

Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model

Pavel Perezhogin¹, Laure Zanna¹, Carlos Fernandez-Granda^{1,2}

¹Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

²Center for Data Science, New York University, New York, NY, USA

Key Points:

- We propose generative machine learning (ML) models to build stochastic parameterization of subgrid mesoscale eddies
- Generative models produce a flow-dependent estimation of the uncertainty with spatially correlated stochastic forcing
- Generative models demonstrate superior numerical stability and outperform baseline ML models in online simulations at the coarsest grid

arXiv:2302.07984v1 [physics.ao-ph] 15 Feb 2023

Abstract

Subgrid parameterizations of mesoscale eddies continue to be in demand for climate simulations. These subgrid parameterizations can be powerfully designed using physics and/or data-driven methods, with uncertainty quantification. For example, Guillaumin and Zanna (2021) proposed a Machine Learning (ML) model that predicts subgrid forcing and its local uncertainty. The major assumption and potential drawback of this model is the statistical independence of stochastic residuals between grid points. Here, we aim to improve the simulation of stochastic forcing with generative models of ML, such as Generative adversarial network (GAN) and Variational autoencoder (VAE). Generative models learn the distribution of subgrid forcing conditioned on the resolved flow directly from data and they can produce new samples from this distribution. Generative models can potentially capture not only the spatial correlation but any statistically significant property of subgrid forcing. We test the proposed stochastic parameterizations offline and online in an idealized ocean model. We show that generative models are able to predict subgrid forcing and its uncertainty with spatially correlated stochastic forcing. Online simulations for a range of resolutions demonstrated that generative models are superior to the baseline ML model at the coarsest resolution.

Plain Language Summary

The climate system includes physical phenomena on a wide range of scales from millimeter scale in the boundary layer to planetary scale. Numerical models used for climate projections can directly simulate only the largest spatiotemporal scales of the flow, while the missing physics due to unresolved (or subgrid) flows must be parameterized. The prediction of the missing term given only the information about the resolved flow is a difficult task, given in part the uncertainty associated with the state of the unresolved eddies which were discarded. Generative machine learning models have demonstrated exceptional ability to create realistic images obeying complex distributions learned directly from data. In this work, we leverage the generative machine learning approach to build a stochastic parameterization of the subgrid eddies which is able to sample many possible realizations of the missing physics forcing. The new stochastic models have shown excellent performance in predicting the missing physics term and have the promise to improve the simulation of turbulence when implemented online in the idealized ocean model.

1 Introduction

Mesoscale eddies, with a horizontal scale roughly equal to the Rossby deformation radius, play a crucial role in ocean circulation. Mesoscale eddies carry most of the kinetic energy in the ocean and account for a substantial part of the transport of momentum, heat, and salt (Vallis, 2017). The dynamics of mesoscale eddies involve a variety of complex physical processes: potential to kinetic energy conversion, upscale energy transfer, upgradient fluxes, sharpening of jet currents, along-isopycnal mixing and bolus advection. Primitive equations can potentially capture all these processes if all the relevant spatial scales of motion are directly resolved on the computational grid. However, direct simulation of mesoscale eddies remains computationally expensive, especially in high latitudes where the deformation radius decreases (Hewitt et al., 2020).

Modern global ocean models have an eddy-permitting resolution (around $1/4^\circ$, Haarsma et al. (2016)), such that the largest mesoscale eddies are resolved but smaller ones are not; therefore the effect of these smaller unresolved (subgrid) mesoscale eddies is missing and needs to be parameterized. A range of grid resolutions where a phys-

ical process is partially (but not fully) resolved is often referred to as the gray zone (Berner et al., 2017; Christensen & Zanna, 2022). Traditional methods to parameterize mesoscale eddies (Redi, 1982; Gent & McWilliams, 1990) were designed to describe their mean effect on the large-scale flow. These parameterizations are suitable for ocean models with a very coarse horizontal resolution, where there is an approximate scale separation between the grid step and the size of mesoscale eddies, but not for the gray zone.

The "Large eddy simulation" approach (LES, Fox-Kemper and Menemenlis (2008); Sagaut (2006)) is a technique to build a mesoscale eddy parameterization in the gray zone. The LES framework introduces a spatial filtering (and coarse-graining) operator which splits the flow into resolved and subgrid components. The filter mimics the effect of finite resolution and its width is proportional to the grid step of the coarse model. The effect of subgrid eddies on the resolved flow is referred to as a subgrid forcing and is diagnosed from the output of the high-resolution model by applying the spatial filter to the governing equations. A subgrid model or parameterization is a model which relates the subgrid forcing to the resolved flow. In recent years many new mesoscale eddy parameterizations were proposed to better capture the effects of mesoscale eddies in the gray zone using some heuristic (or empirical) physical arguments (Thuburn et al., 2014; Jansen & Held, 2014; Mana & Zanna, 2014; Zanna et al., 2017; Bachman et al., 2017; Pearson et al., 2017; Bachman et al., 2018; Jansen et al., 2019; Bachman, 2019; Grooms et al., 2015; Berloff, 2018; Juricke et al., 2020).

Machine Learning (ML) methods have recently gained traction as a new direction for developing subgrid eddy parameterizations in geophysics and turbulence (Rasp et al., 2018; Bolton & Zanna, 2019; Maulik et al., 2019; Beck et al., 2019; Yuval & O’Gorman, 2020; Guan, Chattopadhyay, et al., 2022; Beucler et al., 2021; Shamekh et al., 2022; Wang et al., 2022). ML parameterizations capture the effect of subgrid eddies on the resolved flow by *training* a model in a data-driven fashion. The most popular approach to train ML subgrid models is to minimize the mean squared error (MSE) between their output and a subgrid forcing obtained by reducing the resolution of a high-resolution model via filtering and coarse-graining (Bolton & Zanna, 2019). Such models typically have excellent *offline* performance: they are able to accurately predict the subgrid forcing. However, the ultimate goal of subgrid parameterizations is to improve *online* performance, once the parameterization is included into the coarse ocean model and the model is integrated for a long time. The coarse parameterized model should then reproduce the statistical properties of the coarse-grained high-resolution model (Sagaut, 2006). Recent work has shown that the offline and online performance of subgrid parameterizations correlate poorly (Ross et al., 2023): models trained with the offline MSE loss may be unstable when applied online (Beck et al., 2019; Maulik et al., 2019) and physically-based parameterizations have very low offline MSE but perform reasonably well online (Ross et al., 2023). Several approaches have been proposed to improve ML parameterizations. Kochkov et al. (2021) and Frezat et al. (2022) proposed an online training procedure that improves numerical stability properties but requires a differential model and has a considerable computational cost. Guan, Chattopadhyay, et al. (2022) suggested gradually enlarging the training dataset until the rare events in subgrid forcing are well captured. In Guan, Subel, et al. (2022) the MSE loss function was modified with an additional constraint involving energy exchange. Frezat et al. (2021); Guan, Subel, et al. (2022); Pawar et al. (2022) proposed to account for physical invariances of subgrid forcing.

Conventional subgrid parameterizations are deterministic and predict a single subgrid forcing for a given input (Berner et al., 2017), which represents the mean or most likely prediction given the resolved flow. However, many possible states of the subgrid eddies are typically consistent with a given resolved flow, so there is inherent uncertainty in the subgrid fluxes (Gerard, 2007; Berner et al., 2017; Christensen &

Zanna, 2022). Quantifying this uncertainty requires characterizing the distribution of the subgrid forcing, conditioned on the resolved variables. The stochastic ML model of Guillaumin and Zanna (2021) performs uncertainty quantification by estimating the pointwise conditional mean and conditional variance of the subgrid forcing, but does not take into account spatial correlations.

Subgrid models incorporating uncertainty quantification (UQ) can be used to build *stochastic* parameterizations, where the subgrid forcing is random. Stochastic parameterizations are widely used in climate models and have been shown to improve the mean state and variability (Palmer, 2000; Berner et al., 2012, 2017; Christensen et al., 2017; Juricke et al., 2017). The two simplest stochastic parameterizations are Stochastically perturbed parameterization tendency (SPPT, Buizza et al. (1999); Andrejczuk et al. (2016); Subramanian et al. (2019)) which multiplies a deterministic subgrid model by a random number with unit mean and non-zero spread and Stochastic kinetic energy backscatter scheme (SKEBS, Berner et al. (2009); Storto and Andriopoulos (2021)) which introduces additive stochastic forcing. The effect of stochastic parameterizations on online performance depends in a complex way on the associated UQ model. There is sensitivity to spatial (Grooms et al., 2015) and temporal (Wilks, 2005; Arnold et al., 2013; Schumann, 1995; Berner et al., 2009) correlations of stochastic forcing, its non-Gaussian distribution (Mana & Zanna, 2014; Zanna et al., 2017) and its dependence on the resolved flow (multiplicative noise, Sura et al. (2005); Arnold et al. (2013); Zacharuk et al. (2018)).

In this work, we propose to leverage two powerful uncertainty-quantification ML frameworks to data-driven subgrid parameterization of mesoscale eddies: variational autoencoder (VAEs, Kingma and Welling (2013)) and generative adversarial networks (GANs, Goodfellow et al. (2014)). These frameworks provide a data-driven characterization of the conditional distribution of the subgrid forcing given the resolved flow. The resulting ML models are *generative*, meaning that they allow us to sample from the conditional distribution, and can be therefore directly deployed as stochastic parameterizations. Our proposed ML models do not contain a-priori assumptions about the structure of the statistical model. These ML models can therefore potentially capture any statistically significant properties of the subgrid forcing such as the spatial correlation of stochastic residuals, dependence on the resolved flow, or probability distribution (Adler & Öktem, 2018; Gagne et al., 2020; Alcalá & Timofeyev, 2021; B. T. Nadiga et al., 2022). In addition, generative models can be trained and tested using the same datasets, as MSE-based ML models.

We implement our generative models in an idealized ocean simulation and evaluate them both offline and online. Our offline analysis shows that the generative models provide a flow-dependent prediction of uncertainty. The resulting stochastic residuals are correlated in space and reproduce stochastic backscatter (Leslie & Quarini, 1979; Chasnov, 1991; Frederiksen & Davies, 1997) in the correct band of scales. Additionally, generative models accurately simulate large-scale kinetic energy backscatter (Thuburn et al., 2014; Jansen & Held, 2014) and properly energize the flow. Our online analysis shows that the generative models have better numerical stability and metrics than the baseline ML model in Guillaumin and Zanna (2021) at coarse resolutions.

2 Idealized ocean model and subgrid eddy forcing

In this section, we describe an idealized numerical ocean model based on quasi-geostrophic (QG) equations of layered fluid written in Python (pyqg, Abernathey et al. (2022)), see Figure 1. The configuration of the QG model and the corresponding definition of subgrid forcing are similar to those in Ross et al. (2023). We use this model to perform offline and online evaluation of the proposed methodology to build subgrid parameterization for a range of resolutions.

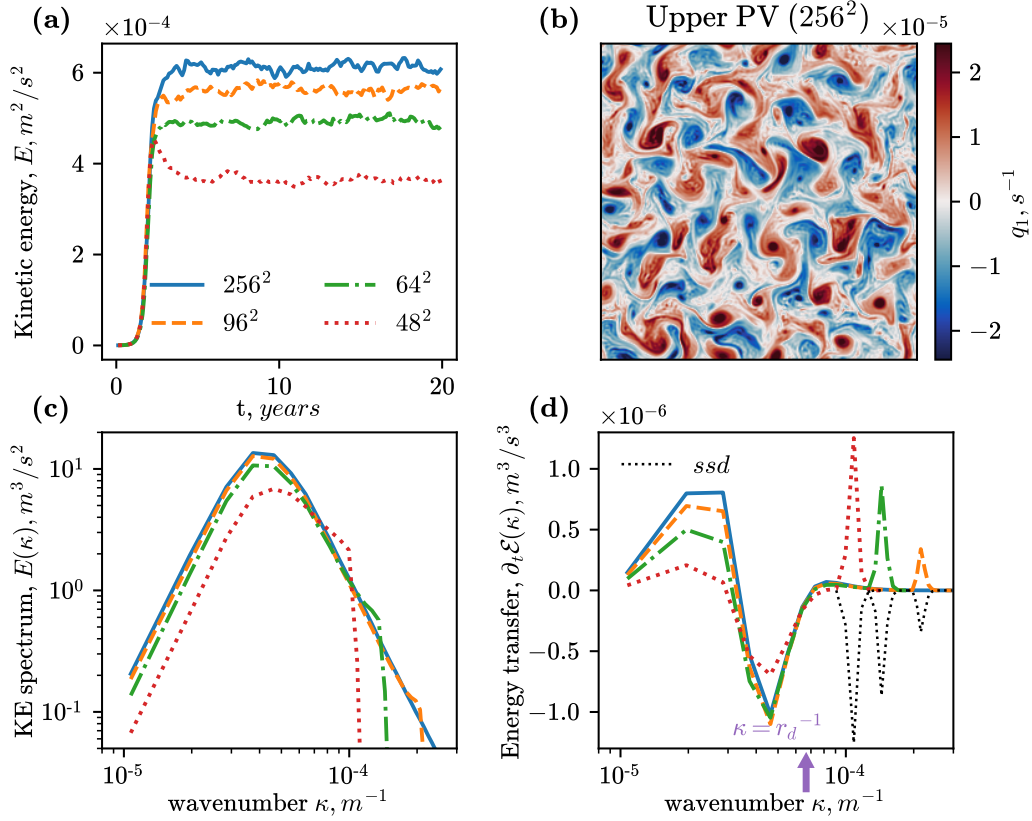


Figure 1. Reference simulations at four different resolutions: (a) kinetic energy (Eq. (3)) as a function of time, (b) snapshot of the potential vorticity in the model with the finest resolution, (c) the spectral density of kinetic energy normalized as $E = \int E(\kappa) d\kappa$, (d) total energy transfer from nonlinear advection (see Eq. (E3)) normalized as $\frac{\partial}{\partial t} \mathcal{E} = \int \partial_t \mathcal{E}(\kappa) d\kappa$. Coarse models fail to reproduce the energy cycle when their resolution is insufficient to resolve the deformation radius $\kappa = r_d^{-1}$ (see arrow). Black dotted lines in panel (d) show the dissipation model (*ssd*), which causes a spurious forward energy cascade (spikes in energy transfer).

2.1 Governing equations

We solve numerically the QG equations for potential vorticity (PV) anomalies relative to the mean flow given by a prescribed vertical shear that plays the role of external forcing driving turbulence.

The two-layer QG equations in Cartesian coordinates (x is zonal, y is meridional) are:

$$\partial_t q_m + \nabla \cdot (\mathbf{u}_m q_m) + \beta_m \partial_x \psi_m + U_m \partial_x q_m = -\delta_{m,2} r_{ek} \nabla^2 \psi_m + \text{ssd} \circ q_m, \quad (1)$$

$$q_m = \nabla^2 \psi_m + (-1)^m \frac{f_0^2}{g' H_m} (\psi_1 - \psi_2), \quad m \in \{1, 2\} \quad (2)$$

where m is the index of the fluid layer (1 for the upper layer and 2 for the lower layer); q_m is the potential vorticity (PV) which is conserved on Lagrangian trajectories in absence of forcing and dissipation; ψ_m is the streamfunction, related to velocity as $\mathbf{u}_m = (u_m, v_m) = (-\partial_y \psi_m, \partial_x \psi_m)$; U_m is the prescribed mean zonal flow (in the x direction); $\beta_m = \beta + (-1)^{m+1} \frac{f_0^2}{g' H_m} (U_1 - U_2)$ is the meridional gradient of potential vorticity due to differential rotation (in β -plane approximation) and prescribed mean flow; r_{ek} is the bottom drag coefficient; $\delta_{m,2}$ is a Kronecker delta which indicates that drag is applied only to the lower layer; f_0 is the reference Coriolis frequency; g' is the reduced gravity and H_m is the fluid layer thickness, $H = H_1 + H_2$ is the total depth; $\nabla = (\partial_x, \partial_y)$ is a horizontal Nabla operator, where ∂_x, ∂_y are partial derivatives w.r.t. x, y . The numerical schemes and how the small-scale dissipation (ssd) is applied to the governing equations are described in Appendix A. The kinetic and total energy per unit mass are respectively given by:

$$E = \frac{1}{2H} \sum_{m=1}^2 H_m \langle |\mathbf{u}_m|^2 \rangle \quad (3)$$

$$\mathcal{E} = -\frac{1}{2H} \sum_{m=1}^2 H_m \langle \psi_m q_m \rangle \quad (4)$$

where $\langle \cdot \rangle$ is 2D spatial averaging. The QG system, described by Eq. (1) and (2) is initially perturbed from rest with random noise in the upper PV field, with a subsequent evolution over the next 2–5 years exhibiting a transition to turbulence. The initial random perturbations are limited to the range of scales of the coarsest model, and it allows to simulate similar energy growth in the transition from laminar to turbulent regimes at different grid resolutions, see Figure 1(a). Model parameters are given in Table 1 and correspond to the "eddy" configuration in Ross et al. (2023).

Mesoscale eddies emerge on a spatial scale determined by the deformation radius $r_d = \frac{g'}{f_0^2} \frac{H_1 H_2}{H}$ (Salmon, 1980; Vallis, 2017), denoted by the arrow in Figure 1. We choose the resolution of the reference simulation (256^2) in order to accurately reproduce the spectral energy transfer. Coarse-resolution models do not resolve the deformation radius properly and fail to reproduce various statistical characteristics (Hallberg, 2013; Hewitt et al., 2020), including kinetic energy (KE), spectrum of KE and energy transfer. In this work, we aim to improve the simulation of turbulence in coarse models by incorporating a subgrid parameterization model, which compensates for the missing physics.

2.2 Filtered equations

In this section, we derive the governing equations for the coarse model which follows the trajectory of the filtered and coarsegrained high-resolution simulation. These equations contain a new term that describes the interaction with unresolved eddies, the term that is not available at the coarse resolution and needs to be parameterized.

Table 1. Parameters of the quasi-geostrophic model in online simulations (eddy configuration from Ross et al. (2023)).

| Common parameters | | | |
|---|--|--------------------------|--------------------------|
| Integration time | 20 years | | |
| Ensemble size | 10 runs | | |
| Domain size ($L \times W$) | 1000km \times 1000km | | |
| Boundary conditions | periodic | | |
| Ocean depth ($H = H_1 + H_2$) | 2500m | | |
| Upper layer thickness (H_1) | 500m | | |
| Bottom drag (r_{ek}) | $5.787 \cdot 10^{-7} \text{s}^{-1}$ | | |
| Differential rotation (β) | $1.5 \cdot 10^{-11} (\text{m s})^{-1}$ | | |
| Deformation radius ($r_d = \frac{g'}{f_0^2} \frac{H_1 H_2}{H}$) | 15km | | |
| Mean flow (U_1, U_2) | (0.025m/s, 0m/s) | | |
| Velocity scale ($\sqrt{2E}$) | $\approx 0.035 \text{m/s}$ | | |
| Grid parameters | resolution | grid step (Δx) | time step (Δt) |
| High resolution | 256 \times 256 | 3.9km | 1hour |
| Coarse models | 96 \times 96 | 10.4km | 2hour |
| | 64 \times 64 | 15.6km | 4hour |
| | 48 \times 48 | 20.8km | 1,2,4,8 hour |

We follow the Large eddy simulation (LES, Sagaut (2006)) approach to split the prognostic variables (ϕ) into resolved ($\bar{\phi}$) and subgrid (ϕ') components by applying a spatial convolutional filter with kernel $G(\mathbf{y})$ such that

$$\phi = \bar{\phi} + \phi', \quad (5)$$

$$\bar{\phi}(\mathbf{x}) = \int G(\mathbf{y} - \mathbf{x}) \phi(\mathbf{y}) d\mathbf{y} \quad \text{with} \quad \int G(\mathbf{y}) d\mathbf{y} = 1. \quad (6)$$

We use two spectral filters from Ross et al. (2023): one filter is a combination of a cut-off and a model filter ("Sharp"), the other is a combination of a cut-off and a Gaussian filter (we denote it as "Gaussian"). Precise definitions are provided in Appendix B.

Applying the filter $G(\mathbf{y})$ to the governing equations (1), (2), we obtain a set of governing equations for the filtered solution:

$$\partial_t \bar{q}_m + \nabla \cdot (\bar{\mathbf{u}}_m \bar{q}_m) + \beta_m \partial_x \bar{\psi}_m + U_m \partial_x \bar{q}_m = -\delta_{m,2} r_{ek} \nabla^2 \bar{\psi}_m + S + \text{ssd} \circ \bar{q}_m, \quad (7)$$

$$\bar{q}_m = \nabla^2 \bar{\psi}_m + (-1)^m \frac{f_0^2}{g' H_m} (\bar{\psi}_1 - \bar{\psi}_2), \quad m \in \{1, 2\}. \quad (8)$$

S is the additional subgrid forcing produced by the unresolved eddies on the resolved scales,

$$S = \nabla \cdot (\bar{\mathbf{u}} \bar{q} - \bar{\mathbf{u}} \bar{q}), \quad (9)$$

which needs to be parameterized. We will omit the index m for the subgrid forcing and related variables to simplify notation. The dissipation term ssd on a coarse grid in Eq. (7) is added a-posteriori to ensure the numerical stability of the simulations. In deriving Eq. (7), we used commutativity between derivatives and spatial filtering, which holds for spectral numerical schemes and spectral filters (Ghosal, 1996). Both subgrid forcing and numerical advection scheme are formulated in flux form, so we include numerical approximation errors into the definition of subgrid forcing (Ghosal, 1996; Chow & Moin, 2003; Gullbrand & Chow, 2003).

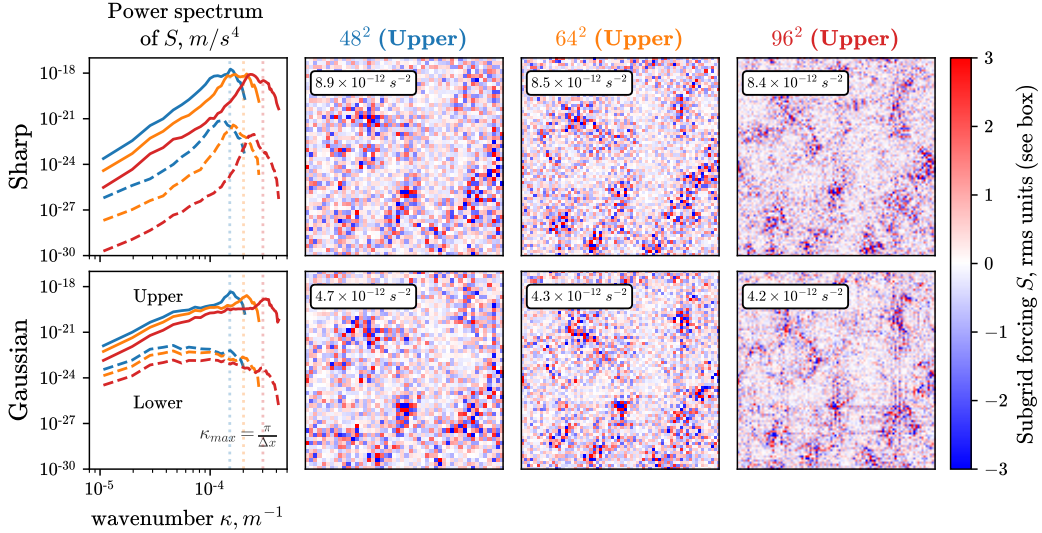


Figure 2. Subgrid forcing (S , Eq. (9)) at different resolutions diagnosed using a Sharp filter (top row), or a Gaussian filter (bottom row). Left: power spectrum of S for the upper fluid layer (solid lines) and lower fluid layer (dashed lines). Colors: 48^2 (blue), 64^2 (orange), 96^2 (red). Vertical lines show grid cut-off for coarse mesh $\kappa_{max} = \pi/\Delta x$. Right: Snapshots of S at three different resolutions for the upper layer.

2.3 Subgrid forcing dataset

The solution to the governing equations (1), (2) for the high-resolution model is denoted by q . The filtered quantities (\bar{q}) are defined on a coarse mesh.

The dataset to train ML subgrid parameterization models is obtained as follows. We integrate the governing equations in time for 10 years at high resolution (256^2) with time step 1 hour and save snapshots every 1000 hours, for a total of 86 snapshots. The training dataset consists of 250 runs, each corresponding to a different random initial condition, for a total of 21500 snapshots. The validation and testing datasets consist of 25 runs each. For each coarse resolution (48^2 , 64^2 , 96^2), we compute a filtered solution represented on a coarse mesh (\bar{q} , $\bar{\mathbf{u}}$) and subgrid forcing (Eq. (9)) using Sharp or Gaussian filters. The spectral content of the resulting subgrid forcing greatly depends on the scale selectivity of the filter, see Figure 2.

3 Data-driven stochastic subgrid models

In this section, we introduce a probabilistic approach for the prediction of subgrid forcing, which can be used to build data-driven stochastic parameterizations.

Conventional subgrid parameterizations establish a functional relationship between the subgrid forcing (S , Eq. (9)) and the resolved flow (\bar{q}) in the form of $S \approx \tilde{S}(\bar{q})$. Such parameterizations are typically deterministic; they produce a single prediction for a given input. However, there is inherent uncertainty in the prediction of subgrid forcing, because many possible states of the subgrid eddies are consistent with a given resolved flow. Therefore, we propose to instead generate a *probabilistic* prediction, by attempting to sample from the conditional distribution of the subgrid forcing given the coarse-grained flow ($S \sim \rho(S|\bar{q})$).

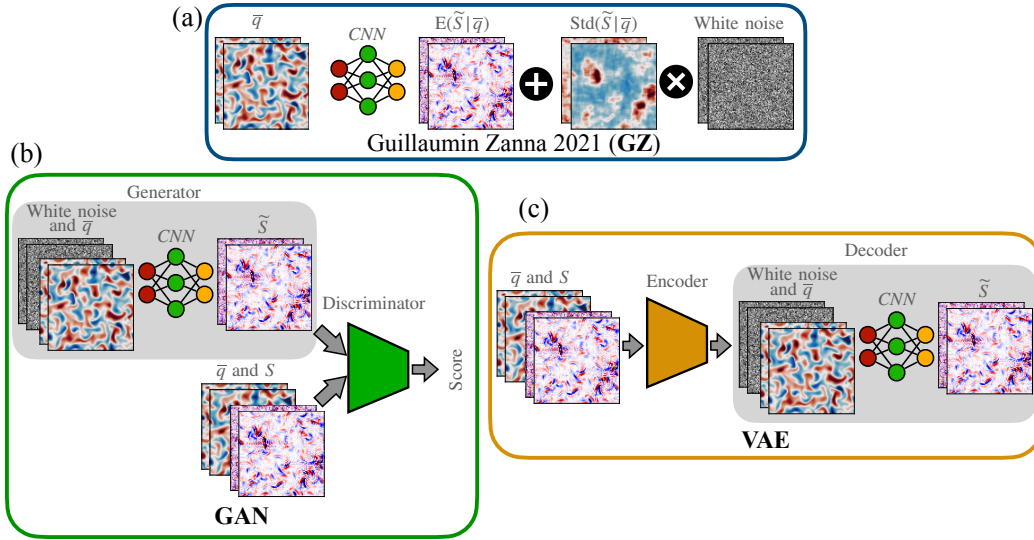


Figure 3. Schematic of three stochastic subgrid models attempting to sample from conditional distribution $\rho(S|\bar{q})$: (a) GZ model, (b) GAN model and (c) VAE model. GZ model predicts uncorrelated stochastic residuals, but generative models (GAN and VAE) transform white noise using a mapping learned directly from data (gray-shaded box). Discriminator and Encoder are supplementary networks that allow training of the mapping but are not required for subgrid forcing prediction.

In order to generate a probabilistic prediction of the subgrid forcing, we propose to apply a generative ML framework, where samples from a desired distribution are obtained by transforming white noise using a mapping learned directly from data (Kingma & Welling, 2013; Goodfellow et al., 2014). We design and compare three different approaches, depicted in Figure 3, to learn this transformation: (a) A model based on Guillaumin and Zanna (2021), which predicts the pointwise mean and pointwise standard deviation of the conditional distribution of the subgrid forcing. (b) A generative adversarial network (GAN), consisting of a *generator* that generates subgrid-forcing samples by trying to fool a *discriminator*, trained to distinguish between these samples and the true high-resolution data. (c) A variational autoencoder (VAE) consisting of an *encoder*, which maps the input signal to a latent space, and a *decoder*, which decodes the latent variables to produce subgrid-forcing samples. The remainder of this section provides a more detailed description of each approach.

3.1 Guillaumin and Zanna model (GZ)

Guillaumin and Zanna (2021) presented a probabilistic ML parameterization, where the mean and variance of the subgrid forcing are estimated at every grid point using a neural network. The original formulation in (Guillaumin & Zanna, 2021) minimizes an i.i.d. Gaussian likelihood cost function to optimize the parameters of the network. Here, we propose an alternative training procedure, which we have found to be more efficient. Following the approach of Adler and Öktem (2018), we estimate the pointwise means and variances sequentially.

First, we estimate the conditional mean at each grid point by minimizing the MSE loss function:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2n^2} \|S - \tilde{S}_\theta^{\text{mean}}(\bar{q})\|_2^2, \quad (10)$$

where $\tilde{S}_\theta^{\text{mean}}(\bar{q})$ is the output of a neural network with parameters denoted by θ , which receives \bar{q} as an input. $S, \tilde{S}_\theta^{\text{mean}}, \bar{q} \in \mathbb{R}^{2 \times n \times n}$ are tensors representing two layers of fluid, each layer having $n \times n$ points. The norm in the cost function is the ℓ_2 norm of the vectorized tensor, which for the vector of length D is $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \dots + x_D^2}$. The loss function is minimized over a training set consisting of samples of the resolved flow \bar{q} and the corresponding high-resolution forcing S obtained as described in Section 2.3. Minimization of \mathcal{L}_{MSE} yields an optimal set of parameters θ^* and a corresponding model which we denote as $\tilde{S}^{\text{mean}}(\bar{q})$.

Second, we estimate the conditional variance at each grid point, based on the residual of the conditional-mean estimate $r = S - \tilde{S}^{\text{mean}}(\bar{q})$. To this end, we minimize the cost function

$$\mathcal{L}_{\text{VAR}} = \frac{1}{2n^2} \|r^2 - \tilde{S}_\phi^{\text{var}}(\bar{q})\|_2^2, \quad (11)$$

where $\tilde{S}_\phi^{\text{var}}(\bar{q})$ is the output of a neural network with parameters denoted by ϕ , which receives \bar{q} as an input. The final layer of the network is a softplus activation function $\ln(1 + e^x)$ to ensure that the variance estimates are nonnegative. The loss function is minimized over the training set fixing the residual r . The resulting model is denoted by $\tilde{S}^{\text{var}}(\bar{q})$. Additional training information is given in Appendix C.

The conditional-mean and conditional-variance models are used to implement a stochastic parameterization with white noise, as follows (see Figure 3(a)):

$$\tilde{S}(z, \bar{q}) = \tilde{S}^{\text{mean}}(\bar{q}) + (\tilde{S}^{\text{var}}(\bar{q}))^{1/2} \cdot z, \quad (12)$$

where $z \in \mathbb{R}^{2 \times n \times n}$ is sampled from a standard normal distribution.

3.2 Generative adversarial network model (GAN)

We propose to leverage the framework of generative adversarial networks (GANs) to build a probabilistic model (Goodfellow et al., 2014), which generates samples from the distribution of possible subgrid forcings (S) at a given resolved flow (\bar{q}) denoted as $\rho(S|\bar{q})$, where both variables are considered as 3D fields, $S, \bar{q} \in \mathbb{R}^{2 \times n \times n}$. The mentioned distribution is defined implicitly by the dataset of pairs of S and \bar{q} . The GAN framework consists of two networks, generator and discriminator, playing an adversarial game: the generator attempts to *fool* the discriminator, which is trained to discriminate between the output of the generator and actual data sampled from a desired distribution.

Sampling from the conditional distribution is possible with the conditional GAN model (cGAN, Mirza and Osindero (2014)), which informs both networks with the conditional variable. Specifically, the generator transforms the latent noise variable z and PV field to the subgrid forcing, see Figure 3(b):

$$\tilde{S} = G(z, \bar{q}). \quad (13)$$

The discriminator returns a score given a pair of subgrid forcing and PV field denoted as $D(S, \bar{q})$. There are many options to define the adversarial loss function (Lucic et al., 2018). We leverage a popular approach of Wasserstein GAN (WGAN, Arjovsky et al. (2017)) with the following optimization problem:

$$\min_G \max_D \mathbb{E} \left[D(S, \bar{q}) - D(G(z, \bar{q}), \bar{q}) \right], \quad (14)$$

where \mathbb{E} is the mathematical expectation over the training samples. The discriminator D is optimized to estimate the Wasserstein-1 distance between the distributions $\rho(S|\bar{q})$ and $\rho(\tilde{S}|\bar{q})$, while the generator learns the true distribution by minimizing this distance.

Solving the optimization problem (14) may lead to the mode collapse phenomenon when the generator ignores the latent variable z : for every coarse field \bar{q} the model

may produce a single fixed subgrid forcing (Isola et al., 2017; Ohayon et al., 2021; Mao et al., 2019; Yang et al., 2019). To overcome mode collapse, we apply a technique proposed by Adler and Öktem (2018): feeding multiple generator outputs \tilde{S} to the discriminator for a given input \bar{q} . Identical outputs are readily detected and penalized by the discriminator. We provide additional details in Appendix C.

Once trained, the GAN generator (13) can be used as a stochastic parameterization by sampling the latent variable $z \in \mathbb{R}^{2 \times n \times n}$ from a standard normal distribution.

3.3 Variational autoencoder model (VAE)

As an alternative to the GAN framework, we propose leveraging the variational autoencoder (VAE, Kingma and Welling (2013)) to sample from the conditional distribution $\rho(S|\bar{q})$. The VAE framework consists of two networks: the encoder and the decoder. The encoder produces a latent representation and the decoder reconstructs the subgrid forcing from this representation. A regularization term constrains the latent vector to be close to a simple distribution, chosen a priori.

The conditional VAE (cVAE) is obtained by feeding a conditional variable to the encoder and decoder (Sohn et al., 2015; Doersch, 2016; Zhang et al., 2016; Mishra et al., 2018; Pagnoni et al., 2018): the decoder maps the latent noise and conditional variable \bar{q} to the subgrid forcing, see Figure 3(c):

$$\tilde{S} \sim \rho_{\theta}(\tilde{S}|z, \bar{q}), \quad (15)$$

where free parameters are denoted as θ and we emphasize that the mapping is probabilistic. The probabilistic encoder with free parameters ϕ is denoted as $z \sim q_{\phi}(z|S, \bar{q})$. The encoder and decoder are trained jointly to maximize the lower bound of the likelihood of observing the training sample (also known as evidence lower bound, ELBO):

$$\ln \rho_{\theta}(S|\bar{q}) \geq \underbrace{\mathbb{E}_{q_{\phi}(z|S, \bar{q})} \ln \rho_{\theta}(S|z, \bar{q})}_{\text{reconstruction}} - \underbrace{D_{\text{KL}}(q_{\phi}(z|S, \bar{q})||\rho(z))}_{\text{regularization}} = -\mathcal{L}_{\text{VAE}}, \quad (16)$$

where $D_{\text{KL}}(p(x), q(x)) = \mathbb{E}_{p(x)} \ln \frac{p(x)}{q(x)}$ is Kullback–Leibler divergence, a measure of the difference between two distributions. The reconstruction term encourages the encoder to seek an accurate latent representation of the subgrid forcing and encourages the decoder to assign a high probability to the training samples. The regularization term constrains the encoder to be close to the prior distribution $\rho(z)$. We parameterize all probabilities with Gaussian distributions and replace the mathematical expectation with one sample from the encoder (reparameterization trick, Kingma and Welling (2013)). The resulting loss function is equivalent to a regularized MSE as explained in more detail in Appendix C.

The mean channel of the Gaussian decoder (15) can be used as a stochastic parameterization by sampling the latent variable $z \in \mathbb{R}^{2 \times n \times n}$ from a standard normal distribution.

4 Offline analysis of stochastic subgrid models

In this section, we perform an offline evaluation of the stochastic subgrid models described in Section 3 using the dataset described in Section 2.3. We show spatial maps and spectra of the predicted subgrid forcing. We propose metrics for the evaluation of the predicted subgrid forcing and stochastic residuals and compare them for a range of resolutions.

For every combination of filter (Sharp and Gaussian) and resolution of coarse mesh (48^2 , 64^2 , 96^2), we train three machine learning models: GZ, GAN and VAE, for

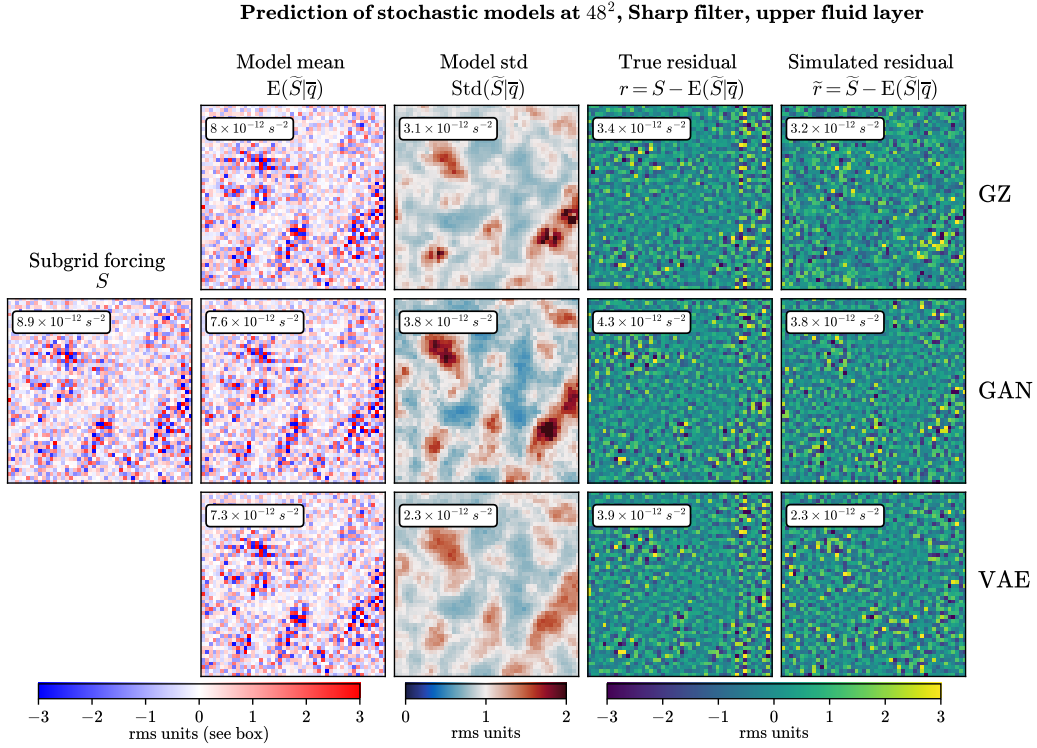


Figure 4. Prediction of subgrid forcing (S) by stochastic models at a fixed conditional variable \bar{q} on the testing dataset: GZ in upper row, GAN in middle row and VAE in lower row. "Model mean" is the deterministic prediction, and "Simulated residual" is the stochastic part of the model, which should be similar to "True residual". "Model std" is the standard deviation of the subgrid forcing prediction. The root mean square (rms) value of each field is shown in the box.

a total of 18 different models of subgrid forcing. The baseline deterministic subgrid model is trained with the MSE loss function and it is referred to as "MSE" (we simply take the mean channel of GZ model). Following Kochkov et al. (2021), every model was trained five times with different random seeds. Three training instances failed and were excluded from the subsequent analysis: 2 realizations of VAE models at resolution 64^2 experienced the posterior collapse problem (zero spread, Dai et al. (2020)), and one realization of GZ model at resolution 48^2 had a large generalization error. In this section, we show results for Sharp filter, and similar plots for Gaussian filter are shown in Appendix D.

4.1 Analysis of stochastic predictions

In this section, we compare stochastic predictions of subgrid forcing to the true subgrid forcing. We suggest to split the stochastic prediction into the deterministic part and the stochastic residual. We define the deterministic part as a mean prediction of subgrid forcing at a fixed resolved field \bar{q} – it is conditional mean denoted as $E(\tilde{S}|\bar{q})$. The deterministic part of the GZ model is given by the mean channel $\tilde{S}^{\text{mean}}(\bar{q})$, and for GAN and VAE models we estimate it by sampling 1000 realizations of the latent vector (Adler & Öktem, 2018). The predicted stochastic residuals ($\tilde{r} = \tilde{S} - E(\tilde{S}|\bar{q})$) should be compared to the true residuals ($r = S - E(\tilde{S}|\bar{q})$), and for an accurate stochastic model they should be statistically similar, see Wilks (2005); Arnold et al. (2013); Agarwal et al. (2021); Berner et al. (2009); Mana and Zanna (2014); Shutts and Palmer (2007); Guillaumin and Zanna (2021); Gagne et al. (2020).

In Figure 4 we show predictions of the stochastic subgrid models. The deterministic part ($E(\tilde{S}|\bar{q})$) is similar for three stochastic models. The predicted stochastic residual (rightmost column) for GZ model looks like uncorrelated spatial white noise in contrast to the true residual. The predicted residuals for the other two models (GAN and VAE) are more visually similar to the true one. The pointwise standard deviation is a measure of the local uncertainty in the deterministic prediction and can be related to the second moment of residuals as:

$$\text{Std}(\tilde{S}|\bar{q}) = \sqrt{E\left(\left(\tilde{S} - E(\tilde{S}|\bar{q})\right)^2 \middle| \bar{q}\right)} = \sqrt{E(\tilde{r}^2|\bar{q})}.$$

It is directly accessible for the GZ model, and for GAN and VAE models it can be estimated similarly to the conditional mean. The standard deviation fields have similar spatial structure for all three stochastic models.

We use the spatial power spectrum to analyze the spatial correlation. In Figure 5(a) we show the power spectrum of stochastic residuals. The true residuals are concentrated near the grid cut-off (Nyquist frequency, $\kappa_{max} = \pi/\Delta x$) and near the spatial frequency of *ssd* filter ($\kappa = 0.65\pi/\Delta x$). The GZ model does not reproduce the two-hill shape of the power spectrum of residuals. The GAN model accurately reproduces the power spectrum of residuals and improves the power spectrum of subgrid forcing (Figure 5(b)) compared to the deterministic and stochastic baselines (MSE, GZ). Note that accurate prediction of the power spectrum of subgrid forcing is a challenging task for deterministic models (Guan, Subel, et al., 2022) because optimization of the mean squared error leads to the loss of details in small scales (Isola et al., 2017). The VAE model predicts the correct shape of the spectrum of residuals, but the total variance of residuals is underestimated. The power spectrum of subgrid forcing for the VAE model is also lower compared to the other models. We explain it by the well-known issue of VAE architecture to predict oversmoothed images (Takida et al., 2022).

An important property of subgrid forcing in QG turbulence is an ability to energize turbulence on a coarse grid, i.e. kinetic energy backscatter (Jansen & Held, 2014). There are two popular approaches to simulate backscatter: stochastic residuals

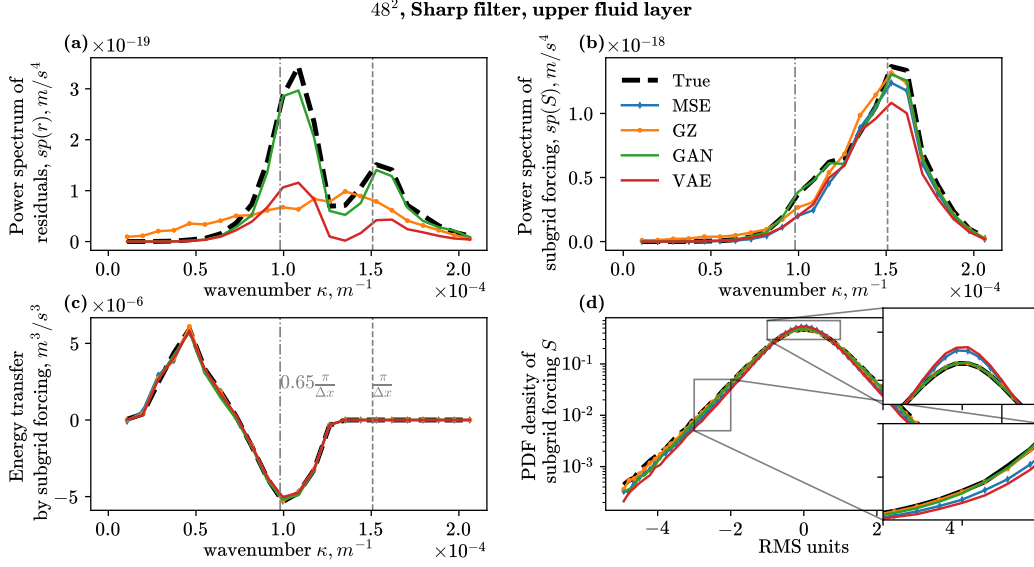


Figure 5. Offline analysis of stochastic subgrid models (GZ, GAN, VAE): (a) power spectrum of stochastic residuals and (b) subgrid forcing; (c) energy transfer (see Eq. (E4)) and (d) marginal PDF of subgrid forcing. MSE is the deterministic subgrid model given by the mean channel of GZ model.

near the grid scale (Leslie & Quarini, 1979; Chasnov, 1991; Schumann, 1995; Frederiksen & Davies, 1997; Grooms et al., 2015) and mean energy injection in large scales (Kraichnan, 1976; Frederiksen et al., 2003; Graham & Ringler, 2013; Thuburn et al., 2014; Jansen & Held, 2014; Juricke et al., 2020). These two types of backscatter result from physical processes of a very different nature: stochastic backscatter simulates the loss of information about unresolved degrees of freedom but energy injection in large scales compensates for the unresolved inverse energy cascade. All the stochastic models are accurate in predicting large-scale energy injection (Figure 5(c)), and GAN model is the best in predicting stochastic residuals near the grid scale.

Marginal PDF of subgrid forcing is often used to evaluate subgrid models (Pawar et al., 2020; Maulik & San, 2017). Both GZ and GAN models improve this PDF in the high-probability region and in the tails compared to the baseline MSE model, see Figure 5(d). The VAE model is similar to the baseline MSE in this characteristic.

4.2 Quantitative offline analysis and metrics

Above we presented a qualitative analysis of the stochastic subgrid models, and here we propose metrics for their quantitative evaluation. We consider three classes of metrics, which demonstrate: the quality of the subgrid forcing, its deterministic part and stochastic residuals, see Table 2. We include spectral metrics for the subgrid forcing and residuals (\mathcal{L}_S and \mathcal{L}_r) in order to evaluate to what extent the models capture the corresponding spatial structure.

In Figure 6 we report the evaluation of the offline metrics for the different models. The upper row provides metrics on the test dataset with the same turbulence regime as the training set. We observe that the generative models (GAN and VAE) have slightly greater deterministic error ($\mathcal{L}_{\text{rmse}}$) compared to the model optimizing this metric directly (GZ and MSE). The GAN and GZ models correctly predict spread of

| Metric | $\mathcal{L}_{\text{rmse}}$ | \mathcal{L}_S | \mathcal{L}_r | σ_{spread}^2 |
|-----------------------|--|---|---|---------------------------------------|
| Expression | $\frac{\ S - E(\tilde{S} \bar{q})\ _2}{\ S\ _2}$ | $\frac{\ sp(S) - sp(\tilde{S})\ _2}{\ sp(S)\ _2}$ | $\frac{\ sp(r) - sp(\tilde{r})\ _2}{\ sp(r)\ _2}$ | $\frac{\ \tilde{r}\ _2^2}{\ r\ _2^2}$ |
| Optimal value | 0 | 0 | 0 | 1 |
| Unparameterized model | 1 | 1 | 1 | 0 |
| Quality of | deterministic part | full forcing | residuals | residuals |

Table 2. Metrics for offline analysis of stochastic subgrid model \tilde{S} : RMSE of deterministic part ($\mathcal{L}_{\text{rmse}}$), RMSE in spectrum of the full subgrid forcing (\mathcal{L}_S), RMSE in spectrum of stochastic residuals (\mathcal{L}_r) and spread of the samples of conditional distribution (σ_{spread}^2). We denote computation of isotropic power spectrum as $sp(\cdot)$.

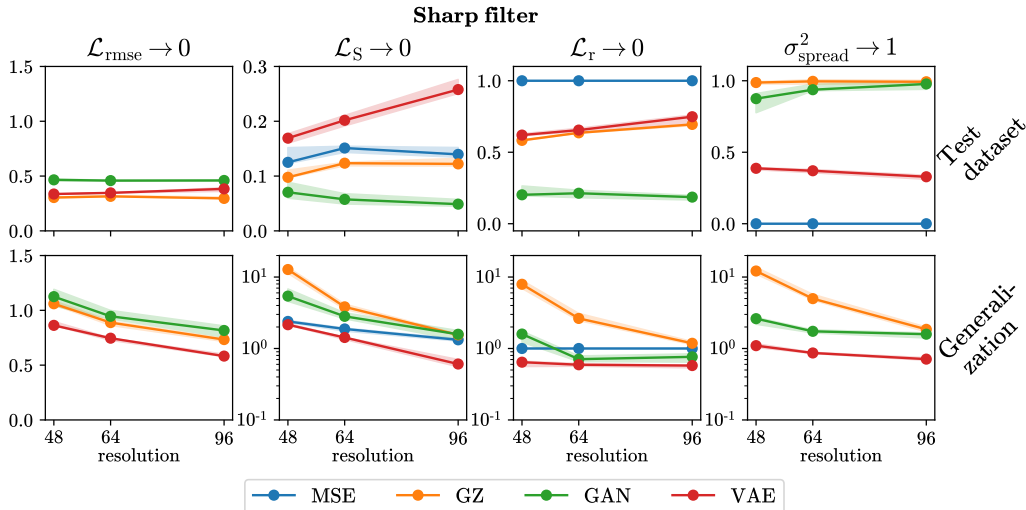


Figure 6. Offline metrics from Table 2 on testing dataset in the upper row and generalization to configuration with jets (Ross et al., 2023) in the lower row. Optimal values are given with arrows. Each model is trained 5 times with different random seeds. The shading area shows min-max values among training realizations, markers show median value.

stochastic residuals $\sigma_{\text{spread}}^2 \approx 1$, but the VAE model underestimates spread $\sigma_{\text{spread}}^2 \approx 0.35$. The GAN model clearly outperforms the rest in predicting the spectra of the subgrid forcing \mathcal{L}_S and the residuals \mathcal{L}_r . The VAE model on the contrary has high errors \mathcal{L}_S and \mathcal{L}_r because it predicts oversmoothed samples with reduced diversity.

In the lower row of Figure 6, we evaluate the generalization ability of the models by computing the offline metrics on dataset corresponding to a different turbulence regime, where flow is dominated by meandering jets, and which is therefore systematically different from the training data (see Ross et al. (2023) for description). GZ model considerably overestimates the spread of the residuals ($2 < \sigma_{\text{spread}}^2 < 10$), and it deteriorates the spectral metrics (\mathcal{L}_S and \mathcal{L}_r). Although the VAE model had various suboptimal metrics on the eddy turbulence configuration, it demonstrates the best generalization capabilities to the jet configuration: VAE model has reasonable spread $\sigma_{\text{spread}}^2 \approx 0.8$, and outperforms other models in the error of the deterministic prediction $\mathcal{L}_{\text{rmse}}$, the quality of the subgrid forcing \mathcal{L}_S and residuals \mathcal{L}_r . The GAN model generalizes better than GZ for most of the metrics, without reaching the performance of the VAE model. We observe similar results for the Gaussian filter, see Appendix D.

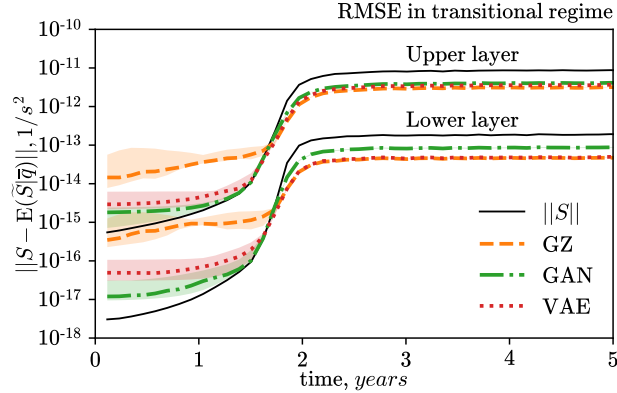


Figure 7. Offline analysis of RMSE of deterministic part in transitional regime ($t < 2$ years) on test dataset. Norm $\| \cdot \|$ is given per one grid point. Shading corresponds to different training realizations of the same model. Sharp filter, resolution 48^2 . Each line is given twice: for lower and upper fluid layer.

During the first few years of simulation, QG model undergoes a transition from a laminar to a turbulent flow regime. Generalization to the transitional regime is a difficult test for subgrid models (Frezat et al., 2021) because the subgrid forcing is a few orders of magnitude smaller compared to the developed turbulence regime. Although we include the transitional regime in the training set, the relative importance of these samples is small due to their small norm. As a result, all subgrid models have large errors compared to the norm of the subgrid forcing during the first few years of simulation ($t < 2$ years), see Figure 7. The generative models (GAN and VAE) demonstrate the best performance in the transitional regime: error is one order of magnitude smaller compared to GZ. In the next section, we show that generative models are also superior to the baseline in the online simulation of transitional flow.

5 Online simulations with subgrid models

In the previous section, we demonstrated the encouraging ability of generative models GAN and VAE to simulate various statistical characteristics of subgrid forcing. In this section, we evaluate the performance of trained subgrid models in online simulations. In more detail, we use the output of the subgrid model \tilde{S} to replace the true subgrid forcing S in the governing equation for the coarsegrained dynamics (7), and perform numerical time integration. Our goal is to study how the subgrid parameterizations impact the dynamics of mesoscale eddies in a statistical equilibrium regime.

Our online experiments are summarized in Table 1. Compared to the generation of the training data, we run experiments for twice as long (20 years). Before passing the subgrid forcing prediction into governing equation we subtract the spatial mean in each fluid layer to ensure the conservation of PV. Recall that we train 5 different models (differing only in the initialization of the weights) for every combination of resolution, filter and type of subgrid model. Each of these models is evaluated in an ensemble of 10 online runs, with different random initial conditions. The total number of runs is approximately 1200. The statistical characteristics of the turbulence are averaged over the 10 ensemble members (and the last 15 years if applicable). We provide the confidence bounds for every averaged statistic defined by the minimum, maximum and median values over 5 realizations of the training algorithm.

Among mentioned experiments, there were a few unstable simulations at resolution 96²: one run of the VAE model for Sharp filter, and 3 GAN models out of 5 training realizations for Gaussian filter. We exclude mentioned experiments from the analysis. In these experiments, an eddy emerges which is constantly amplified by the parameterization, and in a spectral space it corresponds to the overestimated energy injection on the largest scale. This effect is possible because we do not control the amplitude of the parameterization as it is usually done in energetically-consistent physical parameterizations of backscatter (Jansen & Held, 2014).

5.1 Metrics for online analysis

We compare the solution of the coarse parameterized model to the filtered and coarse-grained fields of the high-resolution model similarly to B. Nadiga and Livescu (2007); Beck et al. (2019); Frezat et al. (2022); Guan, Chattopadhyay, et al. (2022); Guan, Subel, et al. (2022).

Following Ross et al. (2023), we consider an error in PDFs of the turbulence fields. Define the Wasserstein distance between distributions as $\mathcal{W}_1(F_1, F_2) = \int |F_1(\xi) - F_2(\xi)| d\xi$, where F_1 and F_2 are cumulative distribution functions (CDF) of some variable ξ . In computing CDF, we aggregate spatial directions, 15 years of simulation, and 10 ensemble members. We consider 5 variables in place of ξ : potential vorticity (q_m), velocity (u_m and v_m), kinetic energy ($\frac{1}{2}|\mathbf{u}_m|^2$) and relative enstrophy ($\frac{1}{2}|\text{curl}(\mathbf{u}_m)|^2$), and each fluid layer is accounted independently. The online distributional metric between the coarse-grid model (F_{model}) and the filtered and coarse-grained high-resolution simulation (F_{hires}) is given by the average of normalized errors:

$$\mathcal{W}(\text{model}, \overline{\text{hires}}) = \frac{1}{10} \sum_{m=1}^2 \sum_{\xi \in \text{Vars}_m} \frac{\mathcal{W}_1(F_{\text{model}}(\xi), F_{\text{hires}}(\xi))}{\sqrt{\int \xi^2 dF_{\text{hires}}}}, \quad (17)$$

where $\text{Vars}_m = \{q_m, u_m, v_m, \frac{1}{2}|\mathbf{u}_m|^2, \frac{1}{2}|\text{curl}(\mathbf{u}_m)|^2\}$ and the normalization constant is the square root of the uncentered second moment.

An additional metric based on spectral characteristics is reported in Appendix E.

5.2 Sensitivity to the correlation time of latent variable

In order to leverage the proposed subgrid-forcing models in a stochastic parameterization, we sample the latent variable z independently at every time step (discrete white noise) similar to Zanna et al. (2017) and Guillaumin and Zanna (2021).

Following (Gagne et al., 2020), we also tested the sensitivity of the online simulation results to the correlation time of the latent variable. The time correlation is introduced with the autoregressive model of order one (AR1), which has covariance function $E(z(t)z(t+n\Delta t)) = (1 - \Delta t/\tau)^n$ (Schumann, 1995), where n denotes the number of time layers between two time moments, $\tau \geq \Delta t$ is correlation time, and at $\tau = \Delta t$ we restore the discrete white noise process. The online distributional metric (17) as a function of correlation time is reported in Figure 8. The optimal online metric corresponds to $\tau = \Delta t$, which justifies our method of sampling (white noise).

5.3 Results

In Figure 9 we show online simulations with subgrid models at the coarsest resolution. The unparameterized model ("lores") has underestimated kinetic energy (a) and underestimated KE spectrum in large scales (b). This is due to the poor representation of the inverse energy cascade on the coarse grid (c). The deterministic subgrid model (MSE) improves inverse energy cascade and KE spectrum in large scales, but

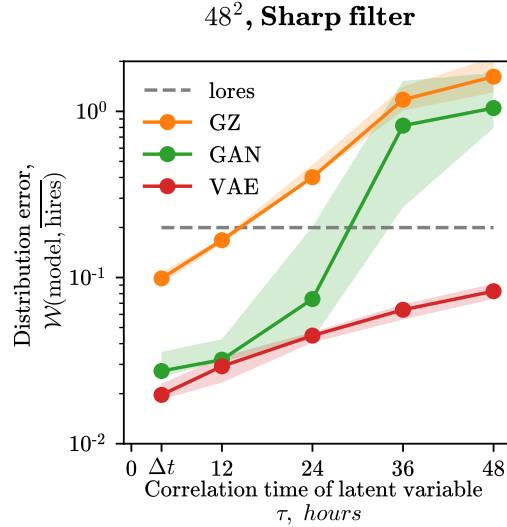


Figure 8. Online distributional metric (17) as a function of correlation time τ of latent variable z for three stochastic subgrid models (GZ, GAN and VAE). lores is model on a coarse grid without parameterization. Shading area shows min-max values among training realizations, and markers show median value. Time step of the numerical integration Δt is 4 hours.

small eddies near the grid scale are energized too much, see KE spectrum in small scales, KE level and tails of PDFs. The GZ model does not prevent overamplification of the small eddies. In contrast, the generative stochastic models (GAN and VAE) improve the simulation of the small eddies: see spectral characteristics in small scales, tails of PDFs and kinetic energy. Note that generative models (GAN and VAE) accurately reproduce kinetic energy growth in transitional flow (panel (a), $t < 2$ years) in agreement with the offline analysis.

Snapshots of the velocity modulus are shown in Figure 10. At time step $\Delta t = 2$ hours baseline models (MSE and GZ) have too many small eddies, and at time step $\Delta t = 1$ hour the flow becomes unphysical and overenergized. GAN and VAE models at both time steps produce physical solutions which look similar to the filtered and coarse-grained high-resolution simulation ($\overline{\text{hires}}$). In Figure 11(a) we show distributional metric as a function of the time step. While baseline models (MSE and GZ) are very sensitive to the time step, the generative models (GAN and VAE) are relatively insensitive to the time step and have the smallest errors. This suggests that the generative stochastic models have better numerical stability properties. See Appendix E for further discussion on numerical stability.

In Figure 11(b) we show the distributional metric as a function of resolution. At the coarsest resolution 48², the generative stochastic models (GAN and VAE) have 5–10 times lower error compared to the unparameterized simulation (lores) and 3–5 times lower error compared to the baseline models (GZ and MSE). For intermediate and higher resolutions (64² and 96²) all ML-based models (GZ, MSE, GAN, VAE) improve distributional error compared to the unparameterized model, but the confidence intervals (shading area) exceed the difference between the median values. So we conclude that the effect of stochastic subgrid models (GZ, GAN, VAE), as opposed to the deterministic one (MSE), at these resolutions is negligible. The discrepancy between the offline and online analysis may be due to the inclusion of *ssd* term, time sampling method of the stochastic parameterization, and time integration scheme. Overall, gen-

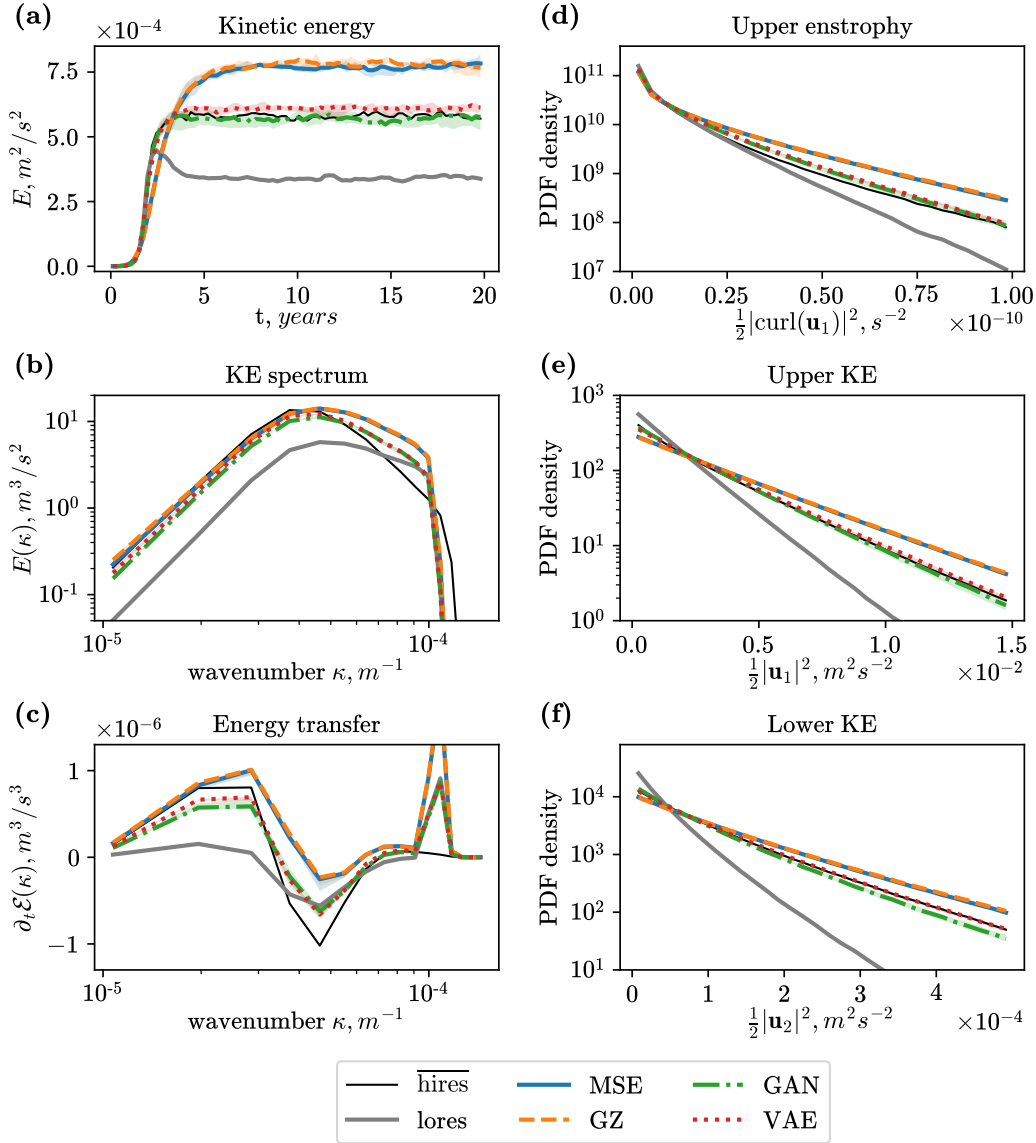
48², Sharp filter


Figure 9. Online simulations with parameterized models (MSE, GZ, GAN and VAE). $\overline{\text{lores}}$ is a coarse unparameterized model. $\overline{\text{hires}}$ is filtered and coarsegrained high-resolution model. MSE is a deterministic subgrid model given by the mean channel of the GZ model. Energy transfer on panel (c) gives a sum of contributions from the resolved advection and subgrid model (see Eq. (E4)). Shading area shows min-max values among training realizations, and lines show median value. The time step Δt is 2 hours.

erative models (GAN and VAE) improve simulation if there are issues with numerical stability, and perform as well as the baseline deterministic model in other cases. The spectral-error metric reported in Appendix E yields similar conclusions.

In Figure 12 we show the online results for the subgrid models trained on the dataset produced using the Gaussian filter. The subgrid models cannot substantially

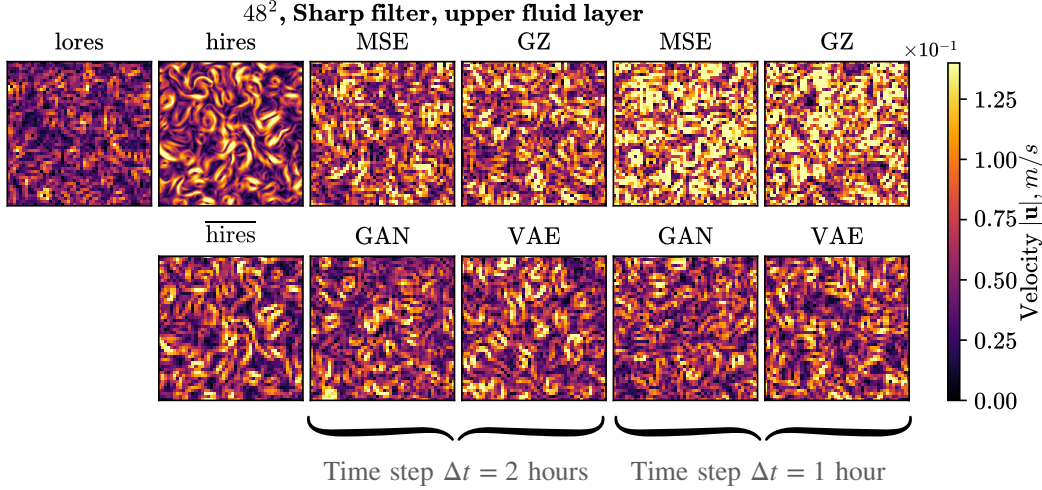


Figure 10. Snapshots of the modulus of velocity. Two columns with $\Delta t = 2$ hours correspond to Figure 9. The smaller the time step, the smaller the effective eddy viscosity, see Appendix E.

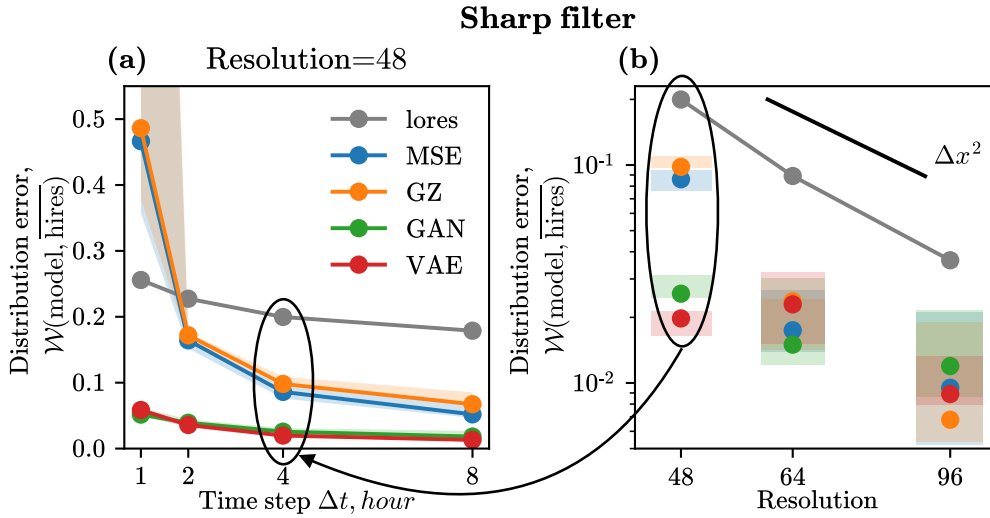


Figure 11. Online distributional metric (Eq. (17)): (a) as a function of time step at the coarsest spatial resolution and (b) as a function of spatial resolution. The shading area shows min-max values among training realizations, and markers show median value. The smaller the time step, the smaller the effective eddy viscosity, see Appendix E.

improve the KE spectrum on large scales with respect to the unparameterized model (panel (b)), and it results in little or no improvement in the other statistical characteristics. At higher resolutions (64^2 and 96^2) we observe the improvement in reproducing the KE spectrum on small scales, but not the large ones (not shown). Similar to Zanna and Bolton (2020), we report in Appendix E how the kinetic energy in online simulation changes when the subgrid model is multiplied by the adjustable parameter. This characteristic clearly demonstrates that subgrid models trained for the Gaussian filter are less efficient in energizing the flow. The same issues for the models trained to predict the subgrid forcing diagnosed with the Gaussian filter were reported in Ross et

48², Gaussian filter

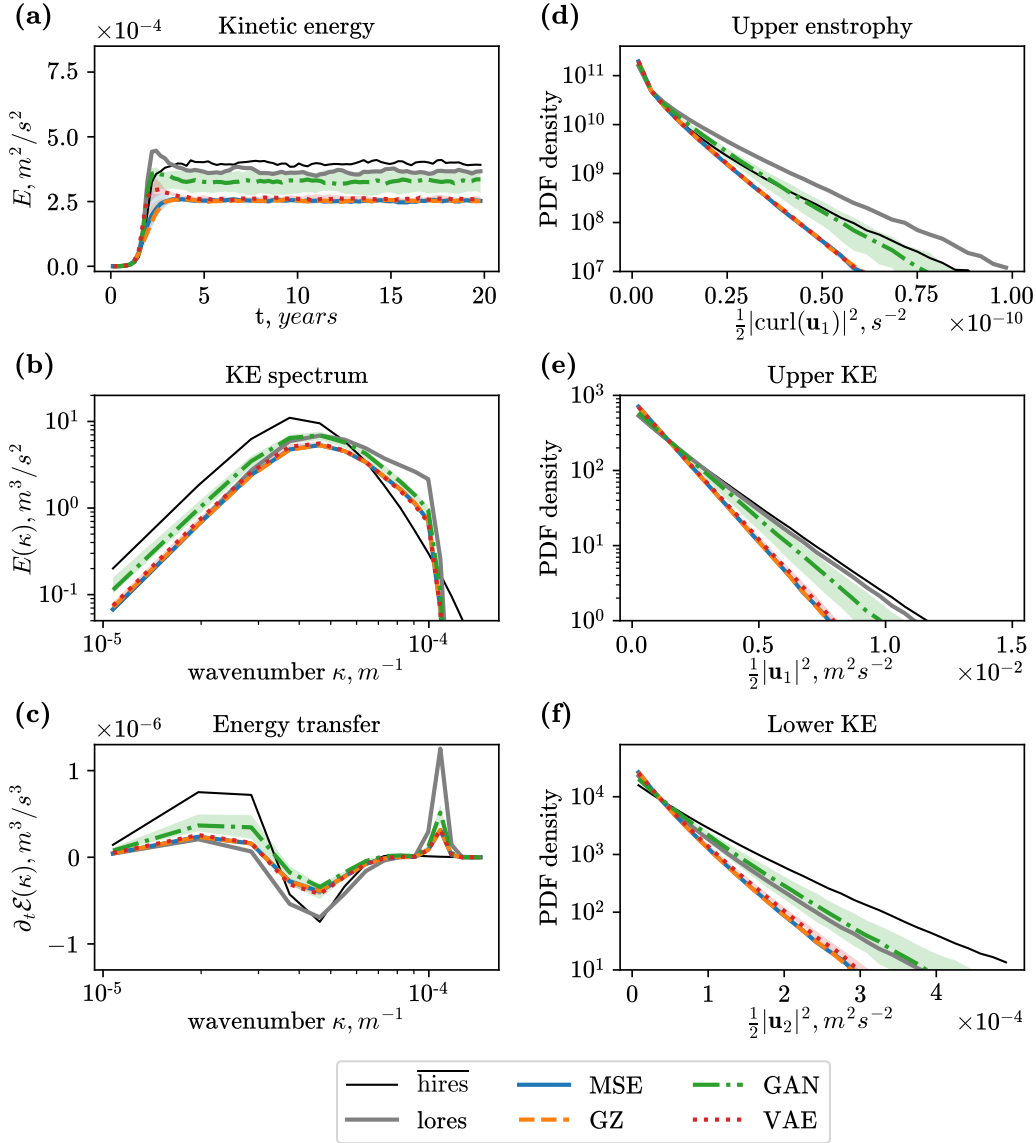


Figure 12. Online simulations with parameterized models. Similar to Figure 9, but for models trained on the dataset with Gaussian filter.

al. (2023) and these may be caused by the mentioned discrepancies between the offline and online analysis.

In Appendix E we include additional online results. The online generalization to the turbulence configuration with jets shows that generative models clearly improve the simulation of the transitional flow, but at a later time, all the subgrid models including baselines experience numerical stability issues. Runtime for the generative models is the same as for the deterministic baseline.

6 Conclusions and discussion

In this work, we propose to leverage generative machine-learning models (GAN and VAE) to build stochastic subgrid parameterizations of mesoscale eddies. Generative models allow to sample from the conditional distribution of subgrid forcing given resolved variables. We performed offline and online evaluations of the proposed subgrid models, and compared them against baseline deterministic and stochastic ML models in an idealized ocean simulation for a range of resolutions.

Our main findings can be summarized as follows:

- Generative models are able to simulate the stochastic residuals of subgrid forcing with spatial structure similar to the true residuals.
- Generative models accurately represent the energy transfer spectrum and thus reproduce the large-scale kinetic energy backscatter missing at coarse-resolution.
- The GAN model is superior to others according to the offline metrics for subgrid forcing; however, the VAE model demonstrates better offline generalization to the unseen turbulence configuration (meandering jets).
- Both generative models (GAN and VAE) improve the numerical stability properties and prevent overamplification of the unphysical flows in online simulations at the coarsest resolution compared to the baseline ML models.

In spite of the different performance of GAN and VAE models in the offline analysis, their performance is similar in online simulations. Therefore, offline metrics or loss functions may be bad proxies for the online performance (Frezat et al., 2022; Ross et al., 2023). The energy transfer spectrum is one of the main properties of subgrid forcing which is essential to properly energize the flow and could be considered as an alternative loss function. However, the spatial structure of the subgrid forcing and stochastic residuals may be important to ensure the development of the physical solution.

Our online simulations are optimal when the time correlation of the latent variable sampling is equal to the model timestep, which is equivalent to a white noise model and consistent with our offline training methodology. The effect of the parameterization can be analyzed by decomposing it into deterministic and stochastic parts. The deterministic part is defined as the conditional mean; while the stochastic part as a white noise model. The white noise process model implies that the energy injection by the stochastic part of the parameterization approaches zero in the limit of the small time steps (Alvelius, 1999). In addition, the average energy injection is fully described by the deterministic part of the parameterization (i.e., conditional mean, see Moser et al. (2021)). One can modify the definition of the subgrid model, for example by including memory effects, to generate a stochastic model with non-vanishing energy input (Chorin & Lu, 2015; Gagne et al., 2020; Agarwal et al., 2021; DelSole, 2000; Berner, 2005; Bhourri & Gentine, 2022).

The important property of the proposed generative models: they do not introduce new limitations to be trained on the global ocean data compared to our baseline model (GZ). Moreover, compared to GZ, both generative models do not require an explicit expression for the likelihood function, and thus slightly more complicated architecture of the stochastic model can be used, for example, the final divergence layer (Zanna & Bolton, 2020) which allows building conservative parameterizations. We expect that the application of generative models for complex flows may greatly improve the quality of the generated stochastic residuals compared to the traditional methods.

Appendix A Numerical schemes and small-scale dissipation

Integration in time is performed with third-order Adams-Bashford scheme (Durran, 1991). Equations (1), (2) are approximated in space on a uniform collocated grid with the use of the pseudo-spectral method, i.e. all derivatives are computed in Fourier space, but the only nonlinear operation ($\mathbf{u}_m q_m$) is computed in physical space (Fox & Orszag, 1973).

In our numerical solver, aliasing errors are reduced with the use of "exponential cut-off" filter (Canuto et al., 2012) denoted as ssd , which attenuates highest spatial frequencies and additionally removes enstrophy near the grid scale thus ensuring numerical stability. Application of the filter $ssd(\cdot)$ is equivalent to multiplication by the following function in Fourier space:

$$\widehat{ssd}(k, l) = \begin{cases} 1, & \kappa < \kappa_c \\ \exp(-23.6(\Delta x)^4(\kappa - \kappa_c)^4), & \text{otherwise} \end{cases} \quad (\text{A1})$$

where $\kappa = \sqrt{k^2 + l^2}$ is radial wavenumber, k and l are zonal and meridional wavenumbers, respectively, Δx is grid step of the model, $\kappa_{max} = \frac{\pi}{\Delta x}$ is maximum wavenumber, $\kappa_c = 0.65\kappa_{max}$. We simplify notation when placing ssd into the right-hand side of the equation (1). Instead, every time step from layer n to a new layer ($q_m^n \rightarrow q_m^*$) is followed by the application of the filter $q_m^{n+1} = ssd(q_m^*)$ (LaCasce, 1996; Arbic & Flierl, 2003).

We do not add molecular viscosity into governing equations (1), and thus formally have "infinite" Reynolds number. However, the dissipation is provided by the term ssd which depends on the grid step (see Eq. (A1)). Grid-dependent small-scale dissipation is a typical choice in ocean modeling (Griffies & Hallberg, 2000). Dissipation introduced by ssd is relatively small, which is important for the simulation of quasi-two dimensional turbulence (Thuburn et al., 2014). An example of the undesirable effect of the dissipative model is shown in Figure 1(d), where ssd is balanced with energy transfer, and thus induces spurious forward energy cascade.

Appendix B LES filters

Spatial filter $\overline{(\cdot)}$ (Eq. (6)) consists of two operations: spectral coarsegraining which reduces the resolution of the image and spectral smoothing. Because both are defined as a pointwise function in Fourier space, they commute and can be composed into a single operator, which we often call a "filter".

Filter $\overline{(\cdot)}$ is applied as multiplication in Fourier space by the following function:

- "Gaussian"

$$\widehat{G}(k, l) = \begin{cases} \exp(-\kappa^2(2\Delta x)^2/24), & \text{otherwise} \\ 0, & \kappa_{max} \leq k \text{ or } \kappa_{max} \leq l \end{cases} \quad (\text{B1})$$

Where $\kappa_{max} = \frac{\pi}{\Delta x}$ and Δx is the grid step of the coarse model. After discarding the frequencies above κ_{max} , the filtered signal is represented on a coarse mesh (Ghosal, 1996). According to the definition of filter width given by Lund (1997), the width of this Gaussian filter is $2\Delta x$, which is twice as large as the grid step of the coarse model.

- "Sharp"

$$\widehat{G}(k, l) = \begin{cases} 1, & \kappa < \kappa_c \\ \exp(-23.6(\Delta x)^4(\kappa - \kappa_c)^4), & \text{otherwise} \\ 0, & \kappa_{max} \leq k \text{ or } \kappa_{max} \leq l \end{cases} \quad (\text{B2})$$

where $\kappa_c = 0.65\kappa_{max}$. This filter is given by a combination of sharp cut-off coarsegraining and model filter (Eq. (A1)).

Motivation for using these filters is given in Ross et al. (2023).

Appendix C Training of the Machine Learning Models

C1 GAN loss function

The presented below training algorithm closely resembles paper of Adler and Öktem (2018), where the discriminator analyzes two generated images.

We generate two images $\tilde{S}_1 = G(z_1, \bar{q})$ and $\tilde{S}_2 = G(z_2, \bar{q})$ for a given \bar{q} with two samples from standard normal distribution $z_1, z_2 \in \mathbb{R}^{2 \times n \times n}$ and stack them in layer dimension:

$$\mathbf{S}_1 = (\tilde{S}_1, S), \mathbf{S}_2 = (S, \tilde{S}_2), \tilde{\mathbf{S}} = (\tilde{S}_1, \tilde{S}_2),$$

where $\mathbf{S}_1, \mathbf{S}_2, \tilde{\mathbf{S}} \in \mathbb{R}^{4 \times n \times n}$. The WGAN loss (Eq. (14)) for a single data sample transforms to:

$$\mathcal{L}_W = \left[\frac{1}{2} (D(\mathbf{S}_1, \bar{q}) + D(\mathbf{S}_2, \bar{q})) - D(\tilde{\mathbf{S}}, \bar{q}) \right].$$

The discriminator D should be 1-Lipschitz in the first argument, and we enforce it with the gradient penalty (WGAN-GP, Gulrajani et al. (2017)):

$$\mathcal{L}_{\text{grad}} = \left(\|\nabla_{\hat{\mathbf{S}}} D(\hat{\mathbf{S}}, \bar{q})\|_2 - 1 \right)^2,$$

where $\hat{\mathbf{S}} = \epsilon \mathbf{S} + (1 - \epsilon) \tilde{\mathbf{S}}$. The random number ϵ is uniformly distributed on $[0, 1]$ and chosen uniquely for every training sample. For every batch we choose randomly \mathbf{S} from set $\{\mathbf{S}_1, \mathbf{S}_2\}$. Note that $D(\hat{\mathbf{S}}, \bar{q}) \in \mathbb{R}$, $\nabla_{\hat{\mathbf{S}}} D(\hat{\mathbf{S}}, \bar{q}) \in \mathbb{R}^{4 \times n \times n}$ and norm $\|\cdot\|_2$ for tensor is defined above. Regularization preventing drift of discriminator:

$$\mathcal{L}_{\text{drift}} = [D(\mathbf{S}_1, \bar{q})]^2.$$

We minimize the following loss for the discriminator:

$$\mathcal{L}_D = -\mathcal{L}_W + 10\mathcal{L}_{\text{grad}} + 10^{-3}\mathcal{L}_{\text{drift}}, \tag{C1}$$

and the loss to be minimized for the generator is:

$$\mathcal{L}_G = -D(\tilde{\mathbf{S}}, \bar{q}). \tag{C2}$$

In the original paper $\mathcal{L}_G = \mathcal{L}_W$ (Adler & Öktem, 2018), but we follow a typical approach when only generated samples constitute the generator loss (Dong & Yang, 2019).

Discriminator D is parameterized by DCGAN discriminator (Radford et al., 2015) with two modifications: we remove the activation function in the final layer and remove batch normalization because it is necessary for proper use of gradient penalty (Gulrajani et al., 2017). Following Arjovsky et al. (2017), we optimize the discriminator loss (Eq. (C1)) for five batches in a row, and then we optimize the generator loss (Eq. (C2)) for one batch.

C2 VAE loss function

To train the VAE model we parameterize every probability density in the VAE loss function (Eq. (16)) with Gaussian distributions.

The distributions for encoder, decoder and prior, respectively:

$$q_\phi(z|S, \bar{q}) = \mathcal{N}(\mu_\phi(S, \bar{q}), \text{diag}(\sigma_\phi^2(S, \bar{q}))) \quad (\text{C3})$$

$$\rho_\theta(S|z, \bar{q}) = \mathcal{N}(\mu_\theta(z, \bar{q}), \gamma I) \quad (\text{C4})$$

$$\rho(z) = \mathcal{N}(0, I), \quad (\text{C5})$$

where $I \in \mathbb{R}^{2n^2 \times 2n^2}$ is identity matrix, γ is free parameter and $S, \bar{q}, z \in \mathbb{R}^{2 \times n \times n}$. The mappings μ_ϕ , $\ln(\sigma_\phi^2)$ and μ_θ are deterministic.

The loss function to be minimized (Eq. (16)) for one training sample transforms to:

$$\mathcal{L}_{\text{VAE}} = \frac{1}{2\gamma} \|S - \mu_\theta(\hat{z}, \bar{q})\|_2^2 + \frac{1}{2} \sum_{m,i,j} (\sigma_\phi^2 + \mu_\phi^2 - 1 - \ln(\sigma_\phi^2))_{m,i,j}, \quad (\text{C6})$$

where \hat{z} is one sample from encoder distribution, i.e. $\hat{z} = \mu_\phi + \epsilon\sigma_\phi$, $\epsilon \sim \mathcal{N}(0, I)$. Note that as suggested by Rybkin et al. (2021), we sum values of MSE loss and KL loss across dimensions. The variance of decoder distribution γ is a parameter regulating the relative importance of reconstruction and regularization terms. According to Takida et al. (2022), common problems of VAE such as posterior collapse and smoothness of generated images may result from the incorrect choice of parameter γ . Following Rybkin et al. (2021), we estimate the variance of the decoder as a mean squared error: $\gamma = \frac{1}{2n^2} \|S - \mu_\theta\|_2^2$. We compute γ uniquely for every batch and do not differentiate it.

C3 Additional training information

All image-to-image mappings (mean and variance prediction in GZ, generator in GAN, encoder and decoder in VAE) are based on the same convolutional neural network (CNN) similar to Guillaumin and Zanna (2021); Ross et al. (2023) with parameters given in Table C1. We follow a common approach with the normalization of input and output variables before passing them to neural networks. Each channel representing a different physical quantity or different fluid layer is normalized by a unique standard deviation computed over the training dataset. Note that the variance channel of GZ model is normalized by the squared standard deviation of the mean channel. Normalization constants become part of the model and they are not adjusted in offline or online tests.

Models are trained in Pytorch (Paszke et al., 2019), batch size is 64, training algorithm is Adam (Kingma & Ba, 2014) with standard parameters $(\beta_1, \beta_2) = (0.9, 0.999)$ for GZ and VAE, and $(\beta_1, \beta_2) = (0.5, 0.999)$ for GAN (Radford et al., 2015). The learning rate is $lr = 0.001$ for GZ and $lr = 0.0002$ for GAN and VAE. GAN and VAE models are optimized for 200 epochs, and in GZ model each channel (mean and variance) is optimized for 50 epochs. Early stopping or any other criteria for choosing the best epoch was not used. Weight decay was not used. We use the following scheduler of the learning rate for GZ and VAE: on every milestone $[1/2, 3/4, 7/8] \cdot N_{\text{epoch}}$ multiply learning rate by $\gamma = 0.1$, for GAN $\gamma = 0.5$. Weights of the discriminator and generator of GAN are initialized with zero mean and standard deviation 0.02 (Radford et al., 2015). During inference, neural networks are switched to evaluation mode so that batch normalization layers use parameters accumulated during training.

Table C1. Configuration of convolutional neural network (CNN) parameterizing image to image mapping.

| | |
|--|--|
| Number of input/output images | arbitrary (n_{in}, n_{out}) |
| Resolution of input/output/hidden layers | arbitrary, but the same |
| Number of filters | 128, 64, 32, 32, 32, 32, 32, n_{out} |
| Filter width | 5, 5, 3, 3, 3, 3, 3, 3 |
| Boundary conditions | periodic ("circular padding") |
| Activation function | ReLU, in hidden layers |
| Batch normalization | after ReLU, in hidden layers |

Appendix D Additional offline results

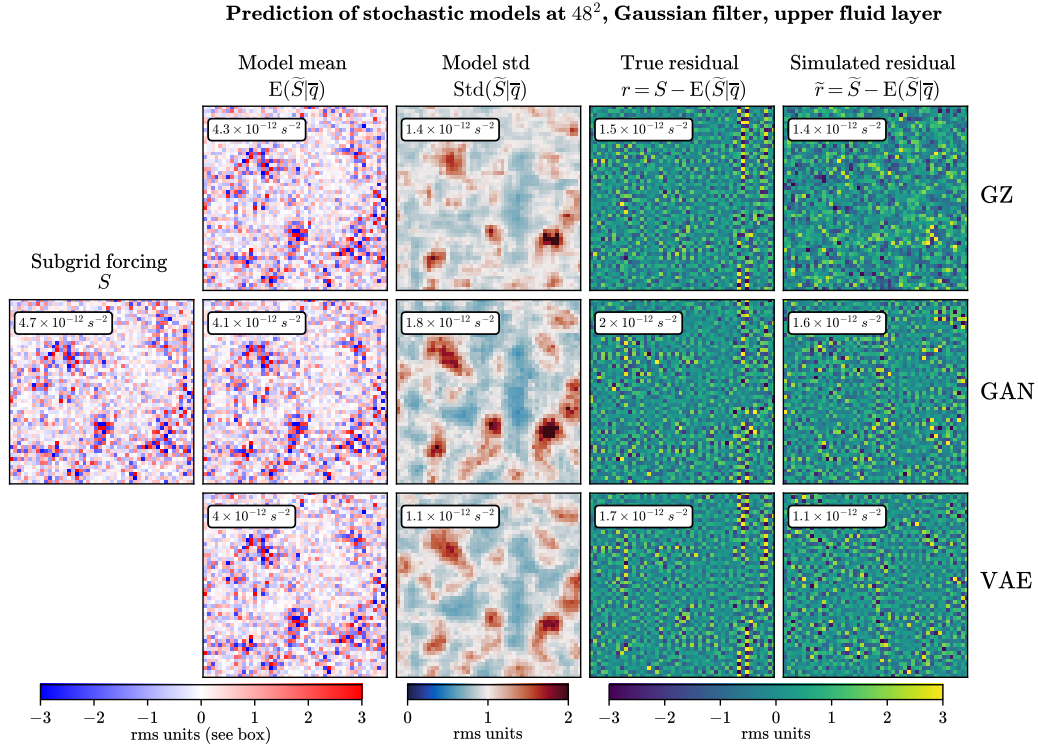


Figure D1. Prediction of subgrid forcing on the testing dataset. Same as Figure 4 but for Gaussian filter.

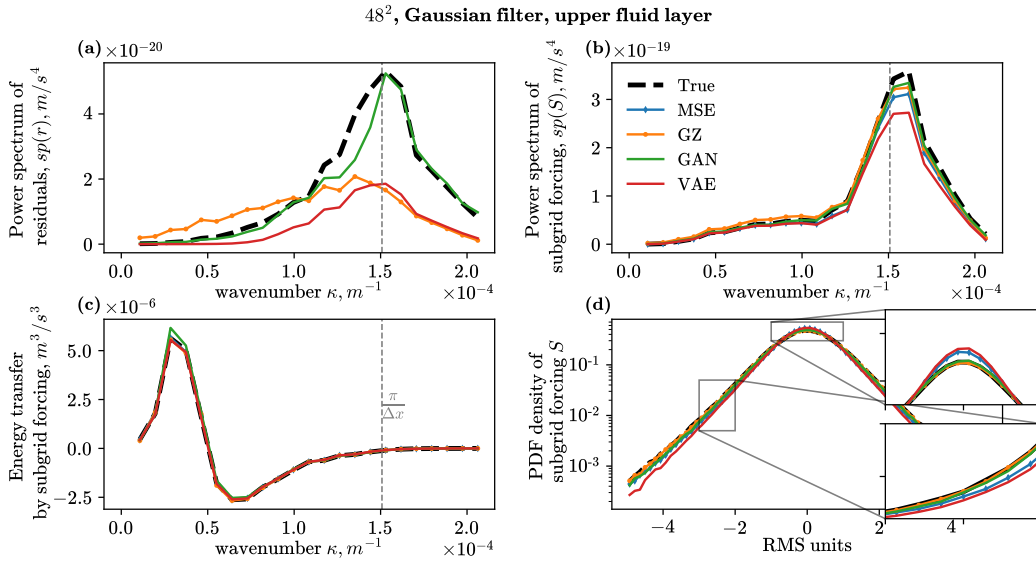


Figure D2. Offline analysis of subgrid models. Same as Figure 5 but for Gaussian filter.

In Figures D1, D2 and D3 we show the results of offline analysis for the Gaussian filter. Our main conclusions about the performance of the stochastic models are the same as for the Sharp filter.

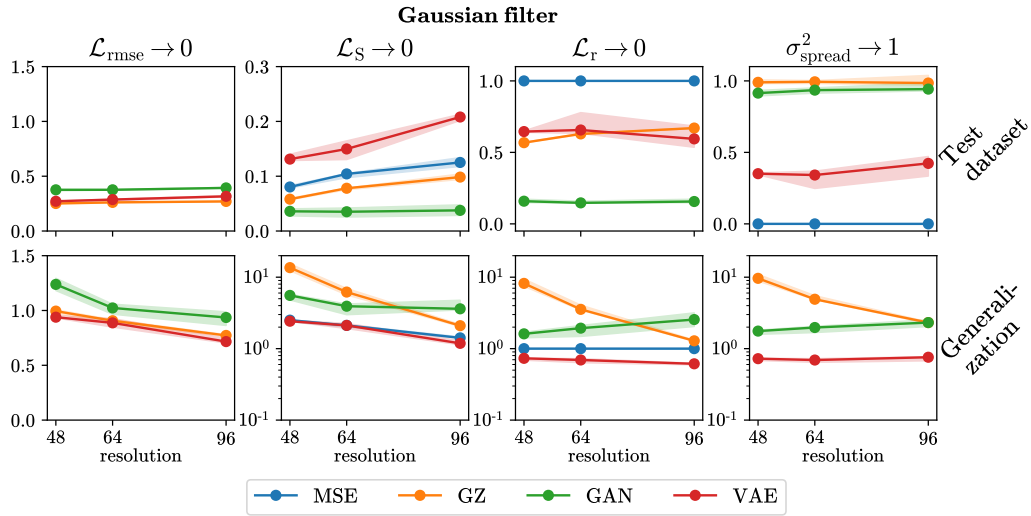


Figure D3. Offline metrics from Table 2. Same as Figure 6 but for Gaussian filter.

Appendix E Additional online results

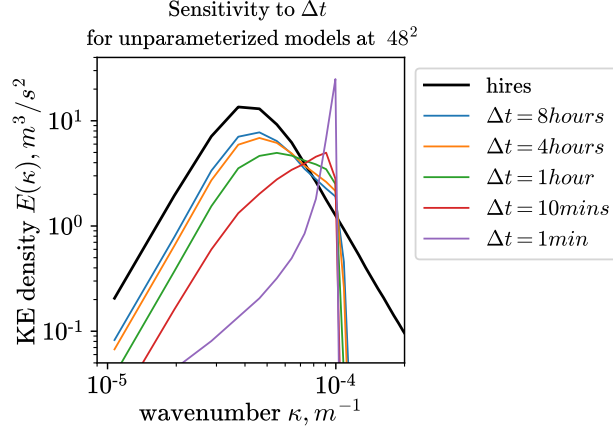


Figure E1. Here we show that the smaller the time step, the smaller the effective eddy viscosity and the closer the unparameterized coarse model to inviscid simulation, i.e. energy accumulates near the cut-off for a small time step.

Below we show that the time step is connected to the effective eddy viscosity in our particular numerical scheme, and thus the sensitivity of the online simulation results to the time step reveals the numerical stability properties. Small-scale dissipation (*ssd*) was formulated not as a tendency in RHS of the governing equation, but as a postprocessing operation following every time step, see Appendix A. Because *ssd* does not contain a time step explicitly, the effect of dissipation accumulates over several time steps (Lund, 2003). An effective filter that we apply to the solution per unit time interval (if solution is steady) is $\widehat{ssd}^{1/\Delta t}$, which converges at every radial wavenumber κ to Heaviside step function

$$(\widehat{ssd}(\kappa))^{1/\Delta t} \rightarrow H(0.65\kappa_{max} - \kappa) \text{ as } \Delta t \rightarrow 0. \quad (\text{E1})$$

Filtering with this Heaviside step function approximately corresponds to 2/3-dealiasing scheme (Orszag, 1971), which is the discretization of inviscid equations and conserves energy and enstrophy. So, we expect that the smaller the time step, the smaller the effective eddy viscosity produced by the *ssd* term. By refining the time step, we established that unparameterized coarse models behave as inviscid simulation at $\Delta t \rightarrow 0$, i.e. energy accumulates near the grid scale, see Figure E1.

Below we derive a spectral metric for the analysis of online simulations. The rate of change of total energy (Eq. (4)) is defined as

$$\partial_t \mathcal{E} = -\frac{1}{H} \sum_{m=1}^2 H_m \langle \psi_m \partial_t q_m \rangle. \quad (\text{E2})$$

Applying this formula to the governing equation (1) and using Parseval theorem, the rate of change of total energy in Fourier space is

$$\begin{aligned} \partial_t \mathcal{E}(k, l) = & -\frac{1}{H} \sum_{m=1}^2 H_m \text{Re} \left(\widehat{\psi}_m^* \partial_t \widehat{q}_m \right) = \\ & \frac{1}{H} \sum_{m=1}^2 H_m \text{Re} \left(\underbrace{\widehat{\psi}_m^* \nabla(\widehat{\mathbf{u}}_m \widehat{q}_m)}_{\text{energy transfer}} + \underbrace{\widehat{\psi}_m^* U_m \partial_x \widehat{q}_m}_{\text{energy source}} + \underbrace{\widehat{\psi}_m^* \delta_{m,2} r_{ek} \nabla^2 \widehat{\psi}_m}_{\text{energy dissipation}} \right). \end{aligned} \quad (\text{E3})$$

Here we neglected the contribution from the small-scale dissipation term ssd which acts on a limited set of wavenumbers. The corresponding energy balance equation for filtered and coarsegrained system (7) is

$$\begin{aligned} \partial_t \bar{\mathcal{E}}(k, l) = & -\frac{1}{H} \sum_{m=1}^2 H_m \text{Re} \left(\widehat{\psi}_m^* \partial_t \widehat{q}_m \right) = \\ & \frac{1}{H} \sum_{m=1}^2 H_m \text{Re} \left(\underbrace{\widehat{\psi}_m^* \nabla(\widehat{\mathbf{u}}_m \widehat{q}_m)}_{\text{energy transfer}} + \underbrace{\widehat{\psi}_m^* U_m \partial_x \widehat{q}_m}_{\text{energy source}} + \underbrace{\widehat{\psi}_m^* \delta_{m,2} r_{ek} \nabla^2 \widehat{\psi}_m}_{\text{energy dissipation}} \right). \end{aligned} \quad (\text{E4})$$

In Eq. (E4) the energy transfer is proportional to $\sim \overline{\mathbf{u}_m q_m}$ and can be split into the resolved transfer $\sim \overline{\mathbf{u}_m \bar{q}_m}$ and unresolved transfer $\sim \overline{\mathbf{u}_m q_m} - \overline{\mathbf{u}_m \bar{q}_m}$ which is parameterized by the subgrid model. Every term in equation (E4) can be obtained from the corresponding term in the energy balance of high-resolution simulation (Eq. (E3)) by multiplying twice by the filter transfer function, i.e. by $(\widehat{G}(k, l))^2$.

We define distance between two isotropic spectra $\mathcal{E}_1, \mathcal{E}_2$ as:

$$L_2(\mathcal{E}_1, \mathcal{E}_2) = \sqrt{\frac{1}{\kappa_c} \int_0^{\kappa_c} (\mathcal{E}_1(\kappa) - \mathcal{E}_2(\kappa))^2 d\kappa}, \quad (\text{E5})$$

where κ_c is the truncation wavenumber of the exponential filter (Eq. (B2)). And average normalized distance between model on a coarse grid ("model") and filtered and coarsegrained high resolution simulation ("hires")

$$\frac{L_2(\mathcal{E}_{\text{model}}, \mathcal{E}_{\text{hires}})}{L_2(0, \mathcal{E}_{\text{hires}})} \quad (\text{E6})$$

over "energy transfer", "energy source" and kinetic energy spectrum in upper and lower fluid layers. The described spectral error is shown in Figure E2.

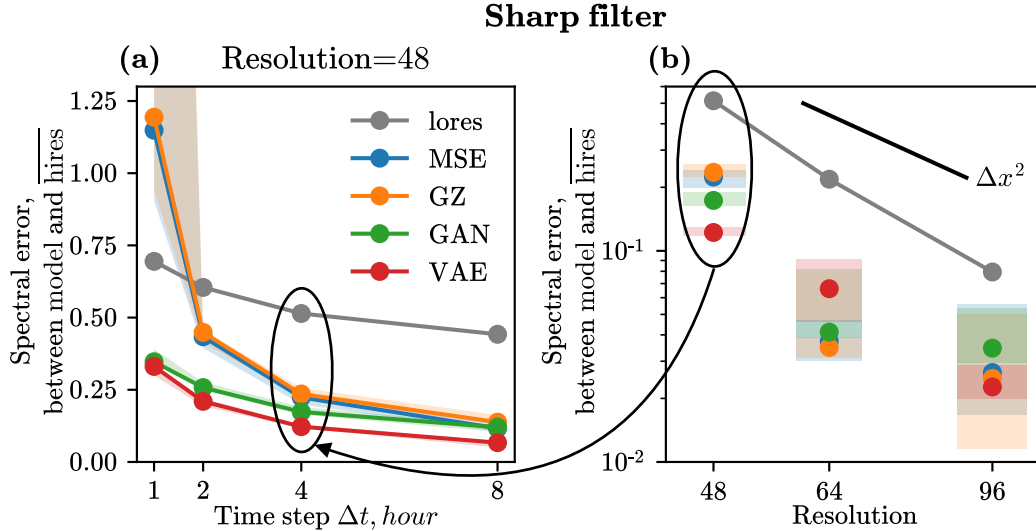


Figure E2. Online metric similar to Figure 11, but for spectral error. The shading area shows min-max values among training realizations, and markers show median values.

Figure E3 shows the sensitivity to the amplitude of the parameterization and Figure E4 shows the online generalization to the jet dataset. Table E1 shows runtime of the parameterized models.

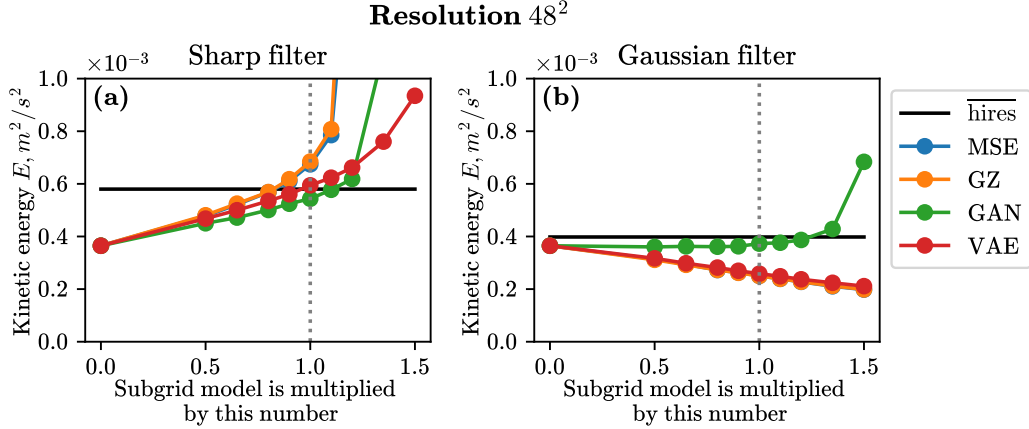


Figure E3. We multiply the subgrid model by a parameter $\alpha \in [0, 1.5]$ as $\tilde{S} \rightarrow \alpha\tilde{S}$ and show the kinetic energy after spin-up. Subgrid models which efficiently simulate backscatter are able to energize the flow when the amplitude is increased, see supplemental Figure S9 in Zanna and Bolton (2020). All models trained for the Sharp filter efficiently energize the flow, but for the Gaussian filter they mostly do not energize the flow. Time step Δt is 4 hours.

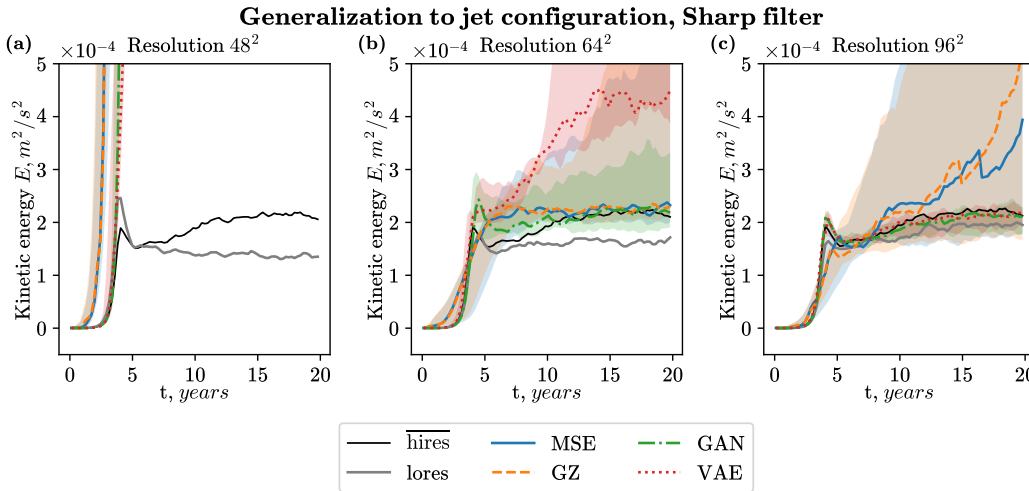


Figure E4. Online generalization to turbulence configuration with jets (Ross et al., 2023). Generative models (GAN and VAE) clearly better reproduce transitional flow ($t < 2$ years), however many of the presented models have problems with numerical stability at a larger time. Improving the generalization capabilities of presented models requires further research. The shading area shows min-max values among training realizations, lines show median value. The time step is 2 hours.

| Δt | 1 hour | 2 hour | 4 hour | | | | |
|--------------|------------------|----------------|----------------|-----|------|-----|-----|
| $n \times n$ | 256×256 | 96×96 | 48×48 | | | | |
| Model | – | – | – | MSE | GZ | GAN | VAE |
| Runtime, sec | 1300 | 130 | 25.4 | 756 | 1480 | 784 | 782 |

Table E1. Runtime on one CPU core for unparameterized model (“–”) and ML-based parameterizations to integrate QG model in time for 20 years. Theoretically, we expect that the runtime for MSE, GAN, and VAE models should be the same, and for GZ is twice as large. Runtime for the GZ model can be reduced if aggregate mean and variance channels into one CNN network, as it is done in Guillaumin and Zanna (2021).

Data Availability Statement

Python software used for training and evaluation of the subgrid models is available via https://github.com/m2lines/pyqg_generative (see the archived version on Zenodo, Perekhgin (2023b)). We provide training and simulation data on Zenodo (Perekhgin, 2023a).

Acknowledgments

This research is supported by the generosity of Eric and Wendy Schmidt by recommendation of Schmidt Futures, as part of its Virtual Earth System Research Institute (VESRI). C.F.G. was partially supported by NSF DMS Grant 2009752. This research was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise and by the National Science Foundation under Grant No. NSF PHY-1748958. The authors would like to thank the members of M²LInES for their helpful comments and discussions.

References

- Abernathy, R., Rocha, C. B., Ross, A., Jansen, M., Li, Z., Poulin, F. J., . . . Tobias (2022, May). *pyqg/pyqg: v0.7.2*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.6563667> doi: 10.5281/zenodo.6563667
- Adler, J., & Öktem, O. (2018). Deep bayesian inversion. *arXiv preprint arXiv:1811.05910*.
- Agarwal, N., Kondrashov, D., Dueben, P., Ryzhov, E., & Berloff, P. (2021). A comparison of data-driven approaches to build low-dimensional ocean models. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002537.
- Alcala, J., & Timofeyev, I. (2021). Subgrid-scale parametrization of unresolved scales in forced burgers equation using generative adversarial networks (gan). *Theoretical and Computational Fluid Dynamics*, *35*(6), 875–894.
- Alvelius, K. (1999). Random forcing of three-dimensional homogeneous turbulence. *Physics of Fluids*, *11*(7), 1880–1889.
- Andrejczuk, M., Cooper, F., Juricke, S., Palmer, T., Weisheimer, A., & Zanna, L. (2016). Oceanic stochastic parameterizations in a seasonal forecast system. *Monthly Weather Review*, *144*(5), 1867–1875.
- Arbic, B. K., & Flierl, G. R. (2003). Coherent vortices and kinetic energy ribbons in asymptotic, quasi two-dimensional f-plane turbulence. *Physics of Fluids*, *15*(8), 2177–2189.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arxiv 2017. *arXiv preprint arXiv:1701.07875*, *30*(4).
- Arnold, H., Moroz, I., & Palmer, T. (2013). Stochastic parametrizations and model uncertainty in the lorenz’96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1991), 20110479.
- Bachman, S. D. (2019). The gm+ e closure: A framework for coupling backscatter with the gent and mcwilliams parameterization. *Ocean Modelling*, *136*, 85–106.
- Bachman, S. D., Anstey, J. A., & Zanna, L. (2018). The relationship between a deformation-based eddy parameterization and the lans- α turbulence model. *Ocean Modelling*, *126*, 56–62.
- Bachman, S. D., Fox-Kemper, B., & Pearson, B. (2017). A scale-aware subgrid model for quasi-geostrophic turbulence. *Journal of Geophysical Research: Oceans*, *122*(2), 1529–1554.
- Beck, A., Flad, D., & Munz, C.-D. (2019). Deep neural networks for data-driven les closure models. *Journal of Computational Physics*, *398*, 108910.

- Berloff, P. (2018). Dynamically consistent parameterization of mesoscale eddies. part iii: Deterministic approach. *Ocean Modelling*, *127*, 1–15.
- Berner, J. (2005). Linking nonlinearity and non-gaussianity of planetary wave behavior by the fokker–planck equation. *Journal of the atmospheric sciences*, *62*(7), 2098–2117.
- Berner, J., Achatz, U., Batte, L., Bengtsson, L., De La Camara, A., Christensen, H. M., . . . others (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588.
- Berner, J., Jung, T., & Palmer, T. (2012). Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *Journal of Climate*, *25*(14), 4946–4962.
- Berner, J., Shutts, G., Leutbecher, M., & Palmer, T. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ecmwf ensemble prediction system. *Journal of the Atmospheric Sciences*, *66*(3), 603–626.
- Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., . . . others (2021). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.
- Bhour, M. A., & Gentine, P. (2022). History-based, bayesian, closure for stochastic parameterization: Application to lorenz’96. *arXiv preprint arXiv:2210.14488*.
- 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. doi: 10.1029/2018MS001472
- Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908.
- Canuto, C., Hussaini, M. Y., Quarteroni, A., Thomas Jr, A., et al. (2012). *Spectral methods in fluid dynamics*. Springer Science & Business Media.
- Chasnov, J. R. (1991). Simulation of the kolmogorov inertial subrange using an improved subgrid model. *Physics of Fluids A: Fluid Dynamics*, *3*(1), 188–200.
- Chorin, A. J., & Lu, F. (2015). Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proceedings of the National Academy of Sciences*, *112*(32), 9804–9809.
- Chow, F. K., & Moin, P. (2003). A further study of numerical errors in large-eddy simulations. *Journal of Computational Physics*, *184*(2), 366–380.
- Christensen, H., Berner, J., Coleman, D. R., & Palmer, T. (2017). Stochastic parameterization and el niño–southern oscillation. *Journal of Climate*, *30*(1), 17–38.
- Christensen, H., & Zanna, L. (2022). Parametrization in weather and climate models. In *Oxford research encyclopedia of climate science*.
- Dai, B., Wang, Z., & Wipf, D. (2020). The usual suspects? reassessing blame for vae posterior collapse. In *International conference on machine learning* (pp. 2313–2322).
- DelSole, T. (2000). A fundamental limitation of markov models. *Journal of the atmospheric sciences*, *57*(13), 2158–2168.
- Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.
- Dong, H.-W., & Yang, Y.-H. (2019). Towards a deeper understanding of adversarial losses. *arXiv preprint arXiv:1901.08753*.
- Durran, D. R. (1991). The third-order adams-bashforth method: An attractive alternative to leapfrog time differencing. *Monthly weather review*, *119*(3), 702–720.
- Fox, D. G., & Orszag, S. A. (1973). Pseudospectral approximation to two-dimensional turbulence. *Journal of Computational Physics*, *11*(4), 612–619.
- Fox-Kemper, B., & Menemenlis, D. (2008). Can large eddy simulation techniques improve mesoscale rich ocean models? *Washington DC American Geophysical Union Geophysical Monograph Series*, *177*, 319–337.

- Frederiksen, J. S., & Davies, A. G. (1997). Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *Journal of the atmospheric sciences*, *54*(20), 2475–2492.
- Frederiksen, J. S., Dix, M. R., & Davies, A. G. (2003). The effects of closure-based eddy diffusion on the climate and spectra of a gcm. *Tellus A: Dynamic Meteorology and Oceanography*, *55*(1), 31–44.
- Frezat, H., Balarac, G., Le Sommer, J., Fablet, R., & Lguensat, R. (2021). Physical invariance in neural networks for subgrid-scale scalar flux modeling. *Physical Review Fluids*, *6*(2), 024607.
- Frezat, H., Sommer, J. L., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *arXiv preprint arXiv:2204.03911*.
- Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz’96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896.
- Gent, P. R., & McWilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, *20*(1), 150–155.
- Gerard, L. (2007). An integrated package for subgrid convection, clouds and precipitation compatible with meso-gamma scales. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *133*(624), 711–730.
- Ghosal, S. (1996). An analysis of numerical errors in large-eddy simulations of turbulence. *Journal of Computational Physics*, *125*(1), 187–206.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.
- Graham, J. P., & Ringler, T. (2013). A framework for the evaluation of turbulence closures used in mesoscale ocean large-eddy simulations. *Ocean Modelling*, *65*, 25–39.
- Griffies, S. M., & Hallberg, R. W. (2000). Biharmonic friction with a smagorinsky-like viscosity for use in large-scale eddy-permitting ocean models. *Monthly Weather Review*, *128*(8), 2935–2946.
- Grooms, I., Lee, Y., & Majda, A. J. (2015). Numerical schemes for stochastic backscatter in the inverse cascade of quasigeostrophic turbulence. *Multiscale Modeling & Simulation*, *13*(3), 1001–1021.
- Guan, Y., Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2022). Stable a posteriori les of 2d turbulence using convolutional neural networks: Backscattering analysis and generalization to higher re via transfer learning. *Journal of Computational Physics*, *458*, 111090.
- Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2022). Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate les. *Physica D: Nonlinear Phenomena*, 133568.
- Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534.
- Gullbrand, J., & Chow, F. K. (2003). The effect of numerical errors and turbulence models in large-eddy simulations of channel flow, with and without explicit filtering. *Journal of Fluid Mechanics*, *495*, 323–341.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, *30*.
- Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., ... others (2016). High resolution model intercomparison project (highresmip v1. 0) for cmip6. *Geoscientific Model Development*, *9*(11), 4185–4208.

- 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., ... others (2020). Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, *6*(4), 137–152.
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).
- Jansen, M. F., Adcroft, A., Khani, S., & Kong, H. (2019). Toward an energetically consistent, resolution aware parameterization of ocean mesoscale eddies. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2844–2860.
- Jansen, M. F., & Held, I. M. (2014). Parameterizing subgrid-scale eddy effects using energetically consistent backscatter. *Ocean Modelling*, *80*, 36–48.
- Juricke, S., Danilov, S., Koldunov, N., Oliver, M., & Sidorenko, D. (2020). Ocean kinetic energy backscatter parametrization on unstructured grids: Impact on global eddy-permitting simulations. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001855.
- Juricke, S., Palmer, T. N., & Zanna, L. (2017). Stochastic subgrid-scale ocean mixing: impacts on low-frequency variability. *Journal of Climate*, *30*(13), 4997–5019.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., & Hoyer, S. (2021). Machine learning-accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, *118*(21), e2101784118.
- Kraichnan, R. H. (1976). Eddy viscosity in two and three dimensions. *Journal of Atmospheric Sciences*, *33*(8), 1521–1536.
- LaCasce, J. H. (1996). *Baroclinic vortices over a sloping bottom* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Leslie, D., & Quarini, G. (1979). The application of turbulence theory to the formulation of subgrid modelling procedures. *Journal of fluid mechanics*, *91*(1), 65–91.
- Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are gans created equal? a large-scale study. *Advances in neural information processing systems*, *31*.
- Lund, T. (1997). On the use of discrete filters for large eddy simulation. *Annual Research Briefs*, 83–95.
- Lund, T. (2003). The use of explicit filters in large eddy simulation. *Computers & Mathematics with Applications*, *46*(4), 603–616.
- Mana, P. P., & Zanna, L. (2014). Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling*, *79*, 1–20.
- Mao, Q., Lee, H.-Y., Tseng, H.-Y., Ma, S., & Yang, M.-H. (2019). Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1429–1437).
- Maulik, R., & San, O. (2017). A neural network approach for the blind deconvolution of turbulent flows. *Journal of Fluid Mechanics*, *831*, 151–181.
- Maulik, R., San, O., Rasheed, A., & Vedula, P. (2019). Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, *858*, 122–144.
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Mishra, A., Krishna Reddy, S., Mittal, A., & Murthy, H. A. (2018). A generative

- model for zero shot learning using conditional variational autoencoders. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 2188–2196).
- Moser, R. D., Haering, S. W., & Yalla, G. R. (2021). Statistical properties of subgrid-scale turbulence models. *Annual Review of Fluid Mechanics*, *53*, 255–286.
- Nadiga, B., & Livescu, D. (2007). Instability of the perfect subgrid model in implicit-filtering large eddy simulation of geostrophic turbulence. *Physical Review E*, *75*(4), 046303.
- Nadiga, B. T., Sun, X., & Nash, C. (2022). Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, *1*, e22.
- Ohayon, G., Adrai, T., Vaksman, G., Elad, M., & Milanfar, P. (2021). High perceptual quality image denoising with a posterior sampling cgan. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1805–1813).
- Orszag, S. A. (1971). On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of Atmospheric Sciences*, *28*(6), 1074–1074.
- Pagnoni, A., Liu, K., & Li, S. (2018). Conditional variational autoencoder for neural machine translation. *arXiv preprint arXiv:1812.04405*.
- Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on progress in Physics*, *63*(2), 71.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.
- Pawar, S., San, O., Rasheed, A., & Vedula, P. (2020). A priori analysis on deep learning of subgrid-scale parameterizations for kraichnan turbulence. *Theoretical and Computational Fluid Dynamics*, *34*(4), 429–455.
- Pawar, S., San, O., Rasheed, A., & Vedula, P. (2022). Frame invariant neural network closures for kraichnan turbulence. *arXiv preprint arXiv:2201.02928*.
- Pearson, B., Fox-Kemper, B., Bachman, S., & Bryan, F. (2017). Evaluation of scale-aware subgrid mesoscale eddy models in a global eddy-rich model. *Ocean Modelling*, *115*, 42–58.
- Perezhogin, P. (2023a, February). *Dataset for paper Pavel Perezhogin, Laure Zanna, Carlos Fernandez-Granda "Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model" submitted to JAMES*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.7622683> doi: 10.5281/zenodo.7622683
- Perezhogin, P. (2023b, February). *Software for paper Pavel Perezhogin, Laure Zanna, Carlos Fernandez-Granda "Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model" submitted to JAMES*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.7641961> doi: 10.5281/zenodo.7641961
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Rasp, S., Pritchard, M. S., & Gentile, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689.
- Redi, M. H. (1982). Oceanic isopycnal mixing by coordinate rotation. *Journal of Physical Oceanography*, *12*(10), 1154–1158.
- Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, *15*(1), e2022MS003258.

- Rybkin, O., Daniilidis, K., & Levine, S. (2021). Simple and effective vae training with calibrated decoders. In *International conference on machine learning* (pp. 9179–9189).
- Sagaut, P. (2006). *Large eddy simulation for incompressible flows: an introduction*. Springer Science & Business Media.
- Salmon, R. (1980). Baroclinic instability and geostrophic turbulence. *Geophysical & Astrophysical Fluid Dynamics*, 15(1), 167–211.
- Schumann, U. (1995). Stochastic backscatter of turbulence energy and scalar variance by random subgrid-scale fluxes. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 451(1941), 293–318.
- Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2022). Implicit learning of convective organization explains precipitation stochasticity. *Authorea Preprints*.
- Shutts, G., & Palmer, T. (2007). Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *Journal of climate*, 20(2), 187–202.
- Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.
- Storto, A., & Andriopoulos, P. (2021). A new stochastic ocean physics package and its application to hybrid-covariance data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 147(736), 1691–1725.
- Subramanian, A., Juricke, S., Dueben, P., & Palmer, T. (2019). A stochastic representation of subgrid uncertainty for dynamical core development. *Bulletin of the American Meteorological Society*, 100(6), 1091–1101.
- Sura, P., Newman, M., Penland, C., & Sardeshmukh, P. (2005). Multiplicative noise and non-gaussianity: A paradigm for atmospheric regimes? *Journal of the atmospheric sciences*, 62(5), 1391–1409.
- Takida, Y., Liao, W.-H., Lai, C.-H., Uesaka, T., Takahashi, S., & Mitsufuji, Y. (2022). Preventing oversmoothing in vae via generalized variance parameterization. *Neurocomputing*, 509, 137–156.
- Thuburn, J., Kent, J., & Wood, N. (2014). Cascades, backscatter and conservation in numerical models of two-dimensional turbulence. *Quarterly Journal of the Royal Meteorological Society*, 140(679), 626–638.
- Vallis, G. K. (2017). *Atmospheric and oceanic fluid dynamics*. Cambridge University Press.
- Wang, P., Yuval, J., & O’Gorman, P. A. (2022). Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, 14(10), e2022MS002984.
- Wilks, D. S. (2005). Effects of stochastic parametrizations in the lorenz’96 system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(606), 389–407.
- Yang, D., Hong, S., Jang, Y., Zhao, T., & Lee, H. (2019). Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*.
- Yuval, J., & O’Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature communications*, 11(1), 3295.
- Zacharuk, M., Dolaptchiev, S. I., Achatz, U., & Timofeyev, I. (2018). Stochastic subgrid-scale parameterization for one-dimensional shallow-water dynamics using stochastic mode reduction. *Quarterly Journal of the Royal Meteorological Society*, 144(715), 1975–1990.
- Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376.
- Zanna, L., Mana, P. P., Anstey, J., David, T., & Bolton, T. (2017). Scale-aware

deterministic and stochastic parametrizations of eddy-mean flow interaction.

Ocean Modelling, 111, 66–80.

Zhang, B., Xiong, D., Su, J., Duan, H., & Zhang, M. (2016). Variational neural machine translation. *arXiv preprint arXiv:1605.07869*.