JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003268

Key Points:

- High-resolution numerical simulation data are used to extract small-scale features of the 2D Euler equations
- An empirical orthogonal function (EOF)-based stochastic forcing is proposed, where the EOF time series serve to define data-driven stochastic processes for each EOF
- The data-driven processes are found to produce ensembles with reduced mean error and spread, compared to using Gaussian noise

Correspondence to:
S. R. Ephrati,
s.r.ephrati@utwente.nl

Citation:
Ephrati, S. R., Cifani, P., Luesink, E., & Geurts, B. J. (2023). Data-driven stochastic Lie transport modeling of the 2D Euler equations. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003268. https://doi.org/10.1029/2022MS003268

Received 1 JUL 2022
Accepted 29 DEC 2022

Data-Driven Stochastic Lie Transport Modeling of the 2D Euler Equations

Sagy R. Ephrati[1], Paolo Cifani[1,2], Erwin Luesink[1], and Bernard J. Geurts[1,3]

[1]Mathematics of Multiscale Modeling and Simulation, Faculty EEMCS, University of Twente, Enschede, The Netherlands, [2]Gran Sasso Science Institute, L'Aquila, Italy, [3]Multiscale Energy Physics, CCER, Faculty Applied Physics, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract In this paper, we propose and assess several stochastic parametrizations for data-driven modeling of the two-dimensional Euler equations using coarse-grid SPDEs. The framework of Stochastic Advection by Lie Transport (SALT) (Cotter et al., 2019, https://doi.org/10.1137/18m1167929) is employed to define a stochastic forcing that is decomposed in terms of a deterministic basis (empirical orthogonal functions, EOFs) multiplied by temporal traces, here regarded as stochastic processes. The EOFs are obtained from a fine-grid data set and are defined in conjunction with corresponding deterministic time series. We construct stochastic processes that mimic properties of the measured time series. In particular, the processes are defined such that the underlying probability density functions (pdfs) or the estimated correlation time of the time series are retained. These stochastic models are compared to stochastic forcing based on Gaussian noise, which does not use any information of the time series. We perform uncertainty quantification tests and compare stochastic ensembles in terms of mean and spread. Reduced uncertainty is observed for the developed models. On short timescales, such as those used for data assimilation (Cotter et al., 2020a, https://doi.org/10.1007/s10955-020-02524-0), the stochastic models show a reduced ensemble mean error and a reduced spread. Particularly, using estimated pdfs yields stochastic ensembles which rarely fail to capture the reference solution on small time scales, whereas introducing correlation into the stochastic models improves the quality of the coarse-grid predictions with respect to Gaussian noise.

Plain Language Summary Turbulent flows often contain small-scale fluctuations that behave in a seemingly random way. Predicting the behavior of such a flow is challenging, since simulating the flow in full detail is computationally expensive. To reduce the computational costs, one can initially ignore the small-scale fluctuations and subsequently try to include the effects of these scales by including an additional term into the equations that describe the flow. We propose and assess various models that represent the influence of the small-scales through a stochastic (random) forcing term. We compare three types of stochastic processes that use information from high-resolution data. It is found that using more information from the data leads to a reduced spread and ensemble mean error.

1. Introduction

A major challenge in geophysical and observational sciences is the representation and quantification of uncertainty in numerical predictions. Uncertainty stems from various sources, most relevantly from incomplete inclusion of all relevant physical mechanisms in the models and uncertainty in the initial and boundary conditions (T. N. Palmer, 2000). Important models for geophysical fluid dynamics, such as the two-dimensional Euler equations, quasi-geostrophic equations or rotating shallow water equations are derived from the three-dimensional Navier-Stokes equations. A sequence of simplifying assumptions is applied in order to reduce the complexity of the model to a more manageable level, while retaining main flow physics (Zeitlin, 2018). Stochastic extensions to these models have also been derived (Holm & Luesink, 2021). These approximate models are nevertheless rich in dynamics and contain a wide range of spatial and temporal scales. Numerically resolving the entire spectrum of scales is often not computationally feasible, meaning that either the complexity of the model should be reduced even further such that the resulting model is simple enough the be solvable numerically, or the complex model is represented on a coarse computational grid and unresolved scales are replaced by a sub-grid model. The latter option may be combined with stochastic forcing, which provides an effective way to represent unresolved scales in numerical simulations (Buizza et al., 1999; Frederiksen & Davies, 1997; Majda et al., 2001). The use of stochasticity as a means to represent the unresolved scales serves to restore some of the missing small-scale

dynamics and at the same time probes an ensemble of solutions and hence also investigates uncertainty. In this paper, we embrace these ideas and develop and assess stochastic data-driven models for the two-dimensional Euler equations on the unit square.

Data-driven stochastic models in dynamical systems have been studied actively in recent years. For weather and climate models, stochasticity was used as a tool to represent uncertainty in initial conditions and in the model, as shown by T. Palmer (2019). A commonly used example to illustrate the data-driven stochastic approach is the two-scale Lorenz '96 (L96) system, introduced by Lorenz (1996) and originally proposed as a simplified model of the atmosphere that incorporates interactions between slow and fast scales. Data of the fast scales, interpreted as unresolved sub-grid scales, may serve to construct a data-informed stochastic model. Examples are given in Arnold et al. (2013) where sub-grid features are modeled using different types of noise including additive, multiplicative and state-dependent noise. This study established that stochastic parametrizations could accurately account for modeling error, with a considerably improved forecasting skill when temporal correlation was included in the noise. The correlated noise was modeled as a one-step autoregressive model with parameters fitted from data. Alternative ideas such as stochastic parametrization based on Markov chains inferred from data are presented by Crommelin and Vanden-Eijnden (2008), where unresolved processes are represented as stochastic processes dependent on the state of the resolved variables and an assumed probability density. Using this approach, good agreement was found for the probability density functions (pdfs) and autocorrelation functions of resolved state variables.

Data-driven machine learning has also been adopted to represent small-scale dynamics for a large range of parameters (Gagne et al., 2020). It was found that several configurations of machine learning accurately reconstructed spatio-temporal correlations of the original system. These methods are not limited to simplified models such as the L96 system, but have also been successfully applied to more complete geophysical models. Examples include oceanic flows as considered in Bolton and Zanna (2019) and atmospheric processes as investigated in O'Gorman and Dwyer (2018). Both studies obtain a parametrization using machine learning based on off-line computed high-resolution model output. This machine learning approach could accurately predict the relation between resolved and unresolved turbulent processes, although a reliable generalization is principally not guaranteed. Here, we follow another data-driven 'offline/online' route and express the differences between a fully resolved model and a coarsened model in terms of a converging series of empirical orthogonal functions (EOFs) and introduce explicit forcing to update the coarsened model to high accuracy. This direct forcing strategy can also be extended to structure-preserving stochastic models as will be clarified below.

In a seminal work (Holm, 2015), stochastic partial differential equations are derived for fluid dynamics by means of a variational principle. As a result, the solution of the SPDE is compliant with the geometry of the underlying equations. This means that conservation laws are maintained under the inclusion of stochastic perturbations. This approach goes by the name of stochastic advection by Lie transport (SALT). In a similar approach, stochastic fluid models can be derived following the framework of location uncertainty (LU) (Mémin, 2014), in which the kinetic energy is conserved. In these approaches, spatial correlations of observational data can be used to model the unresolved scales in a numerical simulation. The spatial correlations can be decomposed into EOFs (Hannachi et al., 2007; Lumley, 1967). These are coupled to noise generated from stochastic processes in a separate modeling step. Together, these terms constitute a stochastic forcing term for the coarse PDE, which models unresolved scales. The conservation properties in the framework of SALT require a calculus in which the chain rule coincides with those of deterministic calculus. When the stochastic integration is of the Stratonovich type, it has been shown that integration with respect to semimartingales preserves the conservation properties (Street & Crisan, 2021). For processes with unbounded variance one should resort to pathwise approaches. Recently, conservation properties for SALT and LU have been established for geometric rough paths (Crisan et al., 2022).

The SALT approach finds meaningful applications within geophysical fluid dynamics, since these models are directly based on a variational point of view. To illustrate the SALT approach, (Cotter et al., 2019) apply it to the two-dimensional Euler equations. In this study a fine-grid simulation is performed from which the Lagrangian trajectories are extracted and compared to the corresponding trajectories acquired after filtering the velocity field. The difference between these trajectories is a measure of the unresolved scales to which the EOF decomposition is applied to form an optimal basis for this term. Subsequently, a coarse SPDE is constructed according to SALT where the amplitude of the EOF basis is modeled as a decorrelated stochastic Gaussian process. It is shown that an ensemble of stochastically forced flows captures the mean values of the true solution over considerable time

intervals. In a follow-up study (Cotter et al., 2020c), a particle filter was added to the SALT two-dimensional Euler equations and data assimilation was motivated this way. It was demonstrated that significant model reduction is possible, reducing the number of degrees of freedom by two orders of magnitude without losing reliability of the results. Similar studies on the quasi-geostrophic equations have been done (Cotter et al., 2020b), with a focus on data assimilation (Cotter et al., 2020a).

Stochastic forcing allows for the use of data-driven models outside of the data set from which the EOFs are obtained and the parametrization of the stochastic forcing ultimately remains a modeling choice. Global basis methods adopting for example, EOFs, Fourier modes or spherical harmonics can be motivated by the nonlocality of turbulence. Such approaches for stochastic forcing based on DNS data have been applied to, for example, barotropic flow on the sphere (Frederiksen & Kepert, 2006) and to three-dimensional atmospheric flows (Kitsios & Frederiksen, 2019). A review of parametrizations for atmospheric flows using stochastic models based on DNS data is provided in Frederiksen et al. (2017). In Resseguier et al. (2020) a data-driven parametrization was compared to a self-similar parametrization, using SALT in the quasi-geostrophic equations. It was found that both parametrizations accurately predict numerical errors and possess good uncertainty skills. The work by Agarwal et al. (2021) adopts EOFs and compares several dependent stochastic models and found that models that include the dynamics and time-delay effects perform well.

In this paper, we extend the work presented by Cotter et al. (2019) of stochastic forcing for the two-dimensional Euler equations. The extension presented in this work consists of the inclusion of additional information in the data-driven approach. This information is readily available from the EOF procedure and is used to define two additional types of stochastic processes. Providing a space-time array of measurements to the algorithm yields the EOFs, which are spatial profiles, and the amplitudes of the EOFs in order to reconstruct the input measurements. The amplitudes of each EOF are a time series and provide the data that are used in this paper to calibrate stochastic processes for each of the EOFs. In order to mimic the measurements, we generate signals that have the same probability distribution function as the measured time series or have similar temporal correlation. By retaining these statistical quantities in the modeled time series, the forcing stays true to characteristic features of the measurements.

The following numerical experiments and findings are reported in this paper. We perform a direct numerical simulation (DNS) of the two-dimensional Euler equations on the unit square, subject to impenetrable boundary conditions. We measure the difference between trajectories of particles advected by the fully-resolved velocity field and the corresponding filtered velocity field. The EOFs and time series that represent the amplitudes of the EOFs are obtained from this data. Stochastic ensembles are generated using three stochastic processes: Gaussian noise, noise based on the underlying pdf of the EOF time series, and noise with a temporal correlation similar to that of the EOF time series. The process of developing the time series into stochastic processes is explained in detail in a subsequent section of the paper. The results presented in this paper show that using the developed stochastic processes leads to a reduction of the ensemble mean error and ensemble spread, compared to using Gaussian noise. This is further explored by performing statistical tests for ensemble solutions. The latter is done for time scales on which data may be assimilated, where the numerical SPDE results may serve as input (Cotter et al., 2020c).

The paper is structured as follows. In Section 2.1 we introduce the deterministic and stochastic governing equations and describe the numerical experiment in detail. This is followed by a description of the data acquisition procedure in Section 2.2. The method used for generating random signals as a model for the measured data is described in Section 2.3. The results of the numerical experiments are presented in Section 3. In Section 3.1 a maximal prediction horizon is established and in Section 3.2 an adapted reference solution defined. These results aid the uncertainty quantification of ensemble predictions, presented in Section 3.3. Predictions on much shorter timescales are further assessed in Section 3.4, comparing additional ensemble statistics. In Section 3.5 we assess the forecast quality for different lengthscales in the flow by analyzing the results in spectral space. We conclude the paper in Section 4 and specify future challenges.

2. SPDE Formulation and Stochastic Models

This section presents the formulation of the stochastic Euler equations using the SALT approach (Section 2.1), the data acquisition procedure (Section 2.2) and the derivation of the stochastic models (Section 2.3).

2.1. Governing Equations and Flow Conditions

The two-dimensional Euler equations are central to this work. These equations are determined fully by the evolution of the vorticity dynamics (Zeitlin, 2018). The behavior of the vorticity ω in terms of the velocity $\mathbf{u}$ and streamfunction ψ is given by

$$\partial_t \omega + (\mathbf{u} \cdot \nabla)\omega = Q - r\omega, \tag{1}$$

$$\mathbf{u} = \nabla^{\perp}\psi, \tag{2}$$

$$\Delta\psi = \omega, \tag{3}$$

which are solved on the unit square, denoted by $\mathcal{D}$. The perpendicular gradient $\nabla^{\perp}$ is defined as $(-\partial_y, \partial_x)$. A forcing and a damping term are added to the equations in order to drive the flow to a nontrivial statistically steady state. In particular, $Q(x, y) = 0.1\sin(8\pi x)$ and $r = 0.01$, which enforce eight spatial gyres that are constant in time. An impenetrable boundary condition is applied via

$$\psi|_{\partial\mathcal{D}} = 0 \tag{4}$$

along the boundary $\partial\mathcal{D}$ of $\mathcal{D}$. For this system a characteristic time scale is the large eddy turnover time, here estimated to be 2.5 time units (Cotter et al., 2019).

The stochastic equations associated with the Euler equations follow from the principle of SALT for ideal fluid dynamics (Holm, 2015). In this approach, SPDEs are derived from a variational principle. In fact, a stochastically constrained functional is minimized to obtain an SPDE which retains the geometric properties equivalent to the corresponding PDE. The result is that quantities that are advected along an infinitesimal vector field $\mathbf{u}\mathrm{d}t$ in the deterministic setting are advected along an infinitesimal vector field $\bar{\mathbf{u}}\mathrm{d}t + \sum_i \boldsymbol{\xi}_i \circ \mathrm{d}B_t^i$ in the stochastic setting. In this paper, $\bar{\cdot}$ denotes a filtered field representative of scales that can be resolved accurately on a coarse numerical grid. As a rough rule of thumb, the resolved scales would comprise of structures for which $\Delta \gtrsim kh$ where h denotes the uniform grid spacing and k is a factor that quantifies the desired accuracy requirements. Typically, one may think of $k \gtrsim 4$ for second order accurate methods (Geurts & Fröhlich, 2002). The velocity fields $\boldsymbol{\xi}_i$ are defined as the eigenvectors of the velocity-velocity correlation tensor (Holm, 2015), B_t^i is a Wiener process. The symbol $\circ$ implies that the stochastic integral should be understood in the Stratonovich sense. This means that the integral is approximated by Riemann sum defined on the midpoints of the subintervals. For a good introduction to this material (Kloeden & Platen, 1992) and (Higham, 2001) can be consulted.

Since the velocity field $\mathbf{u}$ is divergence-free, each velocity field $\boldsymbol{\xi}_i$ is divergence-free (Cotter et al., 2019) and can be expressed by a potential function ζ_i via $\boldsymbol{\xi}_i = \nabla^{\perp}\zeta_i$. The advection velocity can then be written in terms of the potential as

$$\bar{\mathbf{u}}(t)\mathrm{d}t + \sum_i \boldsymbol{\xi}_i \circ \mathrm{d}B_t^i = \nabla^{\perp}\bar{\psi}(t)\mathrm{d}t + \sum_i \nabla^{\perp}\zeta_i \circ \mathrm{d}B_t^i. \tag{5}$$

Numerically, the velocity fields are projected to divergence-free fields to guarantee non-divergence. In this equation the filtered variables are used since the aim of the stochastic model is to represent the components of the fine-grid solution that are not resolvable on the coarse grid. The resulting SPDE then reads (Cotter et al., 2019)

$$\mathrm{d}\bar{\omega} + \nabla^{\perp}\left(\bar{\psi}\mathrm{d}t + \sum_i \zeta_i \circ \mathrm{d}B_t^i\right) \cdot \nabla\bar{\omega} = (Q - r\bar{\omega})\mathrm{d}t, \tag{6}$$

$$\Delta\bar{\psi} = \bar{\omega}. \tag{7}$$

2.2. Data Acquisition

The numerical method for the solution of Equations 6 and 7 and the flow parameters are the same as those used in earlier studies (Cotter et al., 2019, 2020c). A full description of the numerical implementation can be found in the former references. Here, for completeness, we illustrate the key aspects. A finite element method is employed

to solve the system of Equations 6 and 7. The Poisson equation for the streamfunction is discretized using a continuous Galerkin scheme. The vorticity Equation 1, including the stochastic terms, is discretized using a discontinuous Galerkin scheme. The space of discontinuous test functions guarantees numerical conservation of energy in the absence of source terms (Bernsen et al., 2006).

Numerical time integration is performed by applying a third-order strong stability preserving Runge-Kutta (SSPRK3) method (Shu & Osher, 1988). Writing the stochastic advection Equation 6 in the general Stratonovich SPDE form

$$\mathrm{d}\bar{\omega} = L(\bar{\omega})\mathrm{d}t + \sum_{i=1}^{m} G^i(\bar{\omega}) \circ \mathrm{d}B_t^i, \tag{8}$$

where

$$\begin{aligned} L(\bar{\omega}) &= -\nabla^{\perp}\bar{\psi} \cdot \nabla\bar{\omega} + (Q - r\bar{\omega}), \\ G^i(\bar{\omega}) &= -\nabla^{\perp}\zeta_i \cdot \nabla\bar{\omega}, \end{aligned} \tag{9}$$

the SPDE Equation 8 is integrated in time via

$$\begin{aligned} \bar{\omega}_{(1)} &= \bar{\omega}_n + \Delta t L(\bar{\omega}_n) + \sum_{i=1}^{m} G^i(\bar{\omega}_n)\Delta B_n^i, \\ \bar{\omega}_{(2)} &= \frac{3}{4}\bar{\omega}_n + \frac{1}{4}\left[\bar{\omega}_{(1)} + \Delta t L(\bar{\omega}_{(1)}) + \sum_{i=1}^{m} G^i(\bar{\omega}_{(1)})\Delta B_n^i\right], \\ \bar{\omega}_{n+1} &= \frac{1}{3}\bar{\omega}_n + \frac{2}{3}\left[\bar{\omega}_{(2)} + \Delta t L(\bar{\omega}_{(2)}) + \sum_{i=1}^{m} G^i(\bar{\omega}_{(2)})\Delta B_n^i\right]. \end{aligned} \tag{10}$$

The subscript n denotes the nth numerical time step. The stages of the Runge-Kutta algorithm are denoted by the subscripts (1) and (2). The time step size is given by Δt and is chosen such that the CFL number does not exceed 1/3. Here ΔB_n^i denote random samples drawn from an assumed probability distribution with variance Δt. For deterministic systems, the functions G^i equal zero.

The term $\nabla^{\perp}\left(\sum_i \zeta_i \circ \mathrm{d}B_t^i\right)$ in Equation 6 is unknown in the coarsened description and needs to be modeled. The latter is approximated as follows:

$$\mathbf{f}(x,t)\sqrt{\Delta t} = (\mathbf{u} - \bar{\mathbf{u}})\Delta t \approx \sum_i \boldsymbol{\xi}_i(x)\Delta B_n^i. \tag{11}$$

The forcing $\mathbf{f}$ in Equation 11 is computed as the difference of the Lagrangian trajectories originating by the velocity fields $\mathbf{u}$ and $\bar{\mathbf{u}}$ projected onto the coarse grid. As such, the forcing is a large-scale correction to $\bar{\mathbf{u}}$ which measures the part of the velocity fluctuation resolved by the coarse grid. The right hand side incorporates these fluctuations as a stochastic forcing. The measurements are approximated by coarse-grid resolved fields. This approximation may become inaccurate in the case of severe coarsening, in which case a large number of terms must be introduced in the approximation of $\mathbf{f}$ to properly capture the effects of the small scales.

The process of measuring $\mathbf{f}(x, t)$ is as follows. A grid with 512^2 computational cells is adopted for the DNS and all subsequent stochastic results are obtained on a coarse grid of 64^2 computational cells. The filtered fields are derived from the fine-grid DNS results and are obtained by applying a Helmholtz operator to the streamfunction. Given a streamfunction ψ, the filtered streamfunction $\bar{\psi}$ is obtained by solving

$$\left(I - c\nabla^2\right)\bar{\psi} = \psi, \tag{12}$$

where $c = 1/64^2$ to filter out length scales smaller than the coarse grid size. The numerical resolutions and the filter width coincide with those adopted in Cotter et al. (2019). The filtered vorticity $\bar{\omega}$ and filtered velocity $\bar{\mathbf{u}}$ are recovered from applying the relations Equations 2 and 3 to $\bar{\psi}$. Over the entire simulation interval, this filter was found to remove approximately 12% of the kinetic energy.

The initial vorticity is prescribed, as

$$\omega_0 = \sin(8\pi x)\sin(8\pi y) + 0.4\cos(6\pi x)\cos(6\pi y) + 0.3\cos(10\pi x)\cos(4\pi y) + 0.02\sin(2\pi y) + 0.02\sin(2\pi x), \tag{13}$$

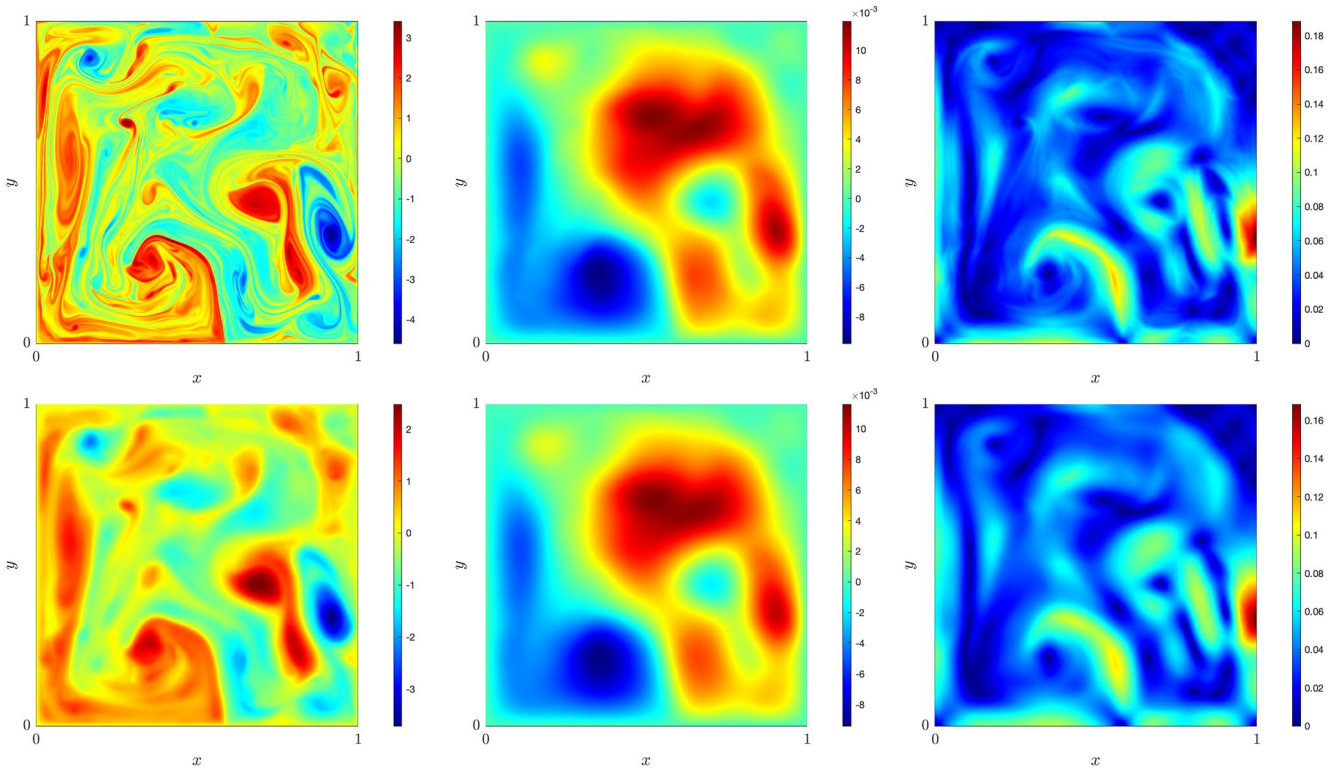

Figure 1. Fine-grid fields of the vorticity (left), streamfunction (middle) and velocity magnitude (right) after the spin-up interval. The top row shows the unfiltered fields, the bottom row shows the corresponding fields after applying the filter Equation 12.

from which the system will be spun-up during an interval of 100 time units so that a statistical equilibrium is reached. The time at which this is reached is denoted by $t = 0$. The initial fields and corresponding filtered fields at the end of the spin-up interval are found in Figure 1.

The method to estimate the eddy velocity is the same as presented by Cotter et al. (2019) and is based on measuring the trajectories of fluid parcels. A decomposition of the true trajectories into a drift and a stochastic perturbation is assumed, which is equivalent to the SALT equations (Cotter et al., 2017). Here, the drift is computed as the filtered velocity field. Thus, a space-time sequence of measurements for determining $\mathbf{f}$ from Equation 11 is obtained by computing the difference of Lagrangian trajectories of particles advected by the velocity field $\mathbf{u}$ and those advected by the filtered velocity field $\bar{\mathbf{u}}$. The difference is measured over a single coarse-grid time step. The particles are released on the coarse grid points and thus a difference in traveled distance can be related to each grid point. A correction field is subsequently obtained by dividing the difference in trajectories by the square root of the coarse-grid time step, in a manner analogous to particle image velocimetry measurement techniques in experimental fluid flow analysis (Adrian & Westerweel, 2011). The measurements are done at each coarse-grid time step, from $t = 0$ (the point at which the simulation is spun-up) until $t = 365$. By doing so, an array $\mathbf{f}(x, t)$ of fields is constructed. This space-time array of measurements is decomposed into empirical orthogonal functions (EOFs or EOF modes) (Hannachi et al., 2007; Lumley, 1967). Here, a total of 4,096 EOFs are available (64^2 degrees of freedom), of which the first 225 are used. These EOFs account for 90% of the energy of the measurements. The fact that only a small portion of EOFs is needed for an accurate reconstruction of the forcing is attributed to the Helmholtz filter being a graded filter. The latter, by construction, filters out not only small-scale dynamics but also part of the large-scale field, which can be captured by the first 5.5% of EOF modes. A potential concern in this general framework is that errors might become too large in case of further coarsening.

Application of the EOF algorithm to a flow that has a definite statistically steady state yields

$$\mathbf{f}(x,t) = \boldsymbol{\xi}_0(x) + \sum_{i=1}^{N} a_i(t)\boldsymbol{\xi}_i(x), \tag{14}$$

where $\boldsymbol{\xi}_0(x)$ is the time-mean of the measurements, $\boldsymbol{\xi}_i(x)$ are normalized spatial EOF modes, also referred to as "topos", and $a_i(t)$ are the corresponding coefficients with reference to the measurements, also referred to as "chronos." These are recorded as time series and may be written instead as $\sqrt{\lambda_i}\bar{a}_i(t)$. Here, $\sqrt{\lambda_i}$ is the standard deviation of $a_i(t)$ and carries the same dimension as the measurements. The time series $\bar{a}_i(t)$ has unit variance and is dimensionless. The EOF modes are orthonormal with respect to the inner product, thus $\left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j\right) = \delta_{ij}$, where

$$(\boldsymbol{f}, \boldsymbol{g}) = \int_{\Omega} \boldsymbol{f}(x)\boldsymbol{g}(x)\, \mathrm{d}x \tag{15}$$

with Ω the flow domain. Due to the orthonormality, the coefficients $a_i(t)$ are readily obtained by projecting the measured velocity fields onto the basis of EOFs by

$$a_i(t) = \left(\mathbf{f}(x,t) - \boldsymbol{\xi}_0(x), \boldsymbol{\xi}_i(x)\right). \tag{16}$$

In order to have a self-contained model which allows to obtain predictions, for example, beyond the time span of the data set, or as surrogate statistical sample of the flow, we choose to model the time traces $a_i(t)$ as independent stochastic processes. This will be described in the next section, where also the possible connection to the available data will be elaborated.

2.3. Generating Random Signals

We will now introduce the models for the time traces Equation 16 and subsequently describe how random signals are generated using these models. By comparing Equation 14 with Equation 11 it is clear that modeling $B_t^i(t)$ amounts to modeling $a_i(t)$. The following models are employed:

1. The stochastic process B_t^i in Equation 11 is modeled by Gaussian noise. For its discrete increments ΔB^i in Equation 10 we use $\Delta B^i = \sqrt{\lambda_i}\sqrt{\Delta t} r_i$, where $r_i \sim \mathcal{N}(0,1)$ (Higham, 2001).
2. The pdf of $\bar{a}_i(t)$ in is estimated from the measured signals Equation 16 and is subsequently used to draw uncorrelated samples. Thus, the increments ΔB^i in Equation 10 are computed as $\Delta B^i = \sqrt{\lambda_i}\sqrt{\Delta t} r_i$, where r_i is randomly drawn from the estimated pdf.
3. The time series $a_i(t)$ in Equation 14 is approximated by an Ornstein-Uhlenbeck (OU) process, using the correlation time obtained from the measurements Equation 16. The constructed OU process is then used to compute ΔB^i in Equation 10.

The probability distributions of model 2 are estimated by fitting a histogram to the values of the corresponding time series, yielding a separate distribution for each EOF. The histograms are fully determined by the smallest and largest measurements and the number of measurements. The number of bins is chosen as the smallest integer larger than $\sqrt[3]{2N_M}$, where N_M denotes the number of measurements, that is, the length of the time series. This choice minimizes the asymptotic mean squared error of the histogram as an estimator of the underlying pdf (Wilks, 2011). Moreover, the measurements are finite due to the spatial continuity of the numerical solution and the finite time step size, resulting in histograms with compact support. This guarantees bounded quadratic variation and finite moments, at the discrete level. Uncorrelated samples from these distributions are drawn using inverse transform sampling. In the latter a random number x is drawn from a uniform distribution between 0 and 1, which can intuitively be thought of as a probability of an event happening, and subsequently the largest value X is found such that $P(X \leq x)$ holds for the estimated distribution (Devroye, 2006). It is expected that the results obtained from model 2 will converge to those obtained from model 1 when the time step size is decreased, due to the central limit theorem.

In model 3, the noise generated using the OU process mimics the temporal correlation of the measured time series. Denoting by B_t^i the approximation of the time series $a_i(t)$, the OU process is defined as (Pope, 2001)

$$\mathrm{d}B_t^i = -B_t^i \frac{\mathrm{d}t}{T_i} + \left(\frac{2\mathrm{d}t}{T_i}\right)^{1/2} \sqrt{\lambda_i}\sqrt{\mathrm{d}t} r_i, \tag{17}$$

where $r_i \sim \mathcal{N}(0,1)$. We set T_i to be the correlation time of the measured time series. These variables are determined for each EOF separately. Here, the correlation time is defined as the smallest time at which the autocorrelation function of the time series is smaller than the computed 95% confidence bound.

A consistent choice for a fourth model is one that incorporates the measured temporal correlation, whilst retaining the estimated probability distribution of measurements. However, for this approach no tractable algorithm to generate the stochastic processes was found.

The conservation properties of SALT hold for the proposed stochastic processes, since these are semimartingales. Conservation of advected quantities then follows from the results of Street and Crisan (2021). Convergence of model 2 for decreasing time step sizes is guaranteed because the histograms have finite moments. Convergence of model 3 is established in Kloeden and Platen (1992) since the processes are semimartingales.

In the next section, we assess the proposed stochastic models by comparing simulations on the SPDE models to findings from deterministic reference solutions.

3. Assessment of Forecast Ensembles

In this section, we provide results of forecast ensembles using the aforementioned methods to generate stochastic signals that serve to force the coarsened dynamics. We first identify a maximal prediction horizon for assessing the forecast ensembles. An adapted reference solution is defined based on the measurements, incorporating on the coarse numerical grid the measured effects of small-scale motions. Subsequently, we show results of forecast ensembles. Statistics are computed and compared to the filtered DNS and the adapted reference solution to quantitatively compare the different stochastic forcing methods. Finally, the results are compared in terms of EOF coefficients, to distinguish between the forecast quality for different lengthscales present in the flow.

3.1. Establishing a Maximal Prediction Horizon

In order to define the maximal prediction horizon until which stochastically forced coarse numerical solutions can reasonably be compared to the DNS results, we set up the following numerical experiment. Starting from an initial condition on the fine grid, we generate a set of perturbed initial conditions of which we then follow the evolution over time. The perturbations are applied in Fourier space by shifting the phase of the Fourier coefficients, while keeping the amplitudes the same. The phase shift is applied only to modes of wave lengths smaller than the smallest scale resolved by the corresponding coarse grid, leaving the resolved modes unaltered. Specifically, a value l is chosen and all Fourier modes with wave numbers $|k| = \left(k_x^2 + k_y^2\right)^{1/2} \in [l, l+1)$ are affected by the additional phase shift. Here k_x and k_y denote the wave numbers in the x- and y- direction, respectively, and l is chosen as 64, 128, and 256. The phase shift is set to π to satisfy the boundary conditions.

As time evolves, the initial perturbation increasingly affects the resolved scales, up to the point where the instantaneous resolved fields will be entirely different from each other. We define this point of no longer truthfully following the unperturbed solution as the maximal prediction horizon $T_{\max}$, after which no model can be expected to consistently give accurate point-wise predictions owing to the sensitivity of the evolving solution to the initial conditions. The value of $T_{\max}$ is expected to depend on the choice of perturbed modes and choice of simulation parameters. However, in this numerical experiment it serves to provide an estimate of the maximal predication horizon.

The observed behavior following the small-scale phase-shift perturbations is illustrated in Figure 2, together with the results obtained from the unperturbed solution. The evolution of the vorticity for the various initial conditions has been measured on four illustrative points in the domain, at (0.25, 0.25), (0.25, 0.75), (0.75, 0.25) and (0.75, 0.75), of which two points are shown in the figure. It can be seen that the evolution of the vorticity values at the measured points in the domain is initially indistinguishable. At $t = 10$ slight differences are visible and at $t = 20$ the measured values are markedly different. The latter result is especially clear at the point (0.25, 0.25), in the left figure. Thus, we conclude that subsequent stochastic realizations can not be reasonably assessed after $t = 20$, which we set as the value for the maximal prediction time $T_{\max}$.

3.2. Defining the Reference Solution

In order to compare the different stochastic models one has to define a reference solution. The choice of the latter is not unique. In this work we define two reference solutions that are employed to measure performance of a given forcing model. The first one is the filtered fine-grid solution, employing the filter Equation 12, and is indicative

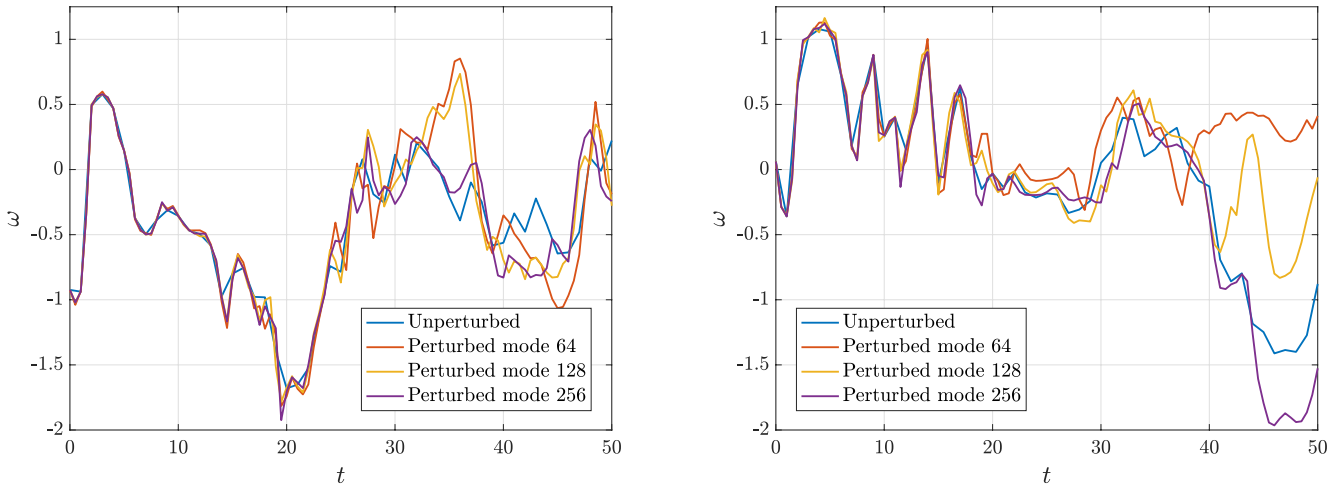

Figure 2. Development of the vorticity in two points of the domain, obtained by direct numerical simulation of perturbed fine-grid initial conditions. On the left, the vorticity is shown at (0.25, 0.25) and on the right at (0.75, 0.75). The results for the unperturbed and three perturbed initial conditions are shown. The perturbations are defined by phase-shifting small-scale modes of the streamfunction field. In the results shown here, the Fourier modes with wave numbers $|k| = \left(k_x^2 + k_y^2\right)^{1/2} \in [l, l+1)$, $l = 64, 128, 256$ are phase-shifted by π.

of flow scales that can be resolved on the coarse grid. Next to the filtered fine-grid solution, we define a reference solution as the numerical solution of Equations 6 and 7 where the reconstructed signal Equation 14, Equation 16 is used in Equation 11 instead of the stochastic forcing. This provides a prescribed deterministic forcing for the coarse numerical simulation. We call this the adapted reference solution. We note that the structure of the closure term Equation 11 does not account for discretization error and is itself not an exact closure since the noise is introduced only in the advection velocity. The inclusion of discretization error is what sets the filtered DNS and the adapted reference solution apart. Therefore, by comparing the stochastic ensembles against the adapted reference solution, one is able to distinguish between modeling error from the proposed stochastic models and the discretization error.

The adapted reference solution at $t = 0, 10, 20, 30$ is shown in the top row of Figure 3. At the same points in time, a single realization of each of the stochastically forced solutions is shown. The second row shows a realization using Gaussian noise, the third row using estimated pdfs and the bottom row using OU processes. The various realizations show no qualitative difference, suggesting that a more detailed, quantitative comparison of the methods is required. This is provided in the following subsections.

3.3. Uncertainty Quantification of Ensemble Predictions

The evolution of the vorticity and streamfunction is used for uncertainty quantification. First, the ensemble predictions are compared globally to the reference solution. In this subsection, the ensembles are compared quantitatively only to the adapted reference solution so that accumulation of discretization error in the coarse numerical solutions is not included in the comparison. Subsequently, similar to (Cotter et al., 2019) four points in the domain are picked for pointwise uncertainty quantification. For each point one ensemble standard deviation around the ensemble mean solution is shown and compared to the reference solution at the same point. In these tests, the ensemble is initialized from a single initial condition in order to isolate the effects of the stochastic processes on the uncertainty of the numerical solution. The initial condition is obtained by injecting the DNS vorticity field onto the coarse grid. Each SPDE is simulated up to $T_{\max} = 20$, and every ensemble is composed of 200 realizations of the SPDE. Our interest here lies in comparing the errors and spreads for the different types of stochastic processes used in the forcing Equation 11. Different error measures will be monitored as outlined next.

For global comparison to the reference solution, we define the pattern correlation

$$\frac{(\omega, \omega_{\text{ref}})}{\sqrt{(\omega, \omega)(\omega_{\text{ref}}, \omega_{\text{ref}})}} = \frac{\int \omega \omega_{\text{ref}} \, \mathrm{d}x}{\sqrt{\int \omega\omega \, \mathrm{d}x \times \int \omega_{\text{ref}} \omega_{\text{ref}} \, \mathrm{d}x}}, \tag{18}$$

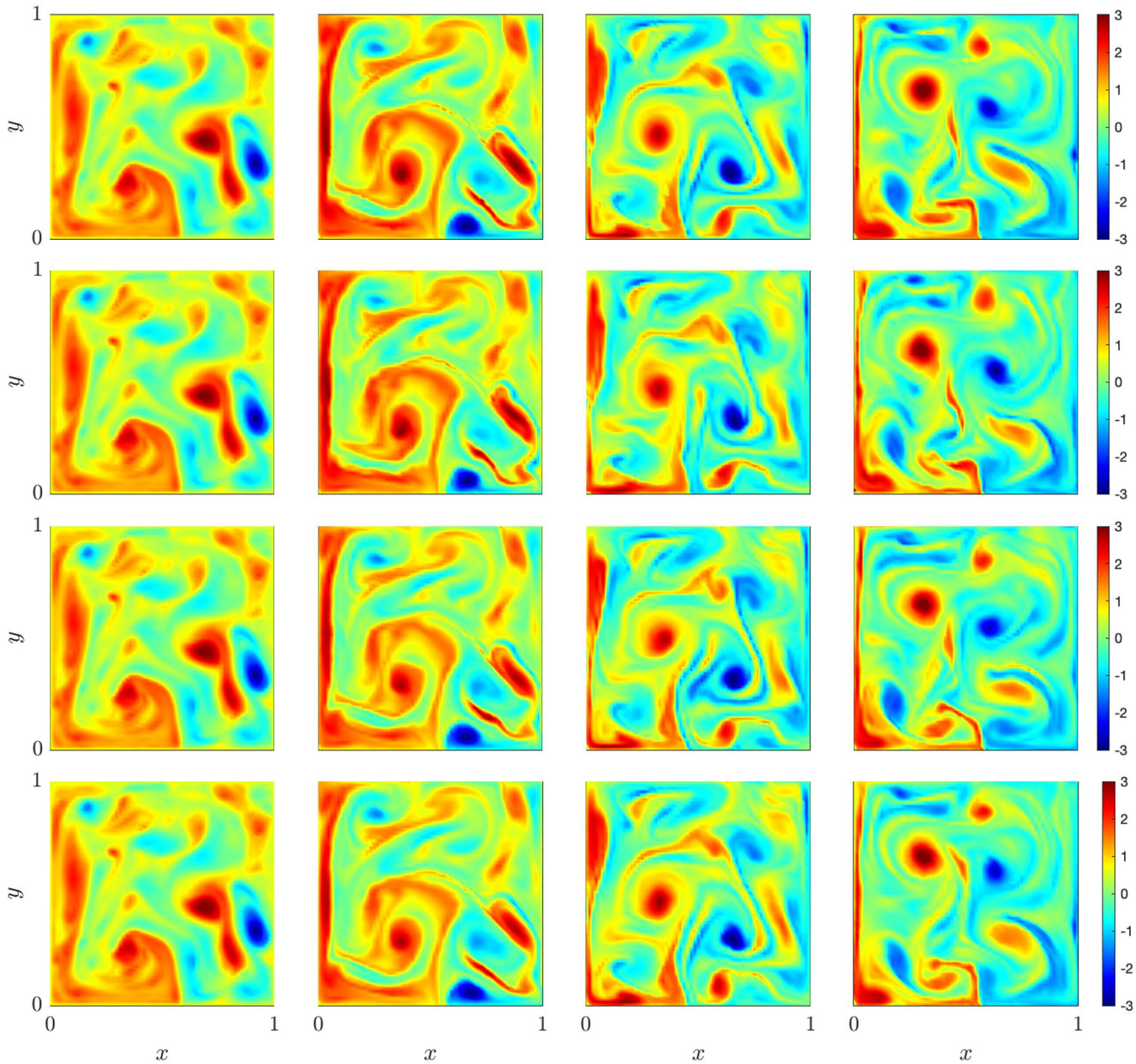

Figure 3. Coarse-grid fields of the vorticity at various points in time. The top row shows, from left to right, the adapted reference solution at $t = 0$, $t = 10$, $t = 20$ and $t = 30$. The other rows show realizations of stochastically forced numerical solutions at the aforementioned times. The second row uses Gaussian noise, the third row uses random samples from estimated distributions and the bottom row uses Ornstein-Uhlenbeck processes.

which can be considered a global measure of likeness between the vorticity ω obtained from the stochastically forced numerical solution and the vorticity ω_{ref} obtained from the reference solution. The same quantity is computed for the streamfunction. The pointwise comparisons are acquired by measuring the instantaneous vorticity and streamfunction at several grid points.

The stochastic ensembles are assessed using the ensemble mean, ensemble standard deviation and ensemble mean error. Here, we denote an ensemble of N stochastic realizations by $\{X_{i,j}\}$, where $i = 1, \ldots, N$ denotes the realization and $j = 0, \ldots, T$ denotes the time index. Then, the ensemble mean at time instance j is defined as

$$\langle X_j \rangle = \frac{1}{N} \sum_{i=1}^{N} X_{i,j}, \tag{19}$$

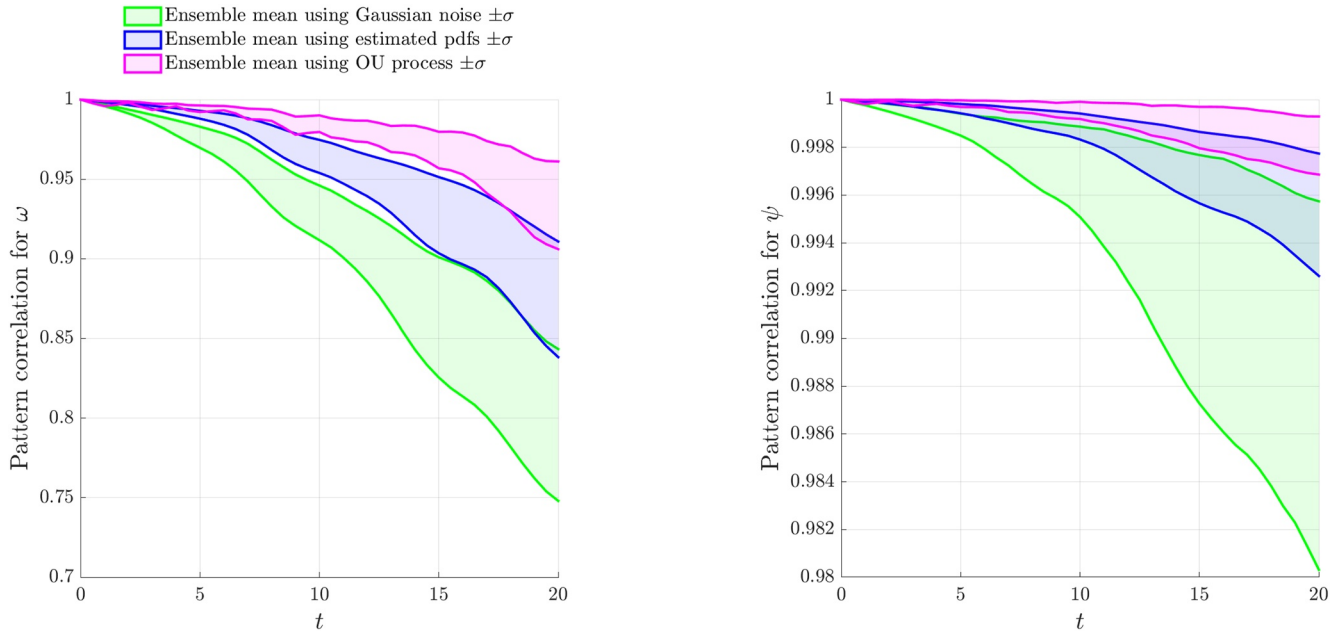

Figure 4. Pattern correlation Equation 18 between the forecast ensembles and the adapted reference solution for the vorticity (left) and the streamfunction (right). Each band is defined as one ensemble standard deviation around the ensemble mean. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses Ornstein-Uhlenbeck processes. The results for each method are generated for an ensemble of 200 realizations.

and the standard deviation, here referred to as spread, is defined as

$$\text{Spread}(X_{i,j}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{i,j} - \langle X_j \rangle)^2}. \tag{20}$$

A small spread indicates a sharp ensemble forecast and a large spread suggests an increased uncertainty in the forecast. The reference solution Y_j, $j = 0, \ldots, T$ is computed at the same time instances as $\{X_{i,j}\}$. The ensemble mean error of $\{X_{i,j}\}$ is then defined as

$$\text{ME}(X_{i,j}, Y_j) = |\langle X_j \rangle - Y_j|. \tag{21}$$

A small ensemble mean error indicates that the ensemble closely follows the reference solution, whereas a large value implies that the ensemble and the reference solution have deviated considerably from each other.

The correlation measure Equation 18 is shown in Figure 4 for the vorticity and the streamfunction. Using estimated pdfs or OU processes show favorable results when compared to using Gaussian noise, for both quantities. A clear difference between the methods can be observed for the vorticity on the time scale of $T_{\max}$. At this point, using estimated pdfs or OU processes yields a smaller spread than using Gaussian noise, and the results of the latter show a smaller correlation with the adapted reference solution. A significant increase in the correlation can also be observed for the streamfunction. The results using estimated pdfs or OU processes, as opposed to using Gaussian noise, exhibit both a larger likeness with the reference solution as well as a smaller spread. Compared to the ensemble obtained using Gaussian noise, at $t = 20$ the ensemble standard deviation of the pattern correlation of the vorticity was found to be reduced by 24% and 42% when using estimated pdfs and OU processes, respectively. For the streamfunction, these values were correspondingly observed to be 67% and 84%. Moreover, the results for the estimated pdfs and the OU processes are nearly indistinguishable before $t = 5$.

The evolution of the vorticity in four points of the domain is shown in Figure 5. The locations considered are (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), and (0.75, 0.75). In each of these plots, the colored bands present are the ensemble standard deviations around the corresponding ensemble mean. In all measured points, forcing based on Gaussian noise produces the largest spread. It is clearly visible that using the OU process yields the smallest ensemble spread and using the estimated pdfs only slightly increases the spread compared to using the OU process.

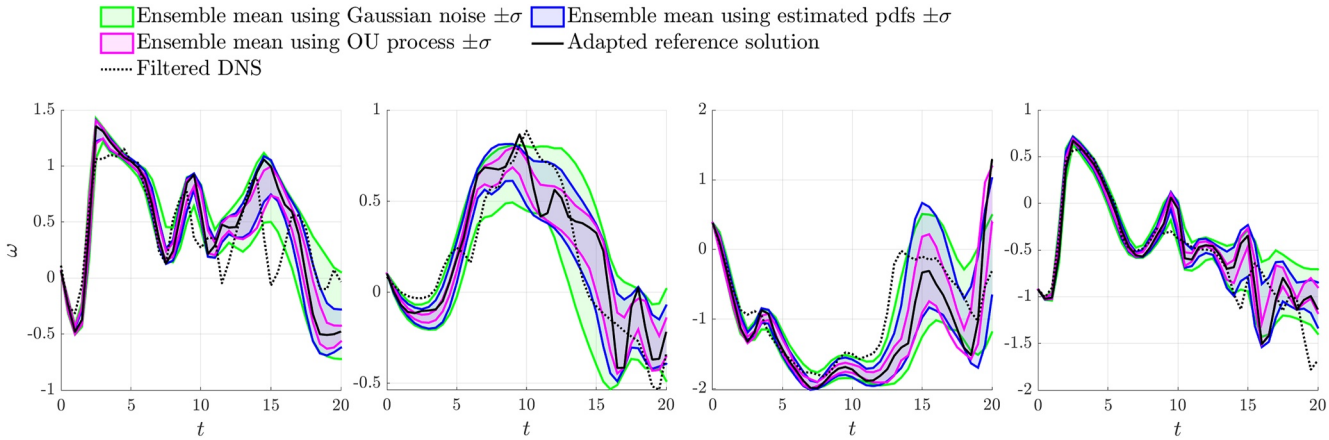

Figure 5. Vorticity measured on four points in the domain. From left to right, (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75). The solid and dotted black lines show the development of adapted reference solution and filtered direct numerical simulation, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses Ornstein-Uhlenbeck processes. The results for each method are generated from an ensemble of 200 realizations.

The ensemble mean error and the ensemble standard deviation are shown in Figure 6, where the ensemble mean error Equation 21 is taken with respect to the adapted reference solution. It becomes evident that the mean error develops similarly for each ensemble. The mean errors for ensembles using the estimated pdfs and the OU process are nearly indistinguishable until $t = 10$, after which some smaller differences can be observed. In contrast, using Gaussian noise results in a much larger spread.

Figure 7 shows the development of the streamfunction in the aforementioned points of the domain. The streamfunction is a smoother function than the vorticity, which is reflected in the smooth evolution of the former. In this figure it can also be observed that all ensembles accurately capture the adapted reference solution, with the OU model performing slightly better. The plots in Figure 8 show the ensemble mean error and the ensemble standard deviation for the same points in the domain. Analogously to the vorticity, we find that the ensembles using the OU process and the estimated pdfs result in a smaller spread than the ensemble using Gaussian noise. Furthermore, it is observed that the ensemble mean error does not exceed the ensemble standard deviation before $t = 10$ and only does so occasionally after this point in time, indicating the reference solution is captured well by the ensembles.

In this subsection we have shown that the three considered stochastic processes accurately follow the adapted reference solution for multiple characteristic time units. Compared to Gaussian noise, using estimated pdfs or OU processes to define the stochastic forcing yielded a smaller spread of the ensemble forecast. Using a global measure, it is found that the latter two types of forcing yield ensembles that better resemble the adapted reference solution. In the next subsection, we perform additional statistical tests to assess short-time predictions.

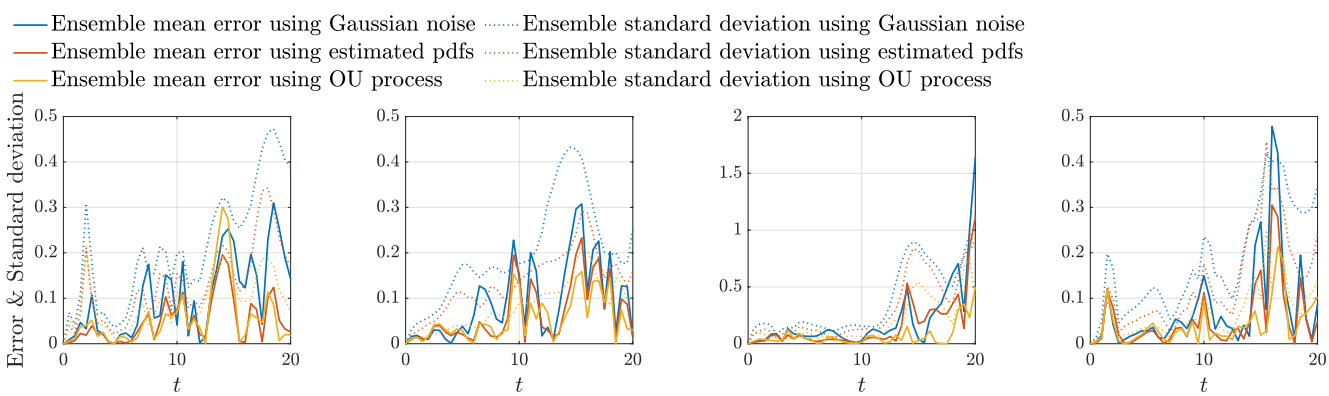

Figure 6. Ensemble mean error with respect to the adapted reference solution and ensemble standard deviation for the vorticity on four points in the domain. From left to right, (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75). The ensemble mean error is depicted by the solid lines, the ensemble standard deviation by the dotted lines.

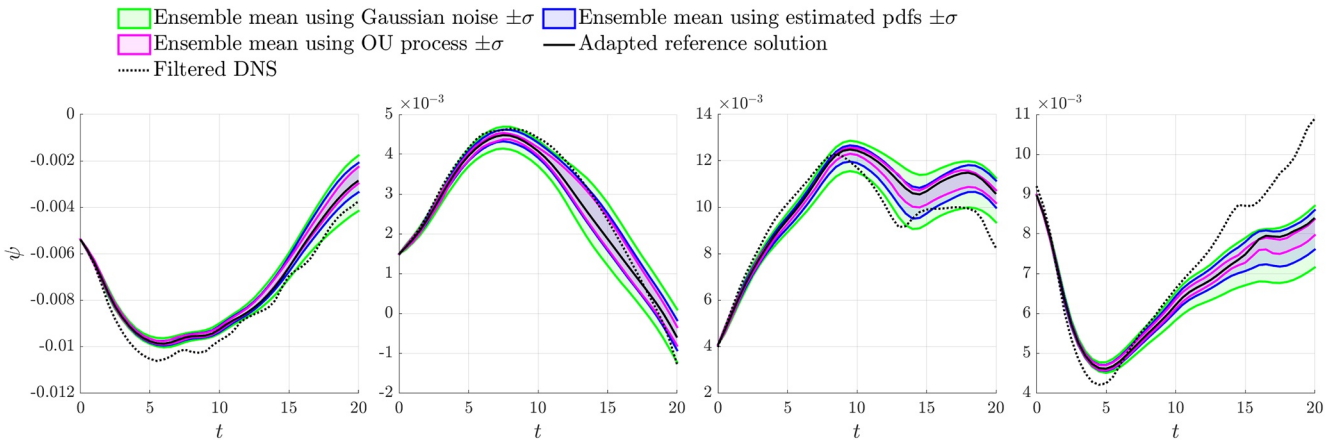

Figure 7. Streamfunction measured on four points in the domain. From left to right, (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75). The solid and dotted black lines show the development of the adapted reference solution and filtered direct numerical simulation, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses Ornstein-Uhlenbeck processes. The results for each method are generated from an ensemble of 200 realizations.

3.4. Statistical Tests for Ensemble Forecasts

Additional ensemble statistics are collected in order to further assess the numerical results of the SPDEs. In particular, forecast ensembles are generated for short lead times.

Two sets of initial conditions are generated to assess the stochastic models by sampling from two reference solutions: the filtered DNS and the adapted reference solution as presented in Section 3.1. The filtered DNS does not contain discretization and modeling error, whereas the adapted reference solution does. Therefore the use of both reference methods provides insight into the effects of these errors on the statistical quantities. Two distinct sets of initial conditions are acquired by sampling the reference solutions at $t = 0, 5, 10, \ldots, 350$, measured after the spin-up time. An ensemble forecast consisting of one hundred stochastic realizations is computed for each initial condition. Every stochastic realization is run for two time units and stored every 0.04 time units in order to study the results for short lead times. This time interval is similar to time intervals at which data may be assimilated (Cotter et al., 2020c). Subsequently, the statistics are computed by comparing the ensembles to the corresponding reference solution. The statistics are provided below for both sets of initial conditions separately.

As a first quantity we compute the root mean square error (RMSE). Recall that $\{X_{i,j}\}$, $i = 1, \ldots, N$, $j = 0, \ldots, T$ denotes an ensemble of N realizations measured at $T + 1$ times. The RMSE between the ensemble mean of the SPDE and the reference solution is computed from

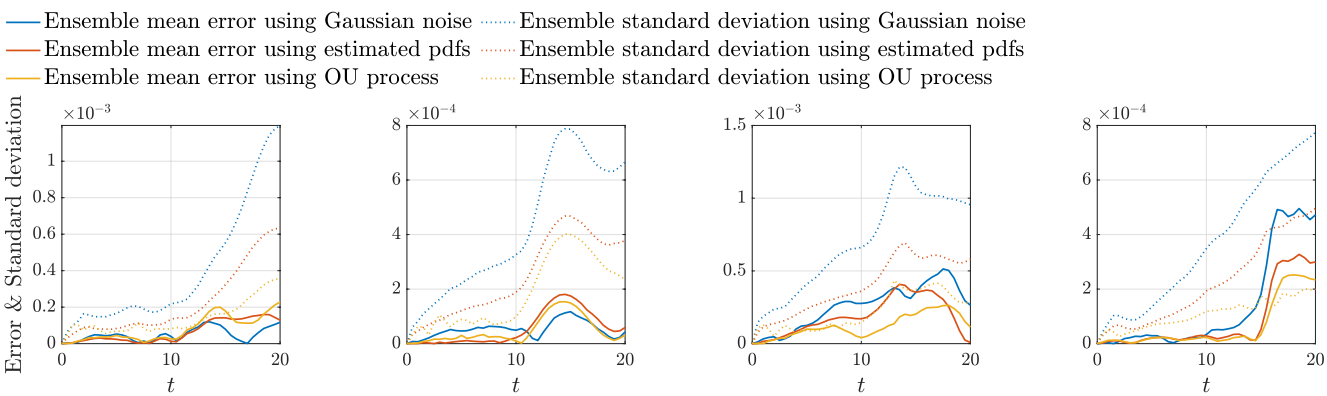

Figure 8. Ensemble mean error with respect to the adapted reference solution and ensemble standard deviation for the streamfunction on four points in the domain. From left to right, (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75). The ensemble mean error is depicted by the solid lines, the ensemble standard deviation by the dotted lines.

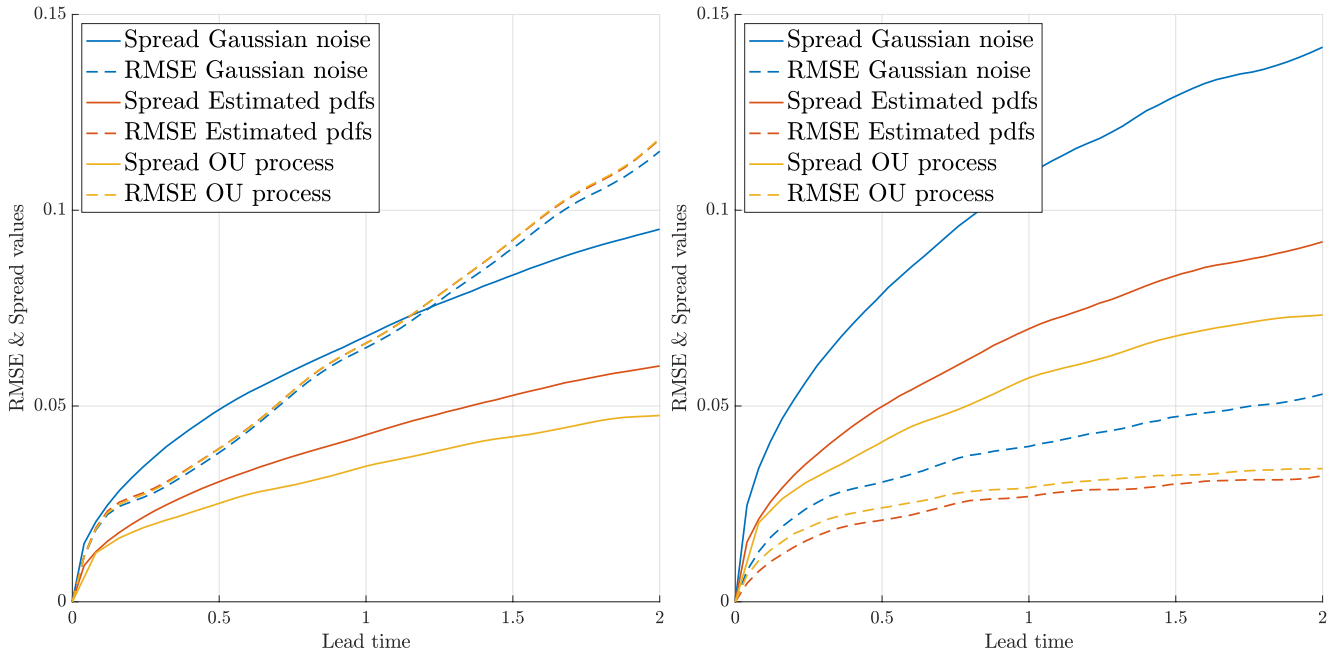

Figure 9. Root mean square error and spread as a function of time when comparing the stochastic ensembles to two different reference solutions. On the left, the filtered direct numerical simulation is regarded as the reference solution and on the right the coarse simulation including the measured $\boldsymbol{\xi}_i$ is used. The data for each figure consists of 71 ensembles of 100 stochastic realizations each.

$$\mathrm{RMSE}(X_{i,j}, Y_j) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\langle X_j\rangle - Y_j)^2}. \tag{22}$$

This provides a measure for the average error of the ensemble (Leutbecher, 2009). The plots in Figure 9 show the development of the RMSE and the spread Equation 20 for increasing lead time for the different stochastic processes. In the left figure the stochastic ensembles are compared to the filtered DNS, in the right figure the ensembles are compared to the adapted reference solution. The RMSE values in the left graph of Figure 9 show rapid growth, indicating that the ensemble mean deviates quickly from the filtered DNS. In contrast, the RMSE values obtained using the adapted reference solution show a significant error reduction. This suggests that the rapid error growth in the left figure is due to the fact that the gap between the coarse-grid SPDE and the filtered DNS contains not only the modeling error but also the discretization error. In addition, the right plot in Figure 9 shows that using the estimated pdfs and the OU process yield similar values of the RMSE and the spreads develop comparably as well.

The second statistical quantity that we compute are rank histograms, which are a tool for measuring the reliability of an ensemble of forecasts (Hamill, 2001). A rank histogram is obtained by plotting the number of occurrences of particular outcomes of the rank function. Here, the rank function R keeps track of where the reference solution appears in the list of sorted ensemble members. That is, given a reference value Y_j and a list of N sorted ensemble members $\{X_{i,j}\}$, R is equal to the integer r that identifies the position of Y_j in the sorted list. It is defined as follows:

$$R(Y_j, \{X_{i,j}\}) = \begin{cases} r & \text{if } Y_j \geq X_{r,j}, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

If the forecast is reliable, then the reference value and the stochastic realizations are indistinguishable. This means that the underlying distributions of the reference value and the stochastic realizations are the same, which implies that the reference value is equally likely to be larger than any number of ensemble members. Thus, the rank function is equally likely to take on any value between 1 and N for reliable forecast ensembles and should therefore produce a rank histogram which approximates a uniform distribution.

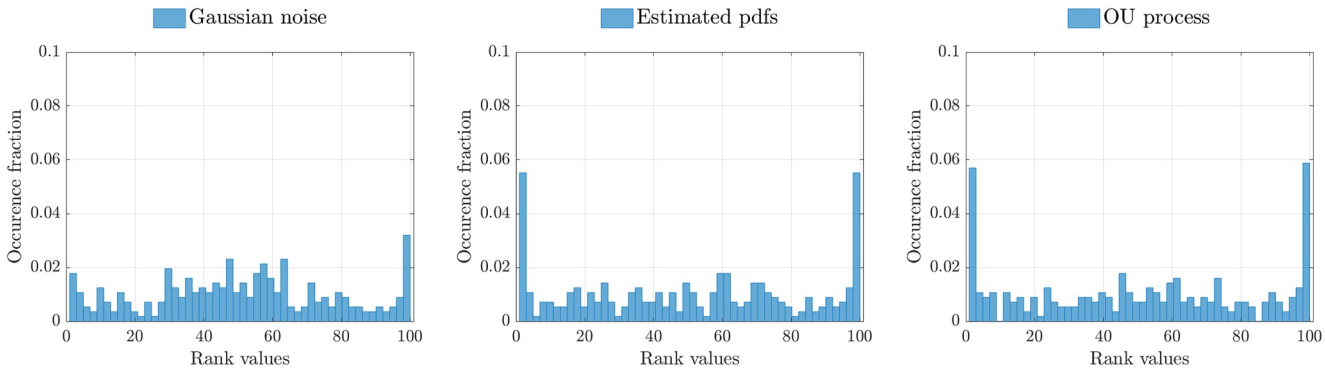

Figure 10. Rank histograms using measurements at the points (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), and (0.75, 0.75) at a lead time of $t = 0.2$. A total of 71 ensembles are computed and measured at the specified points, each consisting of 100 stochastic realizations and compared to the filtered direct numerical simulation at the corresponding time.

Figures 10 and 11 show the rank histograms when using the filtered DNS and the adapted reference solution, respectively, as reference. The measurements at the points (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), and (0.75, 0.75) at a lead time of 0.2 time units are used to generate the histograms. For each ensemble forecast the point values are compared to the reference solutions, leading to 284 ensemble outcomes that are compared to reference values. Only the rank histograms at this particular lead time are shown here, rank histograms at different lead times displayed similar results.

The rank histograms using the filtered DNS (Figure 10) show clear peaks at the edges, caused by all ensemble members either overestimating or underestimating the truth. This effect is least pronounced when applying Gaussian noise, due to the larger spread in the ensemble. The rank histograms obtained when comparing the ensembles to the adapted reference solution (Figure 11) show peaks around the center. This is indicates that the reference solution ranks within the middle range of the ensembles. This is a direct result of the small mean error. The peaks at the edges are significantly reduced when using the adapted reference solution. This is especially clear when using estimated pdfs, which indicates that these ensembles, while showing a small spread, more accurately capture the reference solution. Overall, the differences between the rank histograms of the different methods are small. This indicates that reliability of the ensembles does not seem to depend on the choice of stochastic forcing.

The third statistical quantity that is presented here is the evolution of the vorticity over different time spans, conditioned on the vorticity value at a reference time. That is, the conditional probability distribution

$$P[\omega(t+\tau) - \omega(t) | \omega(t) = \omega_{\text{ref}}] \tag{24}$$

is estimated for different values of τ. This quantity describes the statistical evolution of the vorticity over a time interval of length τ, given a fixed initial configuration.

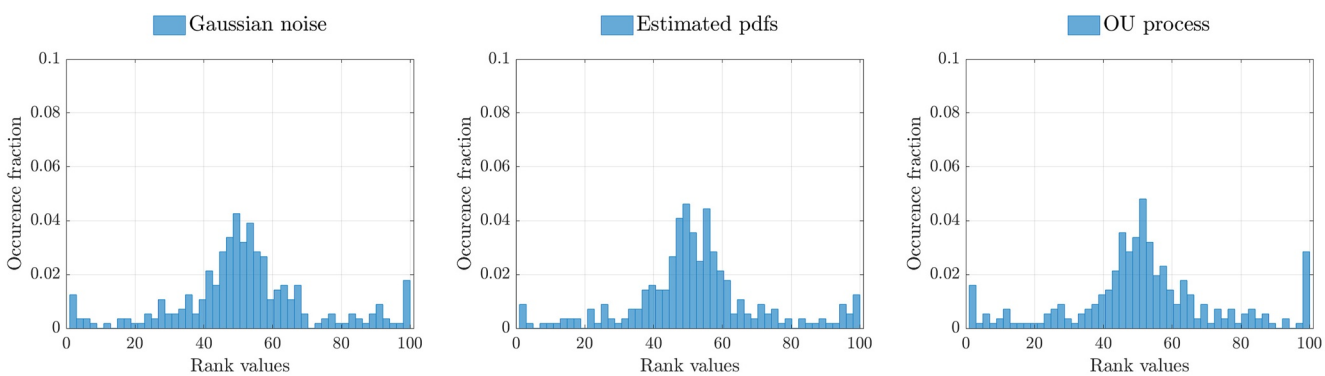

Figure 11. Rank histograms using measurements at the points (0.25, 0.25), (0.25, 0.75), (0.75, 0.25), and (0.75, 0.75) at a lead time of $t = 0.2$. A total of 71 ensembles are computed and measured at the specified points, each consisting of 100 stochastic realizations and compared to the adapted reference solution at the corresponding time.

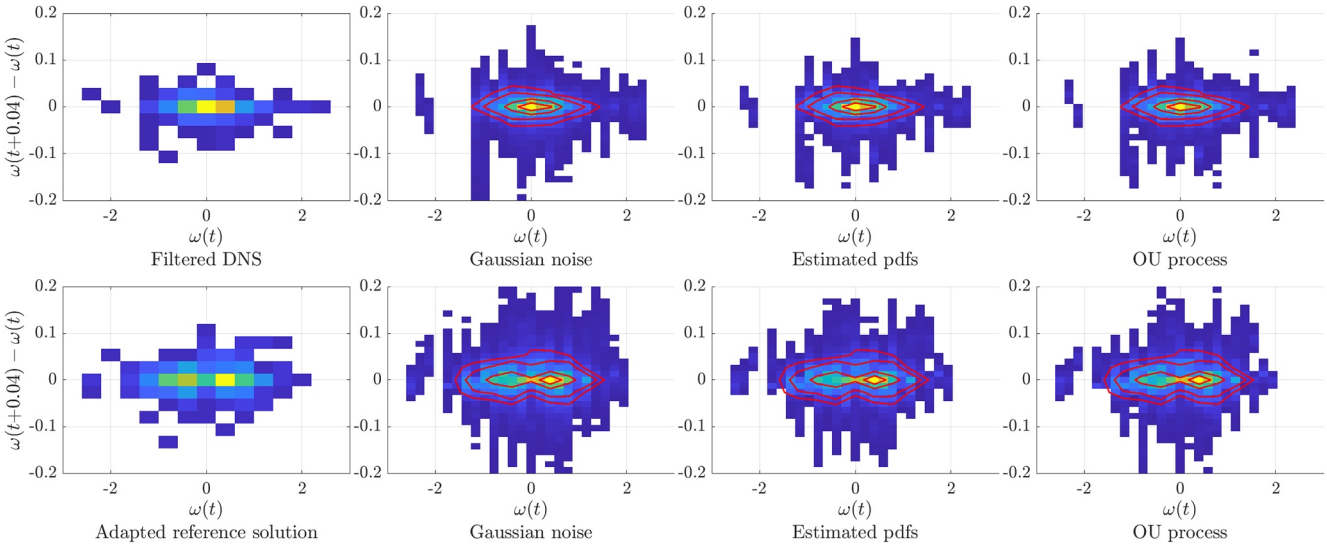

Figure 12. Conditional probability Equation 24 for lead time $\tau = 0.04$. The top row shows the distributions using the filtered direct numerical simulation as a reference, the bottom row uses the adapted reference solution. The contour lines of the reference conditional distributions are overlaid on the distributions obtained from the stochastic ensembles for easier qualitative comparison.

The conditional distributions are shown in Figure 12, at lead time $\tau = 0.04$, and in Figure 13, at lead time $\tau = 1$ to illustrate both short-time and long-time evolution. In both figures, the conditional distributions obtained from the reference solutions are shown in the left panel. For comparison, contour lines of these distributions have been overlaid in the conditional distributions obtained from the stochastic models. The filtered DNS provides the reference for the top row of distributions, the adapted reference solution is used in the bottom row. In particular, the distributions of the stochastic models have been computed from a set of initial conditions sampled along the filtered DNS and the adapted reference solution, respectively. In these figures, a large spread in the vertical direction indicates large uncertainty. This becomes especially clear for the shortest lead times considered. On such short timescales, the stochastic forcing adds considerable variance to the numerical solution. Applying Gaussian noise yields the largest spread, whereas using the estimated pdfs and the OU produce a smaller spread, in accordance with previously presented results. At lead time $\tau = 1$ (Figure 13), the stochastic conditional distributions

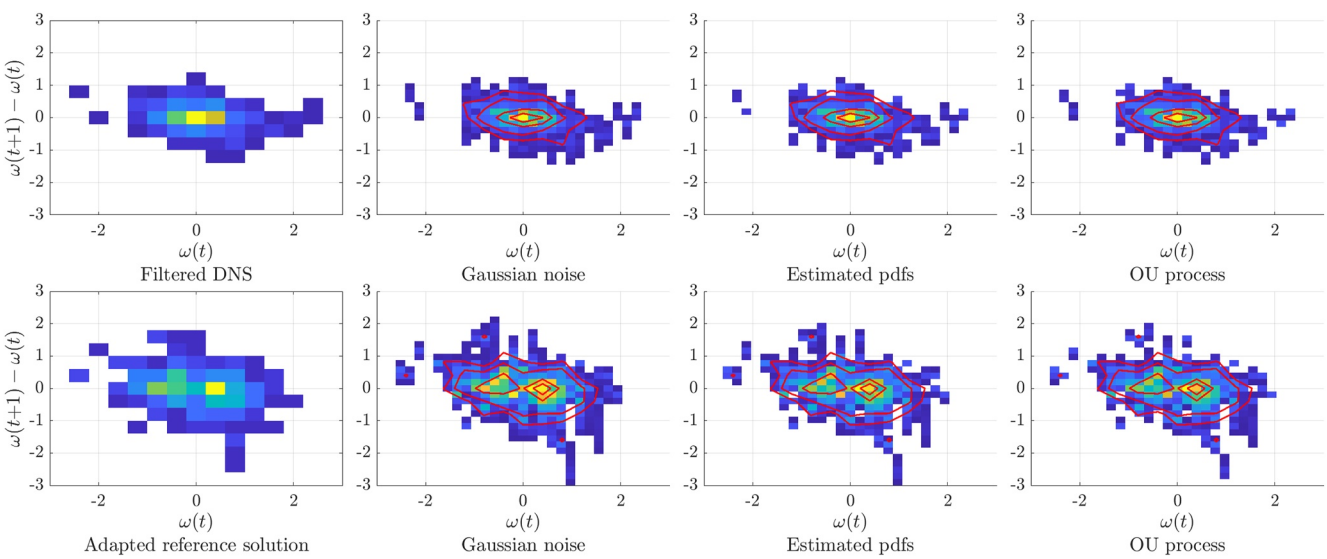

Figure 13. Conditional probability Equation 24 for lead time $\tau = 1$. The top row shows the distributions using the filtered direct numerical simulation as a reference, the bottom row uses the adapted reference solution. The contour lines of the reference conditional distributions are overlaid on the distributions obtained from the stochastic ensembles for easier qualitative comparison.

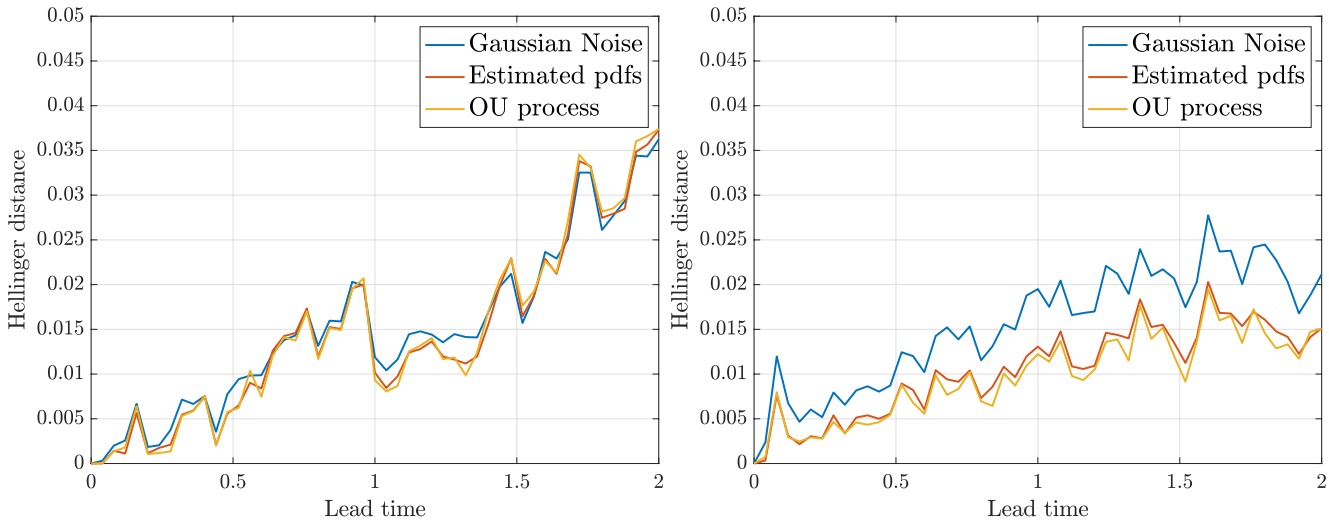

Figure 14. Hellinger distances as a function of time between the reference solution and the stochastic ensembles of distribution Equation 24. On the left, the filtered direct numerical simulation is used as a reference solution, on the right, the adapted reference solution provides the reference.

do not show significant differences. To better judge the agreement between the stochastic conditional distributions and the reference distributions, we compute the Hellinger distance. This measure allows for a quantitative comparison between the different distributions. Given two discrete probability distributions $p = (p_1, \ldots, p_K)$ and $q = (q_1, \ldots, q_K)$, we compute the Hellinger distance (Hellinger, 1909)

$$H^2(p, q) = \frac{1}{2} \sum_{i=1}^{K} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2. \tag{25}$$

The distance $H^2(p, q)$ of Equation 24 is shown in Figure 14 for the filtered DNS (left figure) and for the adapted solution (right figure). The initial conditions of the stochastic ensemble and the reference solutions are the same, therefore the Hellinger distance at $\tau = 0$ is zero. As τ increases, $\omega(t)$ deviates from its reference value and accumulation of error leads to larger values of $H^2(p, q)$. Using the filtered DNS as reference solution yields a comparable Hellinger distance for each method. In contrast, the comparison of the stochastic ensembles to the adapted reference solution clearly favors the models obtained using the estimated pdfs and OU processes over those where Gaussian noise is employed. Despite the quantitative difference in the Hellinger distance, the qualitative behavior is the same for each of the stochastic models.

An overall smaller rate of increase is observed when comparing to the adapted reference solution with respect to the filtered DNS. The latter findings underpin once more the benefits of using the adapted reference solution when assessing the quality of different stochastic models.

3.5. Quantitative Assessment of Results in Spectral Space

To distinguish the quality of the proposed models across different lengthscales we assess the outcomes of the models in spectral space. The EOF modes with a large energy content and a small energy content are representative of large lengthscales and small lengthscales, respectively. Therefore, one might discriminate between the lengthscales that are present in the solution by projecting the latter onto the basis of EOF modes. This translates into applying the projection established in Equation 16 for the reference solutions and the stochastic realizations and subsequently examining the obtained temporal coefficients.

We perform uncertainty quantification at different lengthscales by comparing the EOF coefficients of the stochastic realizations to those of the reference solutions, computed for specified modes. The evolution of the coefficients of four modes, $i = 1, 10, 50, 150$, representative of large, intermediate and small scales, is shown in Figure 15. It is found that the stochastic models accurately follow the adapted reference solution, but deviate somewhat from the filtered DNS result, independently of their lengthscale. Similar to the results in previous subsections, using the OU processes yields the smallest spread, followed by the estimated pdfs and the Gaussian noise.

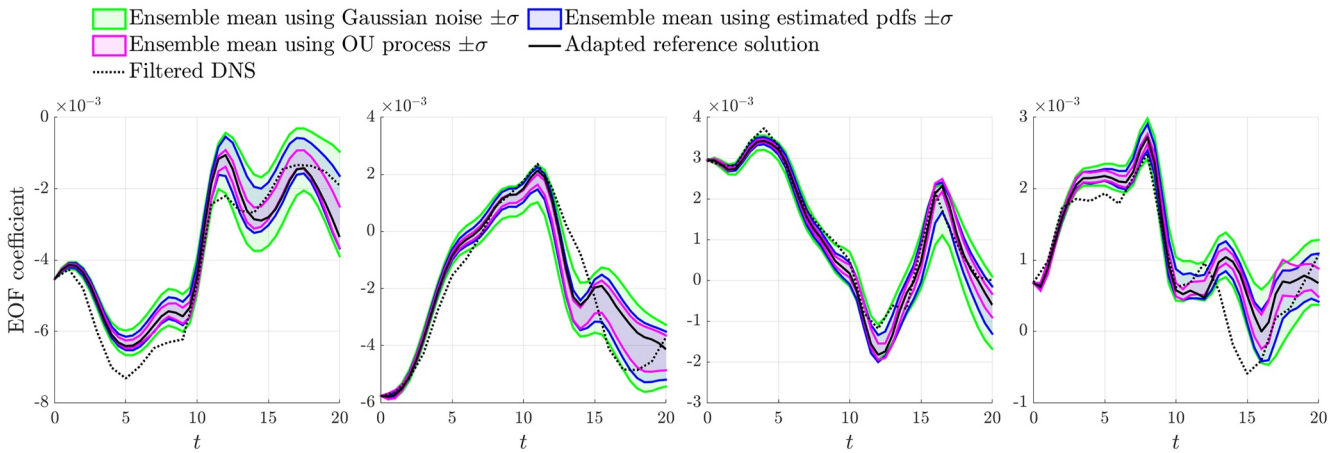

Figure 15. From left to right, empirical orthogonal function coefficients for modes 1, 10, 50, 150. The solid and dotted black lines show the development of the adapted reference solution and filtered direct numerical simulation, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses Ornstein-Uhlenbeck processes. The results for each method are generated from an ensemble of 200 realizations.

To quantify the forecast quality as a function of time, for each EOF mode separately, we define the root integrated mean-squared error (RIMSE):

$$\mathrm{RIMSE(t)} = \frac{1}{t}\frac{\left(\int_0^t \frac{1}{N}\sum_{i=1}^{N}(a_i - a_{\mathrm{ref}})^2\,\mathrm{d}\tau\right)^{1/2}}{\left(\int_0^t a_{\mathrm{ref}}^2\,\mathrm{d}\tau\right)^{1/2}}. \tag{26}$$

This quantity is a measure of the difference between the EOF coefficient a_i of each stochastic realization in the ensemble and the coefficient a_{ref} of the reference solution, integrated over the specified time interval. The values of the RIMSE are shown in Figure 16 for the four modes considered and compared to the adapted reference solution. The results suggest that using OU processes and estimated pdfs is in general favored over using Gaussian noise, largely independent of the lengthscale. For the higher modes, an initial rapid increase in the RIMSE is observed regardless of the employed method. The results obtained using Gaussian noise and OU processes show little difference for short lead times. For increased lead times the latter shows favorable results.

An additional verification of the accuracy of the stochastic realizations is provided by comparing the means and variances of the EOF time series to those of the reference solutions. These values are shown in Figure 17 for all EOF modes. Only the results using Gaussian noise are shown to keep the figures comprehensible. No significant differences were found for the other proposed models. The mean values of the stochastic realizations and the adapted reference solution are found to be nearly indistinguishable, whereas slight deviations from the mean

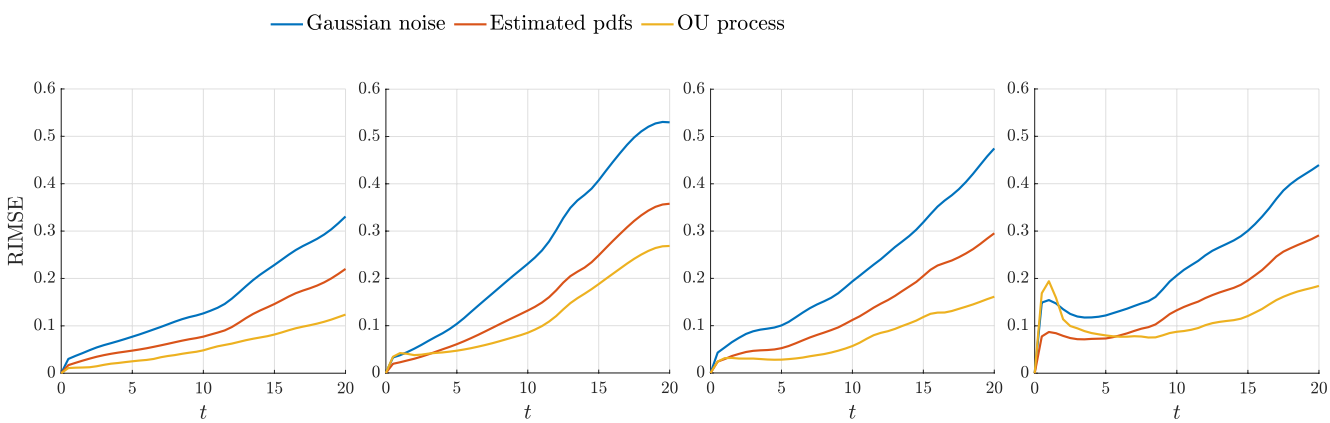

Figure 16. From left to right, root integrated mean-squared error Equation 26 for modes 1, 10, 50, 150, compared to the adapted reference solution. The results for each method are generated from an ensemble of 200 realizations.

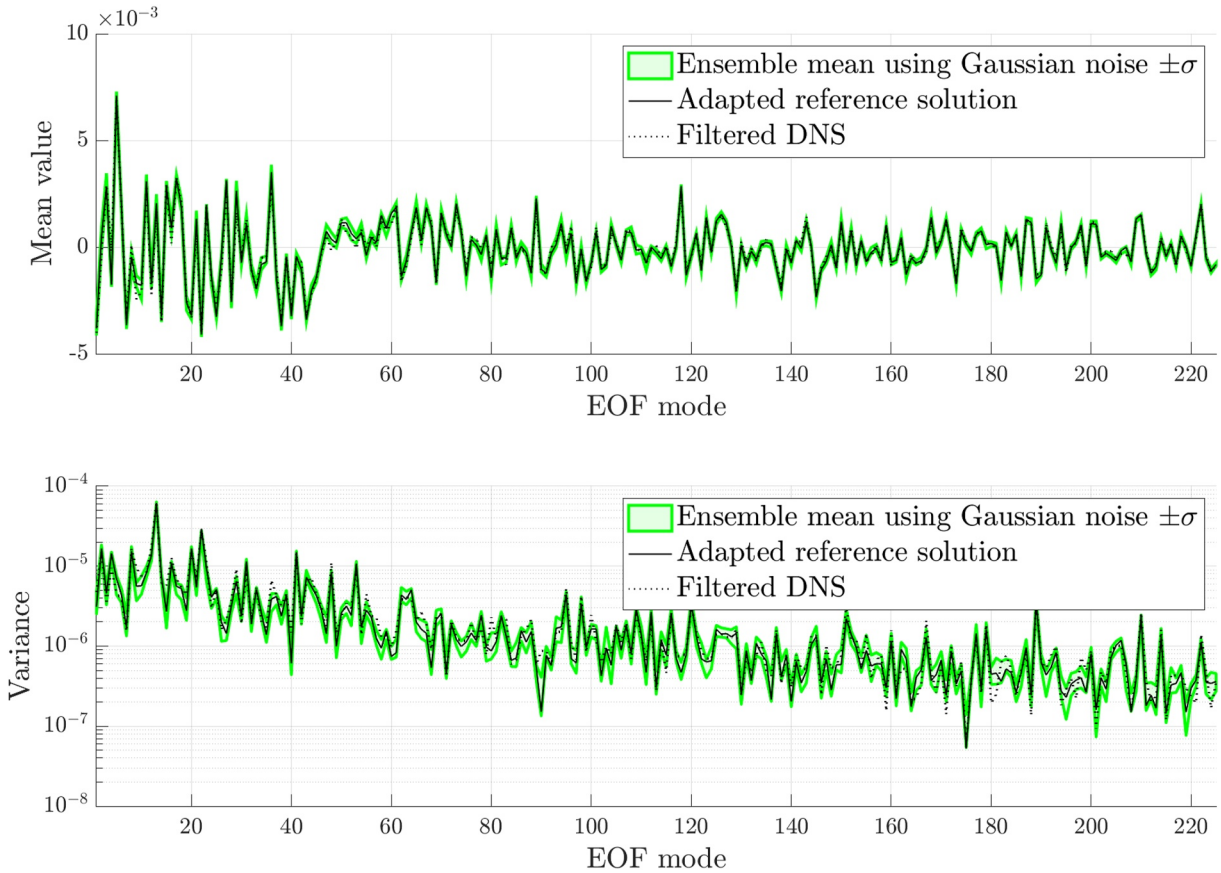

Figure 17. Mean values (top) and variances (bottom) of the empirical orthogonal function time series over the simulated time interval. The solid and dotted show the development of the adapted reference solution and the filtered direct numerical simulation, respectively. The green bands are generated from an ensemble of 200 realizations using Gaussian noise and show the ensemble mean of the quantity of interest ± one ensemble standard deviation.

values of the filtered DNS may be observed. The variances of the time series are also found to be in good agreement with those of the adapted reference solution, but differ marginally from the variances of the time series of the filtered DNS.

4. Conclusions

In this paper, we have assessed three stochastic models for the simulation of the coarse-grained two-dimensional Euler equations. The closure is based on the so-called SALT approach. The resulting SPDE contains a stochastic forcing term that requires to be modeled to close the equations. In particular, the forcing is decomposed into a deterministic basis (EOF, or EOFs) multiplied by stochastic temporal traces. This decomposition is, by construction, fully determined from a fine-grid (DNS) data set. However, to simulate outside the available data set one is required to model the time traces. In the framework of SALT (Cotter et al., 2019) the latter are regarded as Gaussian processes. Here we extend the stochastic forcing to more general processes, sampling from the data-estimated probability distribution functions (pdfs) and introducing correlation through OU processes. The latter two methods use additional data already available from the EOF time series. Between the methods no qualitative differences in the flow realizations were observed. However, the latter methods generally show favorable results compared to the former Gaussian method, in terms of ensemble mean and ensemble spread.

To meaningfully compare the different stochastic models we defined a maximal prediction horizon and an adapted reference solution. The prediction horizon sets the point in time beyond which a bundle of fine-grid solutions, starting from the same initial condition on the coarse grid, deviates on order 1 due to high sensitivity to the initial conditions. This defines the time frame on which to assess the statistical quality of the coarse-grid predictions. The adapted reference solution was defined as the coarse-grid solution using the exact measured time

series of the EOFs for the forcing. The latter allowed us to isolate the modeling error from other sources of error not taken into account in the considered model formulation, such as discretization error. The stochastic ensembles were compared to this reference solution using a global measure and pointwise values. For both the global and local measures, using either estimated pdfs or OU processes to define the forcing term yielded a smaller ensemble mean error and a smaller spread compared to using Gaussian noise.

Stochastic prediction ensembles on timescales relevant for data assimilation were further investigated by performing statistical tests, comparing ensembles of stochastic realizations to the adapted reference solution and the filtered DNS. A significantly smaller ensemble spread was found when using estimated pdfs or OU processes, compared to using Gaussian noise. Additionally, the observed mean ensemble error was lower for the former two methods. All three methods showed rapid growth in ensemble error when compared to the filtered DNS, suggesting that the filtered DNS contains not only the modeling error but also the discretization error and the closure error. These results were further substantiated by rank histograms, showing that the ensembles were biased with respect to the filtered DNS, but were overdispersive compared to the adapted reference solution. In particular, using the estimated pdfs to define the stochastic forcing rarely resulted in the adapted reference solution not being contained in the ensemble. Finally, conditional distributions of the vorticity were computed and compared using the Hellinger distance. Here, using estimated pdfs or OU processes resulted in a smaller distance to the reference solution than using Gaussian noise, indicating a better statistical characterization of the vorticity dynamics.

The ensemble forecasts were assessed in spectral space to discriminate between the different lengthscales present in the solution. The stochastic ensembles were found to accurately capture the adapted reference solution on the considered scales. The overall prediction quality using OU processes was found to be favorable over using Gaussian noise, independent of the lengthscale. Additionally, the stochastic ensembles were found to accurately reproduce the mean values and variances of the EOF time series over the entire simulation period.

The methods presented in this paper may be used in other flows where EOF-based stochastic modeling is relevant. These approaches are particularly appealing since all information used in these methods is readily available from the EOF decomposition and no additional data is required to construct the models. The presented techniques are purely data-driven, they require no further assumption about the governing equations and can therefore be applied to other geophysical fluids. The short-time results indicate that a mean error reduction and smaller ensemble spread can be obtained using these methods, which can complement methods employed in data assimilation. Furthermore, the definition of the adapted reference solution motivates further research of the SALT method using different closure models and incorporating the discretization error.

Data Availability Statement

The computational code, data and post-processing scripts used in this study can be accessed via https://doi.org/10.5281/zenodo.6719311. Computations were performed in the open-source Python Finite Element Package Firedrake (McRae et al., 2016). All post-processing was done in Matlab (MATLAB version 9.7.0.1190202 (R2019b), 2019).

Acknowledgments
The authors would like to thank the associate editor and the two anonymous referees for their valuable input and suggestions that have helped to improve the paper. In addition, the authors would like to thank Wei Pan, at the Department of Mathematics, Imperial College London, for his help preparing the numerical experiments. We are grateful to thank Darryl Holm and James-Michael Leahy, at the Department of Mathematics, Imperial College London, and Arnout Franken, at the University of Twente, for the many inspiring discussions we had in the context of the SPRESTO project, funded by the Dutch Science Foundation (NWO) in their TOP1 program.

References

Adrian, R. J., & Westerweel, J. (2011). *Particle image velocimetry (no. 30)*. Cambridge University Press.

Agarwal, N., Kondrashov, D., Dueben, P., Ryzhov, E., & Berloff, P. (2021). A comparison of data-driven approaches to build low-dimensional ocean models. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002537. https://doi.org/10.1029/2021ms002537

Arnold, H., Moroz, I., & Palmer, T. (2013). Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *371*(1991), 20110479. https://doi.org/10.1098/rsta.2011.0479

Bernsen, E., Bokhove, O., & van der Vegt, J. J. (2006). A (dis) continuous finite element model for generalized 2D vorticity dynamics. *Journal of Computational Physics*, *211*(2), 719–747. https://doi.org/10.1016/j.jcp.2005.06.008

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018ms001472

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908. https://doi.org/10.1002/qj.49712556006

Cotter, C. J., Crisan, D., Holm, D. D., Pan, W., & Shevchenko, I. (2019). Numerically modeling stochastic Lie transport in fluid dynamics. *Multiscale Modeling and Simulation*, *17*(1), 192–232. https://doi.org/10.1137/18m1167929

Cotter, C. J., Crisan, D., Holm, D. D., Pan, W., & Shevchenko, I. (2020a). Data assimilation for a quasi-geostrophic model with circulation-preserving stochastic transport noise. *Journal of Statistical Physics*, *179*(5), 1186–1221. https://doi.org/10.1007/s10955-020-02524-0

Cotter, C. J., Crisan, D., Holm, D. D., Pan, W., & Shevchenko, I. (2020b). Modelling uncertainty using stochastic transport noise in a 2-layer quasi-geostrophic model. *Foundations of Data Science*, *2*(2), 173–205. https://doi.org/10.3934/fods.2020010

Cotter, C. J., Crisan, D., Holm, D. D., Pan, W., & Shevchenko, I. (2020c). A particle filter for stochastic advection by lie transport: A case study for the damped and forced incompressible two-dimensional Euler equation. *SIAM/ASA Journal on Uncertainty Quantification*, *8*(4), 1446–1492. https://doi.org/10.1137/19m1277606

Cotter, C. J., Gottwald, G. A., & Holm, D. D. (2017). Stochastic partial differential fluid equations as a diffusive limit of deterministic Lagrangian multi-time dynamics. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *473*(2205), 20170388. https://doi.org/10.1098/rspa.2017.0388

Crisan, D., Holm, D. D., Leahy, J.-M., & Nilssen, T. (2022). Variational principles for fluid dynamics on rough paths. *Advances in Mathematics*, *404*, 108409. https://doi.org/10.1016/j.aim.2022.108409

Crommelin, D., & Vanden-Eijnden, E. (2008). Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences*, *65*(8), 2661–2675. https://doi.org/10.1175/2008jas2566.1

Devroye, L. (2006). Nonuniform random variate generation. In *Handbooks in operations research and management science* (Vol. 13, pp. 83–121).

Frederiksen, J. S., & Davies, A. G. (1997). Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *Journal of the Atmospheric Sciences*, *54*(20), 2475–2492. https://doi.org/10.1175/1520-0469(1997)054<2475:evasbp>2.0.co;2

Frederiksen, J. S., & Kepert, S. M. (2006). Dynamical subgrid-scale parameterizations from direct numerical simulations. *Journal of the Atmospheric Sciences*, *63*(11), 3006–3019. https://doi.org/10.1175/jas3795.1

Frederiksen, J. S., Kitsios, V., O'Kane, T. J., & Zidikheri, M. J. (2017). Stochastic subgrid modelling for geophysical and three-dimensional turbulence. In *Nonlinear and stochastic climate dynamics* (pp. 241–275). Cambridge University Press.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896. https://doi.org/10.1029/2019ms001896

Geurts, B. J., & Fröhlich, J. (2002). A framework for predicting accuracy limitations in large-eddy simulation. *Physics of Fluids*, *14*(6), L41–L44. https://doi.org/10.1063/1.1480830

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(9), 1119–1152. https://doi.org/10.1002/joc.1499

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die Reine und Angewandte Mathematik*, *1909*(136), 210–271. https://doi.org/10.1515/crll.1909.136.210

Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, *43*(3), 525–546. https://doi.org/10.1137/s0036144500378302

Holm, D. D. (2015). Variational principles for stochastic fluid dynamics. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *471*(2176), 20140963. https://doi.org/10.1098/rspa.2014.0963

Holm, D. D., & Luesink, E. (2021). Stochastic wave–current interaction in thermal shallow water dynamics. *Journal of Nonlinear Science*, *31*(2), 1–56. https://doi.org/10.1007/s00332-021-09682-9

Kitsios, V., & Frederiksen, J. S. (2019). Subgrid parameterizations of the eddy–eddy, eddy–mean field, eddy–topographic, mean field–mean field, and mean field–topographic interactions in atmospheric models. *Journal of the Atmospheric Sciences*, *76*(2), 457–477. https://doi.org/10.1175/jas-d-18-0255.1

Kloeden, P. E., & Platen, E. (1992). Stochastic differential equations. In *Numerical solution of stochastic differential equations* (pp. 103–160). Springer.

Leutbecher, M. (2009). Diagnosis of ensemble forecasting systems. In *Seminar on diagnosis of forecasting and data assimilation systems* (pp. 235–266).

Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proceedings of the seminar on predictability* (Vol. 1).

Lumley, J. L. (1967). The structure of inhomogeneous turbulent flows. Atmospheric turbulence and radio wave propagation.

Majda, A. J., Timofeyev, I., & Vanden Eijnden, E. (2001). A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, *54*(8), 891–974. https://doi.org/10.1002/cpa.1014

MATLAB version 9.7.0.1190202 (R2019b). (2019). (Computer software manual), Natick, Massachusetts.

McRae, A. T. T., Bercea, G.-T., Mitchell, L., Ham, D. A., & Cotter, C. J. (2016). Automated generation and symbolic manipulation of tensor product finite elements [Software]. SIAM Journal on Scientific Computing, *38*(5), S25–S47. https://doi.org/10.1137/15M1021167

Mémin, E. (2014). Fluid flow dynamics under location uncertainty. *Geophysical and Astrophysical Fluid Dynamics*, *108*(2), 119–146. https://doi.org/10.1080/03091929.2013.836190

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018ms001351

Palmer, T. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), 463–471. https://doi.org/10.1038/s42254-019-0062-2

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, *63*(2), 71–116. https://doi.org/10.1088/0034-4885/63/2/201

Pope, S. B. (2001). *Turbulent flows*. IOP Publishing.

Resseguier, V., Pan, W., & Fox-Kemper, B. (2020). Data-driven versus self-similar parameterizations for stochastic advection by lie transport and location uncertainty. *Nonlinear Processes in Geophysics*, *27*(2), 209–234. https://doi.org/10.5194/npg-27-209-2020

Shu, C.-W., & Osher, S. (1988). Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, *77*(2), 439–471. https://doi.org/10.1016/0021-9991(88)90177-5

Street, O. D., & Crisan, D. (2021). Semi-martingale driven variational principles. *Proceedings of the Royal Society A*, *477*(2247), 20200957. https://doi.org/10.1098/rspa.2020.0957

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Zeitlin, V. (2018). *Geophysical fluid dynamics: Understanding (almost) everything with rotating shallow water models*. Oxford University Press.