-\bar{\gamma}\bar{u} - \overline{u'\gamma'} + F^M(t))dt, \\
d\bar{\gamma} &= -d_\gamma^M(\bar{\gamma} - \hat{\gamma}^M)dt, \\
d\overline{u'^2} &= (-2\bar{u}\,\overline{u'\gamma'} - 2\overline{u'^2}\bar{\gamma} + (\sigma_u^M)^2)dt, \\
d\overline{\gamma'^2} &= (-2d_\gamma^M\overline{\gamma'^2} + (\sigma_\gamma^M)^2)dt, \\
d\overline{u'\gamma'} &= [-(\bar{\gamma} + d_\gamma^M)\overline{u'\gamma'} - \bar{u}\,\overline{u'^2}]dt.
\end{aligned}
\tag{15}
$$

In Figure 4, it is clear from panel (b) that even using exactly the same parameters in GCm as in the perfect one and without optimizing σ_u^M, the variance recovered from the GCm is much improved compared with that from the MSm. This indicates the fact that the information barrier can be largely overcome by taking into account the nonlinearity in the imperfect model. The remaining error comes from ignoring the third order moments $\overline{u'\gamma'^2}$ and $-2\overline{u'^2\gamma'}$, which are nonzero in the intermittent non-Gaussian regime. More elaborate closure model techniques involve calibrating the third order moments using various approximations [11,44,104], which are not necessary in this simple example but have been shown to be crucial for more complex turbulent dynamical systems. Such topics will be discussed in detail in Section 6.3. Notably, the periodic behavior in the variance using the GCm is captured due to the nonlinear interactions between mean and variance. For example, the time-periodic mean $\bar{u}$ appears in the equation of $\overline{u'^2}$. This is a significant difference compared with the linear MSm. Finally, with the optimal choice of σ_u^M (panel (f)), the information model error here becomes negligible.

2.2.2. Second Information Barrier: Using Single Point Correlation to Approximate Full Correlation Matrix

The strategy with single point statistics is widely used in climate science [105]. The single point statistics takes into account only the variance at each grid point and ignores the correlations between different grids. Despite the fact that both equilibrium consistency and sensitivity in response in single point statistics can be achieved by turning at most one parameter of the imperfect model, such a strategy is not enough for desirable model performance, which can be measured by the information barrier [10,26,27]. Such an information barrier was first quantified in [44,50]. In the following, we show this information barrier.

Let the PDF from the true model be $\pi(\mathbf{u})$ with $\mathbf{u} = (u_0, u_1, \ldots u_{J-1})^T$ as before. Consider a Gaussian imperfect model where we only measure pointwise marginal PDF $\pi_{1pt}^M(\mathbf{u}_j) \equiv \pi_{1pt,j}^M$ at each grid pint $j = 0, \ldots, J - 1$. Then, we construct the PDF with only single point statistics from the marginal distribution as $\pi_{1pt}^M = \prod_{j=0}^{J-1}\pi_{1pt,j}^M$. According to [34], the information distance between the truth and imperfect model prediction has the form:

$$
\mathcal{P}(\pi, \pi_{1pt}^M) = \mathcal{S}(\pi_G) - \mathcal{S}(\pi) + \mathcal{P}\left(\pi_G, \prod_{j=0}^{J-1}\pi_{1pt,j}^G\right) + \sum_{j=0}^{J-1}\mathcal{P}(\pi_{1pt,j}^G, \pi_{1pt,j}^M),
\tag{16}
$$

with $\pi_{1pt,j}^G = \mathcal{N}(\bar{\mathbf{u}}_j, R_j)$. The first part on the right-hand side of (16) is the intrinsic information barrier in Gaussian approximation. The third part is the model error in the imperfect model as compared to the single point statistics of the perfect model Gaussian fit, which can be vanished (or at least be minimized) by calibrating the imperfect model. The error from single point approximation by ignoring the cross-covariance then comes only from the information barrier in the marginal approximation as shown in the second part on the right-hand side of (16).

Below, we assume the true system is statistically homogeneous, which means the statistics is invariant at different grid points. This is in fact a typical feature in many real applications [2,106,107]. With statistical homogeneity, it is straightforward to show [50] that the diagonal entries of the covariance matrix corresponding to the single point approximation are all the same R_{1pt}. In addition, the covariance matrix $\hat{R}$ in the spectral space (associated with the Fourier modes of $\mathbf{u}$) is also a diagonal matrix. Denote $\hat{R}_j$ the j-th diagonal entry of $\hat{R}$:

$$\hat{R}_j = J \sum_{n=0}^{J-1} \langle u'_0 u'_n \rangle e^{2\pi i n j / J}, \tag{17}$$

where u'_n is the n-th component of **u** by subtracting the mean. Therefore, $\hat{R}_{1pt} = \frac{\sum_{j=0}^{J-1} \hat{R}_j}{J}$. By further assuming the same pointwise mean for π_G and its single point approximation, the information barrier due to single point statistics approximation becomes

$$
\begin{aligned}
\mathcal{P}\left(\pi_G, \prod_{j=0}^{J-1} \pi_{1pt,j}^G\right) &= \sum_{k=J/2+1}^{J/2} \left[-\log \det(\hat{R}_k \hat{R}_{1pt}^{-1}) + \mathrm{tr}(\hat{R}_k \hat{R}_{1pt}^{-1} - I) \right] \\
&= -\sum_{k=J/2+1}^{J/2} \log \det(\hat{R}_k \hat{R}_{1pt}^{-1}) + \mathrm{tr}\left[\sum_{k=J/2+1}^{J/2} (\hat{R}_k \hat{R}_{1pt}^{-1} - I) \right] \\
&= -\log \left(\prod_{k=J/2+1}^{J/2} \frac{\det \hat{R}_k}{\det \hat{R}_{1pt}} \right) \\
&= J \log \left[\frac{\det \left(\sum_{j=0}^{J-1} \hat{R}_j / J \right)}{\left(\prod_{j=0}^{J-1} \det \hat{R}_j \right)^{1/J}} \right],
\end{aligned} \tag{18}
$$

where the second equality just applies the definition of R_{1pt} such that $\sum_{k=-J/2+1}^{J/2} \left(R_k R_{1pt}^{-1} - I \right) = 0$.

In addition to compute the information barrier explicitly using (18), the following result provides an effective estimation of such an information barrier,

$$\mathcal{P}\left(\pi_G, \prod_{j=0}^{J-1} \pi_{1pt,j}^G\right) \sim O\left((\sigma_{max} - \sigma_{min})^2 \right), \tag{19}$$

where σ_{max} and σ_{min} are the largest and smallest variances in $\hat{R}$. See [50] for more details.

Below, we construct a simple linear system to illustrate the information barrier due to the single point statistics approximation and showing the calculation of the formula in (18). An illustration of such information barrier based on a more sophisticated (40-dimensional) turbulent system is included in [50]. Here, the truth is given by a two-dimensional linear model,

$$
\begin{aligned}
\frac{du_0}{dt} &= L_{00} u_0 + L_{01} u_1 + \sigma_{u_0} \dot{W}_{u_0}, \\
\frac{du_1}{dt} &= L_{10} u_0 + L_{11} u_1 + \sigma_{u_1} \dot{W}_{u_1},
\end{aligned} \tag{20}
$$

with the following parameters

$$\mathbf{L} = \begin{pmatrix} L_{00} & L_{01} \\ L_{10} & L_{11} \end{pmatrix} = \begin{pmatrix} -1 & 0.5 \\ 0.5 & -1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_{u_0} & \\ & \sigma_{u_1} \end{pmatrix} = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}. \tag{21}$$

It is easy to check that the two eigenvalues of **L** are $\lambda_1 = -1.5$ and $\lambda_2 = -0.5$ and therefore the linear system (20) is stable. In addition, due to the non-zero coefficients L_{01} and L_{10}, u_0 and u_1 are correlated. Figure 7 shows the sample trajectories of the two-dimensional model (20) with parameters (21). The covariance matrix and statistical equilibrium can be written down explicitly [15],

$$R = \frac{1}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{22}$$

The non-zero off-diagonal entries clearly indicate the cross-covariance between u_0 and u_1. Now, we implement the single point statistics approximation, which ignores the off-diagonal entries in (22) and the result is

$$R_{1pt} = \frac{1}{3} \begin{pmatrix} 2 & \\ & 2 \end{pmatrix}.$$ (23)

Since this example is extremely simple, it is straightforward to compute the information barrier by plugging the covariance matrices (22) and (23) as well as the zero mean into the explicit formula of the relative entropy (6), which yields

$$\mathcal{P}\left(\pi_G, \prod_{j=0}^{J-1} \pi_{1pt,j}^G\right) = 0.1438.$$ (24)

Alternatively, according to (17), it is also easy to show that $\hat{R}$ is a diagonal matrix and is given by

$$\hat{R} = \begin{pmatrix} \hat{R}_0 & \\ & \hat{R}_1 \end{pmatrix},$$ (25)

where

$$\begin{aligned} \hat{R}_0 &= 2\left(\langle u_0 u_0 \rangle + \langle u_0 u_1 \rangle\right) = 2, \\ \hat{R}_1 &= 2\left(\langle u_0 u_0 \rangle - \langle u_0 u_1 \rangle\right) = \frac{2}{3}. \end{aligned}$$ (26)

Plugging (26) into (18) gives the same result as in (24). This clearly shows the information barrier due to single point statistics approximation. Panels (c) and (d) in Figure 7 show true joint probability density function (PDF) $\pi(u_0, u_1)$ and the one with single point statistics approximation, the difference between which indicates the information barrier.
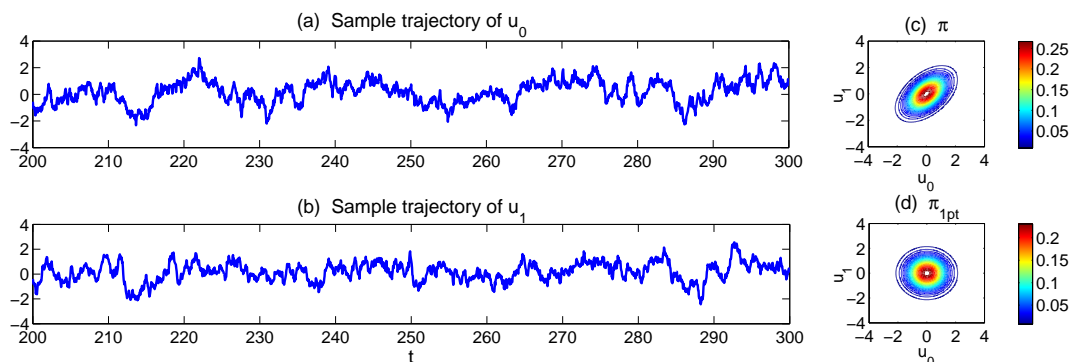
Figure 7. (**a**,**b**): sample trajectories of the two-dimensional model (20) with parameters (21); (**c**): true joint PDF associated with (**a**,**b**); (**d**): joint PDF with single point statistics approximation.

2.3. Intrinsic Information Barrier in Predicting Mean Response to the Change of Forcing

In Section 2.2, we have demonstrated the information barrier in the model fidelity via simple but illustrative examples. In this section, we aim at using a simple example to illustrate the intrinsic information barrier in the model sensitivity. Ref. [27] is a good reference for this topic. The example shown here also reveals the following fact. Even if the model fidelity represented by the equilibrium PDF is captured, the dynamical feature in the perfect model can still be missed if the model sensitivity is not recovered by the imperfect model. Therefore, both the model fidelity and model sensitivity are required in calibrating the imperfect model, which will be discussed with more details in Sections 5 and 6.

Here, the focus is the mean response to the change of forcing in linear models. A general framework of any system response (e.g., variance or higher order moments) to different types of external perturbations

(e.g., forcing, dissipation or phase) in complex nonlinear model will be developed in Section 4 using the so-called fluctuation–dissipation theorem (FDT).

Consider a general linear system with noise

$$\frac{d\mathbf{u}}{dt} = \mathbf{L}\mathbf{u} + \mathbf{F} + \sigma \dot{\mathbf{W}}. \tag{27}$$

In (27), $\mathbf{L}$ is a linear operator whose eigenvalues all have a negative real part, which guarantees the existence of a Gaussian statistical steady state of $\mathbf{u}$. Here $\mathbf{F}$ is an external forcing, which can be a function of time t, and $\dot{\mathbf{W}}$ is stochastic white noise. Now, we impose a forcing perturbation $\delta \mathbf{F}$ to the original system in (27)

$$\frac{d\mathbf{u}^{\delta}}{dt} = \mathbf{L}\mathbf{u}^{\delta} + \mathbf{F} + \delta\mathbf{F} + \sigma\dot{\mathbf{W}}. \tag{28}$$

Since both (27) and (28) are linear models, the mean values $\langle \mathbf{u} \rangle$ and $\langle \mathbf{u}^{\delta} \rangle$ at the statistical steady state can be written explicitly,

$$\langle \mathbf{u} \rangle = \mathbf{L}^{-1}\mathbf{F}, \qquad \text{and} \qquad \langle \mathbf{u}^{\delta} \rangle = \mathbf{L}^{-1}\left(\delta\mathbf{F} + \mathbf{F}\right).$$

Therefore, the mean response of $\mathbf{u}$ to the forcing perturbation $\delta\mathbf{F}$ is given by

$$\langle \delta\mathbf{u} \rangle = \langle \mathbf{u}^{\delta} - \mathbf{u} \rangle = \mathbf{L}^{-1}\delta\mathbf{F}. \tag{29}$$

In practice, model error is usually inevitable. A suitable imperfect model is expected to generate at least the same mean response as in the perfect model in (29) in addition to the model fidelity.

A typical situation with model error for complex systems arises when the true system has additional degrees of freedom that are hidden from the family of imperfect models due to either the lack of scientific understanding or practical lack of computational resolution. The simple example below involves such features.

Consider the following perfect model with linear stochastic equations,

$$\begin{aligned} \frac{du}{dt} &= au + v + F, \\ \frac{dv}{dt} &= qu + Av + \sigma\dot{W}, \end{aligned} \tag{30}$$

where $\dot{W}$ is white noise. The system in (30) has a smooth Gaussian statistical steady state provided that

$$a + A < 0, \qquad aA - q > 0. \tag{31}$$

In (30), u can be treated as a resolved variable while v is a hidden one. All the imperfect models are given by the scalar stochastic equation that involves only the process of the observed variable,

$$\frac{du_M}{dt} = -\gamma_M u_M + F_M + \sigma_M \dot{W}_M. \tag{32}$$

It is natural to require $\gamma_M > 0$ such that the imperfect model (32) has a Gaussian statistical steady state. Next, the imperfect model (32) is tuned to capture the model fidelity in the perfect system (30) by matching the equilibrium mean and variance of u and u_M. This implies

$$\frac{F_M}{\gamma_M} = -\frac{AF}{aA - q}, \qquad \frac{\sigma_M^2}{2\gamma_M} = \frac{\sigma^2}{2(a + A)(aA - q)} \equiv E \tag{33}$$

with a suitable choice of the three tuning parameters F_M, σ_M and $\gamma_M (> 0)$, it is clear that the conditions in (33) can always be satisfied.

In addition to the model fidelity, an important and practical issue is to understand the response of the system to the external forcing perturbation δF. Therefore, it is crucial to have an imperfect model that has the same forcing response as the perfect model. To test the response of the external forcing, we replace F and F_M by $F + \delta F$ and $F_M + \delta F$ in the two linear systems (30) and (32), respectively. Note that the external forcing will not change the variance in linear systems. The only change in the equilibrium response is through the change in mean,

$$\delta u = -\frac{A}{aA - q}\delta F, \qquad \delta u_M = \frac{1}{\gamma_M}\delta F. \tag{34}$$

Now assume $A > 0$ in the perfect model (30). We claim that no model in the family (32) can match the response of u correctly. In fact, with $A > 0$ and $aA - q > 0$ as required in (31), $\delta u \propto -\delta F$. However, $\gamma_M > 0$ implies $\delta u_M \propto \delta F$. In other words, the responses in the perfect and imperfect models are always anti-correlated, which implies an information barrier. To quantify this information barrier, we insert the response of mean (35) into (10) and make use of the fact that the response in the variance is always zero. Then, (10) yields

$$\mathcal{P}(\pi_\delta, \pi_\delta^M) = \frac{1}{2}E^{-1}\left|-\frac{A}{aA - q} - \frac{1}{\gamma_M}\right||\delta F|^2. \tag{35}$$

It is clear that with $A > 0$ there is no finite minimum over γ_M of (35) and necessarily $\gamma_M \to \infty$ in the approach to this minimum value. Thus, there is an intrinsic information barrier to skill in the mean response that cannot be overcome with the imperfect models in (32) even if they satisfy perfect model fidelity (33). On the other hand, if $A < 0$, then (35) indicates a unique minimum with $\gamma_M^* = -A^{-1}(aA - q)$, in which case both the model fidelity and mean response are captured.

2.4. Slow-Fast System and Reduced Model

An important practical issue for complex dynamical systems is how to account for the indirect influence of the unresolved variables $\mathbf{u_{II}}$ on the response of the resolved variables $\mathbf{u_I}$ beyond bare truncation formulae. The importance of this has already been in Sections 2.2 and 2.3 for calibrating the model fidelity and predicting the mean response using linear imperfect models. Understanding this issue also has practical significance since simplified models are always preferred to decrease the computational cost in solving the complex multiscale dynamical systems. Therefore, developing reduced stochastic models for the variables $\mathbf{u_I}$ with high skill for the low-frequency response is a central issue. While nature can be highly nonlinear and non-Gaussian, the focus here is linear systems with slow-fast multiscale features. Below, we will show that the stochastic mode reduction techniques [40,108–110] are able to produce a reduced stochastic model for the low-frequency variables $\mathbf{u_I}$. Despite its simplicity, such a reduced stochastic model has exactly the same mean response operator as that in the complete stochastic system! More Gaussian and non-Gaussian examples can be found in [111].

Consider a linear multiscale stochastic model for variables $\mathbf{u} = (\mathbf{u_I}, \mathbf{u_{II}})^T$ given by

$$
\begin{aligned}
\frac{d\mathbf{u_I}}{dt} &= \mathbf{L}_{11}\mathbf{u_I} + \mathbf{L}_{12}\mathbf{u_{II}} + \mathbf{F_I}, \\
\frac{d\mathbf{u_{II}}}{dt} &= \mathbf{L}_{21}\mathbf{u_I} - \frac{\Gamma}{\epsilon}\mathbf{L}_{22}\mathbf{u_{II}} + \mathbf{F_{II}} + \frac{\sigma}{\epsilon^{1/2}}\dot{\mathbf{W}},
\end{aligned}
\tag{36}
$$

which can also be written in a compact form

$$\frac{d\mathbf{u}}{dt} = \mathbf{L}^\epsilon\mathbf{u} + \sigma^\epsilon\dot{\mathbf{W}} + \mathbf{F}, \qquad \text{with} \qquad \mathbf{L}^\epsilon = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & -\frac{\Gamma}{\epsilon} \end{pmatrix}. \tag{37}$$

The parameter $\epsilon > 0$ in (36) can be large or small. Here, we require that $\mathbf{L}^\epsilon$ has eigenvalues with a negative real part for all ϵ and in particular

$$(\mathbf{L}_{11}\mathbf{u}, \mathbf{u}) < 0, \qquad (\Gamma\mathbf{u}, \mathbf{u}) > 0. \tag{38}$$

for $\mathbf{u} \neq 0$. These requirements guarantee that $\mathbf{L}^\epsilon$ is invertible and the climate mean state is given by $\langle \mathbf{u} \rangle = -(\mathbf{L}^\epsilon)^{-1}\langle \mathbf{F} \rangle$. This together with (37) and (38) implies in particular that the change in the first components of the climate mean state, $\delta\langle \mathbf{u_I} \rangle$, in response to a change in forcing, $\delta\mathbf{F}_1$, is given exactly by

$$\begin{aligned}
\delta\langle \mathbf{u_I} \rangle &= (\mathbf{L}^\epsilon)_{11}^{-1}\delta\mathbf{F_I}, \\
(\mathbf{L}^\epsilon)_{11}^{-1} &= (\mathbf{L}_{11} + \epsilon\mathbf{L}_{12}\Gamma^{-1}\mathbf{L}_{21})^{-1}.
\end{aligned} \tag{39}$$

Stochastic mode reduction techniques [40,108–110] systematically produce a reduced stochastic model for the variables $\mathbf{u_I}$ alone, which is a valid model in the limit $\epsilon \to 0$; such models often have significant skill for moderate variables of ϵ [112,113]. Here, we focus on their skill in producing infinite time-mean response in (39) of the full dynamics from (36) independent of ϵ.

First, the local equations in (36) can be rewritten exactly as an equivalent equation with memory in time for the $\mathbf{u_I}$ variable alone [114] given by

$$\frac{d\mathbf{u_I}}{dt} = \mathbf{L}_{11}\mathbf{u_I} + \mathbf{F_I} + \mathbf{L}_{12}\int_0^t e^{-(\Gamma/\epsilon)(t-s)}\big[\mathbf{L}_{12}\mathbf{u_I}(s) + \mathbf{F_{II}}(s)\big]ds + \mathbf{L}_{12}\epsilon^{-1/2}\int_0^t e^{-(\Gamma/\epsilon)(t-s)}\sigma d\mathbf{W}(s). \tag{40}$$

For simplicity in exposition, zero initial data are assumed for $\mathbf{u_{II}}$. As discussed in detail [40,109], the second and third terms in (40) simplify in the limit $\epsilon \to 0$ and yield reduced simplified local stochastic dynamics for $\mathbf{u_I}$ alone given by

$$\frac{d\tilde{\mathbf{u}}_{\mathbf{I}}}{dt} = \big(\mathbf{L}_{11} + \epsilon\mathbf{L}_{12}\Gamma^{-1}\mathbf{L}_{21}\big)\tilde{\mathbf{u}}_{\mathbf{I}} + \epsilon^{1/2}\mathbf{L}_{12}(-\Gamma)^{-1}\sigma\dot{\mathbf{W}} + \mathbf{F_I} + \epsilon\mathbf{L}_{12}(-\Gamma)^{-1}\mathbf{F_{II}}. \tag{41}$$

This is an explicit example of stochastic mode reduction where the variables $\mathbf{u_{II}}$ have been eliminated and there is a reduced local stochastic equation for $\tilde{u}_{\mathbf{I}}$ alone with explicit corrections that reflect the interaction with the unresolved variables. Here, we address the skill of the approximation in (41) in recovering the exact mean climate response in (39) independent of ϵ. Reasoning as discussed earlier in general below (38), the response of the climate mean in (41) to a change in forcing is given exactly by

$$\delta\langle \tilde{u}_{\mathbf{I}} \rangle = \big(\mathbf{L}_{11} + \epsilon\mathbf{L}_{12}\Gamma^{-1}\mathbf{L}_{21}\big)^{-1}\delta\mathbf{F_I}. \tag{42}$$

Remarkably, the mean response operator in (42) coincides exactly with the projected mean climate response operator in (39) for the complete stochastic system in (36) for any value of $\epsilon > 0$! This general result points to the high skill of the response for the reduced stochastic model in calculating the mean climate response. Note that the asymptotic behavior and the filtering skill of the linear multiscale stochastic model in (36) have both been studied in [115].

We wrap up this subsection with the following remark. Unlike the linear models as shown in Sections 2.3 and 2.4, direct calculations of the response in general nonlinear models become a great challenge. Nevertheless, fluctuation–dissipation theorem (FDT) provides an efficient and practical way for computing the response in nonlinear systems. The general framework of FDT will be developed in Section 4. Note that low-frequency regimes of general circulation models (GCMs) typically exhibit subtle but systematic departures from Gaussianity. In [41], the stochastic mode reduction technique is applied to a simple prototype nonlinear stochastic model that mimics structural features of low-frequency variability of GCMs with non-Gaussian features [116]. FDT is then used to study the skill of the resulting reduced nonlinear stochastic models.

2.5. Fitting Autocorrelation Function of Time Series by a Spectral Information Criteria

As was seen in the previous subsections, the model sensitivity is applied to quantify the information of the temporal evolution of the system. In fact, the autocorrelation function of a given stochastic system is a simple and easily computed measurement that can be used for accessing the model sensitivity. In this subsection, we make use of the autocorrelation to quantify the model sensitivity based on a new information-theoretic framework. Autocorrelation is the correlation of a signal with a delayed copy of itself, as a function of delay. For a zero mean and stationary random process u, the autocorrelation function can be calculated as

$$R(t) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{u(t+\tau)u^*(\tau)}{\text{Var}(u)} d\tau. \tag{43}$$

Clearly, a linear Gaussian process is completely determined by its mean and autocorrelation function, where the autocorrelation function characterizes the memory of the system. Therefore, an accurate estimation of the autocorrelation function in the imperfect models plays an important role in prediction. In many applications, the integral of the autocorrelation function,

$$\tau = \int_0^\infty R(t)dt, \tag{44}$$

which is known as the decorrelation time, is used for model calibration. Although fitting the decorrelation time in the imperfect model is a simpler strategy, it is however insufficient for pointwise agreement with the true autocorrelation. In particular, if the underlying nonlinear turbulent dynamics has a slow mixing rate and involves wave-like behavior, then the profile of the true autocorrelation function is very likely to be a damped oscillation. As a consequence, fitting only the decorrelation time in the imperfect model results in a large model error due to the failure of capturing the detailed oscillation structures of the autocorrelation function, which severely deteriorates the prediction skill. Thus, it is of practical importance to calibrate the autocorrelation function in imperfect models in order to capture the dynamical features beyond the equilibrium statistics of the truth. The autocorrelation function is also directly linked with the model sensitive in terms of the mean response as well as the prediction skill.

Information theory provides a rigorous and practical way to quantify the error in the two autocorrelation functions associated with the perfect and imperfect models respectively [81,117]. However, direct application of the information distance in (1) is not suitable for measuring the difference between the two autocorrelation functions. This is because the autocorrelation function $R(t)$ may oscillate with negative values while π and π^M have to be positive in (1). To generalize the information-theoretic framework to include the autocorrelation functions, the theory of spectral representation of stationary random fields [118] is exploited here. It is proved by Khinchin's formula [118] that if the autocorrelation function $R(t)$ is smooth and rapid-decay, which is the typical property for most systems, then there exists a non-negative function $E(\lambda) \geq 0$ such that

$$R(t) = \int_{-\infty}^\infty e^{i\lambda t} dF(\lambda), \tag{45}$$

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

$$u(t) = \int_{-\infty}^\infty e^{i\lambda t} \hat{Z}(d\lambda). \tag{46}$$

The exact spectral random measure $\hat{Z}(d\lambda)$ has independent increments whose energy spectrum can be represented by $E(\lambda)$ or $dF(\lambda)$

$$dF(\lambda) = E(\lambda)d\lambda = \mathbb{E}\left|\hat{Z}(d\lambda)\right|^2.$$

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

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

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

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

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

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

Then, the normalized relative entropy between these two Gaussian fields becomes

$$\mathcal{P}(p_G, p_G^M) = \sum_j \mathcal{P}\big(p_G(x;\lambda_j), p_G^M(x;\lambda_j)\big)\Delta\lambda,$$

$$\to \int_{-\infty}^{\infty} \mathcal{P}\big(p_G(x;\lambda), p_G^M(x;\lambda)\big)d\lambda, \qquad \text{as } \Delta\lambda \to 0$$

Therefore, given spectral density $E(\lambda)$ and $E^M(\lambda)$, the spectral relative entropy is given by

$$\mathcal{P}(p_G, p_G^M) = \mathcal{P}(E(\lambda), E^M(\lambda)) := \int_{-\infty}^{\infty} \mathcal{P}\big(p_G(x;\lambda), p_G^M(x;\lambda)\big)d\lambda, \tag{47}$$

where we slightly abuse the notion above by using the spectra $E(\lambda)$ to denote density functions. Since E and E^M are variances for the spectral random variables, it is well-defined in the last part of the above formula (47) using the information distance formula (1). Through measuring the information distance in the spectral coefficients $\hat{Z}(\lambda)$, we arrive at the lack of information in the autocorrelation function $R(t)$ from the model. See [81] for more details as well as an efficient algorithm of solving (47).

Finally, let there be the set of parameters θ for the imperfect model. Minimum relative entropy criterion implies the process of minimizing the lack of spectral information distance by picking the optimal parameter set θ^* for the imperfect model so that

$$\mathcal{P}(E(\lambda), E^M(\lambda, \theta^*)) = \min_{\theta} \mathcal{P}(E(\lambda), E^M(\lambda, \theta)). \tag{48}$$

The following example makes use of the above spectral information criteria to reveal the importance in calibrating the autocorrelation function in the imperfect linear prediction model. The perfect model considered here is a noisy version of the so-called Lorenz 84 model [119,120],

$$
\begin{aligned}
\frac{dx}{dt} &= -(y^2 + z^2) - a(x - f) + \sigma_x \dot{W}_x, \\
\frac{dy}{dt} &= -bxz + xy - y + g + \sigma_y \dot{W}_y, \\
\frac{dz}{dt} &= bxy + xz - z + \sigma_z \dot{W}_z.
\end{aligned}
\tag{49}
$$

This model is an extremely simple analogue of the global atmospheric circulation and the noise-free version can be derived as a Galerkin truncation of the two-layer quasigeostrophic potential vorticity equations in a channel [121]. In (49), x represents the intensity of the mid-latitude westerly wind current while y and z represent the cosine and sine phases of a chain of vortices superimposed on the zonal flow.

With $g = 0$, the processes y and z in (49) form a pair of stochastic nonlinear oscillator through the skew-symmetric terms $-bxz$ and bxy, where the frequency of the oscillation is stochastic and it depends on the amplitude of x. Meanwhile, x also plays the role of stochastic damping, which can be seen in the nonlinear terms xy and xz that modify the wave amplitudes.

The parameters used in this test are as follows:

$$
a = 5, \qquad b = 2, \qquad f = 1, \qquad g = 0, \qquad \text{and} \qquad \sigma_x = \sigma_y = \sigma_z = 0.1.
\tag{50}
$$

In Figure 8, sample trajectories and the corresponding autocorrelation functions associated with x, y and z in Equation (49) with parameters (50) are shown. It is clear that the autocorrelation functions associated with y and z oscillate and decay to zero, which satisfies the features of wave pairs.
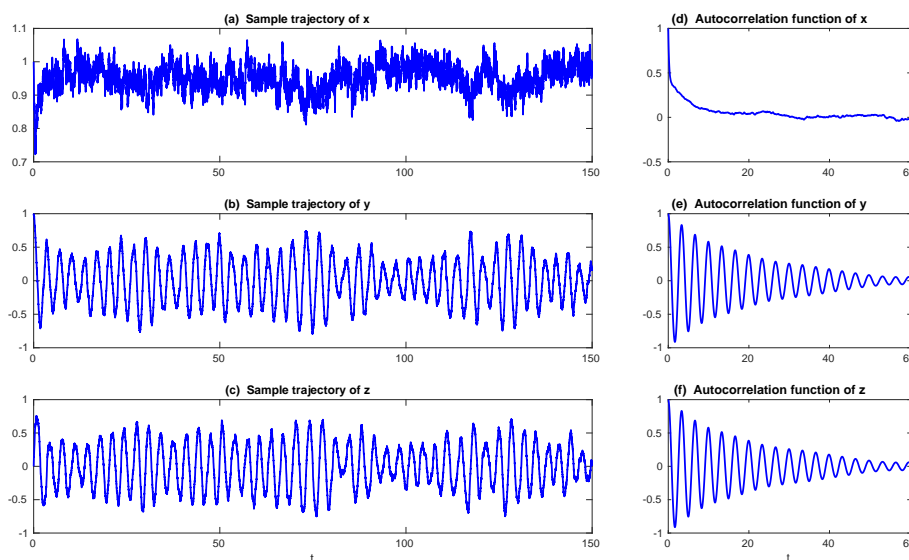

Figure 8. (**a**–**c**): sample trajectories of noisy Lorenz 84 model in (49); (**d**–**f**): the corresponding autocorrelation functions.

The goal here is to predict the vortex variables y and z. The imperfect model for prediction is a mean stochastic model (MSm) with a constant phase,

$$
\frac{du^M}{dt} = (-d_u^M + i\omega_u^M)u^M + \sigma_u^M \dot{W}.
\tag{51}
$$

Note that u^M is a complex process and its real and imaginary parts correspond to the pair of vortex y and z in (49), respectively. Due to the simple linear structure, the autocorrelation function $R^M(t)$ and spectra density $E^M(\lambda)$ of the MSm in (51) can be written down explicitly,

$$R^M(t) = \exp\left((-d_u^M + i\omega_u^M)t\right), \quad \text{and} \quad E^M(\lambda) = \frac{2d_u^M}{(d_u^M)^2 + (\lambda - \omega_u^M)^2}. \tag{52}$$

The model (51) has three parameters to determine: d_u^M, ω_u^M and σ_u^M. Now, we apply the spectral relative entropy in (47) and make use of the analytic formula in (52) for $E^M(\lambda)$ to implement the optimization (48). Note that the spectra density $E^M(\lambda)$ in (52) does not depend on the stochastic forcing coefficient σ_u^M. Therefore, the optimization in (48) is over all the possible choices of d_u^M and ω_u^M. This gives the following results:

$$\text{Optimal parameters by fitting the autocorrelation function:}$$
$$d_u^M = 0.1, \quad \text{and} \quad \omega_u^M = 1.85. \tag{53}$$

For comparison, we also adopt the traditional parameter estimation strategy in MSm by fitting only the decorrelation time (44) [15]:

$$\text{Optimal parameters by fitting only the decorrelation time:}$$
$$d_u^M = 5.0, \quad \text{and} \quad \omega_u^M = 1.85. \tag{54}$$

The remaining parameter σ_u^M is calibrated by matching the variance of y and $\mathrm{Re}(u^M)$ at the statistical steady state, which results in $\sigma_u^M = 0.205$ in Case (53) and $\sigma_u^M = 1.850$ in Case (54). In Figure 9, two sample trajectories of $\mathrm{Re}(u^M)$ and the corresponding autocorrelation functions with parameters in (53) and (54) are shown. It is clear that the sample trajectory of $\mathrm{Re}(u^M)$ and the corresponding autocorrelation function with parameters in (53) highly resemble those of y in Figure 8. On the other hand, the decorrelation time of y is very short, which is due to the canceling out of the oscillation patterns with positive and negative values by integrating the autocorrelation function. Thus, the oscillation patterns in the trajectory of $\mathrm{Re}(u^M)$ are overwhelmed by the noise due to the strong mixing property when the model is calibrated using only the decorrelation time (Panels (c) in Figure 9). This example indicates the necessity of using the information criterion developed here for calibrating the autocorrelation function rather than simply matching the decorrelation time as used in many earlier works.

Finally, in Figure 10, we show the prediction of the time evolution of the mean and variance of y and $\mathrm{Re}(u^M)$ starting from the same initial value $y = \mathrm{Re}(u^M) = 1$ and others $= 0$. In column (a), the mean evolution of $\mathrm{Re}(u^M)$ from the linear model (51) captures that of y in the nonlinear Lorenz 84 model (49) quite accurately with a significant oscillation structure. The trend of the variance using the linear model is also quite similar to that using the Lorenz 84 model, though the oscillation structure in the variance due to the fact that the nonlinearity in the Lorenz 84 model is not predicted by the linear MSm. The latter has already been discussed in [111] and in Section 2.2.1 as an information barrier. On the other hand, by calibrating only the decorrelation time (Column (b) in Figure 10), the time evolutions of the mean and variance have dramatically fast relaxations towards the statistical steady state, which are completely different from the truth.

In [81], the information theory as shown above has been applied to calibrate more complicated reduced order models. The calibrated models succeed in predicting fat-tailed intermittent PDFs in passive scalar turbulence.
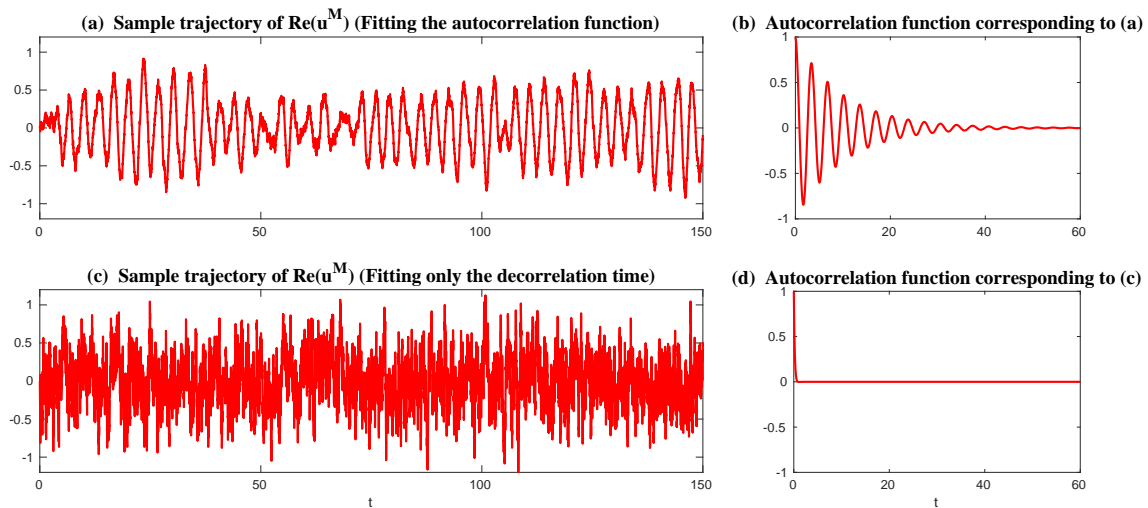
Figure 9. Panels (**a,b**): Sample trajectories and the corresponding autocorrelation functions of $\mathrm{Re}(u^M)$ with parameters in (53). Panels (**c,d**) those with parameters in (54).
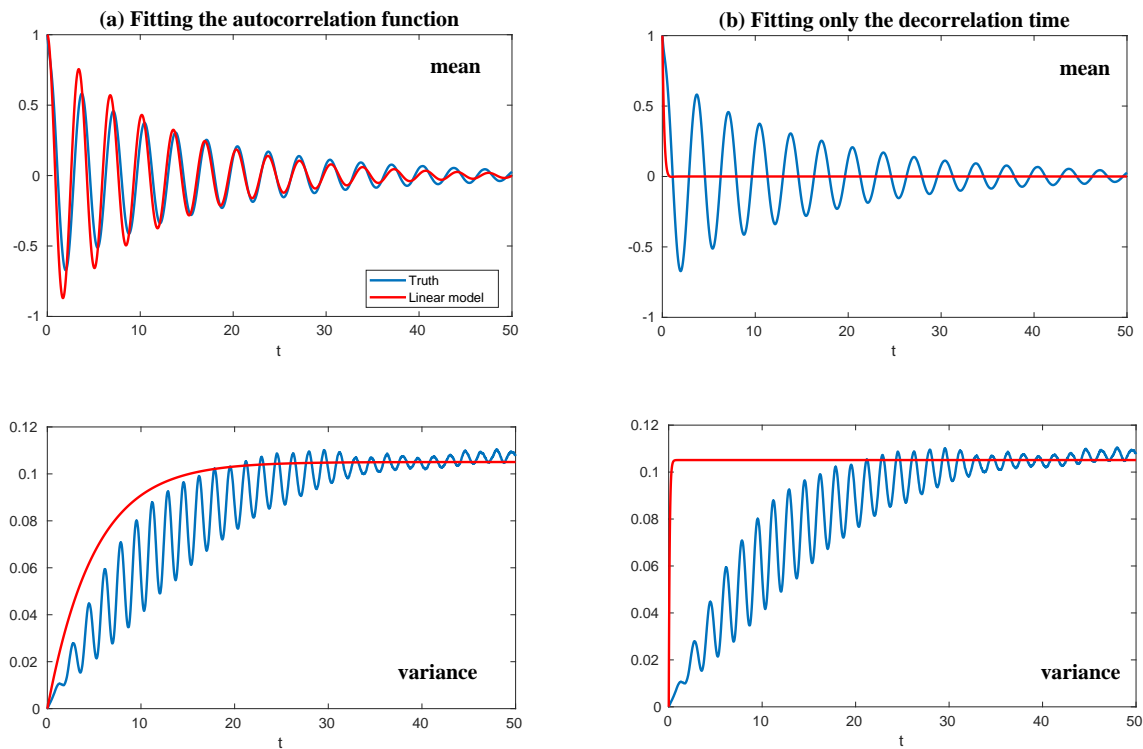
Figure 10. Time evolutions of mean and variance in the perfect model of y and imperfect model of the real part of u^M. (**a**): the parameters in the imperfect model are calibrated by matching the autocorrelation function (53); (**b**): the parameters in the imperfect model are calibrated by matching only the decorrelation time (54).

3. Quantifying Model Error with Information Theory in State Estimation and Prediction

3.1. Kalman Filter, State Estimation and Linear Stochastic Model Prediction

Filtering (also known as data assimilation or state estimation) is the process of obtaining the optimal statistical estimate (based on a Bayesian framework for example) of a natural system from partial observations of the true signal. Important contemporary examples involve the real-time

filtering and prediction of weather and climate as well as the spread of hazardous plumes or pollutants [13–16,122–124].

The general procedure of filtering complex turbulent dynamical systems with partial and noisy observations contains two steps at each time step $t = m\Delta t$. The first step involves a statistical prediction of a probability distribution $u_{m+1|m}$ starting from the initial value $u_{m|m}$ using the given dynamical model. Then, in the second step, $u_{m+1|m}$ is corrected on the basis of the statistical input of noisy observation v_{m+1}, which results in $u_{m+1|m+1}$. See the illustration of Figure 11.
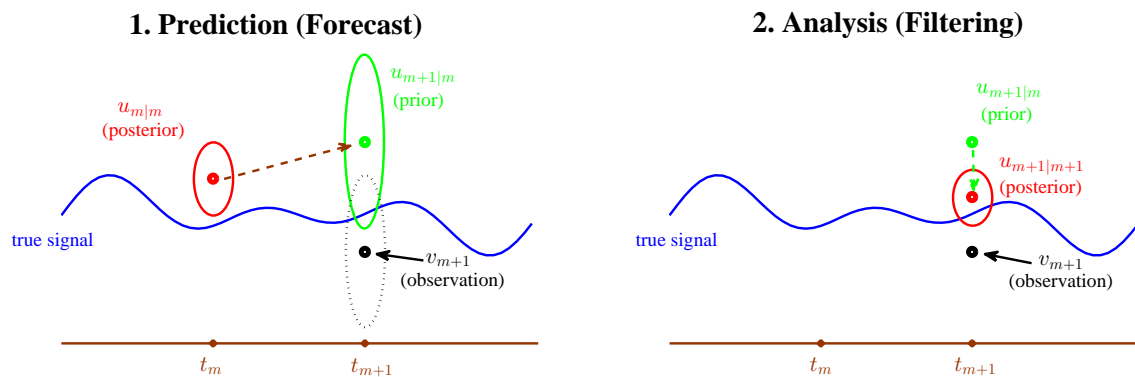
Figure 11. Illustration of the prediction-filtering procedure.

For linear system with Gaussian noise, the above procedure is known as the Kalman filter [125–127]. Below, we summarize the Kalman filter for a one-dimensional complex variable [13,15,17].

Let $u_m \in \mathbb{C}$ be a complex random variable whose dynamics are given by the following:

$$u_{m+1} = Fu_m + \mathcal{F}_{m+1} + \sigma_{m+1}, \tag{55}$$

where σ_{m+1} is a complex Gaussian noise with $\sigma_{m+1} = (\sigma_{1,m+1} + i\sigma_{2,m+1})/\sqrt{2}$ and it has zero mean and variance $r = \langle \sigma_{m+1}\sigma_{m+1}^* \rangle = \frac{1}{2}\sum_{j=1}^{2}\langle \sigma_{j,m+1}^2 \rangle$. Here, F is a complex number known as the forward operator and $\mathcal{F}$ is an external forcing which can vary in time. The goal of the Kalman filter is to estimate the unknown true state u_{m+1}, given noisy observations

$$v_{m+1} = gu_{m+1} + \sigma_{m+1}^o, \tag{56}$$

where g is a linear observation operator and $\sigma_m^o \in \mathbb{C}$ is an unbiased Gaussian noise with variance $r^o = \langle \sigma_m^o(\sigma_m^o)^* \rangle$. The Kalman filter is the optimal (in the least-squares sense) solution found by assuming that the model and the observation operator that relates the model state with the observation variables are both linear and both the observation and prior forecast error uncertainties are Gaussian, unbiased and uncorrelated. In particular, the observation error distribution of v at time t_{m+1} is a Gaussian conditional distribution

$$p(v_{m+1}|u_{m+1}) \sim \mathcal{N}(gu_{m+1}, r^o), \tag{57}$$

which depends on the true state u_{m+1} through (55). In (57), $p(v_{m+1}|u_{m+1})$ is known as the likelihood of estimating u_{m+1} given observation v_{m+1}.

Assume the filter model is perfectly specified [128]. An estimate of the true state prior to knowledge of the observation at time t_{m+1}, which is known as the prior state or forecast state, is given by

$$u_{m+1|m} = Fu_{m|m} + \mathcal{F}_{m+1} + \sigma_{m+1}. \tag{58}$$

See the first step in Figure 11. From the probabilistic point of view, we can represent this prior estimate with a probability density $p(u_{m+1})$. This prior distribution acounts only for the earlier observations up to time t_m,

$$p(u_{m+1}) \sim \mathcal{N}(\bar{u}_{m+1|m}, r_{m+1|m}),\tag{59}$$

where the prior mean and prior variance

$$\bar{u}_{m+1|m} \equiv \langle u_{m+1|m} \rangle,$$
$$r_{m+1|m} \equiv \langle (u_{m+1} - \bar{u}_{m+1|m})(u_{m+1} - \bar{u}_{m+1|m})^* \rangle,\tag{60}$$

are solved via

$$\bar{u}_{m+1|m} = F\bar{u}_{m|m} + \mathcal{F}_{m+1},$$
$$r_{m+1|m} = Fr_{m|m}F^* + r,\tag{61}$$

with $r_{m|m} = \langle (u_m - \bar{u}_{m|m})(u_m - \bar{u}_{m|m})^* \rangle$. Note that in order to solve the prior distribution $p(u_{m+1|m})$, the posterior information in the previous step $\bar{u}_{m|m}, r_{m|m}$ has been used.

Next, we derive the posterior state (or the filtered state) that combines the prior information $p(u_{m+1|m})$ with the observation v_{m+1} at t_{m+1}. This estimate is given in the probabilistic sense by the Bayesian update through maximizing the following conditional density,

$$p(u_{m+1}|v_{m+1}) \sim p(u_{m+1})p(v_{m+1}|u_{m+1}) = e^{-\frac{1}{2}J(u_{m+1})},\tag{62}$$

which is equivalent to minimizing

$$J(u) = \frac{(u - \bar{u}_{m+1|m})^*(u - \bar{u}_{m+1|m})}{r_{m+1|m}} + \frac{(v_{m+1} - gu)^*(v_{m+1} - gu)}{r^o}.$$

The value of u at which $J(u)$ attains its minimum is the estimate for the mean and is given by

$$\bar{u}_{m+1|m+1} = (1 - K_{m+1}g)\bar{u}_{m+1|m} + K_{m+1}v_{m+1},\tag{63}$$

where

$$K_{m+1} = \frac{gr_{m+1|m}}{r^o + g^2 r_{m+1|m}}\tag{64}$$

is the Kalman gain. Note that $0 \leq K_{m+1}g \leq 1$. The filter fully weighs to the model or prior forecast when $K_{m+1}g = 0$ and fully weighs to the observation when $K_{m+1}g = 1$. Such weights depend on the ratio of the uncertainty (reflected by the noise) in the observations and the model. Finally, the posterior variance is calculated via the following:

$$u_{m+1} - \bar{u}_{m+1|m+1} = u_{m+1} - \bar{u}_{m+1|m} - K_{m+1}(v_{m+1} - gu_{m+1} - g(\bar{u}_{m+1} - u_{m+1}))$$
$$e_{m+1|m+1} = (1 - K_{m+1}g)e_{m+1|m} - K_{m+1}\sigma^o_{m+1}.\tag{65}$$

These result in the expression of the posterior variance

$$r_{m+1|m+1} = (1 - K_{m+1}g)r_{m+1|m}.\tag{66}$$

Note that the above Kalman filter is designed for a linear system with Gaussian noise. In practice, different generalizations of the Kalman filter and various nonlinear filters such as the ensemble Kalman filter, particle filter and blended filtering techniques are applied to nonlinear and non-Gaussian systems. See [13–17,101,129,130] for more details.

3.2. Asymptotic Behavior of Prediction and Filtering in One-Dimensional Linear Stochastic Models with Model Error

Recall in Section 3.1 that the true underlying linear stochastic model is given by

$$u_{m+1} = Fu_m + \mathcal{F}_{m+1} + \sigma_{m+1}. \tag{67}$$

However, the true underlying dynamics is typically unknown in practice. Therefore, imperfect forecast models are used in the prediction stage. Now, let's assume the forecast model has the following form

$$u^M_{m+1} = F^M u^M_m + \mathcal{F}^M_{m+1} + \sigma^M_{m+1}, \tag{68}$$

where the model error comes from the imperfect forward operator, forcing and noise coefficient. Due to the appearance of such model errors, the updates of prediction and filtering distributions become

$$\bar{u}^M_{m+1|m} = F^M(1 - gK^M_m)\bar{u}^M_{m|m-1} + F^M K^M_m g u_m + F^M K^M_m \sigma^o_m + \mathcal{F}^M_{m+1},$$
$$r^M_{m+1|m} = (1 - K^M_m g)|F^M|^2 r^M_{m|m-1} + r^M, \tag{69}$$

and

$$\bar{u}^M_{m+1|m+1} = F^M(1 - gK^M_m)\bar{u}^M_{m|m} + F^M K^M_{m+1} g u_{m+1} + F^M K^M_{m+1} \sigma^o_{m+1} + \mathcal{F}^M_{m+1},$$
$$r^M_{m+1|m+1} = (1 - K^M_m g)(|F^M|^2 r^M_{m|m} + r^M), \tag{70}$$

respectively.

Now, we study the asymptotic behavior of the updates (69) and (70) with model error compared with the truth based on (67). The detailed calculations are included in Appendix C, which exploit the augmented system involving the truth and the prediction/filtering with model error [31]. Here, we summarize the results.

From (67) and (68), it is easy to show the equilibrium mean estimates of the perfect and imperfect model

$$\bar{u}_{eq} = \frac{\mathcal{F}_\infty}{1 - F}, \qquad \bar{u}^M_{eq} = \frac{\mathcal{F}^M_\infty}{1 - F^M}. \tag{71}$$

3.2.1. Prediction

The asymptotic prediction mean of $\bar{u}^M_{m+1|m}$ is given by

$$\lim_{m\to\infty} \mathbb{E}(\bar{u}^M_{m+1|m}) = \frac{1}{(1 - F^M) + F^M K^M_\infty g}\left(F^M K^M_\infty g \bar{u}_{eq} + (1 - F^M)\bar{u}^M_{eq}\right). \tag{72}$$

Clearly, the asymptotic mean of the prediction state is a linear combination of the equilibrium mean of original true model $\bar{u}_{eq}$ and that of the forecast model of the mean $\bar{u}^M_{eq}$. With (67), the error in the asymptotic prediction mean of $\bar{u}^M_{m+1|m}$ yields,

$$\lim_{m\to\infty} \mathbb{E}(u_{m+1} - \bar{u}^M_{m+1|m}) = \frac{1 - F^M}{(1 - F^M) + F^M K^M_\infty g}(\bar{u}_{eq} - \bar{u}^M_{eq}). \tag{73}$$

The asymptotic prediction mean of $\bar{u}^M_{m+1|m}$ is equal to the equilibrium mean of the perfect model if and only if the imperfect model has the same equilibrium mean as the perfect model, namely $\bar{u}_{eq} = \bar{u}^M_{eq}$. On the other hand, the asymptotic prediction variance $r^P_\infty = \lim_{m\to\infty} r_{m+1|m}$ is given by

$$r^P_\infty = |F^M|^2(1 - K^M_\infty g)r^P_\infty + r^M = |F^M|^2 \frac{r^o r^P_\infty}{g^2 r^P_\infty + r^o} + r^M, \tag{74}$$

where the asymptotic Kalman gain is

$$K_\infty^M = \frac{g r_\infty^P}{g^2 r_\infty^P + r^o}.$$

Therefore, the asymptotic prediction variance simplifies to

$$r_\infty^P = \frac{|F^M|^2 r^o}{g} K_\infty^M + r^M. \tag{75}$$

3.2.2. Filtering

The asymptotic filtering mean of $\bar{u}_{m+1|m+1}^M$ is given by

$$\lim_{m\to\infty} \bar{u}_{m+1|m+1}^M = \frac{1}{(1 - F^M) + F^M K_\infty^M g} \left(K_\infty^M g \bar{u}_{eq} + (1 - F^M)(1 - K_\infty^M g) \bar{u}_{eq}^M \right) \tag{76}$$

with (67), the error in the asymptotic filtering mean of $\bar{u}_{m+1|m+1}^M$ yields,

$$\lim_{m\to\infty} \mathbb{E}(u_{m+1} - \bar{u}_{m+1|m+1}) = \frac{(1 - F^M)(1 - K_\infty^M g)}{(1 - F^M) + F^M K_\infty^M g} (\bar{u}_{eq} - \bar{u}_{eq}^M). \tag{77}$$

The asymptotic prediction mean of $\bar{u}_{m+1|m+1}^M$ is equal to the equilibrium mean of the perfect model provided that (1) the imperfect model has the same equilibrium mean as the perfect model, namely $\bar{u}_{eq} = \bar{u}_{eq}^M$, or (2) $K_\infty^M g = 1$, namely the observational noise is zero and the filter trusts completely towards the observations.

On the other hand, the dynamics of prediction variance $r_\infty^A = \lim_{m\to\infty} r_{m+1|m+1}$ are given by

$$r_\infty^A = (1 - K_{m+1}^M g)(|F^M|^2 r_\infty^A + r^M) \tag{78}$$

with direct manipulations, the asymptotic analysis variance becomes

$$r_\infty^A = \frac{r^o}{g} K_\infty^M. \tag{79}$$

3.2.3. Comparison

Comparing the asymptotic prediction mean (73) and filtering mean (77), we have

$$\lim_{m\to\infty} \mathbb{E}(u_{m+1} - \bar{u}_{m+1|m+1}) = (1 - K_\infty^M g) \lim_{m\to\infty} \mathbb{E}(u_{m+1} - \bar{u}_{m+1|m}), \tag{80}$$

which indicates that the error in the filtering mean is always smaller than that in the prediction mean in the sense of standard mean deviation, with the existence of observational error.

Next, comparing the asymptotic prediction variance (75) and filtering variance (79), we have

$$r_\infty^A - r_\infty^P = \frac{r^o}{g} K_\infty^M - \frac{|F^M|^2 r^o}{g} K_\infty^M - r^M = (1 - |F^M|^2) r_\infty^P - r^M < 0. \tag{81}$$

See Appendix C for the detailed derivations. This implies that filtering state always results in a smaller uncertainty (variance) than the prediction state. Such an uncertainty reduction is due to the extra information in the noisy observations.

Note that the conclusions made from (80) and (81) are valid only when full observations are available. In high dimensional situations, if the observations are only available on part of the variables (known as partial observations), then the prediction error can be smaller than filtering error. See Section 3.5.1 for simple examples.

3.3. An Information Theoretical Framework for State Estimation and Prediction

3.3.1. Motivation Examples

To illustrate the importance and necessity of developing an information theoretical framework in assessing the filtering and predicting skill, let us first review the two traditional path-wise measurements that are widely used in filtering and prediction [13,131–135]. Denote u_i, $i = 1, \ldots, n$ the true signal and $\hat{u}_i$ the filtering/prediction estimate. These measurements are given by

1. The root-mean-square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{u}_i - u_i)^2}{n}}. \tag{82}$$

2. The pattern correlation (PC):

$$\text{PC} = \frac{\sum_{i=1}^{n}(\hat{u}_i - \overline{\hat{u}}_i)(u_i - \overline{u}_i)}{\sqrt{\sum_{i=1}^{n}(\hat{u}_i - \overline{\hat{u}}_i)^2}\sqrt{\sum_{i=1}^{n}(u_i - \overline{u}_i)^2}}, \tag{83}$$

where $\overline{\hat{u}}_i$ and $\overline{u}_i$ denotes the mean of $\hat{u}_i$ and u_i, respectively.

While these two path-wise measurements are easy to implement and are able to quantify the filtering/prediction skill to some extent, they have fundamental limitations. To see this, let us consider a simple motivation example, where the true dynamics is given by

$$\frac{du}{dt} = -\gamma u + f_0 + f_1 e^{i\omega_1 t} + \sigma \dot{W}. \tag{84}$$

with parameters

$$\gamma = 1, \quad f_0 = 0, \quad f_1 = 1, \quad \omega_1 = 1, \quad \sigma = 2. \tag{85}$$

For the imperfect forecast model, we assume the same ansatz as the prefect one in (84) but the parameters related to the forcing amplitudes contain model errors. Consider the following two imperfect models:

$$\begin{array}{llllll}
\text{Imperfect forecast model } (a): & \gamma = 1, & f_0^M = 0, & f_1^M = 0, & \omega_1 = 1, & \sigma = 2, \\
\text{Imperfect forecast model } (b): & \gamma = 1, & f_0^M = 0, & f_1^M = 2, & \omega_1 = 1, & \sigma = 2.
\end{array} \tag{86}$$

In panels (a) and (b) of Figure 12, we show the predictions using these two imperfect forecast model (green curve) and they are compared with the truth (blue curve). Here, the observational time step is $\Delta t_{obs} = 0.5$ which is less than the decorrelation time $\tau = 1/\gamma = 1$, and the observational noise level is $r^o = 1$. In terms of the RMSE and PC, the two predictions are comparable with each other and the one in panel (a) is even slightly more skillful. However, the prediction in panel (a) by intuition is worse than that in panel (b). In fact, the amplitude of the prediction using the imperfect model (a) is severely underestimated. The consequence is that the prediction fails to capture all the important extreme events in the true signal. On the other hand, the prediction using the imperfect model (b) results in a time series which has the same amplitude as the truth. See the PDFs associated with the time series in panel (d) for estimating the amplitudes. Therefore, the two traditional path-wise measurements—RMSE and PC—are misleading here in providing the prediction skill. In addition, according to the definitions of the RMSE and the PC in (82) and (83), both the measurements take into account the information only up to the second-order statistics. Therefore, they are not able to capture the information beyond Gaussian statistics and are not suitable to assess the filtering/prediction skill for any non-Gaussian models as in nature.

Due to the above fundamental limitations of these two traditional path-wise measurements, various information measurements become useful in assessing the filtering/predictino skill.

In [34,56,136–138], an information measurement called Shannon entropy difference was introduced and was used to assess the filtering/prediction skill. The Shannon entropy difference is defined as

$$\mathcal{S}(\pi) - \mathcal{S}(\pi^M) = -\int \pi \log \pi + \int \pi^M \log \pi^M. \tag{87}$$

In particular, if both $\pi \equiv \pi_G$ and $\pi^M \equiv \pi_G^M$ are Gaussian (as in the linear models), then the Shannon entropy difference has the following explicit formula:

$$\begin{aligned}
\mathcal{S}(\pi_G) - \mathcal{S}(\pi_G^M) &= \left(\frac{1}{2} \log \det R + \frac{1}{2}(1 + \log(2\pi)) \right) - \left(\frac{1}{2} \log \det R^M + \frac{1}{2}(1 + \log(2\pi)) \right) \\
&= \frac{1}{2} \log \det \left(R R^M \right),
\end{aligned} \tag{88}$$

where R and R^M are the covariance of π_G and π_G^M. Intuitively, the Shannon entropy difference quantifies the uncertainty between π and π^M. For Gaussian distributions, the uncertainty is reflected by the variance. Connecting the Shannon entropy difference with the two predictions in panels (a) and (b) of Figure 12, it is expected that the Shannon entropy difference is able to distinguish the two predictions since the associated PDFs of the two predictions have different variances. In fact, the Shannon entropy difference in the imperfect model (a) (0.7122) is much larger than that in the imperfect model (b) (0.1502), which indicates that the prediction in panel (b) is more skillful than that in panel (a).

However, relying solely on the Shannon entropy difference in assessing the filtering/prediction skill is also misleading. Consider an imperfect model with the following parameters:

$$\text{Imperfect forecast model } (c): \qquad \gamma = 1, \qquad f_0^M = 2, \qquad f_1^M = 2, \qquad \omega_1 = 1, \qquad \sigma = 2. \tag{89}$$

Comparing with the perfect model and the other two imperfect models in (86), a non-zero constant forcing $f_0^M = 2$ is imposed in (89). The prediction results are shown in panel (c) of Figure 12. Since the Shannon entropy difference in Gaussian framework (88) takes into account only the the variance but completely ignores the information in the mean, the resulting Shannon entropy difference using the imperfect models (b) and (c) give exactly the same value. In addition, the pattern correlations in these two models are also identical to each other. However, the prediction using the imperfect model (c) has an obvious mean bias and therefore the prediction is not as skillful as that using the imperfect model (b).

From these simple motivation examples, it seems that the combination of the RMSE, the PC and the Shannon entropy difference can overcome the fundamental limitations as discussed above. However, there are at least two extra shortcomings even in the combination of these three measurements. First, two different PDFs associated with the imperfect model, namely $\pi_{\{1\}}^M$ and $\pi_{\{2\}}^M$, may result in the same Shannon entropy difference compared with the truth. For example, such situation happens when $\pi_{\{2\}}^M$ has a mean shift compared with $\pi_{\{1\}}^M$. This is because the Shannon entropy difference computes the uncertainty of the two distributions separately rather than considering the pointwise difference between the two PDFs. Therefore, a more sophisticated measurement should take into account the relative difference between the PDFs associated with the perfect and imperfect models. Second, as has been discussed above, the RMSE and PC only make use of the information up to the second order statistics. The important non-Gaussian features as they appeared in many realistic applications are not reflected in these path-wise measurements.
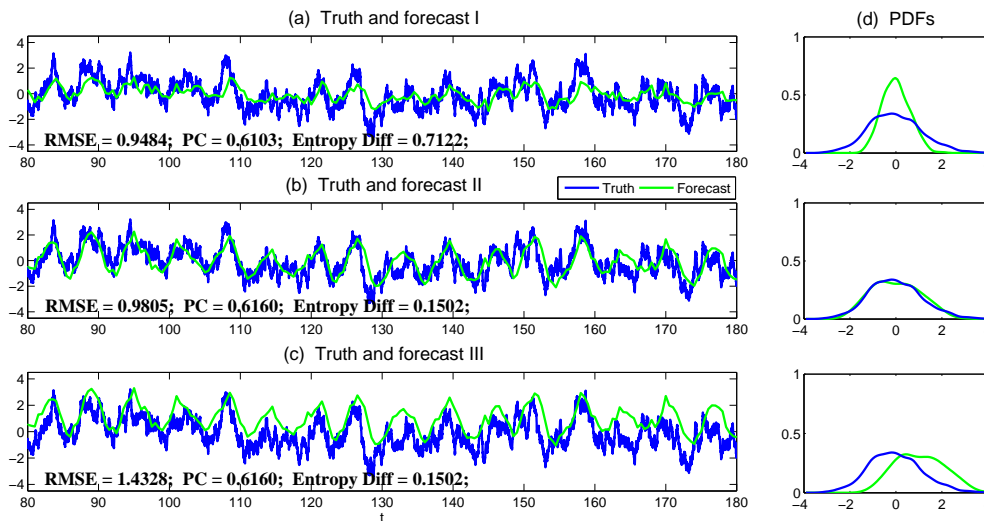
Figure 12. Motivation examples for the limitations of assessing the prediction error based only the Root-mean-square error (RMSE) and the pattern correlation (**a**,**b**) and based only on the Shannon entropy difference (**b**,**c**). Here, the truth is the same in (**a**–**c**), which is generated from (84) and (85). The three imperfect forecast models are given by (86) and (89). In column (**d**), the associated PDFs are shown. In all the panels, only the real part of u is shown.

3.3.2. Assessing the Skill of Estimation and Prediction Using Information Theory

Due to the fundamental limitations in the two classical path-wise measurement, RMSE and PC, as well as those in the Shannon entropy difference, a new information-theoretic framework [102] has developed to assess the filtering/prediction skill. Again, denote $\pi \equiv \pi(u)$ and $\pi^M \equiv \pi(u^M)$ the PDFs associated with truth u and the filtering/prediction estimate u^M, respectively. Denote $p(u, u^M)$ the joint PDF of u and u^M. Let $\mathcal{U} = u - u^M$ be the residual between the truth and the estimate. This information-theoretic framework involves three information measurements:

1. The Shannon entropy residual,

$$\mathcal{S}(\mathcal{U}) = -\int p(\mathcal{U}) \log p(\mathcal{U}). \tag{90}$$

2. The mutual information,

$$\mathcal{M}(\pi, \pi^M) = \iint p(u, u^M) \log \left(\frac{p(u, u^M)}{\pi(u)\pi(u^M)} \right). \tag{91}$$

3. The relative entropy,

$$\mathcal{R}(\pi, \pi^M) = -\int \pi \log \left(\frac{\pi}{\pi^M} \right). \tag{92}$$

The Shannon entropy residual quantifies the uncertainty in the point-wise difference between u and u^M. It is an information surrogate of the RMSE in the Gaussian framework. The mutual information quantifies the dependence between the two processes. It measures the lack of information in the factorized density $\pi(u)\pi(u^M)$ relative to the joint density $p(u, u^M)$, which follows the identity,

$$\mathcal{M}(\pi, \pi^M) = \mathcal{P}\Big(p(u, u^M), \pi(u)\pi(u^M)\Big). \tag{93}$$

The mutual information is an information surrogate of the PC in the Gaussian framework. On the other hand, the relative entropy quantifies the lack of information in π^M related to π and it is a good indicator of the skill of u^M in capturing the peaks and extreme events of u. It also

takes into account the pointwise discrepancy between π and π^M rather than only computing the difference between the uncertainties associated with the two individual PDFs (as in the Shannon entropy difference). Therefore, the combination of these three information measurements is able to capture all the features in assessing the filtering/prediction skill and overcomes the shortcomings as discussed in the previous subsection.

Note that when $\pi \sim \mathcal{N}(\overline{u}, R)$ and $\pi^M \sim \mathcal{N}(\overline{u}^M, R_M)$ are both Gaussian, then the above three information measurements have explicit expressions:

1. The Shannon entropy residual (Gaussian framework),

$$\mathcal{S}(\mathcal{U}) = \frac{1}{2} \log \det \left(R + R^M - 2Re[Cov(u, u^M)] \right). \tag{94}$$

2. The mutual information (Gaussian framework),

$$\mathcal{M}(\pi, \pi^M) = -\frac{1}{2} \log \det \left(I - R_M^{-1} Cov^*(u, u^M) R^{-1} Cov(u, u^M) \right). \tag{95}$$

3. The relative entropy (Gaussian framework),

$$\mathcal{R}(\pi, \pi^M) = \frac{1}{2}(\overline{u} - \overline{u}^M)^* R_M^{-1}(\overline{u} - \overline{u}^M) + \frac{1}{2} \left(-\log \det(RR_M^{-1}) + \mathrm{tr}(RR_M^{-1}) - N \right). \tag{96}$$

In (94)–(96), N is the dimension of π and π^M, I is the identity matrix with size $N \times N$, and $Cov(u, u^M)$ is the covariance between u and u^M. More discussions of Gaussian and non-Gaussian cases can be found in [56] and [54], respectively.

The information-theoretic framework (90)–(92) or (94)–(96) is usually defined in the super-ensemble sense [31] in accessing the data assimilation and prediction skill of the imperfect model given the perfect one. However, in some more realistic situations, although the imperfect models can be run in the ensemble mode, the ensemble run of the perfect model or the truth is never available. This is because the perfect model that describes nature is unknown. The only available information is one realization from observations (e.g., satellites). Nevertheless, the information-theoretic framework can also be used in a path-wise way, where the statistics are computed by collecting all the sample points in the given realization. Some realistic applications of the information-theoretic framework for filtering and prediction can be found in [31,61,139].

3.4. State Estimation and Prediction for Complex Scalar Forced Ornstein–Uhlenbeck (OU) Processes

Now, we study the state estimation (filtering) and prediction. The focus here is a complex scalar forced Ornstein–Uhlenbeck (OU) process,

$$\frac{du}{dt} = (-\gamma + i\omega_0)u + f_0 + f_1 e^{i\omega_1 t} + \sigma \dot{W}, \tag{97}$$

where γ and ω_0 are the damping and oscillation frequency while f_0 and $f_1 e^{i\omega_1 t}$ are a constant and time-periodic large-scale forcing, respectively, and $\sigma \dot{W}$ is stochastic noise. Despite the simplicity of this model in (97), it can be used to mimic some climate physics [15,120]. For example, the deterministic forcing $f_0 + f_1 e^{i\omega_1 t}$ can be regarded as the annual cycle while ω_0 can be treated as internal oscillation which may occur in the intreaseasonal time scale. The damping term γ measures the system memory and the stochastic term represents the input to the system from small or unresolved scales. The model in (97) can also be regarded as one Fourier mode of complex spatially-extended systems.

The information-theoretic framework developed above will be used to assess the filtering/prediction skill and quantify the model error. The complex scalar forced OU process in (97) will be used to generate the true signal. The imperfect forecast model has the same structure as (97) but with model errors in the

parameters. The goal here is to systematically study the model error as functions of the observational time step, observational noise and the forcing amplitude.

The exact solution of (97) can be written down explicitly,

$$
u(t) = u(t_0)e^{(-\gamma+i\omega_0)(t-t_0)} + \frac{f_0}{\gamma - i\omega_0}\left(1 - e^{(-\gamma+i\omega_0)(t-t_0)}\right)
$$
$$
+ \frac{f_1 e^{i\omega_1 t}}{\gamma + i(-\omega_0 + \omega_1)}\left(1 - e^{-(\gamma+i\omega_1-i\omega_0)(t-t_0)}\right) + \sigma \int_{t_0}^{t} e^{(-\gamma+i\omega_0)(t-s)}dW(s),
\tag{98}
$$

which provides the analytical forms of the time evolution of the forecast mean $\overline{u}(t)$ and the forecast variance $r(t)$,

$$
\overline{u}(t) = \overline{u}(t_0)e^{(-\gamma+i\omega_0)(t-t_0)} + \frac{f_0}{\gamma - i\omega_0}\left(1 - e^{(-\gamma+i\omega_0)(t-t_0)}\right)
$$
$$
+ \frac{f_1 e^{i\omega_1 t}}{\gamma + i(-\omega_0 + \omega_1)}\left(1 - e^{-(\gamma+i\omega_1-i\omega_0)(t-t_0)}\right),
\tag{99}
$$
$$
r(t) = r(t_0)e^{-2\gamma(t-t_0)} + \frac{\sigma^2}{2\gamma}\left(1 - e^{-2\gamma(t-t_0)}\right).
$$

With the explicit expressions in (98) and (99), it is easy to write down the corresponding operators $F, \mathcal{F}_{m+1}$ and σ_{m+1} in (67) or those in the imperfect forecast model (68) in each prediction/filtering assimilation step.

To understand the prediction and filtering skill of the complex scalar forced OU process (97), we start with a simple case which involves only a constant forcing. We also adopt the perfect forecast model in this example. The parameters of (97) are given as follows:

$$
\gamma = 0.4, \qquad \omega_0 = 1, \qquad f_0 = 2, \qquad f_1 = 0, \qquad \omega_1 = 0, \qquad \sigma = 1,
\tag{100}
$$

and the observational operator $g = 1$. Here, we take $\Delta t_{obs} = 0.5$ and $r^o = 0.5$ as the default values of the observational time step and the observational noise level, respectively. Since the phase $\omega_0 = 1$ is nonzero, the autocorrelation function has an oscillation structure. Therefore, the decorrelation time which includes the cancellation of the positive and negative values in the autocorrelation function may be misleading for measuring the system memory. Here, we have checked both the standard decorrelation time, namely the integral of the autocorrelation function (ACF), and the integral of the absolute value of the ACF. These two quantities have values $\tau_{ACF} = 0.34$ and $\tau_{|ACF|} = 1.63$. Thus, $\Delta t_{obs} = 0.5$ is a reasonable observational time step that includes some memory of the information in the previous assimilation step. On the other hand, $r^o = 0.5$ here results in a polluted signal with roughly 10% observational noise.

Figure 13 show the prediction and filtering skill in terms of the three information measurements, namely the Shannon entropy residual, the mutual information and the relative entropy, as a function of Δt_{obs} (panels (a)–(c)) and r^o (panels (d)–(f)), respectively. First, the filtering estimate is always more skillful than the prediction one. This is consistent with the theoretical analysis in Section 3.2 in that the filter estimate combines the prediction result with the information observation and the error is therefore reduced. Next, it is clear from all the panels in Figure 13 that with the increase of either Δt_{obs} or r^o, both the prediction and filtering skill deteriorates, which is as expected. Nevertheless, the prediction skill decreases more quickly with the increase of the observational time step Δt_{obs}. In particular, the model error in terms of the relative entropy increases exponentially. As comparison, the filtering skill only has a slight deterioration with the change of Δt_{obs}. On the other hand, the error in the filter estimate increases quickly with the observational noise r^o when r^o is small. When r^o becomes moderate to large, the error in the filter estimate increases steadily. The error in the prediction estimate always increases steadily as a function of r^o.

To have a more intuitive understanding of these results, the time series of the truth, the filter estimate and the prediction estimate are shown in Figure 14 with three different Δt_{obs} or r^o. Comparing with panels (a) and (b), it is clear that a long observational time step Δt_{obs} leads to a smaller fluctuation of the prediction estimate around its steady state mean value. In fact, the signal due to the memory effect in the previous assimilation step is strongly damped with a long observational time step and the resulting signal is dominated by the constant forcing. The consequence is that the PDF associated with the predicted time series has a much smaller variance compared with the truth and the prediction fails to capture the extreme events and large variabilities in the true signal. On the other hand, due to the incorporation of the information from the observations, the filter estimate even with a long observational time step provides a quite skillful result in terms of both the correlation and the signal amplitude. Note that the asymptotic Kalman gain in the filtering in panel (b) is $K_\infty = 0.9$, which means the observations play an important role in regaining the skill in the filter estimate. In panel (c), the observational time step $\Delta t_{obs} = 0.5$ is the same as that in panel (a) but the observational noise level is increased from $r^o = 0.5$ to $r^o = 3$. Both the filtering and prediction skill becomes worse as compared with those in panel (a). Nevertheless, the deterioration is not quite significant which is consistent with the statistics shown in Figure 13.

Figure 13. The three information measurements, namely the Shannon entropy residual, the mutual information and the relative entropy, as a function of Δt_{obs} (**a**–**c**) and r^o (**d**–**f**). Here, the experiments are based on the perfect model (97) with parameters in (100). The left small panel shows the autocorrelation function of $Re[u]$.
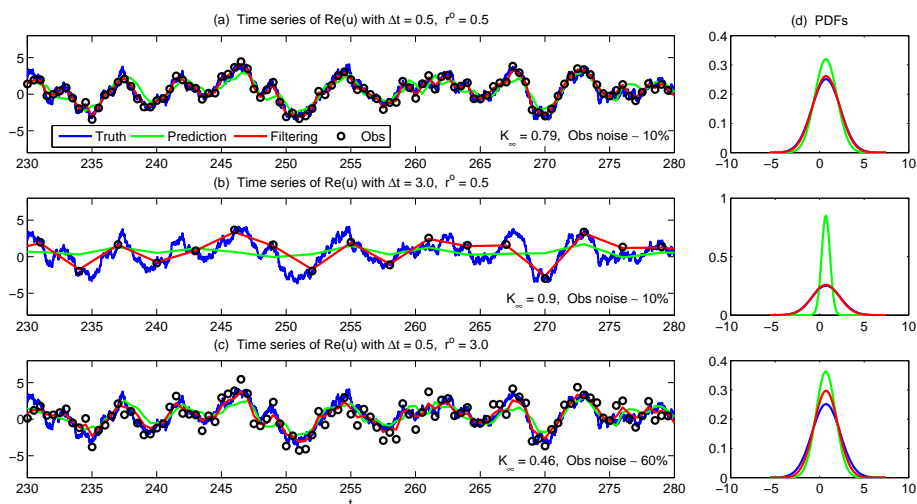
Figure 14. Comparison of the time series of the truth, the prediction estimate and the filter estimate. (**a**): $\Delta t_{obs} = 0.5, r^o = 0.5$; (**b**): $\Delta t_{obs} = 3.0, r^o = 0.5$; (**c**): $\Delta t_{obs} = 0.5, r^o = 3.0$; (**d**): the associated PDFs. Here, the experiments are based on the perfect model (97) with parameters in (100).

Next, we consider the complex forced scalar system with a time-periodic forcing in (97). The parameters are as follows:

$$\gamma = 0.4, \qquad f_0 = 0, \qquad f_1 = 2, \qquad \omega_1 = 1, \qquad \sigma = 2, \tag{101}$$

and the observational operator $g = 1$. Two dynamical regimes are studied here:

$$\text{Regime I:} \qquad \omega_0 = 0.5, \qquad \text{and} \qquad \text{Regime II:} \qquad \omega_0 = 1. \tag{102}$$

It is important to note that, in Regime II, the phase ω_0 and the forcing period ω_1 are equal to each other, which means this dynamical regime has a resonance forcing. On the other hand, the dynamical Regime I has a non-resonance forcing. We take $\Delta t_{obs} = 0.5$ and $r^o = 9$ as the default values of the observational time step and the observational noise level, respectively. Note that although $r^o = 9$ is much larger than that in the previous example, the signal amplitude due to the periodic forcing also increases. The observational noise here is about 25% compared with the true signal. See Figure 15 for the true signal and the noisy observations.

Now, let us assume the imperfect model shares the same model structure as the perfect one in (97). The imperfect part comes from the parameter ω_0^M. This comes from the motivation that the situation that the large-scale forcing $f_0 + f_1 e^{i\omega_1 t}$ is in general known quite well but measuring the internal oscillation ω_0 usually contains error. In Figure 16, we show the model error in terms o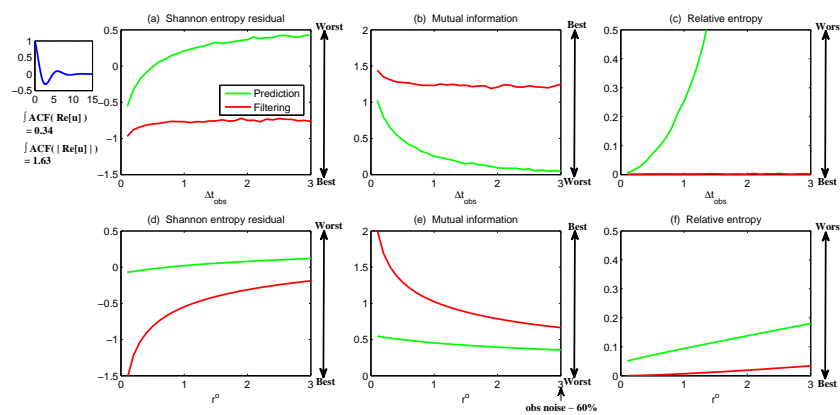f the three information measurements, namely the Shannon entropy residual, the mutual information and the relative entropy, as a function of ω_0^M. Here, the information measurements are computed based on the Gaussian framework (94)–(96) due to their simplicity from a practical point of view, where the statistics here are averaged directly over the time series. Note that such direct average results in a bimodal distribution in the true model due to the large amplitude of the periodic forcing such that the mixing and ergodicity are not satisfied [108] (see Figure 15 and a detailed discussion in Appendix D). Nevertheless, the Gaussian approximation in the information measurements here provides a qualitatively accurate estimate of the model error as can be seen in Figures 15–17. There is an alternative way of computing the model error by first collecting all the points $t' + mT, m = 1, 2, \ldots$ with t' fixed and T being the period of the time series. These points appear in the same location within a period and the collection forms a Gaussian distribution. Compute the information measurements for these Gaussian distributions. Then, let t' vary within $t' \in (0, T]$ and repeat the above procedure. Eventually, take the average of the information measurements within one period to finalize the results. This alternative method is more rigorous in terms of applying the Gaussian framework of the information measurements. However, it requires a very long time series to guarantee the sampling size (the number of period) is sufficient. It also requires knowing the perfect information of the period, which may not be realistic in practice if the period contains randomness.

In Figure 16, the minimum of the model error appears around the perfect value. The small bias in the optimal value compared with the truth comes from applying the Gaussian framework of the information measurements. When the discrepancy between ω_0^M and the truth ω_0 increases, the model error in all the three information measurements becomes large as well. Despite the similar profiles in the model error curves in the two regimes, the model error increases significantly faster in Regime II (the resonance regime). In fact, according to (98) or (99), the contribution of the time-periodic forcing to the forecast solution is given by

$$\text{Contribution of the time-periodic forcing} = \frac{f_1 e^{i\omega_1 t}}{\gamma + i(-\omega_0 + \omega_1)} \left(1 - e^{-(\gamma + i\omega_1 - i\omega_0)(t - t_0)}\right). \tag{103}$$

In addition to the error appears in the phase $e^{-(\gamma + i\omega_1 - i\omega_0)}$ due to an imperfect ω_0^M, the resonance forcing also greatly modifies the amplitude of the contribution in (103). With a resonance forcing $\omega_0 = \omega_1$, the amplitude in the contribution (103) reduces to f_1/γ which can be much larger than

$f_1/(\gamma + i(-\omega_0 + \omega_1))$ if ω_0 is quite different from ω_1 and γ is small. Recall in (101) that $\gamma = 0.4$ here. If in the imperfect model $\omega_0^M = 0 \neq 1 = \omega_0$, then a large error in the amplitude of the forcing contribution using the imperfect is expected. This is shown in Panel (b) of Figure 15. It is clear that in addition to the phase shift in both the filtering and prediction estimates, the amplitudes of these estimates are severely underestimated as well. Notably, the relative entropy here unambiguously indicates such underestimations of the amplitudes and extreme events, which cannot be captured by the RMS error and the pattern correlation.

In order to reduce the model error in the imperfect model, a typical strategy is to optimize the noise coefficient σ^M [1,13,15,101,140]. In Figure 17, we show the model error in the perfect model as a function of σ^M, where $\omega_0^M = 0$. Comparing to the non-optimized values $\sigma^M = \sigma = 2$ as indicated by the blue 'x', the model error with the optimal value $\sigma^M = 7$ for the filter estimate (which is also nearly the optimal value for the prediction estimate) has a significant decrease. This noise inflation strategy is in fact consistent with that in dealing with many operational models or complex dynamical systems [18,141,142]. Figure 17 also confirms that noise inflation leads to a much smaller model error than the underdispersion [15,18,100,140]. In Panel (c) of Figure 15, the filter and prediction estimates with this optimized noise are shown. The amplitudes of the true signal are recaptured by the imperfect model estimates. Interestingly, the filter estimates are now almost perfectly in phase with the true signal and even the discrepancy between prediction estimates and the truth is greatly decreased. See Panel (b) for a comparison. In fact, we note that the Kalman gain has increased from $K_\infty = 0.27$ to $K_\infty = 0.73$, which implies that the observations now play a more important role in obtaining the filter estimates. This is the underlying reason that the filtering becomes more skillful, which also increases the skill of the prediction since the filter estimate now provides a much more accurate initial value of the prediction.
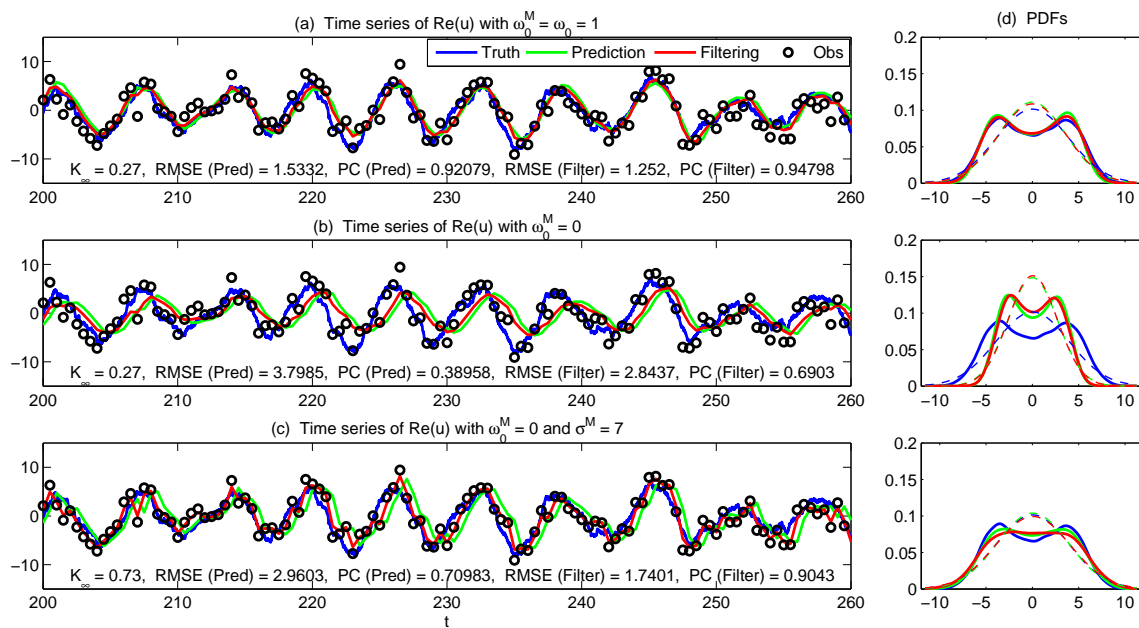
Figure 15. The true signal and filter and prediction estimates. (**a**): perfect model simulation; (**b**): imperfect forecast model with $\omega_0^M = 0 \neq 1 = \omega_0$; (**c**): imperfect forecast model with $\omega_0^M = 0 \neq 1 = \omega_0$ and optimized noise coefficient $\sigma^M = 7$. Here, the true model is given by (97) and the parameters are shown in (101) and (102); (**d**): the associated PDFs formed by directly collecting all the points in the time series (solid curves) and the Gaussian fits (dashed curves).
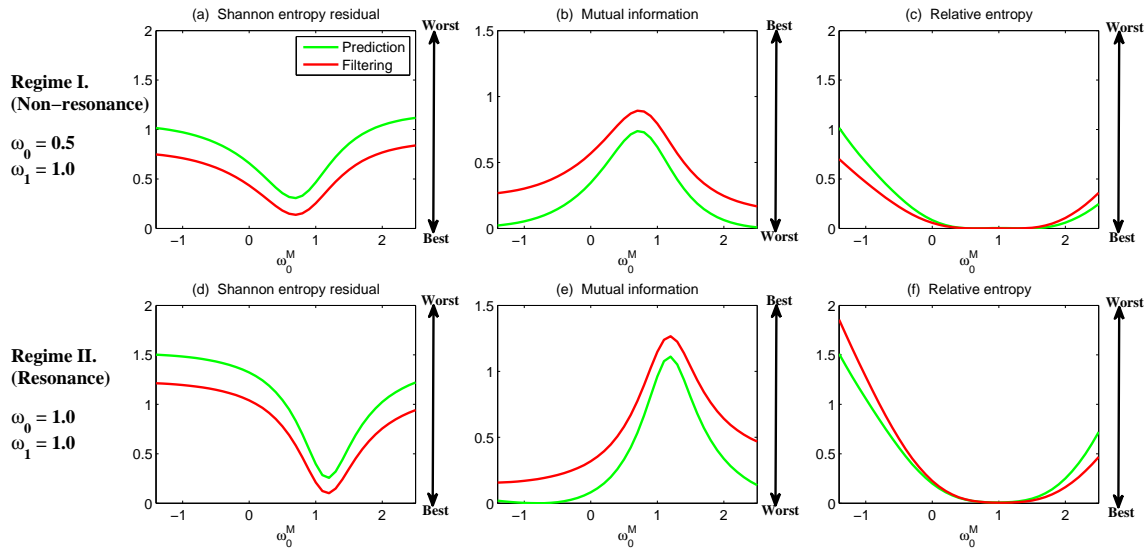
Figure 16. The three information measurements, namely the Shannon entropy residual, the mutual information and the relative entropy, as a function of ω_0^M in the imperfect model. The information measures are given using the Gaussian approximation (94)–(96) and the statistics here are averaged directly over the time series. (**a**–**c**): Regime I (the non-resonance regime); (**d**–**f**): Regime II (the resonance regime). The true model is given by (97) and the parameters are shown in (101) and (102). The imperfect model has the same structure and the same other parameters expect ω_0^M.
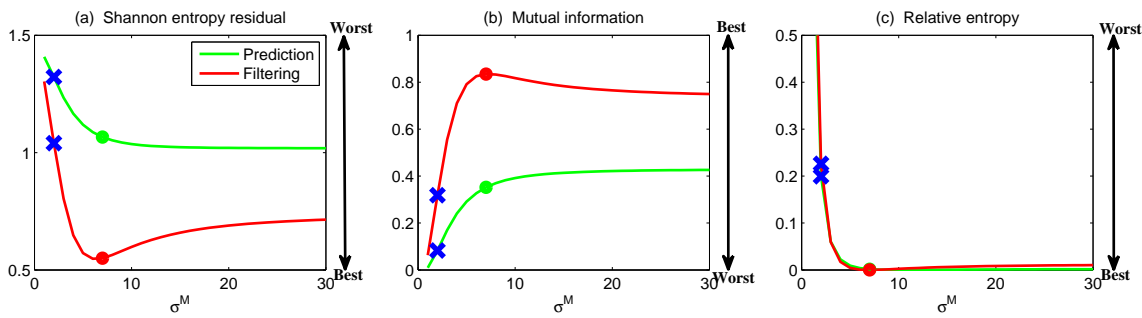
Figure 17. Model error as a function of σ^M in the imperfect model where $\omega_0^M = 0 \neq 1 = \omega_0$. The blue 'x' shows the non-optimized values $\sigma^M = \sigma = 2$. The dot $\sigma^M = 7$ indicates the optimal value for the filter estimate which is also nearly the optimal value for the prediction estimate. (**a**) Shannon entropy residual, (**b**) Mutual information, (**c**) Relative entropy.

3.5. State Estimation and Prediction for Multiscale Slow-Fast Systems

Multiscale slow-fast systems are commonly seen in many geophysical and engineering turbulent flows [18,143–146]. A concrete example involves the coupling of random incompressible geostrophically balanced (GB) flows and random rotating compressible gravity waves in the middle latitude atmosphere [8]. Under the situation with a small Rossby number, the coupled system becomes a multiscale slow-fast system where the GB component dominates the slow-varying geophysical flows [8,147–149].

3.5.1. A 3×3 Linear Coupled Multiscale Slow-Fast System

Here, we start with a simple 3×3 linear coupled multiscale slow-fast system,

$$
\begin{aligned}
\frac{du_1}{dt} &= -d_{u_1} u_1 + L_{12} u_2 + L_{13} u_3 + F_1(t) + \sigma_1 \dot{W}_1, \\
\frac{du_2}{dt} &= L_{21} u_1 - d_{u_2} u_2 + \frac{L_{23}}{\epsilon} u_3 + F_2(t) + \sigma_2 \dot{W}_2, \\
\frac{du_3}{dt} &= L_{31} u_1 + \frac{L_{32}}{\epsilon} u_2 - d_{u_3} u_3 + F_3(t) + \sigma_3 \dot{W}_3.
\end{aligned}
\tag{104}
$$

In (104), we assume the linear coefficients $L_{12} = -L_{21}, L_{13} = -L_{31}$ and $L_{23} = -L_{32}$ such that the L_{ij} forms a skew-symmetric matrix. The three damping coefficients $-d_{u_1}, -d_{u_2}, -d_{u_3} < 0$ to guarantee the mean stability. $F_1(t), F_2(t)$ and $F_3(t)$ are external forcing that can depend on time t. Here, ϵ is a controllable parameter. With $\epsilon \ll 1$, the coupled system has a fast oscillation structure in u_2 and u_3 while u_1 remains as a slow variable. All the variables here are real.

The coupled system in (104) can be regarded as one Fourier mode of the shallow water equations, where u_1 mimics the large-scale GB flow while u_2 and u_3 represent the analogies of the real and imaginary parts of the gravity waves. Note that the gravity waves appear in pairs and therefore the linear combinations of u_2 and u_3 in the complex plane are good surrogates of the two components of the gravity waves associated with one Fourier mode in the shallow water equation. These three variables are coupled in a linear way in (104).

Below, we study the filtering/prediction skill. The following parameters are taken:

$$
\begin{aligned}
d_{u_1} = d_{u_2} = d_{u_3} = 1, \quad \sigma_1 = \sigma_2 = \sigma_3 = 1, \quad L_{12} = L_{13} = 1, \quad L_{21} = L_{31} = -1, \\
L_{23} = 1, \quad L_{32} = -1, \quad F_1 = 2\cos(0.5t), \quad F_2 = F_3 = 0.
\end{aligned}
\tag{105}
$$

Here, we only impose the deterministic time-periodic forcing to u_1. This is because we denote u_1 as the slow (or large) scale variable, which is typically driven by external forcing, such as the seasonal cycle or annual cycle [15]. On the other hand, the other two variables mostly occur in a faster time scale and the forcing is basically stochastic.

To understand the filtering/prediction skill, the following four setups are adopted:

1. *Full observations, full forecast model (F/F).* The observational operator g is an identity such that

$$
\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \begin{pmatrix} \sigma_1^o \\ \sigma_2^o \\ \sigma_3^o \end{pmatrix}.
\tag{106}
$$

The forecast model is the same as in (105). Although this straightforward setup may not be practical (see below) and can be expensive when a much larger dimension of the system is considered (see next subsection), the results from such a setup can be used as a baseline for testing various modifications and reduced models as will be presented below.

2. *Partial observations, full forecast model (P/F).* The real observations typically involve the superposition of different wave components. It is usually impossible to artificially separate these components from the noisy observations. Therefore, here we let the observational operator be $g = (1, 1, 1)$, namely the observation is the combination of the three variables,

$$
v = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \sigma^o.
\tag{107}
$$

The forecast model remains the same as that in (105).

3.   *Partial observations, reduced forecast model (P/R).* In practice, only part of the state variables are of particular interest in filtering and prediction. These state variables usually lie in large or resolved scales, such as the GB flow. Therefore, simple reduced forecast models are typically designed to reduce the computational cost and retain the key features in filtering and predicting these variables. To this end, the following reduced forecast model is used

$$\frac{du_1^M}{dt} = -d_{u_1} u_1^M + F_1(t) + \sigma_1 \dot{W}_1, \tag{108}$$

and the observation remains the same as that in (107). Here, we have completely dropped the dependence of u_1 on u_2 and u_3 since their mean is zero according to the setup above.

4.   *Partial observations, reduced forecast model and tuned observational noise level with inflation (P/R tuned).* It is easy to notice that in the previous setup (P/R), the signals of u_2 and u_3 actually become part of the observational noise in filtering and predicting u_1. This is known as the representation error [53,100,150–154]. However, if the original observational noise level r^o is still used in updating the Kalman gain, then the filtering and prediction skill may be affected by the representation error. To resolve this issue, we utilize an inflated r_M^o in the analysis step to compute the Kalman gain while the other setups remain the same as in the P/R case. Here, the inflated r_M^o is given by

$$r_M^o = r^o + \text{var}(u_2) + \text{var}(u_3), \tag{109}$$

where $\text{var}(u_2)$ and $\text{var}(u_3)$ are the variance of u_2 and u_3 respectively at the statistical steady state. The inflation in (109) is the most straightforward one. More elaborate inflation techniques can be reached by applying the information theory in the training phase. Nevertheless, with such a simple inflation of the observational noise, the signals of u_2 and u_3 are treated as part of the observational noise. The estimation of the Kalman gain using the imperfect forecast model (108) is therefore expected to be improved.

Below, we consider two dynamical regimes with $\epsilon = 0.1$ and $\epsilon = 1$, respectively. The two variables u_2 and u_3 evolve in a much faster time scale than u_1 in the regime with $\epsilon = 0.1$ while the three variables lie in the same time scale with $\epsilon = 1$.

Now, we compare the filtering and prediction skill using the four setups as discussed above. In Figures 18 and 19, the skill as a function of the observational time step Δt_{obs} is shown in Regime $\epsilon = 0.1$. The following conclusions are reached. First, both the filtering and prediction skill overall deteriorates with the increase of the observational time step Δt_{obs}. Second, the filter estimates are almost always more accurate than the prediction estimates since the former contains extra information from observations. Third, the results with F/F is the best among all the four setups, as expected. Nevertheless, the filtering and prediction results of u_1 based on the other three setups remain comparable to that of F/F. However, the predictions of u_2 and u_3 using both the full and partial observations (F/F and P/F) contain a large error when the observational time step becomes large. Such an error is not reflected by the RMSE and PC but is clearly indicated by the relative entropy. In fact, since u_2 and u_3 both lie in faster time scales, their decorrelation times become much shorter than the observational time step when the latter increases. The consequence is that, regardless of the initial value, the prediction estimates always relax to the equilibrium mean and the amplitudes are thus severely weakened. Despite the success in capturing the pattern correlation, the prediction fails to catch any extreme events. On the other hand, the observations help the state estimation of filtering. In fact, the filter estimates with full observations (F/F) can almost perfectly capture the amplitudes of the truth while the partial observations (P/F) at least allow the filter estimates to reach some of the events with large amplitudes, which is nevertheless more skillful than the prediction. See Figures 20 and 21 for the true time series as well as the prediction and filtering estimates.

Next, in Figures 22 and 23, the filtering and prediction skill in Regime $\epsilon = 1$ is shown. Now, the difference in the results between using different setups becomes more significant. The filtering

and prediction skill of u_1 using P/F remains good but the gap compared with that using F/F is more obvious. Interestingly, the reduced strategy P/R now becomes much worse and the filtering results are even worse than the predictions especially with short observational time step Δt_{obs} (see also the time series in Figures 24 and 25). In fact, there are two sources of error that bring about such unskillful results. First, the variances of u_2 and u_3 with $\epsilon = 1$ are now much larger, which leads to a large representation error. Such representation error leads to a larger error in the filtering than prediction (For comparison, see Section 3.2 for the conclusion with no representation error). Second, recall that in Regime $\epsilon = 0.1$, u_2 and u_3 evolve in a much faster time scale and therefore they can be treated as noise. Ignoring them in the dynamics of u_1 provides a good approximation in (108). This is however not true in Regime $\epsilon = 1$ where the long memory of u_2 and u_3 plays an important role in the reduced dynamics of u_1. In other words, the reduced forecast model in (108) results in a large model error in Regime $\epsilon = 1$ due to the ignorance of u_2 and u_3. Thus, the combined effect from both the representation error and the imperfect model leads to the large error in filtering as well as prediction. With a short observational time step (for example, $\Delta t_{obs} = 0.2$ in Figure 24), the representation error becomes dominant. Therefore, an inflation of the observational noise to compensate the representation error (P/R tuned in Figure 23) improves the filtering and prediction skill. However, with a much longer observational time step, say $\Delta t_{obs} = 2$, an inflation of the observational noise (P/R tuned) will reduce the Kalman gain and the model rather than the observation provide more information to the filter results. When Δt_{obs} is large, the model will relax towards its equilibrium mean and thus the amplitude will be underestimated. See Figure 23. This again indicates the importance of using the relative entropy as one of the quantification criteria. Note that, with $\Delta t_{obs} = 2$, Figure 23 clearly states that both the RMSE and PC of the filtering estimates in P/R tuned setup are better than those in P/R setup, but the relative entropy in P/R tuned setup is much larger. This is a good example to show the importance and necessity of using the information-theoretic framework in quantifying the filtering and prediction skill instead of using only the path-wise RMSE and PC. Finally, we note that the signals of u_2 and u_3 here do not behave as a pair of oscillator as in the Regime with $\epsilon = 0.1$. This is because all the three variables now lie in the same time scale and they interact with each other. Here, the large-scale time-periodic forcing in u_1 leads to a time-periodic pattern in u_2 as well. However, the strong anti-correlation in u_1 and u_2 provides a cancelation in the feedback to u_3, which makes the signal of u_3 more noisy than u_2. The consequence is that the filtering and prediction skill in u_3 is much worse than those in u_2 due to the much larger noise to signal ratio in u_3.

To summarize, in the Regime with $\epsilon = 0.1$, all four of the setups lead to comparable results for both filtering and predicting the slow variable u_1. In particular, the most efficient strategy P/R works quite well. The filtering and prediction of the two fast variables u_2 and u_3 using F/F and P/F also show skillful results when the observational time step Δt_{obs} is short. When Δt_{obs} exceeds the decorrelation time of u_2 and u_3, the filter estimates tend to miss some large events while the prediction results fail to capture all the extreme events. In the Regime with $\epsilon = 1$, the reduced strategy (P/R) for u_1 does not work well especially with small observational time step Δt_{obs}. Nevertheless, if an observational noise inflation is adopted (P/R tuned), then both the filtering and prediction skill can be improved and becomes nearly comparable to those to the full filter with full observations (F/F) when Δt_{obs} is small to moderate. When the Δt_{obs} is large, the model error in the reduced forecast model (108) becomes dominant. In such a situation, only a full forecast model provides skillful prediction and filtering results while a partial observation (P/F) is allowed for retaining the skill. The partial observation (P/F) also gives a comparable skill as the full observation (F/F) in filtering and predicting u_2 but only in the setup with both the full forecast model and the full observations (F/F) leads to skillful results for u_3. A summary is shown in Table 1.

Table 1. Summary of the four setups in filtering the 3×3 system in (104). The four setups are: Full observations, full forecast model (F/F); partial observations, full forecast model (P/F); partial observations, reduced forecast model (P/R); and partial observations, reduced forecast model and tuned observational noise level with inflation (P/R tuned). Here, $\sqrt{}$ means the strategy works for small, moderate and moderately large Δ_{obs}. Small Δ_{obs} implies $\Delta_{obs} \leq 0.4$ which is roughly the decorrelation time of u_2 and u_3 in $\epsilon = 0.1$ regime. Moderate Δ_{obs} means $0.4 \leq \Delta_{obs} \leq 1.2$ and moderately large Δ_{obs} is up to $\Delta_{obs} \leq 2$, which is nevertheless below the decorrelation time of u_1 since u_1 has a slow-varying time-periodic forcing.

	F/F	P/F	P/R	P/R Tuned
$\epsilon = 0.1$				
Filter u_1	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$
Pred. u_1	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$
Filter u_2, u_3	small and moderate Δ_{obs}	small and moderate Δ_{obs}	N/A	N/A
Pred. u_2, u_3	small and moderate Δ_{obs}	small Δ_{obs}	N/A	N/A
$\epsilon = 1.0$				
Filter u_1	$\sqrt{}$	small to moderate Δ_{obs}	moderate Δ_{obs}	small to moderate Δ_{obs}
Pred. u_1	$\sqrt{}$	small to moderate Δ_{obs}	moderate Δ_{obs}	small to moderate Δ_{obs}
Filter u_2, u_3	small to moderate Δ_{obs}	small Δ_{obs} for u_2	N/A	N/A
Pred. u_2, u_3	small to moderate Δ_{obs} for u_2 and small Δ_{obs} for u_3	small Δ_{obs} for u_2	N/A	N/A
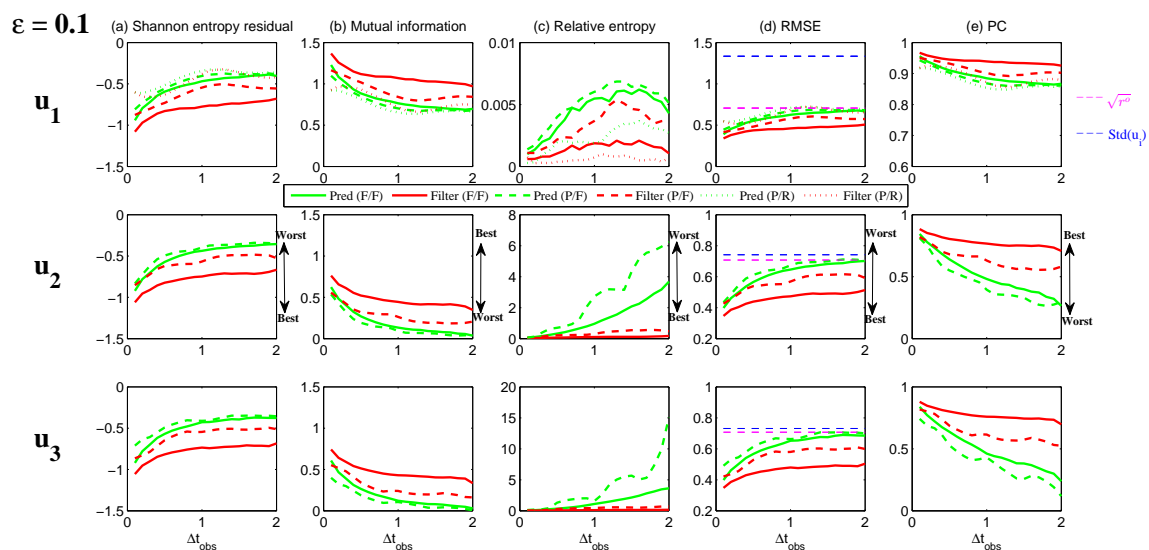
Figure 18. Regime $\epsilon = 0.1$. Prediction and filtering skill as a function of the observational time step Δt_{obs} using the three information measures: (**a**) Shannon entropy of residual, (**b**) mutual information and (**c**) relative entropy as well as the two traditional path-wise measures (**d**) root-mean-square error (RMSE) and (**e**) pattern correlation (PC). The green curves are for prediction and the red curves are for filtering. The solid curves correspond to the situation with full observations and full forecast model (F/F); the dashed curves correspond to the situation with partial observations and full forecast model (P/F); and the dotted curves are for that with partial observations and reduced forecast model (P/R). The three rows are shown for the skill of u_1, u_2 and u_3, respectively. The numerical simulation is based on time series with total length $T_{total} = 5000$ while the largest observational time step here is $\Delta t_{obs} = 2$.
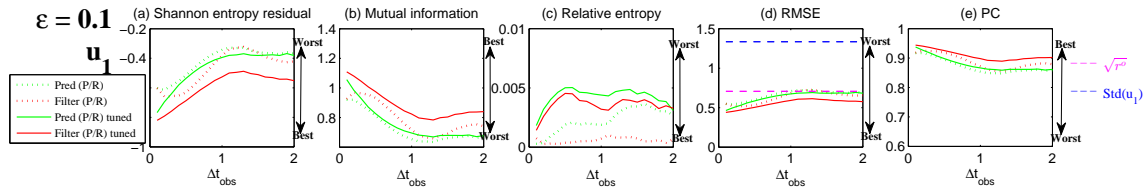
Figure 19. Regime $\epsilon = 0.1$. Similar to Figure 18 but the comparison of the skill of filtering and predicting u_1 based on the setup with partial observations and reduced forecast model (P/R) (dotted line) and that with partial observations, reduced forecast model and tuned observational noise level with inflation (P/R tuned) (thin solid line).
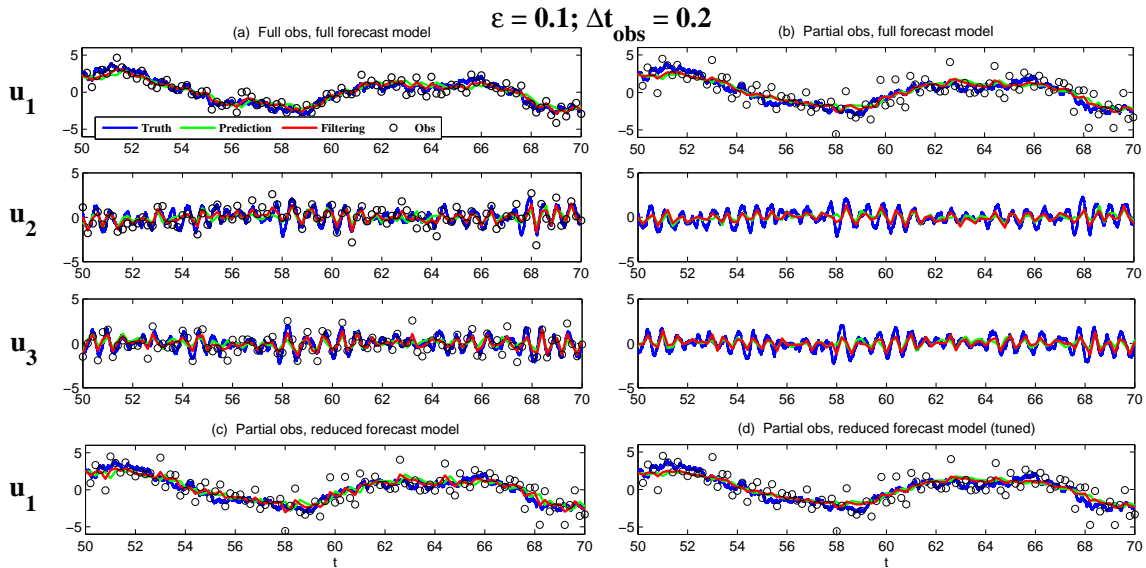
Figure 20. Regime $\epsilon = 0.1$ and $\Delta t_{obs} = 0.2$. Comparison of the filtering and prediction skill in different setups. (**a**): full observations and full forecast model (F/F); (**b**): partial observations and full forecast model (P/F); (**c**): partial observations and reduced forecast model (P/R); and (**d**): partial observations, reduced forecast model and tuned observational noise level (P/R tuned).
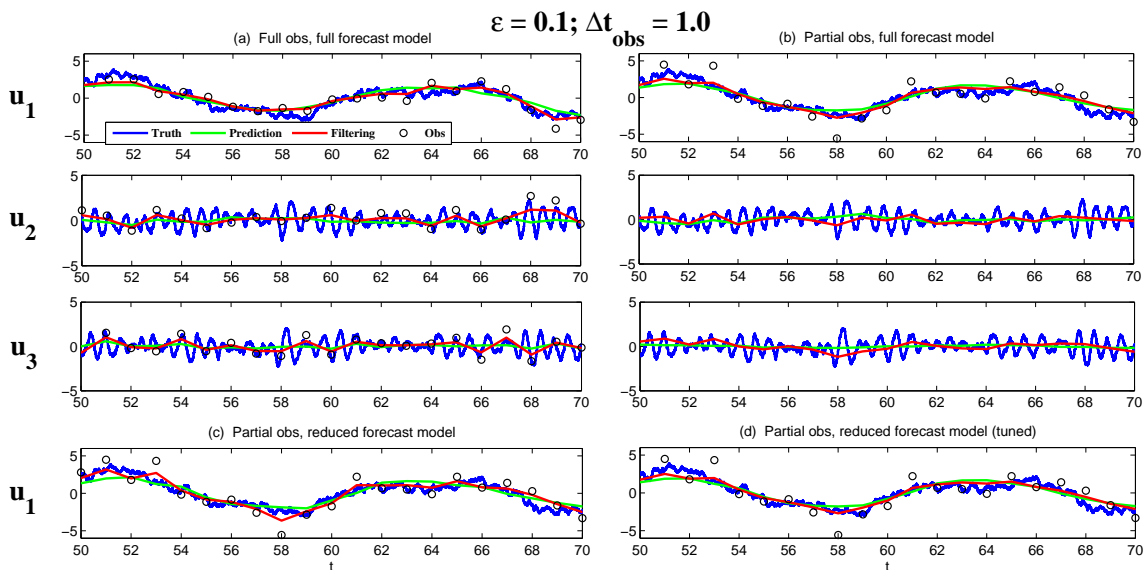
Figure 21. Similar to in Figure 20 but for Regime $\epsilon = 0.1$ and $\Delta t_{obs} = 1.0$.
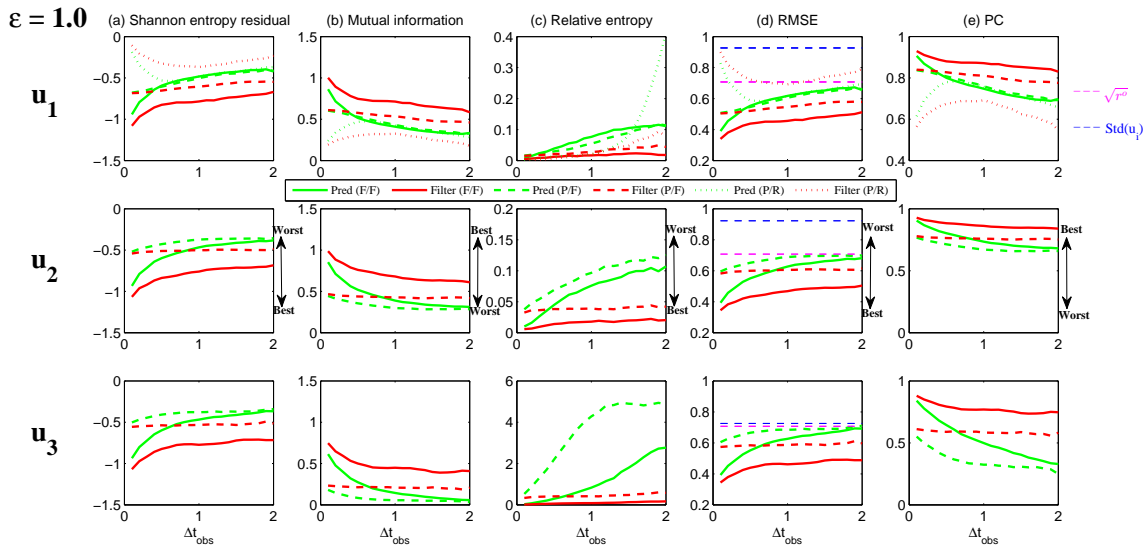
Figure 22. Similar to Figure 18 but for Regime $\epsilon = 1$.
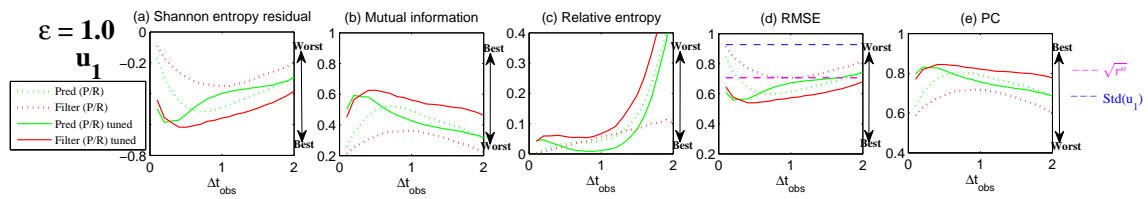
Figure 23. Similar to Figure 19 but for Regime $\epsilon = 1$.
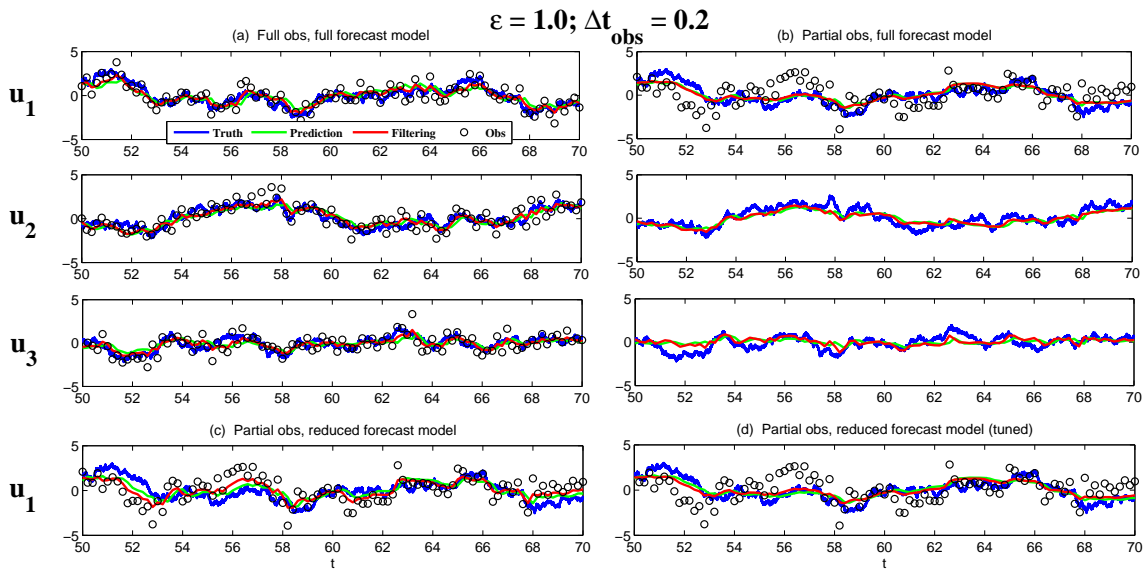
Figure 24. Similar to in Figure 20 but for Regime $\epsilon = 1.0$ and $\Delta t_{obs} = 0.2$.
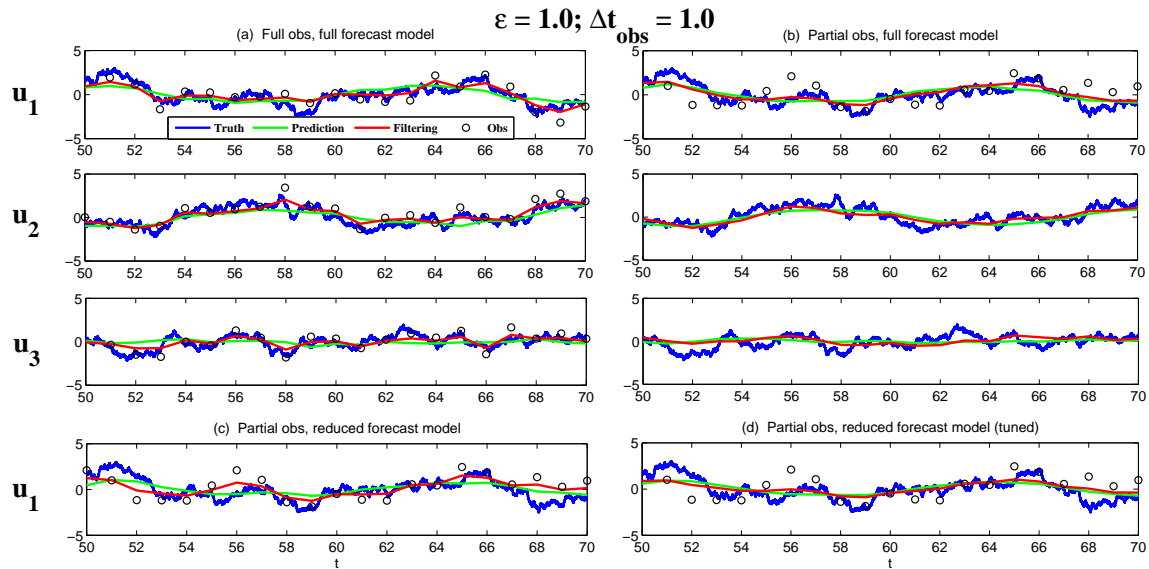
Figure 25. Similar to in Figure 20 but for Regime $\epsilon = 1.0$ and $\Delta t_{obs} = 1.0$.

3.5.2. Shallow Water Flows

Finally, let us study the filtering and prediction for spatially-extended systems. Consider the linearized two-dimensional rotating shallow water equation [8,143]

$$
\begin{aligned}
\frac{\partial \mathbf{u}}{\partial t} + \epsilon^{-1}\mathbf{u}^{\perp} &= -\epsilon^{-1}\nabla \eta, \\
\frac{\partial \eta}{\partial t} + \epsilon^{-1}\nabla \cdot \mathbf{u} &= 0,
\end{aligned}
\tag{110}
$$

where $\mathbf{u} = (u, v)^T$ is the two-dimensional velocity field and η is the geophysical height. Here, ϵ is the Rossby number representing the ratio between the Coriolis term and the advection term. We also set the Froude number equal to the Rossby number, which is the typical case in realistic geophysical flows [8]. Applying the Fourier decomposition method (See Section 4.4 in [8]) to (110), a 3×3 system is obtained for each Fourier wavenumber. In particular, associated with each Fourier wavenumber, there are:

1.   One geostrophically balanced (GB) mode with eigenvalue

$$
\omega_{\mathbf{k},B} = 0.
\tag{111}
$$

   The GB mode is incompressible.

2.   Two gravity modes with eigenvalues

$$
\omega_{\mathbf{k},\pm} = \pm \epsilon^{-1}\sqrt{|\mathbf{k}|^2 + 1}.
\tag{112}
$$

   The gravity modes are compressible.

Therefore, the solution of the shallow water equation in (110) can be written as a superposition of different Fourier modes,

$$
\begin{bmatrix} \mathbf{u}(\mathbf{x},t) \\ \eta(\mathbf{x},t) \end{bmatrix} = \sum_{\mathbf{k}\in\mathbf{K},\alpha\in\{B,\pm\}} \hat{u}_{\mathbf{k},\alpha}(t)\exp(i\mathbf{k}\cdot\mathbf{x})\mathbf{r}_{\mathbf{k},\alpha},
\tag{113}
$$

where the eigenvectors associated with the GB and gravity modes, i.e., $\mathbf{r}_{\mathbf{k},0}$ and $\mathbf{r}_{\mathbf{k},\pm}$, are given by

$$
\mathbf{r}_{\mathbf{k},B} = \frac{1}{\sqrt{|\mathbf{k}|^2+1}} \begin{pmatrix} -ik_2 \\ ik_1 \\ 1 \end{pmatrix}, \qquad \mathbf{r}_{\mathbf{k},\pm} = \frac{1}{|\mathbf{k}|^2\sqrt{2|\mathbf{k}|^2+2}} \begin{pmatrix} ik_2 \pm k_1\sqrt{|\mathbf{k}|^2+1} \\ -ik_1 \pm k_2\sqrt{|\mathbf{k}|^2+1} \\ |\mathbf{k}|^2 \end{pmatrix}, \tag{114}
$$

respectively, for $|\mathbf{k}| \neq 0$ and

$$
\mathbf{r}_{\mathbf{k},B} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{r}_{\mathbf{k},\pm} = \frac{1}{\sqrt{2}} \begin{pmatrix} \pm i \\ 1 \\ 0 \end{pmatrix}, \tag{115}
$$

respectively, for $|\mathbf{k}| = 0$. Here, in (111)–(115), $\mathbf{k} = (k_1, k_2)$ and $\mathbf{x} = (x, y)$.

The time evolution of the random Fourier amplitudes $\hat{u}_{\mathbf{k},\alpha}(t)$ associated with each Fourier wavenumber $\mathbf{k}$ can be described by the 3×3 system as introduced in (104)

$$
\begin{aligned}
\frac{d\hat{u}_{\mathbf{k},1}}{dt} &= -d_{\hat{u}_{\mathbf{k},1}}\hat{u}_{\mathbf{k},1} + L_{\mathbf{k},12}\hat{u}_{\mathbf{k},2} + L_{\mathbf{k},13}\hat{u}_{\mathbf{k},3} + F_{\mathbf{k},1}(t) + \sigma_{\mathbf{k},1}\dot{W}_{\mathbf{k},1}, \\
\frac{d\hat{u}_{\mathbf{k},2}}{dt} &= L_{\mathbf{k},21}\hat{u}_{\mathbf{k},1} - d_{\hat{u}_{\mathbf{k},2}}\hat{u}_{\mathbf{k},2} + \frac{L_{\mathbf{k},23}}{\epsilon}\hat{u}_{\mathbf{k},3} + \sigma_{\mathbf{k},2}\dot{W}_{\mathbf{k},2}, \\
\frac{d\hat{u}_{\mathbf{k},3}}{dt} &= L_{\mathbf{k},31}\hat{u}_{\mathbf{k},1} + \frac{L_{\mathbf{k},32}}{\epsilon}\hat{u}_{\mathbf{k},2} - d_{\hat{u}_{\mathbf{k},3}}\hat{u}_{\mathbf{k},3} + \sigma_{\mathbf{k},3}\dot{W}_{\mathbf{k},3}.
\end{aligned} \tag{116}
$$

Note that the variables $\hat{u}_{\mathbf{k},1}$, $\hat{u}_{\mathbf{k},2}$ and $\hat{u}_{\mathbf{k},3}$ are all real variables while the gravity modes are a pair of complex conjugate. Nevertheless, we can make use of a combination of $\hat{u}_{\mathbf{k},2}$ and $\hat{u}_{\mathbf{k},3}$ to form the two gravity waves:

$$
\begin{aligned}
\hat{u}_{\mathbf{k},+} &= \hat{u}_{\mathbf{k},2} + i\hat{u}_{\mathbf{k},3}, \\
\hat{u}_{\mathbf{k},-} &= \hat{u}_{\mathbf{k},3} + i\hat{u}_{\mathbf{k},2}.
\end{aligned} \tag{117}
$$

On the other hand, $\hat{u}_{\mathbf{k},1} = \hat{u}_{\mathbf{k},B}$. Without $L_{\mathbf{k},12}, L_{\mathbf{k},21}, L_{\mathbf{k},13}$ and $L_{\mathbf{k},31}$, these setups are similar to those in [139,155] except that the starting 3×3 systems in [139,155] are complex and there are two extra freedoms for noise in the pair of the gravity modes. In (116), the GB and gravity modes are coupled with each other linearly through nonzero coefficients $L_{\mathbf{k},12}, L_{\mathbf{k},21}, L_{\mathbf{k},13}$ and $L_{\mathbf{k},31}$.

Next, the noisy observations are given by the velocity fields u and v at each grid point in physical space. This is known as the Euler observations. Note that Lagrangian observations (via Lagrangian tracers) are also widely used in filtering the shallow water flows or more generally the geophysical flows [49,139,155–159]. Here, the Fourier expansion is applied to the noisy observational data of u and v. We assume the observational noise is white. Therefore, the noise level associated with each Fourier wavenumber is the same [108]. Note that the observations are not the Fourier coefficients in (116). They are the summation of the three Fourier components $\hat{u}_{\mathbf{k},\alpha}$ for $\alpha = \{B, \pm\}$ multiplying by the associated eigenvectors $\mathbf{r}_{\mathbf{k},\alpha}$ in (114) and (115) according to the expression of the velocity in (113). These correspond to the setups of P/F, P/R and P/R as discussed in Section 3.5. We will also report the filtering and prediction skill using the F/F as introduced in Section 3.5, which assumes that the observation for each GB and gravity mode is available. Although such a setup is idealized, it provides the optimal filtering and prediction results and can be used to examine the skill in the other setups. Once the results are obtained for each Fourier mode associated with the 3×3 system in (116), the summation of different Fourier modes are taken to recover the velocity field in the physical space. In practice, recovering and predicting the GB flow are of particular interest since GB flows lie in a longer time scale. Therefore, we focus on the study of the GB flow in different setups (F/F, P/F, P/R and P/R tuned). Since the GB modes are incompressible, it is more convenient to show the stream function ψ instead of the velocity field, where $(u, v) = (\partial\psi/\partial y, -\partial\psi/\partial x)$.

In the following, we consider the Fourier wavenumbers $\mathbf{k}$ in $[-2,2]^2$, where there are 25 GB modes and 50 gravity modes. The modes with $\mathbf{k} = (0,0)$ are the background modes, which are usually deterministic. Thus, we filter and predict the other 24 wavenumbers. Note that the mode $\mathbf{k}$ and $-\mathbf{k}$ are complex conjugate. The following parameters are taken in rotating shallow water Equation (110),

$$
\begin{aligned}
d_{\mathbf{k},u_1} = d_{\mathbf{k},u_2} = d_{\mathbf{k},u_3} = 1, \quad & \sigma_{\mathbf{k},1} = 3, \quad \sigma_{\mathbf{k},2} = \sigma_{\mathbf{k},3} = 2, \\
L_{\mathbf{k},12} = L_{\mathbf{k},13} = 1, \quad & L_{\mathbf{k},21} = L_{\mathbf{k},31} = -1, \\
L_{\mathbf{k},23} = \sqrt{|\mathbf{k}|^2 + 1}, \quad L_{\mathbf{k},32} = -\sqrt{|\mathbf{k}|^2 + 1}, \quad & F_{\mathbf{k},1} = 2\cos(0.5t), \quad F_{\mathbf{k},2} = F_{\mathbf{k},3} = 0.
\end{aligned}
\tag{118}
$$

Two dynamical regimes will be studied. They are $\epsilon = 0.1$ (fast rotation regime) and $\epsilon = 1.0$ (moderate rotation regime). The observational noise level is $r_{\mathbf{k}}^0 = 1.5$. The noise to signal ratio varies in different Fourier modes, but the noise is about 30% to 40% compared with the amplitude of the true signals multiplying by the eigenvectors (which is also the observational operator here) when the mode has observability [15,140,160,161]. The observability issue will be discussed at the end of this section.

The statistical behavior in filtering and predicting each Fourier wavenumber based on the information-theoretic framework are quite similar to those in Section 3.5.1. Therefore, in the following, we focus only on the comparison in the physical space, the results of which are given by taking the summation of different Fourier modes. In Figures 26 and 27, the prediction and filtering results in Regime $\epsilon = 0.1$ are shown. With a short observational time step Δt_{obs} (shorter than the decorrelation time of the gravity waves), both the filtering and prediction estimates are quite accurate, despite the fact that the prediction estimates in Figure 26 contain small errors in recovering the vortex in the right bottom corner. When Δt_{obs} is increased to $\Delta t_{obs} = 1$, which is longer than the memory time of the gravity waves, obvious errors are found in the predicted GB flows as shown in Figure 27. Nevertheless, the overall patterns and the amplitudes of the predicted GB flows in all the setups remain acceptable. The filtering estimates are more accurate than the predictions, especially in recovering the vortex near the left edge. On the other hand, in Regime $\epsilon = 1$, even with a short observational time step $\Delta t_{obs} = 0.1$, the prediction is inaccurate. See Figure 28. The error comes from both the pattern and the amplitude, the latter of which is quantified by the relative entropy. When Δt_{obs} becomes $\Delta t_{obs} = 1$, the filtering skills using the three practical setups (P/F, P/R and P/R tuned) all contain significant errors while the prediction estimates provide completely wrong patterns such as those at the right bottom corner. See Figure 29.

One interesting question to ask is that whether the observations of both the velocity fields u and v are needed in filtering and predicting the rotation shallow water flows since these two velocity components are strongly linked through the eigenvectors (114). To answer this question, we show the filtering and prediction estimates in physical space by observing both u and v (Panels (b) and (d)) and observing only u (Panels (c) and (e)). See the first two rows of Figure 30. Here, the fast rotation regime $\epsilon = 0.1$ is chosen and a short observational time step $\Delta t_{obs} = 0.1$ is adopted. It is clear that by observing only u, both the filtering and prediction estimates contain significant errors under the setups of both F/F and P/R (and others, not shown here). In fact, it is expected from Section 3.5 that with such a small Δt_{obs} and in the small ϵ regime, both the filtering and prediction results are accurate. This is true for most of the Fourier modes, such as $\mathbf{k} = (1,1)$ as shown in the last row of Figure 30. However, it is seen in the third row that the estimates of mode $\mathbf{k} = (1,0)$ for both filtering and predictions are quite different from the truth by observing only u. The reason is that the first component of $\mathbf{r}_{\mathbf{k},B}$ in (114) which multiplies $\hat{u}_{\mathbf{k},B}$ in obtaining the observation of u is zero for all modes with $\mathbf{k} = (k_1, 0)$. This means any mode $\hat{u}_{\mathbf{k},B}$ with $\mathbf{k} = (k_1, 0)$ has no observability. In other words, the observation u plays no role in the filtering process. The consequence is that both the filtering and prediction estimates of $\hat{u}_{\mathbf{k},B}$ follow exactly the mean evolution of the dynamics. This is clearly demonstrated in column (e) for P/R. On the other hand, the small fluctuations in the estimates of $\hat{u}_{\mathbf{k},B}$ in F/F are due to the coupling between $\hat{u}_{\mathbf{k},B}$ and $\hat{u}_{\mathbf{k},\pm}$ where the latter is observable. These findings indicate the importance and necessity of observing both u and v.

There are a few issues that have not been fully addressed here but can be good directions for future works. First, there might not be necessary to observe all the components of u and v. For example, observing v only for the modes that u has no observability may provide a cheaper strategy. Second, comparing the Euler and Lagrangian observations is an interesting topic. In fact, it has been shown in [49] that there exists an information barrier in recovering the velocity field using the Lagrangian observations. Whether this information barrier can be rigorously quantified by using Euler measurements and how to combine Euler and Lagrangian observations to maximize the information are both important topics that deserve further explorations. Finally, as has been noticed here, the P/R tuned setup does not significantly reduce the biases due to the representation error. Therefore, a more systematical study of understanding and improving the representation error is a good future direction.
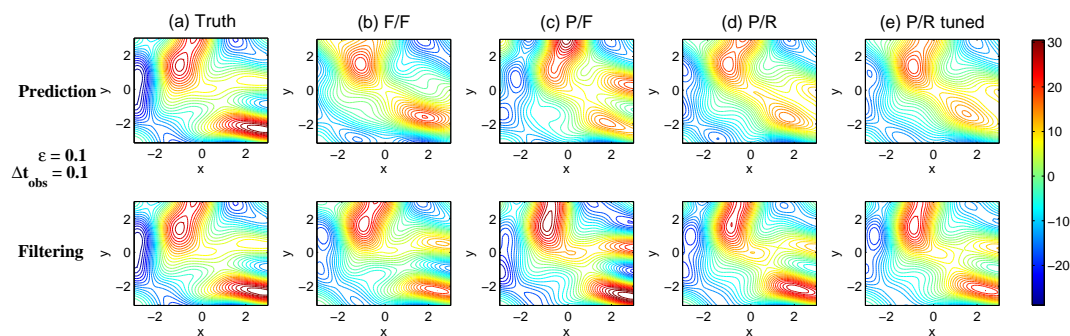
Figure 26. Regime $\epsilon = 0.1$ with observational time $\Delta t_{obs} = 0.1$. Comparison of the prediction and filtering estimates in physical space at a fixed time $t = 26$ using different setups: (**b**) F/F (full observations, full forecast model); (**c**) P/F (partial observations, full forecast model); (**d**) P/R (partial observations, reduced forecast model) and (**e**) P/R tuned (partial observations, reduced forecast model and tuned observational noise level with inflation). The truth is shown in column (**a**).
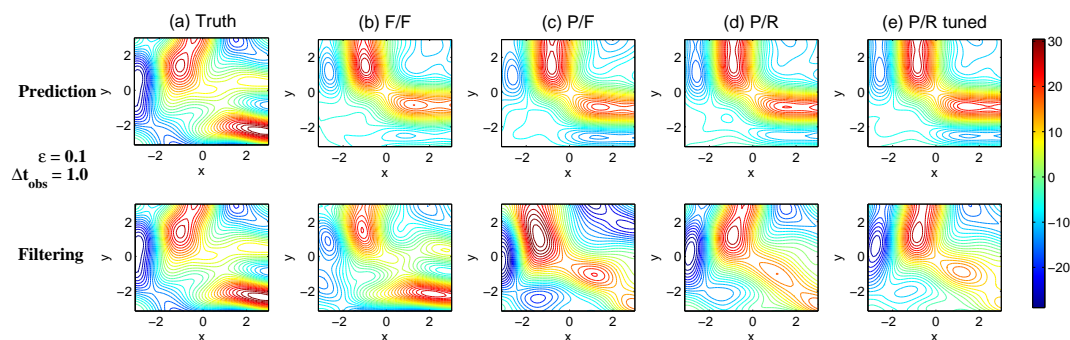
Figure 27. Regime $\epsilon = 0.1$ with observational time $\Delta t_{obs} = 1.0$. The caption is similar to that in Figure 26.
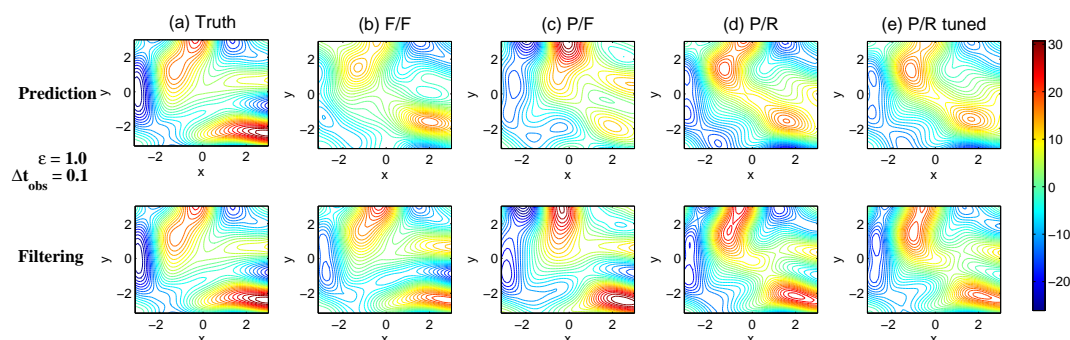
Figure 28. Regime $\epsilon = 1.0$ with short observational time $\Delta t_{obs} = 0.1$. The caption is similar to that in Figure 26.
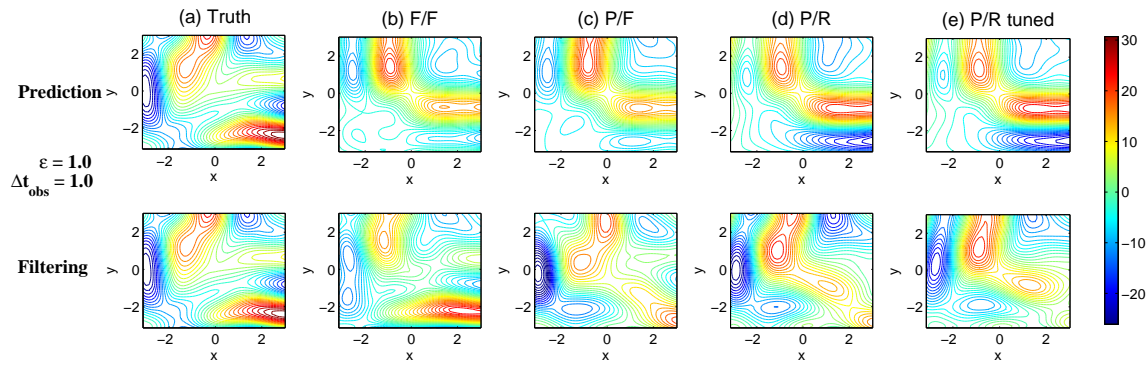
Figure 29. Regime $\epsilon = 1.0$ with observational time $\Delta t_{obs} = 1.0$. The caption is similar to that in Figure 26.
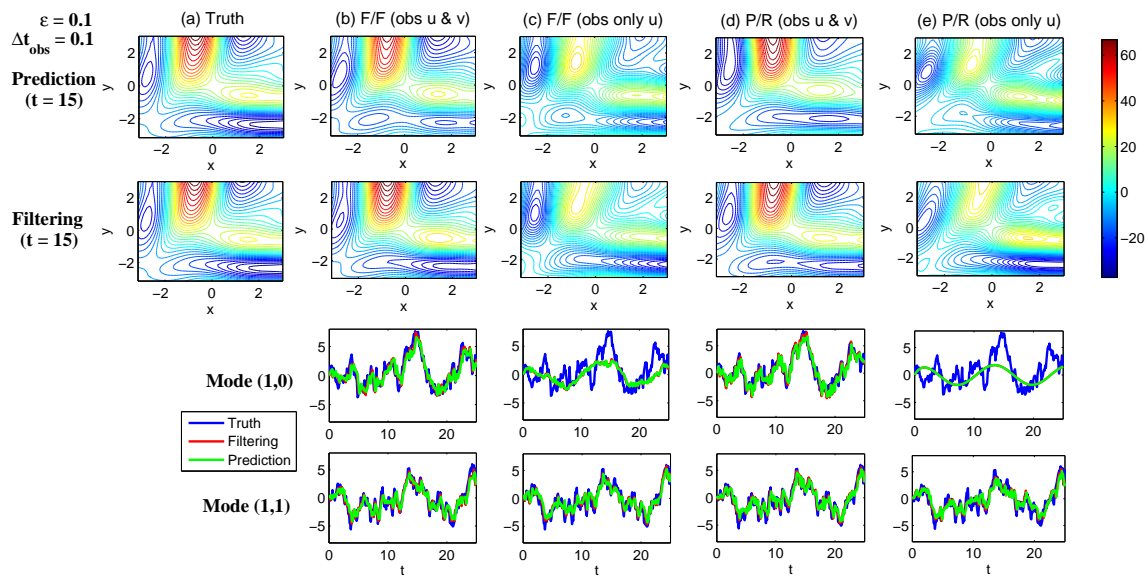
Figure 30. Comparison of the filtering and prediction estimates in physical space using F/F and P/R with two different types of observations. (**a**): the truth; (**b**,**d**): observing both u and v; (**c**,**e**): observing only u. Here, $\epsilon = 0.1$ and $\Delta t_{obs} = 0.1$. The first two rows show the snapshots of the steam function at a fixed time $t = 15$. The third and fourth rows show the time series of the two Fourier modes $\mathbf{k} = (1, 0)$ and $\mathbf{k} = (0, 1)$, where the filtering and prediction estimates are largely overlapped with each other. Note that the first component of vector $\vec{r}_{\mathbf{k},B}$ with $\mathbf{k} = (1, 0)$ is zero.

4. Information, Sensitivity and Linear Statistical Response—Fluctuation–Dissipation Theorem (FDT)

In Sections 2.3 and 2.4, we have shown the response in the statistical mean as a function of the external forcing perturbation in linear models, where analytic formulae were available and they were used to explicitly illustrate the response. For complex nonlinear dynamical systems, computing the system response due to different types of external perturbations is an important issue in many areas including climate change in climate science and feedback control in engineering. These external perturbations can be forcing (as in the examples shown in Sections 2.3 and 2.4), dissipation, phase as well as all other types of perturbations. In addition, the response function of interest is not only the statistical mean but also the energy (variance) and many other nonlinear functions of the state variables. Clearly, for most of the nonlinear systems, analytic formulae for the statistical response are not available and direct numerical methods are too expensive to adopt. Therefore, it is important to develop a general strategy of efficiently computing the system response to any external perturbation in complex nonlinear dynamical systems.

The fluctuation–dissipation theorem (FDT) [38–40,162] is an attractive way to assess the system response by using the statistics of the present states. For example, the important practical and

conceptual advantages for climate change science when a skillful FDT algorithm can be established is that the linear statistical response operator produced by FDT can be used directly for multiple climate change scenarios, multiple changes in forcing, dissipation and other parameters and inverse modelling directly [163,164] without the need of running the complex climate model in each individual case, often a computational problem of overwhelming complexity. With systematic approximations, FDT has been shown to have high skill for suitable regimes of general circulation models (GCMs), which are extremely complicated with an order of a million degrees of freedom [163,164].

4.1. Fluctuation–Dissipation Theorem (FDT)

4.1.1. The General Framework

Here, we summarize the general framework of the FDT [40]. Consider a general nonlinear dynamical system with noise

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}) + \sigma(\mathbf{u})\dot{\mathbf{W}}, \tag{119}$$

where $\mathbf{u} \in \mathbb{R}^N$ is the state variables, σ is an $N \times K$ noise matrix and $\dot{\mathbf{W}} \in \mathbb{R}^K$ is a *K*-dimensional white noise. The evolution of the PDF $p(\mathbf{u})$ associated with $\mathbf{u}$ is driven by the so-called Fokker–Planck equation [108],

$$\frac{\partial p}{\partial t} = -\mathrm{div}_{\mathbf{u}}[\mathbf{F}(\mathbf{u})p] + \frac{1}{2}\mathrm{div}_{\mathbf{u}}\nabla_{\mathbf{u}}(\Sigma p) \equiv L_{FP}p, \tag{120}$$

where $\Sigma = \sigma\sigma^T$ and $p|_{t=0} = p_0(\mathbf{u})$. Let $p_{eq}(\mathbf{u})$ be the smooth equilibrium PDF that satisfies $L_{FP}p_{eq} = 0$. The statistics of some function $A(\mathbf{u})$ are determined by

$$\langle A(\mathbf{u}) \rangle = \int A(\mathbf{u})p_{eq}(\mathbf{u})d\mathbf{u}. \tag{121}$$

Now, consider the dynamical in (119) by a small external forcing perturbation $\delta\mathbf{F}(\mathbf{u}, t)$. The perturbed system reads

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}) + \delta\mathbf{F}(\mathbf{u}, t) + \sigma(\mathbf{u})\dot{\mathbf{W}}. \tag{122}$$

We further assume an explicit time-separable structure for $\delta\mathbf{F}(\mathbf{u}, \mathbf{t})$, which occurs in many applications [40,97,165], namely

$$\delta\mathbf{F}(\mathbf{u}, t) = \delta\mathbf{w}(\mathbf{u})f(t). \tag{123}$$

Then, the Fokker–Planck equation associated with the perturbed system (122) is given by

$$\frac{\partial p^{\delta}}{\partial t} = L_{FP}p^{\delta} + \delta L_{ext}p^{\delta},$$
$$\text{where } \delta L_{ext}p^{\delta} = \mathbf{L}_{ext}p \cdot \delta\mathbf{F}(t), \qquad \mathbf{L}_{ext}p = -\frac{\partial}{\partial\mathbf{u}_i}(\mathbf{w}_i(\mathbf{u})p), \qquad 1 \leq i \leq N. \tag{124}$$

Similar to (121), for the perturbed system (124) the expected value of the nonlinear functional $A(\mathbf{u})$ is given by

$$\langle A(\mathbf{u}) \rangle^{\delta} = \int A(\mathbf{u})p^{\delta}(\mathbf{u})d\mathbf{u}. \tag{125}$$

The goal here is to calculate the change in the expected value

$$\delta\langle A(\mathbf{u}) \rangle = \langle A(\mathbf{u}) \rangle^{\delta} - \langle A(\mathbf{u}) \rangle. \tag{126}$$

To this end, let's take the difference between (120) and (124),

$$\frac{\partial}{\partial t}\delta p = L_{FP}\delta pp + \delta L_{ext}p_{eq} + \delta L_{ext}\delta p, \tag{127}$$

where $\delta p = p^\delta - p_{eq}$ is the small perturbation in the PDF. Ignoring the higher order term $\delta L_{ext}\delta p$ assuming δ is small, (127) reduces to

$$\frac{\partial}{\partial t}\delta p = L_{FP}\delta pp + \delta L_{ext}p_{eq},$$
$$\delta p|_{t=0} = 0. \tag{128}$$

Since L_{FP} is a linear operator, with the semigroup notation, $\exp[tL_{FP}]$, for this solution operator, the solution of (128) is written concisely as

$$\delta p = \int_0^T \exp\left[(t - t')L_{FP}\right]\left(\delta L_{ext}(t')p_{eq}\right)dt'. \tag{129}$$

Now, combining (129) with (124) and (126), we arrive at the *linear response formula*

$$\delta\langle A(\mathbf{u})\rangle(t) = \int_{\mathbb{R}^N} A(\mathbf{u})\delta p(\mathbf{u}, t)d\mathbf{u} = \int_0^t \mathbf{R}(t - t') \cdot \delta\mathbf{F}(t')dt', \tag{130}$$

where the vector linear response operator is given by

$$\mathbf{R}(t) = \int_{\mathbb{R}^N} A(\mathbf{u})\left(\exp[tL_{FP}][\mathbf{L}_{ext}p_{eq}]\right)(\mathbf{u})d\mathbf{u}. \tag{131}$$

This general calculation is the first step in the FDT. However, for nonlinear systems with many degrees of freedom, direct use of the formula in (131) is completely impractical because the exponential $\exp[tL_{FP}]$, cannot be calculated directly.

FDT states that, if δ is small enough, then the leading-order correction to the statistics in (121) becomes [40]

$$\delta\langle A(\mathbf{u})\rangle(t) = \int_0^t \mathbf{R}(t - s)\delta f(s)ds, \tag{132}$$

where $\mathbf{R}(t)$ is the linear response operator, which is calculated through correlation functions in the unperturbed climate:

$$\mathbf{R}(t) = \langle A[\mathbf{u}(t)]B[\mathbf{u}(0)]\rangle, \qquad B(\mathbf{u}) = -\frac{\mathrm{div}_\mathbf{u}(\mathbf{w}p_{eq})}{p_{eq}}. \tag{133}$$

See [40] for a rigorous proof of (132) and (133). Clearly, calculating the correlation functions in (133) via FDT is much cheaper and practical than directly computing the linear response operator (131).

Before we move on to the more specific FDT algorithms, let's comment on the perturbation function in (122) and (123). In fact, if $\mathbf{w}$ has no dependence on $\mathbf{u}$, then $\delta\mathbf{F}(t)$ naturally represents the forcing perturbation. If $\mathbf{w}(\mathbf{u})$ is a linear function of $\mathbf{u}$, then $\delta\mathbf{F}(\mathbf{u}, t)$ represents the perturbation in dissipation. It is also clear that if the functional $A(\mathbf{u})$ in (132) is given by $A(\mathbf{u}) = \mathbf{u}$, then the response computed is for the statistical mean. Likewise, $A(\mathbf{u}) = (\mathbf{u} - \bar{\mathbf{u}})^2$ is used for computing the response in the variance.

Notably, despite the small perturbation, FDT (132) and (133) does not require any linearization of the underlying dynamics in (119). Therefore, it captures the nonlinear features in the underlying turbulent systems.

4.1.2. Approximate FDT Methods

One major issue in applying FDT directly in the form of (133) is that the equilibrium measure $p_{eq}(\mathbf{u})$ is not known exactly. Therefore, different approximate methods have been proposed to compute the linear response operator.

Quasi-Gaussian (qG) FDT. Among all the approximate methods, the quasi-Gaussian (qG) approximation is one of the most effective approaches. It uses the approximate equilibrium measure

$$p_{eq}^G = C_N \exp\left[-\frac{1}{2}(\mathbf{u} - \bar{\mathbf{u}})^* R^{-1}(\mathbf{u} - \bar{\mathbf{u}})\right], \tag{134}$$

where the mean $\bar{\mathbf{u}}$ and covariance matrix R match those in the equilibrium p_{eq}. One then calculates

$$B^G(\mathbf{u}) = -\frac{\text{div}_{\mathbf{u}}(\mathbf{w}p_{eq}^G)}{p_{eq}^G} \tag{135}$$

and replaces $B(\mathbf{u})$ by $B^G(\mathbf{u})$ in the qG FDT. The correlation in (133) with this approximation is calculated by integrating the original system in (119) over a long trajectory or an ensemble of trajectories covering the attractor for shorter times assuming mixing and ergodicity for (133).

For the special case of changes in external forcing $\mathbf{w}(\mathbf{u})_i = \mathbf{e}_i$, $i \leq i \leq N$, the response operator for the qG FDT is given by the matrix

$$\mathbf{R}^G(t) = \langle A(\mathbf{u}(t))\mathbf{C}^{-1}(\mathbf{u} - \bar{\mathbf{u}})(0)\rangle. \tag{136}$$

The qG FDT will be applied in the simple example in Section 4.2.

Kicked FDT. One strategy to approximate the linear response operator which avoids direct evaluation of π_{eq} through the FDT formula is through the *kicked response* of an unperturbed system to a perturbation $\delta\mathbf{u}$ of the initial state from the equilibrium measure [30], that is,

$$\pi \mid_{t=0} = \pi_{\text{eq}}(\mathbf{u} - \delta\mathbf{u}) = \pi_{\text{eq}} - \delta\mathbf{u} \cdot \nabla\pi_{\text{eq}} + O\left(\delta^2\right). \tag{137}$$

One important advantage of adopting this kicked response strategy is that higher order statistics due to nonlinear dynamics will not be ignored (compared with other linearized strategy using only Gaussian statistics [162]). Then, the kicked response theory gives the following fact [28,40] for calculating the linear response operator:

Fact: For δ small enough, the linear response operator $\mathbf{R}(t)$ can be calculated by solving the unperturbed system (119) with a perturbed initial distribution in (137). Therefore, the linear response operator can be achieved through

$$\mathbf{R}(t) = \int A(\mathbf{u})\delta\pi + O\left(\delta^2\right). \tag{138}$$

Here, $\delta\pi$ is the resulting leading order expansion of the transient density function from unperturbed dynamics using initial value perturbation. The straightforward Monte Carlo algorithm to approximate (138) is sketched elsewhere [40,50]. The use of kicked FDT in calibrating the reduced-order models will be illustrated in Section 6.3.

4.2. Information Barrier for Linear Reduced Models in Capturing the Response in the Second Order Statistics

In this subsection, we use a simple 2D example to systematically illustrate the procedure of the FDT as introduced above. We also aim at showing the information barrier for linear reduced models in capturing the response beyond the first-order statistics. Note that such an information barrier was first pointed out in [41] with detailed discussions and more complicated examples.

The perfect model here is the SPEKF type of non-Gaussian model as discussed in (11), except that for simplicity we adopt a constant forcing f_u in the equation of u,

$$\begin{aligned}
\frac{du}{dt} &= -\gamma u + f_u + \sigma_u \dot{W}_u, \\
\frac{d\gamma}{dt} &= -d_\gamma (\gamma - \hat{\gamma}) + \sigma_\gamma \dot{W}_\gamma.
\end{aligned} \tag{139}$$

The following parameters are used in (139) in order to generate non-Gaussian statistics of u,

$$\sigma_u = 0.5, \qquad d_\gamma = 1.3, \qquad \sigma_\gamma = 1, \qquad \hat{\gamma} = 1, \qquad f_u = 1. \tag{140}$$

In Figure 31, sample trajectories and the associated PDFs of the SPEKF type non-Gaussian model (139) with parameters (140) are shown. Since γ frequently crosses zero and becomes negative, the corresponding signal of u is intermittent. Consequently, u has a skewed non-Gaussian PDF with a one-side fat tail.

With this constant forcing $f_u \equiv 1$, the time evolutions of the mean $\langle u \rangle$ and variance $\mathrm{Var}(u)$ of u are shown in Figure 32. For simplicity of the discussion below, the initial time here is set to be $t_0 = -12$. It is clear that after t reaches around $t = -6$, the model (139) arrives at the statistical equilibrium.

Now, we add a forcing perturbation $\delta f_u(t)$ to the model in (139),

$$\begin{aligned}
\frac{du}{dt} &= -\gamma u + f_u + \delta f_u(t) + \sigma_u \dot{W}_u, \\
\frac{d\gamma}{dt} &= -d_\gamma (\gamma - \hat{\gamma}) + \sigma_\gamma \dot{W}_\gamma.
\end{aligned} \tag{141}$$

The function $\delta f_u(t)$ is a ramp-type perturbation with the following form

$$\delta f_u(t) = A_0 \frac{\tanh(a(t - t_c)) + \tanh(a t_c)}{1 + \tanh(a t_c)}, \tag{142}$$

with

$$A_0 = 0.1, \qquad a = 1, \qquad t_c = 2. \tag{143}$$

The profile of $\delta f_u(t)$ is shown in panel (c) of Figure 32. The forcing perturbation $\delta f_u(t)$ starts from 0 at time $t = 0$ and it reaches 0.1 at roughly $t = 5$. After $t = 5$, $\delta f_u(t)$ stays at $\delta f_u(t) = 0.1$. Due to this forcing perturbation, the mean $\langle u \rangle$ and variance $\mathrm{Var}(u)$ also have corresponding changes, which are shown in panels (a) and (b) of Figure 32. Note that these responses are computed by using the analytical formulas of the time evolutions of the statistics, which are accurate. They are known as the idealized responses.

In most realistic scenarios, the true dynamics is unknown or it is too expensive to run the full perfect model. Therefore, simplified or reduced models are widely used in computing the responses. One type of the simple models that are widely adopted is the linear model,

$$du^M = -d_u^M u^M + f_u^M + \sigma_u^M \dot{W}. \tag{144}$$

Note that adopting such a linear model to compute the responses shares the same philosophy as one of the ad-hoc-FDT procedures [166], where linear regression approximate stochastic model [87] is used for the variables of interest before applying FDT.

The three parameters in (144) are calibrated by matching the equilibrium mean, equilibrium variance and decorrelation time with those of u in the perfect model (139), where

$$\langle u \rangle_{eq} = \frac{f_u^M}{d_u^M}, \qquad \mathrm{Var}(u)_{eq} = \frac{(\sigma_u^M)^2}{2 d_u^M}, \qquad \tau_{corr} = \frac{1}{d_u^M}. \tag{145}$$

Note that the autocorrelation function of u in (139) with parameters in (140) does not have a strong oscillation decaying structure, and therefore matching the decorrelation is sufficient for the calibration purpose here. With such calibrations, the linear model (144) automatically fit the unperturbed mean and variance at $t = 0$. Now, we add the same forcing perturbation to the linear model,

$$du^M = -d_u^M u^M + f_u^M + \delta f_u(t) + \sigma_u^M \dot{W}. \tag{146}$$

Since the statistics in the linear model is Gaussian, the formulas in (134)–(136) become rigorous with no approximation. In computing the responses to the forcing perturbation $\delta f_u(t)$ in the mean and variance of u, the functional $A(\mathbf{u}(t))$ is set to be

$$
\begin{aligned}
\text{Response in the mean :} \quad & A(\mathbf{u}(t)) = u^M, \\
\text{Response in the variance :} \quad & A(\mathbf{u}(t)) = (u^M - \bar{u}^M)^2,
\end{aligned}
\tag{147}
$$

respectively. These responses using such a linear model are shown in Figure 33 (green colors) while the idealized responses are shown in blue for reference. It is clear that the response in the mean using the linear model captures the trend of the truth, but the amplitude is severely overestimated. On the other hand, the response in the variance using the linear model is identically zero and therefore it completely misses the truth. In fact, inserting the second Equation (147) into (136) yields solving a third-order centered moment. However, all odd-centered moments automatically vanish for Gaussian distribution and therefore the response in the variance using the linear model is zero [41], which has already been mentioned in Sections 2.2.1 and 2.3. These results unambiguously indicate the insufficiency of using linear approximate models as well as the ad hoc-FDT [166] to compute the responses when the underlying dynamics is highly nonlinear.

As a comparison, we also show the responses using the qG FDT based on the perfect model (139). Since the forcing perturbation is only on the direction of u, $\mathbf{w}(\mathbf{u})$ in (135) is given by $\mathbf{w}(\mathbf{u}) = [1,0]^T$. In Figure 33, it is clear that the qG FDT based on the perfect model (red) captures the response in the mean quite accurately. In addition, this qG FDT also results in a response in the variance and the skill in recovering the time evolution of the variance response is pretty good. Notably, although the response operator $\mathbf{R}(t)$ in (132) is linear and the Gaussian approximation (134) is used in computing the equilibrium PDF of the unperturbed system, the underlying nonlinear dynamics is used in computing the functional $A(\mathbf{u}(t))$ in (147). Therefore, the nonlinear interaction is included in the FDT and the response in the variance is captured to a large extent. It is of importance to keep in mind that FDT does not implement linearization on the original underlying nonlinear system. Thus, the nonlinear dynamical features are reflected in the FDT. The linearization is applied only in the response operator due to small perturbations.

Although the simple test example here deals with a constant forcing in the unperturbed system, the FDT technique can be easily generalized to the systems with time-periodic settings, which usually corresponds to annual or seasonal cycles in climate, atmosphere and ocean sciences. Mathematical theories of the generalizations of FDT to time-dependent ensembles can be found in [162]. In [167], a triad nonlinear stochastic model with time-periodic setting was developed, which mimics the nonlinear interaction of two Rossby waves forced by baroclinic processes with a zonal jet forced by a polar temperature gradient. Systematical studies showed that qG FDT has surprisingly high skill for the mean response to the changes in forcing. The performance of qG FDT for the variance response to the perturbations of dissipation is good in the nearly Gaussian regime and deteriorates in the strongly non-Gaussian regime. More examples can be found in [15,40].
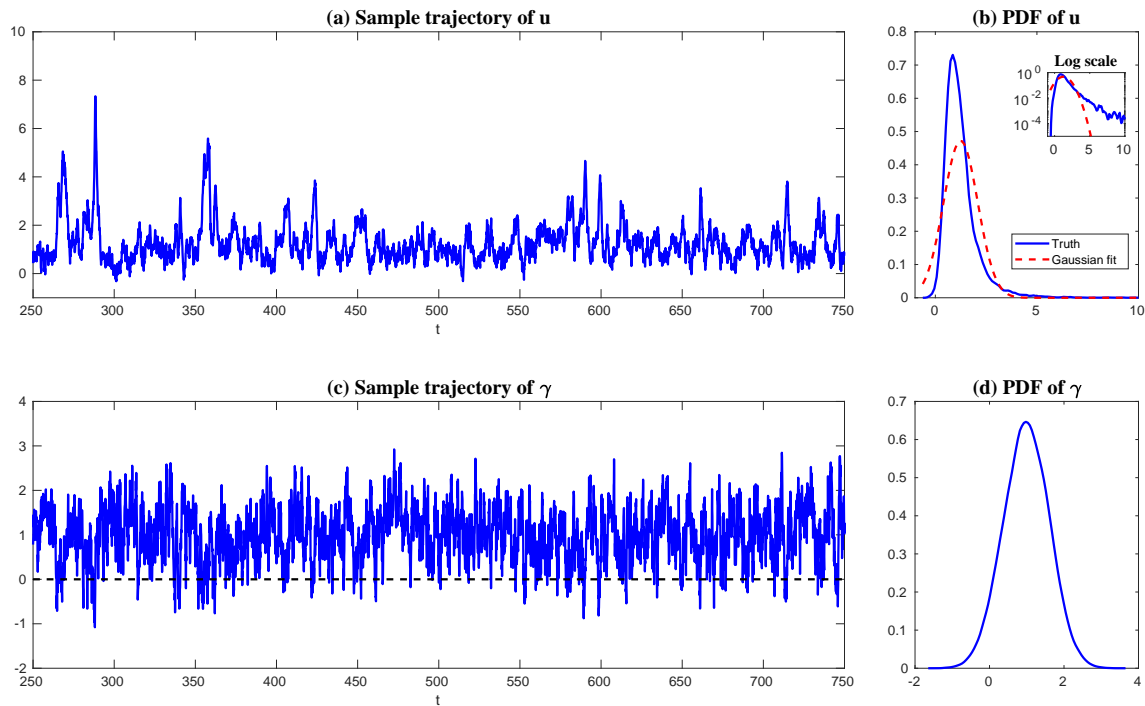
Figure 31. (**a**,**b**): Sample trajectories of u and γ in the stochastic parameterized extended Kalman filter (SPEKF) type of non-Gaussian model (139); (**c**,**d**): the corresponding PDFs. The subpanel within (**b**) shows the PDF in logarithm scale, with the red curves representing the Gaussian fit. The parameters associated with these figures are given in (140).
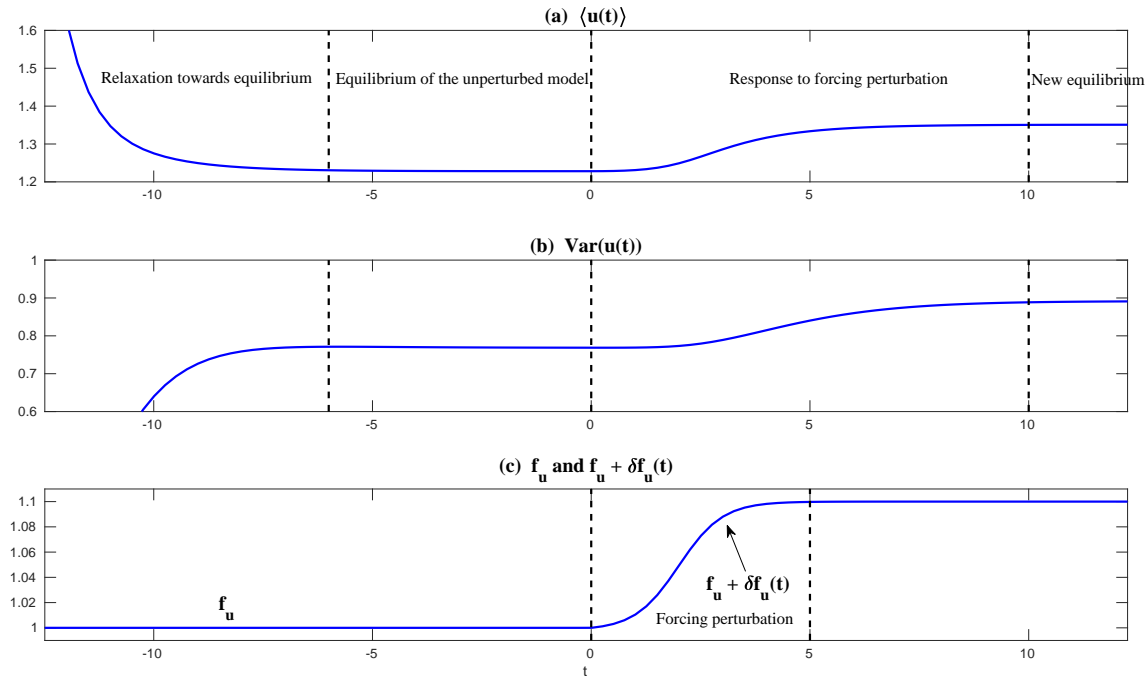
Figure 32. Time evolution of the mean $\langle u \rangle$ and variance $\text{Var}(u(t))$ of u (**a**,**b**) and the corresponding forcing $f_u(t)$ (panel (**c**)). The forcing $f_u(t)$ is perturbed at time $t = 0$ with $\delta f(t)$ given in (142). The mean and variance of u have corresponding responses and eventually arrive at a new equilibrium.
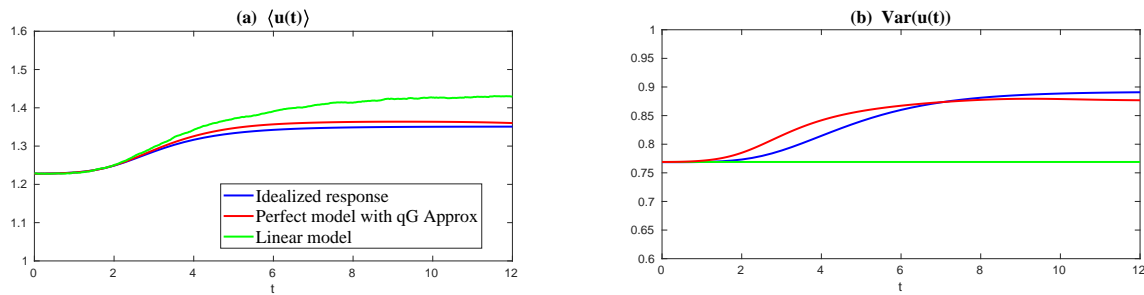
Figure 33. Responses to the mean $\langle u \rangle$ (panel (**a**)) and variance $\text{Var}(u(t))$ (panel (**b**)) of u with the forcing perturbation $\delta f(t)$ given in (142). The perturbation starts at time $t = 0$, which is consistent with that in Figure 32.

Other FDT techniques that have skillful performance in dealing with complex nonlinear dynamical systems includes blended response algorithms [168,169] and kicked FDT [30]. FDT has been demonstrated to have high skill for the mean and variance response in the upper troposphere for changes in tropical heating in a prototype atmospheric GCM and can be utilized for complex multiple forcing and inverse modeling issues of interest in climate change science [163,164]. Note that GCMs usually have a huge number of state variables and applying FDT on the entire phase space is impractical due to the limitations in calculating the covariance matrix. Practical strategies involve computing the response operator on a reduced subspace. Mathematical principles of applying FDT on reduced subspaces can be found in [41].

4.3. Information Theory for Finding the Most Sensitive Change Directions

An important question in climate change is how to find the most sensitive directions for climate change given the present climate. To quantify these most sensitive directions, consider a family of parameters $\lambda \in \mathbb{R}^p$ with π_λ the PDF of the true climate as a function of λ. Here $\lambda = 0$ corresponds to the unperturbed state or the present climate π. Note that λ can consist of external parameters such as changes in forcing or parameters of internal variability such as a change in dissipation. In light of the information theoretic framework, the most sensitive perturbed climate is the one with the largest uncertainty related to the unperturbed one,

$$\mathcal{P}(\pi_{\lambda_*}, \pi) = \max_{\lambda \in \mathbb{R}^p} \mathcal{P}(\pi_\lambda, \pi). \tag{148}$$

The calculation of the most sensitive perturbation for the present climate in (148) is through the information theoretical framework. Assume that π_λ is differentiable with respect to the parameter λ [90,162,170]. Since $\pi_\lambda|_{\lambda=0} = \pi$, for small values of λ, we have

$$\mathcal{P}(\pi_\lambda, \pi) = \lambda \cdot I(\pi)\lambda + O(|\lambda|^3), \tag{149}$$

where $\lambda \cdot I(\pi)\lambda$ is the quadratic form in λ given by the Fisher information [40,93,162,171]

$$\lambda \cdot I(\pi)\lambda = \int \frac{(\lambda \cdot \nabla_\lambda \pi)^2}{\pi}, \tag{150}$$

and the elements of the matrix of this quadratic form are given by

$$I_{kj}(\pi) = \int \frac{\frac{\partial \pi}{\partial \lambda_k} \frac{\partial \pi}{\partial \lambda_j}}{\pi}. \tag{151}$$

Detailed derivations of (149)–(151) are included in Appendix A. Note that the gradients are calculated at the unperturbed state $\lambda = 0$. Therefore, if both the unperturbed state π and the gradients

$\boldsymbol{\lambda} \cdot \nabla_{\lambda} \pi$ are known, then the most sensitive perturbation direction occurs along the unit direction $\mathbf{e}_{\pi}^{*} \in \mathbb{R}^{p}$ which is associated with the largest eigenvalue λ_{π}^{*} of the quadratic form in (150).

Below, we use two simple examples to provide insights of the above information theoretical framework in finding the most sensitive change direction in the underlying models. We will start with a linear example, where all the results using the direct calculation method can be written down explicitly. We aim at comparing the results using the direct method and using the Fisher information in (150). The analytic formulae associated with this linear example also allow us to understand the contributions of the uncertainty in the perturbation from the signal and the dispersion parts, respectively. Then, we will use a more complicated nonlinear example with non-Gaussian noise to show the efficiency and accuracy of using the information criterion in (150).

The first example is an one-dimensional linear model,

$$\frac{du}{dt} = -au + f + \sigma \dot{W}, \tag{152}$$

the equilibrium PDF of which is Gaussian and is given by $\mathcal{N}(\bar{u}, C)$,

$$\pi(u) = N_C \exp\left(-\frac{(u - \bar{u})^2}{2C}\right), \tag{153}$$

with

$$\bar{u} = \frac{f}{a}, \qquad C = \frac{\sigma^2}{2a}. \tag{154}$$

The two-dimensional parameters $\boldsymbol{\lambda} = (f, a)^T \in R^2$ for external forcing and dissipation are the natural parameters which are varied in this model. Therefore, the corresponding $I(\boldsymbol{\lambda})$ in (150) is a 2×2 matrix with entries $I_{ij}, i, j = 1, 2$. Using (153), it is straightforward to compute the first-order derivatives of π with respect to f and a,

$$\begin{aligned}
\frac{\partial \pi}{\partial f} &= \frac{u - \bar{u}}{aC}\pi, \\
\frac{\partial \pi}{\partial a} &= \frac{\sigma^2}{4a^2 C}\pi - \frac{f(u - \bar{u})}{a^2 C}\pi - \frac{\sigma^2(u - \bar{u})^2}{4a^2 C^2}\pi.
\end{aligned} \tag{155}$$

In light of (151) and (155), the four elements of I have the following explicit expressions:

$$\begin{aligned}
I_{11} &= \int \frac{\left(\frac{\partial \pi}{\partial f}\right)^2}{\pi} du = \int \frac{(u - \bar{u})^2}{C^2}\frac{\pi}{a^2} du = \frac{1}{Ca^2}, \\
I_{12} &= I_{21} = \int \frac{\frac{\partial \pi}{\partial f}\frac{\partial \pi}{\partial a}}{\pi} du = \int \frac{u - \bar{u}}{aC}\left(\frac{\sigma^2}{4a^2 C} - \frac{f(u - \bar{u})}{a^2 C} - \frac{\sigma^2(u - \bar{u})^2}{4a^2 C^2}\right)\pi du = -\frac{f}{a^3 C}, \\
I_{22} &= \int \frac{\left(\frac{\partial \pi}{\partial a}\right)^2}{\pi} = \int \left(\frac{\sigma^2}{4a^2 C} - \frac{f(u - \bar{u})}{a^2 C} - \frac{\sigma^2(u - \bar{u})^2}{4a^2 C^2}\right)^2 \pi du \\
&= \int \left[\left(\frac{\sigma^2}{4a^2 C}\right)^2 + \left(\frac{f(u - \bar{u})}{a^2 C}\right)^2 + \left(\frac{\sigma^2(u - \bar{u})^2}{4a^2 C^2}\right)^2 - 2\frac{\sigma^2}{4a^2 C}\frac{\sigma^2(u - \bar{u})^2}{4a^2 C^2}\right]\pi du \\
&= -\frac{f^2}{Ca^4} + \frac{\sigma^4}{8C^2 a^4}.
\end{aligned} \tag{156}$$

Now, let's implement numerical experiments. The following two groups of parameters are used:

$$\begin{aligned}
(a): &\quad a = 1, \quad f = 1, \quad \sigma = 1, \\
(b): &\quad a = 1, \quad f = 1, \quad \sigma = 3.
\end{aligned} \tag{157}$$

Since I is a 2×2 matrix, there are only two eigenmodes. The eigenvector $\mathbf{w}$ associated with the larger eigenvalue corresponds to the most sensitive direction with respect to the perturbation $(\delta f, \delta a)^T$.

By plugging the model parameters (157) into the I matrix in (156), we find the most sensitive direction in both of the cases:

$$(a): \quad \mathbf{e}_\pi^* = \begin{pmatrix} -0.6618 \\ 0.7497 \end{pmatrix}, \qquad (b): \quad \mathbf{e}_\pi^* = \begin{pmatrix} -0.3554 \\ 0.9347 \end{pmatrix}. \tag{158}$$

To gain more intuition on the results of these most sensitive directions, we make use of the simple structure of (152) to solve this problem in an alternative way. In fact, given small perturbations $(\delta f, \delta a)^T$ to $(f, a)^T$, the corresponding perturbed mean and variance can be written down explicitly

$$\bar{u}^\delta = \frac{f + \delta f}{a + \delta a}, \qquad C^\delta = \frac{\sigma^2}{2(a + \delta a)}. \tag{159}$$

Since both the unperturbed and perturbed PDFs are Gaussian, we can easily make use of the explicit formula of the relative entropy in (6) to compute the uncertainty due to the perturbation $\mathcal{P}(\pi, \pi^\delta)$ and find the most sensitive direction in the two-dimensional parameter space. Recall in (6) that the total uncertainty can be decomposed into signal and dispersion parts. Making use of (154) and (159), we have

$$
\begin{aligned}
\text{Signal} &= \frac{1}{2} \left(\frac{f}{a} - \frac{f + \delta f}{a + \delta a} \right)^2 \left(\frac{\sigma^2}{2a} \right)^{-1} = \frac{1}{2} \frac{(fa + f\delta a - fa - a\delta f)^2}{a^2(a + \delta a)^2} \frac{2a}{\sigma^2} \\
&= \frac{(f\delta a - a\delta f)^2}{a\sigma^2} + o\left(\delta a^3\right) + o\left(\delta a^2 \delta f\right) + o\left(\delta a \delta f^2\right) \\
\text{Dispersion} &= -\frac{1}{2} \ln \left(\frac{a + \delta a}{a} \right) + \frac{1}{2} \left(\frac{a + \delta a}{a} - 1 \right) = -\frac{1}{2} \ln \left(1 + \frac{\delta a}{a} \right) + \frac{1}{2} \frac{\delta a}{a} \\
&= -\frac{1}{2} \left(\frac{\delta a}{a} - \frac{1}{2} \left(\frac{\delta a}{a} \right)^2 + o\left(\frac{\delta a}{a} \right)^3 \right) + \frac{1}{2} \frac{\delta a}{a} \\
&= \frac{1}{4} \left(\frac{\delta a}{a} \right)^2 + o\left(\frac{\delta a}{a} \right)^3.
\end{aligned}
\tag{160}
$$

Note that the dispersion part depends only on the perturbation in the dissipation δa since f has no effect on the variance. In addition, it is clear that δa and δf should have opposite signs in order to maximize the relative entropy in the signal part.

Figure 34 shows the total relative entropy as well as its two components, namely signal and dispersion, as a function of the perturbations in the two-dimensional parameter space $(\delta f, \delta a)^T$ using the direct formula (160). The numerical simulation here assumes $\sqrt{\delta f^2 + \delta a^2} \leq 0.05$ to guarantee the perturbation is small enough. In both cases, the most sensible direction with respect to only the dispersion part lies in the direction $(\delta a, \delta f)^T = (1, 0)^T$, due to the fact that δf has no effect on the dispersion part. In the signal part, the most sensitive direction satisfies $a\delta f = -f\delta a$. The overall most sensitive direction depends naturally on the weights of signal and dispersion parts. When σ becomes larger, the weight on the signal part reduces since the signal part is proportional to the inverse of the model variance. It is easy to see that the most sensitive directions as indicated by the black dashed lines in Figure 34 are consistent with the theoretical prediction from (158) using the Fisher information (148)–(151).
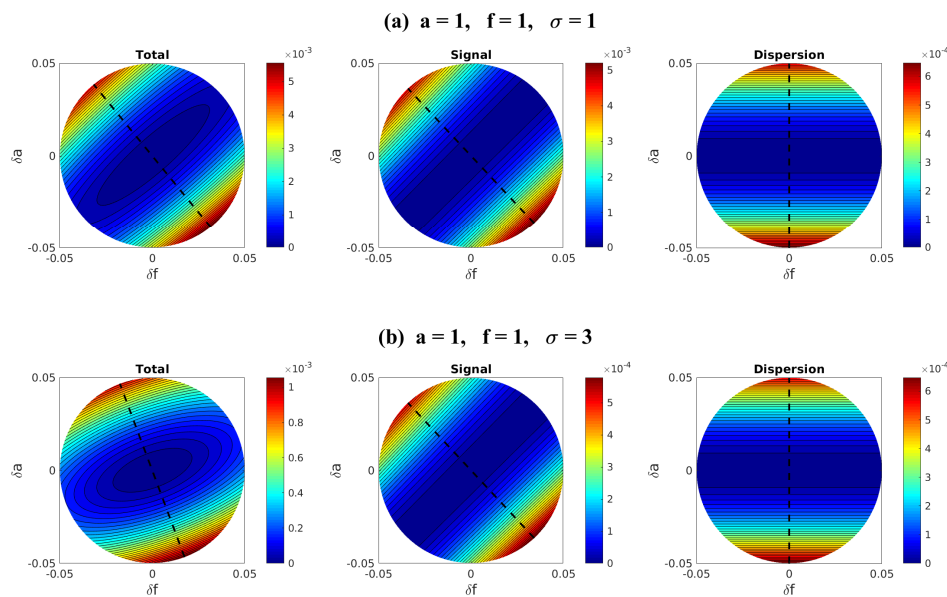
(a) a = 1, f = 1, σ = 1

(b) a = 1, f = 1, σ = 3

Figure 34. Uncertainty dependence of the perturbation in the two-dimensional parameter space $(\delta a, \delta f)^T$ in the linear model (152). (**a**): $a = 1, f = 1, \sigma = 1$; (**b**): $a = 1, f = 1, \sigma = 3$. The total uncertainty (**left column**) is decomposed into signal (**middle column**) and dispersion (**right column**) parts according to (6). The black dashed line shows the direction of the maximum uncertainty, namely the most sensitive direction of perturbation.

Now, we consider a second example with a nonlinear model [116,170],

$$\frac{du}{dt} = (f + au + bu^2 - cu^3) + (A - Bu)\dot{W}_C + \sigma \dot{W}_A. \tag{161}$$

The nonlinear model in (161) is a canonical model for low frequency atmospheric variability and was derived based on stochastic mode reduction strategies. This one-dimensional, normal form was applied in a regression strategy in [116] for data from a prototype AOS model [112] to build one-dimensional stochastic models for low-frequency patterns such as the North Atlantic Oscillation (NAO) and the leading principal component (PC-1) that has features of the Arctic Oscillation. Note that the model in (161) has both correlated additive and multiplicative noise $(A - Bu)\dot{W}_C$ as well as an extra uncorrelated additive noise $\sigma \dot{W}_A$. The nonlinearity interacting with noise allows a rich dynamical features in the model such as strongly non-Gaussian PDFs and multiple attractors. Different from the previous example with linear dynamics, the direct method has no explicit solution for the nonlinear system (161). The goal here is to find the most sensitive directions using the information theory developed above in different dynamical regimes.

Here, we consider a simple case with $A = B = 0$ such that the model has only additive noise. Nevertheless, the cubic nonlinearity still allows the model to have strong non-Gaussian characteristics. With $A = B = 0$, the equilibrium PDF of (161) is given by the following explicit formula

$$\pi(u) = N_0 \exp\left(\frac{2}{\sigma^2} \left(fu + \frac{a}{2}u^2 + \frac{b}{3}u^3 - \frac{c}{4}u^4 \right) \right). \tag{162}$$

We again look at the perturbation in the two-dimensional parameter space $\lambda = (f, a)^T$, which represent the changes in forcing and damping. Following (149)–(151), we aim at solving the eigenvectors of the 2×2 matrix $I(\lambda)$. To explicitly write down the elements in $I(\lambda)$, we define

$$H_k = \int u^k \psi(u)du, \quad k \geq 0 \quad \text{with} \quad \psi(u) = \exp\left(\frac{2}{\sigma^2} \left(fu + \frac{a}{2}u^2 + \frac{b}{3}u^3 - \frac{c}{4}u^4 \right) \right). \tag{163}$$

Straightforward calculations show that

$$
I_{11} = \int \frac{(\frac{\partial \pi}{\partial f})^2}{\pi} du = \frac{4}{\sigma^4 H_0^4}(H_0 H_2 - H_1^2),
$$

$$
I_{12} = I_{21} = \int \frac{\frac{\partial \pi}{\partial f}\frac{\partial \pi}{\partial a}}{\pi} du = \frac{2}{\sigma^4 H_0^4}(H_0 H_3 - H_1 H_2), \tag{164}
$$

$$
I_{22} = \int \frac{(\frac{\partial \pi}{\partial a})^2}{\pi} du = \frac{1}{\sigma^4 H_0^4}(H_0 H_2 - H_2^2).
$$

Now, we focus on the case studies in the following three regimes,

$$
\begin{aligned}
\text{Regime I}: & \quad f = 1.8, & a = 0, & \quad b = -5.4, & c = 4, & \quad \sigma = \sqrt{0.5}, \\
\text{Regime II}: & \quad f = -0.005, & a = -0.018, & \quad b = 0.006, & c = 0.003, & \quad \sigma = 0.226, \\
\text{Regime III}: & \quad f = -1.44, & a = -0.55, & \quad b = -0.073, & c = 0.003, & \quad \sigma = 0.253.
\end{aligned} \tag{165}
$$

The sample trajectories and equilibrium PDFs associated with these regimes are shown in Figure 35.

The PDF in Regime I is unimodal with skewness and an one-side fat tail. Interestingly, the time series in Regime I shows a distinct regimes of behavior [172,173]. Regimes II and III correspond to PC-1 and NAO for the low frequency data as discussed above, where Regime II has a slight skewed PDF with sub-Gaussian tails while Regime III is nearly Gaussian.

With the parameters in (165) and the explicit expression of $I(\lambda)$ in (164), the most sensitive direction of the parameter perturbation in the two-dimensional space $(\delta f, \delta a)^T$ is given by respectively

$$
\begin{aligned}
\text{Regime I}: & \quad \mathbf{e}_\pi^* = (0.9545, 0.2981)^T, \\
\text{Regime II}: & \quad \mathbf{e}_\pi^* = (0.9685, 0.2488)^T, \\
\text{Regime III}: & \quad \mathbf{e}_\pi^* = (-0.0760, 0.9971)^T.
\end{aligned} \tag{166}
$$

The results in (166) imply that the forcing perturbation leads to more significant changes of the system in Regimes I and II while damping perturbation is more crucial in Regime III for the NAO. In column (c) of Figure 35, we show the numerical simulations of the relative entropy in (1) with perturbations in all the directions within the entire two-dimensional parameter space $(\delta f, \delta a)^T$. Here, we take smaller $(\delta f, \delta a)^T$ in Regime II than those in Regimes I and III due to the smaller parameter values $(f, a)^T$ in Regime II. These numerical results, which are more expensive to compute, are consistent with the theoretical predictions in (166). Note that, although the most sensitive directions in Regimes I and II are close to each other, the ratio of the larger to smaller eigenvalues in the two regimes are quite different with 18.2979 in Regime I and 2.5307 in Regime II. This means that there is a direction of $(\delta f, \delta a)^T$ in Regime I in which the perturbation results in almost no change in the PDF, which can also be seen in column (c) of Figure 35.

Note that both the simple examples shown above contain the perfect knowledge of the present climate given by the unperturbed equilibrium PDFs. However, it is often quite difficult in practice to know the exact expression of these PDFs or it is computationally unaffordable to compute the gradient in high dimensions. Therefore, many approximations are combined with the information theoretical framework developed above. One common practical strategy is to adopt some approximated PDFs based on a few measurements such as the mean and covariance. It is also common to use imperfect or reduced models from a practical point of view, where FDT can also be incorporated to calculate the gradient of the present climate. Then, quantifying the model error in finding the most sensitive directions using imperfect models is an important issue. For detailed discussions of these topics, please see the reference [26].
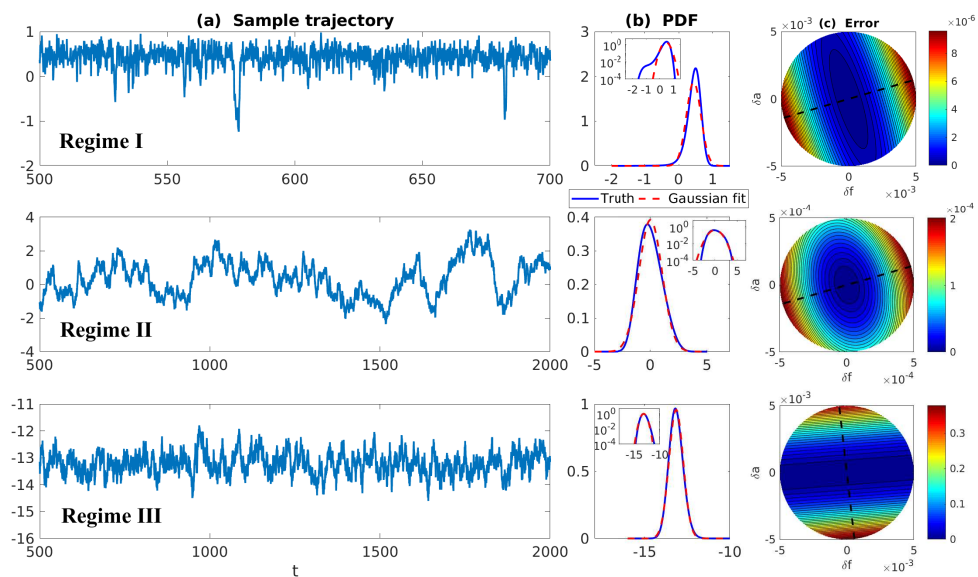
Figure 35. Time series (column (**a**)), equilibrium PDF (column (**b**)) and error due to the parameter perturbation $(\delta f, \delta a)^T$ in the two-dimensional parameter space (column (**c**)). The subpanels in column (**b**) are the PDFs in logarithm scale and the red dashed curves are the Gaussian fits. The parameter values of the three regimes are given in (166). The black dashed line shows the direction of the maximum uncertainty, namely the most sensitive direction of perturbation.

5. Given Time Series, Using Information Theory for Physics-Constrained Nonlinear Stochastic Model for Prediction

5.1. A General Framework

A central issue in contemporary science is the development of data-driven statistical dynamical models for the time series of a partial set of observed variables which arise from suitable observations from nature ([174] and references therein). Examples involve multi-level linear autoregressive models as well as ad hoc quadratic nonlinear regression models. It has been established recently [111] that ad hoc quadratic multilevel regression models can have finite time blow up of statistical solutions and pathological behavior of their invariant measure even though they match the data with high precision. Recently, a new class of physics-constrained multi-level nonlinear regression models was developed which involve both memory effects in time as well as physics constrained energy conserving nonlinear interactions [47,48] and completely avoid the above pathological behavior with full mathematical rigor.

The physics-constrained multi-level nonlinear regression models have the following forms:

$$\frac{d\mathbf{u}}{dt} = \mathbf{L}\mathbf{u} + \mathbf{B}(\mathbf{u},\mathbf{u}) + \mathbf{F} + r_1,$$
$$\frac{d\mathbf{r}}{dt} = \mathbf{Q}\mathbf{u} + \mathbf{A}\mathbf{r} + \sigma\dot{\mathbf{W}},$$

(167)

where $\mathbf{B}(\mathbf{u},\mathbf{u})$ is a quadratic nonlinearity which imposes the physical constraint of energy conservation on the nonlinear terms, namely

$$\mathbf{u} \cdot \mathbf{B}(\mathbf{u},\mathbf{u}) = 0.$$

(168)

In (167), the noise has the form $\mathbf{r} = (r_1, \ldots, r_p)^T$ where p denotes the number of memory levels and these noises are characterizes by the triangular matrix $\mathbf{A}$. The situation with $p = 0$ denotes the special zero-memory level model

$$\frac{d\mathbf{u}}{dt} = \mathbf{L}\mathbf{u} + \mathbf{B}(\mathbf{u}, \mathbf{u}) + \mathbf{F} + \sigma\dot{\mathbf{W}}. \tag{169}$$

See [47,48] for more details.

The ideas of developing physics-constrained nonlinear regression models can be combined with information calibration for predicting strongly nonlinear time series. The general procedure is shown in Figure 36. Here, the observed time series are divided into two parts, namely the training phase and the prediction phase. In the first step, physics-constrained nonlinear stochastic models are developed based on the characteristics of the given time series in the training phase. The second step involves applying information theory for model calibration using the time series again in the training phase. Then, the remaining time series is used for testing the prediction skill of the calibrated model.
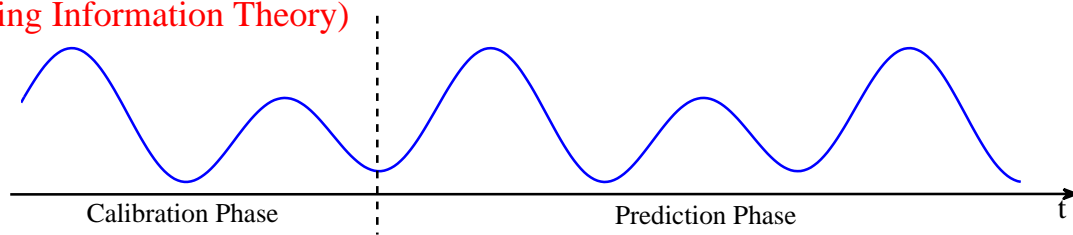
Figure 36. A general procedure for predicting time series.

5.2. Model Calibration via Information Theory

The key step above is the model calibration. As has been seen in Section 2, an effective model is expected to capture both the fidelity and sensitivity of nature. Therefore, two objective functions are utilized here for model calibration. The first one aims at capturing the model fidelity, which is given by minimizing the information distance between the PDF associated with the time series $\pi(\mathbf{u})$ and that associated with the model $\pi^M(\mathbf{u})$. The model fidelity guarantees the model's ability in recovering the long-term statistics of nature. However, the model fidelity does not necessarily provide skillful predictions at short and medium ranges. See examples in Figures 9 and 10. Thus, a second objective function is launched, which aims at minimizing the distance between the two autocorrelation functions associated with the observed time series and the model, respectively. As has been shown in Section 2.5, the autocorrelation function is associated with the mean response of the system. In fact, autocorrelation function characterizes the overall time-evolving patterns of the underlying dynamical system. Capturing the autocorrelation function ensures the dynamical consistency and is crucial for skillful short and medium-range forecasts using the proposed model.

Denote $\boldsymbol{\theta}$ the parameters in the physics-constrained nonlinear stochastic model. If both the model and nature are stationary, then the model calibration is given by the following optimization problem:

$$L = \min_{\boldsymbol{\theta}} \left(w_1 \mathcal{P}(\pi(\mathbf{u}), \pi^M(\mathbf{u})) + w_2 \mathcal{P}(E(\lambda), E^M(\lambda)) \right), \tag{170}$$

where w_1 and w_2 are weight functions. In (170), $E(\lambda)$ and $E^M(\lambda)$ are the energy spectra corresponding to the autocorrelation functions $R(t)$ and $R^M(t)$ of nature and the model, respectively, as studied in Section 2.5. In practice, time-periodic forcing may be involved in both the observed time series and

the physics-constrained model. In such a situation, both π^M and $R^M(t)$ can be formed by making use of the sample points in a long trajectory from the model. Since the stationary assumption is broken, the target function in (170) can be modified as the average value of the minimizations at different points within one period. Alternatively, an even cruder but practically useful target function involves a modified version of the first part in (170) given by the empirical measurements based on the time-averaged PDFs while the second part in (170) is replaced by directly computing the difference between the two autocorrelation functions. The important issue here is that both the PDF and the temporal correlation must be included in the target function.

The model calibration based on (170) or its modified versions has several salient features. First, the information distance $\mathcal{P}(\pi, \pi^M)$ is able to quantify the difference in the non-Gaussian statistics between the model and nature. Particularly, it is able to assess the skill of the model in recovering extreme events. Second, the two target functions play the role of improving long-term and short-term prediction skill, respectively. Therefore, the calibrated model can be used for predicting both transit phases and the statistical equilibrium state. Third, although the number of the parameters, namely the dimension of $\boldsymbol{\theta}$, can be large, the cost function L is in general robust with respect to the perturbation of $\boldsymbol{\theta}$ around the optimal values with a suitable choice of the physics-constrained nonlinear model. This is crucial in practice because it requires only a crude estimation of the parameters for the model, which greatly reduces the computational cost for searching in high-dimensional parameter space. In fact, as has been shown in [46], the energy-conserving nonlinear interaction in these physics-constrained nonlinear models is the underlying mechanism for such robustness property even in the presence of strong nonlinearity and intermittency. Finally, the physics-constrained nonlinear stochastic models require only a short training period [61,175] because the model development automatically involves a large portion of the information of nature. Thus, the data-driven physics-constrained modeling framework as discussed above is much cheaper and more practical than most non-parametric methods where a massive training data is typically required.

5.3. Applications: Assessing the Predictability Limits of Time Series Associated with Tropical Intraseasonal Variability

A striking application combining physics-constrained nonlinear model strategy with the above procedure is to assess the predictability limits of time series associated with the tropical intraseasonal variability such as the the Madden–Julian oscillation (MJO) and monsoon [46,61,176]. They yield an interesting class of low-order turbulent dynamical systems with extreme events and intermittency. Denote by u_1 and u_2 the two observed large-scale components of tropical intraseasonal variability. Here, we focus on the MJO time series [46], which are measured by outgoing longwave radiation (OLR; a proxy for convective activity) from satellite data [177]. See panel (a) of Figure 37. The PDFs for u_1 and u_2 (panel (c)) are highly non-Gaussian with fat tails indicative of the temporal intermittency in the large-scale cloud patterns. To describe the variability of the time series u_1 and u_2, the following family of low-order stochastic models are proposed:

$$\begin{aligned}
\frac{du_1}{dt} &= \left(-d_u u_1 + \gamma(v + v_f(t))u_1 - (a + \omega_u)u_2 \right) + \sigma_u \dot{W}_{u_1}, \\
\frac{du_2}{dt} &= \left(-d_u u_2 + \gamma(v + v_f(t))u_2 + (a + \omega_u)u_1 \right) + \sigma_u \dot{W}_{u_2}, \\
\frac{dv}{dt} &= \left(-d_v - \gamma(u_1^2 + u_2^2) \right) + \sigma_v \dot{W}_v, \\
\frac{d\omega_u}{dt} &= (-d_\omega \omega_u) + \sigma_\omega \dot{W}_\omega,
\end{aligned} \tag{171}$$

where

$$v_f(t) = f_0 + f_t \sin(\omega_f t + \phi). \tag{172}$$

In (171), in addition to the two observed variables u_1 and u_2, the other two variables v and ω_u are hidden and unobserved, representing the stochastic damping and stochastic phase, respectively. Here, $\dot{W}_{u_1}, \dot{W}_{u_2}, \dot{W}_v$ and $\dot{W}_\omega$ are independent white noise. The constant coefficients d_u, d_v, d_ω represent damping for each stochastic process, and the non-dimensional constant γ is the coefficient of the nonlinear interaction. The time periodic damping $v_f(t)$ in the Equation (171) is utilized to crudely model the active winter and the quiescent summer in the annual cycle. The constant coefficients ω_f and ϕ in (172) are the frequency and phase of the damping, respectively. All of the model variables are real. The energy conserving nonlinear interactions between u_1, u_2 and v, ω_u are seen in the following way. First, by dropping the linear and external forcing terms in (171), the remaining equations involving only the nonlinear parts of (171) read

$$
\begin{aligned}
\frac{du_1}{dt} &= \gamma v u_1 - \omega_u u_2, \\
\frac{du_2}{dt} &= \gamma v u_2 + \omega_u u_1, \\
\frac{dv}{dt} &= -\gamma(u_1^2 + u_2^2), \\
\frac{d\omega_u}{dt} &= 0.
\end{aligned}
\tag{173}
$$

To form the evolution equation of the energy from nonlinear interactions $E = (u_1^2 + u_2^2 + v^2 + \omega_u^2)/2$, we multiply the four equations in (173) by u_1, u_2, v, ω_u respectively and then sum them up. The resulting equation yields

$$
\frac{dE}{dt} = 0.
\tag{174}
$$

The vanishing of the right-hand side in (174) is due to the opposite signs of the nonlinear terms involving v multiplying u_1 and u_2 in (174) and those in (174) multiplying by v as well as the trivial cancellation of skew-symmetric terms involving ω_u.

Further motivation for the models in (171) is provided by the stochastic skeleton model which predicts key features of the MJO [178–181]. These are coupled nonlinear oscillator models of the MJO where if we identify the OLR variables with the envelope of synoptic scale convective activity, the hidden variables v, ω_u, and their dynamics become phenomenological surrogates for the energy-conserving interactions in the skeleton model involving the synoptic scale convective activity and the equatorial dynamic equations for temperature, velocity, and moisture.

It is shown in Figure 37 that, with the optimized parameters, the model in (171) almost perfectly capture the highly non-Gaussian fat-tailed PDFs, the autocorrelation functions (up to three months) and the power spectrums. In addition, the wiggles around one year in the autocorrelation functions, representing the annual cycle, are also recovered. Importantly, these parameters are pretty robust around the optimal values. In panel (b), a sample trajectory of u_1 from the model is shown, which shares many salient features as those of the observed MJO time series in panel (a). Another notable advantage of the physics-constrained nonlinear low-order stochastic models developed here is that the model structure allows an efficient nonlinear data assimilation scheme to determine the initial values of the hidden variables v, ω_u [140]. This facilitates the ensemble prediction algorithm since no direct observation is available for these hidden variables. In [46], significant prediction skill of these MJO indices using the physics-constrained nonlinear stochastic model (171) was shown. The prediction based on ensemble mean can have skill even up to 40 days. In addition, the ensemble spread accurately quantify the forecast uncertainty in both short and long terms. In light of a twin experiment, it was also revealed in [46] that the model in (171) is able to reach the predictability limit of the large-scale cloud patterns of the MJO.
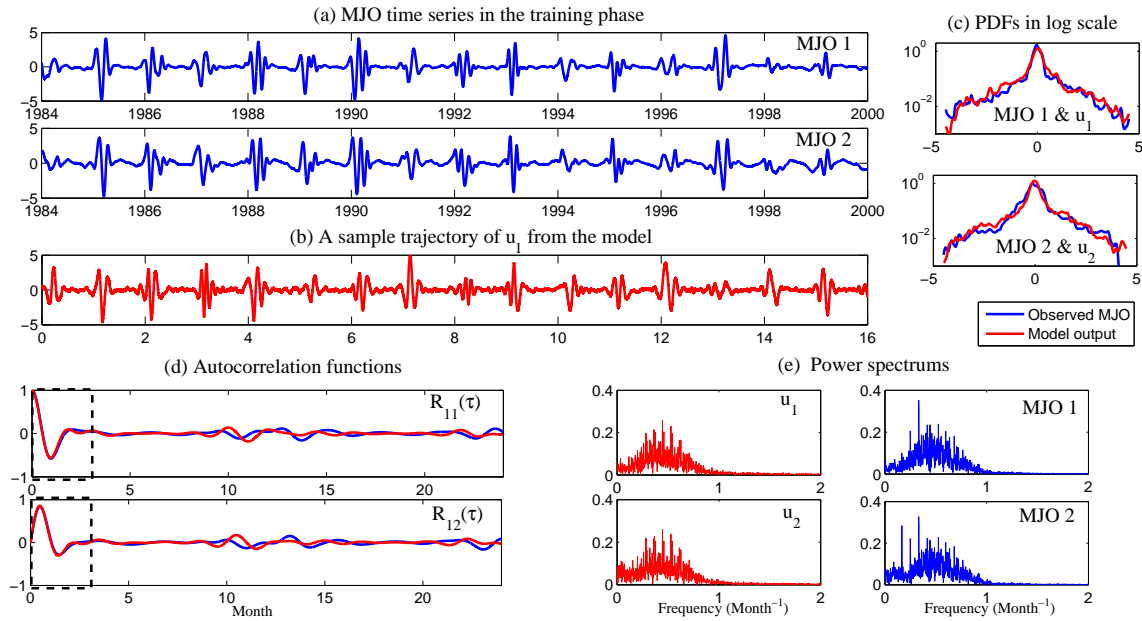
Figure 37. Calibrating the physics-constrained nonlinear stochastic model (171) for the Madden-Julian oscillation (MJO) time series. (**a**): the pair of MJO time series from observations; (**b**): a sample trajectory of u_1 from the model (171), which has the same length as the MJO time series in panel (**a**); (**c**): comparison of the highly intermittent PDFs in logarithm scale; (**d**): comparison of the autocorrelation functions, where the black dashed box indicates the time range within the first three months; (**e**): comparison of the power spectrums.

6. Reduced-Order Models (ROMs) for Complex Turbulent Dynamical Systems

6.1. Strategies for Reduced-Order Models for Predicting the Statistical Responses and UQ

6.1.1. Turbulent Dynamical System with Energy-Conserving Quadratic Nonlinearity

Let's consider a general framework of turbulent dynamical system [1],

$$\frac{d\mathbf{u}}{dt} = (L + D)\,\mathbf{u} + B\,(\mathbf{u}, \mathbf{u}) + \mathbf{F}\,(t) + \sigma_k\,(t)\,\dot{\mathbf{W}}_k\,(t; \omega)\,. \tag{175}$$

The model in (175) has the following properties:

- $\mathcal{L} = L + D$ is a linear operator representing dissipation and dispersion. Here, L is skew symmetric representing dispersion and D is a negative definite symmetric operator representing dissipative process such as surface drag, radiative damping, viscosity, etc.
- $B\,(\mathbf{u}, \mathbf{u})$ is a bilinear term and it satisfies energy conserving property with $\mathbf{u} \cdot B\,(\mathbf{u}, \mathbf{u}) = 0$.

The energy-conserving quadratic nonlinearity is one of the representative features in many turbulent dynamical systems in nature. The energy is transferred from the unstable modes to stable modes where the energy is dissipated resulting in a statistical steady state.

We use a finite-dimensional representation of the stochastic field consisting of a fixed-in-time, N-dimensional, orthonormal basis $\{\mathbf{v}_i\}_{i=1}^N$

$$\mathbf{u}\,(t) = \bar{\mathbf{u}}\,(t) + \sum_{i=1}^N Z_i\,(t; \omega)\,\mathbf{v}_i, \tag{176}$$

where $\bar{\mathbf{u}}(t) = \langle \mathbf{u}(t) \rangle$ represents the ensemble average of the response, i.e., the mean field, and $Z_i(t; \omega)$ are stochastic processes. By taking the average of (175) and using (176), the *mean equation* is given by

$$\frac{d\bar{\mathbf{u}}}{dt} = (L + D)\,\bar{\mathbf{u}} + B\,(\bar{\mathbf{u}}, \bar{\mathbf{u}}) + R_{ij}B\,(\mathbf{v}_i, \mathbf{v}_j) + \mathbf{F}(t), \tag{177}$$

with $R = \langle \mathbf{Z}\mathbf{Z}^* \rangle$ the covariance matrix. Moreover, the random component of the solution, $\mathbf{u}' = Z_i(t; \omega)\,\mathbf{v}_i$ satisfies

$$\frac{d\mathbf{u}'}{dt} = (L + D)\,\mathbf{u}' + B\,(\bar{\mathbf{u}}, \mathbf{u}') + B\,(\mathbf{u}', \bar{\mathbf{u}}) + B\,(\mathbf{u}', \mathbf{u}') - R_{ij}B\,(\mathbf{v}_i, \mathbf{v}_j) + \sigma_k(t)\,\dot{\mathbf{W}}_k(t; \omega). \tag{178}$$

By projecting the above equation to each basis element $\mathbf{v}_i$, we obtain

$$\begin{aligned}
\frac{dZ_i}{dt} = Z_j\big[\,(L + D)\,\mathbf{v}_j + B\,(\bar{\mathbf{u}}, \mathbf{v}_j) + B\,(\mathbf{v}_j, \bar{\mathbf{u}})\,\big] \cdot \mathbf{v}_i \\
+ \big[\,B\,(\mathbf{u}', \mathbf{u}') - R_{ij}B\,(\mathbf{v}_i, \mathbf{v}_j)\,\big] \cdot \mathbf{v}_i + \sigma_k(t)\,\dot{\mathbf{W}}_k(t; \omega) \cdot \mathbf{v}_i.
\end{aligned} \tag{179}$$

From the last equation, we directly obtain the exact *evolution equation of the covariant matrix* $R = \langle \mathbf{Z}\mathbf{Z}^* \rangle$

$$\frac{dR}{dt} = L_v R + R L_v^* + Q_F + Q_\sigma, \tag{180}$$

where we have:

1. The linear dynamical operator expressing energy transfers between the mean field and the stochastic modes (effect due to B), as well as energy dissipation (effect due to D) and non-normal dynamics (effect due to L)

$$\{L_v\}_{ij} = \big[(L + D)\,\mathbf{v}_j + B\,(\bar{\mathbf{u}}, \mathbf{v}_j) + B\,(\mathbf{v}_j, \bar{\mathbf{u}})\big] \cdot \mathbf{v}_i. \tag{181}$$

2. The positive definite operator expressing energy transfer due to the external stochastic forcing

$$\{Q_\sigma\}_{ij} = (\mathbf{v}_i \cdot \sigma_k)\,(\sigma_k \cdot \mathbf{v}_j). \tag{182}$$

3. The energy flux between different modes due to non-Gaussian statistics (or nonlinear terms) modeled through third-order moments

$$\{Q_F\}_{ij} = \langle Z_m Z_n Z_j \rangle\,B\,(\mathbf{v}_m, \mathbf{v}_n) \cdot \mathbf{v}_i + \langle Z_m Z_n Z_i \rangle\,B\,(\mathbf{v}_m, \mathbf{v}_n) \cdot \mathbf{v}_j. \tag{183}$$

We note that the energy conservation property of the quadratic operator B is inherited by the matrix Q_F since

$$\mathrm{tr}\,(Q_F) = 2\,\langle Z_m Z_n Z_i \rangle\,B\,(\mathbf{v}_m, \mathbf{v}_n) \cdot \mathbf{v}_i = 2B\,(\mathbf{u}', \mathbf{u}') \cdot \mathbf{u}' = 0. \tag{184}$$

Based on the observation that the eigenvalues are effectively changed by the existence of the nonlinear energy transfer mechanism, we propose a special form of the flux Q_F that will make the correct steady state statistics a stable equilibrium. More specifically, we split the nonlinear fluxes into a positive semi-definite part $Q_{F,+}$ and a negative semi-definite part $Q_{F,-}$:

$$Q_F = Q_{F,-} + Q_{F,+}.$$

As in (184), the nonlinear fluxes should always satisfy the conservative property of B, namely,

$$\mathrm{tr}[Q_F] = 0 \implies \mathrm{tr}[Q_{F,-}] = -\mathrm{tr}[Q_{F,+}].$$

The positive fluxes $Q_{F,+}$ indicate the energy being 'fed' to the stable modes in the form of external stochastic noise. On the other hand, the negative fluxes $Q_{F,-}$ should act directly on the linearly unstable modes of the spectrum, effectively stabilizing the unstable modes.

6.1.2. Modeling the Effect of Nonlinear Fluxes

The first idea here is to model the effect of the nonlinear energy transfers on each mode by adding additional damping balancing the linearly unstable character of these modes, and adding additional (white) stochastic excitation with standard deviation which will model the energy received by the stable modes,

$$Q_F^M = Q_{F,-}^M + Q_{F,+}^M = -D_M(R)R_M - R_M D_M^*(R) + \Sigma_M(R). \tag{185}$$

In (185), (D_M, Σ_M) are $N \times N$ matrices that replace the original nonlinear unstable and stable effects from the original dynamics. Here $Q_{F,-}^M = -D_M(R)R_M - R_M D_M^*(R)$ represents the additional damping effect to stabilize the unstable modes with positive Lyapunov coefficients, while $Q_{F,+}^M = \Sigma_M(R)$ is the positive-definite additional noise to compensate for the overdamped modes. Now, the problem is converted to finding expressions for D_M and Σ_M. In the following, by gradually adding more detailed characterization about the statistical dynamical model, we display the general procedure of constructing a hierarchy of the closure methods step by step. Below is a review about several model closure ideas [1,11,50,117] with increasing complexity:

1. *Quasilinear Gaussian closure model:* The simplest approximation for the closure methods at the first stage should be simply neglecting the nonlinear part entirely [182–184]. That is, set

    $$D_M(R) \equiv 0, \qquad \Sigma_M(R) \equiv 0, \qquad Q_F^{QG} \equiv 0. \tag{186}$$

    Thus, the nonlinear energy transfer mechanism will be entirely neglected in this Gaussian closure model. This is the similar idea in the eddy-damped Markovian model where the moment hierarchy is closed at the level of second moments with Gaussian assumption and a much larger eddy-damped parameter is introduced to replace the molecular viscosity [121,185]. Obviously, this crude Gaussian approximation will not work well in general due to the cutoff of the energy flow when strong nonlinear interactions between modes occur. Actually, the deficiency of this crude approximation has been shown under the Lorenz 96 framework, and in a final equilibrium state, there exists only one active mode with a critical wavenumber [11,50]. Such closures are only useful in the weakly nonlinear case where the quasi-linear effects are dominant.

2. *Models with consistent equilibrium statistics:* The next strategy is to construct the simplest closure model with consistent equilibrium statistics. Thus, the direct way is to choose constant damping and noise term scaled with the total variance. We propose two possible choices as in [50] for the damping and noise in (185) below.

    Gaussian closure 1 (GC1): let

    $$D_M(R) = \epsilon_M I_N \equiv \text{const.}, \quad \Sigma_M(R) = \sigma_M^2 I_N \equiv \text{const.}, \quad Q_F^{GC1} = -(\epsilon_M R + R\epsilon_M) + \sigma_M^2 I_N. \tag{187}$$

    Gaussian closure 2 (GC2): let

    $$D_M(R) = \epsilon_M \left(\frac{\text{tr}R}{\text{tr}R_{eq}} \right)^{1/2} I_N, \qquad \Sigma_M(R) = \sigma_M^2 \left(\frac{\text{tr}R}{\text{tr}R_{eq}} \right)^{3/2} I_N,$$
    $$Q_F^{GC2} = - \left(\frac{\text{tr}R}{\text{tr}R_{eq}} \right)^{1/2} (\epsilon_M R + R\epsilon_M) + \sigma_M^2 \left(\frac{\text{tr}R}{\text{tr}R_{eq}} \right)^{3/2} I_N. \tag{188}$$

    Above, only two scalar model parameters (ϵ_M, σ_M) are introduced, and I_N represents the $N \times N$ identity matrix. GC1 is the familiar strategy of adding constant damping and white noise forcing to represent nonlinear interaction; GC2 scales with the total variance $\text{tr}R$ (or total statistical

energy) so that the model sensitivity can be further improved as the system is perturbed. From both GC1 and GC2, we introduce uniform additional damping rate for each spectral mode controlled by a single scalar parameter ϵ_M, while the additional noise with variance σ_M^2 is added to make sure climate fidelity in equilibrium.

The statistical model closure Q_F^M is used to approximate the third-order moments in the true dynamics, thus the exponents of the total energy tr$R$ in GC2 should be consistent in scaling dimension. In the positive-definite part Q_{F+}^M, it calibrates the rate of energy injected into the spectral mode due to nonlinear effect in the order $|u'|^3$. The factor scales with the total energy with exponent $3/2$ so that the corrections keep consistent with the third-order moment approximations; In the negative damping rate $Q_{F,-}^M$, the scaling function is used to characterize the amount of energy that flows out the spectral mode due to nonlinear interactions. Scaling factor with a square-root of the total energy with exponent $1/2$ is applied for this damping rate multiplying the variance in order $|u'|^2$ to make it consistent in scaling dimension with third moments.

However, the damping and noise are chosen empirically without consideration about the true dynamical features in each mode. A more sophisticated strategy with slightly more complexity in computation is to introduce the damping and noise judiciously according to the linearized dynamics. Then, climate consistency for each mode can be satisfied automatically.

3. *Modified quasi-Gaussian (MQG) closure with equilibrium statistics:* In this modified quasi-Gaussian closure model originally proposed in [11,45], we exploit more about the true nonlinear energy transfer mechanism from the equilibrium statistical information. Thus, the additional damping and noise proposed as before are calibrated through the equilibrium nonlinear flux by letting

$$D_M(R) = -N_{M,eq}, \qquad \Sigma_M(R) = Q_{F,eq}^+, \qquad Q_F^{MQG} = -(N_M R + R N_M^*) + Q_F^+. \qquad (189)$$

$N_{M,eq}$ is the effective damping from equilibrium, and $Q_{F,eq}^+$ is the effective noise from the positive-definite component. Unperturbed equilibrium statistics in the nonlinear flux $Q_{F,eq}$ are used to calibrate the higher-order moments as additional energy sink and source. The true equilibrium higher-order flux can be calculated without error from first and second order moments in $(\bar{\mathbf{u}}_{eq}, R_{eq})$ from the unperturbed true dynamics (180) in steady state following the steady state statistical solution relation:

$$Q_{F,eq} = Q_{F,eq}^- + Q_{F,eq}^+ = -L_v(\bar{\mathbf{u}}_{eq})R_{eq} - R_{eq}L_v(\bar{\mathbf{u}}_{eq})^* - Q_\sigma, \qquad N_{M,eq} = \frac{1}{2}Q_{F,eq}^- R_{eq}^{-1}, \qquad (190)$$

where $Q_{F,eq}^-, Q_{F,eq}^+$ are the negative and positive definite components in the unperturbed equilibrium nonlinear flux $Q_{F,eq}$. Since exact model statistics are used in the imperfect model approximations, the true mechanism in the nonlinear energy transfer can be modeled under this first correction form. This is the similar idea used for measuring higher order interactions in [45], where more sophisticated and expensive calibrations are required to make that model work there.

6.1.3. A Reduced-Order Statistical Energy Model with Optimal Consistency and Sensitivity

The above closure model ideas, especially (187)–(189), have advantages of their own. Models in (187) and (188) are simple and efficient to construct with consistent equilibrium consistency, while (189) involves the true information about the higher-order statistics in equilibrium so that the energy mechanism can be characterized well. The validity of these approaches has been tested and compared from several papers [11,45,50] using the simplified triad model and L96 model. The methods have also been applied to the two-layer quasi-geostrophic equation [44,117], where the phase space of the original system is 5×10^5 while the ROM contains only 0.1% of the large scale modes and can cope with the changes in external forcing. Still, when it comes to the more complicated and realistic flow systems such as the quasi-geostrophic equations, more detailed calibration for model consistency and sensitivity is required to achieve the optimal performance. A preferred approach for the nonlinear

flux Q_F^M combining both the detailed model energy mechanism and control over model sensitivity is proposed in the form

$$Q_F^M = Q_F^{M,-} + Q_F^{M,+} = f_1(R) \left[-(N_{M,eq} + d_M I_N) R_M \right] + f_2(R) \left[Q_{F,eq}^+ + \Sigma_M \right]. \tag{191}$$

The closure form (191) consists of three indispensable components:

(i). *Higher-order corrections from equilibrium statistics:* In the first part of the correction using the damping and noise operator as $\left(N_{M,eq}, Q_{F,eq}^+ \right)$, unperturbed equilibrium statistics in the nonlinear flux $Q_{F,eq}$ are used to calibrate the higher-order moments as additional energy sink and source following the procedure in (189). Therefore, the equilibrium statistics can be guaranteed to be consistent with the truth, and the true energy mechanism can be restored.

(ii). *Additional damping and noise to model changes in nonlinear flux:* The above corrections in step (i) by using equilibrium information for nonlinear flux is found to be insufficient for accurate prediction in the reduced-order methods since the scheme is only marginally stable and the energy transferring mechanism may change with large deviation from the equilibrium case when external perturbations are applied. Thus, we also introduce the additional damping and noise (d_M, Σ_M) as from (187). d_M is just a constant scalar parameter to add uniform dissipation on each mode, and Σ_M is the further correction as an additional energy source to maintain climate fidelity.

(iii). *Statistical energy scaling to improve model sensitivity:* Still note that these additional parameters are added regardless of the true nonlinear perturbed energy mechanism where only unperturbed equilibrium statistics are used. To capture the responses to a specific perturbation forcing, it is better to make the imperfect model parameters change adaptively according to the total energy structure. Considering this, the additional damping and noise corrections are scaled with factors $f_1(R), f_2(R)$ related with the total statistical variance $\mathrm{tr}R$ as

$$f_1(R) = \left(\frac{\mathrm{tr}R}{\mathrm{tr}R_{eq}} \right)^{1/2}, \qquad f_2(R) = \left(\frac{\mathrm{tr}R}{\mathrm{tr}R_{eq}} \right)^{3/2}. \tag{192}$$

6.1.4. Calibration Strategy

As discussed in the previous sections, the calibration should involve two criteria: (1) the model fidelity (consistency) with the same equilibrium statistics as the truth, and (2) the optimal model sensitivity.

Let's denote $\pi_G(\mathbf{u})$ and $\pi_G^M(\mathbf{u})$ the Gaussian distributions of the truth and the imperfect model, as in Section 2.1. Here, the Gaussian approximation of the PDFs is adopted since the above reduced-order model strategy only involves the evolution of mean and variance in the imperfect model. According to the information-theoretic framework in Section 2.1, the statistical equilibrium fidelity means that the Gaussian relative entropy satisfies

$$\mathcal{P}(\pi_G(\mathbf{u}), \pi_G^M(\mathbf{u})) = 0. \tag{193}$$

Practically, we can make use of the explicit form (6) to calibrate the parameters such that (193) is satisfied. Statistical equilibrium fidelity is a natural necessary condition to tune the mean and variance of the imperfect model to match those of the perfect model; it is far from a sufficient condition. To see this, recall from Section 2.5 that different dynamical systems can have the same Gaussian invariant measure and therefore statistical equilibrium fidelity among the models is obviously satisfied (see [40] for several concrete examples). Thus, the condition in (193) should be regarded as an important necessary condition. UQ requires an accurate assessment of both the mean and variance and at least (193) guarantees calibration of this on a subspace, $\mathbf{u} \in \mathbb{R}^M$, for the unperturbed model. Climate scientists often just tune only the means (see [26] and references therein).

Next, the prediction skill of imperfect models can be improved by comparing the information distance through the linear response operator with the true model. The response in the Gaussian framework $\mathcal{P}(\pi_\delta, \pi_\delta^M)$ can be computed by making use of (10). This condition has been shown to be crucial in calibrating the parameters (see examples in Sections 2.5 and 5). The optimal model $M^* \in \mathcal{M}$ that ensures best information consistent responses to various kinds of perturbations is characterized with the smallest additional information in the linear response operator $\mathcal{R}$ among all the imperfect models, such that

$$\left\| \mathcal{P} \left(\pi_\delta, \pi_\delta^{M^*} \right) \right\|_{L^1([0,T])} = \min_{M \in \mathcal{M}} \left\| \mathcal{P} \left(\pi_\delta, \pi_\delta^M \right) \right\|_{L^1([0,T])}, \tag{194}$$

where π_δ^M can be achieved through a kicked response procedure (138) in the training phase compared with the actual observed data π_δ in nature, and the information distance between perturbed responses $\mathcal{P} \left(\pi_\delta, \pi_\delta^M \right)$ can be calculated with ease through the expansion formula (10). For the time-periodic cases, the information distance $\mathcal{P} \left(\pi_\delta \left(t \right), \pi_\delta^M \left(t \right) \right)$ is measured at each time instant, so the entire error is averaged under the L^1-norm inside proper time window $[0, T]$ (such as one period). Some low dimensional examples of this procedure for turbulent systems can be found in [10,30,186].

Below, in the example of predicting passive tracer extreme events using low-order reduced model (Section 6.3), the imperfect models are all linear Gaussian models. As we have already seen in Sections 2.2.1 and 2.3, the linear Gaussian models are only able to capture the response in the mean statistics. Therefore, minimizing the model sensitivity is based on optimizing the mean response in the imperfect models compared with that in the truth. Note that optimizing the mean response is equivalent to optimizing the autocorrelation function in the linear Gaussian models. Thus, instead of applying the general strategy with FDT for the optimization of the response, minimizing the spectral density (for autocorrelations) using information theory as discussed in Section 2.5 is an efficient alternative approach, which will be adopted below. The readers should keep in mind that these two methods share the same goal of optimizing the sensitivity in imperfect models.

6.2. Physics-Tuned Linear Regression Models for Hidden (Latent) Variables

Before we show concrete applications of the reduced-order modeling framework developed in Section 6.1, let's briefly discuss the physics-tuned linear regression modeling strategy. These physics-tuned linear regression models are particularly useful for simplifying the hidden or latent processes in complex dynamical systems while they preserve the important feedback from the latent variables to the resolved variables. Thus, such physics-tuned linear regression technique allows the dynamics and statistical structure of the models to become much more tractable and facilitates the application of the reduced-order modeling strategy in Section 6.1.

Consider the following general nonlinear system,

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= \mathbf{F}_1(\mathbf{u}, \mathbf{v}) + \sigma_{\mathbf{u}} \dot{\mathbf{W}}_{\mathbf{u}}, \\ \frac{d\mathbf{v}}{dt} &= \mathbf{F}_2(\mathbf{u}, \mathbf{v}) + \sigma_{\mathbf{v}} \dot{\mathbf{W}}_{\mathbf{v}}, \end{aligned} \tag{195}$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are nonlinear functions of $\mathbf{u}$ and $\mathbf{v}$. The model in (195) is usually written as a collection of Fourier modes. In (195), the state variables $\mathbf{u}$ are resolved variables that are our primary interest. The state variables $\mathbf{v}$ are the latent or unresolved variables, which nevertheless play an important role in contributing to the variability of $\mathbf{u}$ through nonlinear interactions. Here, the goal is to develop physics-tuned linear regression models of $\mathbf{v}$ such that their dynamics and statistical structure become much more tractable. Meanwhile, the feedback from $\mathbf{v}$ to $\mathbf{u}$ under the physics-tuned linear regression modeling framework are expected to be preserved to a large extent. The physics-tuned linear regression modeling framework for the latent variables v is given as follows:

$$\frac{d\mathbf{u}^M}{dt} = \mathbf{F}_1(\mathbf{u}^M, \mathbf{v}^M) + \sigma_{\mathbf{u}}\dot{\mathbf{W}}_{\mathbf{u}},$$

$$\frac{d\mathbf{v}^M}{dt} = \mathbf{\Lambda}^M \mathbf{v}^M + (\mathbf{v}^M - \hat{\mathbf{v}}^M) + \sigma_{\mathbf{v}}^M \dot{\mathbf{W}}_{\mathbf{v}}^M, \tag{196}$$

where $\mathbf{\Lambda}^M$ are $\sigma_{\mathbf{v}}^M$ both diagonal matrices. The i-th diagonal entry of $\mathbf{\Lambda}^M$ usually has the form $\lambda_j^M = -d_j^M + i\omega_j^M$, where the real part of each diagonal entry of λ_j^M, namely $-d_j^M$, is negative. In other words, each component v_j of $\mathbf{v}$ satisfies a one-dimensional OU process:

$$\frac{dv_j^M}{dt} = \lambda_j^M v_j^M + (v_j^M - \hat{v}_j^M) + \sigma_{\mathbf{v},j}^M \dot{W}_{\mathbf{v},j}^M. \tag{197}$$

The physics-tuned linear regression modeling strategy requires that each v_j^M in (197) satisfies both the model fidelity and model sensitivity compared with the i-th component of the truth $\mathbf{v}$, namely $\mathbf{v}_j$, in (195). Therefore, the model parameters $d_j^M, \omega_j^M, \hat{v}_j^M$ and $\sigma_{\mathbf{v},j}^M$ in (197) are calibrated by matching the equilibrium Gaussian PDF and the autocorrelation function with those associated with $\mathbf{v}_j$ in (195). See Section 2.5 for more technical details. Note that the goal here is to let $\mathbf{v}^M$ provide the least biased statistical feedback to $\mathbf{u}$ instead of recovering all the point-wise details of the latent random process $\mathbf{v}^M$.

Below, we use a simple example to illustrate physics-tuned linear regression modeling strategy and emphasize the importance of capturing the feedback from $\mathbf{v}$ to $\mathbf{u}$. Note that such ideas will be adopted in Section 6.3 for predicting the extreme events in passive tracers, where $\mathbf{v}$ and $\mathbf{u}$ can be thought of as the surrogates of the advection flow and the passive tracer fields there, respectively.

The true model is a two-dimensional nonlinear model:

$$\frac{du}{dt} = (-v + i\omega_u)u + \sigma_u \dot{W}_u,$$

$$\frac{dv}{dt} = (f + av + bv^2 - cv^3) + (A - Bv)\dot{W}_C + \sigma_v \dot{W}_A. \tag{198}$$

In (198), u is complex but v is real. The unresolved process v is given by the canonical model for low frequency atmospheric variability, derived based on stochastic mode reduction strategies [116,170]. It has been used in Section 4.3. The unresolved variable v serves as the stochastic damping for the resolved variable u. The feedback mechanism of v with suitable parameters can result in the intermittent instability of u. The parameters in this coupled model are all constants. We consider the following two dynamical regimes:

Fat-tailed regime: $a = -2.20, \ b = -2.6, \ c = 0.8, \ A = 1.0, \ B = -2.0, \ f = \ \ 2.0, \ \sigma_v = \sqrt{2}.$

Bimodal regime: $a = -2.64, \ b = \ \ 7.8, \ c = 4.0, \ A = 0.1, \ B = \ \ 0.1, \ f = -0.2, \ \sigma_v = \sqrt{2}, \tag{199}$

where the fat-tailed regime is a typical non-Gaussian regime for the unresolved process while the bimodal one is an extremely tough test regime. The other two parameters in the dynamics of u are the same in the two regimes,

$$\sigma_u = 0.1, \qquad \omega_u = 2. \tag{200}$$

The reduced model with simplified process of v is given by

$$\frac{du^M}{dt} = (-v^M + i\omega_u)u^M + \sigma_u \dot{W}_u,$$

$$\frac{dv^M}{dt} = -d_v^M(v^M - \hat{v}^M) + \sigma_v^M \dot{W}_v^M. \tag{201}$$

Since v is real in the true model (198), v^M is also real in the reduced model (201). Thus, $\lambda_j^M \equiv -d_v^M$ here, which is a special case of the general framework in (197). The three parameters $d^M, \hat{v}^M$ and σ_v^M

are tuned by capturing the model sensitivity (autocorrelation function) and model fidelity (equilibrium PDF) of the those associated with the truth of v in (198).

The sample trajectories and the statistics of both the truth and the reduced model with the physics-tuned parameters are shown in Figures 38 and 39 for the fat-tailed and the bimodal regimes, respectively. In both figures, despite the failure in capturing the non-Gaussianity in the hidden process v, the non-Gaussian fat-tailed PDFs as well as the intermittent trajectories of the resolved variable u are both recovered with high accuracy in the reduced model with physics-tuned parameters. One crucial reason for the high skill in the reduced model is that the autocorrelation function of v^M resembles that of v in the truth. Therefore, the duration and frequency of the intermittent phases of u^M are statistically similar to those of u. In fact, even in the bimodal regime which is an extremely tough test case (Figure 39), where the PDF of v^M is highly biased from that of v, the feedback mechanism from v to u is well recovered by the reduced model due to the capturing of both the model fidelity and model sensitivity. In Figure 40, we show that only matching the equilibrium PDF of v^M with that of v but ignoring the autocorrelation function in the calibration process is insufficient in recovering the key features of u. This emphasizes the importance of physics-tuned calibration strategy and the skill of using the resulting linear regression model for the hidden unresolved variables in capturing the non-Gaussian intermittent behavior of the resolved variable u.
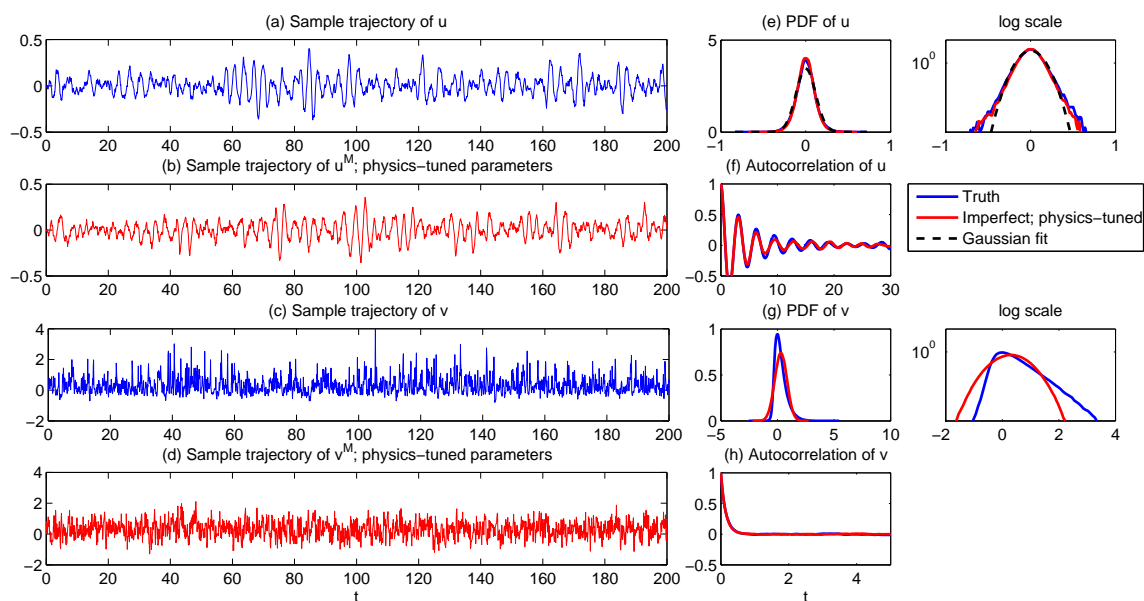
Figure 38. (**a**,**c**): sample trajectories of the truth of u and v in the regime where v has a fat-tailed PDF. (**b**,**d**): sample trajectories of u^M and v^M from the reduced model with physics-tuned parameters; (**e**,**g**): comparison of the PDFs of u and v in the truth and physics-tuned model. The right panels show the PDFs in the logarithm scale; (**f**,**h**): comparison of the autocorrelation function. All the trajectories and statistics with respect to u means the real part of u.
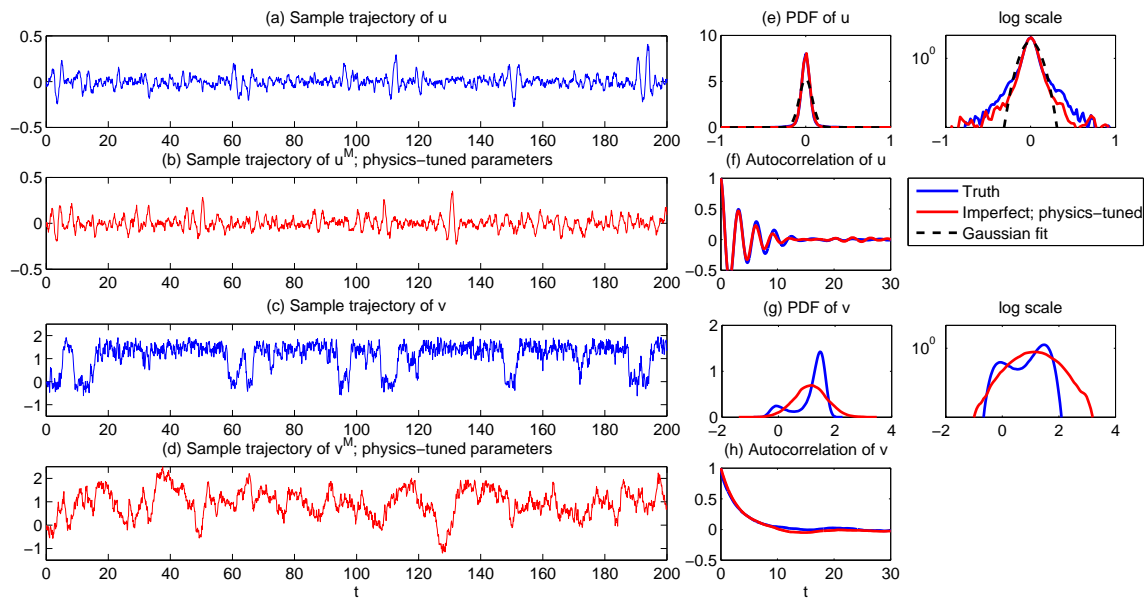
Figure 39. Same as Figure 38 but in bimodal regime of v.
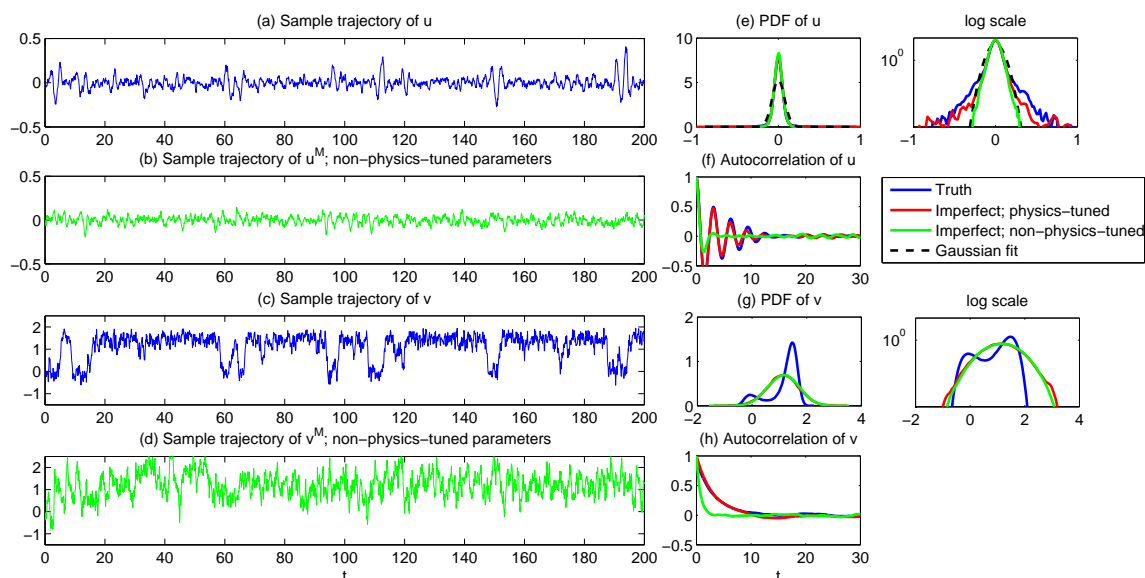
Figure 40. As in Figure 39 where the process of the truth of v is in a bimodal regime. In the reduced model, the parameters are not physical-tuned. Thus, a large model error is seen from both the trajectories in (**b**,**d**) compared with the truth (**a**,**c**) and the statistics in (**e**–**h**).

6.3. Predicting Passive Tracer Extreme Events

Turbulent diffusion models of passive tracers have numerous applications in geophysical science and engineering. These applications range from, for example, the spread of contaminants or hazardous plumes in environmental science, to the behavior of anthropogenic and natural tracers in climate change science, and many related areas [187–190]. The scalar field $T(\mathbf{x}, t)$ describes the concentration of the passive tracer immersed in the fluid which is carried with the local fluid velocity but which does not itself significantly influence the dynamics of the fluid. The evolution of the passive tracer is driven by turbulent advection, diffusion and usually uniform damping,

$$\frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T = -d_T T + \kappa \Delta T, \tag{202}$$

where $\mathbf{v}(x, y, t)$ is a velocity field. One key feature of great interest in the tracer turbulent model (202) is the existence of intermittency, which can be found in atmosphere observational data [190], laboratory experiments [191], and numerical simulations of idealized models [15,189,192,193].

A special form of the velocity field $\mathbf{v}$, which is a superposition of a spatially uniform but possibly temporally fluctuating cross-sweep in the x-direction and a random shear flow (with fluctuations possibly in both time and spatial x-direction) in the y-direction

$$\mathbf{v}(x, y, t) = (U(t), v(x, t)), \tag{203}$$

has been proposed by Majda et al. [189,193] and tested on simple mathematical models [26,29,194]. Assume the existence of a background mean gradient for the tracer varying in only y-variable and a tracer fluctuation component dependent with only the x-variable

$$T(\mathbf{x}, t) = T'(x) + \alpha y. \tag{204}$$

Combining (204) with the tracer dynamics (202) and the simplified flow field (203), the fluctuation part of the tracer T' satisfies the following equation:

$$\frac{\partial T'}{\partial t} + U(T)\frac{\partial T'}{\partial x} = -\alpha v(x, t) - d_T T' + \kappa \frac{\partial^2 T'}{\partial x^2}. \tag{205}$$

Despite their simplicity, the model (205) in random shear flow with a mean sweep can capture and preserve key features for various inertia range statistics for turbulent diffusion [192,195–198] including intermittency. Even for roughly Gaussian velocity fields v in (203) as observed in turbulent flows, the linear scalar field can experience rare but very large fluctuations in amplitude, and its statistics can depart significantly from Gaussianity displaying fatter tails representing the intermittency [199–203]. Explicit formulations about the statistical solutions of the tracers have been derived in [193] under this simplified flow system, and a rigorous mathematical proof about the intermittent fat tails in tracer distributions has been achieved recently in [195].

Complex nonlinear and non-Gaussian features in the flow components are unavoidable and ubiquitous especially in realistic turbulent flows. The complexity and large computational expense in resolving the highly turbulent advection flow equations require the introduction of simpler and more tractable imperfect models while still maintaining the ability in capturing the key intermittent features in the tracer field. Below, we investigate the effects from a nonlinear advection flow on the steady state passive tracer intermittency, and especially the errors and performances of imperfect approximation models are tested in a variety of turbulent regimes. In particular, the following two issues will be addressed in the remaining of this section:

1. Whether a linear Gaussian dynamics in approximating the advection flow is able to capture tracer non-Gaussian statistical structures?
2. How to design an unambiguous reduced-order stochastic modeling strategy with high prediction skill of the tracer field?

There are at least two motivations for using linear Gaussian imperfect models for describing advection flow v. First, the dynamics and statistical structure become much more tractable with explicit solutions that enable us to design the model and tune parameters with ease. Second, the computational difficulty and cost are also greatly reduced considering the simple and controllable structure of the linear model. However, it is challenging by applying these linear Gaussian models with no positive Lyapunov exponents to estimate the non-Gaussian flow field including various degrees of internal instabilities. Therefore, a systematic procedure in calibrating the imperfect model parameters is required.

Here, the information-theoretic framework developed in Section 6.1.4 is applied to train the imperfect model parameters in a training phase so that the model predicted stationary process can

possess the least biased estimation in energy and autocorrelation function, the latter of which plays an particularly important role in determining the structure of tracer statistics. With such a systematical calibration strategy, these linear stochastic models can be greatly improved through this proposed tuning strategy under a proper information metric. On the other hand, the reduced-order strategy aims at using low order equilibrium statistics in the model calibration as a correction to the nonlinear small scale feedback, which avoids high computational cost.

6.3.1. Approximating Nonlinear Advection Flow Using Physics-Tuned Linear Regression Model

Here, we aim at answering the question that whether a linear Gaussian dynamics is able to approximate the advection flow such that the non-Gaussian statistical structures in the tracer field is preserved. One good reference of this topic is [50].

To this end, we consider a background flow, which is driven by the 40-dimensional Lorenz 96 (L96) system [204]. The L96 model is able to mimic baroclinic turbulence in the midlatitude atmosphere with the effects of energy conserving nonlinear advection and dissipation, displaying a wide range of distinct dynamical regimes from Gaussian to extremely non-Gaussian statistics. Therefore, the L96 model is a representative testbed for studying the model error here.

The L96 system with state variables $\mathbf{u}(x, t) = (u_0, u_1, \ldots u_{J-1})^T$ is given by

$$\frac{du_j}{dt} = (u_{j+1} - u_{j-2})u_{j-1} - d_u(t)u_j + F(t), \qquad j = 0, 1, \ldots, J-1, \qquad J = 40, \tag{206}$$

with periodic boundary condition $u_0 = u_J$. Nonlinearity comes from the bilinear quadratic form $B_j(\mathbf{u}, \mathbf{u}) = (u_{j+1} - u_{j-2})u_{j-1}$, which conserves energy through $\mathbf{u} \cdot \mathbf{B}(\mathbf{u}, \mathbf{u}) = 0$. By changing the amplitude of the external forcing F, the L96 system displays a wide range of different dynamical regimes ranging from weakly chaotic ($F = 5$), strongly chaotic ($F = 8$), to finally full turbulence ($F = 16$) with varying statistics.

The advection flow field $\mathbf{v} = (U(t), v(x, t))$ is constructed from the L96 model solution. Note that the system is homogeneous and translation invariant along each grid point, so standard Fourier basis $\mathbf{e}_k = \left\{ e^{2\pi ikj/J} \right\}_{j=1}^{J-1}$ naturally becomes the empirical orthogonal functions (EOFs) of the system [50]. The state variables of the system can be decomposed under Fourier basis as

$$\mathbf{u}(x, t) = \bar{\mathbf{u}}(t) + \sum_{k=-J/2+1}^{J/2} \hat{u}_k(t)\mathbf{e}_k(x), \qquad \langle \hat{u}_k \rangle = 0, \qquad \hat{u}_{-k} = \hat{u}_k^*, \tag{207}$$

where $\langle \cdot \rangle$ is the ensemble average. We construct the passive tracer fields nonlinearly advected by the flow generated through the L-96 system. The gradient cross-sweeping component $U(t)$ is from the mean state with randomness from zero mode, while the shearing component $v(x_j, t)$ simulated by the flow fluctuation modes with varying values at each grid point,

$$U(t) = \bar{u}(t) + \hat{u}_0(t), \qquad v(x_j, t) = \sum_{k \neq 0} \hat{u}_k(t)e^{2\pi ikx_j}. \tag{208}$$

Below, we will focus on the statistical features of the scalar tracer field in stationary steady state. To make sure the system converges to the final stationary state, that is, $\bar{u}(t) \to \bar{u}_\infty$, $r_k(t) = \langle |\hat{u}_k|^2(t) \rangle \to r_{k,\infty}$ as $t \to \infty$, we consider the simplified dynamics of (206) with constant damping and forcing terms $d_u \equiv d_u(t)$, $F \equiv F(t)$.

The exact dynamical equations for each mode in the shearing flow $\hat{u}_k$ and the mean gradient U can be derived from the L96 system (206) as

$$\frac{dU}{dt} = -d_u U(t) + \sum_{k \neq 0} \Gamma_k |\hat{u}_k|^2(t) + F, \tag{209}$$

$$\frac{d\hat{u}_k}{dt} = -d_u \hat{u}_k + \left(e^{2\pi i \frac{k}{J}} - e^{-2\pi i \frac{2k}{J}} \right) U(t)\hat{u}_k + \sum_{m \neq 0} \hat{u}_{k+m} \hat{u}_m^* \left(e^{2\pi i \frac{2m+k}{J}} - e^{-2\pi i \frac{m+2k}{J}} \right), \tag{210}$$

where $k = 1, \ldots, J/2$ and the energy transfer rate is $\Gamma_k = \cos\frac{4\pi k}{J} - \cos\frac{2\pi k}{J}$. The cross-sweep field U is forced by the combined effects from each fluctuation mode $\sum_{k \neq 0} \Gamma_k |\hat{u}_k|^2$, and conversely the shearing flow is advected by the mean drift through the second term in the first line in (210).

The accuracy in the steady state passive tracer statistics is limited by the modeling and computation skill of the complex background advection flow v. However, several difficulties cannot be bypassed if we directly go with the true flow system with nonlinearity. First, simple Galerkin truncation of high frequency wave-numbers in the dynamical equations may introduce large errors to the flow system due to strong nonlinear interactions between the (truncated) small scale and large scale modes. Second, even with a low dimensional Galerkin truncation model, large ensemble size may still be required to resolve the flow if non-Gaussian features and intermittencies are important and of interest. On the other hand, returning to our original problem, the central issue of major interest is the turbulent fluctuation and statistical structure of the passive tracer T rather than the background flow field v. Considering both sides of the problem, the question that is worth asking is whether we can predict the crucial features (such as, intermittency) in steady state tracer statistics advected and forced by nonlinear non-Gaussian background flow v using simpler imperfect models with error for the background dynamical field.

Now, we adopt the simplest approximation about the advection flow with imperfect models using linear stochastic dynamics along each spectral mode from the Ornstein–Uhlenbeck process [15,193,196,205]. With the simple structures in these linear Gaussian models, the dynamics and statistical structure become much more tractable with explicit solutions that enable us to design the model and tune parameters with ease. The linear stochastic models for each mode can be written as

$$\frac{d\hat{u}_k^M}{dt} = (-\gamma_{u_k} + i\omega_{u_k})\hat{u}_k^M + \sigma_{u_k}\dot{W}_k, \tag{211}$$

with $\gamma_{u_k}, \omega_{u_k}$ and σ_{u_k} as parameters to be determined, together with the dynamics for the mean

$$\frac{d\bar{u}^M}{dt} = -d_u \bar{u}^M + \sum_{k \neq 0} \Gamma_k r_k^M + \hat{F}, \tag{212}$$

with $r_k^M = \langle |\hat{u}_k^M|^2 \rangle$. Note that, in both (211) and (212), we consider all the Fourier modes. In practice, Galerkin truncation is naturally applied to these imperfect models, which greatly reduces the dimension of the imperfect system [81]. Since the goal of this subsection is to understand the role of these linear models with optimized parameters, we do not include the Galerkin truncation here. In the next subsection, we will apply the Galerkin truncation for $\hat{u}_k^M$ and reach a suite of low-order models in approximating both the velocity and the tracer fields.

Under the approximations in (211) and (212), the background flow $v^M = (U^M(t), v^M(x_j, t))$ can be constructed as before for the mean cross-sweep U^M and the shearing flow v^M in the tracer model (202),

$$U^M(t) = \bar{u}^M(t) + \hat{u}_0^M(t), \qquad v^M(x_j, t) = \sum_{k \neq 0} \hat{u}_k^M(t) e^{2\pi i k x_j}. \tag{213}$$

Now, the problem is converted to finding systematic strategies of assigning values to the three undetermined coefficients $\gamma_{u_k}, \omega_{u_k}, \sigma_{u_k}$ so that the tracer structure (intermittency) can be reconstructed

from this imperfect model. They should be chosen in an unambiguous way according to the true steady state statistics of the system (which is available from observations). In comparison with the original equation for each mode described in (210), the linear Gaussian approximation of L96 system replaces the nonlinear interaction part in the second line of (210) by linear damping and rotation together with a white noise

$$\sum_{m \neq 0} \hat{u}_{k+m} \hat{u}_m^* \left(e^{2\pi i \frac{2m+k}{J}} - e^{-2\pi i \frac{m+2k}{J}} \right) \sim (-\gamma_{u_k} + i\omega_{u_k})\hat{u}_k^M + \sigma_{u_k}\dot{W}_k. \tag{214}$$

The white noise $\sigma_{u_k}\dot{W}_k$ is added to each Fourier mode in order to make sure that the system converges to the consistent equilibrium steady state spectra. γ_{u_k} represents the damping that neutralizes the additional energy from the white noise. The imaginary component ω_{u_k} is the additional degree of freedom for tuning the autocorrelation function (or in other words, to control the 'memory' of this mode of its previous history). Note that the quasi-linear part with $U(t)$ in the first line of the formula (210) is also included in the coefficients $\gamma_{u_k}, \omega_{u_k}$. It is discovered that even under this linear flow field with Gaussian statistics, intermittency with fat-tailed distributions can be generated in the steady state tracer distributions [193,195]. Here, the challenge is whether we can still capture the correct structure in the tracer spectra and density functions, especially for the intermittency, under these imperfect linear models. Therefore, judicious choice of the model parameters needs to be investigated.

One of the simplest and most direction way to estimate the undetermined coefficients $\gamma_{u_k}, \omega_{u_k}, \sigma_{u_k}$ is through the mean stochastic model (MSM) [15,41] by calibrating the energy (variance) $E_k = \langle|\hat{u}_k(t) - \langle\hat{u}_k\rangle|^2|\rangle$ and decorrelation time τ in (44) of the truth (known as "MSM tuned parameters"). Note that the decorrelation time $\tau = T_k + i\theta_k$ here contains real and imaginary parts, fitting both as well as the energy provides three conditions. Despite the simplicity in this mean stochastic model, reasonably skillful prediction in uncertainty quantification as well as filtering under this strategy have been obtained for some turbulent systems [193,195]. However, MSM still suffers several shortcomings when strong nonlinearity takes place in the system. Most importantly, the decorrelation time $\tau = T_k + i\theta_k$ involves only the time-integrated effects in each mode. This works well when the system is strongly mixing within a nearly Gaussian regime, whereas, when non-Gaussian features become crucial in the system, the pointwise decaying process of the entire autocorrelation function $R(t)$ becomes important and we need take into account the temporal performance of the autocorrelation in the linear model approximation. This has already been seen in the simple example in Section 2.5. In fact, the autocorrelation function becomes strongly oscillatory when $F = 5$ in the L96 model, which shows the insufficiency of fitting only the decorrelation time.

Therefore, following the physics-tuned linear regression modeling strategy in Section 6.2 and the information-theoretic framework of calibrating the autocorrelation function in Section 2.5, we fit the autocorrelation function of each $\hat{u}_k$ by the spectral information criteria (47) and (48). Note that the linear Gaussian model in (211) has explicit solution for the autocorrelation function and power spectrum (52), which provides an extremely efficient way of calibrating the two parameters $\gamma_{u_k}, \omega_{u_k}$. The remaining parameter σ_{u_k} is calibrated by fitting the energy. Finally, we keep the tracer equation to be the same in this example. Finding a reduced order model for the tracer equation following the general strategy in Section 6.1 will be discussed in the next subsection.

In Figure 41, the statistical features of both the advection field v and the tracer T are shown. Here, the parameters in the true model (205), (209) and (210) are as follows:

$$d_T = 0.1, \qquad \alpha = 2, \qquad \kappa = 0.001, \qquad d_u = 1, \qquad F = 5. \tag{215}$$

Note that $F = 5$ corresponds to the weakly chaotic regime in L96 model, which results in a very slow mixing and therefore the autocorrelation function in a certain modes decays quite slowly with strong oscillations. See the black curves Panel (a) of Figure 41. It is expected from Section 2.5 that using the strategy of MSM by fitting only the decorrelation time results in a large bias, which is

clearly indicated by the blue curve in Panel (a). On the other hand, with the physics-tuned parameters calibrated based on fitting the autocrrelation function via the spectral density, the imperfect model provides a significantly accurate estimation of the autocorrelation function even in this tough regime. Next, the comparison of the PDFs associated with both v and T are shown in Panels (b)–(g). Clearly, the linear Gaussian models of v fail to capture the sub-Gaussian PDFs of the velocity field, which indicates an information barrier. Nevertheless, the nonlinear interaction between U and T allows the imperfect model to capture the non-Gaussian features in the tracer field with fat-tailed PDFs in both physical space (Panel (e)) and spectrum space (Panels (f)–(g)). The sample time series using the linear Gaussian velocity model (209) and (210) with the physics-tuned parameters also resembles that of the truth with significant intermittency (Panels (h) and (i)). On the other hand, the linear model with MSM tuned parameters (fitting only the decorrelation time) fails to capture these features (not shown here). See [81] for more discussions and numerical tests in other regimes ($F = 8$ and $F = 16$).
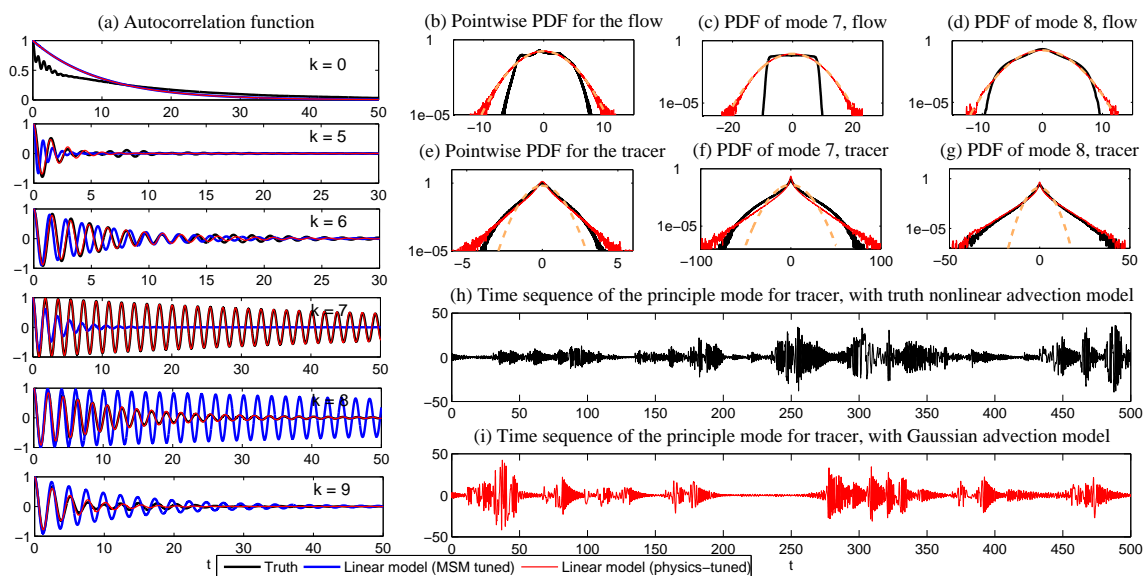

Figure 41. Comparison of the statistical feature of both the advection field v and the tracer T. The true advection model is given by L96 system (41) with $F = 5$ (weakly chaotic regime). (**a**): fitting of the autocorrelation functions for representative Fourier modes $k = 0$ and $5 \leq |k| \leq 9$ of the flow state variables $\hat{u}_k$, where only the real part is shown. The autocorrelation function from the true system is plotted in thick black lines, and the results from MSM is in blue lines, while the optimal model results from tuning parameters in spectral density functions are shown in red lines. Note that black and red lines are largely overlapped together; (**b**–**g**): comparison of probability density functions in logarithm scale, where the brown dashed curves are the Gaussian fit. Note that modes 7 and 8 are the most energetic modes in L96 model with $F = 5$; (**h**,**i**): sample trajectories of the tracer principal mode from the perfect model and linear Gaussian model with physics-tuned parameters.

6.3.2. Predicting Passive Tracer Extreme Events with Low-Order Stochastic Models

Now, we aim at answering the second question proposed at the beginning of this section. That is, how to design an unambiguous reduced-order modeling (ROM) strategy with high prediction skill of the tracer field [82]. Here, we consider a more realistic and complicated advection flow $\mathbf{v}(\mathbf{x}, t)$, which is described from the solution of the two-layer quasi-geophysics (QG) equation [121,143]

$$
\frac{\partial q_j}{\partial t} + \mathbf{v}_j \cdot \nabla q_j + (\beta + k_d^2 U_j)\frac{\partial \psi_j}{\partial x} = -\delta_{2j} r \Delta \psi_j - \nu \Delta^s q_j,
$$

$$
q_j = \Delta \psi_j + \frac{k_d^2}{2}(\psi_{3-j} - \psi_j), \qquad \mathbf{v}_j = (U_j, 0) + \mathbf{v}_j'. \tag{216}
$$

Above, the subindex $j = 1, 2$ is used to represent the upper and lower layer of the two-layer flow model. The two-dimensional incompressible velocity field $\mathbf{v}_j$ is decomposed into a zonal mean cross sweep, $(U_j, 0)$, and a fluctuating shear flow $\mathbf{v}'_j = \nabla^\perp \psi_j = (-\partial_y \psi_j, \partial_x \psi_j)$. For the passive tracer field, now we assume the background mean gradient varying in both x and y directions together with a tracer fluctuation component

$$T_j(\mathbf{x}, t) = \boldsymbol{\alpha} \cdot \mathbf{x} + T'_j(\mathbf{x}, t), \tag{217}$$

where $\boldsymbol{\alpha} = (\alpha_x, \alpha_y)$. Plugging (217) into (202), the fluctuation part of the passive tracer model yields

$$\epsilon \frac{\partial T'_j}{\partial t} + \mathbf{v}'_j(\mathbf{x}, t) \cdot \nabla T'_j + U_j \frac{\partial T'_j}{\partial x} = -(\alpha_x u_j + \alpha_y v_j)(\mathbf{x}, t) - d_T T'_j + \kappa \Delta T'_j. \tag{218}$$

In (218), $\mathbf{v}'_j = (u_j, v_j)$ is the fluctuating advection flow field from the solution of (216) together with a zonal mean flow U_j. In addition, a scale separation in the tracer Equation (218) with order ϵ is introduced. The difference in time scale in the tracer is through a different time scale, $\tilde{t} = \epsilon^{-1} t$, in the tracer time as in various previous works [189,193,195]. As $\epsilon < 1$, the velocity field is varying at a faster time scale than the passive tracer process, while on the other hand with $\epsilon > 1$ the tracer evolves in a more rapid rate than the advection field. A long time rescaling limit with explicit analytic tracer solutions is derived in [195] and numerical simulations for varying values of ϵ among a wide range are investigated in [192] under a much simpler linear model. In general, different intermittent features will be generated from near Gaussian statistics to distributions with fat tails as the scale separation parameter value changes [189,192].

Given periodic boundary condition in both the two-layer flow and the tracer field, we formulate the flow and tracer fields with Galerkin truncation to finite number of Fourier modes. Spatial Fourier decomposition in flow potential vorticity q_j and passive tracer disturbance T'_j can be written in the expansion under modes $\exp(i\mathbf{k} \cdot \mathbf{x})$ as

$$q_j = \sum_{\mathbf{k}} \hat{q}_{j,\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{x}}, \qquad T'_j = \sum_{\mathbf{k}} \hat{T}_{j,\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{x}}. \tag{219}$$

Note that here we focus on the homogeneous flow on mesoscale and therefore the periodic condition is reasonable. By projecting the tracer and flow Equations (218) and (216) to each Fourier spectral mode, equations for the spectral coefficients in each wavenumber of the two-layer tracer field $\vec{T}_{\mathbf{k}} = (\hat{T}_{1,\mathbf{k}}, \hat{T}_{2,\mathbf{k}})^T$, and two-layer advection flow field $\vec{q}_{\mathbf{k}} = (\hat{q}_{1,\mathbf{k}}, \hat{q}_{2,\mathbf{k}})^T$, form the set of ODEs in the spectral domain as

$$\frac{d\vec{T}_{\mathbf{k}}}{dt} + \epsilon^{-1} \sum_{\mathbf{m}+\mathbf{n}=\mathbf{k}} \left(A_{\mathbf{km}} \vec{q}_{\mathbf{m}} \circ \vec{T}_{\mathbf{n}} + A_{\mathbf{kn}} \vec{q}_{\mathbf{n}} \circ \vec{T}_{\mathbf{m}} \right) = -\epsilon^{-1} (\gamma_{T,\mathbf{k}} + i\omega_{T,\mathbf{k}}) \vec{T}_{\mathbf{k}} + \epsilon^{-1} G_{\mathbf{k}} \vec{q}_{\mathbf{k}}, \tag{220}$$

$$\frac{d\vec{q}_{\mathbf{k}}}{dt} + \sum_{\mathbf{m}+\mathbf{n}=\mathbf{k}} \left(A_{\mathbf{km}} \vec{q}_{\mathbf{m}} \circ \vec{T}_{\mathbf{n}} + A_{\mathbf{kn}} \vec{q}_{\mathbf{n}} \circ \vec{T}_{\mathbf{m}} \right) = -(\gamma_{q,\mathbf{k}} + i\omega_{q,\mathbf{k}}) \vec{q}_{\mathbf{k}}, \tag{221}$$

where '$\circ$' is used to denote the pointwise produce, namely $\mathbf{a} \circ \mathbf{b} = (a_i b_i)$. The potential vorticities $\vec{q}_{\mathbf{k}}$ and stream function $\vec{\psi}_{\mathbf{k}}$ in two layers are related by the transform matrix $H_{\mathbf{k}}$,

$$\vec{q}_{\mathbf{k}} = H_{\mathbf{k}} \vec{\psi}_{\mathbf{k}} = -\begin{pmatrix} |\mathbf{k}|^2 + \frac{k_d^2}{2} & -\frac{k_d^2}{2} \\ -\frac{k_d^2}{2} & |\mathbf{k}|^2 + \frac{k_d^2}{2} \end{pmatrix} \vec{\psi}_{\mathbf{k}} \tag{222}$$

through the relation $q_j = \nabla^2 \psi_j + \frac{k_d^2}{2}(\psi_{3-j} - \psi_j)$ in (216). The other operators and terms in the nonlinear dynamics (220) and (221) are given by

$$A_{\mathbf{km}} = \frac{1}{2}(k_x m_y - k_y m_x)H_m^{-1}, \qquad G_{\mathbf{k}} = -i\boldsymbol{\alpha} \cdot \mathbf{k}^\perp H_{\mathbf{k}}^{-1} = \boldsymbol{\Gamma}_{\mathbf{k}} H_{\mathbf{k}}^{-1}, \qquad \gamma_{T,\mathbf{k}} = d_t + \kappa |\mathbf{k}|^2,$$

$$\omega_{T,\mathbf{k}} = k_x \vec{U}, \qquad \gamma_{q,\mathbf{k}} = (0,1)^T \circ r|\mathbf{k}|^2 H_{\mathbf{k}}^{-1} + \nu |\mathbf{k}|^{2s}, \qquad \omega_{q,\mathbf{k}} = k_x \left(\vec{U} + (\beta + k_d^2 \vec{U})H_{\mathbf{k}}^{-1} \right). \tag{223}$$

In (223), the linear dissipation $\gamma_{T,\mathbf{k}}$ is due to the Ekman friction applied only on the bottom layer and the hyperviscosity. The dispersion $\omega_{T,\mathbf{k}}$ is from the rotational β-effect as well as the background zonal mean flow advection from the original Equation (216) applied on the vorticity modes.

The advection terms in the tracer and flow Equations (220) and (221) involve interactions between modes of different scales along the entire spectrum in a large dimensional phase space, thus usually high computational cost is required in achieving accurate statistical results from direct numerical simulations. In general, intermittency in a tracer field is dominated by the variability in largest scales, thus we will concentrate on the large-scale modes with wavenumber $|\mathbf{k}| \leq M \ll N$, where M is the number of resolved modes and N is the full dimensionality of the system. Usually, we could choose M much smaller than N that only covers the essential most energetic directions in the flow system. Below, we first develop the simple strategy with linear corrections to approximate the advection flow field in the leading modes, which is similar to that in Section 6.3.1. Then, the calibration and improvement of the imperfect models due to model errors from this approximation will be discussed.

As in Section 6.3.1, in order to approximate the advection flow, the simple Gaussian approximation is adopted to replace the quadratic interactions $(\mathbf{v} \cdot \nabla q)_{\mathbf{k}}$ in the flow equations by additional linear damping and random Gaussian noise. Thus, the reduced-order advection flow equations are given by

$$\frac{d\vec{q}_{M,\mathbf{k}}}{dt} = -(\gamma_{q,\mathbf{k}} + i\omega_{q,\mathbf{k}})\vec{q}_{M,\mathbf{k}} - D_{q,\mathbf{k}}^M \vec{q}_{M,\mathbf{k}} + \Sigma_{q,\mathbf{k}}^M \dot{\vec{W}}_{q,\mathbf{k}}, \qquad 1 \leq |\mathbf{k}| \leq M,$$

$$\mathbf{v}_M = \nabla^\perp \vec{\psi}_M, \qquad \vec{q}_{M,\mathbf{k}} = H_{\mathbf{k}} \vec{\psi}_{M,\mathbf{k}}, \tag{224}$$

with only Gaussian statistics generated. Only the first M large-scale modes $1 \leq |\mathbf{k}| \leq M$ are resolved in the reduced-order model (224). In addition to the linear dissipation and dispersion operators $(\gamma_{q,\mathbf{k}}, \omega_{q,\mathbf{k}})$, additional damping and noise $D_{q,\mathbf{k}}^M, \sigma_{q,\mathbf{k}}^M$ are introduced to correct model errors due to the neglected nonlinear interactions in the flow equations. On the other hand, there is no additional model calibrations of the tracer field statistics in case of over fitting of data. The idea here is to improve the reduced-order model prediction skill by optimizing the background advection flow field, thus the reduced order passive tracer equations can be modeled through a direct truncation

$$\frac{d\vec{T}_{M,\mathbf{k}}}{d\tilde{t}} + \left(\tilde{\mathbf{v}}_M \cdot \nabla \vec{T}_M \right)_{\mathbf{k}} = \boldsymbol{\Gamma}_{\mathbf{k}} \vec{\psi}_{M,\mathbf{k}} - (\gamma_{T,\mathbf{k}} + i\omega_{T,\mathbf{k}})\vec{T}_{M,\mathbf{k}}, \qquad 1 \leq |k| \leq M,$$

$$\tilde{\mathbf{v}}_M = \sum_{|\mathbf{k}| \leq M_1} i\mathbf{k}^\perp \vec{\psi}_{M,\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{x}}, \qquad M_1 \leq M, \tag{225}$$

where only the first leading modes of the advection flow $1 \leq |\mathbf{k}| \leq M \ll N$ are resolved in the tracer approximation model.

Again, the major difficulty in modeling the tracer dynamics is from the accurate approximation of the tracer advection $A_{\mathbf{km}} \vec{q}_{\mathbf{m}} \circ \vec{T}_{\mathbf{n}}$ in (220). Exact modeling about this nonlinear interaction term requires the flow mode solution $\vec{q}_{M,\mathbf{k}}$ along the entire spectrum $0 < |\mathbf{k}| \leq N$, while only the first M leading modes are available through the reduced-order model. One crude approximation idea could be to replace the nonlinear advection in the tracer field $\mathbf{v}(\mathbf{x},t) \cdot \nabla \vec{T}(\mathbf{x},t)$ with damping and noise in a similar fashion as the flow approximation model (224). However, as discussed in previous works [50,195], the nonlinear advection in the tracer equation is crucial in the generation of many important statistical features including the intermittency. Thus, including of nonlinear effects from the

flow solution is essential, at least for the large scale modes. On the left-hand side of the Equation (225), the nonlinear advection $\tilde{\mathbf{v}}_M \cdot \nabla \vec{T}_M$ is modeled explicitly, but only the first $M_1 \leq M$ largest scale flow modes in the model velocity solution $\tilde{\mathbf{v}}_M$ are used to calculate the imperfect model tracer advection. This nonlinear advection is essential in generating the accurate spectra in tracer statistics, while it is also not expensive to calculate since only leading modes are involved. The idea for this approximation is through the assumption that the dominant features in tracer statistics (such as intermittency and equilibrium spectrum in large scales) are due to the leading advection flow modes with largest energy.

Now, we calibrate the imperfect low-order linear Gaussian model for advection flow system (225) using equilibrium statistics and information theory. Such calibration procedure is divided into two steps:

1. Properly reflecting the nonlinear energy mechanism from the true system.
2. Imperfect stochastic model consistency in equilibrium statistics and autocorrelation functions.

In the first step, we aim at making sure that the imperfect model calibration parameters $(D^M_{q,\mathbf{k}}, \Sigma^M_{q,\mathbf{k}})$ can properly reflect the true nonlinear energy mechanism from the true system. The consistent imperfect model then can be proposed by consulting the model statistical dynamics. Therefore, it is useful to investigate the statistical equations for the second order moments from the fluctuation equations of (221). The dynamics for the covariance matrix $R^q_{\mathbf{k}} = \langle \vec{q}_{\mathbf{k}} \vec{q}^*_{\mathbf{k}} \rangle$ of flow vorticity can be derived as a 2×2 blocked system for each wavenumber [117],

$$\frac{dR^q_{\mathbf{k}}}{dt} + \mathcal{L}_{\mathbf{k}}(\bar{q}) R^q_{\mathbf{k}} + R^q_{\mathbf{k}} \mathcal{L}^*_{\mathbf{k}}(\bar{q}) + Q^q_F = (\mathcal{L}^q_{\mathbf{k}} + \mathcal{D}^q_{\mathbf{k}}) R^q_{\mathbf{k}} + R^q_{\mathbf{k}} (\mathcal{L}^q_{\mathbf{k}} + \mathcal{D}^q_{\mathbf{k}})^*, \qquad |\mathbf{k}| \leq N. \tag{226}$$

The linear operators $(\mathcal{L}^q, \mathcal{D}^q)$ represent the skew-symmetric dispersion and dissipation effects from the right-hand side of (221). The additional operator $\mathcal{L}_{\mathbf{k}}(\bar{q})$ represents the interactions with a non-zero statistical mean state, where internal instability occurs with positive growth rate. Most importantly, the nonlinear interactions between different spectral modes introduce the additional nonlinear flux term Q^q_F indicating higher-order interactions, that is,

$$Q^q_F(\vec{q}_{\mathbf{k}}) = \frac{1}{2} \sum_{\mathbf{m}+\mathbf{n}=\mathbf{k}} \langle (A_{\mathbf{km}} \vec{q}_{\mathbf{m}} \circ \vec{q}_{\mathbf{n}} + A_{\mathbf{kn}} \vec{q}_{\mathbf{n}} \circ \vec{q}_{\mathbf{m}}) \vec{q}^*_{\mathbf{k}} \rangle. \tag{227}$$

Therefore, the small and large scale modes are linked through third-order moments $\langle \vec{q}_{\mathbf{m}} \vec{q}_{\mathbf{n}} \vec{q}^*_{\mathbf{k}} \rangle$ in (227) between the triad modes $\mathbf{m} + \mathbf{n} = \mathbf{k}$. The nonlinear flux Q^q_F plays the central role in the energy mechanism that balances the unstable directions due to internal instability from the linear operators. Here, our focus is on the low-order stochastic realization in (224) of the statistical closure model of (226), thus solving the statistical Equation (226) directly is not favorable considering its complexity.

Below, we follow the general framework developed in Section 6.1 to determine the reduced order model. The nonlinear flux Q^q_F in (227) corresponds to the unresolved nonlinear effects in the stochastic model in (224). Thus, it is useful to exploit the nonlinear flux Q^q_F so that the imperfect model parameters (D^M_q, Σ^M_q) in (224) can be proposed according to the true model energy transfer mechanism. Especially in statistical equilibrium, as $t \to \infty$ the nonlinear fluxes can be calculated easily from the localized lower-order moments

$$Q^q_{F,eq} = (\mathcal{L}^q_{\mathbf{k}} + \mathcal{D}^q_{\mathbf{k}} - \mathcal{L}_{\mathbf{k}}(\bar{q}_{eq})) R^q_{\mathbf{k},eq} + R^q_{\mathbf{k},eq} (\mathcal{L}^q_{\mathbf{k}} + \mathcal{D}^q_{\mathbf{k}} - \mathcal{L}_{\mathbf{k}}(\bar{q}_{eq}))^*. \tag{228}$$

Next, we further decompose the matrix $Q^q_F = Q^{q,+}_F + Q^{q,-}_F$ by singular value decomposition into positive-definite and negative-definite components. The positive definite part $Q^{q,+}_F$ illustrates the additional energy that injected into this mode from other scales, while the negative definite part $Q^{q,-}_F$ shows the extraction of energy through nonlinear transfer to other scales. In adopting the true equilibrium statistics from $Q^q_{F,eq}$, the true model energy transfer mechanism is respected and the least

artificial effect is introduced into the imperfect approximation model. Considering all these aspects, the first proposal for the linear damping and Gaussian random noise correction can be introduced as

$$D_{q,\mathbf{k}}^{eq} = -\frac{1}{2}Q_{F,eq,\mathbf{k}}^{q,-}(R_{\mathbf{k},eq}^{q})^{-1}, \qquad \Sigma_{q,\mathbf{k}}^{eq} = (Q_{F,eq,\mathbf{k}}^{q,+})^{1/2}. \tag{229}$$

The additional damping is from the negative definite equilibrium flux $Q_{F,eq}^{q,-}$ and the positive definite equilibrium flux $Q_{F,eq}^{q,+}$ acts as additional noise to the system. The above additional damping and noise (229) offer a desirable quantification for the minimum amount of corrections to stabilize the system with consistent equilibrium statistics for the mean and variance. This is the same idea applied to the statistical modified quasi-linear Gaussian closures developed in [45].

As discussed in Section 6.1.3, the above estimation of parameters (229) may not be optimal for the reduced-order Gaussian model considering that: (i) it only guarantees marginal stability in the unstable modes for equilibrium; and more importantly (ii) the time mixing scale in each mode (represented by the autocorrelation functions) may still lack the accuracy in the approximation using only equilibrium information. The nonlinear energy transferring mechanism may change with large deviation from the equilibrium case when intermittent fluctuations are present. The shortcomings for purely using the approximation (229) only from equilibrium statistics can be observed from the various numerical simulations [82]. As a further correction, we propose additional terms on top of (229) with a simple constant damping for all the spectral modes and an additional noise accordingly to make sure consistency in energy,

$$Q_{M,\mathbf{k}}^{add} = -D_M^{add}R_{M,\mathbf{k}} + \left(\Sigma_{M,\mathbf{k}}^{add}\right)^2, \qquad D_M^{add} = diag\{d_M + i\omega_M, d_M - i\omega_M\}. \tag{230}$$

The correction term in (230) is aimed to offer stabilizing effects in the marginal stable equilibrium form (229), and to offer further corrections in modeling the autocorrelation function that is important for the mixing rate in each spectral mode. Combining (229) and (230), the additional damping and noise corrections for the reduced-order flow vorticity model (224) are given in the following form

$$D_{q,\mathbf{k}}^{M} = -\frac{1}{2}Q_{F,eq,\mathbf{k}}^{q,-}\left(R_{\mathbf{k},eq}^{q}\right)^{-1}, \qquad \Sigma_{q,\mathbf{k}}^{M} = \left(Q_{F,eq,\mathbf{k}}^{q,+} + \left(\Sigma_{M,\mathbf{k}}^{add}\right)^2\right)^{1/2}. \tag{231}$$

Comparing with the exact true system (221), the reduced-order approximation is equivalent to replacing the nonlinear interaction terms with the judiciously calibrated damping and noise in consideration with both the equilibrium energy transfer mechanism and further sensitivity correction.

Now, we move to the second step. Here, we tune the undermined model parameters $(D_M^{add}, \Sigma_M^{add})$ to guarantee the imperfect stochastic model consistency in equilibrium statistics (the leading two moments) and autocorrelation functions. The procedure here is exactly the same as that in Section 6.3.1, where information theory developed in Section 2.5 is used for calibrating the autocorrelation function. Thus, we neglect the details here.

Finally, let us show a simple example for predicting the tracer statistics using the low-order model prediction. The example here has the same setup as one of the regimes considered in [81], that is, the high latitude atmosphere regime, where the parameters are given as follows:

$$\begin{aligned} N = 128, \quad &\beta = 1, \quad F = 4, \quad U = 0.1, \quad r = 0.2, \\ \nu = 10^{-13}, \quad &s = 4, \quad d_T = 0.1 \quad \kappa = 10^{-3}, \quad \alpha = 1. \end{aligned} \tag{232}$$

Here, $N = 128$ is the number of grid points in each direction. The true statistics are calculated by a pseudo-spectra code with $128 \times 128 \times 2$ grid points in total. The zonal mean flow $\vec{U} = (U, -U)$ is taken as the same strength with opposite directions in the two layers. In the tracer simulations, for simplicity, we consider the mean gradient along y direction, that is to assume, $T = T' + \alpha y$. This assumption is

representative in many previous investigations [117,193,195]. The scale separation parameter ϵ in this example is chosen to be $\epsilon^{-1} = 5$ such that intermittency is prominent. In the reduced-order model, we only compute the modes $\mathbf{k} \leq M = 10$ in largest scales, compared with the true system resolution $N = 128$.

The autocorrelation functions of the first four leading modes $(1,0), (0,1), (1,1)$ and $(-1,1)$ in both the flow stream functions and the tracer fields are plotted in Figure 42. It is clear that for both the flow and tracer fields, the reduced order model with optimized parameters calibrated by information theory succeeds in capturing the autocorrelation function of the truth. As a comparison, equipped with the parameters with no additional corrections $d_M = 0, \sigma_M = 0$, the reduced order model has a huge bias in recovering the autocorrelation function of the flow field. Next, we test the prediction skill of the turbulent tracer statistics in the reduced-order model. As we have seen in Section 6.3.1 and the discussions above, the nonlinear advection in the tracer equation $\mathbf{v}_M \cdot \nabla T_M$ is important for the final tracer statistical structure. The goal here is to see whether the intermittent and other features in the tracer field can be accurately predicted using only principal modes with largest variance in $\mathbf{v}_M$ in calculating the nonlinear term. Figure 43 compares the representative time-series and tracer PDFs of the leading modes in statistical steady state. Despite only 0.6% of the modes being involved in the flow field, the fat-tails in the distribution functions of the tracer can be captured, and similar characteristic structures can be seen in the truth and reduced model time series. In fact, the high skill of recovering the non-Gaussian features is due to the fact that the advection term $\mathbf{v}_M \cdot \nabla T_M$ is captured quite well even with such a crude truncation of the flow field. The results in Figures 42 and 43 imply the skillful predictions using the reduced order model with the optimized parameters. In [82], the recovered tracer field using different truncation size M has been explored. It is important to note that with only the first two modes $M = 2$ being included in calculating the nonlinear advection, larger errors appear due to the insufficient quantification for flow advection. The recovering skill of other statistical features such as the power spectrum and eddy diffusivity approximations for the tracers in this regime as well as the test examples in other regimes have also been systematically discussed in [82].
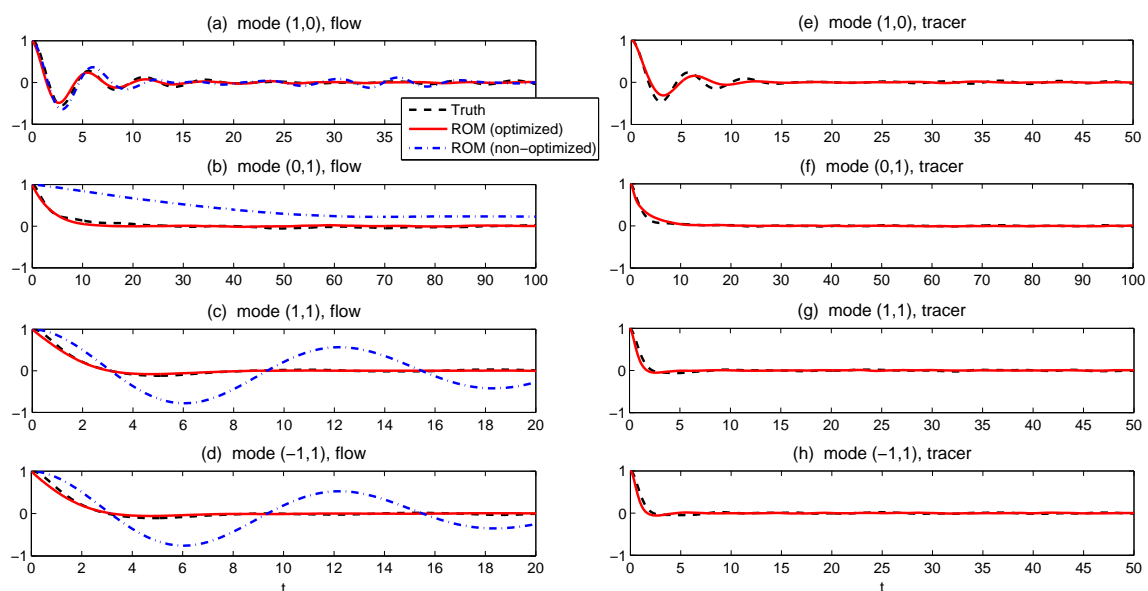

Figure 42. Prediction of the reduced-order autocorrelation functions in the flow (**a–d**) and tracer fields (**e–h**) in the first four most energetic modes in high latitude atmosphere regime with parameters $(\epsilon^{-1}, d_T) = (5, 0.1)$. The truth is shown in black dashed lines and the reduced-order model (ROM) prediction in red lines.
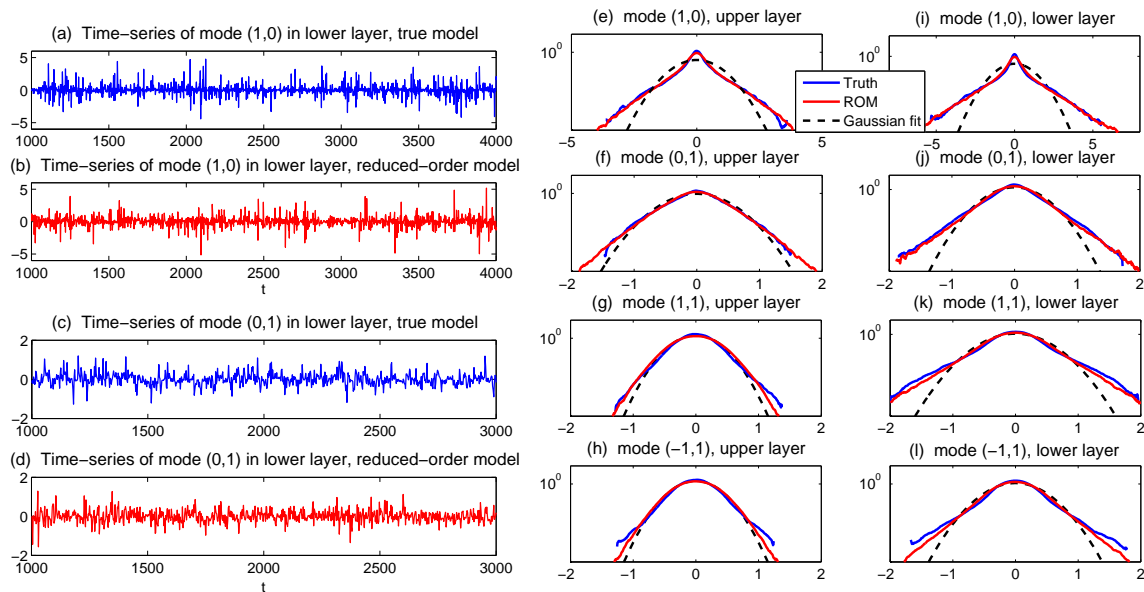
Figure 43. Prediction of tracer intermittency in high latitude atmosphere regime with parameters $(\epsilon^{-1}, d_T) = (5, 0.1)$. (**a**–**d**): the time-series for the first two leading modes $(1, 0)$ and $(0, 1)$ between true model and reduced-order model (ROM) results; (**e**–**l**): comparison of the PDFs in the first four modes between the truth in blue and reduced model prediction in red with the Gaussian fit in dashed black lines.

7. Conclusions

This research expository article discusses various important topics related to model error, information barriers, state estimation and prediction in complex multiscale systems. A recent information-theoretic framework is developed and summarized, which is applied together with other mathematical tools to study all these topics. It is also combined with novel reduced-order nonlinear modeling strategies for understanding and predicting complex multiscale systems. The contents of this article include the general mathematical framework and theory, effective numerical procedures, instructive qualitative models, and concrete models from climate, atmosphere and ocean science. The information-theoretic framework is developed in Section 2 and is applied to understand various information barriers in the presence of model error via instructive stochastic models. In Section 3, the information-theoretic framework is adopted to assess model error in state estimation and prediction with examples coming from both complex scalar models and spatially-extended multiscale turbulent systems. The advantage of the information-theoretic framework over the traditional path-wise measurements are illustrated. Section 4 deals with sensitivity and linear statistical response using the fluctuation–dissipation theorem. An efficient and effective algorithm in finding the most sensitive change directions using information theory is also included in this section. In Section 5, a novel framework of data-driven physics-constrained nonlinear stochastic models and predictions is developed and is applied to predicting the MJO which contains strong intermittent instabilities and extreme events. Section 6 includes the development of the new effective reduced-order models that involve higher order statistical features but nevertheless remain computationally efficient. These new models together with the information-optimization model calibration strategy are applied to predicting passive tracers extreme events.

The simple imperfect models used in Sections 2 and 4 are all motivated from the strategies that are commonly used in practice for approximating extremely complicated systems such as GCMs. Therefore, the information barriers shown in this article clearly indicate the deficiency of these strategies and point out the directions of improving the imperfect models. The computationally efficient reduced-order modeling framework developed in Section 6 is promising in dealing with

many complicated real-world issues. In particular, including higher order statistical features using the novel approach allows these new reduced-order models to capture the nonlinear evolution and non-Gaussian characteristics in both the dynamics and statistics. Therefore, these models are able to overcome those information barriers resulting from the linear tangential or Gaussian closure approximations as well as the ignorance of the cross-correlations between different modes or grid points. In addition to studying the spatially-extended systems associated with predicting passive tracers extreme events, the other applications of these low-order modeling strategies are good future directions. Note that these low-order modeling strategies combined with FDT can also be powerful tools to study the effective statistical control of complex turbulent dynamical systems [206]. On the other hand, although great efforts have been put in understanding the sources of model errors in data assimilation (or filtering), the representation error was nevertheless overlooked in the past. In Section 3, the issue of representation error is emphasized and some practical strategies have been proposed and tested. More systematic studies are required in this area as future works. It is also of great importance to study filtering and prediction as a whole and understand the model error and improved strategies for both procedures instead of focusing solely on the filtering part. In addition, as is briefly discussed in Section 3.5.2, combining the Euler and Lagrangian observations is another interesting topic in improving the skill of data assimilation and prediction of spatially-extended systems as well as quantifying the uncertainty reduction. Finally, it has been shown in Section 5 that the data-driven physics-constrained nonlinear stochastic modeling framework has several salient advantages over the purely data-driven non-parametric methods in terms of both understanding the underlying physics and obtaining skillful predictions. These advantages include a much shorter training phase, a systematic calibration strategy, gaining clear physical insights and reaching model robustness. Applying the data-driven physics-constrained nonlinear stochastic modeling framework to many other complex real-world problems is potentially important. Many related issues remain as future work.

Author Contributions: Conceptualization, A.J.M. and N.C.; Data curation, A.J.M. and N.C.; Formal analysis, A.J.M. and N.C.; Funding acquisition, A.J.M.; Methodology, A.J.M. and N.C.; Visualization, N.C.; Writing—Original draft, N.C.; Writing—Review and editing, A.M.

Appendix A. Derivations of Fisher Information from Relative Entropy

Here, we include the details of the derivations of Fisher information (149) and (150) from the relative entropy. Let's consider the two PDFs $\pi_{\boldsymbol{\theta}}(u)$ and $\pi_{\boldsymbol{\theta}'}(u)$ with $\boldsymbol{\theta}' = \boldsymbol{\theta} + \delta\boldsymbol{\theta}$ and $\delta\boldsymbol{\theta}$ is a small increment. Applying the Talyor's expansion to $\pi_{\boldsymbol{\theta}}$ and $\ln \pi_{\boldsymbol{\theta}}$ yields

$$
\begin{aligned}
\pi_{\boldsymbol{\theta}'} &= \pi_{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}\delta\boldsymbol{\theta} + \frac{1}{2}\delta\boldsymbol{\theta}^T \nabla_{\boldsymbol{\theta}}^2 \pi_{\boldsymbol{\theta}}\delta\boldsymbol{\theta} + O(\delta\boldsymbol{\theta}^3), \\
\ln \pi_{\boldsymbol{\theta}'} &= \ln \pi_{\boldsymbol{\theta}} + \frac{1}{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}\delta\boldsymbol{\theta} + \frac{1}{2}\delta\boldsymbol{\theta}^T\left(-\frac{1}{\pi_{\boldsymbol{\theta}}^2}(\nabla_{\boldsymbol{\theta}})^2 + \frac{1}{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}^2\pi_{\boldsymbol{\theta}}\right)\delta\boldsymbol{\theta} + O(\delta\boldsymbol{\theta}^3).
\end{aligned} \tag{A1}
$$

With (A1) in hand, now we compute the relative entropy in (149):

$$
\begin{aligned}
\mathcal{P}(\pi_{\theta'}, \pi_{\theta}) &= \int \pi_{\theta'} \ln \frac{\pi_{\theta'}}{\pi_{\theta}} \\
&= \int \pi_{\theta'} \ln \pi_{\theta'} - \int \pi_{\theta'} \ln \pi_{\theta} \\
&= \int \left(\pi_{\theta} + \nabla_{\theta}\pi_{\theta}\delta\theta + \frac{1}{2}\delta\theta^T \nabla_{\theta}^2 \pi_{\theta}\delta\theta \right) \left(\ln \pi_{\theta} + \frac{1}{\pi_{\theta}}\nabla_{\theta}\pi_{\theta}\delta\theta + \frac{1}{2}\delta\theta^T \left(-\frac{1}{\pi_{\theta}^2}(\nabla_{\theta})^2 + \frac{1}{\pi_{\theta}}\nabla_{\theta}^2\pi_{\theta} \right) \delta\theta \right) \\
&\quad - \int \left(\pi_{\theta} + \nabla_{\theta}\pi_{\theta}\delta\theta + \frac{1}{2}\delta\theta^T \nabla_{\theta}^2 \pi_{\theta}\delta\theta \right) \ln \pi_{\theta} + O(\delta\theta^3) \\
&= \int \pi_{\theta}\ln\pi_{\theta} + \nabla_{\theta}\pi_{\theta}\delta\theta + \frac{1}{2}\pi_{\theta}\delta\theta^T \left(-\frac{1}{\pi_{\theta}^2}(\nabla_{\theta})^2 + \frac{1}{\pi_{\theta}}\nabla_{\theta}^2\pi_{\theta} \right)\delta\theta + \nabla_{\theta}\pi_{\theta}\ln\pi_{\theta}\delta\theta + \delta\theta^T \frac{1}{\pi_{\theta}}(\nabla_{\theta}\pi_{\theta})^2\delta\theta \quad \text{(A2)} \\
&\quad + \frac{1}{2}\ln\pi_{\theta}\delta\theta^T\nabla_{\theta}^2\pi_{\theta}\delta\theta - \pi_{\theta}\ln\pi_{\theta} - \nabla_{\theta}\pi_{\theta}\ln\pi_{\theta}\delta\theta - \frac{1}{2}\ln\pi_{\theta}\delta\theta^T\nabla_{\theta}^2\pi_{\theta}\delta\theta + O(\delta\theta^3) \\
&= \nabla_{\theta}\int\pi_{\theta}\delta\theta + \frac{1}{2}\delta\theta^T \left(\int \frac{1}{\pi_{\theta}}(\nabla_{\theta}\pi_{\theta})^2 \right)\delta\theta + \frac{1}{2}\delta\theta^T \left(\nabla_{\theta}^2 \int\pi_{\theta} \right)\delta\theta + O(\delta\theta^3) \\
&= \frac{1}{2}\delta\theta^T \cdot \left(\int \frac{1}{\pi_{\theta}}(\nabla_{\theta}\pi_{\theta})^2 \right) \cdot \delta\theta + O(\delta\theta^3) \\
&= \frac{1}{2} \left(\int \frac{(\delta\theta \cdot \nabla_{\theta}\pi_{\theta})^2}{\pi_{\theta}} \right) + O(\delta\theta^3),
\end{aligned}
$$

where we have made use of the fact that $\int \pi_{\theta} \equiv 1$ and therefore $\nabla_{\theta} \int \pi_{\theta} = 0$. Some regularity assumptions are also required [207] such that the integral and gradient operator can be interchanged. Clearly, the final result is the Fisher information. Note that $\delta\theta$ here is λ in (150).

Appendix B. Details of the Canonical Model for Low Frequency Atmospheric Variability

Here, we provide more details of the canonical model for low frequency atmospheric variability [116,170] with cubic nonlinearity and correlated additive and multiplicative (CAM) noise, which has been used in Sections 4.3 and 6.2.

The model reads

$$
\frac{dx}{dt} = (f + ax + bx^2 - cx^3) + (A - Bx)\dot{W}_C + \sigma\dot{W}_A. \tag{A3}
$$

The Fokker–Planck equation for the evolution of the PDF is given by

$$
\frac{\partial\pi}{\partial t} = -\frac{\partial}{\partial x}\left[(f + ax + bx^2 - cx^3)\pi \right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[((Bx - A)^2 + \sigma^2)\pi \right].
$$

For the case of nonzero correlated additive and multiplicative (CAM) noise, i.e., $A \neq 0$ and $B \neq 0$, we find the following equilibrium PDF:

$$
\pi(x) = \frac{N_0}{((Bx - A)^2 + \sigma^2)^{a_1}} e^{d \arctan\left(\frac{Bx - A}{\sigma}\right)} e^{\frac{-c_1 x^2 + b_1 x}{B^4}}, \tag{A4}
$$

where N_0 is a normalizing constant to make $\pi(x)$ integrate to one, and the new parameters can be computed via

$$
a_1 = 1 - \frac{-3A^2 c + aB^2 + 2AbB + c\sigma^2}{B^4}, \qquad b_1 = 2bB^2 - 4cAB, \qquad c1 = cB^2,
$$

$$
d = \frac{d_1}{\sigma} + d_2\sigma, \qquad d_1 = 2\frac{A^2 bB - A^3 c + AaB^2 + B^3 f}{B^4}, \qquad d_2 = \frac{6cA - 2bB}{B^4}.
$$

On the other hand, in a special case of additive noise only, i.e., $A = B = 0$, we find the following invariant PDF

$$
\pi(x) = N_0 \exp\left(\frac{2}{\sigma^2}\left(fx + \frac{a}{2}x^2 + \frac{b}{3}x^3 - \frac{c}{4}x^4 \right) \right). \tag{A5}
$$

Appendix C. Augmented System for Prediction and Filtering Distributions

Here, we show the details of using augmented systems for studying the prediction/filtering state estimates compared to the truth as discussed in Section 3.2.

Appendix C.1. Augmented System for Prediction

In light of the truth (67) and the prediction mean (69), the coupled evolution of the augmented state $\mathcal{X}_m := (u_m, \bar{u}_{m+1|m})^T$ is given by

$$\begin{pmatrix} u_{m+1} \\ \bar{u}_{m+1|m} \end{pmatrix} = \begin{pmatrix} F & 0 \\ K_m^M g F^M & (1 - K_m^M g) F^M \end{pmatrix} \begin{pmatrix} u_m \\ \bar{u}_{m|m-1} \end{pmatrix} + \begin{pmatrix} \sigma_{m+1} \\ K_m^M F^M \sigma_m^o \end{pmatrix} + \begin{pmatrix} \mathcal{F}_{m+1} \\ \mathcal{F}_{m+1}^M \end{pmatrix}. \quad (A6)$$

The system in (A6) is a Gaussian system and therefore its behavior is completely characterized by its mean and variance. The evolution of the mean of (A6) is given by

$$\mathbb{E}(\mathcal{X}_m) = \begin{pmatrix} F & 0 \\ K_m^M g F^M & (1 - K_m^M g) F^M \end{pmatrix} \mathbb{E}(\mathcal{X}_{m-1}) + \begin{pmatrix} \mathcal{F}_{m+1} \\ \mathcal{F}_{m+1}^M \end{pmatrix}. \quad (A7)$$

With the equilibrium mean of the perfect and imperfect model $\bar{u}_{eq} = \frac{\mathcal{F}_\infty}{1-F}$ and $\bar{u}_{eq}^M = \frac{\mathcal{F}_\infty^M}{1-F^M}$, the asymptotic mean of $\mathcal{X}_m$ is given by

$$\mathbb{E}(\mathcal{X}_\infty) = \lim_{m \to \infty} \mathbb{E}(\mathcal{X}_m) = \begin{pmatrix} \bar{u}_{eq} \\ \frac{1}{(1 - F^M) + F^M K_\infty^M g} (K_\infty^M g F^M \bar{u}_{eq} + (1 - F^M)\bar{u}_{eq}^M) \end{pmatrix}, \quad (A8)$$

where the asymptotic Kalman gain K_∞^M is given by

$$K_\infty^M = \frac{1}{2g} \left(1 - \frac{g^2 r^M}{|F^M|^2 r^o} - \frac{1}{|F^M|^2} + \left[\left(1 - \frac{g^2 r^M}{|F^M|^2 r^o} - \frac{1}{|F^M|^2} \right)^2 + \frac{4g^2 r^M}{|F^M|^2 r^o} \right]^{1/2} \right).$$

Clearly, the asymptotic mean of the prediction state is a linear combination of the equilibrium mean of original true model $\bar{u}_{eq}$ and that of the forecast model of the mean $\bar{u}_{eq}^M$. With (A8), the mean bias yields

$$\lim_{m \to \infty} \mathbb{E}(u_{m+1} - \bar{u}_{m+1|m}) = \frac{1 - F^M}{(1 - F^M) + F^M K_\infty^M K_\infty^M g} (\bar{u}_{eq} - \bar{u}_{eq}^M). \quad (A9)$$

According to (A9), the asymptotic prediction mean is equal to the equilibrium mean of the perfect model if and only if the imperfect model has the same equilibrium mean as the perfect model, namely $\bar{u}_{eq} = \bar{u}_{eq}^M$.

Next, we derive the covariance of the augmented system. Denote the operators $\mathcal{F}^P$ and $\mathcal{R}^P$ as

$$\mathcal{F}^P = \begin{pmatrix} F & 0 \\ K_m^M g F^M & (1 - K_m^M g) F^M \end{pmatrix}, \qquad \mathcal{R}^P = \begin{pmatrix} r & 0 \\ 0 & (K_m^M)^2 |F^M|^2 r^o \end{pmatrix}. \quad (A10)$$

Denote the covariance of $\mathcal{X}_m$ by $C_m^P = Cov(\mathcal{X}_m, \mathcal{X}_m)$, where

$$\mathcal{C}_m^P = \begin{pmatrix} Cov(u_m, u_m) & Cov(u_m, \bar{u}_{m|m-1}) \\ Cov(\bar{u}_{m|m-1}, u_m) & Cov(\bar{u}_{m|m-1}, \bar{u}_{m|m-1}) \end{pmatrix} \equiv \begin{pmatrix} C_{(11)m}^P & C_{(12)m}^P \\ C_{(21)m}^P & C_{(22)m}^P \end{pmatrix}. \quad (A11)$$

The evolution of the covariance matrix of the augmented system is given by

$$C_{m+1}^P = \mathcal{F}_m^P \mathcal{C}_m^P \mathcal{F}_m^{P*} + \mathcal{R}_m^P, \quad (A12)$$

and the components of the asymptotic limit $\mathcal{C}^P_\infty$ are

$$C^P_{(11)\infty} = \frac{r}{1 - |F|^2}, \qquad C^P_{(12)\infty} = \frac{FF^{M*}K^M_\infty g C^P_{(11)\infty}}{1 - FF^{M*}(1 - K^M_\infty g)}, \qquad C^P_{(21)\infty} = C^P_{(12)\infty},$$

$$C^P_{(22)\infty} = \frac{|F^M|^2}{1 - |F^M|^2(1 - K^M_\infty g)^2}\left((g^2(K^M_\infty)C^P_{(11)\infty} + 2K^M_\infty g(1 - K^M_\infty g)Re(C^P_{(12)\infty})) + (K^M_\infty)^2 r^o\right).$$

$$\text{(A13)}$$

With direct calculation, the dynamics of the prediction variance $r_{m+1|m}$ is

$$r_{m+1|m} = |F^M|^2 r_{m|m} + r^M = |F^M|^2(1 - K^M_m g)r_{m|m-1} + r^M. \tag{A14}$$

Asymptotically, it is given by

$$r^P_\infty = |F^M|^2(1 - K^M_\infty g)r^P_\infty + r^M = |F^M|^2\frac{r^o r^P_\infty}{g^2 r^P_\infty + r^o} + r^M = \frac{|F^M|^2 r^o}{g}K^M_\infty + r^M. \tag{A15}$$

Appendix C.2. Augmented System for Filtering

The coupled evolution of the augmented state $\mathcal{Y}_m := (u_m, \bar{u}_{m|m})^T$ is

$$\begin{pmatrix} u_{m+1} \\ \bar{u}_{m+1|m+1} \end{pmatrix} = \begin{pmatrix} F & 0 \\ K^M_{m+1}gF & (1 - K^M_{m+1}g)F^M \end{pmatrix} \begin{pmatrix} u_m \\ \bar{u}_{m|m} \end{pmatrix} + \begin{pmatrix} \sigma_{m+1} \\ K^M_{m+1}(g\sigma_{m+1} + \sigma^o_{m+1}) \end{pmatrix}$$
$$+ \begin{pmatrix} \mathcal{F}_{m+1} \\ (1 - K^M_{m+1}g)\mathcal{F}^M_{m+1} + K^M_{m+1}g\mathcal{F}_{m+1} \end{pmatrix}. \tag{A16}$$

Again, the Gaussian statistics of the augmented state $\mathcal{Y}_m$ is fully characterized by its mean and covariance. The evolution of its mean is given by

$$\mathbb{E}(\mathcal{Y}_m) = \begin{pmatrix} F & 0 \\ K^M_{m+1}gF & (1 - K^M_{m+1}g)F^M \end{pmatrix} \mathbb{E}(\mathcal{Y}_{m-1}) + \begin{pmatrix} \mathcal{F}_{m+1} \\ (1 - K^M_{m+1}g)\mathcal{F}^M_{m+1} + K^M_{m+1}g\mathcal{F}_{m+1} \end{pmatrix}. \tag{A17}$$

When $m \to \infty$, the asymptotic mean of $\mathcal{Y}_m$ is

$$\mathbb{E}(\mathcal{Y}_\infty) = \lim_{m\to\infty}\mathbb{E}(\mathcal{Y}_m) = \begin{pmatrix} \bar{u}_{eq} \\ \frac{1}{(1-F^M)+F^M K^M_\infty g}(K^M_\infty g\bar{u}_{eq} + (1 - F^M)(1 - K^M_\infty g)\bar{u}^M_{eq}) \end{pmatrix}, \tag{A18}$$

and the deviation of analysis mean from the truth signal is therefore given as follows:

$$\lim_{m\to\infty}\mathbb{E}(u_{m+1} - \bar{u}_{m+1|m+1}) = \frac{(1 - F^M)(1 - K^M_\infty g)}{(1 - F^M) + F^M K^M_\infty g}(\bar{u}_{eq} - \bar{u}^M_{eq}). \tag{A19}$$

Next, the covariance of prediction mean is

$$\mathcal{C}^A_{m+1} = \mathcal{F}^A_m \mathcal{C}^A_m \mathcal{F}^{A*}_m + \mathcal{R}^A_m,$$

where the operator $\mathcal{F}^A$ and $\mathcal{R}^A$ are given respectively by

$$\mathcal{F}^A_m = \begin{pmatrix} F & 0 \\ K^M_{m+1}gF & (1 - K^M_{m+1}g)F^M \end{pmatrix}, \qquad \mathcal{R}^A_m = \begin{pmatrix} r & rgK^M_{m+1} \\ rgK^M_{m+1} & (K^M_{m+1})^2(r^o + g^2 r) \end{pmatrix}.$$

The elements of covariance matrix are

$$
\mathcal{C}_m^A = \begin{pmatrix} Cov(u_m, u_m) & Cov(u_m, \bar{u}_{m|m-1}) \\ Cov(\bar{u}_{m|m-1}, u_m) & Cov(\bar{u}_{m|m-1}, \bar{u}_{m|m-1}) \end{pmatrix} \equiv \begin{pmatrix} C_{(11)m}^A & C_{(12)m}^A \\ C_{(21)m}^A & C_{(22)m}^A \end{pmatrix},
\tag{A20}
$$

with the corresponding components of the asymptotic covariance $\mathcal{C}_\infty^A$,

$$
\mathcal{C}_{(11)\infty}^A = \frac{r}{1 - |F|^2}, \qquad \mathcal{C}_{(12)\infty}^A = \frac{K_\infty^M g(|F|^2 \mathcal{C}_{(11)\infty}^A + r)}{1 - FF^{M*}(1 - K_\infty^M g)}, \qquad \mathcal{C}_{(21)\infty}^A = \mathcal{C}_{(12)\infty}^{A*},
$$
$$
\mathcal{C}_{(22)\infty}^A = \frac{g^2(K_\infty^M)^2 |F|^2 \mathcal{C}_{(11)\infty}^A + 2K_\infty^M g(1 - K_\infty^M g)Re(FF^M \mathcal{C}_{(12)\infty}^A) + (K_\infty^M)^2(r^o + g^2 r)}{1 - |F^M|^2(1 - K_\infty^M g)^2}.
\tag{A21}
$$

Thus, direct calculations result in the dynamics of prediction variance $r_{m+1|m+1}$,

$$
r_{m+1|m+1} = (1 - K_{m+1}^M g)(|F^M|^2 r_{m|m} + r^M).
\tag{A22}
$$

As discussed in [15], the asymptotic analysis variance is

$$
r_\infty^A = \frac{r^o}{g} K_\infty^M.
\tag{A23}
$$

Finally, we compare the asymptotic prediction and filtering variance. We have the following conclusion:

$$
r_\infty^A - r_\infty^P = \frac{r^o}{g} K_\infty^M - \frac{|F^M|^2 r^o}{g} K_\infty^M - r^M = (1 - |F^M|^2) r_\infty^P - r^M < 0.
\tag{A24}
$$

The details are as follows. From [15], r_∞^P satisfies the equation

$$
(r_\infty^P)^2 + \left(\frac{r^M}{|F^M|^2} + \frac{r^o}{g^2|F^M|^2} - \frac{r^o}{g^2} \right) r_\infty^P - \frac{r^o r^M}{g^2 |F^M|^2} = 0,
\tag{A25}
$$

which has a positive and negative solution. For the equilibrium imperfect model variance, we have

$$
\left(\frac{r^M}{1 - |F^M|^2} \right)^2 + \left(\frac{r^M}{|F^M|^2} + \frac{r^o}{g^2|F^M|^2} - \frac{r^o}{g^2} \right) \frac{r^M}{1 - |F^M|^2} - \frac{r^o r^M}{g^2|F^M|^2}
$$
$$
= \frac{(r^M)^2}{1 - |F^M|^2} \left(\frac{1}{1 - |F^M|^2} + \frac{1}{|F^M|^2} \right) > 0.
\tag{A26}
$$

Hence, the equilibrium imperfect model variance is larger than the asymptotic filtering variance. We have

$$
r_\infty^A - r_\infty^P = (1 - |F^M|^2) r_\infty^P - r^M < 0,
\tag{A27}
$$

which indicates the filtering estimate has smaller uncertainty (variance) than the prediction.

Appendix D. Possible Non-Gaussian PDFs of a Linear Model with Time-Periodic Forcing Based on the Sample Points in a Single Trajectory

Recall the complex scalar forced OU process in (97),

$$
\frac{du}{dt} = (-\gamma + i\omega_0)u + f_0 + f_1 e^{i\omega_1 t} + \sigma \dot{W},
\tag{A28}
$$

where the evolution of the statistics is given by (99). When t is sufficiently large, the effect of initial value decays to zero. Thus, at any time instant t at the attractor, the PDF $u(t)$ is Gaussian. However, the PDF computed by taking the time average for a long trajectory at the attractor based on may not be Gaussian

if the time-periodic forcing $f_1 \neq 0$. This can be easily seen in the examples in Figure 15. In other words, the system is not ergodic [108] when the time-periodic forcing is strong. Below, we provide a mathematical quantification of such behavior. We focus on the system at the attractor and ignore the contribution from the initial value. For simplicity, we also assume $f_0 = 0$ since this constant forcing only shifts the mean of the solution by a constant and won't affect the non-Gaussian behavior in the time-averaged PDF. Therefore, the solution of (A28), according to (98), reduces to

$$u(t) = \frac{f_1 e^{i\omega_1 t}}{\gamma + i(-\omega_0 + \omega_1)} + \sigma \int_{t_0}^t e^{(-\gamma + i\omega_0)(t-s)} dW(s). \tag{A29}$$

Clearly, the true signal $u(t)$ can be decomposed into two parts: the deterministic time-periodic mean state $\bar{u}(t)$ and the fluctuation around the time-periodic mean state $u'(t)$:

$$\bar{u}(t) = \frac{f_1 e^{i\omega_1 t}}{\gamma + i(-\omega_0 + \omega_1)},$$

$$u'(t) = \sigma \int_{t_0}^t e^{(-\gamma + i\omega_0)(t-s)} dW(s). \tag{A30}$$

For the simplicity of illustration, in the following, we only consider the real part of the true signal. Thus, $\bar{u}(t)$ and $u'(t)$ become

$$\bar{u}(t) = A \cos(\omega_1 t + \phi), \qquad \phi = \arg \frac{f_1}{\gamma + i(-\omega_0 + \omega_1)}, \qquad A = \left| \frac{f_1}{\gamma + i(-\omega_0 + \omega_1)} \right|,$$

$$u'(t) = \frac{\sigma}{\sqrt{2}} \int_{t_0}^t e^{(-\gamma + i\omega_0)(t-s)} dW(s). \tag{A31}$$

As a simple illustration, we show in Panel (a) of Figure A1 the signal $u(t)$ and its decomposition $\bar{u}(t)$ and $u'(t)$ within one period.

With such a mean-fluctuation decomposition, the PDF based on the samples of the long trajectory can be computed in the following way. Say the total length of the trajectory is NT, where $T = 2\pi/\omega_1$ is the period. Now, we use take n_o grid points with uniform increment within one period, namely $nT, nT + \Delta T, nT + 2\Delta T, \ldots, nT + (n_o - 1)\Delta T$, where $\Delta T = T/n_o$ and $n = 1, \ldots, N$. For different n and fixed i, the time-dependent mean $\bar{u}(nT + i\Delta T)$ is the same while $u'(nT + i\Delta T)$ is different due to the randomness in the fluctuation. It is known from (A31) that the collection of the n points $u'(nT + i\Delta T)$ with $n = 1, \ldots, N$ and fixed i satisfies a Gaussian distribution $\mathcal{N}(\mu_i, R_i)$ with

$$\mu_i = \bar{u}(nT + i\Delta T), \qquad \text{and} \qquad R_i = \langle u'(nT + i\Delta T) u'^*(nT + i\Delta T) \rangle = \frac{\sigma^2}{4\gamma}, \tag{A32}$$

where the variance is computed from the second equation of (A31) and it has no dependence on t. Therefore, the PDF of $u(t)$ is given by the summation of all the Gaussian distributions in (A32) for $i = 1, \ldots, n_o$ with $n_o \to \infty$. See Figure A2 for an illustration. Mathematically, this can be written as a convolution,

$$p(u) = \frac{1}{C} p_{\bar{u}} * p_{u'} = \frac{1}{C} \int_{-\infty}^{\infty} p_{\bar{u}}(v) p_{u'}(u - v) dv, \tag{A33}$$

with C a normalized constant to guarantee $\int_{-\infty}^{\infty} p(u) du = 1$. Here, $p_{u'}(u - t)$ is the Gaussian PDF with mean and variance given by (A32), namely,

$$p_{u'}(u - v) = \frac{1}{\sqrt{\pi \sigma^2/(2\gamma)}} \exp\left(-\frac{(u - v)^2}{\sigma^2/(2\gamma)}\right). \tag{A34}$$

To compute $p_{\bar{u}}(u)$, we refer to Figure A1. Since $\bar{u}$ is bounded by $-A$ and A, the support of $p_{\bar{u}}(u)$ is also within $[-A, A]$. Direct calculation of $p_{\bar{u}}(u)$ is difficult. Nevertheless, the cumulative distribution

function (CDF) can be used, which facilitates the derivation of $p_{\bar{u}}(u)$. Now, consider the interval $[-\phi/\omega_1, (2\pi - \phi)/\omega_1]$. In fact, for any $u \in [-A, A]$, the CDF is given by

$$\mathbb{P}(\bar{u} < u) = \begin{cases} 1 - \frac{2}{\omega_1} \arccos \frac{u}{A} / (\frac{2\pi}{\omega_1}) = 1 - \frac{1}{\pi} \arccos \frac{u}{A}, & |u| \leq A, \\ 0, & |u| > A. \end{cases} \tag{A35}$$

The CDF in (A35) is easily derived by making use of the fact that the samples are uniformly distributed in $t \in [-\phi/\omega_1, (2\pi - \phi)/\omega_1]$, which is of length $2\pi/\omega_1$. With the CDF (A35) in hand, the PDF $p_{\bar{u}}(t)$ is given by the derivative of the CDF. Therefore,

$$p_{\bar{u}} = \begin{cases} \frac{1}{A\pi} \frac{1}{\sqrt{1-(u/A)^2}}, & |u| \leq A, \\ 0, & |u| > A. \end{cases} \tag{A36}$$

Therefore, combining (A34) and (A36) leads to the calculation of (A33).

It is clear that the full PDF depends on the variation of the time-dependent mean $\bar{u}$, and the variance of the Gaussian PDF $p_{u'}$ for the fluctuation part. If the variation of $\bar{u}$ is much smaller than the variance of u', then the full PDF is nearly Gaussian. With the increase of the variation of $\bar{u}$ and a fixed variance of u', then the full PDF becomes bimodal. Thus, the ratio of the bandwidths associated with $p_{\bar{u}}$ and $p_{u'}$ can be used to quantify the Gaussianity of the full PDF. The bandwidth of $p_{\bar{u}}$ is

$$L_{\bar{u}} = 2A = 2 \left| \frac{f_1}{\gamma + i(-\omega_1 + \omega_0)} \right|. \tag{A37}$$

On the other hand, although there is no finite support of the Gaussian distribution $p_{u'}$, the three-sigma rule of thumb [208] is always used as an empirical bandwidth to quantify the "range" of the Gaussian distribution, where the three sigma here means the three standard deviation from the mean of the Gaussian distribution that covers 99.73% of the values of $p_{u'}$. Thus, the bandwidth of $p_{u'}$ is

$$L_{u'} = 2 \left(3\sqrt{\frac{\sigma^2}{4\gamma}} \right). \tag{A38}$$

The ratio of (A37) and (A38) is given by

$$r := \frac{L_{\bar{u}}}{L_{u'}} = \frac{\left| \frac{f_1}{\gamma + i(-\omega_1 + \omega_0)} \right|}{3\sqrt{\frac{\sigma^2}{4\gamma}}}. \tag{A39}$$

Figure A3 shows the full PDFs with different ratio r. In Panel (a), the time-periodic forcing f_1 increases from $f_1 = 0$ to $f_1 = 5$. When $f_1 = 0$, the PDF is Gaussian and the system is ergodic. When f_1 increases, $L_{\bar{u}}$ becomes larger while $L_{u'}$ is fixed. Note that, in this example, $\omega_1 = \omega_0$, which means the forcing is resonant and therefore the bandwidth $L_{\bar{u}}$ is sensitive to the change of forcing amplitude f_1. With $f_1 > 1$, the bimodality in the PDF becomes significant. In Panel (b), we fix $f_1 = 5, \omega_1 = 1$ and let ω_0 change. With the resonant forcing $\omega_0 = \omega_1 = 1$, the PDF is significantly bimodal but with a non-resonant forcing $\omega_0 = 3$, the PDF is nearly Gaussian.

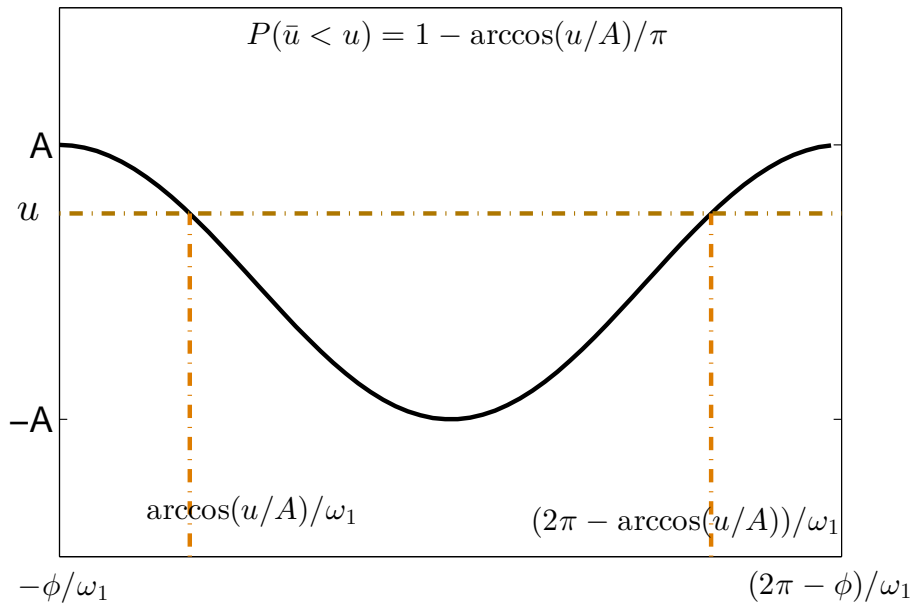## Illustration for the CDF for the time–dependent mean part

$$P(\bar{u} < u) = 1 - \arccos(u/A)/\pi$$

Figure A1. Illustration of computing the cumulative distribution function (CDF) of the deterministic mean $\bar{u}(t)$.
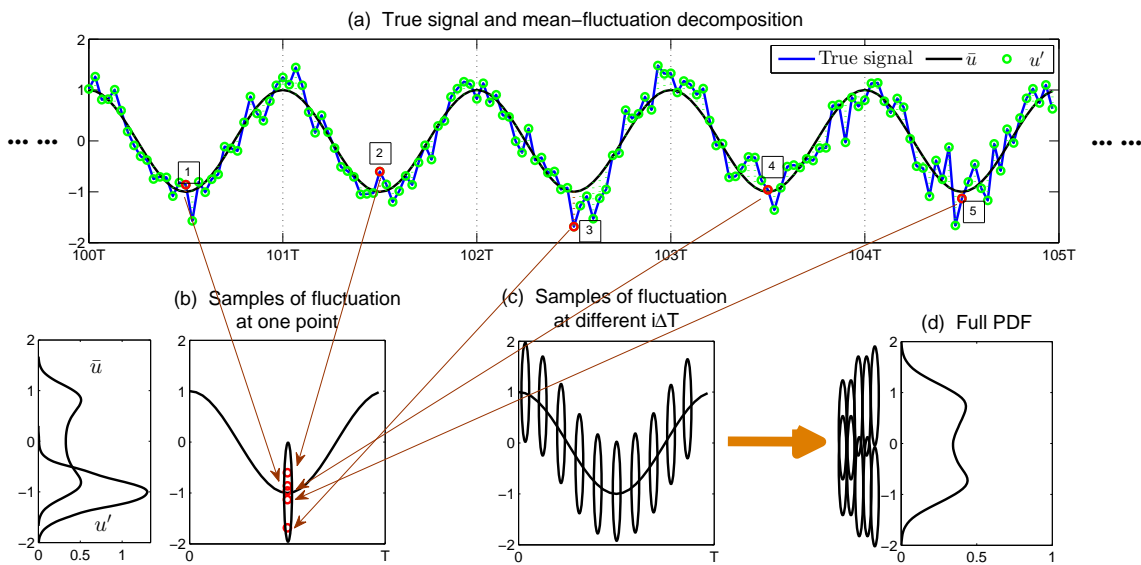
Figure A2. Illustration of the mean-fluctuation decomposition and the infinite Gaussian mixture PDF. (**a**): the true signal and its mean-fluctuation decomposition. Here, the signal within only five periods is shown; (**b**): the collection of the fluctuations at fixed time within each period, namely $nT + i\Delta T$ with i fixed and $n = 1, \ldots, N$, which gives a Gaussian PDF in (A32). The left subpanel shows the non-Gaussian time-averaged PDF of the deterministic mean $\bar{u}(t)$ and the Gaussian PDF of $u'(nT + i\Delta T)$ for a fixed i; (**c**): repeating (**b**) at different i; (**d**): the full PDF given by a Gaussian mixture from all the Gaussian PDFs in (**c**) with $\Delta T \to 0$.

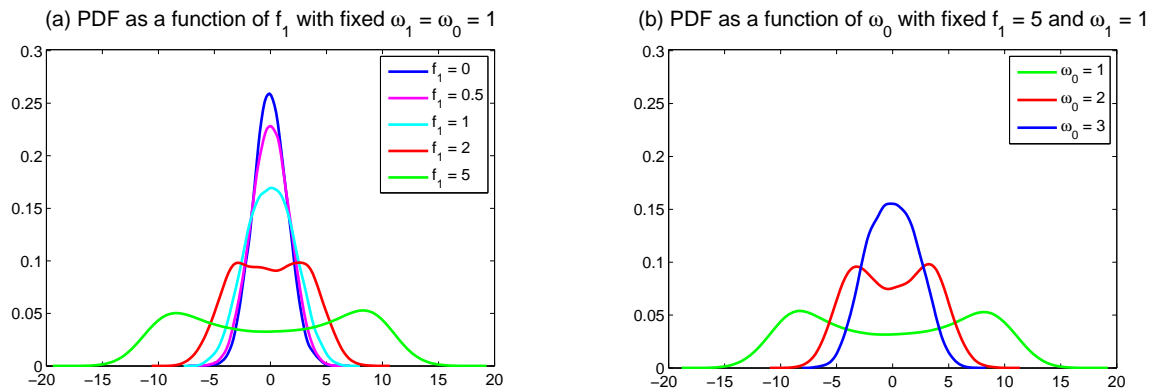**Figure A3.** The PDF of *u* averaged over a single long trajectory. (**a**): the PDF as a function of the amplitude of time-periodic forcing f_1 in a resonant forcing setup $\omega_1 = \omega_0 = 1$; (**b**): the PDF as a function of oscillation frequency ω_0 with fixed $f_1 = 5$ and $\omega_1 = 1$.

References

1. Majda, A.J. *Introduction to Turbulent Dynamical Systems in Complex Systems*; Springer: Berlin, Germany, 2016.
2. Majda, A.; Wang, X. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*; Cambridge University Press: Cambridge, UK, 2006.
3. Strogatz, S.H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*; CRC Press: Boca Raton, FL, USA, 2018.
4. Baleanu, D.; Machado, J.A.T.; Luo, A.C. *Fractional Dynamics and Control*; Springer Science & Business Media: Berlin, Germany, 2011.
5. Deisboeck, T.; Kresh, J.Y. *Complex Systems Science in Biomedicine*; Springer Science & Business Media: Berlin, Germany, 2007.
6. Stelling, J.; Kremling, A.; Ginkel, M.; Bettenbrock, K.; Gilles, E. *Foundations of Systems Biology*; MIT Press: Cambridge, MA, USA, 2001.
7. Sheard, S.A.; Mostashari, A. Principles of complex systems for systems engineering. *Syst. Eng.* **2009**, *12*, 295–311.
8. Majda, A. *Introduction to PDEs and Waves for the Atmosphere and Ocean*; American Mathematical Society: Providence, RI, USA, 2003; Volume 9.
9. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*; Springer Science & Business Media: Berlin, Germany, 2003; Volume 2.
10. Majda, A.J.; Branicki, M. Lessons in uncertainty quantification for turbulent dynamical systems. *Discret. Contin. Dyn. Syst. A* **2012**, *32*, 3133–3221.
11. Sapsis, T.P.; Majda, A.J. A statistically accurate modified quasilinear Gaussian closure for uncertainty quantification in turbulent dynamical systems. *Phys. D Nonlinear Phenom.* **2013**, *252*, 34–45.
12. Mignolet, M.P.; Soize, C. Stochastic reduced order models for uncertain geometrically nonlinear dynamical systems. *Comput. Methods Appl. Mech. Eng.* **2008**, *197*, 3951–3963.
13. Kalnay, E. *Atmospheric Modeling, Data Assimilation and Predictability*; Cambridge University Press: Cambridge, UK, 2003.
14. Lahoz, W.; Khattatov, B.; Ménard, R. Data assimilation and information. In *Data Assimilation*; Springer: Berlin, Germany, 2010; pp. 3–12.
15. Majda, A.J.; Harlim, J. *Filtering Complex Turbulent Systems*; Cambridge University Press: Cambridge, UK, 2012.
16. Evensen, G. *Data Assimilation: The Ensemble Kalman Filter*; Springer Science & Business Media: Berlin, Germany, 2009.
17. Law, K.; Stuart, A.; Zygalakis, K. *Data Assimilation: A Mathematical Introduction*; Springer: Berlin, Germany, 2015; Volume 62.
18. Palmer, T.N. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **2001**, *127*, 279–304.

19. Orrell, D.; Smith, L.; Barkmeijer, J.; Palmer, T. Model error in weather forecasting. *Nonlinear Process. Geophys.* **2001**, *8*, 357–371.

20. Hu, X.M.; Zhang, F.; Nielsen-Gammon, J.W. Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: A real-data study. *Geophys. Res. Lett.* **2010**, *37*, doi:10.1029/2010GL043017 .

21. Benner, P.; Gugercin, S.; Willcox, K. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.* **2015**, *57*, 483–531.

22. Majda, A.J. Challenges in climate science and contemporary applied mathematics. *Commun. Pure Appl. Math.* **2012**, *65*, 920–948.

23. Giannakis, D.; Majda, A.J. Quantifying the predictive skill in long-range forecasting. Part II: Model error in coarse-grained Markov models with application to ocean-circulation regimes. *J. Clim.* **2012**, *25*, 1814–1826.

24. Givon, D.; Kupferman, R.; Stuart, A. Extracting macroscopic dynamics: Model problems and algorithms. *Nonlinearity* **2004**, *17*, R55.

25. Trémolet, Y. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.* **2007**, *133*, 1267–1280.

26. Majda, A.J.; Gershgorin, B. Quantifying uncertainty in climate change science through empirical information theory. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 14958–14963.

27. Majda, A.J.; Gershgorin, B. Improving model fidelity and sensitivity for complex systems through empirical information theory. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10044–10049.

28. Majda, A.J.; Gershgorin, B. Link between statistical equilibrium fidelity and forecasting skill for complex systems with model error. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12599–12604.

29. Gershgorin, B.; Majda, A.J. Quantifying uncertainty for climate change and long-range forecasting scenarios with model errors. Part I: Gaussian models. *J. Clim.* **2012**, *25*, 4523–4548.

30. Branicki, M.; Majda, A.J. Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency. *Nonlinearity* **2012**, *25*, 2543.

31. Branicki, M.; Majda, A. Quantifying Bayesian filter performance for turbulent dynamical systems through information theory. *Commun. Math. Sci* **2014**, *12*, 901–978.

32. Kleeman, R. Information theory and dynamical system predictability. *Entropy* **2011**, *13*, 612–649.

33. Kleeman, R. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **2002**, *59*, 2057–2072.

34. Majda, A.; Kleeman, R.; Cai, D. A mathematical framework for quantifying predictability through relative entropy. *Methods Appl. Anal.* **2002**, *9*, 425–444.

35. Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans* **1994**, *99*, 10143–10162.

36. Tebaldi, C.; Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2007**, *365*, 2053–2075.

37. Reichler, T.; Kim, J. How well do coupled models simulate today's climate? *Bull. Am. Meteorol. Soc.* **2008**, *89*, 303–312.

38. Marconi, U.M.B.; Puglisi, A.; Rondoni, L.; Vulpiani, A. Fluctuation–dissipation: Response theory in statistical physics. *Phys. Rep.* **2008**, *461*, 111–195.

39. Leith, C. Climate response and fluctuation dissipation. *J. Atmos. Sci.* **1975**, *32*, 2022–2026.

40. Majda, A.; Abramov, R.V.; Grote, M.J. *Information Theory and Stochastics for Multiscale Nonlinear Systems*; American Mathematical Society: Providence, RI, USA, 2005; Volume 25.

41. Majda, A.J.; Gershgorin, B.; Yuan, Y. Low-frequency climate response and fluctuation–dissipation theorems: Theory and practice. *J. Atmos. Sci.* **2010**, *67*, 1186–1201.

42. Beniston, M.; Stephenson, D.B.; Christensen, O.B.; Ferro, C.A.; Frei, C.; Goyette, S.; Halsnaes, K.; Holt, T.; Jylhä, K.; Koffi, B.; et al. Future extreme events in European climate: an exploration of regional climate model projections. *Clim. Chang.* **2007**, *81*, 71–95.

43. Palmer, T.; Räisänen, J. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **2002**, *415*, 512.

44. Majda, A.J.; Qi, D. Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Rev.* **2018**, in press.

45. Sapsis, T.P.; Majda, A.J. Statistically accurate low-order models for uncertainty quantification in turbulent dynamical systems. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13705–13710.

46. Chen, N.; Majda, A.J.; Giannakis, D. Predicting the cloud patterns of the Madden-Julian Oscillation through a low-order nonlinear stochastic model. *Geophys. Res. Lett.* **2014**, *41*, 5612–5619.

47. Harlim, J.; Mahdi, A.; Majda, A.J. An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *J. Comput. Phys.* **2014**, *257*, 782–812.

48. Majda, A.J.; Harlim, J. Physics constrained nonlinear regression models for time series. *Nonlinearity* **2012**, *26*, 201.

49. Chen, N.; Majda, A.J.; Tong, X.T. Information barriers for noisy Lagrangian tracers in filtering random incompressible flows. *Nonlinearity* **2014**, *27*, 2133.

50. Majda, A.J.; Qi, D. Improving prediction skill of imperfect turbulent models through statistical response and information theory. *J. Nonlinear Sci.* **2016**, *26*, 233–285.

51. Archie, K.M.; Dilling, L.; Milford, J.B.; Pampel, F.C. Unpacking the 'information barrier': Comparing perspectives on information as a barrier to climate change adaptation in the interior mountain West. *J. Environ. Manag.* **2014**, *133*, 397–410.

52. Dunne, S.; Entekhabi, D. An ensemble-based reanalysis approach to land data assimilation. *Water Resour. Res.* **2005**, *41*, doi:10.1029/2004WR003449.

53. Janjić, T.; Bormann, N.; Bocquet, M.; Carton, J.; Cohn, S.; Dance, S.; Losa, S.; Nichols, N.; Potthast, R.; Waller, J.; et al. On the representation error in data assimilation. *Q. J. R. Meteorol. Soc.* **2017**, doi:10.1002/qj.3130.

54. Fowler, A.; Jan Van Leeuwen, P. Measures of observation impact in non-Gaussian data assimilation. *Tellus A Dyn. Meteorol. Oceanogr.* **2012**, *64*, 17192.

55. Fowler, A.; Jan Van Leeuwen, P. Observation impact in data assimilation: The effect of non-Gaussian observation error. *Tellus A Dyn. Meteorol. Oceanogr.* **2013**, *65*, 20035.

56. Xu, Q. Measuring information content from observations for data assimilation: Relative entropy versus Shannon entropy difference. *Tellus A* **2007**, *59*, 198–209.

57. Roulston, M.S.; Smith, L.A. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **2002**, *130*, 1653–1660.

58. Weisheimer, A.; Corti, S.; Palmer, T.; Vitart, F. Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. *Philos. Trans. R. Soc. A* **2014**, *372*, 20130290.

59. Huffman, G.J. Estimates of root-mean-square random error for finite samples of estimated precipitation. *J. Appl. Meteorol.* **1997**, *36*, 1191–1201.

60. Illian, J.; Penttinen, A.; Stoyan, H.; Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns*; John Wiley & Sons: New York, NY, USA, 2008; Volume 70.

61. Chen, N.; Majda, A.J. Predicting the real-time multivariate Madden—Julian oscillation index through a low-order nonlinear stochastic model. *Mon. Weather Rev.* **2015**, *143*, 2148–2169.

62. Xie, X.; Mohebujjaman, M.; Rebholz, L.; Iliescu, T. Data-driven filtered reduced order modeling of fluid flows. *SIAM J. Sci. Comput.* **2018**, *40*, B834–B857.

63. Lassila, T.; Manzoni, A.; Quarteroni, A.; Rozza, G. Model order reduction in fluid dynamics: Challenges and perspectives. In *Reduced Order Methods for Modeling and Computational Reduction*; Springer: Berlin, Germany, 2014; pp. 235–273.

64. Benosman, M.; Borggaard, J.; San, O.; Kramer, B. Learning-based robust stabilization for reduced-order models of 2D and 3D Boussinesq equations. *Appl. Math. Model.* **2017**, *49*, 162–181.

65. Mehta, P.M.; Linares, R. A methodology for reduced order modeling and calibration of the upper atmosphere. *Space Weather* **2017**, *15*, 1270–1287.

66. Rozier, D.; Birol, F.; Cosme, E.; Brasseur, P.; Brankart, J.M.; Verron, J. A reduced-order Kalman filter for data assimilation in physical oceanography. *SIAM Rev.* **2007**, *49*, 449–465.

67. Farrell, B.F.; Ioannou, P.J. State estimation using a reduced-order Kalman filter. *J. Atmos. Sci.* **2001**, *58*, 3666–3680.

68. Ştefănescu, R.; Sandu, A.; Navon, I.M. POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation. *J. Comput. Phys.* **2015**, *295*, 569–595.

69. Berner, J.; Achatz, U.; Batté, L.; Bengtsson, L.; de la Cámara, A.; Christensen, H.M.; Colangeli, M.; Coleman, D.R.; Crommelin, D.; Dolaptchiev, S.I.; et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 565–588.

70. Dawson, A.; Palmer, T. Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Clim. Dyn.* **2015**, *44*, 2177–2193.

71. Franzke, C.L.; O'Kane, T.J.; Berner, J.; Williams, P.D.; Lucarini, V. Stochastic climate theory and modeling. *Wiley Interdiscip. Rev. Clim. Chang.* **2015**, *6*, 63–78.

72. Bauer, P.; Thorpe, A.; Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **2015**, *525*, 47.

73. Mellor, G.L.; Yamada, T. Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.* **1982**, *20*, 851–875.

74. Sander, J. Dynamical equations and turbulent closures in geophysics. *Contin. Mech. Thermodyn.* **1998**, *10*, 1–28.

75. Wang, Z.; Akhtar, I.; Borggaard, J.; Iliescu, T. Proper orthogonal decomposition closure models for turbulent flows: A numerical comparison. *Comput. Methods Appl. Mech. Eng.* **2012**, *237*, 10–26.

76. Gatski, T.; Jongen, T. Nonlinear eddy viscosity and algebraic stress models for solving complex turbulent flows. *Prog. Aerosp. Sci.* **2000**, *36*, 655–682.

77. Cambon, C.; Scott, J.F. Linear and nonlinear models of anisotropic turbulence. *Annu. Rev. Fluid Mech.* **1999**, *31*, 1–53.

78. Nakanishi, M.; Niino, H. Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteorol. Soc. Jpn. Ser. II* **2009**, *87*, 895–912.

79. Wilcox, D.C. *Turbulence Modeling for CFD*; DCW Industries: La Cañada Flintridge, CA, USA, 1998; Volume 2.

80. Majda, A.J.; Qi, D.; Sapsis, T.P. Blended particle filters for large-dimensional chaotic dynamical systems. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7511–7516.

81. Qi, D.; Majda, A.J. Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory. *Commun. Math. Sci.* **2016**, *14*, 1687–1722.

82. Qi, D.; Majda, A.J. Predicting extreme events for passive scalar turbulence in two-layer baroclinic flows through reduced-order stochastic models. *Commun. Math. Sci.* **2018**, in press.

83. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

84. Kullback, S. Letter to the editor: The Kullback–Leibler distance. *Am. Stat.* **1987**, *41*, 340–341.

85. Kullback, S. *Statistics and Information Theory*; John Wiley Sons: New York, NY, USA, 1959.

86. Branstator, G.; Teng, H. Two limits of initial-value decadal predictability in a CGCM. *J. Clim.* **2010**, *23*, 6292–6311.

87. DelSole, T. Predictability and information theory. Part I: Measures of predictability. *J. Atmos. Sci.* **2004**, *61*, 2425–2440.

88. DelSole, T. Predictability and information theory. Part II: Imperfect forecasts. *J. Atmos. Sci.* **2005**, *62*, 3368–3381.

89. Teng, H.; Branstator, G. Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Clim. Dyn.* **2011**, *36*, 1813–1834.

90. Hairer, M.; Majda, A.J. A simple framework to justify linear response theory. *Nonlinearity* **2010**, *23*, 909.

91. Bernardo, J.M.; Smith, A.F. Bayesian Theory; JohnWiley & Sons: Hoboken, NJ, USA, 2009.

92. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*; Springer Science & Business Media: Berlin, Germany, 2013.

93. Williams, D.; Williams, D. *Weighing the Odds: A Course in Probability and Statistics*; Springer: Berlin, Germany, 2001; Volume 3.

94. Bennett, A.F. *Inverse Modeling of the Ocean and Atmosphere*; Cambridge University Press: Cambridge, UK, 2005.

95. Schneider, S.H. Introduction to climate modeling. *SMR* **1992**, *648*, 6.

96. Gershgorin, B.; Harlim, J.; Majda, A.J. Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation. *J. Comput. Phys.* **2010**, *229*, 32–57.

97. Gershgorin, B.; Harlim, J.; Majda, A.J. Test models for improving filtering with model errors through stochastic parameter estimation. *J. Comput. Phys.* **2010**, *229*, 1–31.

98. Lee, W.; Stuart, A. Derivation and analysis of simplified filters. *Commun. Math. Sci.* **2017**, *15*, 413–450.

99. Mohamad, M.A.; Sapsis, T.P. Probabilistic description of extreme events in intermittently unstable dynamical systems excited by correlated stochastic processes. *SIAM/ASA J. Uncertain. Quantif.* **2015**, *3*, 709–736.

100. Lee, Y.; Majda, A.J.; Qi, D. Preventing catastrophic filter divergence using adaptive additive inflation for baroclinic turbulence. *Mon. Weather Rev.* **2017**, *145*, 669–682.

101. Anderson, J.L. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **2001**, *129*, 2884–2903.

102. Branicki, M.; Gershgorin, B.; Majda, A.J. Filtering skill for turbulent signals for a suite of nonlinear and linear extended Kalman filters. *J. Comput. Phys.* **2012**, *231*, 1462–1498.

103. Sykes, R.; Gabruk, R. A second-order closure model for the effect of averaging time on turbulent plume dispersion. *J. Appl. Meteorol.* **1997**, *36*, 1038–1045.

104. Majda, A.J. State Estimation, Data Assimilation, or Filtering for Complex Turbulent Dynamical Systems. In *Introduction to Turbulent Dynamical Systems in Complex Systems*; Springer: Berlin, Germany, 2016; pp. 65–83.

105. DelSole, T.; Shukla, J. Model fidelity versus skill in seasonal forecasting. *J. Clim.* **2010**, *23*, 4794–4806.

106. Chen, N.; Majda, A.J. Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 12864–12869.

107. Lindner, B.; Garcıa-Ojalvo, J.; Neiman, A.; Schimansky-Geier, L. Effects of noise in excitable systems. *Phys. Rep.* **2004**, *392*, 321–424.

108. Gardiner, C.W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*; Series in Synergetics; Springer: Berlin, Germany 2004; Volume 13.

109. Majda, A.J.; Timofeyev, I.; Eijnden, E.V. Models for stochastic climate prediction. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 14687–14691.

110. Majda, A.J.; Franzke, C.; Khouider, B. An applied mathematics perspective on stochastic modelling for climate. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2008**, *366*, 2427–2453.

111. Majda, A.J.; Yuan, Y. Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. *Discret. Contin. Dyn. Syst. B* **2012**, *17*, 1333–1363.

112. Franzke, C.; Majda, A.J. Low-order stochastic mode reduction for a prototype atmospheric GCM. *J. Atmos. Sci.* **2006**, *63*, 457–479.

113. Majda, A.J.; Timofeyev, I.; Vanden-Eijnden, E. Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.* **2003**, *60*, 1705–1722.

114. Kubo, R.; Toda, M.; Hashitsume, N. *Statistical Physics II: Nonequilibrium Statistical Mechanics*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 31.

115. Gottwald, G.A.; Harlim, J. The role of additive and multiplicative noise in filtering complex dynamical systems. *Proc. R. Soc. A* **2013**, *469*, 20130096.

116. Majda, A.J.; Franzke, C.; Crommelin, D. Normal forms for reduced stochastic climate models. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3649–3653.

117. Qi, D.; Majda, A.J. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Two-layer baroclinic turbulence. *J. Atmos. Sci.* **2016**, *73*, 4609–4639.

118. Yaglom, A.M. *An Introduction to the Theory of Stationary Random Functions*; Courier Corporation: Chelmsford, MA, USA, 2004.

119. Lorenz, E.N. Irregularity: A fundamental property of the atmosphere. *Tellus A Dyn. Meteorol. Oceanogr.* **1984**, *36*, 98–110.

120. Olbers, D. A gallery of simple models from climate physics. In *Stochastic Climate Models*; Springer: Berlin, Germany, 2001; pp. 3–63.

121. Salmon, R. *Lectures on Geophysical Fluid Dynamics*; Oxford University Press: Oxford, UK, 1998.

122. Park, S.K.; Xu, L. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*; Springer Science & Business Media: Berlin, Germany, 2013; Volume 2.

123. Beven, K.; Freer, J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **2001**, *249*, 11–29.

124. Reichle, R.H. Data assimilation methods in the Earth sciences. *Adv. Water Resour.* **2008**, *31*, 1411–1418.

125. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45.

126. Anderson, B.D.; Moore, J.B. *Optimal Filtering*; Dover Publications: Englewood Cliffs, NJ, USA, 1979; Volume 21, pp. 22–95.

127. Chui, C.K.; Chen, G. *Kalman Filtering*; Springer: Berlin, Germany, 2017.

128. Castronovo, E.; Harlim, J.; Majda, A.J. Mathematical test criteria for filtering complex systems: Plentiful observations. *J. Comput. Phys.* **2008**, *227*, 3678–3714.

129. Van Leeuwen, P.J. Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Q. J. R. Meteorol. Soc.* **2010**, *136*, 1991–1999.

130. Lee, Y.; Majda, A.J. Multiscale methods for data assimilation in turbulent systems. *Multiscale Model. Simul.* **2015**, *13*, 691–713.

131. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192.

132. Houtekamer, P.L.; Mitchell, H.L. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **1998**, *126*, 796–811.

133. Lermusiaux, P.F. Data assimilation via error subspace statistical estimation. Part II: Middle Atlantic Bight shelfbreak front simulations and ESSE validation. *Mon. Weather Rev.* **1999**, *127*, 1408–1432.

134. Hendon, H.H.; Lim, E.; Wang, G.; Alves, O.; Hudson, D. Prospects for predicting two flavors of El Niño. *Geophys. Res. Lett.* **2009**, *36*, doi:10.1029/2009GL040100.

135. Kim, H.M.; Webster, P.J.; Curry, J.A. Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Clim. Dyn.* **2012**, *39*, 2957–2973.

136. Barato, A.; Seifert, U. Unifying three perspectives on information processing in stochastic thermodynamics. *Phys. Rev. Lett.* **2014**, *112*, 090601.

137. Kawaguchi, K.; Nakayama, Y. Fluctuation theorem for hidden entropy production. *Phys. Rev. E* **2013**, *88*, 022147.

138. Pham, D.T. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Weather Rev.* **2001**, *129*, 1194–1207.

139. Chen, N.; Majda, A.J. Model error in filtering random compressible flows utilizing noisy Lagrangian tracers. *Mon. Weather Rev.* **2016**, *144*, 4037–4061.

140. Chen, N.; Majda, A.J. Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics. *Mon. Weather Rev.* **2016**, *144*, 4885–4917.

141. Harlim, J.; Majda, A.J. Catastrophic filter divergence in filtering nonlinear dissipative systems. *Commun. Math. Sci.* **2010**, *8*, 27–43.

142. Tong, X.T.; Majda, A.J.; Kelly, D. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *Commun. Math. Sci.* **2016**, *14*, 1283–1313.

143. Vallis, G.K. *Atmospheric and Oceanic Fluid Dynamics*; Cambridge University Press: Cambridge, UK, 2017.

144. Hasselmann, K. Stochastic climate models Part I. Theory. *Tellus* **1976**, *28*, 473–485.

145. Buizza, R.; Milleer, M.; Palmer, T. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **1999**, *125*, 2887–2908.

146. Franzke, C.; Majda, A.J.; Vanden-Eijnden, E. Low-order stochastic mode reduction for a realistic barotropic model climate. *J. Atmos. Sci.* **2005**, *62*, 1722–1745.

147. Rossby, C. On the mutual adjustment of pressure and velocity distributions in certain simple current systems. *J. Mar. Res* **1937**, *1*, 15–27.

148. Gill, A. Atmospheric-ocean dynamics. *Int. Geophys. Ser.* **1982**, *30*, 662.

149. Cushman-Roisin, B.; Beckers, J.M. *Introduction to Geophysical Fluid Dynamics: Physical and Numerical Aspects*; Academic Press: New York, NY, USA, 2011; Volume 10.

150. Grooms, I.; Lee, Y.; Majda, A. Ensemble filtering and low-resolution model error: Covariance inflation, stochastic parameterization, and model numerics. *Mon. Weather Rev.* **2015**, *143*, 3912–3924.

151. Grooms, I.; Lee, Y.; Majda, A.J. Ensemble Kalman filters for dynamical systems with unresolved turbulence. *J. Comput. Phys.* **2014**, *273*, 435–452.

152. Huang, C.; Li, X. Experiments of soil moisture data assimilation system based on ensemble Kalman filter. *Plateau Meteorol.* **2006**, *4*, 013.

153. Oke, P.R.; Sakov, P. Representation error of oceanic observations for data assimilation. *J. Atmos. Ocean. Technol.* **2008**, *25*, 1004–1017.

154. Hodyss, D.; Nichols, N. The error of representation: Basic understanding. *Tellus A Dyn. Meteorol. Oceanogr.* **2015**, *67*, 24822.

155. Chen, N.; Majda, A.J.; Tong, X.T. Noisy Lagrangian tracers for filtering random rotating compressible flows. *J. Nonlinear Sci.* **2015**, *25*, 451–488.

156. Ide, K.; Kuznetsov, L.; Jones, C.K. Lagrangian data assimilation for point vortex systems. *J. Turbul.* **2002**, *3*, doi:10.1088/1468-5248/3/1/053.

157. Kuznetsov, L.; Ide, K.; Jones, C.K. A method for assimilation of Lagrangian data. *Mon. Weather Rev.* **2003**, *131*, 2247–2260.

158. Apte, A.; Jones, C.; Stuart, A.; Voss, J. Data assimilation: Mathematical and statistical perspectives. *Int. J. Numer. Methods Fluids* **2008**, *56*, 1033–1046.

159. Molcard, A.; Piterbarg, L.I.; Griffa, A.; Özgökmen, T.M.; Mariano, A.J. Assimilation of drifter observations for the reconstruction of the Eulerian circulation field. *J. Geophys. Res. Oceans* **2003**, *108*, doi:10.1029/2001JC001240.

160. Ide, K.; Courtier, P.; Ghil, M.; Lorenc, A.C. Unified Notation for Data Assimilation: Operational, Sequential and Variational (gtSpecial IssueltData Assimilation in Meteology and Oceanography: Theory and Practice). *J. Meteorol. Soc. Jpn. Ser. II* **1997**, *75*, 181–189.

161. Dee, D.P.; Da Silva, A.M. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.* **1998**, *124*, 269–295.

162. Majda, A.; Wang, X. Linear response theory for statistical ensembles in complex systems with time-periodic forcing. *Commun. Math. Sci.* **2010**, *8*, 145–172.

163. Gritsun, A.; Branstator, G. Climate response using a three-dimensional operator based on the fluctuation–dissipation theorem. *J. Atmos. Sci.* **2007**, *64*, 2558–2575.

164. Gritsun, A.; Branstator, G.; Majda, A. Climate response of linear and quadratic functionals using the fluctuation–dissipation theorem. *J. Atmos. Sci.* **2008**, *65*, 2824–2841.

165. Risken, H. Fokker-planck equation. In *The Fokker–Planck Equation*; Springer: Berlin, Germany, 1996; pp. 63–95.

166. Penland, C.; Sardeshmukh, P.D. The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **1995**, *8*, 1999–2024.

167. Gershgorin, B.; Majda, A.J. A test model for fluctuation–dissipation theorems with time-periodic statistics. *Phys. D Nonlinear Phenom.* **2010**, *239*, 1741–1757.

168. Abramov, R.V.; Majda, A.J. Blended response algorithms for linear fluctuation–dissipation for complex nonlinear dynamical systems. *Nonlinearity* **2007**, *20*, 2793.

169. Abramov, R.V.; Majda, A.J. A new algorithm for low-frequency climate response. *J. Atmos. Sci.* **2009**, *66*, 286–309.

170. Majda, A.J.; Abramov, R.; Gershgorin, B. High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 581–586.

171. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2012.

172. Franzke, C.; Horenko, I.; Majda, A.J.; Klein, R. Systematic metastable atmospheric regime identification in an AGCM. *J. Atmos. Sci.* **2009**, *66*, 1997–2012.

173. Majda, A.J.; Franzke, C.L.; Fischer, A.; Crommelin, D.T. Distinct metastable atmospheric regimes despite nearly Gaussian statistics: A paradigm model. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8309–8314.

174. Cressie, N.; Wikle, C.K. *Statistics for Spatio-Temporal Data*; John Wiley & Sons: New York, NY, USA, 2015.

175. Chen, N.; Majda, A.J.; Sabeerali, C.T.; Ravindran, A.S. Predicting monsoon intraseasonal precipitation using a low-order nonlinear stochastic model. *J. Clim.* **2018**, in press, doi:10.1175/JCLI-D-17-0411.1.

176. Chen, N.; Majda, A.J. Predicting the Cloud Patterns for the Boreal Summer Intraseasonal Oscillation Through a Low-Order Stochastic Model. *Math. Clim. Weather Forecast.* **2015**, *1*, 1–20.

177. Hodges, K.; Chappell, D.; Robinson, G.; Yang, G. An improved algorithm for generating global window brightness temperatures from multiple satellite infrared imagery. *J. Atmos. Ocean. Technol.* **2000**, *17*, 1296–1312.

178. Majda, A.J.; Stechmann, S.N. The skeleton of tropical intraseasonal oscillations. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8417–8422.

179. Majda, A.J.; Stechmann, S.N. Nonlinear dynamics and regional variations in the MJO skeleton. *J. Atmos. Sci.* **2011**, *68*, 3053–3071.

180. Thual, S.; Majda, A.J.; Stechmann, S.N. A stochastic skeleton model for the MJO. *J. Atmos. Sci.* **2014**, *71*, 697–715.

181. Ogrosky, H.R.; Stechmann, S.N. The MJO skeleton model with observation-based background state and forcing. *Q. J. R. Meteorol. Soc.* **2015**, *141*, 2654–2669.

182. Epstein, E.S. Stochastic dynamic prediction. *Tellus* **1969**, *21*, 739–759.

183. Fleming, R.J. On stochastic dynamic prediction. *Mon. Wea. Rev* **1971**, *99*, 1236.

184. Srinivasan, K.; Young, W. Zonostrophic instability. *J. Atmos. Sci.* **2012**, *69*, 1633–1656.

185. Lesieur, M. *Turbulence in Fluids*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 40.

186. Branicki, M.; Chen, N.; Majda, A.J. Non-Gaussian test models for prediction and state estimation with model errors. *Chin. Ann. Math. Ser. B* **2013**, *34*, 29–64.

187. Frierson, D.M. Midlatitude static stability in simple and comprehensive general circulation models. *J. Atmos. Sci.* **2008**, *65*, 1049–1062.

188. Frierson, D.M. Robust increases in midlatitude static stability in simulations of global warming. *Geophys. Res. Lett.* **2006**, *33*, doi:10.1029/2006GL027504.

189. Majda, A.J.; Kramer, P.R. Simplified models for turbulent diffusion: Theory, numerical modelling, and physical phenomena. *Phys. Rep.* **1999**, *314*, 237–574.

190. Neelin, J.D.; Lintner, B.R.; Tian, B.; Li, Q.; Zhang, L.; Patra, P.K.; Chahine, M.T.; Stechmann, S.N. Long tails in deep columns of natural and anthropogenic tropospheric tracers. *Geophys. Res. Lett.* **2010**, *37*, doi:10.1029/2009GL041726.

191. Jayesh; Warhaft, Z. Probability distribution, conditional dissipation, and transport of passive temperature fluctuations in grid-generated turbulence. *Phys. Fluids A Fluid Dyn.* **1992**, *4*, 2292–2307.

192. Bourlioux, A.; Majda, A. Elementary models with probability distribution function intermittency for passive scalars with a mean gradient. *Phys. Fluids* **2002**, *14*, 881–897.

193. Majda, A.J.; Gershgorin, B. Elementary models for turbulent diffusion with complex physical features: Eddy diffusivity, spectrum and intermittency. *Philos. Trans. R. Soc. A* **2013**, *371*, 20120184.

194. Gershgorin, B.; Majda, A. A nonlinear test model for filtering slow-fast systems. *Commun. Math. Sci.* **2008**, *6*, 611–649.

195. Majda, A.J.; Tong, X.T. Intermittency in turbulent diffusion models with a mean gradient. *Nonlinearity* **2015**, *28*, 4171.

196. Gershgorin, B.; Majda, A.J. Filtering a statistically exactly solvable test model for turbulent tracers from partial observations. *J. Comput. Phys.* **2011**, *230*, 1602–1638.

197. Avellaneda, M.; Majda, A.J. Mathematical models with exact renormalization for turbulent transport. *Commun. Math. Phys.* **1990**, *131*, 381–429.

198. Avellaneda, M.; Majda, A. Mathematical models with exact renormalization for turbulent transport, II: Fractal interfaces, non-Gaussian statistics and the sweeping effect. *Commun. Math. Phys.* **1992**, *146*, 139–204.

199. Bronski, J.C.; McLaughlin, R.M. The problem of moments and the Majda model for scalar intermittency. *Phys. Lett. A* **2000**, *265*, 257–263.

200. Vanden Eijnden, E. Non-Gaussian invariant measures for the Majda model of decaying turbulent transport. *Commun. Pure Appl. Math.* **2001**, *54*, 1146–1167.

201. Kraichnan, R.H. Small-scale structure of a scalar field convected by turbulence. *Phys. Fluids* **1968**, *11*, 945–953.

202. Kraichnan, R.H. Anomalous scaling of a randomly advected passive scalar. *Phys. Rev. Lett.* **1994**, *72*, 1016.

203. Kramer, P.R.; Majda, A.J.; Vanden-Eijnden, E. Closure approximations for passive scalar turbulence: A comparative study on an exactly solvable model with complex features. *J. Stat. Phys.* **2003**, *111*, 565–679.

204. Lorenz, E.N. Predictability: A problem partly solved. In Proceedings of the 1996 Seminar on Predictability, Shinfield, UK, 4–8 September 1995; Volume 1.

205. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435.

206. Majda, A.J.; Qi, D. Effective control of complex turbulent dynamical systems through statistical functionals. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 5571–5576.

207. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **1946**, *186*, 453–461.

208. Brue, G.; Howes, R. *The McGraw Hill 36 Hour Six Sigma Course*; McGraw Hill Professional: New York, NY, USA, 2004.