Towards Data-Driven Dynamic Surrogate Models for Ocean Flow

Wouter Edeling
Centrum Wiskunde & Informatica
Amsterdam
Wouter.Edeling@CWI.nl

Daan Crommelin
Centrum Wiskunde & Informatica / Korteweg-de Vries Institute for Mathematics
Amsterdam
Daan.Crommelin@CWI.nl

ABSTRACT

Coarse graining of (geophysical) flow problems is a necessity brought upon us by the wide range of spatial and temporal scales present in these problems, which cannot be all represented on a numerical grid without an inordinate amount of computational resources. Traditionally, the effect of the unresolved eddies is approximated by deterministic closure models, i.e. so-called parameterizations. The effect of the unresolved eddy field enters the resolved-scale equations as a forcing term, denoted as the 'eddy forcing'. Instead of creating a deterministic parameterization, our goal is to infer a stochastic, data-driven surrogate model for the eddy forcing from a (limited) set of reference data, with the goal of accurately capturing the long-term flow statistics. Our surrogate modelling approach essentially builds on a resampling strategy, where we create a probability density function of the reference data that is conditional on (time-lagged) resolved-scale variables. The choice of resolved-scale variables, as well as the employed time lag, is essential to the performance of the surrogate. We will demonstrate the effect of different modelling choices on a simplified ocean model of two-dimensional turbulence in a doubly periodic square domain.

CCS CONCEPTS

• **Mathematics of computing → Stochastic processes**; • **Computing methodologies → Multiscale systems**; • **Applied computing** → *Environmental sciences.*

KEYWORDS

Dynamic surrogate modelling, Multiscale modelling, Ocean flow

ACM Reference Format:
Wouter Edeling and Daan Crommelin. 2019. Towards Data-Driven Dynamic Surrogate Models for Ocean Flow. In *Proceedings of the Platform for Advanced Scientific ComputingConference (PASC '19), June 12–14, 2019, Zurich, Switzerland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3324989.3325713

1 INTRODUCTION

Numerical simulation of turbulent flow problems is a very computationally expensive enterprise, which introduces the need for coarse-grained solutions. Thus, one has to cope with processes which cannot be resolved directly on the numerical grid. The effect of the unresolved eddy field enters the resolved-scale equations as a forcing term, denoted as the eddy forcing, which is highly complex, dynamic, and shows intricate spatio-temporal correlations. Traditionally, the effect of the unresolved processes on the resolved solution is approximated by deterministic closure models, i.e. so-called parameterizations. In the context of geophysical flows, such parameterizations are based on e.g. the work of Gent-McWilliams [7], or through the inclusion of a tunable (hyper) viscosity term meant to damp the smallest resolved scales of the model [12].

Such parameterizations constitute a clear improvement with respect to unparameterized case where the unresolved scales are simply ignored. Still, it is well known that no parameterization scheme is perfect, and attempts have been made to improve their performance. For instance, the authors of [16] analysed the transfer of energy and enstrophy in spectral space for a number of parameterizations, and compared their performance to a reference solution of a two-dimensional turbulent flow case. They proposed a deterministic 'energy fixer' scheme, based on adding a weighted vorticity pattern to the computed vorticity field. Recently, data-driven techniques have been applied as well. For instance the recent work of [11] used artificial neural networks to learn the eddy forcing from a set of reference snapshots. A (deterministic) map between local stencils of resolved variables and the eddy forcing was created, obtaining a dynamic surrogate model for the latter.

However, a general limitation of any deterministic approach is their inability to represent the strong non-uniqueness of the unresolved scales with respect to the resolved scales [1, 13, 17]. Since the resolved scales are generally defined as the convolution of the full-scale solution with some filter, multiple unresolved states can correspond to the same resolved solution. Thus there is no one-to-one correspondence between the two scales, and yet deterministic parameterizations do assume such correspondence.

As a result, stochastic methods for representing the unresolved scales have received an increasing amount of attention. Early contributions to this topic in the context of ocean modelling includes the work of [1], where the eddy-forcing is replaced by a space-time correlated random-forcing process. Other notable examples include the work of [8, 10, 21]. Probability density functions (pdfs) of the eddy forcing were constructed using a reference solution, conditioned on a suitable, resolved-scale variable which showed high-correlation with respect to the reference eddy forcing.

PASC '19, June 12–14, 2019, Zurich, Switzerland

ACM ISBN 978-1-4503-6770-7/19/06…$15.00
https://doi.org/10.1145/3324989.3325713

Here. we consider a stochastic and purely data-driven method, in order to extract a dynamic surrogate model of the eddy forcing from a series of reference snapshots. Thus, by a surrogate we do not mean a method to propagate uncertainty through the solver with the goal of obtaining output statistics, as e.g. polynomial chaos or stochastic collocation methods[6]. Instead, we create a surrogate to replace an unclosed source term in a dynamical system. Besides adding stochasticity, our hypothesis is that the use of a reference solution could lead to (dynamical) improvement upon traditional parameterizations, considering that the reference data contains the exact dynamics we aim to capture. However, this will be conditional on our ability to create a suitable mapping between resolved-scale variables and the reference data. Our approach essentially builds on the work of [17, 18], where a resampling strategy of reference data conditional on (time-lagged) resolved-scale variables was developed. As a test case, we consider a two-dimensional ocean flow model, based on the forced, two-dimensional and incompressible vorticity equations. The purpose of the current article is to examine the impact of the choice of conditional variables, as well as the employed time lag, on the performance of the method. As a performance indicator we will use the degree by which time-averaged energy and enstrophy statistics are captured.

The article is organised as follows. In Section 2 we describe the governing equations and multiscale decomposition, followed by a section describing the method by which we construct the surrogate model for the eddy forcing. Initial results are shown in Section 4, and finally the conclusion and outlook are given in Section 5.

2 GOVERNING EQUATIONS

We study the same model as in [19], i.e. the forced-dissipative vorticity equations for two-dimensional incompressible flow. The governing equations read

$$\frac{\partial \omega}{\partial t} + J(\Psi, \omega) = \nu \nabla^2 \omega + \mu (F - \omega), \qquad \nabla^2 \Psi = \omega. \tag{1}$$

Here, ω is the vertical component of the vorticity, defined from the curl of the velocity field $\mathbf{V}$ as $\omega := \mathbf{e}_3 \cdot \nabla \times \mathbf{V}$, where $\mathbf{e}_3 := (0, 0, 1)^T$. The stream function Ψ relates to the horizontal velocity components by the well-known relations $u = -\partial\Psi/\partial y$ and $v = \partial\Psi/\partial x$. As in [19], the forcing term is chosen as the single Fourier mode $F = 2^{3/2}\cos(5x)\cos(5y)$. The system is fully periodic in x and y directions over a period of $2\pi L$, where L is a user-specified length scale, chosen as the earth's radius ($L = 6.371 \times 10^6 [m]$). The inverse of the earth's angular velocity Ω^{-1} is chosen as a time scale, where $\Omega = 7.292 \times 10^{-5} [s^{-1}]$. Thus, a simulation time period of a single 'day' can now be expressed as $24 \times 60^2 \times \Omega \approx 6.3$ non-dimensional time units. Given these choices, (1) is non-dimensionalized, and solved using values of ν and μ chosen such that a Fourier mode at the smallest retained spatial scale is exponentially damped with an e-folding time scale of 5 and 90 days respectively. We note that our target statistics (energy and enstrophy), are only conserved in the case of $\nu = \mu = 0$ [19]. For more details on the numerical setup we refer to [19]. Furthermore, our Python source code can be found in [3].

Finally, the key term in (1) is the Jacobian, i.e. the nonlinear advection term defined as

$$J(\Psi, \omega) := \frac{\partial \Psi}{\partial x}\frac{\partial \omega}{\partial y} - \frac{\partial \Psi}{\partial y}\frac{\partial \omega}{\partial x}. \tag{2}$$

It is this term that leads to the need for a closure model when (1) the discretized on a relatively coarse grid which lacks the resolution to capture all turbulent eddies.

2.1 Discretization

We solve (1) by means of a spectral method, where we apply a truncated Fourier expansion:

$$\omega_{\mathbf{k}}(x, y, t) = \sum_{\mathbf{k}} \hat{\omega}_{\mathbf{k}}(t) e^{i(k_1 x + k_2 y)},$$
$$\Psi_{\mathbf{k}}(x, y, t) = \sum_{\mathbf{k}} \hat{\Psi}_{\mathbf{k}}(t) e^{i(k_1 x + k_2 y)}. \tag{3}$$

The sum is taken over the components k_1 and k_2 of the wave number vector $\mathbf{k} := (k_1, k_2)^T$, and $-K' \leq k_i \leq K'$, $i = 1, 2$. These decompositions are inserted in (1), and solved for the Fourier coefficients $\hat{\omega}_{\mathbf{k}}$, $\hat{\Psi}_{\mathbf{k}}$ by means of the real Fast Fourier Transform. To avoid the aliasing problem in the nonlinear term (2), we use the pseudo spectral method, such that in practice the maximum resolved wave number is K, where $K \leq 2K'/3$ [14].

To advance the solution in time we use the second-order accurate AB/BDI2 scheme, which results in the following discrete system of equations [14]

$$\frac{3\hat{\omega}_{\mathbf{k}}^{i+1} - 4\hat{\omega}_{\mathbf{k}}^{i} + \hat{\omega}_{\mathbf{k}}^{i-1}}{2\Delta t} + 2\hat{J}_{\mathbf{k}}^{i} - \hat{J}_{\mathbf{k}}^{i-1} = -\nu k^2 \hat{\omega}_{\mathbf{k}}^{i+1} + \mu\left(\hat{F}_{\mathbf{k}} - \hat{\omega}_{\mathbf{k}}^{i+1}\right)$$
$$-k^2 \hat{\Psi}_{\mathbf{k}}^{i+1} - \hat{\omega}_{\mathbf{k}}^{i+1} = 0. \tag{4}$$

Here, $\Delta t = 0.01$ and $\hat{J}_{\mathbf{k}}^{i}$ is the Fourier coefficient of the Jacobian at time level i, computed with the pseudo spectral technique, and $k^2 := k_1^2 + k_2^2$.

2.2 Multiscale decomposition

As in [19], we apply a spectral filter in order to decompose the full reference solution into a resolved ($\mathcal{R}$) and an unresolved component ($\mathcal{U}$), i.e. we use

$$\hat{\omega}_{\mathbf{k}}^{\mathcal{R}} = P^{\mathcal{R}} \hat{\omega}_{\mathbf{k}}, \qquad \hat{\omega}_{\mathbf{k}}^{\mathcal{U}} = P^{\mathcal{U}} \hat{\omega}_{\mathbf{k}}, \tag{5}$$

where the projection operators $P^{\mathcal{R}}$ and $P^{\mathcal{U}}$ are depicted in Figure 1. Note that the full projection operator $P := \mathcal{P}^{\mathcal{R}} + \mathcal{P}^{\mathcal{U}}$ also removes wave numbers due to the use of the pseudo spectral method.

Applying the resolved projection operator to the governing equations (1) results in the following resolved-scale transport equation

$$\frac{\partial \omega^{\mathcal{R}}}{\partial t} + \mathcal{P}^{\mathcal{R}} J(\Psi, \omega) = \nu \nabla^2 \omega^{\mathcal{R}} + \mu\left(F^{\mathcal{R}} - \omega^{\mathcal{R}}\right) \tag{6}$$

As mentioned, the key term is the Jacobian (2), since due to its non linearity, $\mathcal{P}^{\mathcal{R}} J(\Psi,\ \omega) \neq \mathcal{P}^{\mathcal{R}} J\left(\Psi^{\mathcal{R}},\ \omega^{\mathcal{R}}\right)$. We therefore write

$$J(\Psi, \omega) - J\left(\Psi^{\mathcal{R}},\ \omega^{\mathcal{R}}\right) =: r, \tag{7}$$

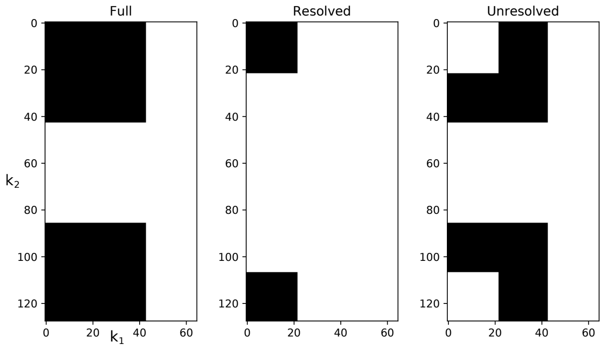


Figure 1: The spectral filter (black=1, white=0) plotted in wave number space (k_1, k_2), of the full, resolved and unresolved solutions. Due to the fact that we use the real FFT algorithm, only part of the spectrum is computed, as Fourier coefficients with opposite values of k are complex conjugates in order to enforce real ω and Ψ fields [14].

such that r is the exact subgrid-scale term, commonly referred to as the 'eddy forcing' [1]. The resolved-scale equation (6) can now be written as

$$\frac{\partial\omega^{\mathcal{R}}}{\partial t} + \mathcal{P}^{\mathcal{R}} J\left(\Psi^{\mathcal{R}}, \omega^{\mathcal{R}}\right) = \nu\nabla^2\omega^{\mathcal{R}} + \mu\left(F^{\mathcal{R}} - \omega^{\mathcal{R}}\right) - \bar{r}. \quad (8)$$

We use the notation $\bar{r} := \mathcal{P}^{\mathcal{R}} r$ for the sake of brevity. A snapshot of the resolved vorticity $\omega^{\mathcal{R}}$ and corresponding resolved eddy forcing $\bar{r}$ is depicted in Figure 2. Notice the fine-grained character of the eddy forcing compared to the vorticity field. From a multiscale point of view, we therefore consider the system (8) an interesting problem, and its spectral approximation is fast enough to allow for the prototyping of a surrogate for $\bar{r}$, which is our main goal. However, from an oceanographic point of view, (8) is admittedly rather simple, as it does not contain e.g. bottom friction or the Coriolis effect. Nonetheless, the problem of parameterizing the eddy forcing remains relevant in more realistic ocean models.

Equation (8) is still unclosed due to the ω and Ψ dependence of (7). As noted, our overall goal is to create a surrogate for $\bar{r}$.

3 SURROGATE EDDY FORCING

For our present purpose, we define an ideal surrogate for the eddy forcing as one which satisfies the following set of requirements:

(1) **Data-driven**: In absence of a single best deterministic parameterization of r, we opt for a model inferred from a pre-computed database of reference data. The generation of this database is described in Section 4.
(2) **Stochastic**: In general, the resolved scales are defined as a convolution of the full solution with some (spatial/spectral) filter. As a result there is no longer just a single unresolved-scale field that is consistent with the resolved-scale solution. This ambiguity provides us with the motivation for a stochastic model for the unresolved, small-scale fields.
(3) **Correlated in space and time**: As demonstrated by Figure 2, the reference eddy forcing shows complex spatial structure. A surrogate would ideally reflect this as well. That is, despite its stochastic nature we would like to prevent the surrogate from producing fields which are too noisy, and lack all spatial correlation.
(4) **Conditional on the resolved variables**: The resolved and unresolved scales are in reality two-way coupled. Hence, the eddy-forcing surrogate should not be independent from the resolved solution.
(5) **Pre-computed & cheap**: While the reference database can be computationally expensive to compute, the resulting data-driven surrogate must be cheap.
(6) **Extrapolates well**: To justify the cost of creating the reference database in the first place, the model must be able to predict the chosen quantity of interest well, *substantially beyond* the time and/or spatial domain of the data.

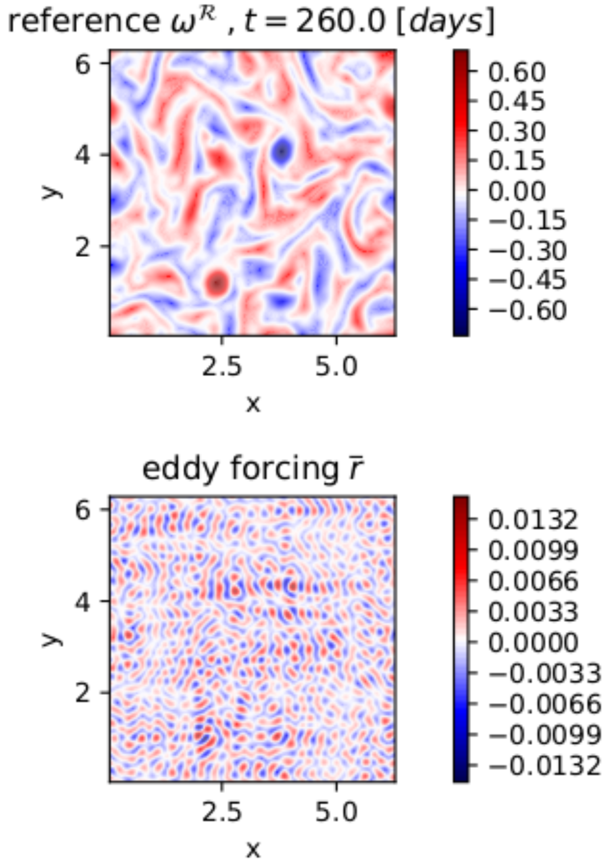


Figure 2: A snapshot of the exact, reference vorticity field $\omega^{\mathcal{R}}$ and the corresponding eddy forcing.

Ultimately, we will measure the performance of a surrogate model by its ability to accurately represent the time-averaged flow statistics. Thus, we do not expect from the model with eddy-forcing surrogate the ability to produce individual flow fields which are in absolute lockstep with the reference data, especially considering the stochastic nature of the surrogate.

We will build on the resampling strategies as developed by [2, 17]. In general, these methods model the unresolved term at time t_{i+1} by sampling from the conditional probability distribution of the reference data:

$$\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid C_i,\ C_{i-1} \cdots . \tag{9}$$

Here $\widetilde{r}_{i+1}$ denotes the surrogate eddy forcing at time t_{i+1}, whereas as $\overline{r}_{i+1}$ represent actual data from the reference model. The set of 'conditioning variables' C_i contains variables that are available from the resolved model, e.g. they can be (functions of) $\omega^{\mathcal{R}}$ or $\widetilde{r}_i$. Examples of these conditional distributions are $\overline{r}_{i+1} \mid \widetilde{r}_i$ and $\overline{r}_{i+1} \mid \widetilde{r}_i, \omega_i^{\mathcal{R}}$. We could assume a Markov property $(r_{i+1} \mid C_i)$, or build in a larger time history. Note that by design, (9) is already data-driven, stochastic, correlated in time, and conditioned on resolved variables. We leave an evaluation of the cost (point 5) of the surrogate for a later study. Most of the cost will be concentrated in an offline phase (i.e. precomputing the reference database), when predicting we just resample reference data. The efficiency gain will grow with a larger difference in grid resolution. In our current setup, the number of grid points differs just by a factor of 2 in each spatial dimension. Moreover, for now we use the same approach as [19], where we evaluate all models on the reference grid, the resolved model just has fewer non-zero Fourier coefficients (see Figure 1). While computationally convenient, this approach complicates a straightforward cost comparison.

The main challenges with this approach, that must be met before the remaining goals could be achieved, are twofold. Clearly, the first challenge concerns the actual formation of the conditional distribution, i.e. how to map the observed C_i to some plausible subset of $\overline{r}_{i+1}$ samples from which $\widetilde{r}_{i+1}$ can be drawn randomly. The second challenge concerns the proper choice of conditioning variables C_i, which is somewhat reminiscent of the choice of 'features' in a machine-learning context. The main focus of this paper is to investigate the effect of these choices, and we show some initial exploratory results in Section 4.

3.1 Building the distribution

At any given point in space and time during iteration of (8), given a (local) value of the conditioning variable C_i, we wish to select a corresponding subset of reference data from which we can sample randomly. This suggest some discretization approach where intervals of C_i are coupled to subsets of $\overline{r}$. We use the so-called 'binning' approach of [17], which starts with a snapshot sequence of the eddy forcing

$$\mathbf{R}_1^S = \{\overline{r}_1, \overline{r}_2, \cdots, \overline{r}_S\}, \tag{10}$$

where i is the time index, and each snapshot $\overline{r}_i$ is an $N \times N$ field, where N is typically 2^7, 2^8 or higher. For our initial calculations we used 2^7. In addition, we also have snapshots of corresponding conditioning variables

$$\mathbf{C}_1^S = \{C_1, C_2, \cdots, C_S\}. \tag{11}$$

Let C be the total number of time-lagged conditioning variables used in (9). We then proceed by creating C-dimensional disjoint bins [1], each bin spanning a unique conditioning variable range, and containing a number of associated *scalar* $\overline{r}(x, y)$ values, i.e. the mapping is done pointwise in spatial domain. For any single location (x_i, y_j) we allow for resampling of all observed $\overline{r}(x, y)$ values from the entire flow domain, provided that they fall in the bin selected by the local value of the conditioning variable at (x_i, y_j). Note that not all bins may contain samples, especially if two or more conditioning variables are used. If during prediction an empty bin is sampled, the data of the nearest bin (in Euclidean sense) is used instead. Once a bin is selected by C_i, the resulting subset of scalar $\overline{r}$ values can be sampled randomly, or one might sample from the local bin average instead, leading to a *deterministic* prediction.

3.2 Choice of conditioning variables

Ideally we would like the conditioning variables of (9) to correlate well with $\overline{r}_{i+1}$. If $\overline{r}_{i+1}$ correlates well with C_i, the range of plausible $\overline{r}$ values in the selected subset is smaller. Consider the two bins depicted in Figure 3, each with 1 conditioning variable $(\overline{r}_{i+1} \mid C_i)$. The binning object of Figure 3(a) shows considerable less correlation between C_i and r_{i+1} than its counterpart in Figure 3(b). As a result, each bin contains a larger spread in possible $\overline{r}$ values, leading to more noisy $\widetilde{r}_{i+1}$ fields.

One possible choice, given the definition of r in (7), is to use the resolved Jacobian $\mathcal{P}^{\mathcal{R}} J\left(\Psi^{\mathcal{R}}, \omega^{\mathcal{R}}\right)$ as conditioning variable. Furthermore, the authors of [10, 21] also constructed a conditional pdf of the eddy forcing for a quasi-geostrophic double-gyre ocean model. Using an argument of frame invariance, and constraining their choice to that of a divergence of a stress, they showed that

$$\nabla \cdot \nabla \left(\frac{\partial \omega^{\mathcal{R}}}{\partial t} + \mathcal{P}^{\mathcal{R}} J\left(\Psi^{\mathcal{R}}, \omega^{\mathcal{R}}\right)\right) = \nabla \cdot \nabla \frac{\mathrm{D}\omega^{\mathcal{R}}}{\mathrm{D}t} \tag{12}$$

displayed very good correlation with $\overline{r}$.

We will systematically investigate the performance of a large set of candidate conditioning variables. Let us define a set of operators $\mathcal{L}$ as

$$\mathcal{L} := \{1, \nabla, \nabla^2\}, \tag{13}$$

and the following set of variables $\phi^{(i)}$, all evaluated at time t_i, as

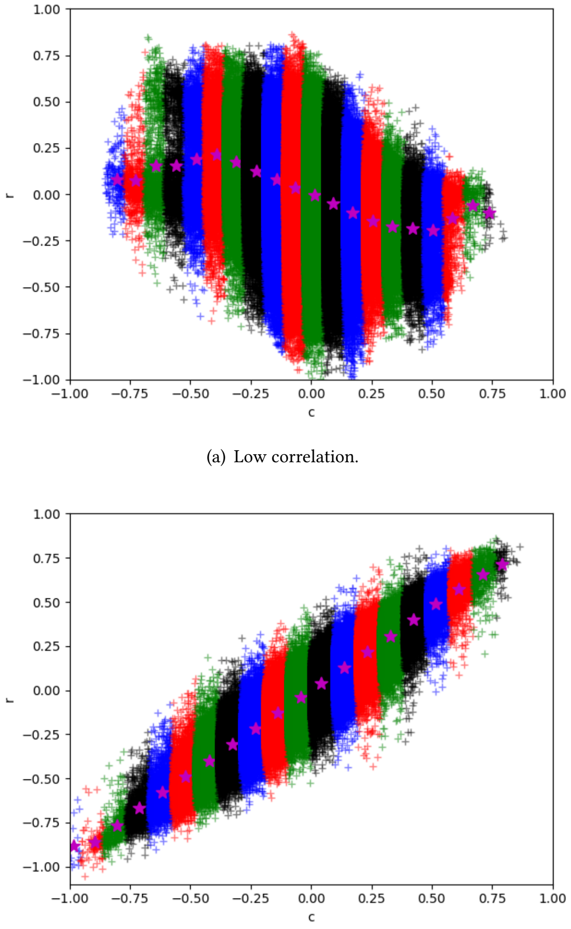
$$\phi^{(i)} := \{\pm\omega_i^{\mathcal{R}},\ \pm\mathcal{P}^{\mathcal{R}} J\left(\Psi_i^{\mathcal{R}}, \omega_i^{\mathcal{R}}\right),\ \pm\mathrm{D}\omega_i^{\mathcal{R}}/\mathrm{D}t\} \tag{14}$$

Let $\mathcal{L}_i$ and $\phi_j^{(i)}$ be members of the corresponding set. We allow for conditioning variables C_i, written as the sum of up to two $\mathcal{L}_i\phi_j^n$ terms, which gives us a total of 120 candidate conditioning variables. Clearly, not all conditioning variables will make sense and some just differ in sign, but (8) is computationally tractable enough for a brute-force computation using all possible combinations. Specifically, we will compute the 120 spatial correlation coefficients

$$\rho\left(\overline{r}_{i+1}, C_i^{(j)}\right) = \frac{Cov\left[\overline{r}_{i+1}, C_i^{(j)}\right]}{\sigma\left(\overline{r}_{i+1}\right)\sigma\left(C_i^{(j)}\right)}, \quad j = 1, \cdots, 120, \tag{15}$$

at a sampling rate of 1 day during a 250 day simulation period. Here, $Cov(\cdot, \cdot)$ and $\sigma(\cdot)$ denote the covariance and the standard deviation. The results are shown in Figure 4. Roughly speaking, we can distinguish three different groups of conditioning variables. The vast majority of C_i yield a correlation coefficient which does not exceed much beyond ± 0.25. The resolved Jacobian is amongst them, with $\rho \approx 0.14$. Taking the Laplacian of the Jacobian raises this value to

[1] We used equidistant bins, but this is not a hard requirement.



(a) Low correlation.

(b) Higher correlation.

Figure 3: The samples from 2 different binning objects with 1 conditioning variable. The vertical axis contains the $\overline{r}$ samples and the horizontal the conditioning variables. Stars denote bin means.

0.24. There are some that hover around the ±0.8 mark. A representative example of those is $\mathrm{D}\omega^{\mathcal{R}}/\mathrm{D}t$. Finally, a number of conditioning variables yield a correlation coefficient of approximately ±0.96. A common term in all of these high-correlation combinations of $\mathcal{L}_i\phi_j^n$ is in fact (12) from [10].

It is important to note that the correlations of Figure 4 were computed using the *exact* $\overline{r}$ in (8). When $\overline{r}$ will be replaced by the surrogate $\widetilde{r}$, these correlations cannot be expected to be maintained. Discrepancies between the exact eddy forcing and the surrogate will build up over time, and the model forced by $\widetilde{r}$ will develop its own dynamics. We reiterate here that our goal is to predict the time-averaged flow statistics, which might still be feasible if we are not in absolute lockstep with the fields of the full-scale equations. Even two full-scale simulations with slightly different initial conditions will diverge from each other (due to their turbulent/chaotic nature), yet can converge in a statistical sense[15].

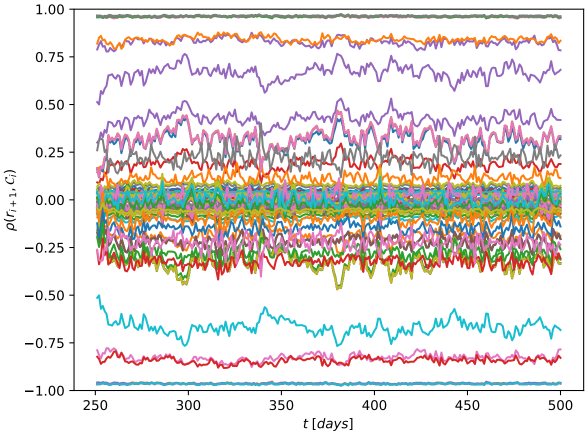


Figure 4: The spatial correlation coefficient (15) **between $\overline{r}_{i+1}$ and the 120 candidates conditioning variables C_i at 250 time instances. The exact $\overline{r}$ was used in** (8)**.**

3.3 Dynamic behaviour of the surrogate

Let us now examine the time evolution of $\widetilde{r}$ as a function of the chosen set of conditioning variables, *when* $\widetilde{r}$ *actually replaces* $\overline{r}$ *in* (8). We denote this as the predictive phase, as opposed to the training phase when (8) is forced by $\overline{r}$.

First, we create the auto-correlated surrogate $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \overline{r}_i$ with 100 bins, i.e. we used the exact eddy forcing at a previous time step as a conditioning variable. Clearly, this conditioning variable will not be available outside the training period, and so this surrogate merely serves as a sanity check. Given such a perfect C_i, we isolate the error introduced due to the random sampling alone. Over a simulation period of 250 days, we computed the $\omega^{\mathcal{R}}$ from the full-scale reference solution, as well as the $\omega^{\mathcal{R}}$ obtained with the surrogate eddy forcing. Figure 5 shows the final two snapshots of the aforementioned two $\omega^{\mathcal{R}}$ fields, which virtually look identical. Since we use the time-lagged exact eddy forcing as conditioning variable, we are never extrapolating the surrogate more than 1 time cycle, leading to the matching results of Figure 5. Hence, this is not a validation of the surrogate, and we view this instead as verification exercise of our implementation.

We also examined the dynamical behaviour of the $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \nabla\cdot\nabla\frac{\mathrm{D}\omega^{\mathcal{R}}}{\mathrm{D}t}$ surrogate. While $\nabla\cdot\nabla\frac{\mathrm{D}\omega^{\mathcal{R}}}{\mathrm{D}t}$ showed extremely good correlation with $\overline{r}_{i+1}$ in the training phase, when used as a conditioning variable in the predictive phase it becomes unstable in our problem. It grows in absolute magnitude, leading to excessive sampling of the two outer bins. This process is exacerbated over time, leading to an unstable solution.

Another sure way to construct a surrogate with a built-in high correlation between the conditioning variable and the reference $\overline{r}$ data is by means of auto correlation with respect to the surrogate,

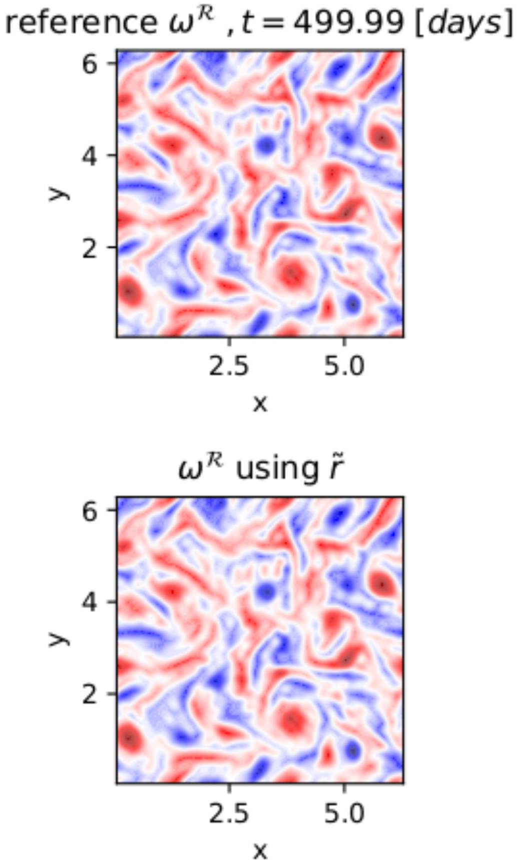


Figure 5: The final snapshot of the reference $\omega^{\mathcal{R}}$ (top) and the $\omega^{\mathcal{R}}$ field computed using the surrogate $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \overline{r}_i$.

i.e.

$$\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \widetilde{r}_i. \qquad (16)$$

For the surrogate construction we still use $\overline{r}_i$ from the training phase to construct the conditioning variable bins. Essentially, we use $\widetilde{r}_{i+1} \sim \overline{r}_{j+1} \mid \overline{r}_j$ with j such that $\overline{r}_j$ is close to (is in the same bin as) $\widetilde{r}_i$. Note that (16) leads to a one-way coupled system, without feedback from the resolved-scale equation (8) back to (16). Whilst this is in contradiction with the 4-th requirement outlined in Section 3, the one-way coupled attribute does allows us to exactly compute, per bin, what we denote as jump probabilities of surrogate (16). Any random sample $\widetilde{r}_{i+1}$ will become the conditioning variable in the next iteration, and thus we can directly determine if the same bin will be selected, or if a bin to the left or right will be sampled instead. Doing so for every sample in a given bin allows us to compute the so-called stay, jump-left and jump-right probabilities, which will add up to 1 per bin. These are the empirical jump probabilities of (16). The surrogate $\overline{r}_{i+1} \mid \overline{r}_i$ samples from the same distribution, which can be computed from all consecutive $\overline{r}_{i+1}, \overline{r}_i$ data snapshot pairs. This is shown in Figure 6, which depicts the jump probabilities computed independently per bin and from the snapshot sequence. However, the match does not necessarily mean that (16) has statistical power, which $\overline{r}_{i+1} \mid \overline{r}_i$ clearly does have (see Figure 5). As mentioned, the latter does not effectively extrapolate and the use of $\overline{r}_i$ as conditioning variable imposes a spatial structure on $\widetilde{r}_{i+1}$ that is very close to $\overline{r}_{i+1}$, which is something that (16) cannot guarantee.

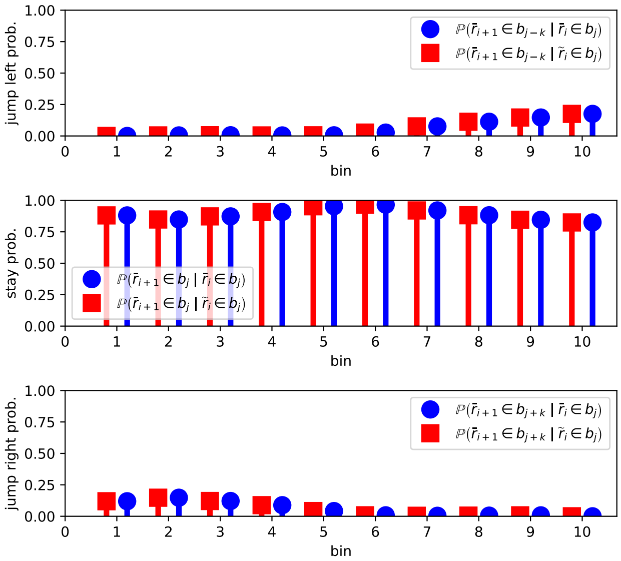


Figure 6: The jump left, stay and jump right probabilities (top, middle, bottom) for $\overline{r}_{i+1} \mid \overline{r}_i$ (circles) and $\overline{r}_{i+1} \mid \widetilde{r}_i$ (squares), using 10 bins and a time lag of 0.05 day. In the legend, k is an integer with $k \geq 1$.

Plotting the jump probabilities shows the dynamic character of the surrogate. The time lag employed in Figure 6 is small, namely 0.05 days. As a result, we see that the surrogate has a fairly 'static' character, i.e. that the stay probabilities are close to 1. Employing a larger time lag (see Figure 7), results in a surrogate with a seemingly more dynamic behaviour, where neighbouring bins are more likely to be sampled in the next iteration.

This makes clear sense, since a larger time lag implies a greater difference between consecutive snapshots. It does raise an important point however, i.e. that the time step embedded in the surrogate can differ from the Δt used to integrate (8) in time [2]. If the $\overline{r}$ data is sub sampled, the surrogate time step δt will be larger than Δt. When running a simulation, we therefore take care to update the surrogate every δt, rather than every Δt in order to ensure dynamical consistency. If multiple conditioning variables with different time lags are present, the value of δt equals the smallest employed time lag.

3.3.1 Data reduction. The reason for sub sampling the data is to reduce the memory requirements of the surrogate technique. If $\delta t \longrightarrow \Delta t$, a potentially very large number of data snapshots must be stored in memory during the entire run of the simulation [18]. Although not actually employed in this article, we briefly discuss another route of much more significant data reduction. Instead of saving N_i data points in any given bin indexed by i, we can also only

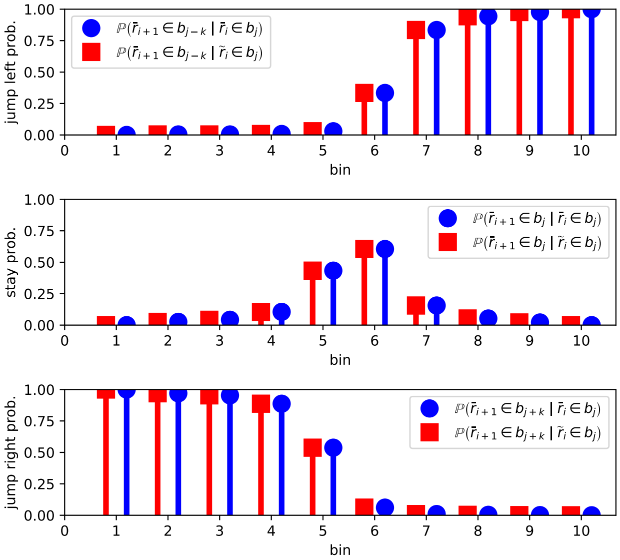


Figure 7: The jump left, stay and jump right probabilities (top, middle, bottom) for $\bar{r}_{i+1} \mid \bar{r}_i$ (circles) and $\bar{r}_{i+1} \mid \tilde{r}_i$ (squares), using 10 bins and a time lag of 1.0 **day.**

store certain bin means during the simulation run. Consider again the jump probabilities of Figures 6-7. In addition to the bin wise computation of the probabilities, we can also compute the mean of all $\bar{r}_{i+1}$ samples that either stay of leave the current bin in a certain direction. We could do this directly from the data, and the approach would not be limited to auto-correlated surrogates only. The local bin means are then sampled with their associated jump probability. As an example, consider the results of Figure 8 which shows the local bins means of the (16) surrogate. For this particular example, the number of data points which must be carried in memory during simulation is reduced from 83.2×10^6 to 28. This method requires further study, but the potential for data reduction is clear. A similar, yet not the same, method is proposed in [2].

4 INITIAL RESULTS

Here we outline some initial results which were obtained with the surrogate model described in the preceding section. To generate the reference database, we run the full-scale and the resolved model forced by $\bar{r}$ for a spin-up period of 250 days. Next, data is collected for another 250 days using a subsampling δt of 0.05 days. For these particular results, the full- and resolved scales were defined on a numerical grid with $2^7 \times 2^7$ and $2^6 \times 2^6$ points respectively.

4.1 Short-time prediction

Figure 9 shows a snapshot of the reference vorticity and a corresponding snapshot from a model forced by the surrogate (16). Clearly, compared to the results of Figure 5 (where we had a perfect conditioning variable), the performance is reduced. As discussed, the model forced by the surrogate develops its own dynamics and the $\omega^{\mathcal{R}}$ fields diverge from the reference solution.

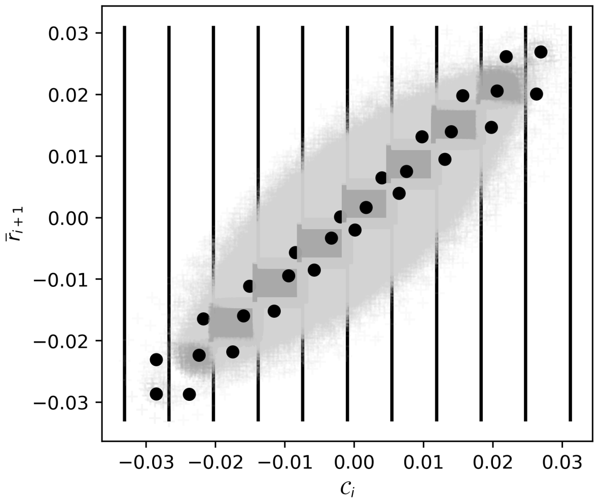


Figure 8: All samples of the autocorrelated surrogate (16) **(in grey) per bin, and the local bin means associated to the three jump probabilities (circles). Per bin, the top circles are the bin means of all samples which jumped right. The middle and bottom circles represent the stay and jump-left means, respectively. The reference data ($\bar{r}_{i+1}$) is plotted along the vertical axis, and the conditioning variable along the horizontal ($C_i = \tilde{r}_i$).**

When we continue to run the reference model in the background during the predictive phase, we can compute how fast a surrogate decorrelates from the exact eddy forcing. For the autocorrelated surrogate (16) with $\delta t = 0.05$ days, the spatial correlation coefficient $\rho\left(\bar{r}_{i+1}, \tilde{r}_{i+1}\right)$ is plotted versus time in Figure 10(a). As expected, we observe a decrease, and after about 15 days all correlation is lost. The opposite behaviour is observed when conditioning on the resolved Jacobian instead, see Figure 10(b). The correlation *increases* from 0.1 to roughly 0.45 in the predictive phase. Due to the definition of the eddy forcing (7) (which directly includes the resolved Jacobian), this increased correlation can be expected to be maintained. This is an indication (yet no guarantee), that $\mathcal{P}^{\mathcal{R}} J\left(\Psi^{\mathcal{R}}, \omega^{\mathcal{R}}\right)$ could be a good conditioning variable to include in the case of a statistical prediction.

4.2 Statistical prediction

From the initial condition at 250 days, the model forced by the surrogate is run for 8 years, during which data for the energy $E^{\mathcal{R}}$ and enstropy $Z^{\mathcal{R}}$ densities are collected. These quantities are defined as the integrated square velocity and vorticity, normalised by the flow domain area, i.e.

$$E^{\mathcal{R}} := \left(\frac{1}{2\pi}\right)^2 \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \mathbf{V}^{\mathcal{R}} \cdot \mathbf{V}^{\mathcal{R}} \mathrm{d}x\mathrm{d}y,$$

$$Z^{\mathcal{R}} := \left(\frac{1}{2\pi}\right)^2 \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \left(\omega^{\mathcal{R}}\right)^2 \mathrm{d}x\mathrm{d}y, \tag{17}$$

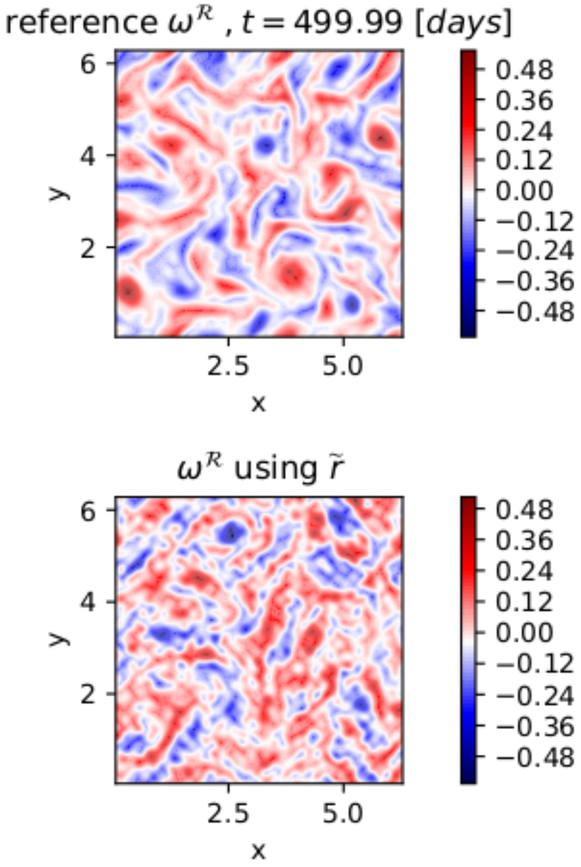


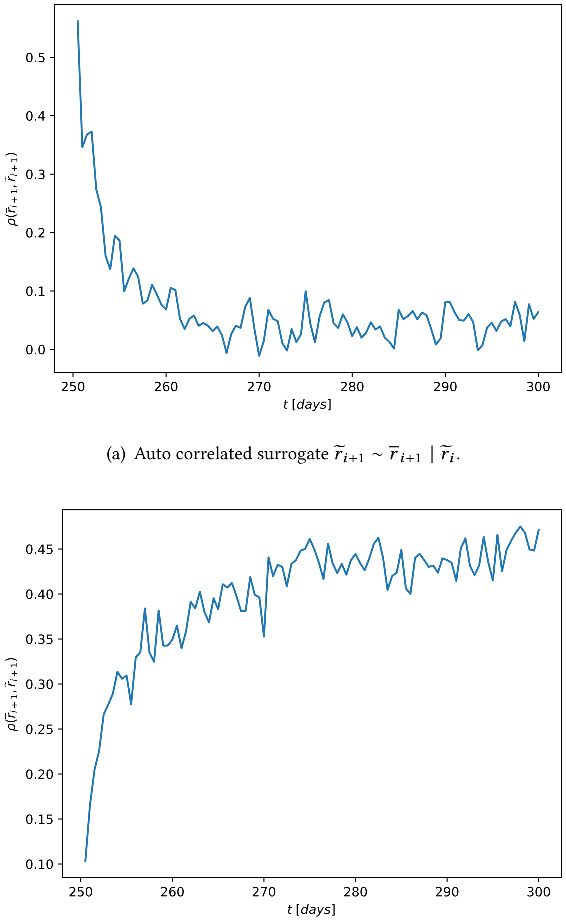
Figure 9: The final snapshot of the reference $\omega^{\mathcal{R}}$ (top) and the $\omega^{\mathcal{R}}$ field computed using the surrogate $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \widetilde{r}_i$. The discrepancy between the two snapshots is expected, as the reference is deterministic and the surrogate is stochastic.

where $\mathbf{V}^{\mathcal{R}}$ is the two-dimensional vector of the resolved velocity components in x and y direction, see also [19].

Let us start by showing the results of purely autocorrelated surrogates in Figure 11. We limit ourselves to one conditioning variable using 100 bins, and investigate the effect of the employed time lag, measured in multiples of δt. The reference energy and enstrophy pdfs, as well as the pdfs from the unparameterized model ($\overline{r} = 0$) are also depicted. The shortest time lag of 1 δt leads to an energy pdf which is too diffuse. This is improved by increasing the time lag, but in all cases the pdf of the enstrophy $Z^{\mathcal{R}}$ has a bias with respect to the reference pdf. Thus overall the performance is quite low, which might be expected since the surrogate acts independently from the resolved state.

To obtain two-way coupling between the surrogate and the resolved state we now condition on the resolved Jacobian. Two surrogates used one conditioning variable, with a time lag of δt and $20\delta t$. To investigate the performance of a surrogate with a larger memory, we also constructed a surrogate with two conditioning variables with a time lag of $(5\delta t, 10\delta t)$. In all cases the number of bins was kept constant at 100. The two single conditioning variable surrogates do not perform very well. Their pdfs of the energy density are worse than the unparameterized result, and the bias in $Z^{\mathcal{R}}$ is comparable to the autocorrelated results. In contrast, the use of two conditioning variables yields a $E^{\mathcal{R}}$ pdf which is close to the reference (blue curve). Furthermore, while it does not eliminate the enstropy bias, it does reduce it to the point where there is now more significant high-probability overlap between the model and reference pdf.

We note that these are preliminary results, which point in the direction of two-way coupled surrogates with larger (non-Markovian)



(a) Auto correlated surrogate $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \widetilde{r}_i$.

(b) Surrogate conditioned on the Jacobian $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \mathcal{P}^{\mathcal{R}} J\left(\Psi_i^{\mathcal{R}}, \omega_i^{\mathcal{R}}\right)$

Figure 10: The spatial correlation coefficient $\rho\left(\overline{r}_{i+1}, \widetilde{r}_{i+1}\right)$ between the exact eddy forcing and two surrogates, constructed with 100 bins and $\delta t = 0.05$ days. These results were computed during the predictive phase, the initial transient behaviour is an initial condition effect, where the state was initialised using the solution forced by $\overline{r}$.

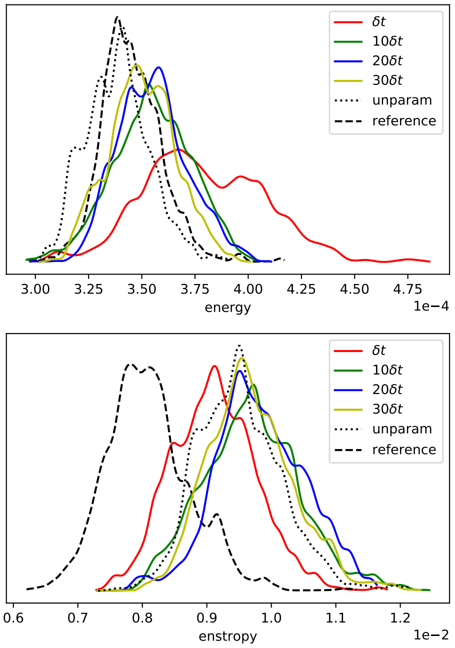


Figure 11: The pdfs of (17), i.e. of the energy (top) and enstrophy (bottom) density, computed over the time integration period of 8 years, using surrogates of the form $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid \widetilde{r}_i$. The surrogates differ via the employed time lag, which are δt, $10\delta t$, $20\delta t$ and $30\delta t$, with $\delta t = 0.05$ days.

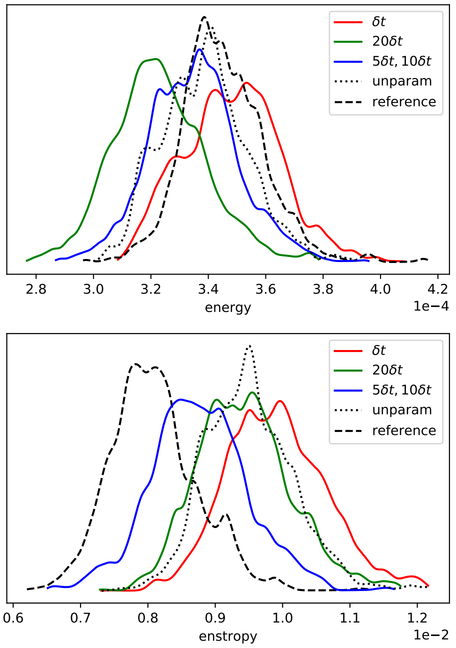


Figure 12: The pdfs of (17), i.e. of the energy (top) and enstrophy (bottom) density, computed over the time integration period of 8 years, using surrogates of the form $\widetilde{r}_{i+1} \sim \overline{r}_{i+1} \mid J\left(\Psi_i^{\mathcal{R}}, \omega_i^{\mathcal{R}}\right)$. The time lags of the conditioning variables are δt and $20\delta t$. For the surrogate with two resolved Jacobians as conditioning variables we used $(5\delta t, 10\delta t)$. Again, $\delta t = 0.05$ days.

memory. A more systematic study of different conditioning variable combinations is currently under way.

5 CONCLUSION & OUTLOOK

We presented a data-driven method to create a stochastic surrogate model, conditioned on time-lagged observable variables, of a set of reference data coming from a dynamical system. In the current application, we focused on recreating the forcing term due to unresolved eddies in a two-dimensional ocean model. This approach is general however, and not restricted to our particular application. We consider the method as validated when the resolved model forced by the stochastic surrogate accurately captures time-averaged statistics of the full-scale simulation, such as integrated energy densities. To that end, our initial results indicates that the choice of a proper set of conditioning variables is of paramount importance. Multiple conditioning variables at different time instances, such that a longer memory is built into the surrogate, outperform the considered surrogates with a Markovian character.

We are currently conducting a systematic study of a wide range of conditioning variables, to seek further gains in performance. Furthermore, as briefly outlined, data-reduction techniques to drastically decrease the computational memory requirements are also under investigation. Finally, an interesting avenue of future research is the *a-priori* incorporation of constraints from mathematical physics. Initial work along these lines showed promising results [4], with no enstrophy bias. Another option involves rewriting the eddy forcing in tensor format, such that certain constraints on the tensor shape can be found [20]. Such an approach would no longer be purely data-driven, and opens up the possibility for efficient, physics-constrained uncertainty quantification, see e.g. [5] for examples in steady flow problems or [9] for large-eddy simulations.

ACKNOWLEDGMENTS

This research is funded by the Netherlands Organization for Scientific Research (NWO) through the Vidi project "Stochastic models for unresolved scales in geophysical flows", and from the European Union Horizon 2020 research and innovation programme under grant agreement #800925 (VECMA project).

We also thank W.T.M. Verkley for making his vorticity equation source code available to us.

REFERENCES

[1] P.S. Berloff. 2005. Random-forcing model of the mesoscale oceanic eddies. *Journal of Fluid Mechanics* 529 (2005), 71–95.

[2] D. Crommelin and E. Vanden-Eijnden. 2008. Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences* 65, 8 (2008), 2661–2675.

[3] W.N. Edeling. 2019. vorticity-solver (GitHub repository). https://github.com/wedeling/vorticity-solver.

[4] W.N. Edeling and D.T. Crommelin. 2019. Reduced model-error source terms for fluid flow. In *UNCECOMP19 Conference proceedings (submitted).*

[5] W.N. Edeling, G. Iaccarino, and P. Cinnella. 2018. Data-free and data-driven rans predictions with quantified uncertainty. *Flow, Turbulence and Combustion* 100, 3 (2018), 593–616.

[6] M. Eldred and J. Burkardt. 2009. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. In *47th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition*. 976.

[7] P.R. Gent and J.C. Mcwilliams. 1990. Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography* 20, 1 (1990), 150–155.

[8] I. Grooms and L. Zanna. 2017. A note on 'Toward a stochastic parameterization of ocean mesoscale eddies'. *Ocean Modelling* 113 (2017), 30–33.

[9] L. Jofre, S.P. Domino, and G. Iaccarino. 2018. A framework for characterizing structural uncertainty in large-eddy simulation closures. *Flow, Turbulence and Combustion* 100, 2 (2018), 341–363.

[10] P. Mana and L. Zanna. 2014. Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling* 79 (2014), 1–20.

[11] R. Maulik, O. San, A. Rasheed, and P. Vedula. 2019. Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics* 858 (2019), 122–144.

[12] J.C. McWilliams. 1984. The emergence of isolated coherent vortices in turbulent flow. *Journal of Fluid Mechanics* 146 (1984), 21–43.

[13] T. Palmer and P. Williams. 2010. *Stochastic physics and climate modelling*. Cambridge University Press Cambridge, UK.

[14] R. Peyret. 2013. *Spectral methods for incompressible viscous flow*. Vol. 148. Springer Science & Business Media.

[15] J. Slingo and T. Palmer. 2011. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369, 1956 (2011), 4751–4767.

[16] J. Thuburn, J. Kent, and N. Wood. 2014. Cascades, backscatter and conservation in numerical models of two-dimensional turbulence. *Quarterly Journal of the Royal Meteorological Society* 679, 140 (2014), 626–638.

[17] N. Verheul and D. Crommelin. 2016. Data-driven stochastic representations of unresolved features in multiscale models. *Commun. Math. Sci* 14, 5 (2016), 1213–1236.

[18] N. Verheul, J. Viebahn, and D. Crommelin. 2017. Covariate-based stochastic parameterization of baroclinic ocean eddies. *Mathematics of Climate and Weather Forecasting* 3, 1 (2017), 90–117.

[19] W.T.M. Verkley, P.C. Kalverla, and C.A. Severijns. 2016. A maximum entropy approach to the parametrization of subgrid processes in two-dimensional flow. *Quarterly Journal of the Royal Meteorological Society* 142, 699 (2016), 2273–2283.

[20] S. Waterman and J.M. Lilly. 2015. Geometric decomposition of eddy feedbacks in barotropic systems. *Journal of Physical Oceanography* 45, 4 (2015), 1009–1024.

[21] L. Zanna, P. Mana, J. Anstey, T. David, and T. Bolton. 2017. Scale-aware deterministic and stochastic parametrizations of eddy-mean flow interaction. *Ocean Modelling* 111 (2017), 66–80.