

# Forecasting Turbulent Modes with Nonparametric Diffusion Models

Tyrus Berry

*Department of Mathematics, The Pennsylvania State University, University Park, PA 16802*

John Harlim\*

*Department of Mathematics and Department of Meteorology, The Pennsylvania State University, University Park, PA 16802*

---

## Abstract

This paper presents a nonparametric diffusion modeling approach for forecasting partially observed noisy turbulent modes. The proposed forecast model uses a basis of smooth functions (constructed with the diffusion maps algorithm) to represent probability densities, so that the forecast model becomes a linear map in this basis. We estimate this linear map by exploiting a previously established rigorous connection between the discrete time shift map and the semi-group solution associated to the backward Kolmogorov equation. In order to smooth the noisy data, we apply diffusion maps to a delay embedding of the noisy data, which also helps to account for the interactions between the observed and unobserved modes. We show that this delay embedding biases the geometry of the data in a way which extracts the most predictable component of the dynamics. The resulting model approximates the semigroup solutions of the generator of the underlying dynamics in the limit of large data and in the observation noise limit. We will show numerical examples on a wide-range of well-studied turbulent modes, including the Fourier modes of the energy conserving Truncated Burgers-Hopf (TBH) model, the Lorenz-96 model in weakly chaotic to fully turbulent regimes, and the barotropic modes of a quasi-geostrophic model with baroclinic instabilities. In these examples, forecasting skills of the nonparametric diffusion model are compared to a wide-range of stochastic parametric modeling approaches, which account for the nonlinear interactions between the observed and unobserved modes with white and colored noises.

*Keywords:* nonparametric forecasting, kernel methods, diffusion maps, diffusion models, time-lagged embedding.

---

## 1. Introduction

A long-standing issue in modeling turbulent dynamics is the so-called turbulent closure problem (see e.g.[1]) where the goal is to find a set of effective equations up to a finite resolution which represent the fundamental features of the turbulent dynamics. In other words, we have fewer equations than unknowns to predict the coarse grained variables of interest. Common closure approaches typically use physical insights to choose a parametric ansatz to represent the feedback from the unresolved scales (see e.g., [2] for various closure approximations for passive scalar turbulence problems). Alternatively, many stochastic modeling approaches have been successful in predicting geophysical turbulence [3, 4, 5, 6]. While these parametric modeling approaches were typically derived from the first principles through some asymptotic analysis, they also use some physical insight to choose an appropriate ansatz to represent the unresolved scales.

Despite these successes, the parametric modeling approaches have practical issues due to model error when the necessary physical insights are not known. If the parametric model (or ansatz) is not chosen appropriately, one can end up with a model with poor predictive skills (or even with solutions which diverge catastrophically) even when the parameters can be obtained by a standard regression fitting procedure [7]. Moreover, even when an appropriate parametric form is chosen, specifying the parameters from noisy observations of the physical variables can be

---

\*Corresponding author.

Email addresses: `tbh11@psu.edu` (Tyrus Berry), `jharlim@psu.edu` (John Harlim)

nontrivial since the parameters are typically not directly observed. Indeed, it was shown that an appropriate parameterization scheme is crucial for accurate filtering and climatology (which represents the long term forecast) even when the parametric forms are appropriately chosen [8].

Recently, a nonparametric modeling approach for forecasting the evolution of the probability density of low-dimensional dynamics was introduced in [9]. The key idea behind this modeling approach is to use a basis of smooth functions (constructed with the diffusion maps algorithm [10, 11]) to represent probability densities, so that the forecast model becomes a linear map in this basis. Numerically, this linear map is estimated by exploiting a rigorous connection between the discrete time shift map and semi-group solution associated to the backward Kolmogorov equation. In [9], it was shown that the resulting model estimates the semigroup solutions of the generator of the underlying dynamics in the limit of large data. In this paper, we test this nonparametric modeling approach as a method of forecasting noisy observed Fourier modes of a selection of well-studied high-dimensional dynamical systems in various turbulent regimes. The novel aspect of this paper is that we consider building a forecasting model given only finite amount of noisy training data set as in practical situation, in contrast to the work in [9] which assumed the data set is noise-free.

We should note that the method of [9] could be applied to noisy signals, however the nonparametric model would implicitly include the noise in the model which would limit the forecast skill compared to treating the noise as a separate process. Treating the noise as a separate process requires first learning the model from the noisy training data set, and then generating ‘clean’ initial conditions for forecasting from the noisy observations. In [12, 13, 14] it was shown that applying diffusion maps to the delay-embedded data reduced the noise in the training data. Building upon the work in [9], we apply the theory of [12] to show that building the nonparametric model using the delay-embedded data biases the geometry of the data in a way which extracts the most predictable component of the dynamics. Having built a nonparametric model from the noisy training data, the next step is to determine the initial state for forecasting from a noisy observation. To find the initial state, we introduce a Bayesian filtering method which iteratively assimilates each observation in order to find the probability density of the current state given all the previous observations.

One interesting question which we address here is whether it is possible to build a skillful nonparametric forecasting model for a turbulent mode given only a small amount of noisy training data, when the true dynamics are solutions of a high-dimensional dynamical systems with chaotic behavior. This question arises because the nonparametric model has a practical limitation in terms of modeling dynamics with high-dimensional attractors, namely: it will require an immense amount of data to unwind the attractors. Moreover, even given a sufficiently large data set, the required computational power would be a limiting factor since the diffusion maps algorithm requires storing and computing eigenvectors of an  $N \times N$  matrix, where  $N$  is the number of data points. Given a small data set, such that we cannot unwind the full attractor of the high-dimensional attractor, we attempt to circumvent the curse-of-dimensionality by decomposing the data into Fourier modes. Of course, the standard theory of embedology [15] suggests that the delay-embedding of a single Fourier mode would reconstruct the entire high-dimensional attractor, which would again be inaccessible to our nonparametric model due to dimensionality. However, while the full attractor is reconstructed in a topological sense, the geometry of the reconstructed attractor is dramatically altered. The nonparametric model of [9] leverages this biased geometry so that each Fourier mode gives a separate component of the dynamics.

The remainder of this paper is organized as follows. In Section 2, we introduce the problems under consideration and establish the necessary background, including a brief overview of the nonparametric modeling approach introduced in [9] as well as discussing of how the theory of [12] is applied to mitigate the effect of noise on the model. We conclude Section 2 by introducing the iterative Bayesian filter which we use to generate initial conditions for forecasting with the nonparametric model. In Section 3, we numerically compare the nonparametric model to a physics constrained nonlinear regression model [16] for predicting an energy-conserving turbulent mode. In Section 4, we numerically compare the nonparametric model with various parametric models, including the persistence model, autoregressive models of order-1 (MSM [17]) and order-3 [18], and the perfect model, for predicting Fourier modes of Lorenz-96 model in various chaotic regimes. In Section 5, we numerically compare the nonparametric model with a stochastic model with additive and multiplicative noises (SPEKF model [19, 20]) for predicting the barotropic modes of a geostrophic turbulence. We close this paper with a short summary in Section 5.

## 2. Nonparametric diffusion modeling

Let  $u(x, t) \in \mathbb{R}^s$  be the solutions of an ergodic system of nonlinear PDEs,

$$\frac{\partial u}{\partial t} = \mathcal{A}(\partial_x u), \quad (1)$$

where  $\mathcal{A}$  denotes nonlinear differential operators, for smooth initial conditions  $u(x, 0)$  and periodic boundary conditions on a non-dimensionalized periodic domain  $x \in [0, 2\pi]^n$ . To simplify the exposition, set  $s, n = 1$  without loss of generality. Here, the solutions of (1) can be described by the infinite Fourier series,

$$u(x, t) = \sum_{k \in \mathbb{Z}} \hat{u}_k e^{ikx}, \quad \hat{u}_k^* = \hat{u}_{-k}, \hat{u}_0 \in \mathbb{R}$$

where the Fourier modes  $\hat{u}_k$  can be utilized in analyzing (1).

Define  $\vec{u}_i = (u(x_0, t_i), \dots, u(x_{2m}, t_i)) \in \mathbb{R}^{2m+1}$  whose components are the solutions of (1) at time  $t_i$  realized at grid point  $x_\ell = \ell h, \ell = 0, \dots, 2m$ , such that  $(2m+1)h = 2\pi$ . Our goal is to construct a forecasting model for the discrete Fourier coefficients,

$$\hat{u}_{i,k} = (\vec{u}_i, \vec{e}^k)_h \equiv \frac{h}{2\pi} \sum_{\ell=0}^{2m} u(x_\ell, t_i) e^{-ikx_\ell}, \quad (2)$$

given the corresponding partially observed modes,  $\hat{v}_{i,k} = (\vec{v}_i, \vec{e}^k)_h$  where the  $\ell$ -th component of  $\vec{v}_i$ ,

$$v(x_\ell, t_i) = u(x_\ell, t_i) + \varepsilon_{\ell,i}, \quad \varepsilon_{\ell,i} \sim \mathcal{N}(0, R), \quad (3)$$

is the solution of (1), corrupted by i.i.d. Gaussian noise, at the grid point  $x_\ell$  and time step  $t_i$ . Notice that  $\{\vec{e}^\ell\}_{\ell=0, \pm 1, \dots, \pm m}$  forms an orthonormal basis of  $\mathbb{C}^{2m+1}$  with respect to the inner product defined in (2). One should also note that  $\hat{v}_{i,k} = \hat{u}_{i,k} + \eta_{i,k}$ , where  $\eta_{i,k} \sim \mathcal{N}(0, \hat{R})$  and  $\hat{R} = R/(2m+1)$  [21, 17].

Given the noisy time series of a single mode  $\{\hat{v}_{i,k}\}_{i=1}^N$ , our goal is to use this noisy data set to train the nonparametric model and to generate initial conditions for forecasting. In Section 2.1, we provide an overview of the construction of the nonparametric model introduced in [9] for fully observed data without observation noise. In Section 2.2, we show that by applying a delay embedding we can compensate for the partially observed noisy data,  $\hat{v}_{i,k}$ . Finally, in Section 2.3, we introduce a simple Bayesian filtering method for generating initial conditions for forecasting from the noisy observations  $\hat{v}_{i,k}$ .

### 2.1. Nonparametric forecasting model

In this section, we review the nonparametric diffusion forecasting model introduced in [9], assuming the data set consists of full observations of the dynamical system with no observation noise. Consider a system of SDEs,

$$d\hat{u} = a(\hat{u}) dt + b(\hat{u}) dW_t, \quad \hat{u}(0) = \hat{u}_0, \quad (4)$$

where  $a(\hat{u})$  is a vector field,  $b(\hat{u})$  is a diffusion tensor, and  $W_t$  is an i.i.d. Wiener process; all defined on a manifold  $\mathcal{M} \subset \mathbb{R}^n$ . Assume that the dynamical system governed by (4) is ergodic and that it has a density function  $p(\hat{u}, t) = e^{t\mathcal{L}^*} p(\hat{u}, 0)$  which converges to the invariant measure  $p_{eq}(\hat{u})$ , where  $\mathcal{L}^*$  denotes the Fokker-Planck (or forward) operator and  $p(\hat{u}, 0)$  denotes an initial density.

Given a time series  $\{\hat{u}_i = \hat{u}(t_i)\}_{i=1}^N$ , where  $\tau = t_{i+1} - t_i$ , our goal is to approximate  $p(\hat{u}, t)$  without knowing or estimating  $a$  and  $b$ . Instead, we will directly estimate the semigroup solution  $e^{t\mathcal{L}}$  associated to the generator  $\mathcal{L}$  of (4) by projecting the density onto an appropriate basis for  $L^2(\mathcal{M}, p_{eq})$ . In particular, we choose eigenfunctions  $\{\varphi_j\}$  of the generator  $\hat{\mathcal{L}}$  of a stochastically forced gradient flow system,

$$d\tilde{u} = -\nabla U(\tilde{u}) dt + \sqrt{2} dW_t, \quad (5)$$

where the potential function  $U \equiv -\log(p_{eq})$  is defined by the invariant measure of the full system (4) so that the invariant measure of (5) is  $\hat{p}_{eq} = e^{-U} = p_{eq}$ . This choice of basis is motivated by several considerations. First, we can

estimate the eigenfunctions  $\{\varphi_j\}$  by the diffusion maps algorithm for data lying on a compact manifold [10] and on a non-compact manifold [11]. In this paper, we will use the variable bandwidth kernel introduced in [11] to construct these eigenfunctions since the sampling measure of the data set may be arbitrarily small, which would imply that the data does not lie on a compact manifold. Second, by projecting the density function on this basis of eigenfunctions, we can accurately forecast each projected component by a discrete representation of the semigroup solution  $e^{\tau\mathcal{L}}$ . We now show how to construct the discrete representation of  $e^{\tau\mathcal{L}}$  using the “shift operator”.

Let  $\{\varphi_j\}$  be the eigenfunctions of the generator  $\hat{\mathcal{L}}$  of the gradient system in (5); these eigenfunctions are orthonormal under  $\langle \cdot, \cdot \rangle_{p_{eq}}$  in  $L^2(\mathcal{M}, p_{eq})$  and the integral here (and in all of the inner product defined below) is with respect to the volume form  $dV$  which  $\mathcal{M}$  inherits from the ambient space. Note that  $\hat{\mathcal{L}}$  is the generator of gradient system in (5) and it is the adjoint of the Fokker-Planck operator  $\hat{\mathcal{L}}^*$  with respect to inner-product  $\langle \cdot, \cdot \rangle$  in  $L^2(\mathcal{M})$ . One can show that  $\{\psi_j = \varphi_j p_{eq}\}$  are eigenfunctions of  $\hat{\mathcal{L}}^*$  which are orthonormal with respect to inner-product  $\langle \cdot, \cdot \rangle_{p_{eq}^{-1}}$  in  $L^2(\mathcal{M}, p_{eq}^{-1})$ . Given an initial density  $p(\hat{u}, 0) = p_0(\hat{u})$ , we can write the forecast density  $p(\hat{u}, t)$  as,

$$p(\hat{u}, t) = e^{t\mathcal{L}^*} p_0(\hat{u}) = \sum_j \langle e^{t\mathcal{L}^*} p_0, \psi_j \rangle_{p_{eq}^{-1}} \psi_j(\hat{u}) = \sum_j \langle p_0, e^{t\mathcal{L}} \varphi_j \rangle_{p_{eq}} \varphi_j(\hat{u}) p_{eq}(\hat{u}). \quad (6)$$

Setting  $t = 0$  in (6) we find,

$$p_0(\hat{u}) = p(\hat{u}, 0) = \sum_j \langle p_0, \varphi_j \rangle_{p_{eq}} \varphi_j(\hat{u}) p_{eq}(\hat{u}) \equiv \sum_j c_j(0) \varphi_j(\hat{u}) p_{eq}(\hat{u}), \quad (7)$$

where we define  $c_j(0) \equiv \langle p_0, \varphi_j \rangle$ . Substituting (7) into (6), we obtain,

$$p(\hat{u}, t) = \sum_j \sum_l c_l(0) \langle \varphi_l, e^{t\mathcal{L}} \varphi_j \rangle_{p_{eq}} \varphi_j(\hat{u}) p_{eq}(\hat{u}). \quad (8)$$

The key idea of the non-parametric forecasting algorithm in [9] is to approximate  $A_{jl} \equiv \langle \varphi_l, e^{\tau\mathcal{L}} \varphi_j \rangle_{p_{eq}}$  by replacing the semi-group solution  $e^{\tau\mathcal{L}}$  with the discrete time shift operator  $S$ , which is defined as  $S f(\hat{u}_i) = f(\hat{u}_{i+1})$  for any  $f \in L^2(\mathcal{M}, p_{eq})$ . In [9], it was shown that  $S$  is a stochastic estimate of  $e^{\tau\mathcal{L}}$  and the error due to the stochastic terms can be minimized by projecting  $S$  on the basis  $\varphi_j$ . Indeed it was shown that  $\hat{A}_{jl} \equiv \langle \varphi_l, S \varphi_j \rangle_{p_{eq}}$  is an unbiased estimate of  $A_{jl}$ , meaning that  $\mathbb{E}[\hat{A}_{jl}] = A_{jl}$ . Furthermore, assuming that  $\hat{u}_i$  are independent samples of  $p_{eq}$ , one can show that the error of this estimate is of order  $\sqrt{\tau/N}$ , which means that one can apply this approximation for any sampling time  $\tau$  given a sufficiently large data set  $N$ . By the ergodicity property, the largest eigenvalue of  $A$  should be equal to 1 (with no complex component), numerically however, we sometimes find that  $\hat{A}$  has some eigenvalues slightly larger than one, due to the finite number of samples in the Monte-Carlo integrals. If this occurs, we can easily ensure the stability of the diffusion forecast by dividing any eigenvalue with norm greater than 1 so that it has norm equal to one. However, In our numerical experiments below this issue rarely occurs.

We should also note that if (4) is a gradient flow as in (5), then  $A_{jl} \equiv \langle \varphi_l, e^{\tau\mathcal{L}} \varphi_j \rangle_{p_{eq}} = e^{\lambda_j \tau} \delta_{j,l}$ , where  $\lambda_j$  are the eigenvalues of  $\mathcal{L} = \hat{\mathcal{L}}$ , which can be obtained directly from the diffusion maps algorithm [10, 11]. Moreover, the matrix  $A$  becomes diagonal (due to the orthonormality of  $\varphi_j$ ) so that the diffusion maps algorithm estimates  $A$  directly and the shift operator approximation is not required. See [22] for various uncertainty quantification applications for this special case.

To conclude, the non-parametric forecasting algorithm (which we refer to as the *diffusion forecast*) for a given initial density  $p_0(\hat{u})$  is performed as follows:

1. *Learning phase*: Given a data set  $\{\hat{u}_i\}_{i=1}^N$ , apply the diffusion maps algorithm [10, 11] to obtain eigenvectors  $\{\tilde{\varphi}_j\}$ , whose  $i$ -th component,  $(\tilde{\varphi}_j)_i = \varphi_j(\hat{u}_i)$ , approximates the eigenfunction  $\varphi_j$  of the gradient flow (5) evaluated at the data point  $\hat{u}_i$ . We implement the diffusion maps algorithm with a variable bandwidth kernel, see the Appendix for the step-by-step algorithm to obtain these eigenfunctions. Also, see the supplementary material of [9] for a short overview of [11].
2. *Initial conditions*: Represent the initial density  $p_0$  in this basis as,

$$c_j(0) = \langle p_0, \varphi_j \rangle = \langle p_0 / p_{eq}, \varphi_j \rangle_{p_{eq}} = \sum_{i=1}^N \frac{p_0(\hat{u}_i)}{p_{eq}(\hat{u}_i)} \varphi_j(\hat{u}_i). \quad (9)$$

which is numerically obtainable from the last inner product as a Monte-Carlo integral since we have  $\varphi_j$  evaluated on the data set  $\hat{u}_i$  which are samples of  $p_{eq}$ .

3. *Forecasting*: Apply a Monte-Carlo integral to approximate the shift operator  $S$  in coordinate basis  $\varphi_j$ ,

$$\hat{A}_{jl} \equiv \langle \varphi_l, S \varphi_j \rangle_{p_{eq}} \approx \frac{1}{N} \sum_{i=1}^N \varphi_l(\hat{u}_i) \varphi_j(\hat{u}_{i+1}). \quad (10)$$

We then approximate  $A_{jl}$  with  $\hat{A}_{jl}$  such that the diffusion forecast is given by

$$p(\hat{u}, m\tau) \approx \sum_j \sum_l (\hat{A}^m)_{jl} c_l(0) \varphi_j(\hat{u}) p_{eq}(\hat{u}) \equiv \sum_j c_j(m\tau) \varphi_j(\hat{u}) p_{eq}(\hat{u}), \quad (11)$$

where in practice, these sums are truncated at a finite number,  $M$ , of eigenfunctions. With this truncation, the coefficient  $c_j(m\tau)$  is the  $j$ -th component of a matrix-vector multiplication,  $\vec{c}(m\tau) = \hat{A}^m \vec{c}(0)$ , of an  $M \times M$  matrix  $\hat{A}^m$  and an  $M$ -dimensional vector  $\vec{c}(0) = (c_1(0), \dots, c_M(0))^T$ .

## 2.2. Time-lagged embedding

The diffusion forecast algorithm of Section 2.1 assumes that the data  $\hat{u}_i$  are sampled directly from the full dynamical system (4). That is, the full system  $\hat{u}$  in (4) is equivalent to the dynamical system for  $\vec{u}$  and the turbulent dynamics considered here will be very high dimensional. For such high dimensional problems, exploring the entire attractor would require a prohibitively large data set; larger than will typically be available in applications. Instead, we attempt to build a low-dimensional model for each mode  $\hat{u}_{i,k}$  individually. An individual mode  $\hat{u}_{i,k}$  is simply a linear projection of the full system  $\hat{u}_i$ , and moreover it is an ‘observation’ of the spatial state  $u$ . While we could apply the diffusion forecast to the one dimensional time series  $\hat{u}_{i,k}$ , this would ignore the dependence of the  $k$ -th mode on all the other modes. The fundamental problem with building a one-dimensional stochastic model for  $\hat{u}_{i,k}$  is that any interaction with the other modes will result in (among other changes to the model) an inflated stochastic component in the closed model for  $\hat{u}_{i,k}$  (see [23, 8] for a rigorous example). Inflating the stochastic component of the model for  $\hat{u}_{i,k}$  will severely limit the predictive skill of the nonparametric model. On the other hand, if we include other modes in the nonparametric model, this would increase the dimensionality and not all of the information in the other modes would be relevant to forecasting the  $k$ -th mode. Instead, we will apply the delay coordinate reconstruction of [24, 15, 25, 26] to implicitly recover only the missing components of the dynamics which are important for forecasting each  $\hat{u}_{i,k}$ . Moreover, we will apply the theory of [12] to show that the delay coordinate reconstruction projects the missing components of the dynamics onto the component most important for the prediction of  $\hat{u}_{i,k}$ . Finally, in practice we will only have access to a noisy data set  $\hat{v}_{i,k}$ , and the theory of [12] shows that the delay coordinate reconstruction also reduces the influence of the noise.

Given a time series  $\hat{u}_{i,k} = \hat{u}_k(t_i)$  of the  $k$ -th mode, we construct the delay embedding coordinates,

$$\vec{\hat{u}}_{i,k} = H(\hat{u}_{i,k}) = (\hat{u}_{i,k}, \hat{u}_{i-1,k}, \dots, \hat{u}_{i-L,k})^T.$$

The theory of embedology shows that if the  $k$ -th mode is a generic observation of the full system (4), then for sufficiently many lags,  $L$ , there exists a diffeomorphism  $\mathcal{F}$  such that  $\vec{\hat{u}}_{i,k} = \mathcal{F}(\hat{u}_i)$ . This statement was first shown for deterministic dynamics on smooth attractors in [24] and subsequently generalized to fractal attractors in [15] and then to non-autonomous systems in [25] and stochastic systems in [26]. The existence of the diffeomorphism  $\mathcal{F}$  says that topologically the attractor of the delay reconstruction  $\vec{\hat{u}}_{i,k}$  is equivalent to the full dynamics  $\hat{u}_i$  and so we have *reconstructed* the hidden variables in the evolution of the  $k$ -th mode. However, the theory of embedology only concerns the topology of the attractor, whereas the basis of eigenfunctions  $\{\varphi_j\}$  will depend on the geometry of the delay reconstruction  $\vec{\hat{u}}_{i,k}$ .

In [12] it was shown that the delay reconstruction severely biases the geometry in the reconstructed coordinates  $\vec{\hat{u}}_{i,k}$ , so that in the limit of infinite delays,  $L \rightarrow \infty$ , the dynamics were projected on the most stable component of the dynamics. Consider the  $k$ -th mode  $\hat{u}_{i,k}$  as an observation of the state  $\hat{u}_i$ , where the observation function is given by  $\hat{u}_{i,k} = h_k(\hat{u}_i) = (0, \dots, 0, 1, 0, \dots, 0)\hat{u}_i$ . Notice that the derivative of the observation function is  $Dh_k = (0, \dots, 0, 1, 0, \dots, 0)^T$  where the 1 occurs in the  $k$ -th entry. Moreover, the previous value of the  $k$ -th mode,  $\hat{u}_{i-1,k}$ , can also be considered

an observation of  $\hat{u}_i$  where the observation function is given by  $\hat{u}_{i-1,k} = h_k(F_{-\tau}(\hat{u}_i))$  and the map  $F_{-\tau}(\hat{u}_i) = \hat{u}_{i-1}$  is given by the reverse time evolution for the discrete time step  $\tau$ . Interpreting each  $\hat{u}_{i-l,k}$  as an observation of the full state  $\hat{u}$  at time  $t_i$  shows that the time-delay embedding  $\vec{\hat{u}}_{i,k}$  is an observation of  $\hat{u}_i$  with observation function  $H$ . For  $L$  sufficiently large and assuming that the  $k$ -th mode is a generic observable on the manifold, the observation  $H$  will have a full rank derivative. Our goal now is to examine the change in metric induced by this derivative by seeing how it acts on tangent vectors on the attractor. Let  $v_1, v_2 \in T_{\hat{u}}\mathcal{M}$  be tangent vectors on the attractor and let  $\hat{v}_1 = DH(\hat{u})v_1$  and  $\hat{v}_2 = DH(\hat{u})v_2$  where  $\hat{v}_1, \hat{v}_2 \in T_{H(\hat{u})}H(\mathcal{M})$  are the transformed tangent vectors in the new geometry given by the time-delay embedding. Then the inner product  $\langle \cdot, \cdot \rangle_{\mathbb{R}^L}$  in the delay embedding space is

$$\begin{aligned} \langle \hat{v}_1, \hat{v}_2 \rangle_{\mathbb{R}^L} &= \sum_{l=0}^L \langle Dh_k(F_{-l\tau}(\hat{u}))DF_{-l\tau}(\hat{u})v_1, Dh_k(F_{-l\tau}(\hat{u}))DF_{-l\tau}(\hat{u})v_2 \rangle_{\mathbb{R}} \\ &= \sum_{l=0}^L (DF_{-l\tau}(\hat{u})v_1)_k (DF_{-l\tau}(\hat{u})v_2)_k, \end{aligned} \quad (12)$$

where  $F_{-l\tau}$  is the reversed shift map which takes  $\hat{u}_i$  to  $\hat{u}_{i-l}$ . If  $v_1$  and  $v_2$  are in the  $m$ -th Oseledets space, with Lyapunov exponent  $\sigma_m$ , then the inner product reduces to,  $\langle \hat{v}_1, \hat{v}_2 \rangle_{\mathbb{R}^L} = \sum_{l=0}^L e^{-2\sigma_m l} (v_1)_k (v_2)_k$ . This shows that the most stable Oseledets space will dominate the geometry in the embedding since  $\sigma_m < 0$  will be most negative (largest in absolute value of the negative Lyapunov exponents) in the most stable direction.

The bias introduced into the geometry by the delay embedding has several advantages for forecasting. First, the most stable components of the dynamics are those that have the most predictability. Since numerically we will not be able to represent the whole attractor, the variable bandwidth kernel will implicitly project away the small components of the geometry. By amplifying the stable components we insure that the most desirable aspects of the dynamics will be well represented in the discrete representation of the geometry. Secondly, when applied to the noisy data,  $\hat{v}$ , the noise will correspond to an unstable component, and so the delay embedding geometry will de-emphasize the noise component of the data. Finally, note that each mode,  $\vec{u}_{i,k}$ , will capture a different aspect of the most stable components of the dynamics since  $(v_1)_k (v_2)_k$  corresponds to the projection of the vectors  $v_1, v_2$  into the  $k$ -th coordinate direction. Thus, the delay embedded modes  $\vec{v}_{i,k} = (\hat{v}_{i,k}, \dots, \hat{v}_{i-L,k})^\top$  represent orthogonal components of the most stable Oseledets directions of the dynamics, which in turn are orthogonal to the noise in the limit of large  $L$ .

We now briefly demonstrate the effect of the biased geometry on a simple example. We will generate a stochastic dynamics on the unit circle by first numerically integrating the one-dimensional SDE,

$$d\theta = (2 + \sin(\theta)) dt + \sqrt{0.1} dW_t, \quad (13)$$

and then mapping the intrinsic variable  $\theta$  onto the unit circle embedded in  $\mathbb{R}^2$  by the map  $\theta \mapsto (x(\theta), y(\theta))^\top = (\cos(\theta), \sin(\theta))^\top$ . In this example we use discrete time step  $\Delta t = 0.1$  to produce 10000 samples  $\theta_i$  of the system (13) which are mapped into the plane as  $(x_i, y_i)^\top = (\cos(\theta_i), \sin(\theta_i))^\top$ . We then generate noisy observations  $(\tilde{x}_i, \tilde{y}_i)^\top = (x_i, y_i)^\top + \eta_i$  where  $\eta_i$  are independently sampled from a two dimensional Gaussian distribution with mean zero and covariance matrix  $\frac{1}{20} I_{2 \times 2}$ .

Our goal in this example is to show how the delay-embedding can reduce the influence of noise on a diffusion model by biasing the geometry. We use the first 5000 noisy data points  $(\tilde{x}_i, \tilde{y}_i)^\top$  to train multiple diffusion models, each time applying a different number of delays before building the diffusion model. We also train a diffusion model using the clean data set  $(x_i, y_i)^\top$ . For each diffusion model, we apply the Bayesian filter developed in Section 2.3 below to generate initial conditions for forecasting from the noisy observations  $(\tilde{x}_i, \tilde{y}_i)^\top$  in the verification period  $i = 5001, \dots, 10000$ . Each initial condition in the verification period is forecast for a total of 500 forecast steps (50 time units) and the RMSE between the forecast and the truth is averaged over the verification period. The RMSE as a function of the number of forecast steps for each model is shown in Figure 1, along with the forecast of the  $x$  coordinate at the 50 step lead time, compared to the true value of  $x$ .

While no amount of delays is able to match the diffusion model built using the clean data set, the improvement in forecast skill is significant as the number of delays increases. We should note that since this model is intrinsically one dimensional, the most stable component of the dynamics represents the entire dynamical system. This implies that projecting on the most stable component can only improve the forecasting skill for this example. For high-dimensional

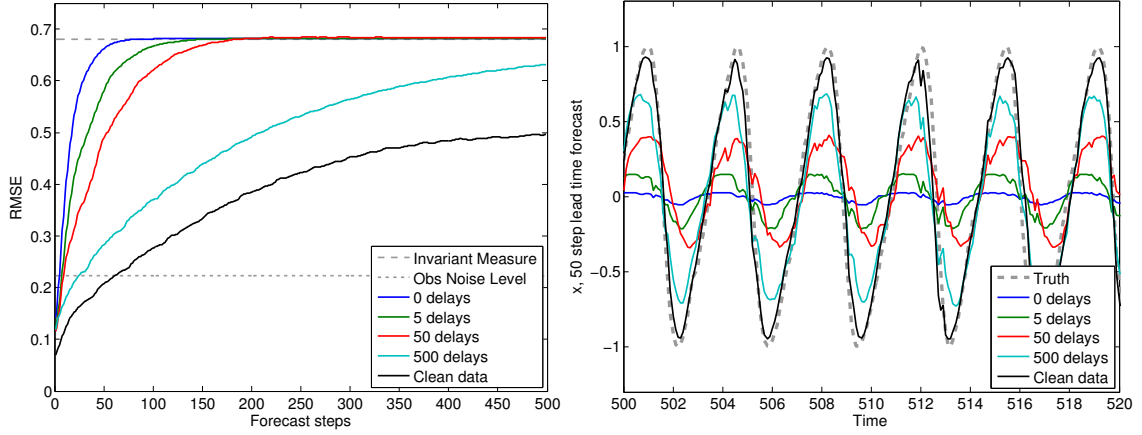


Figure 1: The effect of increasing the number of lags on the diffusion model learned from noisy observations is shown in terms of the forecast RMSE (left) and the 50 step lead time forecast (right).

examples we suspect that as the number of delays becomes very high, the projection onto the most stable component will remove information which is important to the forecast. Therefore, in general we expect to see a tradeoff where small numbers of delays can remove the influence of noise, but extremely large numbers of delays may actually degrade the forecast by projecting away valuable components of the dynamics.

### 2.3. Bayesian filter

In Step 2 of the forecasting algorithm described in Section 2.1, we assume that the initial distributions for forecasting are given. In practice, however, we only have the noisy data  $\hat{v}_{i,k}$ , and the goal of this section is to develop a numerical method to determine the initial distribution. We will use a Bayesian filter to iteratively combine the information from the noisy data with the diffusion forecast. For the remainder of the paper we will be working with a single mode  $\hat{v}_{i,k}$  and so we will drop the subscript  $k$  for convenience.

If the observations were continuous (i.e. for  $\tau$  very small), we could filter the noisy observations  $\hat{v}_i$  using the Zakai equation projected onto the basis  $\{\varphi_j\}$ , and this was the approach taken in [22]. However, since the observation time may be long, we will take a different approach here and apply a Bayesian update at discrete finite time step. Explicitly, we want to find the posterior density  $p(\hat{u}_i | \hat{v}_s, s \leq i)$  by updating the previous posterior density  $p(\hat{u}_{i-1} | \hat{v}_s, s \leq i-1)$  with the current noisy observation  $\hat{v}_i$ .

Our Bayesian filter will follow the standard predictor-corrector approach. Given an initial density at the previous time,  $p(\hat{u}_{i-1} | \hat{v}_s, s \leq i-1)$ , the first step is to compute the prior forecast density with our nonparametric model,

$$p(\hat{u}_i | \hat{v}_s, s < i) = \sum_{j=1}^M c_j(t_i) \varphi_j(\hat{u}) p_{eq}(\hat{u}) = \sum_{j=1}^M (\hat{A}c(t_{i-1}))_j \varphi_j(\hat{u}) p_{eq}(\hat{u}),$$

where the coefficients  $c_j(t_{i-1}) = \langle p(\hat{u}_{i-1} | \hat{v}_s, s \leq i-1), \varphi_j \rangle$  represent the posterior density at time  $t_{i-1}$ , projected in diffusion coordinates,  $\varphi_j$ . Since the noisy observation  $\hat{v}_i$  has the form  $\hat{v}_i = \hat{u}_i + \eta_i$  where  $\eta_i \sim \mathcal{N}(0, \hat{R})$ , the likelihood function for the noisy observation is,

$$p(\hat{v}_i | \hat{u}_i) \propto \exp(-\|\hat{v}_i - \hat{u}_i\|_{\hat{R}}^2/2).$$

We can now assimilate the observation  $\hat{v}_i$  by using Bayes law to combine the prior forecast and the likelihood function above as follows,

$$p(\hat{u}_i | \hat{v}_s, s \leq i) \propto p(\hat{u}_i | \hat{v}_s, s < i) p(\hat{v}_i | \hat{u}_i). \quad (14)$$

By computing the product (14) we can find the desired posterior at time  $t_i$ , up to the normalization factor. We estimate the normalization factor,  $Z$ , as a Monte-Carlo integral,

$$Z = \sum_{l=1}^N \frac{p(\hat{u}_l | \hat{v}_s, s < i) p(\hat{v}_i | \hat{u}_l)}{p_{eq}(\hat{u}_l)} \approx \int_{\mathcal{M}} p(\hat{u} | \hat{v}_s, s < i) p(\hat{v}_i | \hat{u}) dV(\hat{u}),$$

and dividing the product in (14) by  $Z$  we recover the posterior at time  $t_i$ .

In order to initialize this iterative procedure, we spin up the filtering process for few assimilation cycles starting from the invariant measure  $p(\hat{u}_0) = p_{eq}(\hat{u})$ . To obtain initial density  $p(\hat{u}_i | \hat{v}_s, s \leq i)$ , which is used for forecasting starting at time  $t_i$ , we run this Bayesian filter, starting from  $t_{i-S}$  to allow for  $P$  steps of spin-up with the invariant distribution as the initial density,  $p(\hat{u}_{i-S} | \hat{v}_s, s \leq i - P) = p_{eq}(\hat{u})$ . In our numerical simulations below, we used a very short spin-up time of  $P = 10$  steps.

### 3. Forecasting an energy-conserving turbulent mode

As our first numerical example, we consider forecasting the first Fourier component of the Truncated Burgers-Hopf (TBH) model. The TBH model is essentially a Fourier Galerkin approximation to the inviscid Burgers equation which has intrinsic stochastic dynamics with strong evidence of ergodicity and mixing for large enough degrees of freedom [27, 28]. Moreover, it was shown that this model conserves energy and has a family of Gaussian invariant measures with mean zero and constant equipartition energy. The TBH model is described by the following quadratically nonlinear equation for the complex Fourier modes,  $\hat{u}_k$ ,  $1 \leq |k| \leq \Lambda$ :

$$\frac{d\hat{u}_k}{dt} = -\frac{ik}{2} \sum_{\substack{k+p+q=0 \\ 1 \leq |k| \leq \Lambda}} \hat{u}_p^* \hat{u}_q. \quad (15)$$

We simulate the truth, following [28], by setting  $\Lambda = 50$  and numerically integrating the TBH model with a pseudo-spectral method combined with the fourth-order Runge-Kutta time discretization with small enough time step  $\delta t = 2.5 \times 10^{-4}$  to conserve energy with small relative error.

Our choice to model only the first mode is motivated by the fact that this mode has the longest autocorrelation time [27, 28] and, thus, the largest statistical memory of the dynamics. This fact implies that designing reduced models for predicting this mode alone can be difficult since one has to account for the nontrivial memory terms resulting from the interactions between the first mode and remaining unresolved modes. For this test case, it was shown that these memory terms can be modeled with a stochastic parameterization that involves a physics constrained multi-level regression parametric model [16]. In particular, the proposed stochastic parametric model of [16] for this specific problem is given by the following equation,

$$\begin{aligned} \frac{d\hat{u}_1}{dt} &= -ia\hat{u}_1^* \hat{u}_2^* + a_{11}\hat{u}_1 + a_{12}\hat{u}_2, \\ \frac{d\hat{u}_2}{dt} &= ia(\hat{u}_1^*)^2 + a_{11}\hat{u}_1 + a_{22}\hat{u}_2 + r, \\ \frac{dr}{dt} &= a_{31}\hat{u}_1 + a_{32}\hat{u}_1 + a_{33}r + \sigma \dot{W}. \end{aligned} \quad (16)$$

This choice of stochastic model can be obtained by truncating the TBH model up to the first two modes, in order to respect the energy conservation,  $E = \frac{1}{2}(|\hat{u}_1|^2 + |\hat{u}_2|^2)$ , of the quadratic terms. One then adds linear dissipation terms to the first two modes and a colored noise  $r$  to represent the unresolved (or truncated) modes, where  $\dot{W}$  is the standard physicists notation for white noise.

In [16], they applied an adaptive Kalman filtering method to estimate the parameters  $\{a, a_{ij}, \sigma, \hat{R}\}$  in (16) from noisy observations of  $\hat{u}_1$  at time step  $\Delta t = 0.01$  with observations noise variance of  $\hat{R} = 10\% \text{Var}[\hat{u}_1]$ , without knowing  $\hat{R}$ . It was shown in [16] that the resulting model produces remarkably accurate equilibrium statistics when compared to the true signal in a long-time integration of the model in (16). Here, we will compare the forecasting skill of this parametric model and that of the proposed nonparametric model discussed in Section 2.

Given 18000 noisy observations  $\{\hat{v}_i = \hat{u}_{i,1} + \eta_i, \eta_i \sim \mathcal{N}(0, \hat{R})\}$ , at discrete time step  $\Delta t = t_i - t_{i-1} = 0.1$  (10 times longer than the one used to parameterize (16)), we use the first 10000 noisy data points,  $\hat{v}_i$ , to train the diffusion model and the remaining  $V = 8000$  data points to verify the prediction skill. In this numerical experiment, the delay embedding is applied with an arbitrarily chosen lag  $L = 20$  and the nonparametric model is resolved with  $M = 5000$  modes. To see whether this lag size is large enough to smooth out the noise, yet small enough that it does not project away the important unresolved modes that we want to model, we also include an experiment in which we train the diffusion model with the perfect (meaning the noise-free) data set. To diagnose the prediction skill, we use the standard Root-Mean-Squared Error (RMSE) and pattern correlation skill defined as follows,

$$RMS = \left( \frac{1}{V} \sum_{i=1}^V (u_i^t - u_i^f)^2 \right)^{1/2}, \quad (17)$$

$$PC = \frac{\frac{1}{V} \sum_{i=1}^V (u_i^t - \bar{u}^t)(u_i^f - \bar{u}^f)}{\left( \frac{1}{V} \sum_{i=1}^V (u_i^t - \bar{u}^t)^2 \right)^{1/2} \left( \frac{1}{V} \sum_{i=1}^V (u_i^f - \bar{u}^f)^2 \right)^{1/2}}, \quad (18)$$

where  $u_i^t$  denotes the truth and  $u_i^f$  denotes the forecast. Also,  $\bar{u}$  in (17) and (18) denotes the empirical time average over the verification period. For this example, of course  $u_i^t = \hat{u}_{i,1}$  and  $u_i^f$  are the mean forecasts resulting from either the an ensemble forecast of the parametric model in (16) with a relatively small ensemble size of 20 or our nonparametric diffusion model.

In Figure 2, we show the forecasting skill in terms of the measures in (17) and (18). We also show the equilibrium statistical error in the top panel and the true autocorrelation function in the bottom panel (dashes) which can be used to diagnose the predictability limit in the following sense: The equilibrium statistical error is a measure of the average error between the true solution at a randomly selected time and the statistical mean of the steady state; the autocorrelation function shows the rate at which a randomly selected delta function initial condition approaches the invariant measure. Notice that the parametric model gives a relatively accurate forecasting skill, with a pattern correlation which decays exactly like the true autocorrelation function and with error saturating at the invariant measure around a forecast lead time of 1.2. This result gives us confidence that the parametric model in (16) is a reasonable reduced model which can be used to predict this time series. On the other hand, the forecasting skill of the non-parametric model, although not as good as the parametric model, is nearly as accurate in terms of the rate of decay of the RMSE and correlation. First, notice that the forecasting skill of the diffusion model constructed from the noisy data does not differ significantly from that of the noise-free data, the latter having only slightly better skill. This suggests that the choice of lag  $L$  is appropriate in this experiment;  $L$  is large enough to reduce the influence of the noise, but not so large as to project away important information from the reconstructed components of the dynamics. Notice also that the error (and correlation) at the initial time, resulting from the Bayesian filtering, is not as good as the parametric model. This initial error is due to the error in the reconstruction of the density which results from the projection onto the finite dimensional space spanned by  $\{\varphi_j\}_{j=1}^M$ . In particular, since the  $\varphi_j$  are very smooth functions (see Section 2.1), densities which are very close to delta functions will suffer from significant Gibbs phenomenon in the reconstruction. In Figure 3, we compare the mean of the diffusion forecast density to the truth at various forecasting lead times for the time interval [700, 900] within the verification period. Notice that the diffusion forecast has reasonable skill while being less noisy than the forecast produced by the parametric model.

From the numerical results, we conclude that the diffusion forecast is relatively skillful, especially considering that the nonparametric model is effectively a black box and has no knowledge of the true equation whatsoever, whereas the parametric model in (16) was designed using knowledge of the equations of the true dynamics.

#### 4. Forecasting weakly chaotic to fully turbulent modes of Lorenz-96 model

As our second example, we consider Fourier modes of the Lorenz-96 (L96) model [29] in various chaotic regimes. The Lorenz-96 model is defined by the following forced-dissipative nonlinear system of ODEs,

$$\frac{du_\ell}{dt} = (u_{\ell+1} - u_{\ell-2})u_{\ell-1} - u_\ell + F, \quad \ell = 1, \dots, 40, \quad (19)$$

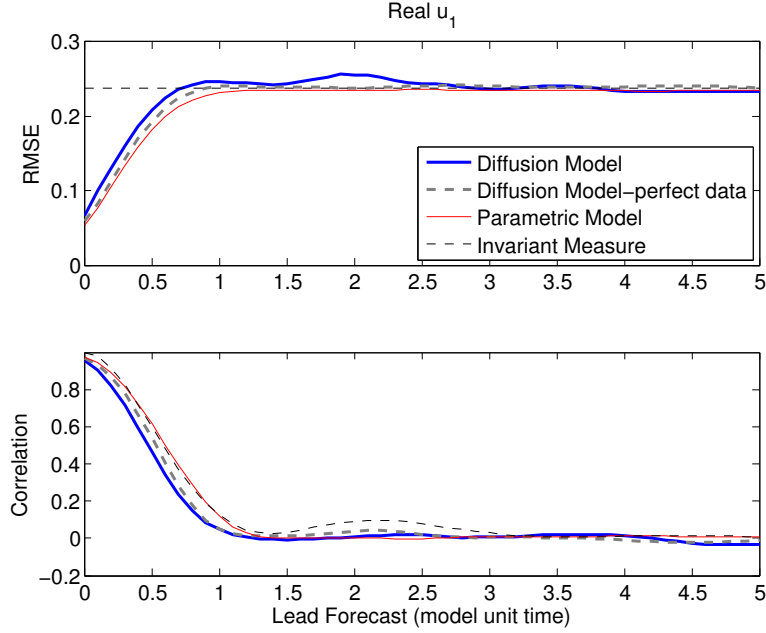


Figure 2: Lead forecasting skill of the real component of the first Fourier mode of TBH model (RMSE *top* and pattern correlation *bottom*). We also show the equilibrium statistical error in the top panel and the true autocorrelation function in the bottom panel (dashes).

The right hand side of (19) consists of an energy conserving quadratic nonlinear advective-like term, a linear dissipation, and a forcing parameter  $F$ . Following [29], we resolve the L96 model at 40 equally spaced grid points with a periodic boundary (represented by the subscript  $l$  being taken modulo 40) in order to mimic weather wave patterns on a midlatitude belt. The statistical behavior of the Fourier modes of this model has been analyzed in [27]; as the forcing parameter  $F$  increases, the model becomes fully turbulent with Gaussian-like modal distributions. In our numerical experiments we will consider three different values of  $F$  following [27]. As reported in [30], for  $F = 6$ , the system is weakly chaotic with largest Lyapunov exponent  $\lambda_1 = 1.02$  and the dimension of the expanding subspace of the attractor is  $N^+ = 12$ . For  $F = 8$ , which is the original choice in [29], the system is strongly chaotic with  $\lambda_1 = 1.74$  and  $N^+ = 13$ . For  $F = 16$ , the system is “fully turbulent” with  $\lambda_1 = 3.94$  and  $N^+ = 16$ . We should also note that when  $F = 6$ , the L96 model has a long memory depth with relatively slow decaying time, this is manifested visibly as a regular dominant westward propagating “weather-like” wave pattern of wavenumber-8. The memory depth becomes shorter as  $F$  increases and the “weather-like” wave pattern becomes less obvious.

In the following numerical experiments we will examine the forecasting skill of the nonparametric modeling approach on a selection of Fourier modes, including those with high and low energies as well as large and small correlation times. For diagnostic purpose, we compare the diffusion forecast on these modes with forecasts from various stochastic modeling approaches that were designed as cheap filter models: the Mean Stochastic Model (MSM) [31, 17] and the stable constrained autoregressive model of order-3 (AR3) [32, 18].

The key idea of MSM is to model each Fourier mode as a complex valued linear stochastic model with additive white noise,

$$du = \lambda u dt + \sigma dW_t, \quad (20)$$

where  $W_t$  is a complex valued Wiener noise with variance  $t$ . The parameters  $\lambda$  and  $\sigma$  are determined by matching the analytical expression for the variance and correlation time at the statistical equilibrium state of the MSM model in (20) with the corresponding empirically estimated statistics from the data (see Chapter 12 of [17] for a detailed formulation). In our implementation below, we will obtain these parameters from the statistics of the noisy dataset,  $\hat{v}$ , during the training period. To generate initial conditions for the forecast, we apply a one-dimensional Kalman filter in order to account for the observation noise in  $\hat{v}$ , and we assume that the noise observation variance,  $\hat{R}$ , is known.

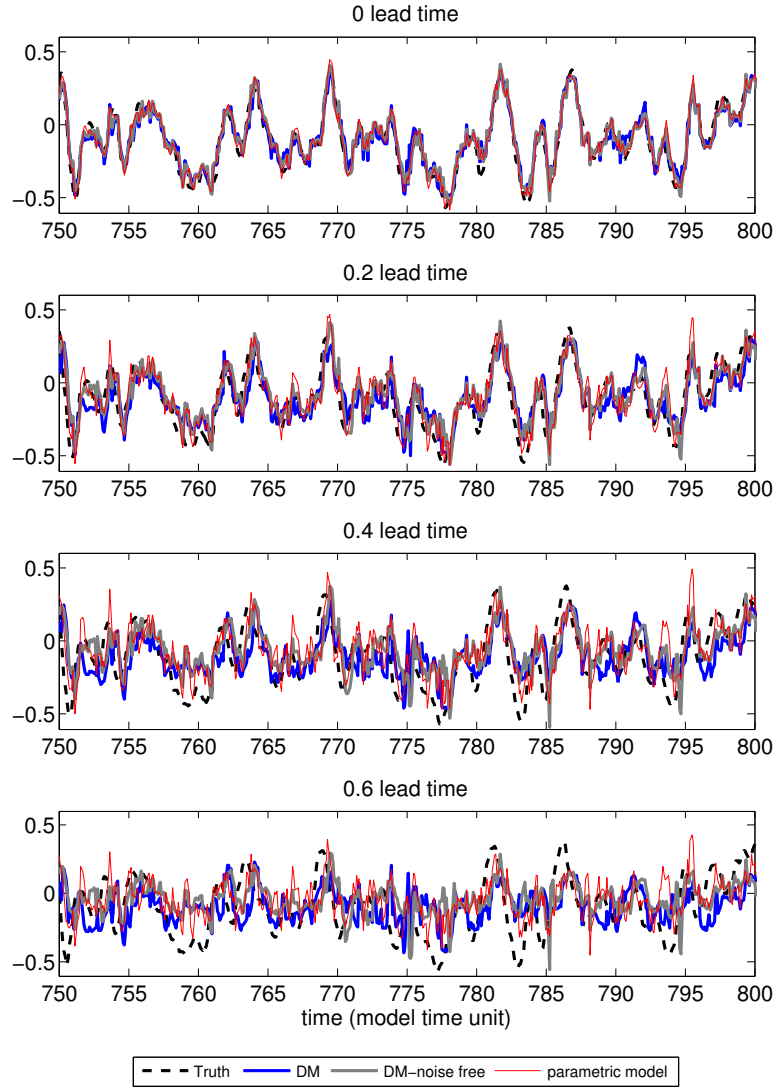


Figure 3: Mean forecast estimates at various lead times at the verification period of [750, 800].

The stable and consistent AR3 model was introduced in [32] as a cheap linear model for filtering turbulent modes with long memory depth, such as the L96 model with  $F = 6$ . The standard AR3 model is given by,

$$u_{m+1} = a_1 u_{m-2} + a_2 u_{m-1} + (1 + a_3) u_m + \eta_m, \quad \eta_m \sim \mathcal{N}(0, Q). \quad (21)$$

The stability of this model is sensitive to the choice of sampling time  $\Delta t = t_{m+1} - t_m$ , especially when a standard linear regression is used to determine the model parameters. In [32], it was shown that one can obtain accurate filtered mean estimates in a linear filtering problem with truth generated from (20) if the parameters  $a_j$  are chosen to satisfy a set of linear algebraic consistency conditions which depends on  $\lambda$  of (20) and the sampling time  $\Delta t$  (see [32] for details). For nonlinear problems, a stable model together with these consistency constraints can be determined by an algebraic method proposed in [18] and this AR3 model is what we use in the examples below. In particular, we will enforce the consistency condition for the AR3 model using the parameter  $\lambda$  obtained from the MSM fit of the noisy data at the training period. To generate initial conditions for the forecast, we apply a three-dimensional Kalman filter to account for the noise observed  $\hat{v}_k$ , assuming that the noise observation variance,  $\hat{R}$ , is known (see e.g. [33, 32, 18] for the details of the AR filter).

In our comparison, we also include the forecast of the perfect model and a simple persistence model for diagnostic purpose. In the perfect model case, we implement a standard ensemble Kalman filter method [34] for the full spatial model (19) using 80 ensemble members, double the dimension of the L96 model. These analysis ensemble provides the initial conditions for an ensemble forecast, again using the true model (19), and the Fourier modes are then computed from the ensemble mean. Finally, we also include the persistence model which fixes the noisy observation,  $\hat{v}_i$ , at the initial time  $t_i$  as the forecast for each lead time.

In our numerical simulations, the true time series are generated by integrating the L96 model with the fourth-order Runge-Kutta method with time step  $\delta t = 1/64$ . Noisy observations are generated by adding random samples of a Gaussian mean zero and variance  $R = 1$  to the true solutions at every grid point at each observation time step  $\Delta t = 1/8$  as in (3), resulting in a noise variance of  $\hat{R} = 1/40$  on each Fourier mode. Given a time series of length 15000 consisting of a single Fourier mode of the noisy observations, we use the first 5000 data points to train the three models (diffusion, MSM, and AR3) and we compute the prediction skill measures on the remaining 10000 data points. For the diffusion forecast model, we used  $L = 5$  lags in the time-delay embedding, and the number of diffusion modes was chosen to be  $M = 2000$ . As in the previous numerical example, we measure the prediction skill in terms of the RMSE and pattern correlation as defined in (17) and (18), respectively.

In Figure 4, we show the RMSE and pattern correlation for forecasting three different Fourier modes of the L96 model in a weakly chaotic regime with  $F = 6$ : mode-8 carries the largest energy with a relatively short correlation time; mode-13 carries low energy with a relatively long correlation time; and mode-18 carries low energy with a relatively short correlation time. Note that the ratios between the observation noise variance  $\hat{R} = 1/40$  and the energy of the modes are very different for these three modes; 4% for mode-8, 11.7% for mode-13, and 30% for mode-18. Therefore, we expect that the noise does not affect the forecasting skill of modes-8 and 13 as much as that of mode-18. To see this, for each mode we include a separate diffusion forecast in which we train the model with perfect (noise-free) data set. Notice that the difference between the forecasting skill of the diffusion model trained with noisy data (solid, black) and that trained with noise-free data (dashes, black) are insignificant for mode-8, still small for mode-13, and somewhat noticeable for mode-18. As discussed in Section 2.2, we expect that it may be possible to improve the estimate for mode-18 by choosing a larger number of lags,  $L$ , but at the same time one should not choose  $L$  too large, since this could project away important components of the dynamics.

Notice that the perfect model provides the best forecast in all modes, as expected. The diffusion forecast performs very well on the energetic mode-8, outperforming the other reduced stochastic models which have very little predictive skill. On mode-13, the diffusion forecast skill is similar to the other reduced stochastic models, and in the very short term forecast none of these models has significantly more skill than the trivial persistence forecast. The MSM and AR3 are not very different; AR3 is slightly better on modes with long correlation time, as expected. The persistence model always produces severely biased forecasts in the long run, since fixing a particular observation will always be worse than predicting the statistical mean in a long term forecast. On mode-18, the MSM and persistence model is slightly better than diffusion forecast (learned from noisy data) in the short term, but then the diffusion forecast performs much better in the intermediate and long-term predictions. On this mode, the AR3 performs the worst.

In Figures 5 and 6, we compare the forecast of each model to the truth for forecast lead times of 1 and 3 model time units respectively. This comparison is made for mode-8 with  $F = 6$  for verification time steps between 1000

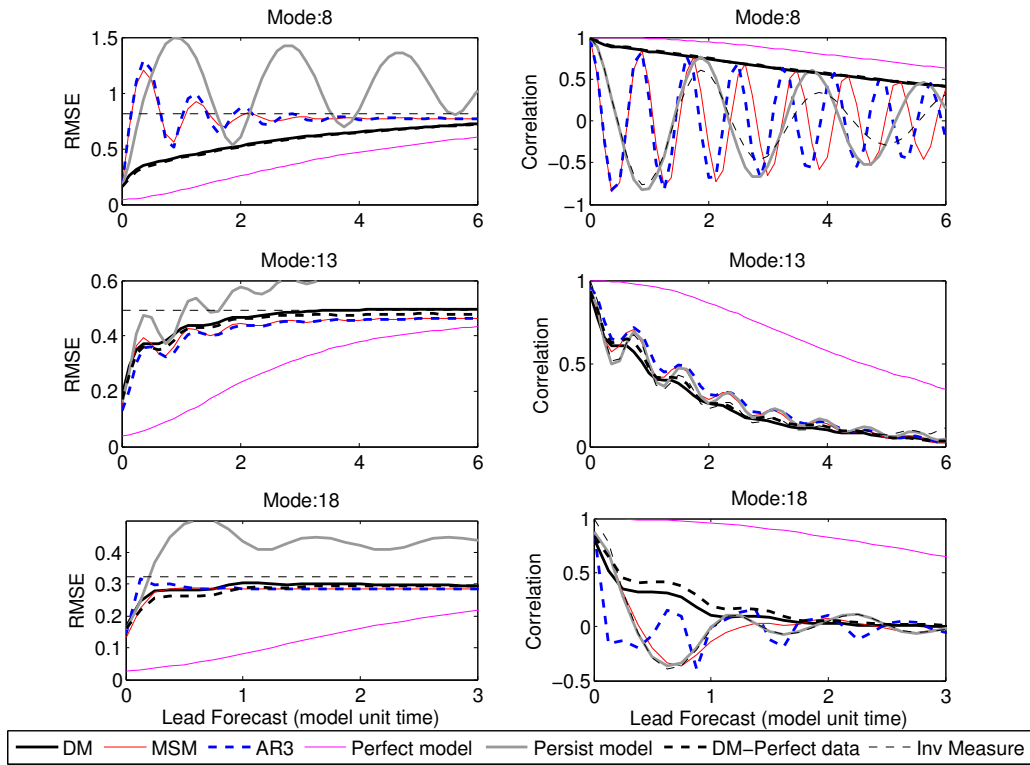


Figure 4: Forecasting skills of L96 models with  $F = 6$  for modes 8 (first row), 13 (second row), and 18 (third row) in terms of RMSE as defined in (17) (first column) and pattern correlation in (18) (second column).

and 1100. Notice that the diffusion model forecasting skill is very similar to that of the true model for both lead times 1 and 3, while the persistence model forecasts are clearly out-of-phase. On the other hand, both the MSM and AR3 models significantly underestimate the amplitude of the signals and in the long term they simply forecast zero, as expected since these forecasts represent the mean of linear unbiased autoregressive models. We also report the forecast skills of the more chaotic regime with  $F = 8$  and the fully turbulent regime with  $F = 16$  in Figure 7. In these cases, the observation noise becomes negligible since the ratio between the observation noise error variance and the energy of this mode decreases as  $F$  increases, namely, 3.2% for  $F = 8$  and 1.5% for  $F = 16$ . Based on the RMSE and pattern correlation measures, these results suggest that the diffusion model still produces the most skillful reduced model compared to the MSM, AR3, and persistence models. While not shown, we also compared the forecast skills on the less energetic modes and found that the forecasting skill of the diffusion, MSM, and AR3 models are more or less comparable, as was found on the corresponding modes shown in Figure 4. We note that other stochastic reduced models exists [35] which would certainly outperform the simple MSM and AR3 models, however those parametric models are derived using knowledge of the true form of the equations (19). Clearly, none of the black box modeling forecasts are close to the perfect model forecast, however the diffusion forecast does offer significant improvement in skill over the other non-physically designed statistical parameterizations considered here.

The numerical experiments in this section suggest that the diffusion model is most skillful when modeling the highly energetic modes. We should note that as  $F$  increases, although mode-8 is still the most energetic mode, the energy is less dominant relative to the other modes in these turbulent regimes [27]. We suspect that this is the main reason that the relative forecasting skill deteriorates when  $F$  is larger, in addition to the more quickly decaying correlation functions in the fully turbulent regimes.

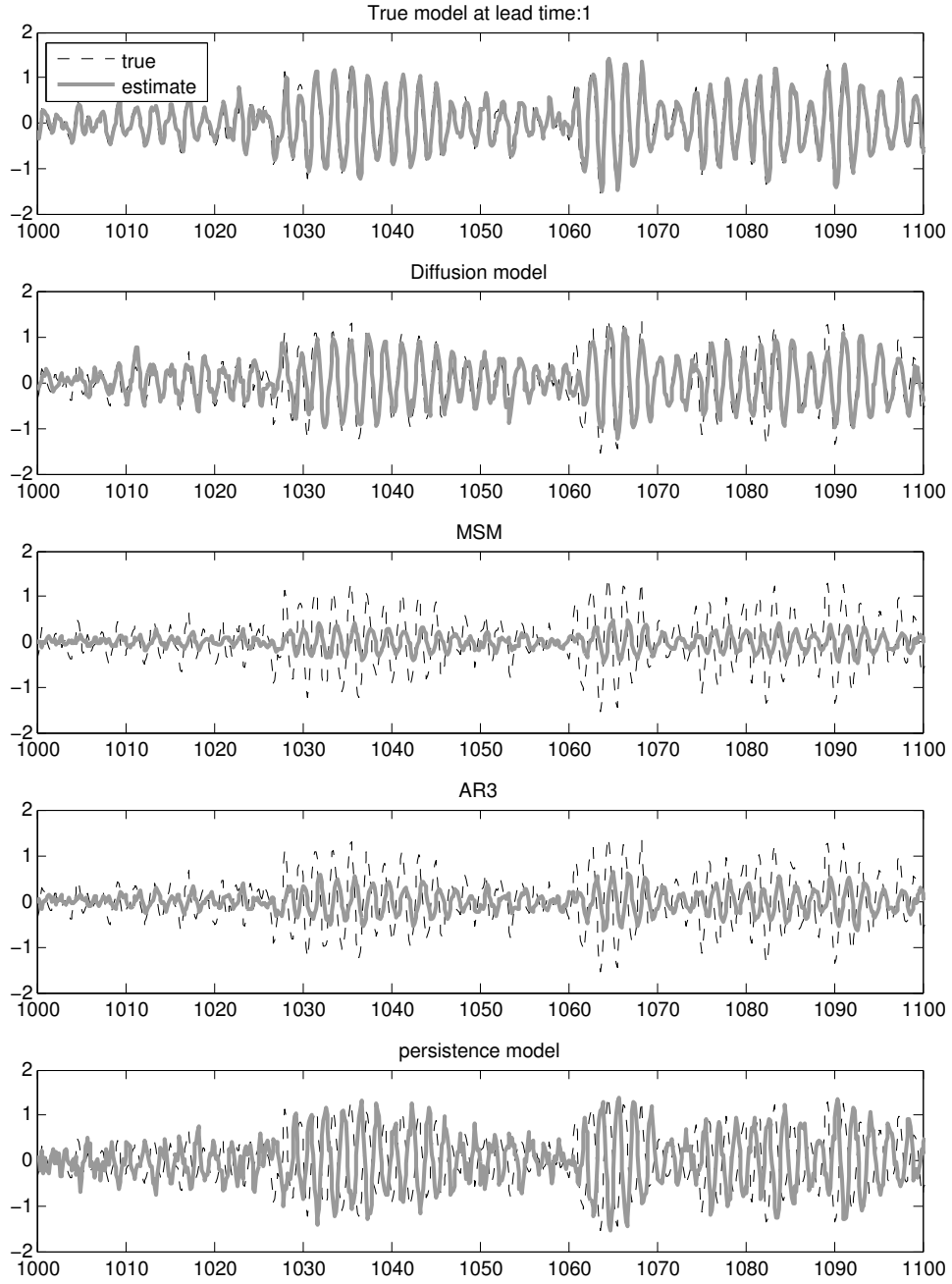


Figure 5: Mean forecast estimates for mode-8 of  $F = 6$  at lead times 1 model time unit at the verification period of  $[1000, 1100]$ . The truth (black dashes) and estimates (gray solid).

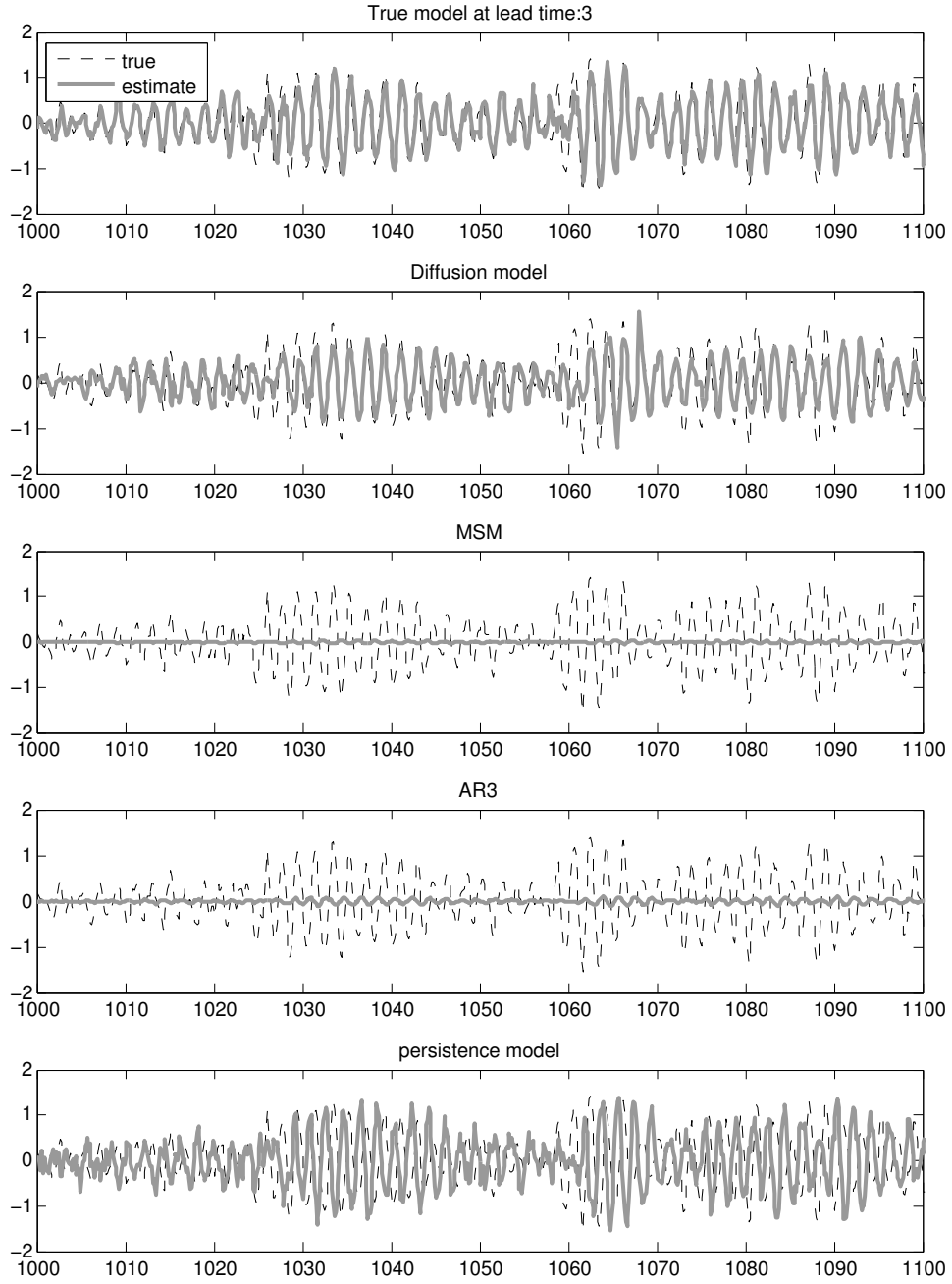


Figure 6: Mean forecast estimates for mode-8 of  $F = 6$  at lead times 3 model time unit at the verification period of [1000, 1100]. The truth (black dashes) and estimates (gray solid).

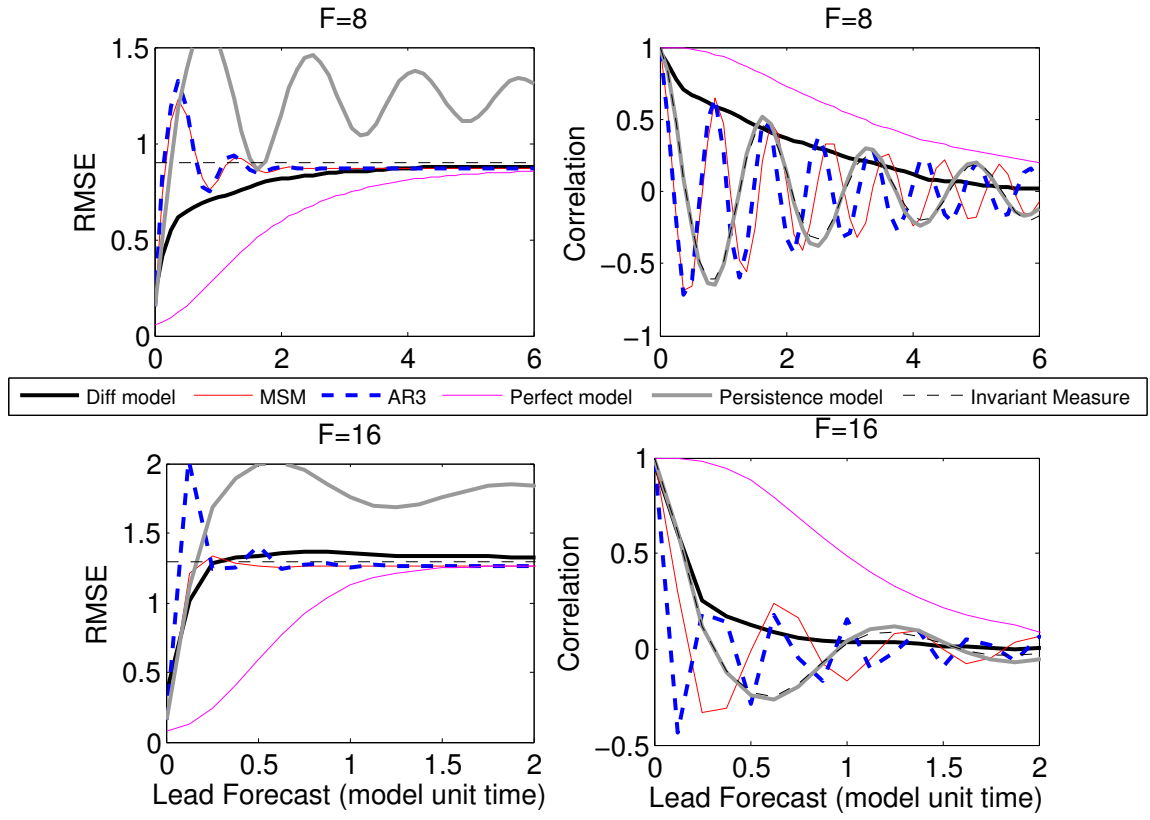


Figure 7: Forecasting skills of L96 models for the most energetic mode 8 of  $F = 8$  (first row) and  $F = 16$  (second row) in terms of RMSE as defined in (17) (first column) and pattern correlation in (18) (second column).

## 5. Forecasting geophysical turbulent modes

In this section, we consider forecasting turbulent modes of the quasigeostrophic model, a prototypical model for midlatitude atmosphere and oceanic dynamics [36]. We consider a two-layer quasigeostrophic (QG) model which is externally forced by a mean shear with streamfunctions

$$\Psi_1 = -Uy, \quad \Psi_2 = Uy, \quad (22)$$

such that it exhibits baroclinic instabilities; the properties of the turbulent cascade has been extensively discussed in this setting (see e.g., [36, 37] and citations in [38]).

The governing equations for the two-layer QG model with a flat bottom, rigid lid, and equal depth  $H$  are given by,

$$\left( \frac{\partial}{\partial t} + U \frac{\partial}{\partial x} \right) q_1 + J(\psi_1, q_1) + \frac{\partial \psi_1}{\partial x} (\beta + k_d^2 U) + \nu \nabla^8 q_1 = 0, \quad (23)$$

$$\left( \frac{\partial}{\partial t} + U \frac{\partial}{\partial x} \right) q_2 + J(\psi_2, q_2) + \frac{\partial \psi_2}{\partial x} (\beta - k_d^2 U) + \kappa \nabla^2 \psi_2 + \nu \nabla^8 q_2 = 0, \quad (24)$$

where  $\psi$  denotes the perturbed streamfunction, subscript 1 corresponds to the upper layer and subscript 2 corresponds to the bottom layer. In (23)-(24),  $\beta$  is the meridional gradient of the Coriolis parameter;  $\kappa$  is the Ekman bottom drag coefficient;  $J(\psi, q) = \psi_x q_y - \psi_y q_x$  is a Jacobian function which acts as nonlinear advection;  $U$  is the zonal mean shear, selected so that the QG equations exhibit baroclinic instability with a turbulent cascade;  $\nu$  is the hyperviscosity coefficient, chosen so that  $\nu \nabla^8 q$  filters out the energy buildup on smaller scales when finite discretization is enforced. The perturbed QG potential vorticity  $q$  is defined for each layer by

$$q_i = \nabla^2 \psi_i + \frac{k_d^2}{2} (\psi_{3-i} - \psi_i). \quad (25)$$

The parameter  $k_d = \sqrt{8}/L_d$  gives the wavenumber associated with radius of deformation, or Rossby radius,  $L_d$  (the scale at which Earth's rotation becomes significant to the dynamics of the system).

In our numerical simulations, the true signal is generated by resolving (23)-(24) with  $128 \times 64 \times 2$  Fourier modes, which corresponds to the  $128 \times 128 \times 2$  grid points. With this resolution, the numerical integration of this turbulent system only needs slightly more than 30,000 state variables because one can always compute  $\psi$  from  $q$  and vice versa, via (25). This model has two important nondimensional parameters:  $b = \beta(L/2\pi)^2/U_o$ , where  $U_o = 1$  is the horizontal non-dimensionalized velocity scale and  $L$  is the horizontal domain size in both directions (we choose  $L = 2\pi$ ), and  $F = (L/2\pi)^2/L_d^2$ , which is inversely proportional to the deformation radius [39]. As in [40, 39], we will consider two cases: one with a deformation radius  $L_d$  such that  $F = 4$  which roughly mimics a turbulent jet in the midlatitude atmosphere and the other case will have a deformation radius  $L_d$  such that  $F = 40$  which roughly mimics a turbulent jet in the ocean. For the ocean case, the QG model in (23)-(24) is numerically very stiff since the term that involves  $k_d^2 = 8/L_d^2 = 8F$  is large.

The large-scale components of this turbulent system are barotropic and for the two-layer model with equal depth, the barotropic streamfunction is defined as an average between the two layers,  $\psi^b = \frac{1}{2}(\psi_1 + \psi_2)$  [36, 40]. In the atmospheric regime with  $F = 4$ , the barotropic flow  $\psi^b$  is dominated by a strong zonal jet while in the ocean case with  $F = 40$  there are transitions between zonal flow and large-scale Rossby waves (e.g., see [41, 39, 42]).

In our numerical examples, we build a diffusion model for the two-dimensional Fourier modes  $\psi_{k,\ell}$  of the barotropic streamfunction,  $\psi_b$ . Following the experiments in [39, 42], our choice to only model this mode is mainly due to the fact that small-scale observations of a turbulent jet are typically not available, especially in the ocean. For diagnostic purposes, we compare the diffusion model to a simple stochastic model with combined white and colored additive and multiplicative noises that was designed as a test filter model for stochastic parameterization in the presence of model error [19, 20]. The governing equation of this model is given as follows,

$$\begin{aligned} d\hat{\psi} &= (-\gamma\hat{\psi} + i\omega\hat{\psi} + b) dt + \sigma dW_t, \\ db &= -(\lambda_b b - \bar{b}) dt + \sigma_b dW_{b,t}, \\ d\gamma &= -(\lambda_\gamma \gamma - \bar{\gamma}) dt + \sigma_\gamma dW_{\gamma,t}, \end{aligned} \quad (26)$$

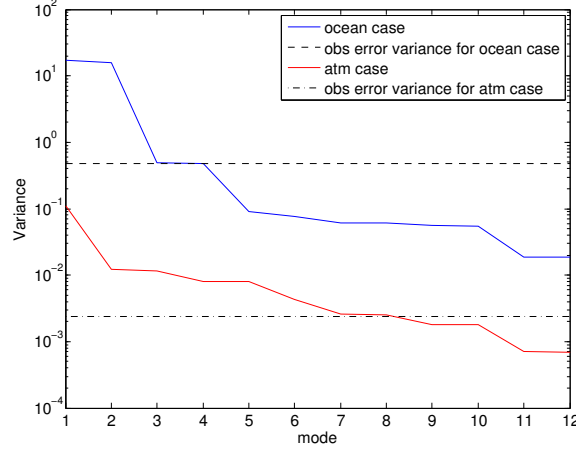


Figure 8: Observation error variance relative to the variances of the 12 Fourier modes corresponding to the 36 regularly spaced observed grid points.

where variable  $\hat{\psi}$  models the Fourier mode of  $\psi_{k,\ell}$ , dropping the horizontal and vertical wave components ( $k, \ell$ ) to simplify the notation. The equation governing  $\hat{\psi}$  represents the interactions between this resolved mode and the remaining unresolved modes by several noise terms. The first noise term,  $\gamma$ , is a real valued multiplicative noise. The second noise term is  $b$ , which is an additive complex valued noise governed by the Ornstein-Uhlenbeck mean reverting SDE. The final noise term is  $\sigma dW_t$ , which is an additive white noise. In (26), the stochastic terms,  $W_t, W_{b,t}$ , are complex valued and  $W_{\gamma,t}$  are real valued Wiener processes. We should note that while the simple system in (26) is nonlinear (due to the multiplicative term  $-\gamma\hat{\psi}$ ), it has analytical statistical solutions [19, 20] which we will utilize in our numerical experiments here.

Implementing a Kalman update on the analytical mean and covariance solutions of (26), is a filtering method known as SPEKF [19, 20, 17], which stands for Stochastic Parameterized Extended Kalman Filter. In our implementation below, we will use SPEKF to generate initial conditions for our forecast at the verification times. In our experiments, we consider two different choices of parameters,  $\{\omega, \sigma, \lambda_b, \bar{b}, \sigma_b, \lambda_\gamma, \bar{\gamma}, \sigma_\gamma\}$ , in (26): The original set of parameters which was empirically chosen to optimize the filtered solutions in [39] and we will refer to this forecasting strategy as SPEKF. The second choice is to use the adaptive Kalman filtering method [43] (which was also used in Section 2 above) in order to extract these parameters from noisy the data set  $\hat{v}$  in the training period. The adaptive Kalman filter will start with the parameters set from the SPEKF as initial guess. We refer to this forecasting strategy, with parameters determined adaptively, as SPEKF-QR.

In our numerical experiments, we generate a time series of 9000 data points at discrete time step  $\Delta t = 1/4$  and added Gaussian noise in the physical space at 36 regularly spaced grid points in the spatial domain as in [39] with noise variance  $R = 25\% \text{Var}(\psi^b)$ . In Fourier modes, the observation error variance is  $\hat{R} = R/36$ ; in Figure 8, we show  $\hat{R}$  relative to the variance of the 12 modes corresponding to the 36 regularly spaced observed grid points for both the atmospheric and ocean regimes. We use the first 5000 noisy data points to train the diffusion and SPEKF-QR models and the remaining 4000 data points to compute the forecasting skill measures. For the diffusion forecast model, we used  $L = 20$  lags in the time-delay embedding and  $M = 2000$  diffusion modes.

In Figures 9 and 10, we show the RMSE and pattern correlation for the first two energetic modes in the atmospheric and ocean regimes, respectively. For the atmospheric regime, the first two energetic modes correspond to the two-dimensional horizontal wave numbers (0, 1) and (1, 1) with explained variances of about 65% and 73%, respectively, so the behavior is dominated by the zonal jet in the horizontal direction [39]. In this regime, notice that the SPEKF-QR produces slightly better forecasts compared to the diffusion model in terms of RMSE and pattern correlation. On the other hand, the SPEKF model, with parameters empirically chosen to optimize the filter accuracy, produces lower forecasting skill after 10 unit times. For the ocean case, the first two energetic modes correspond to the two-dimensional horizontal wave numbers (1, 0) and (0, 1) with explained variance of about 50% and 97%, respectively,

so the behavior is dominated by the Rossby mode  $(1, 0)$  and the zonal jet mode  $(0, 1)$  [39]. In this case, notice that for the first mode both the SPEKF-QR forecasts and the SPEKF forecasts are unstable to the point that the RMSE diverge far beyond the invariant measure. For the second mode, the SPEKF-QR diverges while the SPEKF produces accurate forecasts. The instability in SPEKF-QR and SPEKF were due to the fact that the SPEKF model is statistically stable only if certain algebraic constraints are satisfied (see [44] for the precise definition of these algebraic constraints) and these constraints were not imposed in choosing the SPEKF parameters and also not imposed in the adaptive parameterization scheme which is used to choose the parameters in SPEKF-QR. In contrast, the nonparametric model is always stable and reasonably skillful in terms of RMSE and pattern correlation.

In Figure 11, we show the forecasts of the real component of the second most energetic mode for the ocean case with  $F = 40$  at various lead times 0-5 for the verification period [900, 1000]. Notice that the SPEKF-QR forecasts are very noisy, whereas the diffusion forecast and SPEKF forecast are reasonably accurate. In Figures 12-13, we show the spatial reconstruction of the barotropic streamfunction  $\psi^b$ , based on separately modeling 12 Fourier modes corresponding to the 36 regularly spaced sparse noisy observations in the physical space. We should point out that while the observation noise is relatively small compared to the two most energetic modes (see Figure 8), the less energetic modes are completely overwhelmed by the noise in this numerical experiment and the results below are reported without tuning the lagged embedding dimension,  $L$ . Nevertheless, we find that for the atmospheric case (see Figure 12), the dominant turbulent zonal jet structure of the streamfunction is recovered by all the methods at lead times 4 and 8; the spatial structure of the streamfunctions produced by the diffusion model appears closer to the truth compared to those from the parametric SPEKF-QR and SPEKF models. For the ocean case (see Figure 13, notice that the diffusion model is the only one that produces accurate forecasts.

From the simulations in this section on a nontrivial prototypical example of midlatitude geophysical turbulent flows, we see that the proposed non-parametric diffusion models always produce stable forecasts. Of course, one can always resolve this instability issue by carefully choosing a different parametric form, for example the physics constrained model in (16), in which the stability constraint was implicitly imposed in the parametric equations. Our point here is to show that if the parametric form for the reduced stochastic model is not chosen appropriately, then one can easily run into these instability issues. In contrast, this issue is not encountered in the non-parametric modeling. We also see that the diffusion models can recover the non-trivial structure of streamfunctions despite the fact that each Fourier mode is modeled separately and some of the low energy modes are completely swamped by the observation noise. This encouraging result empirically suggests that predicting the first two most energetic modes that are dominant can overcome the least accurate models of the low energetic modes.

## 6. Summary

In this paper, we apply the nonparametric diffusion modeling approach proposed in [9] to forecast Fourier modes of turbulent dynamical systems given finite set of noisy observations of these modes. Motivated by the results in [12, 13, 14], we build the nonparametric model in the delay embedding coordinates in order to account for the interaction between the resolved and unresolved modes as well as to smooth out the noisy observations. While our theoretical derivation suggests that too large of embedding space dimension,  $L$ , may project away the reconstructed unresolved modes, we empirically verify that larger  $L$  helps to reduce the influence of the observation noise and improves the diffusion model forecasts. The choice of  $L$  should not be so large as to project away the unresolved modes, but large enough to smooth out the noise in the training data set. To initiate each forecast, we apply a simple Bayesian filtering method which accounts for the information of the noisy observations up to the current time. We compare the forecasting skills of the proposed nonparametric model with various modeling approaches including the persistence model, linear autoregressive models of order-1 (MSM) and order-3, the SPEKF model which involves combined additive and multiplicative noise terms, the physics constrained nonlinear regression models which use multi-level colored noises, and the perfect model on various examples. We should note that these numerical results were obtained without any careful tunings of size of the training data set,  $N$ , and the number of eigenfunctions,  $M$  and all the remaining nuisance parameters in the nonparametric model were automatically tuned from the data set as described in [9]. The numerical results empirically suggest that the diffusion model gives more accurate forecasts for modes that are highly energetic with slow decaying correlation functions, which is a characteristic of weakly turbulence.

One important thing to take away from this paper is that although the proposed nonparametric model seems to be competitive with the standard parametric models chosen here, we do not claim that this equation-free approach is

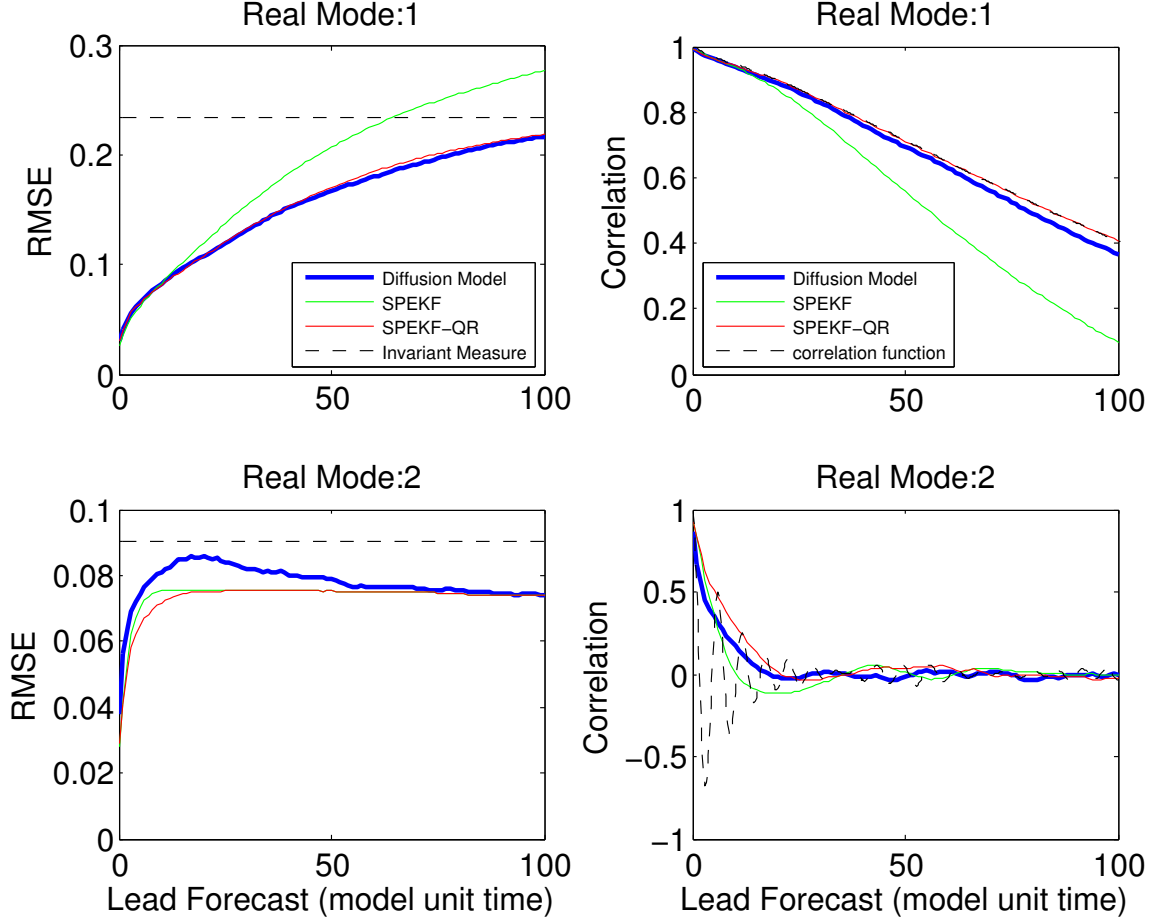


Figure 9: Forecasting skills of the QG models for the real component of the two most energetic modes for the atmospheric regime with  $F = 4$  in terms of RMSE as defined in (17) (first column) and pattern correlation in (18) (second column).

the model to use for everything. If one knows the physics of the underlying dynamics (or the appropriate reduced parametric models, such as equation (16) for modeling the first mode of the TBH example in Section 3 above), then one should use the physics-based model rather than the equation-free diffusion models. Indeed, it was shown rigorously in a simple setup that optimal filtering and accurate statistical prediction can be achieved with an appropriate stochastic parametric modeling of the unresolved scales [8]. However, there are at least two valuable aspects of the nonparametric model we develop here. First, one can use this model to diagnose whether his/her modeling approach is appropriate; we expect that appropriate physics-based models should beat this black box approach. Of course when an appropriate, physically motivated model is not available, then this approach will often outperform adhoc parametric models. Second, in practice it is difficult to guess the appropriate choice of parametric models even if some physics-based knowledge has been imposed in deriving a reduced model (e.g., in turbulent closure problems, one typically uses parameters to represent some small-scale processes or some physical processes which are not well understood). In this situation, we propose to extract the time series of these processes (if possible) and to subsequently build nonparametric models for such processes. This idea is what we refer to as the *semi-parametric modeling* approach and is introduced in the companion paper [45].

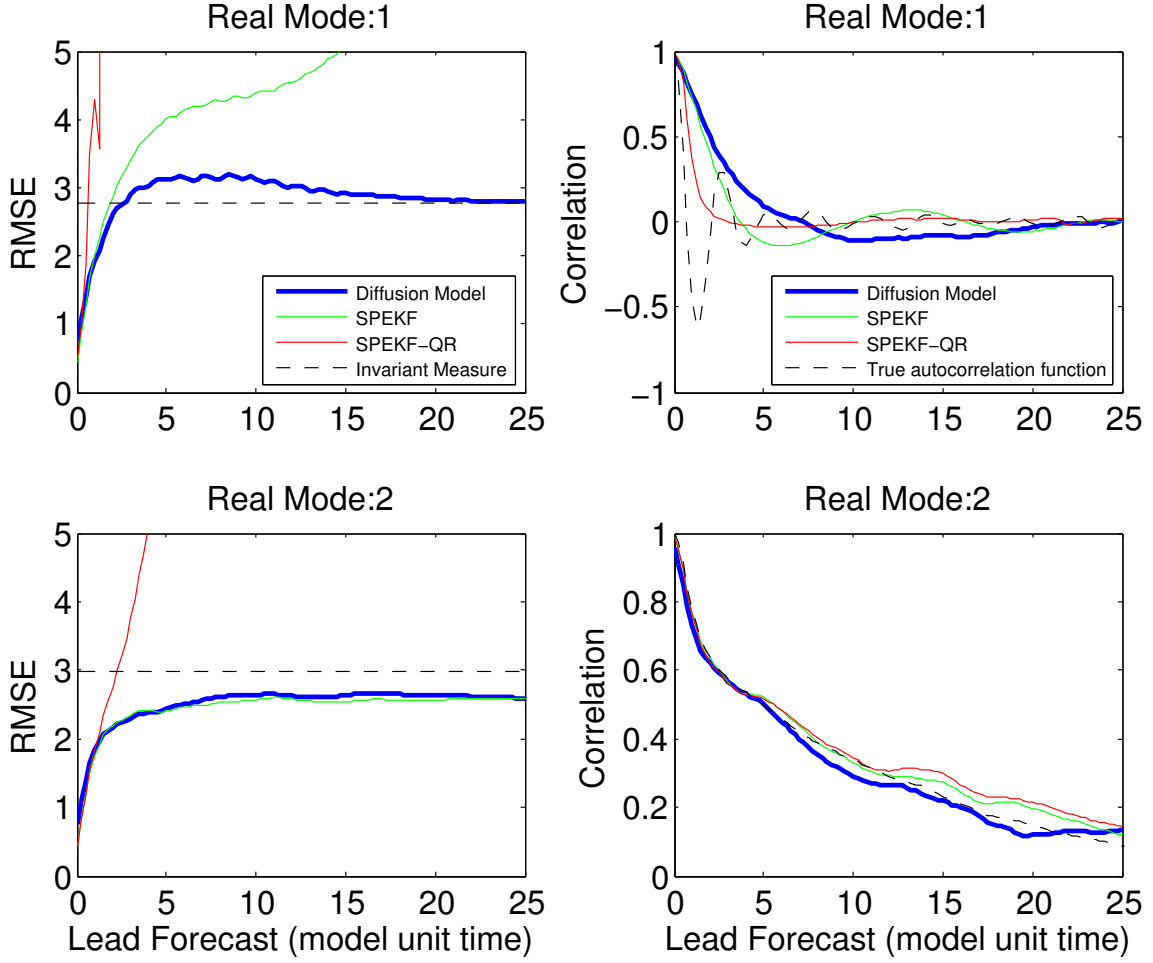


Figure 10: Forecasting skills of the QG models for the real component of the two most energetic modes for the ocean regime with  $F = 40$  in terms of RMSE as defined in (17) (first column) and pattern correlation in (18) (second column).

## Acknowledgment

The research of JH was partially supported by the Office of Naval Research Grants N00014-13-1-0797, MURI N00014-12-1-0912, and the National Science Foundation grant DMS-1317919. TB was partially supported through ONR MURI grant N00014-12-1-0912.

## References

- [1] P. Davidson, *Turbulence: An Introduction for Scientists and Engineers*, OUP Oxford, 2004.
- [2] P. Kramer, A. Majda, E. Vanden-Eijnden, Closure approximations for passive scalar turbulence: A comparative study on an exactly solvable model with complex features, *J. Stat. Phys.* 111 (2003) 565–679.
- [3] T. DelSole, Stochastic model of quasigeostrophic turbulence, *Surveys in Geophysics* 25 (2004) 107–149.
- [4] B. Khouider, J. Biello, A. Majda, A stochastic multicloud model for tropical convection, *Commun. Math. Sci.* 8 (2010) 187 – 216.
- [5] I. Grooms, A. Majda, Efficient stochastic superparameterization for geophysical turbulence, *Proc. Nat. Acad. Sci.* 110 (2013) 4464–4469.
- [6] A. Majda, I. Grooms, New perspectives on superparameterization for geophysical turbulence, *J. Comput. Phys.* 271 (2014) 60 – 77. *Frontiers in Computational Physics Modeling the Earth System*.
- [7] A. Majda, Y. Yuan, Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems, *Discrete Contin. Dyn. Syst. Ser. B* 17 (2012) 1333–1363.

- [8] T. Berry, J. Harlim, Linear theory for filtering nonlinear multi scale systems with model error, *Proc. R. Soc. A* 470 (2014) 2167.
- [9] T. Berry, D. Giannakis, J. Harlim, Nonparametric forecasting of low-dimensional dynamical systems, submitted to *Phys. Rev. Lett.* (2015). <http://arxiv.org/abs/1411.5069v2>.
- [10] R. Coifman, S. Lafon, Diffusion maps, *Appl. Comput. Harmon. Anal.* 21 (2006) 5–30.
- [11] T. Berry, J. Harlim, Variable bandwidth diffusion kernels, *Appl. Comput. Harmon. Anal.* (in press) (2015). doi:10.1016/j.acha.2015.01.001.
- [12] T. Berry, J. R. Cressman, Z. G. Ferenčák, T. Sauer, Time-scale separation from diffusion-mapped delay coordinates, *SIAM J. Appl. Dyn. Syst.* 12 (2013) 618–649.
- [13] D. Giannakis, A. J. Majda, Time series reconstruction via machine learning: Revealing decadal variability and intermittency in the north pacific sector of a coupled climate model, in: *Conference on Intelligent Data Understanding (CIDU)*, Mountain View, California, 2011, pp. 107–117.
- [14] D. Giannakis, A. J. Majda, Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability, *Proc. Nat. Acad. Sci.* 109 (2012) 2222–2227.
- [15] T. Sauer, J. Yorke, M. Casdagli, Embedology, *J. Stat. Phys.* 65 (1991) 579–616.
- [16] A. Majda, J. Harlim, Physics constrained nonlinear regression models for time series., *Nonlinearity* 26 (2013) 201–217.
- [17] A. Majda, J. Harlim, *Filtering Complex Turbulent Systems*, Cambridge University Press, UK, 2012.
- [18] J. Harlim, H. Hong, J. Robbins, An algebraic method for constructing stable and consistent autoregressive filters, *J. Comput. Phys.* 283 (2015) 241 – 257.
- [19] B. Gershgorin, J. Harlim, A. Majda, Test models for improving filtering with model errors through stochastic parameter estimation, *J. Comput. Phys.* 229 (2010) 1–31.
- [20] B. Gershgorin, J. Harlim, A. Majda, Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation, *J. Comput. Phys.* 229 (2010) 32–57.
- [21] A. Majda, M. Grote, Mathematical test models for superparametrization in anisotropic turbulence, *Proc. Nat. Acad. Sci.* 2106 (2009) 5470–5474.
- [22] T. Berry, J. Harlim, Nonparametric uncertainty quantification for stochastic gradient flows, submitted to *SIAM/ASA J. Uncertainty Quantification* (2014). <http://arxiv.org/abs/1407.6972v2>.
- [23] G. Gottwald, J. Harlim, The role of additive and multiplicative noise in filtering complex dynamical systems, *Proc. R. Soc. A* 469 (2013).
- [24] F. Takens, Detecting strange attractors in turbulence, in: D. Rand, L.-S. Young (Eds.), *In: Dynamical Systems and Turbulence*, Warwick, Eds. Rand, D. and Young, L.-S., volume 898 of *Lecture Notes in Mathematics*, Springer Berlin / Heidelberg, 1981, pp. 366–381. URL: <http://dx.doi.org/10.1007/BFb0091924>.
- [25] J. Stark, Delay embeddings for forced systems. I. Deterministic forcing, *J. Nonlin. Sci.* 9 (1999) 255–332.
- [26] J. Stark, D. Broomhead, M. Davies, J. Huke, Delay embeddings for forced systems. II. Stochastic forcing, *J. Nonlin. Sci.* 13 (2003) 519–577.
- [27] A. Majda, R. Abramov, M. Grote, Information theory and stochastics for multiscale nonlinear systems, CRM Monograph Series v.25, American Mathematical Society, Providence, Rhode Island, USA, 2005.
- [28] A. Majda, I. Timofeyev, Statistical mechanics for truncations of the Burgers-Hopf equation: A model for intrinsic behavior with scaling, *Milan Journal of Mathematics* 70 (2002) 39–96.
- [29] E. Lorenz, Predictability - a problem partly solved, in: *Proceedings on predictability*, held at ECMWF on 4-8 September 1995, 1996, pp. 1–18.
- [30] R. Abramov, A. Majda, Blended response algorithm for linear fluctuation-dissipation for complex nonlinear dynamical systems, *Nonlinearity* 20 (2007) 2793–2821.
- [31] J. Harlim, A. Majda, Filtering nonlinear dynamical systems with linear stochastic models, *Nonlinearity* 21 (2008) 1281–1306.
- [32] E. Bakunova, J. Harlim, Optimal filtering of complex turbulent systems with memory depth through consistency constraints, *J. Comput. Phys.* 237 (2013) 320–343.
- [33] E. Kang, J. Harlim, Filtering nonlinear spatio-temporal chaos with autoregressive linear stochastic models, *Phys. D* 241 (2012) 1099 – 1113.
- [34] B. Hunt, E. Kostelich, I. Szunyogh, Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter, *Phys. D* 230 (2007) 112–126.
- [35] T. Sapsis, A. Majda, Statistically accurate low-order models for uncertainty quantification in turbulent dynamical systems, *Proc. Nat. Acad. Sci.* 110 (2013) 13705–13710.
- [36] R. Salmon, *Lectures on geophysical fluid dynamics*, volume 378, Oxford University Press, 1998.
- [37] G. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*, Cambridge University Press, Cambridge, U.K., 2006.
- [38] S. Smith, G. Boccaletti, C. Henning, I. Marinov, C. Tam, I. Held, G. Vallis, Turbulent diffusion in the geostrophic inverse cascade, *J. Fluid Mech.* 469 (2002) 13–48.
- [39] J. Harlim, A. Majda, Filtering turbulent sparsely observed geophysical flows, *Mon. Wea. Rev.* 138 (2010) 1050–1083.
- [40] R. Kleeman, A. Majda, Predictability in a model of geophysical turbulence, *J. Atmos. Sci.* 62 (2005) 2864–2879.
- [41] A. Majda, C. Franzke, B. Khouider, An applied mathematics perspective on stochastic modelling for climate., *Philos. Transact. A Math. Phys. Eng. Sci.* 366 (2008) 2429–2455.
- [42] K. Brown, J. Harlim, Assimilating irregularly spaced sparsely observed turbulent signals with hierarchical bayesian reduced stochastic filters, *J. Comput. Phys.* 235 (2013) 143 – 160.
- [43] J. Harlim, A. Mahdi, A. Majda, An ensemble kalman filter for statistical estimation of physics constrained nonlinear regression models, *J. Comput. Phys.* 257, Part A (2014) 782 – 812.
- [44] M. Branicki, B. Gershgorin, A. Majda, Filtering skill for turbulent signals for a suite of nonlinear and linear extended kalman filters, *J. Comput. Phys.* 231 (2012) 1462 – 1498.
- [45] T. Berry, J. Harlim, Semiparametric forecasting: Correcting low-dimensional model error in high-dimensional parametric models, submitted to *J. Comput. Phys.* (2015).

## Appendix A. Variable Bandwidth Diffusion Map Algorithm

For data  $\{x_i\}_{i=1}^N$  lying on a smooth manifold  $\mathcal{M}$  (with  $L = 0$  lags), or for data given by a generic observation of a dynamical system on a smooth manifold (with  $L$  sufficiently large), the algorithm given below will provably estimate the eigenfunctions  $\varphi_j$  of the generator  $\hat{\mathcal{L}}$  of (5). The estimate  $\hat{\mathbf{L}}$  of the operator

$$\hat{\mathcal{L}} = -\nabla U \cdot \nabla + \Delta, \quad (\text{A.1})$$

where  $U = -\log(p_{eq})$ , is up to a pointwise error of order  $\mathcal{O}\left(\epsilon, \frac{q(x_i)^{(1-d\beta)/2}}{\sqrt{N}\epsilon^{2+d/4}}, \frac{\|\nabla f(x_i)\|q(x_i)^{-c_2}}{\sqrt{N}\epsilon^{1/2+d/4}}\right)$ , where  $q(x)$  denotes the sample distribution, which by ergodicity is exactly the invariant measure  $p_{eq}$  of the underlying dynamics. We note that  $c_2 = d(1/4 - d/2) < 0$  for the choice  $\beta = -1/2$  and  $\alpha = -d/4$  so that the final error term is bounded even when  $q(x_i)$  is arbitrarily close to zero. For a derivation of these facts see [11] and for a brief overview see the supplemental material of [9]. Below, we only give a step-by-step cookbook for constructing the eigenfunctions  $\{\varphi_j\}$  of  $\hat{\mathcal{L}}$ .

1. Choose delay weight  $\kappa > 0$  and number of lags  $L$ . In our numerical experiments, we use  $\kappa = 0$  and  $L = 20$ .
2. Let  $x_i = x(t_i) \in \mathbb{R}^n$  be a time series with  $i = 1, \dots, N + L$  data points. In our case, the data set are noisy Fourier modes  $\hat{v}_{i,k}$ .
3. For  $i = 1 + L, \dots, N + L$  form the state vector,

$$y_i = [x_i, e^{-\kappa}x_{i-1}, \dots, e^{-L\kappa}x_{i-L}]^T \in \mathbb{R}^{n(L+1)}.$$

4. For each  $i$  find the  $k$ -nearest neighbors of  $y_i$  in  $\mathbb{R}^{n(L+1)}$ , let their indices be  $I(i, j)$  for  $j = 1, \dots, k$  ordered by increasing distance. Here we used  $k = 512$ .
5. Form a sparse  $(N - L) \times (N - L)$  matrix with  $(N - L)k$  nonzero entries given by

$$d(i, I(i, j)) = \|y_i - y_{I(i, j)}\|.$$

6. Define the ad hoc bandwidth function  $\hat{\rho}_i = \sqrt{\sum_{j=1}^k d(i, I(i, j))^2}$ .

7. Automatically tune the bandwidth for the kernel density estimate.

(a) Let  $\epsilon_l = 2^l$  for  $l = -30, -29, \dots, 9, 10$ .

(b) Compute  $T_l = \sum_{i,j=1}^N \exp\left(\frac{-d(i,j)}{2\epsilon_l \hat{\rho}_i \hat{\rho}_j}\right)$ .

(c) Estimate the local power law  $T_l = \epsilon_l^a$  at each  $l$  by  $a_l = \frac{\log T_l - \log T_{l-1}}{\log \epsilon_l - \log \epsilon_{l-1}}$ .

(d) Choose  $\epsilon = \arg\max_{\epsilon_l} \{a_l\}$ .

(e) Estimate the intrinsic dimension  $d = 2 \max_{\epsilon_l} \{a_l\}$ .

8. Form the density estimate  $q_i = q(x_i) = (2\pi\epsilon\hat{\rho}_i^2)^{-d/2} N^{-1} \sum_{j=1}^N \exp\left(\frac{-d(i,j)}{2\epsilon\hat{\rho}_i\hat{\rho}_j}\right)$ .

9. Define the bandwidth function  $\rho_i = q_i^\beta$ . We use  $\beta = -1/2$  and  $\alpha = -d/4$ .

10. Repeat step 7 to choose the bandwidth  $\epsilon$  with  $T_l = \sum_{i,j=1}^N \exp\left(\frac{-d(i,j)}{4\epsilon\rho_i\rho_j}\right)$ .

11. Form the sparse kernel matrix  $K(i, I(i, j)) = \exp\left(\frac{-d(i, I(i, j))}{4\epsilon\rho_i\rho_{I(i, j)}}\right)$ .

12. Form the symmetric matrix  $K^S = (K + K^T)/2$ .

13. Form the diagonal normalization matrix  $D_{ii} = \sum_{j=1}^N K_{ij}^S / q_i^{d\beta}$ .

14. Normalize to form matrix  $K_\alpha^S = D^{-\alpha} K^S D^{-\alpha}$ .

15. Form the diagonal normalization matrix  $(D_\alpha)_{ii} = \sum_{j=1}^N (K_\alpha^S)_{ij}$ .

16. Form the diagonal matrices  $\hat{D}_{ii} = 2\epsilon q_i^{d\beta}$  and  $P = \hat{D}^{1/2} D_\alpha^{1/2}$ .

17. Form the symmetric matrix  $\hat{\mathbf{L}} = P^{-1} K_\alpha^S P^{-1} - \hat{D}^{-1}$ .

18. Find the smallest magnitude eigenvalues  $\lambda_j$  and associated eigenvectors  $\varphi_j$  of  $\hat{\mathbf{L}}$ .

19. Normalize the eigenvectors so that  $\frac{1}{N} \sum_{l=1}^N \varphi_j(x_l) = 1$ .

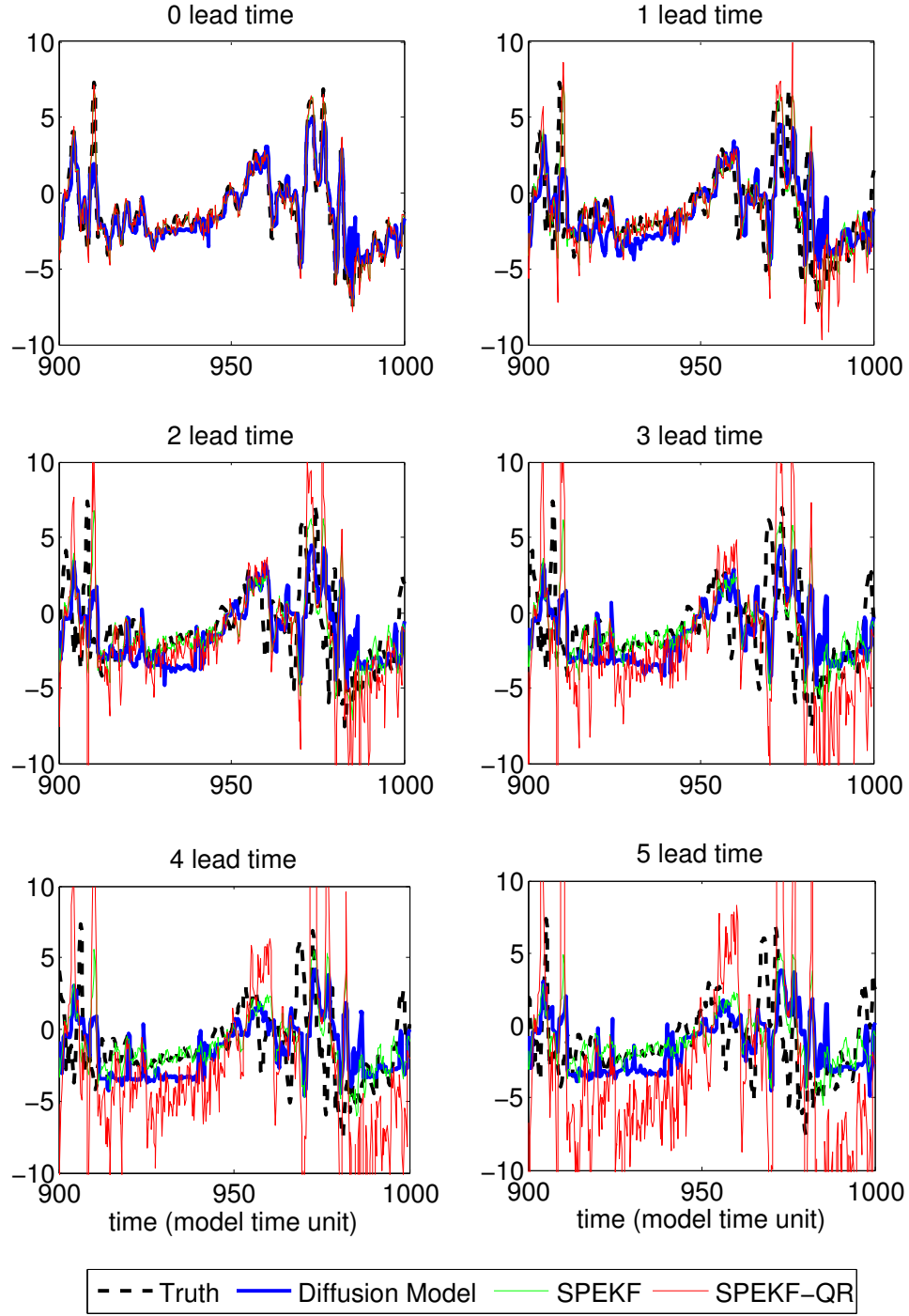


Figure 11: Forecasts of the QG models for the real component of the second most energetic mode for the ocean regime with  $F = 40$  at lead times 0 – 5 on the verification interval [900, 1000]. Truth (black dashes), diffusion model (blue), SPEKF (green), SPEKF-QR (red).

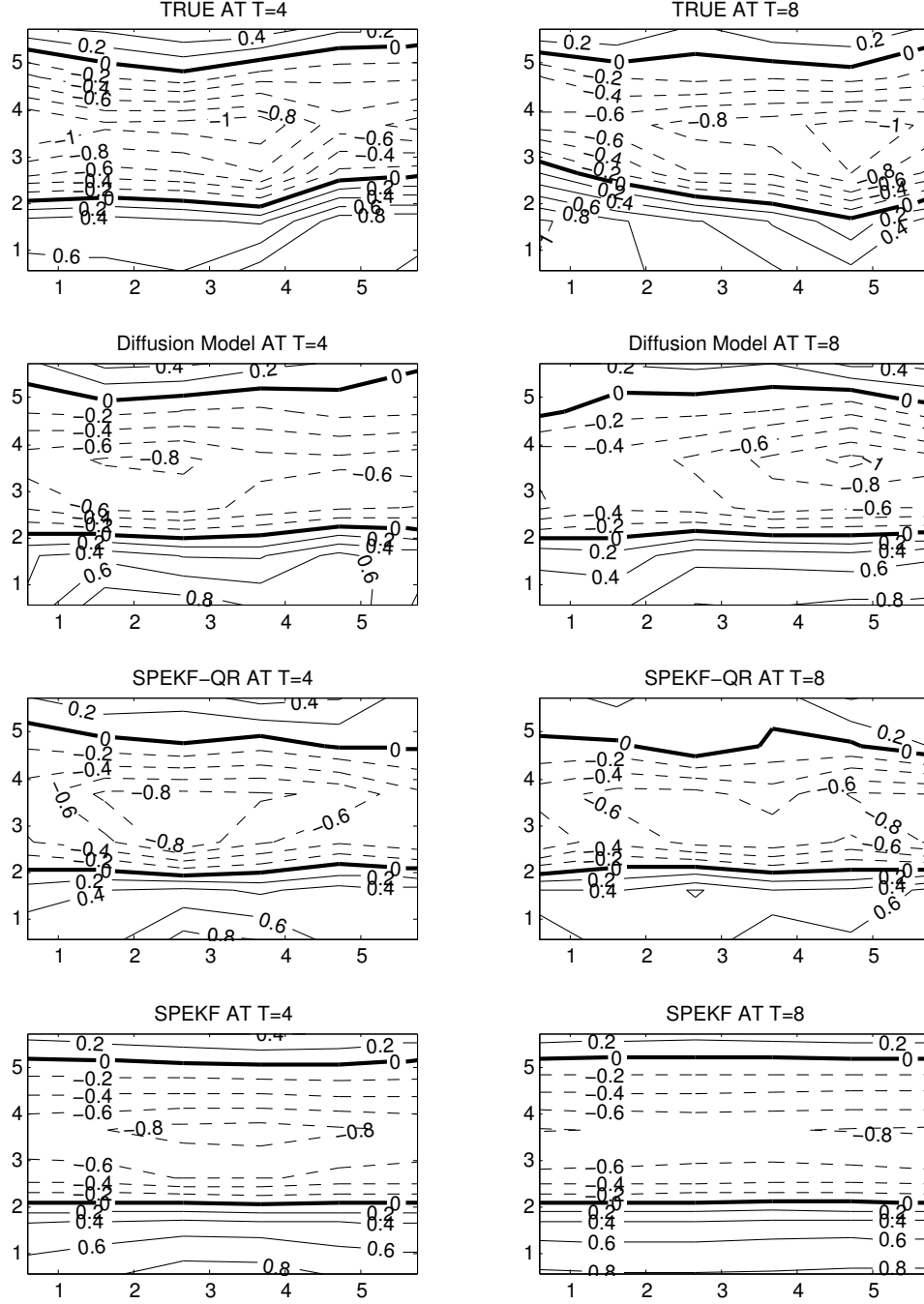


Figure 12: Spatial reconstruction of the lead forecast at times 4 and 8 (model unit time) of the QG models at the sparsely 36 regularly spaced locations for the atmospheric regime with  $F = 4$ .

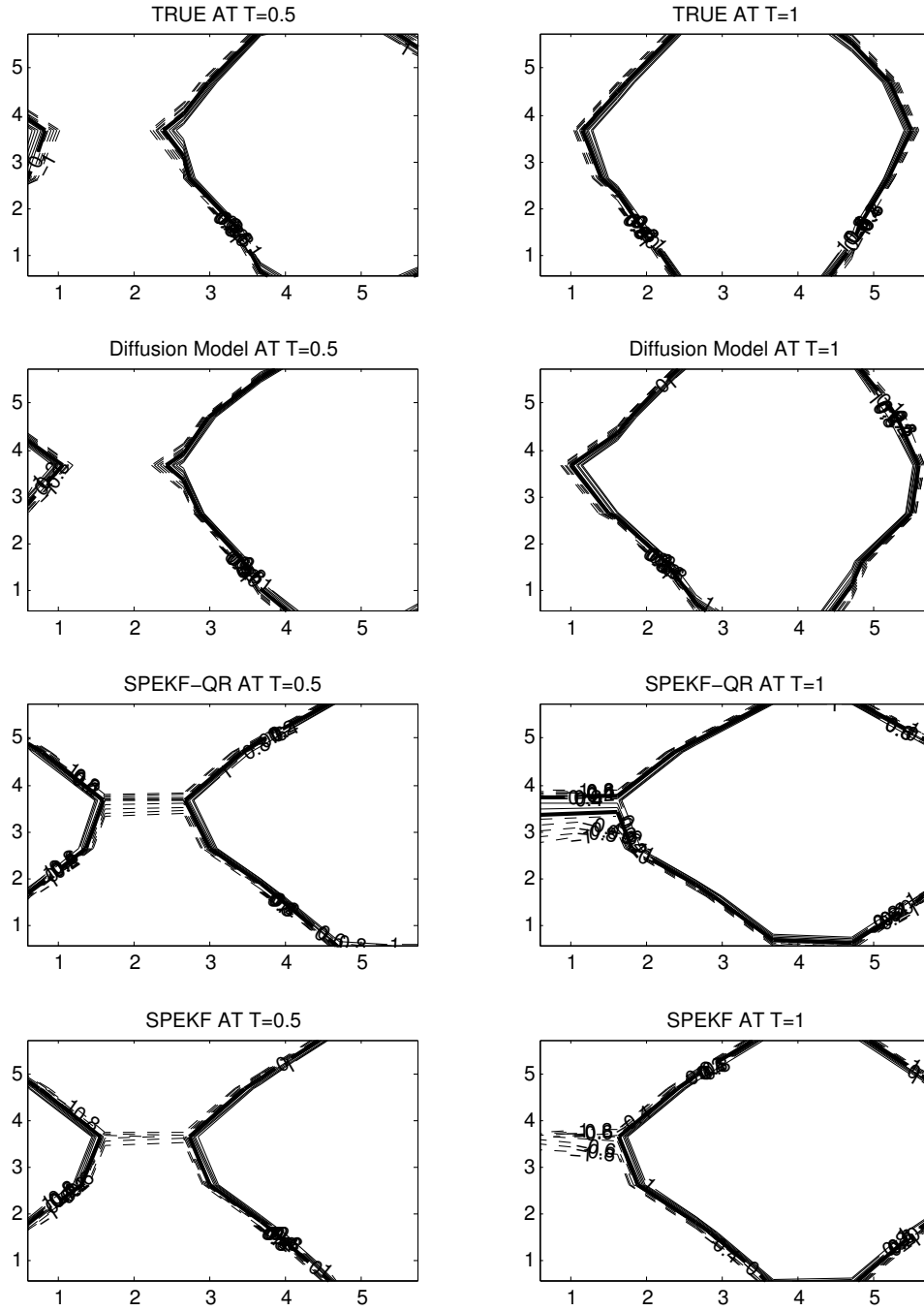


Figure 13: Spatial reconstruction of the lead forecast at times 0.5 and 1 (model unit time) of the QG models at the sparsely 36 regularly spaced locations for the ocean regime with  $F = 40$ .