Symbiotic Ocean Modeling using Physics-Controlled Echo State Networks

Thomas Erik Mulder[1], Sven Baars[2], Fred W. Wubs[2], Inti Pelupessy[3], Merijn Verstraaten[3], and Henk A. Dijkstra[4]

[1]Swedish Meteorological and Hydrological Institute
[2]University of Groningen
[3]Netherlands eScience Center
[4]Institute for Marine and Atmospheric research Utrecht

March 23, 2023

Symbiotic Ocean Modeling using Physics-Controlled Echo State Networks

T.E. Mulder[1,2], S. Baars[1], F.W. Wubs[1], F.I. Pelupessy[3], M. Verstraaten[3], and H.A. Dijkstra[4,5]

[1]Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, the Netherlands.
[2]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden.
[3]Netherlands eScience Center, Amsterdam, the Netherlands.
[4]Institute for Marine and Atmospheric research Utrecht, Department of Physics, Utrecht University, Utrecht, the Netherlands.
[5]Center for Complex Systems Studies, Utrecht University, Utrecht, the Netherlands.

Key Points:

- We demonstrate part of a symbiotic ocean modeling framework where models of different complexities benefit from each other.
- Unresolved processes are represented through hybrid machine learning methods using data from the symbiotic framework.
- Hybrid correction strategies with imperfect physics as control input improve the representation of key long-term flow properties.

Corresponding author: T.E. Mulder, erik.mulder@smhi.se

Abstract

We introduce the concept of 'symbiotic' ocean modeling where high- and low-resolution dynamical models coexist and benefit from each other through data-driven improvements. In this work we specifically focus on how a low-resolution model may benefit from such a symbiotic setup. The broader aim is to improve the efficiency of high-resolution models, while simultaneously enhancing the representation of unresolved processes in low-resolution models. To achieve a symbiosis we use a grid-switching approach together with hybrid modeling techniques that combine linear regression-based methods with nonlinear echo state networks (ESNs). The approach is applied to both the Kuramoto–Sivashinsky equation and a single-layer quasi-geostrophic ocean model, and shown to simulate short-term and long-term behavior better than either purely data-based methods or low-resolution models.

Plain Language Summary

Models of the ocean vary in complexity. Some are very detailed and manage to show oceanic vortices, whereas others are very efficient but coarse, and unable to compute such vortices. The idea in this paper is to let these different model types work together, as if in a symbiosis. With knowledge of differences between the detailed and coarse model we can use machine learning techniques to improve the coarse model, while a coarse model can be used to aid a detailed model computationally. Here we focus on the former part and perform numerous experiments to test different kinds of coarse model improvements. We apply our ideas to the Kuramoto–Sivashinsky (KS) model and a quasi-geostrophic (QG) ocean model, where we show that promising short-term KS results may generalize to models of the ocean. Long-term equilibrium experiments with QG show in addition how the correction strategies let a coarse model produce correct flow properties, where standalone physics- or data-based approaches fail.

1 Introduction

One of the most important spatial scales in the ocean circulation is the internal Rossby radius of deformation L_D; it ranges from 50-100 km at midlatitudes to a few km in the polar regions (Hallberg, 2013). At this scale, perturbations are amplified on mean flows through mixed barotropic/baroclinic instability, giving rise to ocean eddies. Interactions between these eddies and the mean flow can lead to upgradient momentum transport affecting the strength and separation of ocean western boundary currents such as the Kuroshio and Agulhas (Chassignet et al., 2020).

Most climate models, in particularly those used in CMIP5 and CMIP6, do not resolve ocean processes at the scale L_D as the spatial grid size used is too large, e.g. typically 1° (Eyring et al., 2016). The main reason is computational costs, as doubling the horizontal resolution increases these costs roughly by a factor 10. Effects of subgrid-scale processes are hence parameterized in these models. For example, the effect of ocean eddies on tracer transport is represented by the Gent–McWilliams (Gent et al., 1995) scheme, but such a scheme cannot capture, for example, the upgradient momentum transport. Hence, western boundary flows are too weak and diffuse, and do not separate at the correct location (Chassignet et al., 2020).

Over the last few years, first simulations have been performed with global climate models, where the ocean model component has a resolution of 0.1°, which is smaller than L_D for many locations on the globe (Chang et al., 2020; Jüling et al., 2021). We will refer to those models as high-resolution (HR) models to contrast them with the 1° models which we will call low-resolution (LR) models. But also the high-resolution models are not completely eddy-resolving as this requires an even higher spatial resolution. There is now a substantial amount of model data available to compare results on ocean-climate variability and climate change for both types of models. Clearly, high-resolution models reduce biases compared to observations particularly in western boundary currents, sea surface temperature variability patterns and Southern Ocean mean flows (Chang et al., 2020; Jüling et al., 2021).

However, HR model simulations form a great drain on computational resources and hence there are still many efforts to represent the effects of unresolved processes in LR models. This parameterization process has been around for decades and approaches can be grouped into three types. First, semi-empirical parameterizations are used, where observation motivated schemes are implemented (Gargett, 1989; Viebahn et al., 2019). Second, theoretically derived schemes, where specific approximations are made in the underlying equations (Gent et al., 1995) have been used. Third, stochastic schemes derived from sample high-resolution model simulations (Berloff, 2005; Mana & Zanna, 2014) have shown potential in representing unresolved processes in LR models (Hewitt et al., 2020).

To this, recently a new approach has been added, where the subgrid-scale model is derived from a machine learning (ML) model, such as a neural network. In Bolton & Zanna (2019), a convolutional neural network (CNN) was trained with data from a high-resolution model of the midlatitude gyres. This CNN was shown to successfully capture the small-scale processes and the effects of those on the mean flow in the low-resolution version of the same model. Traditional feedforward neural network models (FFNN) have also been used as subgrid-scale representations in both ocean and atmospheric models

(Irrgang et al., 2021; Rasp et al., 2018). Another ML technique that shows promise in the modeling of climate physics is the reservoir computing approach, often referred to as an echo state network (ESN). An ESN is a type of recurrent neural network (RNN) that is especially suited to simulate chaotic dynamics (Jaeger & Haas, 2004; Pathak et al., 2017) and is shown to be capable of emulating EOF interactions (Nadiga, 2021). Where FFNNs are generally regarded as functions, RNNs can be seen as artificial dynamical systems (Lukoševičius & Jaeger, 2009). Computationally there are close relations between ESN-based methods, linear regression and models based on a dynamic mode decomposition (DMD) (Schmid, 2010; Kutz et al., 2016). Theoretical connections between the ESN approach, DMD and also vector autoregression (VAR) have been explored in Bollt (2021).

Recent 'hybrid' (or physics-controlled) ESN advances (Pathak et al., 2017, 2018) provide an elegant approach to correct known model imperfections, such as those due to the lack of eddies in LR ocean models. With training data based on ground truths and imperfect model predictions, model tendencies and nonlinear model mismatches are encoded in an ESN. The result is an artificial dynamical system that can be controlled using an imperfect model. Combining an imperfect model with corrections from a trained ESN creates a hybrid dynamical system that greatly outperforms both the network and the imperfect model (Wikner et al., 2020). This approach was recently applied to an atmospheric model (SPEEDY) and shown to be able to improve the simulations of mean flow and variability considerably on short time scales (Arcomano et al., 2022).

In this paper we use the hybrid modeling framework as key ingredient for a 'symbiotic' ocean modeling approach. The idea is to couple models of different complexities and configure them to solve the same problem, where we distinguish between perfect and imperfect models in terms of differing resolution and parameterizations. This model coexistence can be made mutually beneficial using data-driven techniques. With the symbiotic approach we aim to improve the computational efficiency of HR models, while simultaneously enhancing the parameterizations of unresolved processes in LR models. We will focus on the latter part and use the hybrid modeling strategy with data generated from both LR and HR models to correct imperfect model transients. Model corrections made can then be seen as modeling subgrid effects. To this end, we employ a grid-switching approach and introduce a correction framework that includes models based on linear regression, DMD, ESN and hybrid variants (Section 2). We apply the correction strategy to coupled LR and HR versions of the Kuramoto–Sivashinsky equations (Section 3) and a single-layer quasi-geostrophic ocean model (Section 4). Both short-term predictions and long-term equilibrium runs are performed with the ocean model to compare the available corrective models. A summary and discussion with the main conclusions is provided in Section 5.

2 Methodology

In a general framework, the HR model is defined on a fine grid Ω^f and is regarded as a *perfect* model. An LR model is considered as an *imperfect* model, and is defined on a coarse grid Ω^c. The grids Ω^f and Ω^c have dimensions N_f and N_c, respectively, and cover the same domain. Both models attempt to solve the same problem, but apart from different grids we also allow differences in key parameters and forcings between the perfect and imperfect model. The physics resolved by the perfect model is then used as ground truth and the imperfect model results are considered to be in need of correction.

The perfect model is a system of coupled partial differential equations (PDEs), spatially discretized on Ω^f, which leads to a large system of differential-algebraic equations (DAEs):

$$M_P\dot{\boldsymbol{\xi}} = F_P(\boldsymbol{\xi}), \quad \text{with } \boldsymbol{\xi} \in \mathbb{R}^{N_f}. \tag{1}$$

Here, $\boldsymbol{\xi} = \boldsymbol{\xi}(t)$ is a time dependent state vector and $M_P \in \mathbb{R}^{N_f \times N_f}$ is a mass matrix that determines the dependence on temporal derivatives. The nonlinear operator $F_P : \mathbb{R}^{N_f} \to \mathbb{R}^{N_f}$ is a spatial discretization of the perfect model physics. Similarly, the semi-discretized imperfect model has a coarse state $\mathbf{x} = \mathbf{x}(t)$ that evolves according to

$$M_I\dot{\mathbf{x}} = F_I(\mathbf{x}), \quad \text{with } \mathbf{x} \in \mathbb{R}^{N_c}, \tag{2}$$

where $M_I \in \mathbb{R}^{N_c \times N_c}$ and $F_I : \mathbb{R}^{N_c} \to \mathbb{R}^{N_c}$ are again the mass matrix and spatial discretization operator. For simplicity we only consider models in this form (equations (1) and (2)), but the methodology explained here is not restricted to this formulation.

Transfers between the solutions on the two grids Ω^f and Ω^c are made through a fully weighted restriction $R \in \mathbb{R}^{N_c \times N_f}$ and a prolongation operator $P \in \mathbb{R}^{N_f \times N_c}$. We choose these operators for their convenient (variational) property that they are each other's transpose up to a constant factor: $R = cP^\top$ (Briggs et al., 2000). The perfect model evolves according to $\phi_P : \mathbb{R} \times \mathbb{R}^{N_f} \to \mathbb{R}^{N_f}$. Similarly, the evolution of the imperfect model is given by $\phi_I : \mathbb{R} \times \mathbb{R}^{N_c} \to \mathbb{R}^{N_c}$. The evolution operator ϕ_I solves for the transient state $\widetilde{\mathbf{x}}^{k+1}$ according to a certain time-discretization and hence $\widetilde{\mathbf{x}}^{k+1} = \phi_I(\mathbf{x}^k)$.

The imperfect spatial discretization F_I is incapable of capturing the physics resolved by the perfect model and we therefore attempt to improve the imperfect evolution ϕ_I with a combination of linear and non-linear corrections. We employ an auxiliary (surrogate) model f with auxiliary state $\mathbf{s} \in \mathbb{R}^{N_r}$ of size N_r, that is forced by imperfect and improved predictions. As these corrections are data-driven we divide our approach into a data gathering and a prediction phase.

2.1 Data gathering

We gather data from a trajectory of $\boldsymbol{\xi}(t)$ on Ω^f. From this transient, associated restricted states, imperfect predictions and auxiliary states are computed. Starting at time t_0, we collect N_T+1 snapshots of the evolving state $\boldsymbol{\xi}(t)$:

$$\left\{\boldsymbol{\xi}^0,\ \boldsymbol{\xi}^1,\ \ldots,\ \boldsymbol{\xi}^{N_T}\right\},\quad \boldsymbol{\xi}^k=\boldsymbol{\xi}(t_0+k\Delta t), \tag{3}$$

at fixed time intervals Δt such that we cover the model time $T = N_T\Delta t$. The snapshots are restricted to the coarse grid and combined into two data matrices:

$$\begin{aligned} X &= \left[\mathbf{x}^0,\ \mathbf{x}^1,\ \ldots,\ \mathbf{x}^{N_T-1}\right]=\left[R\boldsymbol{\xi}^0,\ R\boldsymbol{\xi}^1,\ \ldots,\ R\boldsymbol{\xi}^{N_T-1}\right], &(4)\\ X' &= \left[\mathbf{x}^1,\ \mathbf{x}^2,\ \ldots,\ \mathbf{x}^{N_T}\right]=\left[R\boldsymbol{\xi}^1,\ R\boldsymbol{\xi}^2,\ \ldots,\ R\boldsymbol{\xi}^{N_T}\right]. &(5)\end{aligned}$$

Apart from the restricted data matrix $X\in\mathbb{R}^{N_c\times N_T}$ and its shifted version $X'\in\mathbb{R}^{N_c\times N_T}$, we also create a collection of imperfect predictions $\Phi(X)$:

$$\Phi(X)=\left[\phi_I\left(\mathbf{x}^0\right),\ \phi_I\left(\mathbf{x}^1\right),\ \ldots,\ \phi_I\left(\mathbf{x}^{N_T-1}\right)\right]\in\mathbb{R}^{N_c\times N_T}. \tag{6}$$

The elements of X and $\Phi(X)$ serve as forcing to the auxiliary model f, which we evolve and gather snapshots from. We iterate according to

$$\begin{aligned} \mathbf{u}^k &= h\left(\mathbf{x}^k,\phi_I\left(\mathbf{x}^k\right)\right), &(7)\\ \mathbf{s}^{k+1} &= f\left(\mathbf{s}^k,\mathbf{u}^k\right), &(8)\end{aligned}$$

with combined input $\mathbf{u}$ given by a mapping h and with initialization $\mathbf{s}^0=\mathbf{s}_0$ at $t=t_0$. For h we either use a selection, e.g. $h\left(\mathbf{x}^k,\widetilde{\mathbf{x}}^{k+1}\right)=\mathbf{x}^k$ or combine the forcing such that $h\left(\mathbf{x}^k,\widetilde{\mathbf{x}}^{k+1}\right)=\left(\mathbf{x}^k;\widetilde{\mathbf{x}}^{k+1}\right)$, where (;) denotes vertical stacking. These are the most straightforward choices and of course other options are possible here. The surrogate model f comes in the form of an Echo State Network (ESN) and is described in Section 2.3. From the evolution of f we gather N_T+1 state snapshots $\mathbf{s}^0,\ldots,\mathbf{s}^{N_T}$ and combine them into a data matrix, with the exception of the initialization $\mathbf{s}^0$:

$$S=\left[\mathbf{s}^1,\ \mathbf{s}^2,\ldots,\ \mathbf{s}^{N_T}\right]\in\mathbb{R}^{N_r\times N_T}. \tag{9}$$

2.2 Prediction

The data gathered up until time $t=t_0+T$ is used to obtain linear best fit operators. Given data X, X', $\Phi(X)$ and S, these operators optimally combine $\mathbf{x}$, $\phi_I(\mathbf{x})$ and

$\mathbf{s}$ to improve the imperfect evolution given by ϕ_I alone. Here we provide a general transient strategy that covers a number of different corrective methods.

A corrected imperfect transient is started at $t_0 + T$. Now, the models ϕ_I and f operate in isolation from any perfect model data and f augments ϕ_I. Using starting states $\mathbf{x}^{N_T}$ and $\mathbf{s}^{N_T}$, the transient proceeds as follows:

$$\widetilde{\mathbf{x}}^{k+1} = \phi_I\left(\mathbf{x}^k\right) \qquad \text{create an imperfect model prediction,} \tag{10}$$

$$\mathbf{u}^k = h\left(\mathbf{x}^k, \widetilde{\mathbf{x}}^{k+1}\right) \qquad \text{construct a forcing,} \tag{11}$$

$$\mathbf{s}^{k+1} = f\left(\mathbf{s}^k, \mathbf{u}^k\right) \qquad \text{evolve the auxiliary state,} \tag{12}$$

$$\mathbf{x}^{k+1} = A\mathbf{x}^k + B\widetilde{\mathbf{x}}^{k+1} + C\mathbf{s}^{k+1} \qquad \text{create an improved prediction,} \tag{13}$$

for $k = N_T, N_T+1, \ldots$. Hence the trajectory of $\mathbf{x}$ is initialized with a restricted truth ($\mathbf{x}^{N_T} = R\boldsymbol{\xi}^{N_T}$) but continues independently of the perfect model ($\mathbf{x}^{N_T+1} \neq R\boldsymbol{\xi}^{N_T+1}$).

With the general formulation in (10)-(13) we aim to include several methods and their combinations in the same framework. The operators A, B, C have separate interpretations. On its own, A is obtained as a linear best fit of the propagation from X to $X^{'}$. Its eigendecomposition is known as a dynamic mode decomposition (DMD) (Schmid, 2010; Kutz et al., 2016) and A is often called a DMD-operator. The matrix B is the best direct correction of $\Phi(X)$ to $X^{'}$ in the least squares sense. Lastly, as f is a neural net, the operator C is the optimal output layer, i.e., the linear best fit translation of S to $X^{'}$. Hence these different methods can be seen as special cases in (10)-(13).

Combinations of the operators A, B and C are fitted at $t = t_0 + T$ using regularized linear regressions with the data matrices X, $X^{'}$, $\Phi(X)$ and S. Choices for the architecture of f and h and the use of operators A, B, C lead to a variety of predictive methods (Table 1). A *model only* approach uses $B = I$ and ignores A and C. The transient (10)-(13) is reduced to only the imperfect model evolution. In an *ESN* prediction we translate from states of the neural net (ESN) to predictions using a best fit C. Here, f is forced with restricted states only: $\mathbf{x}^k$. A *DMD* prediction is based on the best linear approximation of the propagation from X to $X^{'}$. When the operators B and C are combined and $\mathbf{u}^k = \left(\mathbf{x}^k, \widetilde{\mathbf{x}}^{k+1}\right)$, the auxiliary model f is subjected to a physics-based control $\phi_I\left(\mathbf{x}\right)$, both internally through $\mathbf{u}$ and externally through B. With f an ESN this is referred to as *ESNc*, which is equivalent to the hybrid scheme in Pathak et al. (2018). *DMDc* denotes DMD with control (Proctor et al., 2016) and is obtained by combining operators A and B. In DMDc the imperfect physics assist the DMD model which, on its own, generalizes poorly outside the training data. A basic *correction-only* approach follows from using only B, whereas additional combinations lead to the varieties *ESN+DMD* and *ESN+DMDc*.

Table 1: Overview of corrective methods based on operator configurations in (13). The associated minimizations are linear regression problems for which we do not include the regularization here; $|| \, ||_F$ is the Frobenius norm. Additional variations on these methods rely on the specific architecture chosen for h and f. The choices we make for h are added as a separate column to this table.

Method	$h\left(\mathbf{x}^k, \widetilde{\mathbf{x}}^{k+1}\right)$	Operator choices	Minimization to compute operators
Model only		$A=0,\ B=I,\ C=0$	No minimization necessary
ESN	$\mathbf{x}^k$	$A=0,\ B=0,\ C=?$	$\min\limits_{C} \left\Vert CS - X^{'} \right\Vert_F$
DMD		$A=?,\ B=0,\ C=0$	$\min\limits_{A} \left\Vert AX - X^{'} \right\Vert_F$
ESNc	$\begin{bmatrix} \mathbf{x}^k \\ \widetilde{\mathbf{x}}^{k+1} \end{bmatrix}$	$A=0,\ B=?,\ C=?$	$\min\limits_{[B\ C]} \left\Vert \begin{bmatrix} B & C \end{bmatrix} \begin{bmatrix} \Phi(X) \\ S \end{bmatrix} - X^{'} \right\Vert_F$
DMDc		$A=?,\ B=?,\ C=0$	$\min\limits_{[A\ B]} \left\Vert \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X \\ \Phi(X) \end{bmatrix} - X^{'} \right\Vert_F$
Correction-only		$A=0,\ B=?,\ C=0$	$\min\limits_{B} \left\Vert B\Phi(X) - X^{'} \right\Vert_F$
ESN+DMD	$\mathbf{x}^k$	$A=?,\ B=0,\ C=?$	$\min\limits_{[A\ C]} \left\Vert \begin{bmatrix} A & C \end{bmatrix} \begin{bmatrix} X \\ S \end{bmatrix} - X^{'} \right\Vert_F$
ESN+DMDc	$\begin{bmatrix} \mathbf{x}^k \\ \widetilde{\mathbf{x}}^{k+1} \end{bmatrix}$	$A=?,\ B=?,\ C=?$	$\min\limits_{[A\ B\ C]} \left\Vert \begin{bmatrix} A & B & C \end{bmatrix} \begin{bmatrix} X \\ \Phi(X) \\ S \end{bmatrix} - X^{'} \right\Vert_F$

Connections between ESNs and DMD exist (Bollt, 2021) and within this framework it is straightforward to combine (and consequently isolate) both approaches.

The minimizations shown in Table 1 are computed using Tikhonov regularization, which introduces an additional penalty on the size of the fitted operator. Regularization is crucial as it reduces overfitting and improves the stability of a long-term transient (Lukosevicius, 2012). For instance, the DMD-operator actually minimizes

$$\min_{A} \left(\left\Vert AX - X^{'} \right\Vert_F + \lambda^2 \left\Vert A \right\Vert_F \right), \tag{14}$$

with $\lambda > 0$ a regularization parameter.

2.3 Echo State Network

An echo state network (Jaeger, 2001; Jaeger & Haas, 2004) will act as the auxiliary predictive model f. Here we will roughly outline the organization of an ESN. For a detailed explanation we refer to Pathak et al. (2018), which we follow closely. An ESN is a recurrent neural network that can be viewed as an artificial nonlinear dynamical system with a state $\mathbf{s} \in \mathbb{R}^{N_r}$ of sufficient dimension N_r. The components of $\mathbf{s}$ interact through a sparse, random linear operator $W : \mathbb{R}^{N_r} \to \mathbb{R}^{N_r}$ that is not altered after initialization. The average degree of the adjacency graph associated with W is denoted with $\overline{d}$. Input data $\mathbf{u} \in \mathbb{R}^{N_u}$ is standardized (every unknown has zero mean and unit variance) and is fed as forcing to the system, where it is combined with the state using a fixed linear operator $W_{in} : \mathbb{R}^{N_u} \to \mathbb{R}^{N_r}$. The input operator W_{in} is random and sparse, with only a single element per row that is drawn from a uniform distribution on $[-1, 1]$. The internal state evolves according to

$$\mathbf{s}^{k+1} = f(\mathbf{s}^k, \mathbf{u}^k) = (1 - \alpha)\mathbf{s}^k + \alpha \tanh\left(W\mathbf{s}^k + W_{in}\mathbf{u}^k\right), \quad \mathbf{s}^0 = \mathbf{s}_0 \tag{15}$$

with initialization $\mathbf{s}_0$ and a relaxation parameter $\alpha \in (0, 1]$ (also known as the leaking rate) that controls the 'speed' of the artifical dynamics (Lukoševičius & Jaeger, 2009). Hence the state $\mathbf{s}$ evolves according to a deterministic iteration with internal interactions given by a random (but fixed) W and forcing provided by the input data. The $\tanh(\cdot)$ activation function introduces a nonlinearity that is controlled by the weights in W_{in}. The spectral radius $\rho(W)$ determines the damping (or memory) of the system. From (15) it is apparent that α allows a matching of time-scales between the network and the variability in the training data, which is beneficial to the network's predictive performance (Lukoševičius & Jaeger, 2009). The addition of the relaxation parameter α is the only significant difference between our formulation of f and that in Pathak et al. (2018).

Starting at $t = t_0$ with $\mathbf{s}_0$, the recursion (15) generates N_T new states that are combined into a data matrix S, as described in Section 2.1. A linear operator C provides output predictions by translating the auxiliary state to a prediction. In the standard ESN approach the output operator $C : \mathbb{R}^{N_r} \to \mathbb{R}^{N_c}$ is computed from a regularized minimization problem using $S^\star$ and X', see Section 2.2. Here $S^\star$ is an adapted version of S. As in Pathak et al. (2018), we take the square of the even elements in each state $\mathbf{s}^k \in S$. The motivation for this is largely empirical but related to problems that may originate with capturing symmetry in the model equations (Lu et al., 2017).

3 Results: Kuramoto–Sivashinsky model

In Pathak et al. (2018) a hybrid ESN was applied to the Kuramoto–Sivashinsky (KS) equation. Here we will begin with a test of our framework by replicating these results. We will first consider equal grids ($N_f = N_c$) and introduce an imperfection through a perturbation in one of the parameters. Later we explore a perfect/imperfect model setup with $N_f = 2N_c$ and no parameter perturbation in the KS-equation.

The KS-equation is capable of displaying rich spatiotemporal dynamics and is used to study a variety of phenomena such as flame front dynamics (Sivashinsky, 1977) and reaction-diffusion dynamics (Kuramoto, 1984). In one dimension it is given by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + (1+\epsilon)\frac{\partial^2 u}{\partial x^2} + \frac{\partial^4 u}{\partial x^4} = 0, \tag{16}$$

with $u \in [0, L]$, initial value $u(x, 0) = u_0(x)$ and periodic boundaries $u(x, t) = u(x + L, t)$. The domain size L is also the bifurcation parameter of the problem. In Pathak et al. (2018), the domain size is chosen at $L = 35$, for which the KS-equation has a positive maximum Lyapunov exponent λ_{max} and produces chaotic behavior (Hyman & Nicolaenko, 1986). A perturbation $\epsilon \geq 0$ is introduced to create an imperfection. With $\epsilon = 0$ we obtain the true, 'perfect' evolution whereas our 'imperfect' model will have $\epsilon > 0$.

The KS-equation is discretized on an equidistant grid: $x_i = i/N_f$ with $i = 1, 2, \ldots, N_f = N_c = 64$. We use a fully-implicit time stepping scheme with $\Delta t = 0.25$ and initialize with

$$u_0(x_i) = \begin{cases} 1, & i = 1, \\ 0, & i > 1. \end{cases}$$

Starting at $t = t_0$, a transient is computed up to $T = 6000$ from which we select a large number of training and testing intervals. In the remaining experiments we also use long transients to sample training periods from. This approach is efficient from a data-management perspective but does not guarantee uncorrelated data.

The ESN used closely follows that in Pathak et al. (2018). The spectral radius is set at $\rho(W) = 0.4$, the average degree is $\overline{d} = 3$, we use training intervals of size $T = 5000$ and ignore any relaxation with $\alpha = 1$. The KS-equation and its discretization are also equivalent to Pathak et al. (2018) so, for a coherent interpretation of the predictions, we scale the obtained timings with the same Lyapunov exponent $\lambda_{max} = 0.07$.

The methods summarized in Table 1 are compared in a scaling experiment where the auxiliary state size N_r is doubled several times (see Fig. 1). Only those methods based on an ESN depend on this parameter which leads to constant results for the other pre-

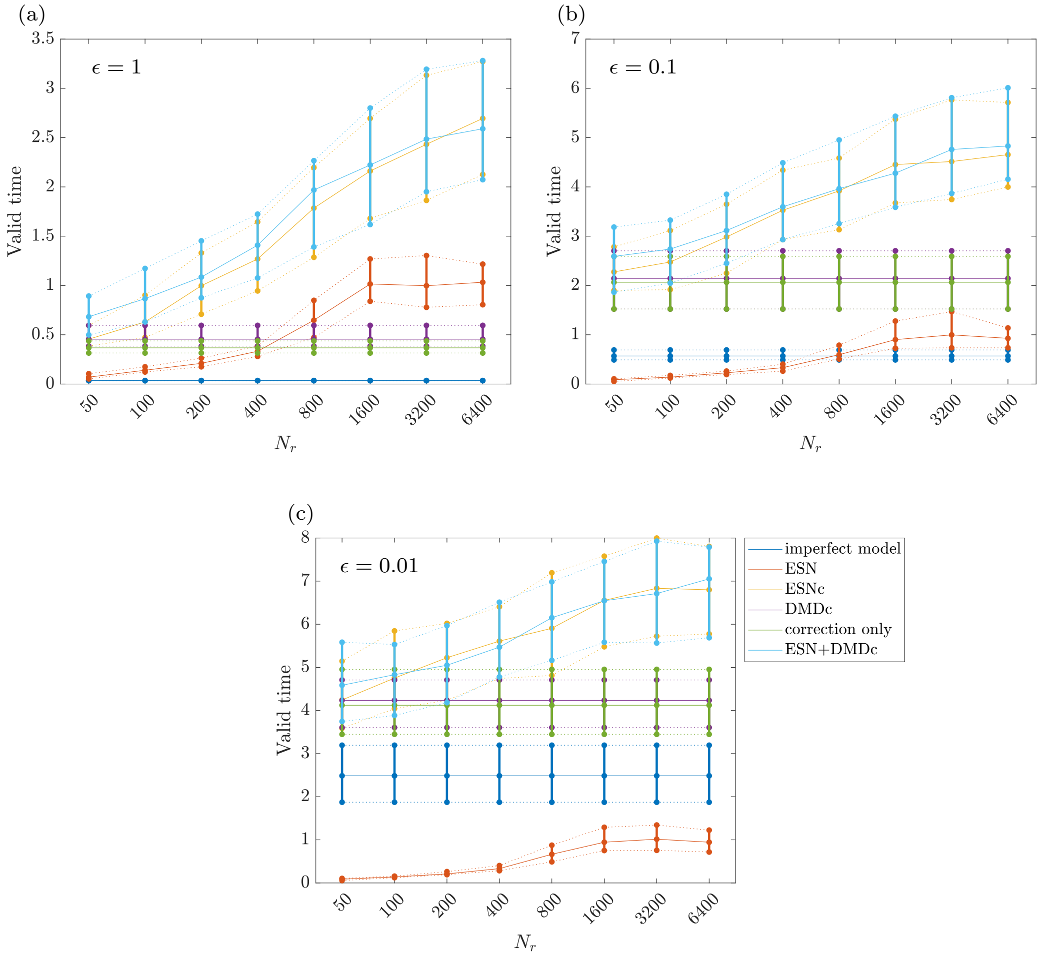

Figure 1: Results for the replication of the experiments in Pathak et al. (2018) where the imperfect model is a perturbed version of the KS-equation with perturbation parameter ϵ. 'Valid time' is the time it takes until the error threshold is passed: $E\left(\mathbf{x}^k, \mathbf{y}^k\right) > 0.4$. These timings are in Lyapunov units ($\lambda_{max}t$). The experiment is repeated for 100 different training intervals and network realizations. For each N_r a box plot is depicted showing the first, second and third quartile.

dictions. For each method we use 100 different training intervals and hence network realizations, as we do not reuse W. We fix the regularization parameter at $\lambda = 1 \cdot 10^{-5}$. The pure DMD-based methods (DMD and ESN+DMD) are not shown as they did not produce meaningful results. This is likely caused by DMD generalizing poorly and showing only valid predictions for a short period after $t_0 + T$.

The short-term prediction accuracy is measured using the normalized error used in Pathak et al. (2018). We compare the k-th prediction $\mathbf{x}^k$ with the restricted truth $\mathbf{y}^k =$

$R\boldsymbol{\xi}^k$ through

$$E\left(\mathbf{x}^k, \mathbf{y}^k\right) = \frac{\left\|\mathbf{x}^k - \mathbf{y}^k\right\|}{\sqrt{\left\langle \left\|\mathbf{y}^k\right\|^2 \right\rangle}}, \tag{17}$$

with $\langle \cdot \rangle$ the mean over a time window up until k.

In Fig. 1 we see a strong resemblance with the results in Pathak et al. (2018). The imperfect model performs poorly on its own and the ESN-based methods improve the prediction as expected. A standalone ESN is able to achieve decent predictions for $\epsilon = 1$ and $\epsilon = 0.1$. For $\epsilon = 0.01$, however, it appears impossible for a standalone ESN to perform better than the imperfect model. In all studied cases it is remarkable how the hybrid variant ESNc stands out. By combining the imperfect model physics with the ESN a significant improved is achieved.

The three additional models in Table 1 further explain the advantage of the hybrid ESNc over the standalone ESN. The correction-only and DMDc predictions do not depend on N_r and show up here as constant solutions. These two regression-based corrections outperform the standalone ESN for $\epsilon = 0.1$ and $\epsilon = 0.01$. The third approach, ESN+DMDc, follows the ESNc performance but with an overall slight advantage for the two largest perturbations ϵ. This advantage is explained by the performance of DMDc and correction-only, as these are the linear components of ESN+DMDc and ESNc, respectively. In experiments where DMDc outperforms the correction we find a similar overall gain between ESN+DMDc and ESNc. From the experiments in Fig. 1 it is apparent that ESN+DMDc and ESNc reduce to their linear components for low N_r, which is what would be expected from the correction equation (13). Hence the performance of the linear models can be seen as a departure point for hybrid variants that add a nonlinear ESN. This largely explains the performance gain of, e.g. ESNc over the standalone ESN.

In a different perfect/imperfect model setup, illustrating the symbiotic modeling approach, the models both use $\epsilon = 0$ and have different spatial resolutions instead. The perfect model is discretized on a grid with twice the resolution, $N_f = 2N_c$. The domain size, ESN parameters and regularization remain unchanged. As explained in Section 2.1, fine grid information is restricted to the coarse grid and any data-driven corrections are made to the imperfect, coarse model evolution. Hence, instead of a model perturbation, it is now the difference in truncation errors and resolved scales between two resolutions that causes a model mismatch. With this setup the approach given by Equations (10)-(13) can be seen as a subgrid modeling technique.

The coarse model is capable of a good prediction in this setup (Fig. 2). DMDc, the correction-only and the standalone ESN are all unable to improve the coarse model. How-

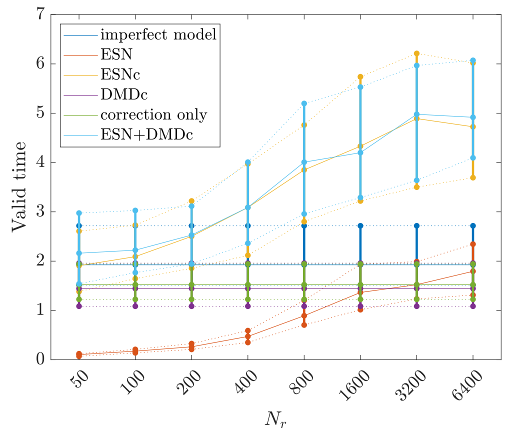

Figure 2: Grid experiment with the KS-equation. The imperfect model consists of the same equations but discretized on a grid half the resolution of the perfect model. Solutions are valid until $E\left(\mathbf{x}^k, \mathbf{y}^k\right) > 0.4$. As in Fig. 1, we repeat the experiment for 100 different training sets and network realizations.

ever, the hybrid variants ESNc and ESN+DMDc do show an overall improvement and an increase in predictive skill for larger N_r, similar to the parameter perturbation results (Fig. 1). For large values of N_r the hybrid methods double the predictive performance. This, again, shows the benefit of introducing the imperfect physical predictions to both force and control the artificial ESN. Hence the hybrid approach in Pathak et al. (2018) shows promise as a nonlinear subgrid modeling technique.

4 Results: quasi-geostrophic model

The barotropic quasi-geostrophic (QG) vorticity equation for a square (length L, constant depth D) ocean basin is solved on a β-plane. The ocean flow is driven by an idealized zonal wind-stress forcing τ^x. Typical horizontal length and velocity scales are denoted L and U, from which the time scale follows as L/U. Using $L = 10^6$ m and $U = 3.17{\cdot}10^{-2}$ ms^{-1}, we obtain a time scale of approximately one year. The equations are solved on a square domain, $x \in [0, 1]$, $y \in [0, 1]$, with periodic boundaries in both directions.

The QG equations in non-dimensional form are given by

$$\left[\frac{\partial}{\partial t} - \frac{\partial \psi}{\partial y}\frac{\partial}{\partial x} + \frac{\partial \psi}{\partial x}\frac{\partial}{\partial y}\right](\omega + \beta y) = \frac{1}{\mathrm{Re}}\nabla^2\omega + \alpha_\tau C_\tau(x, y), \tag{18}$$

$$\omega = \nabla^2\psi, \tag{19}$$

with ω the vertical component of the vorticity vector and streamfunction ψ. The Reynolds number is $\mathrm{Re} = LU/A_H$, where A_H is the horizontal mixing coefficient and $\beta = \beta_0 L^2/U$, with $\beta_0 = 1.6{\cdot}10^{-11}$ $(\mathrm{ms})^{-1}$. Wind forcing enters through the nondimensional parameter $\alpha_\tau = \tau_0 L/(\rho D U^2)$, with forcing amplitude $\tau_0 = 0.3$ Nm^{-2}, density $\rho = 1 \cdot 10^3$ kg m^{-3} and layer depth $D = 6 \cdot 10^2$ m. We use a constant idealized wind-stress curl forcing in the form of a stirring pattern with stirring wavenumber $k_f = 5$ in both directions:

$$C_\tau(x, y) = \cos\left(2k_f\pi\, x\right)\, \cos\left(2k_f\pi\, y\right). \tag{20}$$

This problem setup is a variant of the approach in Edeling & Crommelin (2019), but here we add a rotating frame.

4.1 Approach

Following the perfect/imperfect model approach we discretize the QG equations on two different grids. The perfect model uses a fine discretization on Ω^f with $N_f = 2 \cdot 256^2$ unknowns and the imperfect variant is discretized on Ω^c with $N_c = 2 \cdot 32^2$ unknowns ($N_f = 64 N_c$). Furthermore, for both grids we model a flow with a Reynolds number that does not cause any numerical artifacts. With the perfect model we can run with $\mathrm{Re}_f = 2000$. A stable flow for this Reynolds number and forcing amplitude τ_0 is impossible to achieve on the coarse grid and we therefore choose to use $\mathrm{Re}_c = 500$ for the imperfect model.

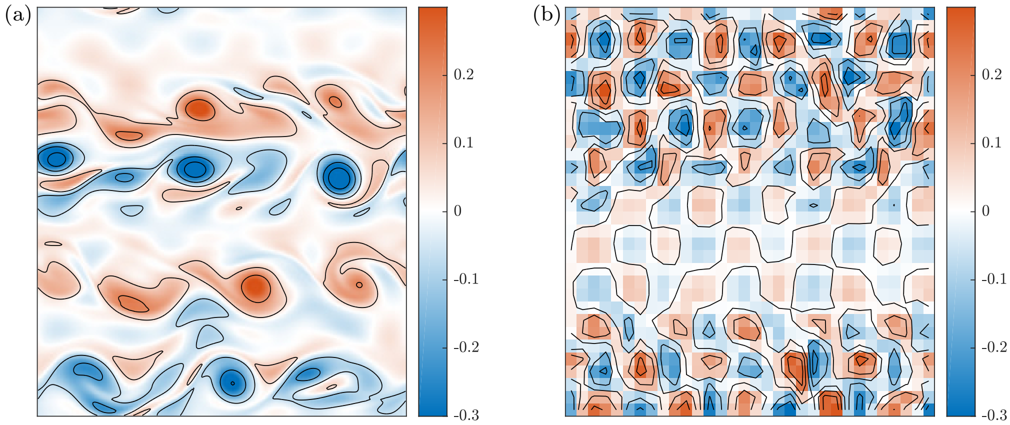

Figure 3: Snapshots of the vorticity fields (in day^{-1}) at the end of the transient depicted in Fig. 5. (a) Perfect model vorticity snapshot from a statistical equilibrium with $N_f = 2 \cdot 256^2$ and $\mathrm{Re}_f = 2000$. (b) Imperfect model vorticity, also in a statistical equilibrium, $N_c = 2 \cdot 32^2$ and $\mathrm{Re}_c = 500$.

For the discretization in time we use a fully implicit time stepping scheme that allows the use of the same time step for both models. In our experiments we will use $\Delta t =$ 1 day. The perfect QG solution (ω, ψ) is randomly initialized and run into a statistical steady state. From the steady state we select training periods of size $T = N_T = 10,000$ days and follow the data gathering process described in Section 2.1. To get an idea of the perfect and imperfect flows we restart the imperfect model from a restricted fine state and run it into a steady regime. Snapshots from the two different statistical steady states are shown in Fig. 3. The imperfect model solution in Fig. 3b is highly diffusion dominated and shows a flow that strongly reflects the forcing pattern. The 'perfect' solution in Fig. 3a is — with 256^2 grid points — a moderately high-resolution flow and the difference in resolved features with the imperfect model is substantial, which makes it an ideal testing ground for the corrective approaches in Table 1.

For the QG flow problem we will investigate the performance of the corrected transients, following (10)-(13), in two different ways. We will make short-term predictions with the methods in Table 1 and compare with the truth using a normalized error, similar to the KS results in the Figs. 1 and 2. Then we let the different corrective models run into a statistical equilibrium regime and compare the probability density functions (PDFs) of key flow properties with the perfect model equilibrium using their Kullback–Leibler divergence D_{KL}.

4.2 Short-term predictions

In Fig. 4 we present a short-term prediction experiment using the methods in Table 1. Only the standalone DMD and DMD+ESN corrections are excluded for their lack of meaningful results. For the ESN operators we again use $\rho(W) = 0.4$ and $\overline{d} = 3$, but with $T = 10,000$ days and $\Delta t = 1$ day we use half the amount of training data. For this problem we find that the optimal relaxation parameter lies around $\alpha = 0.2$ and the regularization is increased to $\lambda = 1 \cdot 10^{-4}$. The number of accurate days is measured using a stricter tolerance $E\left(\mathbf{x}^k, \mathbf{y}^k\right) < 0.2$, allowing only a small departure from the true trajectory.

The poor performance of the imperfect QG model shown in Fig. 4 is improved by all studied methods. The standalone ESN needs at least $N_r = 1600$, while the other methods show a significant improvement for all chosen N_r. From Equation (13) and Table 1 it is evident that ESNc is an ESN combined with the correction-only approach. In the short-term QG predictions we find that these methods coincide for small N_r. A similar observation can be made for DMDc and the combination ESN+DMDc, which also coincide for low N_r. Controlled DMDc has better short-term predictive power than the

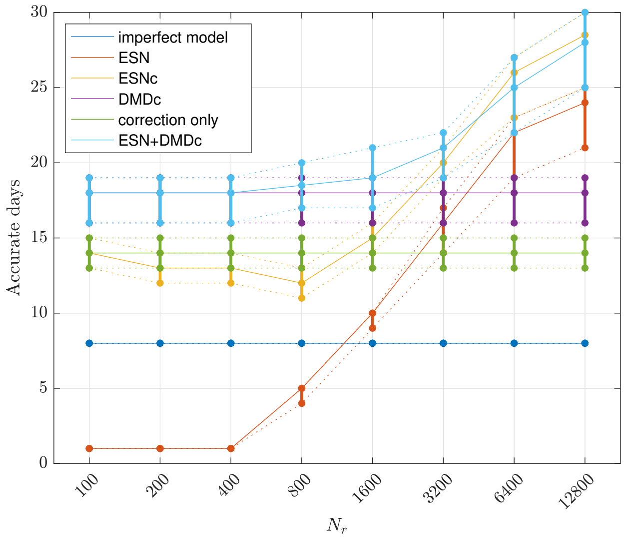

Figure 4: Short-term prediction experiments with the imperfect QG equations in a setup similar to Fig. 2. The experiments are repeated for 50 different network realizations and training sets. 'Accurate days' marks the time steps ($\Delta t = 1$ day) it takes until the error threshold is passed: $E\left(\mathbf{x}^k, \mathbf{y}^k\right) > 0.2$.

correction-only variant, which is also reflected in the behavior of ESN+DMDc and ESNc at low N_r. With this domain setup we expect DMDc to perform reasonably well on short time scales and it can therefore be viewed as a linear benchmark. The nonlinear ESN+DMDc hybrid improves on it immediately and it takes at least $N_r = 3200$ for the other ESN-based methods to take over. For large N_r both hybrid methods (ESN+DMDc and ESNc) almost coincide and any positive influence of the DMD component is negligible.

The N_r doubling results are reminiscent of the findings with the KS-equation here and in Pathak et al. (2018). Similar to the KS scaling results, increasing N_r improves the short-term predictions of ESN-based methods for the QG problem. Based on the experiments with the KS-equation we expect that also here a plateau or a maximum will be reached for $N_r > 12800$. For ESN state sizes ranging between 200 and 1600 the ESN+DMDc combination gives the best results, where ESNc shows a slight decrease in performance. After $N_r = 1600$, the ESN component begins to dominate the results and ESNc becomes comparable to ESN+DMDc. Note, however, that also the standalone ESN is doing remarkably well for large state sizes.

4.3 Long-term dynamical regime

For the short-term results in the previous subsection, we used a normalized error based on the full fields (ω, ψ) for a comparison of the 'hybrid' model results with the (restricted) perfect model truth. Failure in terms of this measure does not imply the predictions are invalid, only that the exact truth is not reproduced. We are therefore also interested in reproducing ergodic properties of long-term time series as in Pathak et al. (2017). In this fashion we will continue here and investigate three flow properties for long-term transient runs: mean kinetic energy K_m, eddy kinetic energy K_e and enstrophy Z.

Horizontal velocities u, v follow from the streamfunction ψ, with $u = -\partial\psi/\partial y,\ v = \partial\psi/\partial x$, and are decomposed into a (time) mean and transient component: $u = \langle u \rangle + u^{'}, v = \langle v \rangle + v^{'}$ with the mean $\langle \cdot \rangle$ taken over a window of 50 days. The quantities K_m, K_e and Z are then given by

$$K_m = \int_\Omega \left(\langle u \rangle^2 + \langle v \rangle^2 \right) d\Omega, \tag{21}$$

$$K_e = \int_\Omega \left(\left\langle u^{'2} \right\rangle + \left\langle v^{'2} \right\rangle \right) d\Omega = \int_\Omega \left(\left\langle u^2 \right\rangle - \langle u \rangle^2 + \left\langle v^2 \right\rangle - \langle v \rangle^2 \right) d\Omega, \tag{22}$$

$$Z = \int_\Omega \omega^2 \, d\Omega, \tag{23}$$

where the integral is approximated with a Riemann sum over the coarse domain Ω^c.

A switch from the perfect ($N_f = 2{\cdot}256^2$, $\mathrm{Re}_f = 2000$) to the imperfect ($N_c = 2{\cdot}32^2$, $\mathrm{Re}_f = 500$) QG model solution will inevitably lead to a different statistical steady state. An example of this process is presented in Fig. 5. The perfect QG model is randomly initialized and runs into a statistical equilibrium. Predictions using imperfect QG, a standalone ESN and the hybrid ESNc then start from a restricted perfect QG state and run for 100 years. For stable long-term transients with the ESN-based methods we need a significantly larger regularization parameter ($\lambda = 1$) compared to the short-term experiments. Vorticity snapshots of the perfect and imperfect model depicted in Fig. 3 are taken at the end of the trajectories in Fig. 5. In Fig. 6 we present vorticity snapshots at the end of the ESN and ESNc trajectories.

The imperfect model reaches a very different statistical equilibrium after a transition period of approximately 10 years. A corrected transient based on (10)-(13) should stay closer to the perfect model's dynamical regime and the presented ESN and ESNc trajectories show that this is feasible. Especially the hybrid ESNc shows a significantly better reproduction of the perfect model's K_m PDF, compared to imperfect model (Fig. 5b).

In Fig. 7 the average energy spectrum over the final 80 years in Fig. 5a is shown. The spectrum provides another demonstration of the improved dynamics given by the

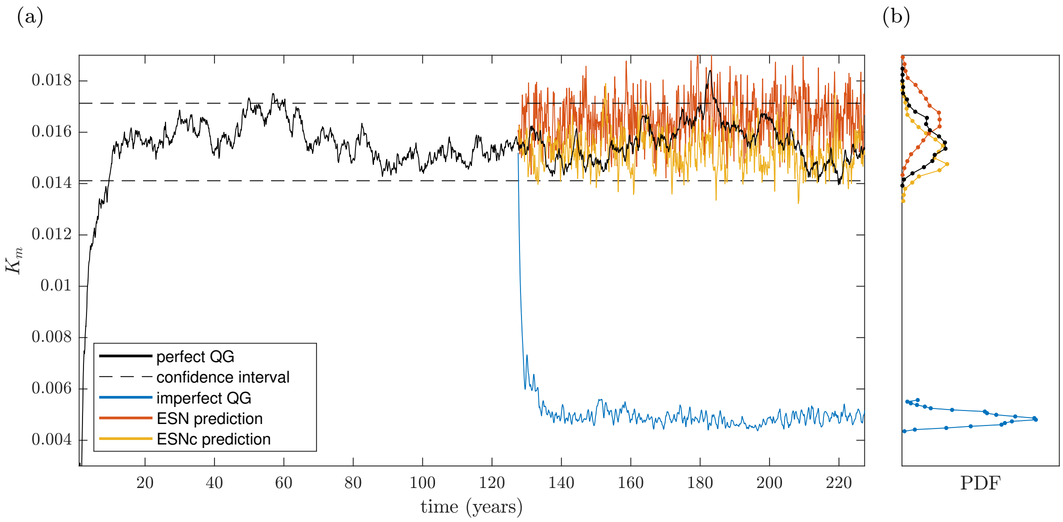

Figure 5: Spinup and long-term transient dynamics indicated by mean kinetic energy K_m. (a) A 100 year spinup with the perfect QG equations using a time step $\Delta t = 1$ day is followed by a training period of $N_T \Delta t = T = 10,000$ days. After the training period, 100 year predictions with imperfect QG, ESN and ESNc are shown, using $N_r = 3200$ and $\lambda = 1$. (b) Approximations of the probability density functions (PDFs) associated with the equilibrium transients, using 20 bins and excluding spinup/transition periods. Vorticity snapshots at the end of the depicted trajectories are shown in Figs. 3 and 6. The ESN results are with single realizations and serve as a demonstration of the corrected dynamics.

standalone ESN and the hybrid ESNc. The imperfect QG solution strongly reflects the forcing, which is also noticeable in the vorticity snapshot (Fig. 3b). In an incompressible 2D flow we expect energy to be transferred from the stirring wavelength to the larger scales, whereas enstrophy is transferred to the smaller scales and dissipated (Vallis, 2019). Both the energy and the enstrophy transfer are poorly represented in the imperfect model. The ESN-based methods are a lot better at producing the correct energy transfer and achieve a good correspondence for the lowest frequencies. Around the stirring frequency ESNc still performs well, whereas the standalone ESN is overestimating. The enstrophy transfer appears even more difficult to capture correctly but still the hybrid ESNc shows a great improvement over the standalone ESN at these scales.

The transients shown in Fig. 5 are specific examples and provide only information for a single realization of the ESN and a single training range. For a more rigorous approach we compute transients for 50 training periods (and hence network realizations).

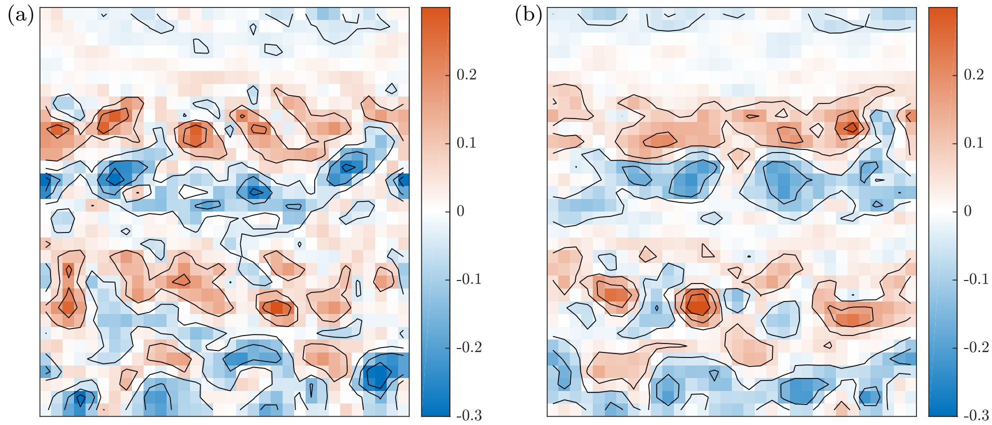

Figure 6: Snapshots of the vorticity fields (in day^{-1}) at the end of the transient in Fig. 5. (a) standalone ESN prediction with $N_r = 3200$ and $\lambda = 1$, (b) hybrid ESNc prediction with $N_r = 3200$ and $\lambda = 1$.

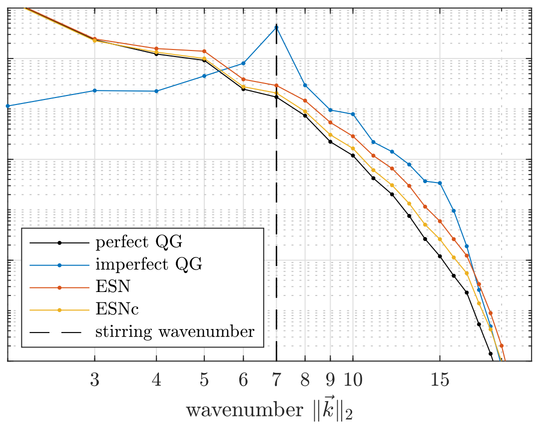

Figure 7: Average equilibrium energy spectrum based on the final 80 years of the trajectories in Fig. 5a. A dashed line is added to mark the frequency of the forcing.

466 We turn to all models studied in the short-term experiment (Fig. 4) and, to maintain a stable iteration, need to increase the regularization parameter λ. For the ESN-based methods we use $\lambda = 1$, for correction-only we will use $\lambda = 5$ and with DMDc we use $\lambda = 10$ to compute stable evolutions. Later in this section we explore how these methods perform for various other λ choices.

From the trajectories we compute flow properties (K_m, K_e, Z) and compare their PDFs to the perfect model using their Kullback–Leibler (KL) divergence (Cover & Thomas,

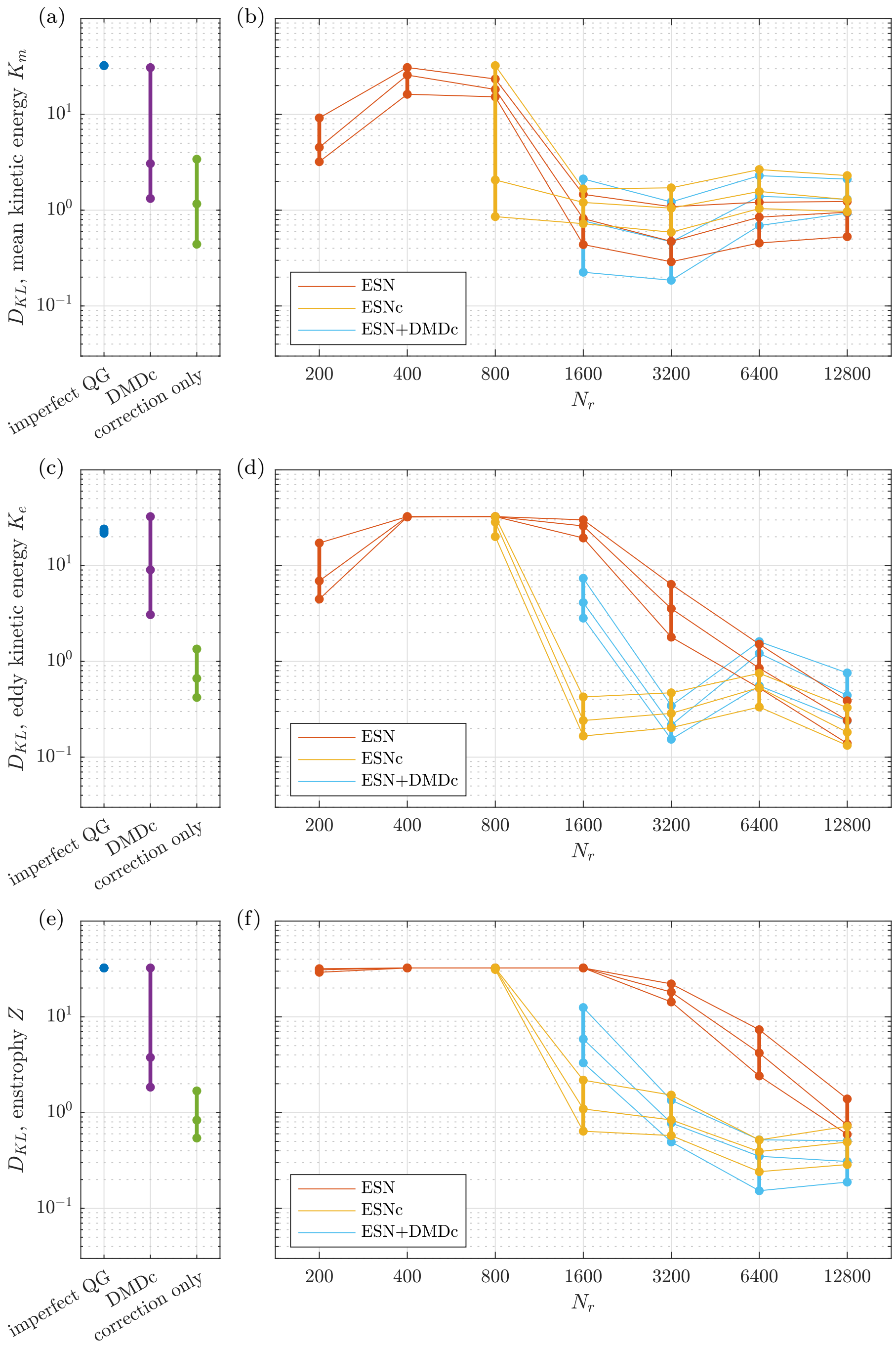

Figure 8: Long-term (100 year) transient results for 50 different training intervals. D_{KL} results from methods that do not depend on an ESN are shown in (a), (c) and (e), for K_m, K_e and Z, respectively. In (b), (d) and (f) the respective scalings with N_r are depicted for models with an ESN-dependence. Missing values in the plots are caused by unstable configurations.

2006): for two discrete distributions P and Q, the divergence of Q from P is given by

$$D_{KL}(P,Q) = \sum_i P_i \ln\left(\frac{P_i}{Q_i}\right). \tag{24}$$

The PDFs are approximated using a domain that ranges beyond the perfect model's PDF with twice the standard deviation. This domain is divided into 100 bins and every transient is truncated to exclude initial spinup effects. For each flow property the divergence of its PDF from the 'truth' is computed and combined into boxplots for different ESN state sizes N_r (Fig. 8). We avoid division by zero in (24) by substituting zero-values with machine precision. This leads to large but finite divergences for non-overlapping distributions (~ 32) . The imperfect model shows a poor representation of the variability, which should be expected from the transient example in Fig. 5. The PDFs for all flow properties show no resemblance with the true PDF, giving D_{KL} results that remain at the maximum divergence value. Controlled DMDc and correction-only methods are better at capturing the variability, although this is higly dependent on the stabilizing regularization. Especially for the correction-only approach it is possible to find a configuration such that PDFs give a reasonable correspondence.

The KL–divergences for ESN-based methods in Fig. 8 are partly missing. For low N_r, ESNc and ESN+DMDc are unstable when $\lambda = 1$. The remaining results show an overall improvement for increasing ESN state size N_r (cf. Fig. 1), although not very clear for all flow properties. Both mean and eddy kinetic energy KL–divergences are somewhat irregular with optima at moderate N_r values. For enstrophy, the ESN-based methods gradually improve with ESN state size. From the energy spectrum in Fig. 7 we know that the enstrophy transfer is difficult to capture and here a similar effect is visible in the correspondence between PDFs. ESNc requires at least $N_r = 1600$ to obtain small KL–divergences from the enstrophy PDF, further improving for larger N_r.

Diverged trajectories show up as non-overlapping with either a maximal KL–divergence or a missing value in the D_{KL} results. Poor performing methods are hence indistinguishable from unstable ones. Especially the combination ESN+DMDc appears to suffer from stability issues for small N_r, leading to missing D_{KL} values. We find that the ESN stabilizes regression-based corrective methods, as already noted in Arcomano et al. (2022). When the regression-based methods run on their own we choose a regularization that stabilizes sufficiently such that divergent trajectories are rare.

To provide an idea of how regularization affects the long-term performance of various methods we perform numerous equilibrium runs for different λ. In Fig. 9 we present the results for enstrophy Z. The correction-only approach gives remarkably good results within a narrow optimal region for λ. It is also only slightly enhanced by the combina-

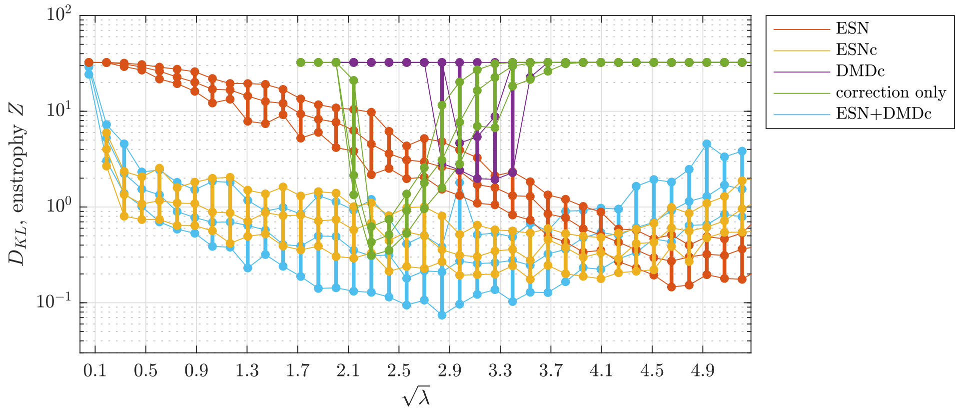

Figure 9: Reproduction of the enstrophy Z variability for different regularization parameters using an equidistant spacing in $\sqrt{\lambda}$. Long-term (100 year) equilibrium runs are performed for 50 different (but partially overlapping) training sets and network realizations. Boxplots show the first, second and third quartile of the resulting spread of divergences D_{KL}. The ESN-based methods have dimension $N_r = 3200$.

tion with an ESN (i.e. ESNc). The hybrid ESNc and ESN+DMDc are, however, much more robust and overall better at reproducing the correct enstrophy variability. From the regularization parameter study it is clear that DMDc needs a stronger regularization than the correction-only approach. The KL–divergences in Fig. 8 show a related problem for the models that incorporate an ESN, where the ESN that combines with DMDc needs a much larger state size N_r to achieve sufficient stabilization. Hence stabilization is achieved through both regularization λ and ESN complexity N_r.

5 Summary and discussion

In this paper we demonstrated part of a symbiotic ocean modeling approach, i.e., a framework in which models with different complexities are coupled in order to benefit from each other. We distinguish between perfect and imperfect models in terms of differing spatial resolutions and key parameterizations, and focus on how an imperfect model can benefit from a symbiotic setup. With data generated from both model types we seek to correct imperfect model transients. To this end, we make use of hybrid modeling techniques that combine linear regression-based methods with nonlinear echo state networks (ESNs). Currently, efforts are under way to demonstrate the second part of the symbiotic framework, i.e., HR models that benefit computationally from LR models.

We establish that our hybrid (or physics-controlled) ESNc implementation reproduces short-term predictions for the Kuramoto–Sivashinsky equation (KS) that are consistent with earlier work in Pathak et al. (2017). Our framework furthermore allows a straightforward comparison with purely regression-based methods. We show how corrections based on linear regression contribute to the success of the hybrid machine learning combinations and serve as a departure point for hybrid methods. When we apply these techniques to a subgrid modeling version of the KS problem, we observe a similar scaling behavior with ESN complexity and departure points rooted in the linear regression techniques. For the subgrid modeling problem of the single-layer quasi-geostrophic potential vorticity equations (QG), short-term predictions give results that are comparable to the findings with the KS-equation. A scaling behavior is found with the size of the ESN. When the ESN complexity is negligible, the hybrid methods reduce to their linear regression components. For the long-term flow development, our comparison of statistical steady states shows that the hybrid combinations are robust and perform well for various flow parameters. The parameter study with long-term statistics also shows how the ESN-based methods improve with ESN state size, reminiscent of the short-term full-field reproductions. For our purposes, however, the comparison of long-term flow characteristics is more informative than an error norm on state differences.

The parameter studies with equilibrium simulations show that subgrid models based on only an ESN or regression are often inaccurate or difficult to stabilize. We control the stabilization through regularizing the regression-problem, which is another (hyper) parameter to tweak. For purely regression-based methods, regularization is the only tunable apart from data choices. A benefit of combining regression with an ESN is apparent from our regularization experiments. Here we observe that the ESN stabilizes its regression-based component, which was also mentioned in Arcomano et al. (2022). We conclude that stabilization is achieved through both regularization λ and ESN state size N_r. However, other parameters such as the spectral radius $\rho(W)$ (controlling damping in the ESN) were not studied in this context and may have similar stabilizing effects. Combinations with a DMD model are available within our framework, which yields interesting comparisons, especially in the short-term QG experiments. Benefits of adding a DMD model are visible for moderate ESN state sizes. For long-term transient runs the advantage of hybrid DMD-ESN models is less pronounced, which is possibly due to the DMD model being valid for only a short period and hence it should be (partially) rebuilt in an online fashion (Pendergrass et al., 2016).

Obviously, the QG ocean model used here is highly idealized compared to state-of-the-art ocean models. However, we think that these ideas are applicable to the general problem of correcting large scale flows, i.e., improving a coarse and more viscous version of the flow problem at hand. For models with a higher dimension than studied here a reduced order version of the corrective transient framework, as defined by (10)-(13), is worth investigating. Here the best choice of reduced coordinates (POD, Fourier, wavelets) in combination with an ESN remains uncertain. Projecting with global POD modes, for instance, greatly reduces the ESN's predictive skill (Vlachas et al., 2020). A localized representation as used in Wan et al. (2021) shows more promise. Another way to tackle high-dimensional problems is through parallelization. A parallel hybrid ESNc based on a local domain decomposition is used in Wikner et al. (2020) and Arcomano et al. (2022). It would be interesting to apply this approach as a subgrid model and reproduce long-term flow characteristics, comparing especially its ability to correctly capture energy and enstrophy transfer at low wave numbers.

Open Research

The software developed for this paper is archived at Zenodo and available through `https://doi.org/10.5281/zenodo.7572246`.

Acknowledgments

This work was supported by funding from the SMCM project of the Netherlands eScience Center (NLeSC) with project number 027.017.G02. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. In addition we thank T.M. Hepkema for helpful discussions and two anonymous reviewers for their thorough and useful comments which greatly improved this work.

References

Arcomano, T., Szunyogh, I., Wikner, A., Pathak, J., Hunt, B. R., & Edward, O. (2022). A Hybrid Approach to Atmospheric Modeling that Combines Machine Learning with a Physics-Based Numerical Model. *Preprint*. doi: 10.1002/essoar.10507548.2

Berloff, P. S. (2005). Random-forcing model of the mesoscale oceanic eddies. *J. Fluid Mech.*, *529*, 71–95. doi: 10.1017/S0022112005003393

Bollt, E. (2021). On explaining the surprising success of reservoir computing forecaster of chaos? The universal machine learning dynamical system with contrast to VAR and DMD. *Chaos*, *31*(1). doi: 10.1063/5.0024890

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001472 doi: https://doi.org/10.1029/2018MS001472

Briggs, W. L., Henson, V. E., & McCormick, S. F. (2000). *A multigrid tutorial* (Second ed.). Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Retrieved from https://doi.org/10.1137/1.9780898719505 doi: 10.1137/1.9780898719505

Chang, P., Zhang, S., Danabasoglu, G., Yeager, S. G., Fu, H., Wang, H., ... Wu, L. (2020). An Unprecedented Set of High-Resolution Earth System Simulations for Understanding Multiscale Interactions in Climate Variability and Change. *Journal of Advances in Modeling Earth Systems*, *12*(12). doi: 10.1029/2020ms002298

Chassignet, E. P., Yeager, S. G., Fox-Kemper, B., Bozec, A., Castruccio, F., Danabasoglu, G., ... Xu, X. (2020). Impact of horizontal resolution on global ocean–sea ice model simulations based on the experimental protocols of the ocean model intercomparison project phase 2 (omip-2). *Geoscientific Model Development*, *13*(9), 4595–4637. Retrieved from https://gmd.copernicus.org/articles/13/4595/2020/ doi: 10.5194/gmd-13-4595-2020

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (Second ed.). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Edeling, W., & Crommelin, D. (2019). Towards data-driven dynamic surrogate models for ocean flow. *Proceedings of the Platform for Advanced Scientific Computing Conference, PASC 2019*. doi: 10.1145/3324989.3325713

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. Retrieved from https://gmd.copernicus.org/articles/9/1937/2016/ doi: 10.5194/gmd-9-1937-2016

Gargett, A. E. (1989). Ocean turbulence. *Annual Review of Fluid Mechanics*. Retrieved from http://www.annualreviews.org/doi/pdf/10.1146/annurev.fl.21.010189.002223

Gent, P. R., Willebrand, J., McDougall, T. J., & McWilliams, J. C. (1995). Parameterizing eddy-induced tracer transports in ocean circulation models. *Journal Of Physical Oceanography*, *25*, 463 – 474.

Hallberg, R. (2013). Using a resolution function to regulate parameterizations of oceanic mesoscale eddy effects. *Ocean Modelling*, *72*, 92–103.

Hewitt, H. T., Roberts, M., Mathiot, P., Biastoch, A., Blockley, E., Chassignet,

E. P., . . . Zhang, Q. (2020). Resolving and Parameterising the Ocean Mesoscale in Earth System Models. *Current Climate Change Reports*, *6*(4), 137–152. doi: 10.1007/s40641-020-00164-w

Hyman, J. M., & Nicolaenko, B. (1986). The kuramoto-sivashinsky equation: A bridge between pde's and dynamical systems. *Physica D: Nonlinear Phenomena*, *18*(1), 113-126. Retrieved from https://www.sciencedirect.com/science/article/pii/0167278986901661 doi: https://doi.org/10.1016/0167-2789(86)90166-1

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. doi: 10.1038/s42256-021-00374-3

Jaeger, H. (2001). *The "echo state" approach to analysing and training recurrent neural networks* (GMD Report No. 148). GMD - German National Research Institute for Computer Science. Retrieved from http://www.faculty.jacobs-university.de/hjaeger/pubs/EchoStatesTechRep.pdf

Jaeger, H., & Haas, H. (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, *304*(5667), 78–80. doi: 10.1126/science.1091277

Jüling, A., Zhang, X., Castellana, D., Heydt, A. S. v. d., & Dijkstra, H. A. (2021). The Atlantic's freshwater budget under climate change in the Community Earth System Model with strongly eddying oceans. *Ocean Science*, *17*(3), 729–754. doi: 10.5194/os-17-729-2021

Kuramoto, Y. (1984). *Chemical oscillations, waves, and turbulence* (Vol. 19). Springer-Verlag, Berlin. Retrieved from https://doi.org/10.1007/978-3-642-69689-3 doi: 10.1007/978-3-642-69689-3

Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). *Dynamic mode decomposition : Data-driven modeling of complex systems.* SIAM. Retrieved from http://www.dmdbook.com/

Lu, Z., Pathak, J., Hunt, B., Girvan, M., Brockett, R., & Ott, E. (2017). Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. *Chaos*, *27*(4). Retrieved from http://dx.doi.org/10.1063/1.4979665 doi: 10.1063/1.4979665

Lukosevicius, M. (2012). A practical guide to applying echo state networks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade (2nd ed.)* (Vol. 7700, p. 659-686). Springer.

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149. doi: 10.1016/j.cosrev.2009.03.005

Mana, P. P., & Zanna, L. (2014, 07). Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling*, *79*(C), 1 – 20. doi: 10.1016/j.ocemod.2014.04.002

Nadiga, B. T. (2021). Reservoir Computing as a Tool for Climate Predictability Studies. *Journal of Advances in Modeling Earth Systems*, e2020MS002290. doi: 10.1029/2020ms002290

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos*, *27*(12). doi: 10.1063/1.5010300

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos*, *28*(4). Retrieved from `https://doi.org/10.1063/1.5028373` doi: 10.1063/1.5028373

Pendergrass, S. D., Kutz, J. N., & Brunton, S. L. (2016). *Streaming gpu singular value and dynamic mode decompositions.* arXiv. Retrieved from `https://arxiv.org/abs/1612.07875` doi: 10.48550/ARXIV.1612.07875

Proctor, J. L., Brunton, S. L., & Kutz, J. N. (2016). Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, *15*(1), 142-161. Retrieved from `https://doi.org/10.1137/15M1013857` doi: 10.1137/15M1013857

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684-9689. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.1810286115` doi: 10.1073/pnas.1810286115

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, *656*, 5–28. doi: 10.1017/S0022112010001217

Sivashinsky, G. I. (1977). Nonlinear analysis of hydrodynamic instability in laminar flames. I. Derivation of basic equations. *Acta Astronaut.*, *4*(11-12), 1177–1206. Retrieved from `https://doi.org/10.1016/0094-5765(77)90096-0` doi: 10.1016/0094-5765(77)90096-0

Vallis, G. K. (2019). *Essentials of atmospheric and oceanic dynamics.* Cambridge University Press. doi: 10.1017/9781107588431

Viebahn, J., Crommelin, D., & Dijkstra, H. (2019). Toward a turbulence closure based on energy modes. *Journal of Physical Oceanography*, *49*(4), 1075 - 1097. Retrieved from `https://journals.ametsoc.org/view/journals/phoc/49/4/`

jpo-d-18-0117.1.xml doi: 10.1175/JPO-D-18-0117.1

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., & Koumoutsakos, P. (2020). Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, *126*, 191–217. Retrieved from https://doi.org/10.1016/j.neunet.2020.02.016 doi: 10.1016/j.neunet.2020.02.016

Wan, Z. Y., Dodov, B., Lessig, C., Dijkstra, H., & Sapsis, T. P. (2021). A data-driven framework for the stochastic reconstruction of small-scale features with application to climate data sets. *Journal of Computational Physics*, *442*, 110484. Retrieved from https://doi.org/10.1016/j.jcp.2021.110484 doi: 10.1016/j.jcp.2021.110484

Wikner, A., Pathak, J., Hunt, B., Girvan, M., Arcomano, T., Szunyogh, I., … Ott, E. (2020). Combining machine learning with knowledge-based modeling for scalable forecasting and subgrid-scale closure of large, complex, spatiotemporal systems. *Chaos (Woodbury, N.Y.)*, *30*(5), 053111. doi: 10.1063/5.0005541