JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2023MS003681

Special Section:
Machine learning application to Earth system modeling

Generative Data-Driven Approaches for Stochastic Subgrid Parameterizations in an Idealized Ocean Model

Pavel Perezhogin[1], Laure Zanna[1], and Carlos Fernandez-Granda[1,2]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA, [2]Center 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

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P. Perezhogin,
pperezhogin@gmail.com

Citation:

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model. *Journal of Advances in Modeling Earth Systems*, *15*, e2023MS003681. https://doi.org/10.1029/2023MS003681

Received 15 FEB 2023
Accepted 28 SEP 2023

Author Contributions:

Conceptualization: Laure Zanna, Carlos Fernandez-Granda
Formal analysis: Pavel Perezhogin
Funding acquisition: Laure Zanna, Carlos Fernandez-Granda
Investigation: Pavel Perezhogin

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, https://doi.org/10.1029/2021ms002534) 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 an 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°, 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 physical 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 (Gent & Mcwilliams, 1990; Redi, 1982) were designed to describe their mean effect on the large-scale flow. These parameterizations are

Methodology: Pavel Perezhogin, Laure Zanna, Carlos Fernandez-Granda
Project Administration: Laure Zanna, Carlos Fernandez-Granda
Software: Pavel Perezhogin
Supervision: Laure Zanna, Carlos Fernandez-Granda
Validation: Pavel Perezhogin
Visualization: Pavel Perezhogin
Writing – original draft: Pavel Perezhogin
Writing – review & editing: Laure Zanna, Carlos Fernandez-Granda

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 & 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 (Bachman, 2019; Bachman et al., 2017, 2018; Berloff, 2018; Grooms et al., 2015; Jansen & Held, 2014; Jansen et al., 2019; Juricke et al., 2020; Mana & Zanna, 2014; Pearson et al., 2017; Thuburn et al., 2014; Zanna et al., 2017).

Machine Learning (ML) methods have recently gained traction as a new direction for developing subgrid eddy parameterizations in geophysics and turbulence (Beck et al., 2019; Beucler et al., 2021; Bolton & Zanna, 2019; Guan, Chattopadhyay, et al., 2022; Maulik et al., 2019; Rasp et al., 2018; Shamekh et al., 2023; Wang et al., 2022; Yuval & O'Gorman, 2020). 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 data set 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), and 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 (Berner et al., 2017; Christensen & Zanna, 2022; Gerard, 2007). 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 (Berner et al., 2012, 2017; Christensen et al., 2017; Juricke et al., 2017; Palmer, 2000). The two simplest stochastic parameterizations are Stochastically perturbed parameterization tendency (SPPT; Andrejczuk et al., 2016; Buizza et al., 1999; 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 & 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 (Arnold et al., 2013; Berner et al., 2009; Schumann, 1995; Wilks, 2005) correlations of stochastic forcing, its non-Gaussian distribution (Mana & Zanna, 2014; Zanna et al., 2017) and its dependence on the resolved flow (multiplicative noise, Arnold et al., 2013; Sura et al., 2005; 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 & 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; Alcala & Timofeyev, 2021; Gagne et al., 2020; B. T. Nadiga et al., 2022). In addition, generative models can be trained and tested using the same data sets, 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 (Chasnov, 1991; Frederiksen & Davies, 1997; Leslie & Quarini, 1979) in the correct band of scales. Additionally, generative models accurately simulate large-scale kinetic energy backscatter (Jansen & Held, 2014; Thuburn et al., 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 evaluations of the proposed methodology to build subgrid parameterization for a range of resolutions.

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:

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

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

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 Kroneker 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 kinetic and total energy per unit mass are respectively given by:

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

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

where $\langle \cdot \rangle$ is 2D spatial averaging.

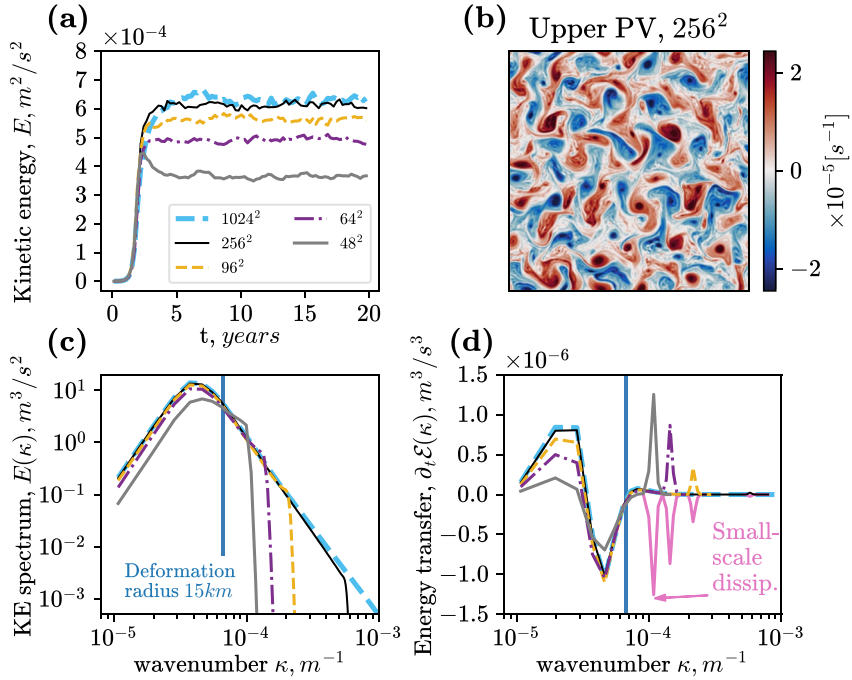

Figure 1. Reference simulations at five different resolutions: (a) kinetic energy (Equation 3) as a function of time, (b) snapshot of the potential vorticity in the model used for diagnosing subgrid forcing, (c) the spectral density of kinetic energy normalized as $E = \int E(\kappa)d\kappa$, (d) total energy transfer from nonlinear advection $\frac{1}{H}\sum_{m=1}^{2} H_m \mathrm{Re}\left(\hat{\psi}_m^* \widehat{\nabla(\mathbf{u}_m q_m)}\right)$; see Text S1 in Supporting Information S1 for the derivation of spectral energy budget. Coarse models fail to reproduce the energy cycle when their resolution is insufficient to resolve the deformation radius $\kappa = r_d^{-1} = (15\mathrm{km})^{-1}$ (see blue vertical line). Pink lines in panel (d) show the dissipation provided by the ssd term, which causes a spurious forward energy cascade (positive spikes in energy transfer).

2.1.1. Small-Scale Dissipation and Numerical Scheme

We ignore "molecular viscosity" in the governing Equations 1 and 2, as typically done in ocean modeling (Griffies & Hallberg, 2000) due to negligible contribution. Numerical noise is suppressed by the small-scale dissipation term (ssd; Ross et al., 2023), which smoothes the potential vorticity after every time step ($q_m \rightarrow \mathrm{ssd}(q_m)$) using an "exponential cut-off" filter (Arbic & Flierl, 2003; Canuto et al., 1988; LaCasce, 1996). The filter is defined by its spatial spectrum $\widehat{ssd}$. The highest wavenumbers ($\kappa > \kappa_c$) are multiplied by the transfer function:

$$\widehat{ssd}(\kappa) = e^{-23.6(\Delta x)^4(\kappa-\kappa_c)^4}, \tag{5}$$

where $\kappa = \sqrt{k^2 + l^2}$ is the radial wavenumber, k and l are zonal and meridional wavenumbers, respectively, Δx is the grid step, $\kappa_c = 0.65\kappa_{\max}$ and $\kappa_{\max} = \pi/\Delta x$ is the Nyquist frequency. The ssd term attenuates the KE spectrum near the grid scale: see the abrupt decrease of KE density in high wavenumbers in Figure 1c. The pink lines in Figure 1d show the corresponding energy dissipation. At statistical equilibrium, the small-scale dissipation is balanced by a spurious positive energy transfer which disappears with an increase in resolution (Figure 1d). The ssd term is kept in the unparameterized and parameterized models.

Equations 1 and 2 are solved using the third-order Adamsh-Bashford time integration scheme and a pseudo-spectral spatial approximation without dealiasing (Ross et al., 2023).

2.1.2. Model Setup

The QG system 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 (48^2), and it allows to simulate similar energy growth in the transition from laminar to turbulent regimes at different grid resolutions, see Figure 1a. Our default

Table 1
Parameters of the Quasi-Geostrophic Model in Online Simulations (Ross et al., 2023)

QG configuration	"Eddy"	"Jet"
Integration time	20 years	–(same)
Ensemble size	10 runs	–
Domain size ($L \times W$)	1,000 km × 1,000 km	–
Boundary conditions	Periodic	–
Upper layer thickness (H_1)	500 m	–
Ocean depth ($H = H_1 + H_2$)	2,500 m	5,500 m
Bottom drag (r_{ek})	5.787×10^{-7} s^{-1}	7×10^{-8}s^{-1}
Differential rotation (β)	1.5×10^{-11} (m s)$^{-1}$	10^{-11}(m s)$^{-1}$
Deformation radius $\left(r_d = \sqrt{\frac{g'}{f_0^2}\frac{H_1 H_2}{H}}\right)$	15 km	–
Mean flow (U_1, U_2)	(0.025, 0 m/s)	–
Velocity scale ($\sqrt{2E}$)	≈0.035 m/s	≈0.02 m/s

Grid parameters	Resolution	Grid step (Δx) (km)	Time step (Δt) (hr)
High resolution	256 × 256	3.9	1
Coarse models	96 × 96	10.4	2
	64 × 64	15.6	4
	48 × 48	20.8	1,2,4,8

Note. In our paper, unless otherwise mentioned, we use the "eddy" configuration.

configuration is called "eddy" (Table 1). An additional "jet" configuration is dominated by meandering jets and used for generalization studies (Ross et al., 2023).

Mesoscale eddies emerge on a spatial scale determined by the deformation radius $r_d = \sqrt{\frac{g'}{f_0^2}\frac{H_1 H_2}{H}}$ (Salmon, 1980; Vallis, 2017). Following Hallberg (2013), the model grid step should be small enough $r_d/\Delta x \geq 2$ to resolve mesoscale eddies on the grid. High-resolution models (256^2, $1{,}024^2$) satisfy this criterion. In the remaining text, we will use the 256^2 model as a reference simulation because a further increase of resolution to $1{,}024^2$ almost does not change the statistics of the large scales $\kappa \lesssim r_d^{-1}$, see spectra left to the blue line in Figure 1. Note that fast convergence of statistics of large scales is achieved due to the application of the pseudo-spectral method with highly scale-selective dissipation.

The resolution of coarse models (48^2, 64^2, 96^2) is chosen such that $r_d/\Delta x < 2$. These models fail to reproduce various statistical characteristics (Hallberg, 2013; Hewitt et al., 2020), including kinetic energy (KE), spectrum of KE, and energy transfer (Figures 1a, 1c, and 1d). 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

We follow the Large eddy simulation (LES; Sagaut, 2006) approach and define a spatial filtering and coarse-graining operator $\overline{(\cdot)}$ which maps the solution of the high-resolution model (q) to a coarse grid. The filtered and coarse-grained field is denoted as $\overline{q}$. The time evolution equation for $\overline{q}$ is derived below and contains a new term that describes the interaction with unresolved eddies. This term is not available at the coarse resolution and needs to be parameterized.

Most of our analysis is performed with the so-called "Sharp" filter introduced in Ross et al. (2023). The Sharp filter combines the spectral cut-off coarse-graining followed by the model filter ssd (Equation 5) applied on a coarse grid. The spectral cut-off coarse-graining discards the wavenumbers above the Nyquist frequency ($\pi/\Delta x$) of the coarse grid.

An additional filtering and coarse-graining operator studied is the combination of the spectral cut-off followed by the Gaussian filter (Guan, Chattopadhyay, et al., 2022). The transfer function of the Gaussian filter is:

$$\widehat{G}(\kappa) = e^{-\kappa^2(2\Delta x)^2/24}, \tag{6}$$

where Δx is the coarse grid step and $2\Delta x$ is the filter width. For brevity, we refer to the combination of cut-off and Gaussian filters as "Gaussian."

Applying the filter $\overline{(\cdot)}$ to the governing Equations 1 and 2, we obtain a set of governing equations for the filtered and coarse-grained solution:

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

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

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

$$S = \nabla \cdot \left(\overline{\mathbf{u}}\ \overline{q} - \overline{\mathbf{u}q}\right), \tag{9}$$

which needs to be parameterized. To simplify notation, we will omit the index m for the subgrid forcing and related variables. The dissipation term ssd on a coarse grid in Equation 7 is added a-posteriori to ensure the numerical stability of the simulations. In deriving Equation 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 the form of divergence of flux, so we include numerical approximation errors into the definition of subgrid forcing (Chow & Moin, 2003; Ghosal, 1996; Gullbrand & Chow, 2003).

2.3. Subgrid Forcing Data Set

The data set 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 hr and save snapshots every 1,000 hr, for a total of 86 snapshots. The training data set consists of 250 runs, each corresponding to a different random initial condition, for a total of 21,500 snapshots. The validation and testing data sets 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 ($\overline{q}$, $\overline{\mathbf{u}}$) and subgrid forcing (Equation 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 predicting subgrid forcing, which can be used to build data-driven stochastic parameterizations.

Conventional subgrid parameterizations establish a functional relationship between the subgrid forcing (S, Equation 9) and the resolved flow $\left(\overline{q}\right)$ in the form of $S \approx \widetilde{S}\left(\overline{q}\right)$. 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 $\left(S \sim \rho\left(S|\overline{q}\right)\right)$.

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 (Goodfellow et al., 2014; Kingma & Welling, 2013). 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

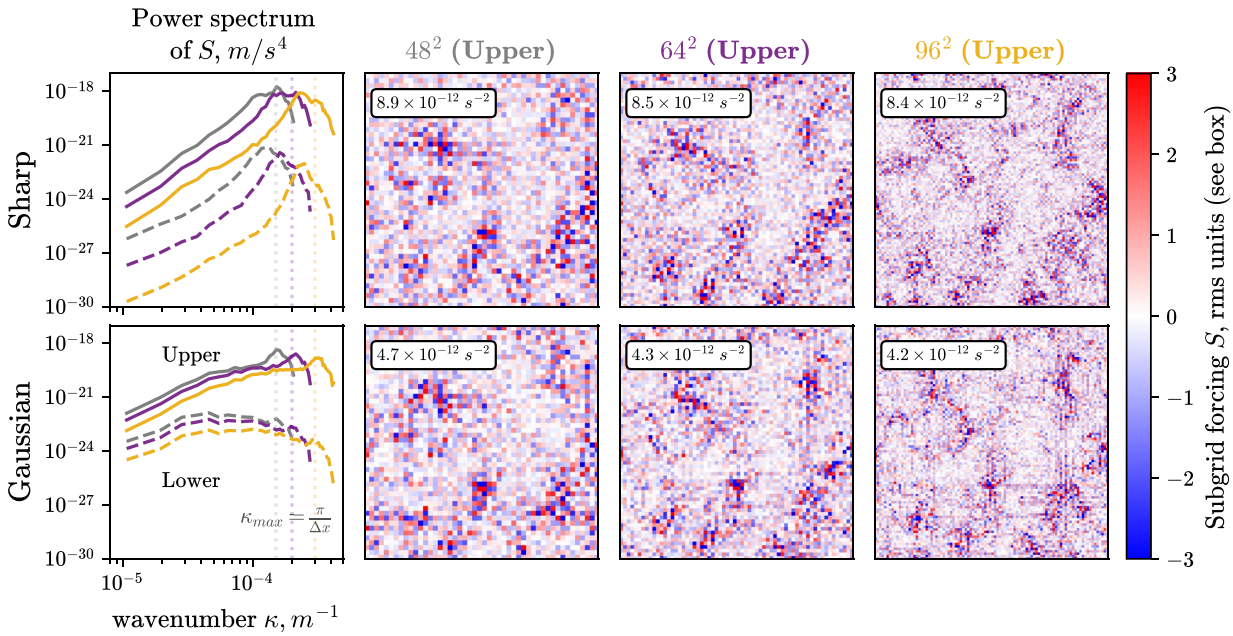

Figure 2. Subgrid forcing (S, Equation 9) at different resolutions was diagnosed using a Sharp filter (top row), or a Gaussian filter (bottom row). Left: power spectrum of S for the upper fluid layer (solid lines) and lower fluid layer (dashed lines). Colors: 48^2 (gray), 64^2 (violet), 96^2 (yellow). 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.

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.

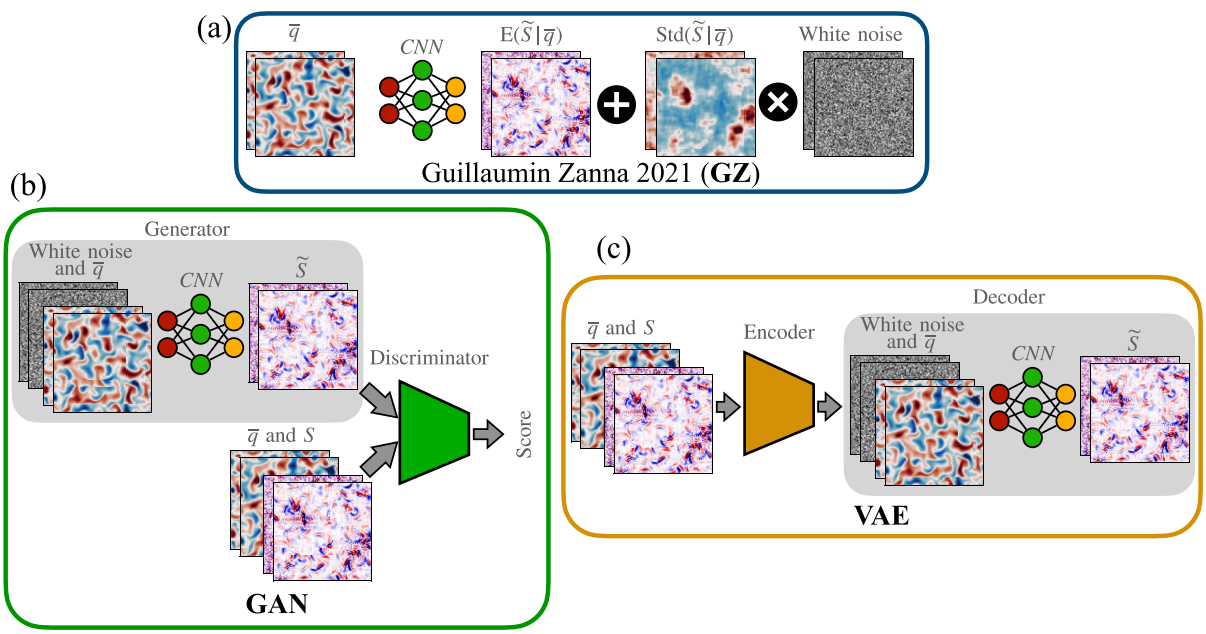

Figure 3. Schematic of three stochastic subgrid models attempting to sample from the conditional distribution $\rho(S|\overline{q})$: (a) GZ model, (b) GAN model and (c) VAE model. GZ model predicts uncorrelated stochastic residuals, but generative models (GAN and VAE) transform white noise using a mapping 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.

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 and 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:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{2n^2}\|S - \widetilde{S}_{\theta}^{\mathrm{mean}}(\overline{q})\|_2^2, \tag{10}$$

where $\widetilde{S}_{\theta}^{\mathrm{mean}}(\overline{q})$ is the output of a neural network with parameters denoted by θ, which receives $\overline{q}$ as an input. $S, \widetilde{S}_{\theta}^{\mathrm{mean}}, \overline{q} \in \mathbb{R}^{2\times n\times n}$ 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 $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_D^2}$. The loss function is minimized over a training set consisting of samples of the resolved flow $\overline{q}$ and the corresponding high-resolution forcing S obtained as described in Section 2.3. Minimization of $\mathcal{L}_{\mathrm{MSE}}$ yields an optimal set of parameters θ^* and a corresponding model which we denote as $\widetilde{S}^{\mathrm{mean}}(\overline{q})$.

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

$$\mathcal{L}_{\mathrm{VAR}} = \frac{1}{2n^2}\|r^2 - \widetilde{S}_{\phi}^{\mathrm{var}}(\overline{q})\|_2^2, \tag{11}$$

where $\widetilde{S}_{\phi}^{\mathrm{var}}(\overline{q})$ is the output of a neural network with parameters denoted by ϕ, which receives $\overline{q}$ 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 $\widetilde{S}^{\mathrm{var}}(\overline{q})$. The architecture of the CNNs (convolutional neural network) used to parameterize the mappings $\widetilde{S}^{\mathrm{mean}}(\overline{q})$ and $\widetilde{S}^{\mathrm{var}}(\overline{q})$ and details of the optimization algorithm are provided in Appendix A3.

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

$$\widetilde{S}(z, \overline{q}) = \widetilde{S}^{\mathrm{mean}}(\overline{q}) + \left(\widetilde{S}^{\mathrm{var}}(\overline{q})\right)^{1/2} \cdot z, \tag{12}$$

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 $(\overline{q})$ denoted by $\rho(S|\overline{q})$, where both variables are considered as 3D fields, $S, \overline{q} \in \mathbb{R}^{2\times n\times n}$. The mentioned distribution is defined implicitly by the data set of pairs of S and $\overline{q}$. 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 & Osindero, 2014), which informs both networks with the conditional variable. Specifically, the generator transforms the latent noise variable $z \in \mathbb{R}^{2\times n\times n}$ and PV field to the subgrid forcing, see Figure 3b:

$$\widetilde{S} = G(z, \overline{q}), \tag{13}$$

where the mapping G is parameterized by the same CNN as in the GZ model, and the tensors z and $\overline{q}$ are concatenated to form a single input tensor of size $\mathbb{R}^{4\times n\times n}$. The discriminator $D(S, \overline{q})$ returns a score (scalar value) given

a pair of subgrid forcing and PV field concatenated to form a single input tensor. The discriminator D is parameterized by a neural network similar to the one used in DCGAN (Radford et al., 2015).

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:

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

where $\mathbb{E}$ is the mathematical expectation over the training samples. The discriminator estimates the Wasserstein-1 ($\mathcal{W}_1$) distance between the distributions $\rho(S|\overline{q})$ and $\rho\left(\widetilde{S}|\overline{q}\right)$, which quantifies how close they are. This distance has a complicated definition and cannot be computed directly but can be shown to satisfy the equality (Adler & Öktem, 2018)

$$\mathcal{W}_1\left(\rho(S|\overline{q}),\rho\left(\widetilde{S}|\overline{q}\right)\right) = \max_D \mathbb{E}\left[D(S,\overline{q}) - D\left(\widetilde{S},\overline{q}\right)\right],$$

which is the inner optimization problem in Equation 14. The optimization of the generator ($\min_G$ in Equation 14) is designed to minimize the distance between the true and generated distributions.

Solving the optimization Problem 14 may lead to the mode collapse phenomenon when the generator ignores the latent variable z: for every coarse field $\overline{q}$ the model may produce a single fixed subgrid forcing (Isola et al., 2017; Mao et al., 2019; Ohayon et al., 2021; Yang et al., 2019). To overcome mode collapse, we apply a technique proposed by Adler and Öktem (2018): feeding multiple generator outputs $\widetilde{S}$ to the discriminator for a given input $\overline{q}$. Identical outputs are readily detected and penalized by the discriminator. We define the GAN loss function accounting for this technique in Appendix A1. The architecture of the networks parameterizing G and D and the optimization algorithm are described in Appendix A3.

Once trained, the GAN generator Equation 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 & Welling, 2013) to sample from the conditional distribution $\rho(S|\overline{q})$. 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 (Doersch, 2016; Mishra et al., 2018; Pagnoni et al., 2018; Sohn et al., 2015; Zhang et al., 2016): the decoder maps the latent noise and conditional variable $\overline{q}$ to the subgrid forcing, see Figure 3c:

$$\widetilde{S} \sim \rho_\theta\left(\widetilde{S}|z,\overline{q}\right), \tag{15}$$

where free parameters are denoted by θ and we emphasize that the mapping is probabilistic. The probabilistic encoder with free parameters ϕ is denoted as $z \sim q_\phi(z|S,\overline{q})$. 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):

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

where $D_{\mathrm{KL}}(p(x),q(x)) = \mathbb{E}_{p(x)}\ln\frac{p(x)}{q(x)}$ 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 the encoder and

decoder with CNNs predicting mean and variance of Gaussian distributions. The resulting loss function is equivalent to a regularized MSE as explained in more detail in Appendix A2.

The mean channel of the Gaussian decoder (Equation 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 data set 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 resolution (48^2, 64^2, 96^2), we train three machine learning models: GZ, GAN, and VAE. 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 subgrid models trained and evaluated on the data set obtained with the Sharp filter. Similar results for the Gaussian filter are shown in Figures S1–S3 of the Supporting Information S1.

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 $\overline{q}$—it is conditional mean denoted as $\mathrm{E}\left(\widetilde{S}|\overline{q}\right)$. The deterministic part of the GZ model is given by the mean channel $\widetilde{S}^{\mathrm{mean}}\left(\overline{q}\right)$. For GAN and VAE models, we fix the conditional variable $\overline{q}$ and sample many realizations of subgrid forcing prediction $\widetilde{S}$ by sampling in the latent space z. We then average over 1,000 realizations to estimate the deterministic part $\mathrm{E}\left(\widetilde{S}|\overline{q}\right)$ similarly to Adler and Öktem (2018). We define the stochastic residual as a deviation of $\widetilde{S}$ from its deterministic part, $\widetilde{r} = \widetilde{S} - \mathrm{E}\left(\widetilde{S}|\overline{q}\right)$. Each realization of $\widetilde{S}$ provides a unique realization of $\widetilde{r}$. The error of the deterministic part $\left(r = S - \mathrm{E}\left(\widetilde{S}|\overline{q}\right)\right)$ can be used to study the statistical properties of the stochastic residuals $\widetilde{r}$. We refer to r as a *true residual*; in the literature, it is often simply called "residual," see Wilks (2005), Arnold et al. (2013), and Gagne et al. (2020). Wilks (2005) proposed a statistical model of stochastic residuals where free parameters were estimated from the time series of true residuals. Additional works that compare the statistical properties of stochastic $\widetilde{r}$ and true r residuals include Shutts and Palmer (2007), Arnold et al. (2013), Mana and Zanna (2014), Gagne et al. (2020), Agarwal et al. (2021), and Guillaumin and Zanna (2021).

In Figure 4 we show predictions of the stochastic subgrid models. The deterministic part $\left(\mathrm{E}\left(\widetilde{S}|\overline{q}\right)\right)$ is similar for three stochastic models. The rightmost column shows a single realization of the stochastic residual. The stochastic residual for GZ model looks like uncorrelated spatial white noise in contrast to the true residual. The stochastic 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:

$$\mathrm{Std}\left(\widetilde{S}|\overline{q}\right) = \sqrt{\mathrm{E}\left(\left(\widetilde{S} - \mathrm{E}\left(\widetilde{S}|\overline{q}\right)\right)^2 \middle| \overline{q}\right)} = \sqrt{\mathrm{E}\left(\widetilde{r}^2|\overline{q}\right)}.$$

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 structures for all three stochastic models.

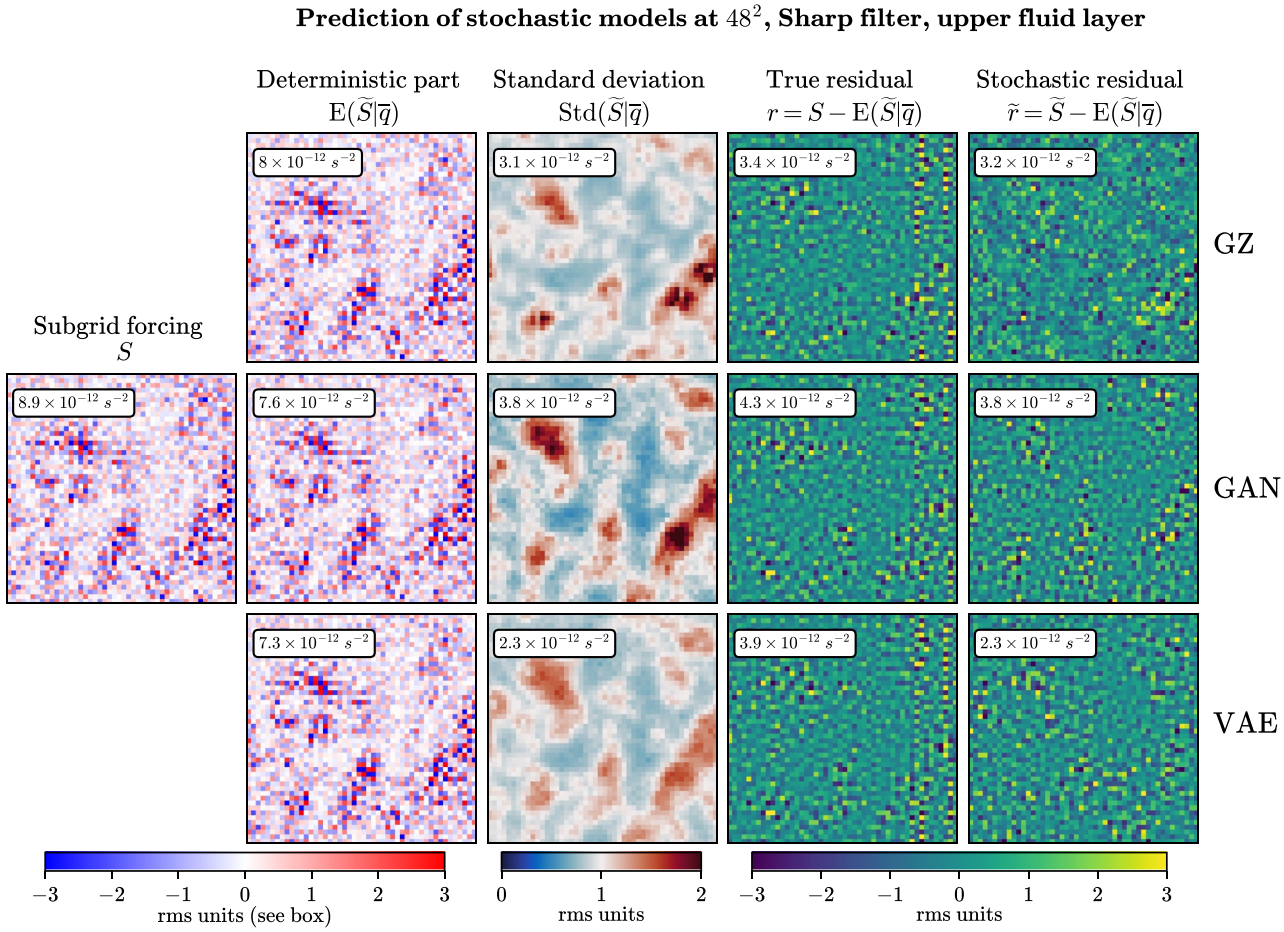

Figure 4. Prediction of subgrid forcing (S) by stochastic models at a fixed conditional variable $\overline{q}$ on the testing data set: GZ in upper row, GAN in middle row and VAE in lower row. Stochastic residual is shown for a single realization of the latent vector z. Stochastic and true residuals should be statistically similar for an accurate model. The root mean square (rms) value of each field is shown in the box.

We use the spatial power spectrum to analyze the spatial correlation. In Figure 5a 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 5b) 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 on small spatial scales, that is, large wavenumbers, see Figure 5b. We explain it by the well-known issue of VAE architecture to predict locally smooth images (Takida et al., 2022).

An important property of subgrid forcing in QG turbulence is an ability to energize turbulence on a coarse grid, that is, kinetic energy backscatter (Jansen & Held, 2014). There are two popular approaches to simulate backscatter: stochastic residuals near the grid scale (Chasnov, 1991; Frederiksen & Davies, 1997; Grooms et al., 2015; Leslie & Quarini, 1979; Schumann, 1995) and mean energy injection in large scales (Frederiksen et al., 2003; Graham & Ringler, 2013; Jansen & Held, 2014; Juricke et al., 2020; Kraichnan, 1976; Thuburn et al., 2014). 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

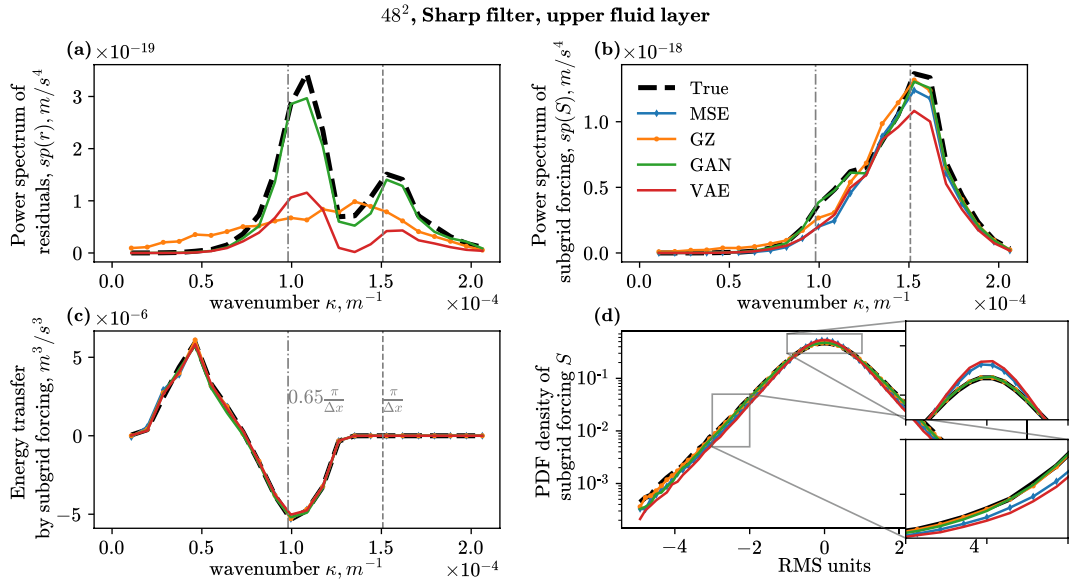

Figure 5. Offline analysis of stochastic subgrid models (GZ, GAN, VAE): (a) power spectrum of stochastic residuals and (b) subgrid forcing; (c) energy transfer $\left(-\mathrm{Re}\left(\widehat{\overline{\psi}}^{*}\,\widehat{S}\right)\right)$ and (d) marginal PDF of subgrid forcing. MSE is the deterministic subgrid model given by the mean channel of GZ model. On panel (a) we show the true residuals for the GAN model in black dashed line; true residuals for the other two models (GZ, VAE) are similar and not shown for conciseness.

large-scale energy injection (Figure 5c), 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 (Maulik & San, 2017; Pawar et al., 2020). 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 5d. The VAE model is similar to the baseline MSE in this characteristic.

We further emphasize that the stochastic residuals of the proposed subgrid models statistically emulate the error of the deterministic prediction $\left(r = S - \mathrm{E}\left(\widetilde{S}|\overline{q}\right)\right)$. Consequently, the spread of the stochastic residuals should be sensitive to the accuracy of the deterministic prediction. The deterministic prediction may be affected by many factors including model expressivity (number of hidden layers) and amount of training data. In Figure 6a we show that by reducing the number of hidden layers in CNNs, the quality of the deterministic part deteriorates (MSE increases), accompanied by an increase of the spread of the stochastic residuals for all three subgrid models (GZ,

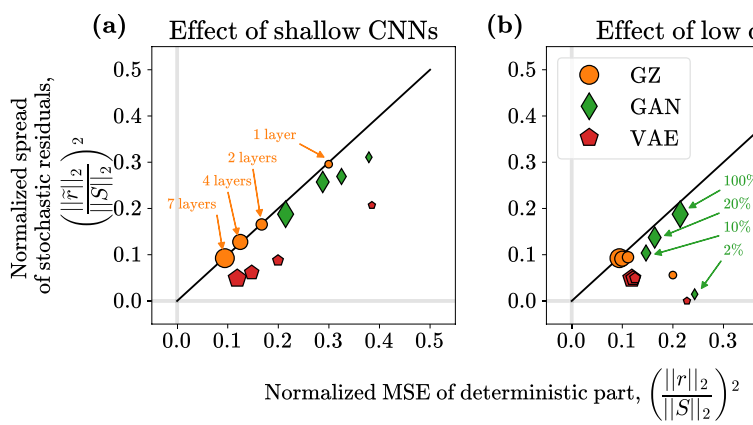

Figure 6. Sensitivity of the model spread to the accuracy of the prediction of the deterministic part. (a) Effect of reducing the number of hidden layers in CNN representing the stochastic subgrid model. (b) Effect of using a fraction of the training data set (in %). The number of (a) hidden layers and (b) amount of training data are indicated by the size of the markers. Data set: 48^2, Sharp filter.

Table 2
Metrics for Offline Analysis of Stochastic Subgrid Model $\widetilde{S}$: RMSE of the Deterministic Part ($\mathcal{L}_{\text{rmse}}$), RMSE in the Spectrum of the Full Subgrid Forcing ($\mathcal{L}_{\text{S}}$), RMSE in the Spectrum of Stochastic Residuals ($\mathcal{L}_{\text{r}}$) and Spread of the Samples of Conditional Distribution $\left(\sigma^2_{\text{spread}}\right)$

Metric	$\mathcal{L}_{\text{rmse}}$	$\mathcal{L}_{\text{S}}$	$\mathcal{L}_{\text{r}}$	σ^2_{spread}
Expression	$\frac{\|S-\mathrm{E}(\widetilde{S}\|\overline{q})\|_2}{\|S\|_2}$	$\frac{\|sp(S)-sp(\widetilde{S})\|_2}{\|sp(S)\|_2}$	$\frac{\|sp(r)-sp(\widetilde{r})\|_2}{\|sp(r)\|_2}$	$\frac{\|\widetilde{r}\|_2^2}{\|r\|_2^2}$
Optimal value	0	0	0	1
Unparameterized model	1	1	1	0
Quality of	Deterministic part	Full forcing	Residuals	Residuals

Note. We denote computation of isotropic power spectrum as $sp(\cdot)$.

GAN, VAE). Under the assumption of statistical similarity between r and , the MSE should be equal to the spread of stochastic residuals, that is, $\|r\|_2^2 = \|\widetilde{r}\|_2^2$, represented by the diagonal black line. Our formulation of the GZ model explicitly fits the empirical variance (r^2) on the training set, which is why the corresponding markers are on the black line. The GAN model slightly underestimates the spread. The VAE model is the least accurate and significantly underestimates the spread, see Figure 6a. In Figure 6b we reduce the size of the training data set and show how the stochastic models perform on the testing data set. The GZ and VAE models can be trained with the least amount of training data: the model quality (MSE and spread) almost does not change when 10% of the training data is used. On the contrary, the MSE and spread of the GAN model change even when 20% of the training data is used. Note that while the MSE error for the GAN model improves with a slight reduction of training data, the prediction of the power spectrum of residuals deteriorates (not shown). Finally, all stochastic models behave similarly in the limit of very low amount of training data (2%): the MSE of the deterministic part is too large, and the spread of stochastic residuals is too low. This indicates that the subgrid parameterizations memorize the training data set and effectively become deterministic.

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 ($\mathcal{L}_{\text{S}}$ and $\mathcal{L}_{\text{r}}$) in order to evaluate to what extent the models capture the corresponding spatial structure.

In Figure 7 we report the evaluation of the offline metrics for the different models for a range of resolutions. The upper row provides metrics on the test data set 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 model). The GAN and GZ models correctly predict spread of stochastic residuals $\sigma^2_{\text{spread}} \approx 1$, but the VAE model underestimates spread $\sigma^2_{\text{spread}} \approx 0.35$. The GAN model clearly outperforms the rest in predicting the spectra of the subgrid forcing $\mathcal{L}_{\text{S}}$ and the residuals $\mathcal{L}_{\text{r}}$. The VAE model on the contrary has high errors $\mathcal{L}_{\text{S}}$ and $\mathcal{L}_{\text{r}}$ because it predicts oversmoothed samples with reduced diversity.

In the lower row of Figure 7, we evaluate the generalization ability of the models by computing the offline metrics on the data set corresponding to a different turbulence regime, where flow is dominated by meandering jets (Table 1), and which is therefore systematically different from the training data. GZ model considerably overestimates the spread of the residuals $\left(2 < \sigma^2_{\text{spread}} < 10\right)$, and it deteriorates the spectral metrics ($\mathcal{L}_{\text{S}}$ and $\mathcal{L}_{\text{r}}$). The VAE model demonstrates the best generalization capabilities to the jet configuration: it has reasonable spread $\sigma^2_{\text{spread}} \approx 0.8$, and outperforms other models in the error of the deterministic prediction $\mathcal{L}_{\text{rmse}}$, the quality of the subgrid forcing $\mathcal{L}_{\text{S}}$ and residuals $\mathcal{L}_{\text{r}}$. The GAN model generalizes better than GZ for most of the metrics, without reaching the performance of the VAE model. As we noted above, the VAE model underestimates the power spectrum of the subgrid forcing producing smooth images. This property likely facilitates the generalization to the jet data set because the RMS value of the subgrid forcing on the jet data set is twice as small as on the eddy data set. We observe similar generalization results for the Gaussian filter, see Figure S3 in Supporting Information S1.

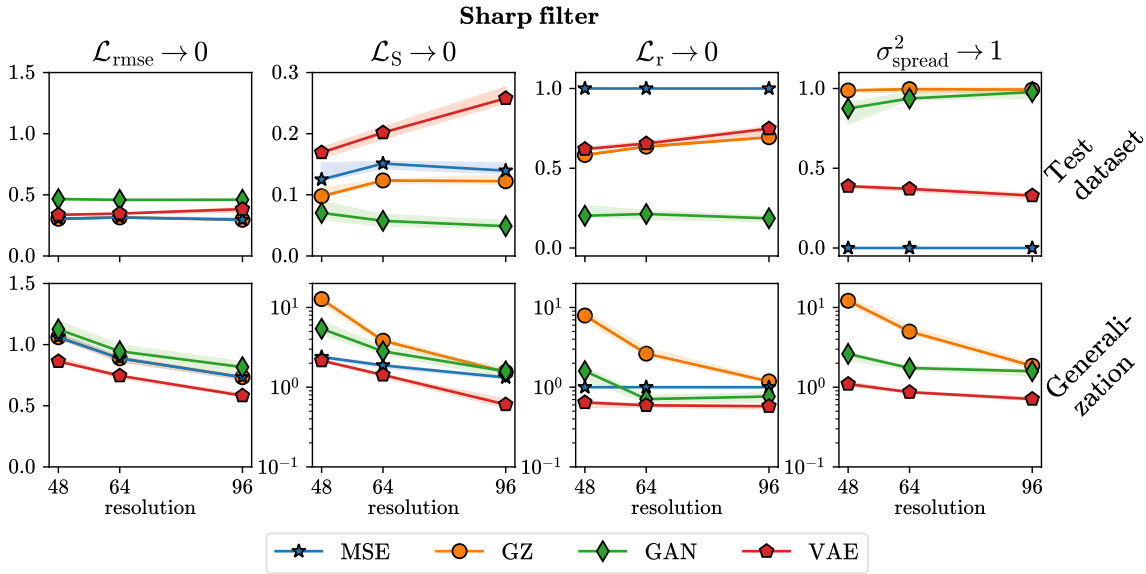

Figure 7. Offline metrics from Table 2 on the testing data set in the upper row and generalization to configuration with jets (Table 1) in the lower row. Optimal values are given with arrows. Each model is trained 5 times with different random seeds. The shading area shows min-max values among training realizations, markers show median value.

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 8. 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.

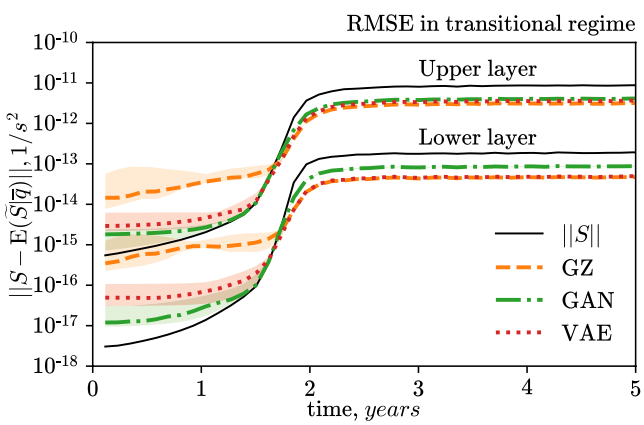

Figure 8. Offline analysis of RMSE of deterministic part in transitional regime ($t < 2$ years) on the test data set. Norm $\|\cdot\|$ is given per one grid point. Shading corresponds to different training realizations of the same model. Sharp filter, resolution 48^2. Each line is given twice: for the lower and upper fluid layer.

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 $\widetilde{S}$ to replace the true subgrid forcing S in the governing equation for the coarsegrained dynamics Equation 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 ("eddy" configuration in this section). Compared to the generation of the training data, we run experiments for twice as long (20 years). 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 1,200. 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.

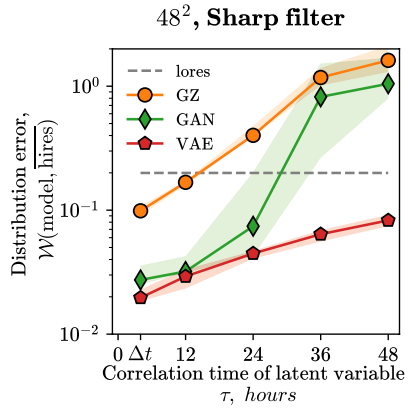

Figure 9. Online distributional metric (Equation 17) as a function of correlation time τ of latent variable z for three stochastic subgrid models (GZ, GAN, and VAE) at coarse resolution 48^2 with respect to high-resolution model (256^2). Lores is model on a coarse grid without parameterization. Shading area shows min-max values among training realizations, and markers show median value. Time step of the numerical integration Δt is 4 hr.

Before passing the subgrid forcing prediction into the governing equation, we subtract the spatial mean in each fluid layer to ensure the conservation of PV. Without this subtraction, some experiments become unstable when the spatial mean exceeds the spatial RMS of the PV of the solution. In the conclusion, we discuss possible options to include the conservation of PV in suggested subgrid models.

Among mentioned experiments, there were a few unstable simulations at resolution 96^2: 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), and 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 $\left(\frac{1}{2}|\mathbf{u}_m|^2\right)$ and relative enstrophy $\left(\frac{1}{2}|\text{curl}(\mathbf{u}_m)|^2\right)$, 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 $\left(F_{\overline{\text{hires}}}\right)$ is given by the average of normalized errors:

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

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

An additional metric based on spectral characteristics is reported in Supporting Information S1 (Text S3 and Figure S4).

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 $\text{E}(z(t)z(t + n\Delta t)) = (1 - \Delta t/\tau)^n$ (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 (Equation 17) as a function of correlation time is reported in Figure 9. The optimal online metric corresponds to $\tau = \Delta t$, which justifies our method of sampling (white noise). This result is consistent with our training procedure where we were sampling the latent variable independently for every time moment. The autoregressive process can be introduced during a training stage if

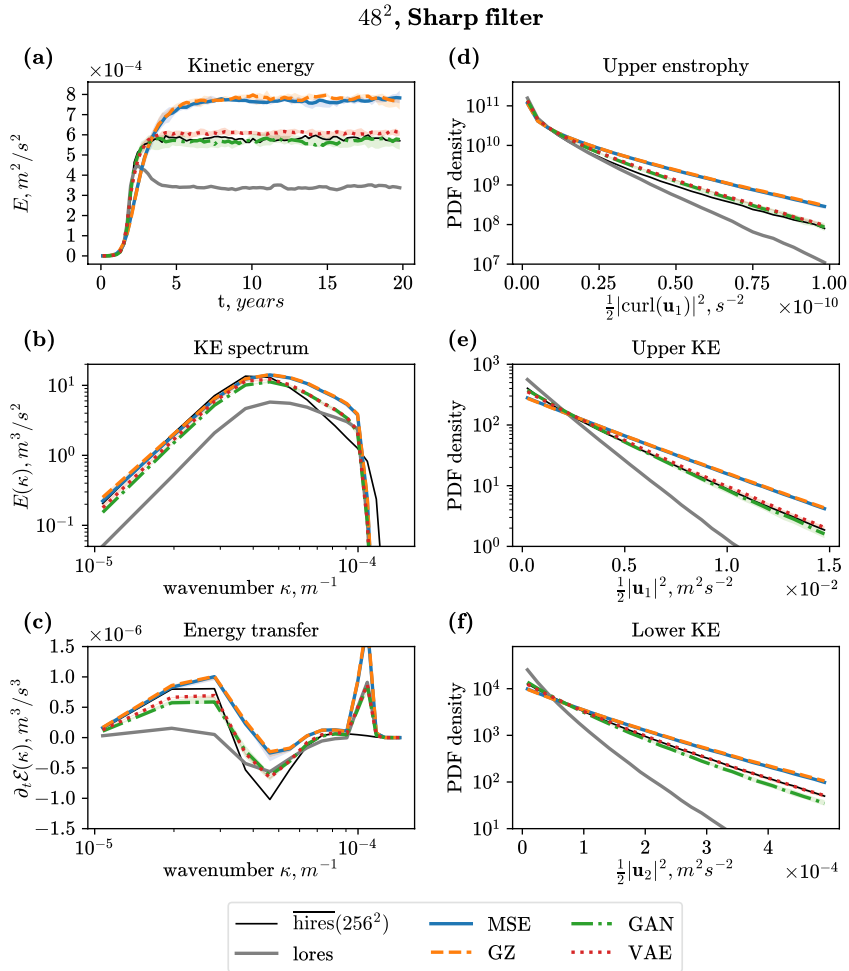

Figure 10. Online simulations with parameterized models (MSE, GZ, GAN, and VAE) and unparameterized model (lores) at coarse resolution 48^2. $\overline{\text{hires}}$ is filtered and coarsegrained high-resolution model (256^2). MSE is a deterministic subgrid model given by the mean channel of the GZ model. Energy transfer on panel (c) gives a sum of contributions from the resolved advection and subgrid model: $\frac{1}{H}\sum_{m=1}^{2} H_m \mathrm{Re}\left(\widehat{\overline{\psi}}_m^* \widehat{\nabla(\overline{\mathbf{u}}_m \overline{q}_m)} - \widehat{\overline{\psi}}_m^* \widehat{S}_m\right)$; see derivation in Text S2 of the Supporting Information S1. Shading area shows min-max values among training realizations, and lines show median value. The time step Δt is 2 hr.

consider sampling from the distribution conditional on the previous time moment, that is, $\rho\left(S(t)|\overline{q}(t), S(t-\Delta t)\right)$ (Gagne et al., 2020). We postpone it for future studies.

5.3. Results

In Figure 10 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 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 11. At time step $\Delta t = 2$ hr baseline models (MSE and GZ) have too many small eddies, and at time step $\Delta t = 1$ hr the flow becomes unphysical and overenergized.

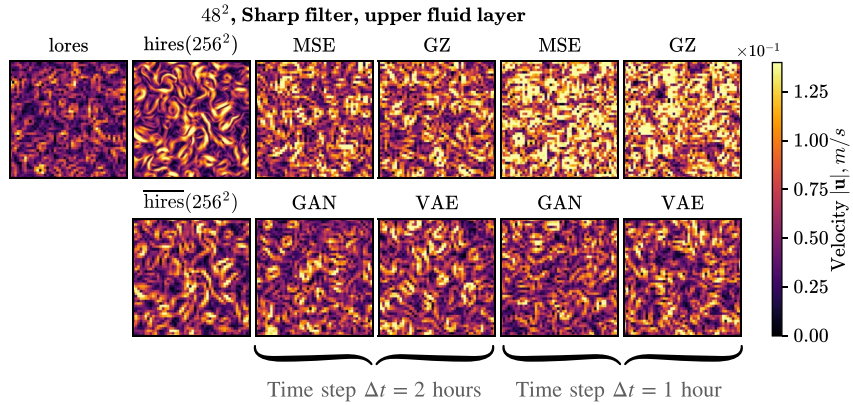

Figure 11. Snapshots of the modulus of velocity in coarse-resolution models (48^2) and high-resolution simulation (hires(256^2)). Two columns with $\Delta t = 2$ hr correspond to Figure 10. The smaller the time step, the smaller the effective eddy viscosity, see Text S4 in Supporting Information S1.

GAN and VAE models at both time steps produce physical solutions which look similar to the filtered and coarse-grained high-resolution simulation $\left(\overline{\text{hires}}\right)$. In Figure 12a 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. In Text S4 of the Supporting Information S1, we further explain that the time step is connected to the effective eddy viscosity in our particular numerical scheme, and thus computations at a small time step reveal numerical stability properties.

In Figure 12b we show the distributional metric as a function of resolution. At the coarsest resolution 48^2, 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 (64^2 and 96^2) 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. Overall, generative 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. Analysis of subgrid models with respect to another metric (error in representation of spectral properties) yields similar conclusions, see Text S3 and Figure S4 in Supporting Information S1.

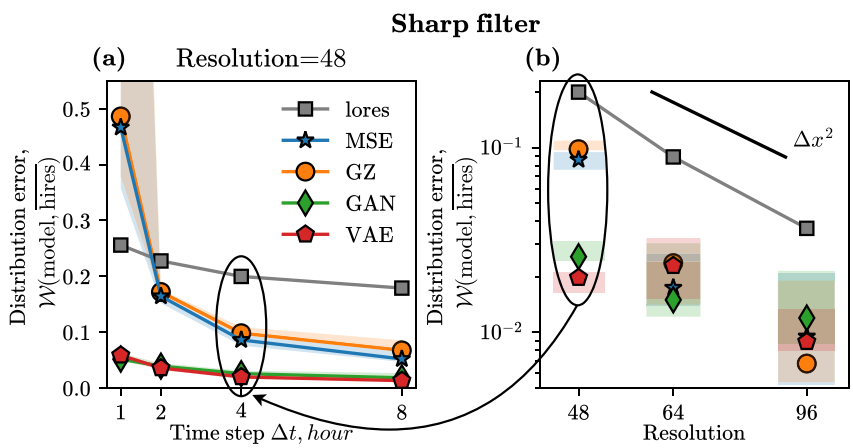

Figure 12. Online distributional metric (Equation 17): (a) as a function of time step at the coarsest spatial resolution and (b) as a function of spatial resolution. The metric is computed with respect to the high-resolution simulation (256^2). The shading area shows min-max values among training realizations, and markers show median value.

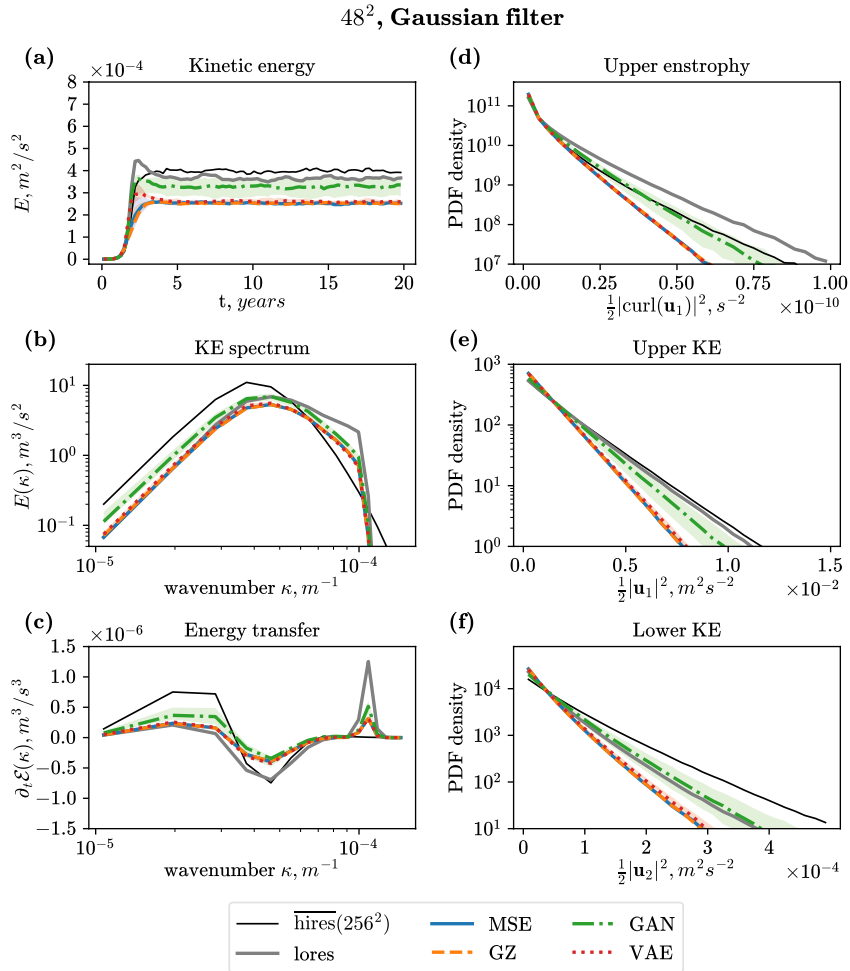

Figure 13. Online simulations with parameterized models. Similar to Figure 10, but for models trained on the data set with Gaussian filter.

The numerical stability issues occurring in MSE and GZ models can be explained by the overestimation of energy backscatter in online simulations (positive energy transfer on large scales in Figure 10c). This is in contrast to the observation that all subgrid models are equally accurate in predicting energy transfer offline (Figure 5c). We expect that GZ and MSE models first fail to reproduce the dynamics of small eddies and it eventually leads to transformation of the flow and inaccurate prediction of the backscatter. The discrepancy between the offline and online analysis may be due to the inclusion of the ssd term, time sampling method of the stochastic parameterization, and time integration scheme. The effect of ssd term on the energy transfer is seen in Figure 10c near the grid scale ($\kappa \approx 10^{-4}\,\mathrm{m}^{-1}$). The spurious positive energy transfer balances the dissipation produced by ssd term on these scales (we already explained it in Section 2.1.1). Removing of ssd term is possible but requires careful treatment of the numerical scheme because simulations become fully inviscid. In Text S5 and Figure S6 of the Supporting Information S1 we show that proposed subgrid models are stable in fully inviscid simulations and VAE model is the most accurate in reproducing energy transfer and other statistical characteristics.

In Figure 13 we show the online results for the subgrid models trained on the data set produced using the Gaussian filter. The subgrid models cannot substantially 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 Figure B1 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 al. (2023) and these may be caused by the mentioned discrepancies between the offline and online analysis.

In Appendix B 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 (Figure B2). It happens because of the overestimation of the positive energy transfer (backscatter) on the large scales which is present even in offline analysis as a result of generalization error (not shown). Runtime for the generative models is the same as for the deterministic baseline (Table B1) because we use the same CNN as a building block in all subgrid models.

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 (Agarwal et al., 2021; Berner, 2005; Bhouri & Gentine, 2022; Chorin & Lu, 2015; DelSole, 2000; Gagne et al., 2020).

The proposed subgrid parameterizations predict the divergence of subgrid PV flux and thus do not conserve PV. Restoring the conservation properties can be done in multiple ways. First, the predicted target can be changed to PV subgrid flux (or momentum subgrid flux) similarly to Ross et al. (2023). Second, the divergence operator can be implemented as a final convolutional layer in GAN and VAE models similarly to Zanna and Bolton (2020) and Srinivasan et al. (2023). Third, a promising approach would be to propose an ML model predicting the free parameters in physical parameterizations (Sane et al., 2023; Zhu et al., 2022) and thus automatically satisfy conservation properties. However, we note that a suitable form of parameterization of mesoscale eddy fluxes remains to be established. The existing approaches may lead to numerical instabilities due to the presence of

negative eddy viscosity or diffusivity. For example, the tracer diffusivity tensor produced by mesoscale eddies contains one negative eigenvalue (Bachman et al., 2020; Haigh & Berloff, 2021, 2022; Haigh et al., 2021; Kamenkovich et al., 2021; Lu et al., 2022; Ryzhov & Berloff, 2022). Parameterization of mesoscale momentum fluxes by the eddy viscosity operator is also challenging: the crucial physical process of kinetic energy backscatter can be captured only with negative eddy viscosity (Bachman, 2019; Jansen & Held, 2014; Jansen et al., 2019; Juricke et al., 2020).

As we noted in Section 4.2, offline generalization results can be partially explained by the difference in the magnitude of subgrid forcing on eddy and jet data sets. Thus, all the proposed subgrid models can benefit from the normalization of input and output features according to physical scalings (Beucler et al., 2021). For example, the input features can be normalized by their RMS values (different for every snapshot), while the RMS value for the subgrid flux is unknown and can be estimated given the prediction of velocity gradient model (Xie et al., 2020). Composing the training data set representing multiple dynamical regimes also should help to improve the generalization ability (O'Gorman & Dwyer, 2018).

Appendix A: Training of the Machine Learning Models

A1. 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 $\widetilde{S}_1 = G\left(z_1, \overline{q}\right)$ and $\widetilde{S}_2 = G\left(z_2, \overline{q}\right)$ for a given $\overline{q}$ 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:

$$\mathbf{S}_1 = \left(\widetilde{S}_1, S\right), \quad \mathbf{S}_2 = \left(S, \widetilde{S}_2\right), \quad \widetilde{\mathbf{S}} = \left(\widetilde{S}_1, \widetilde{S}_2\right),$$

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

$$\mathcal{L}_{\mathrm{W}} = \left[\frac{1}{2}\left(D\left(\mathbf{S}_1, \overline{q}\right) + D\left(\mathbf{S}_2, \overline{q}\right)\right) - D\left(\widetilde{\mathbf{S}}, \overline{q}\right)\right].$$

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):

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

where $\widehat{\mathbf{S}} = \epsilon\mathbf{S} + (1-\epsilon)\widetilde{\mathbf{S}}$. 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 $D\left(\widehat{\mathbf{S}}, \overline{q}\right) \in \mathbb{R}$, $\nabla_{\widehat{\mathbf{S}}} D\left(\widehat{\mathbf{S}}, \overline{q}\right) \in \mathbb{R}^{4\times n\times n}$ and norm $\|\cdot\|_2$ for tensor is defined above. Regularization preventing drift of discriminator:

$$\mathcal{L}_{\mathrm{drift}} = \left[D\left(\mathbf{S}_1, \overline{q}\right)\right]^2.$$

We minimize the following loss for the discriminator:

$$\mathcal{L}_{\mathrm{D}} = -\mathcal{L}_{\mathrm{W}} + 10\mathcal{L}_{\mathrm{grad}} + 10^{-3}\mathcal{L}_{\mathrm{drift}}, \tag{A1}$$

and the loss to be minimized for the generator is:

$$\mathcal{L}_{\mathrm{G}} = -D\left(\widetilde{\mathbf{S}}, \overline{q}\right). \tag{A2}$$

In the original paper $\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{W}}$ (Adler & Öktem, 2018), but we follow a typical approach when only generated samples constitute the generator loss (Dong & Yang, 2019). Following Arjovsky et al. (2017), we optimize the discriminator loss (Equation A1) for five batches in a row, and then we optimize the generator loss (Equation A2) for one batch.

A2. VAE Loss Function

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

The distributions for encoder, decoder and prior, respectively:

$$q_\phi(z|S,\overline{q}) = \mathcal{N}\left(\mu_\phi(S,\overline{q}), \mathrm{diag}\left(\sigma_\phi^2(S,\overline{q})\right)\right) \tag{A3}$$

$$\rho_\theta(S|z,\overline{q}) = \mathcal{N}\left(\mu_\theta(z,\overline{q}), \gamma I\right) \tag{A4}$$

$$\rho(z) = \mathcal{N}(0, I), \tag{A5}$$

where $I \in \mathcal{R}^{2n^2\times 2n^2}$ is identity matrix, γ is free parameter and $S, \overline{q}, z \in \mathbb{R}^{2\times n\times n}$. The encoder is parameterized by a single CNN predicting two features: mean $\mu_\phi \in \mathbb{R}^{2\times n\times n}$ and log-variance $\ln\left(\sigma_\phi^2\right) \in \mathbb{R}^{2\times n\times n}$ of the latent variable. The decoder is parameterized by CNN predicting the mean of subgrid forcing $\mu_\theta \in \mathbb{R}^{2\times n\times n}$.

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

$$\mathcal{L}_{\mathrm{VAE}} = \frac{1}{2\gamma}\|S - \mu_\theta(\hat{z},\overline{q})\|_2^2 + \frac{1}{2}\sum_{m,i,j}\left(\sigma_\phi^2 + \mu_\phi^2 - 1 - \ln(\sigma_\phi^2)\right)_{m,i,j}, \tag{A6}$$

where we replaced the mathematical expectation over the encoder distribution with a single sample from this distribution (reparameterization trick; Kingma & Welling, 2013), that is, $\hat{z} = \mu_\phi + \epsilon\sigma_\phi$, $\epsilon \sim \mathcal{N}(0, I)$. 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.

A3. Neural Networks and Optimization Algorithm

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) and Ross et al. (2023) with parameters given in Table A1. Discriminator D in GAN 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).

Table A1
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}
Kernel size	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

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 data set. 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.

Appendix B: Additional Online Results

See Appendix Figure B1.

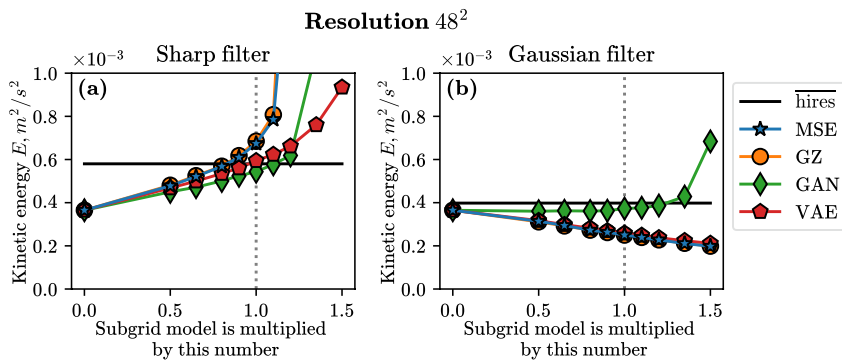

Figure B1. We multiply the subgrid model by a parameter $\alpha \in [0, 1.5]$ as $\widetilde{S} \to \alpha \widetilde{S}$ and show the kinetic energy after spin-up. Subgrid models which efficiently simulate backscatter are able to energize the flow when the amplitude is increased, see supplemental Figure 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 hr.

See Appendix Figure B2.

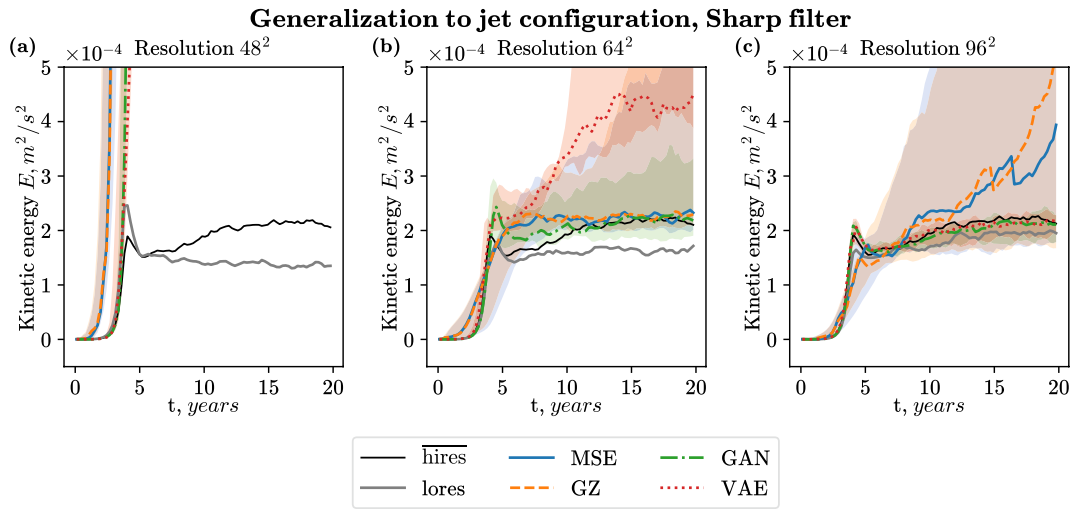

Figure B2. Online generalization to turbulence configuration with jets (Table 1). Generative models (GAN and VAE) clearly better reproduce transitional flow ($t < 2$ years), however many of the presented models have problems with numerical stability at a larger time. Improving the generalization capabilities 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 hr.

See Appendix Table B1.

Table B1
Runtime on One CPU Core for Unparameterized Model ("–") and ML-Based Parameterizations to Integrate QG Model in Time for 20 Years

Δt	1 hr	2 hr	4 hr				
$n \times n$	256 × 256	96 × 96	48 × 48				
Model	–	–	–	MSE	GZ	GAN	VAE
Runtime, sec	1,300	130	25.4	756	1,480	784	782

Note. 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).

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 NSF PHY-1748958. The authors would like to thank the members of M²LInES for their helpful comments and discussions. We also thank the editor, Stephen Griffies, as well as three anonymous reviewers for their constructive comments that helped to improve the quality and presentation of this paper.

Data Availability Statement

The Python software, including the subgrid ML models and plotting scripts, is available at Perezhogin and Zanna (2023). The training and simulation data are available at Perezhogin (2023).

References

Abernathey, R., Rocha, C. B., Ross, A., Jansen, M., Li, Z., Poulin, F. J., et al. (2022). pyqg/pyqg: v0.7.2. *Zenodo*. https://doi.org/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. https://doi.org/10.1029/2021ms002537

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. https://doi.org/10.1007/s00162-021-00581-z

Alvelius, K. (1999). Random forcing of three-dimensional homogeneous turbulence. *Physics of Fluids*, *11*(7), 1880–1889. https://doi.org/10.1063/1.870050

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. https://doi.org/10.1175/mwr-d-15-0245.1

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. https://doi.org/10.1063/1.1582183

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *arxiv 2017. arXiv preprint arXiv:1701*, *30*(4), 07875.

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

Bachman, S. D. (2019). The GM+ E closure: A framework for coupling backscatter with the gent and McWilliams parameterization. *Ocean Modelling*, *136*, 85–106. https://doi.org/10.1016/j.ocemod.2019.02.006

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. https://doi.org/10.1016/j.ocemod.2018.04.007

Bachman, S. D., Fox-Kemper, B., & Bryan, F. O. (2020). A diagnosis of anisotropic eddy diffusion from a high-resolution global ocean model. *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001904. https://doi.org/10.1029/2019ms001904

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. https://doi.org/10.1002/2016jc012265

Beck, A., Flad, D., & Munz, C.-D. (2019). Deep neural networks for data-driven les closure models. *Journal of Computational Physics*, *398*, 108910. https://doi.org/10.1016/j.jcp.2019.108910

Berloff, P. (2018). Dynamically consistent parameterization of mesoscale eddies. Part III: Deterministic approach. *Ocean Modelling*, *127*, 1–15. https://doi.org/10.1016/j.ocemod.2018.04.009

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. https://doi.org/10.1175/jas3468.1

Berner, J., Achatz, U., Batte, L., Bengtsson, L., De La Camara, A., Christensen, H. M., et al. (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588. https://doi.org/10.1175/bams-d-15-00268.1

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. https://doi.org/10.1175/jcli-d-11-00297.1

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. https://doi.org/10.1175/2008jas2677.1

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.

Bhouri, 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. https://doi.org/10.1029/2018MS001472

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

Canuto, C., Hussaini, M. Y., Quarteroni, A., & Zhang, T. A. (1988). *Spectral methods in fluid dynamics*. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-84108-8

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. https://doi.org/10.1063/1.857878

Chorin, A. J., & Lu, F. (2015). Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(32), 9804–9809. https://doi.org/10.1073/pnas.1512080112

Chow, F. K., & Moin, P. (2003). A further study of numerical errors in large-eddy simulations. *Journal of Computational Physics*, *184*(2), 366–380. https://doi.org/10.1016/s0021-9991(02)00020-7

Christensen, H., Berner, J., Coleman, D. R., & Palmer, T. (2017). Stochastic parameterization and El Niño–southern oscillation. *Journal of Climate*, *30*(1), 17–38. https://doi.org/10.1175/jcli-d-16-0122.1

Christensen, H., & Zanna, L. (2022). Parametrization in weather and climate models. In *Oxford research encyclopedia of climate science*. https://doi.org/10.1093/acrefore/9780190228620.013.826

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. https://doi.org/10.1175/1520-0469(2000)057<2158:aflomm>2.0.co;2

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*.

Fox-Kemper, B., & Menemenlis, D. (2008). Can large eddy simulation techniques improve mesoscale rich ocean models? *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. https://doi.org/10.1175/1520-0469(1997)054<2475:evasbp>2.0.co;2

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. https://doi.org/10.1034/j.1600-0870.2003.201329.x

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. https://doi.org/10.1103/physrevfluids.6.024607

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*, *14*(11). https://doi.org/10.1029/2022ms003124

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

Gent, P. R., & Mcwilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, *20*(1), 150–155. https://doi.org/10.1175/1520-0485(1990)020<0150:imiocm>2.0.co;2

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. https://doi.org/10.1002/qj.58

Ghosal, S. (1996). An analysis of numerical errors in large-eddy simulations of turbulence. *Journal of Computational Physics*, *125*(1), 187–206. https://doi.org/10.1006/jcph.1996.0088

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (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. https://doi.org/10.1016/j.ocemod.2013.01.004

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. https://doi.org/10.1175/1520-0493(2000)128<2935:bfwasl>2.0.co;2

Grooms, I., Lee, Y., & Majda, A. J. (2015). Numerical schemes for stochastic backscatter in the inverse cascade of quasigeostrophic turbulence. *Multiscale Modeling and Simulation*, *13*(3), 1001–1021. https://doi.org/10.1137/140990048

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. https://doi.org/10.1016/j.jcp.2022.111090

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*, *443*, 133568. https://doi.org/10.1016/j.physd.2022.133568

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021ms002534

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. https://doi.org/10.1017/s0022112003006268

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., et al. (2016). High resolution model intercomparison project (highresmip v1. 0) for CMIP6. *Geoscientific Model Development*, *9*(11), 4185–4208. https://doi.org/10.5194/gmd-9-4185-2016

Haigh, M., & Berloff, P. (2021). On co-existing diffusive and anti-diffusive tracer transport by oceanic mesoscale eddies. *Ocean Modelling*, *168*, 101909. https://doi.org/10.1016/j.ocemod.2021.101909

Haigh, M., & Berloff, P. (2022). On the stability of tracer simulations with opposite-signed diffusivities. *Journal of Fluid Mechanics*, *937*, R3. https://doi.org/10.1017/jfm.2022.126

Haigh, M., Sun, L., McWilliams, J. C., & Berloff, P. (2021). On eddy transport in the ocean. Part I: The diffusion tensor. *Ocean Modelling*, *164*, 101831. https://doi.org/10.1016/j.ocemod.2021.101831

Hallberg, R. (2013). Using a resolution function to regulate parameterizations of oceanic mesoscale eddy effects. *Ocean Modelling*, *72*, 92–103. https://doi.org/10.1016/j.ocemod.2013.08.007

Hewitt, H. T., Roberts, M., Mathiot, P., Biastoch, A., Blockley, E., Chassignet, E. P., et al. (2020). Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, *6*(4), 137–152. https://doi.org/10.1007/s40641-020-00164-w

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. https://doi.org/10.1029/2019ms001750

Jansen, M. F., & Held, I. M. (2014). Parameterizing subgrid-scale eddy effects using energetically consistent backscatter. *Ocean Modelling*, *80*, 36–48. https://doi.org/10.1016/j.ocemod.2014.06.002

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. https://doi.org/10.1029/2019ms001855

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. https://doi.org/10.1175/jcli-d-16-0539.1

Kamenkovich, I., Berloff, P., Haigh, M., Sun, L., & Lu, Y. (2021). Complexity of mesoscale eddy diffusivity in the ocean. *Geophysical Research Letters*, *48*(5), e2020GL091719. https://doi.org/10.1029/2020gl091719

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 of the United States of America*, *118*(21), e2101784118. https://doi.org/10.1073/pnas.2101784118

Kraichnan, R. H. (1976). Eddy viscosity in two and three dimensions. *Journal of the Atmospheric Sciences*, *33*(8), 1521–1536. https://doi.org/10.1175/1520-0469(1976)033<1521:evitat>2.0.co;2

LaCasce, J. H. (1996). Baroclinic vortices over a sloping bottom. Doctoral dissertation. Massachusetts Institute of Technology and Woods Hole Oceanographic Institution. https://doi.org/10.1575/1912/2457

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. https://doi.org/10.1017/s0022112079000045

Lu, Y., Kamenkovich, I., & Berloff, P. (2022). Properties of the lateral mesoscale eddy-induced transport in a high-resolution ocean model: Beyond the flux–gradient relation. *Journal of Physical Oceanography*, *52*(12), 3273–3295. https://doi.org/10.1175/jpo-d-22-0108.1

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.

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

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. https://doi.org/10.1017/jfm.2017.637

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. https://doi.org/10.1017/jfm.2018.770

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 auto-encoders. 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*(1), 255–286. https://doi.org/10.1146/annurev-fluid-060420-023735

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. https://doi.org/10.1103/physreve.75.046303

Nadiga, B. T., Sun, X., & Nash, C. (2022). Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, *1*, e22. https://doi.org/10.1017/eds.2022.32

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

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).

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), 71–116. https://doi.org/10.1088/0034-4885/63/2/201

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (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. https://doi.org/10.1007/s00162-019-00512-z

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. https://doi.org/10.1016/j.ocemod.2017.05.007

Perezhogin, P. (2023). 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*. https://doi.org/10.5281/zenodo.7622683

Perezhogin, P., & Zanna, L. (2023). 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*. https://doi.org/10.5281/zenodo.8226121

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., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United Sttaes of America*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Redi, M. H. (1982). Oceanic isopycnal mixing by coordinate rotation. *Journal of Physical Oceanography*, *12*(10), 1154–1158. https://doi.org/10.1175/1520-0485(1982)012<1154:oimbcr>2.0.co;2

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. https://doi.org/10.1029/2022ms003258

Rybkin, O., Daniilidis, K., & Levine, S. (2021). Simple and effective vae training with calibrated decoders. In *International conference on machine learning* (pp. 9179–9189).

Ryzhov, E., & Berloff, P. (2022). On transport tensor of dynamically unresolved oceanic mesoscale eddies. *Journal of Fluid Mechanics*, *939*, A7. https://doi.org/10.1017/jfm.2022.169

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. https://doi.org/10.1080/03091928008241178

Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical mixing coefficients in the ocean surface boundary layer using neural networks. *arXiv preprint arXiv:2306.09045*.

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. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences of the United Sttaes of America*, *120*(20), e2216158120. https://doi.org/10.1073/pnas.2216158120

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. https://doi.org/10.1175/jcli3954.1

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, *28*.

Srinivasan, K., Chekroun, M. D., & McWilliams, J. C. (2023). Turbulence closure with small, local neural networks: Forced two-dimensional and β-plane flows. *arXiv preprint arXiv:2304.05029*.

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. https://doi.org/10.1002/qj.3990

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. https://doi.org/10.1175/bams-d-17-0040.1

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. https://doi.org/10.1175/jas3408.1

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. https://doi.org/10.1016/j.neucom.2022.08.067

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. https://doi.org/10.1002/qj.2166

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. https://doi.org/10.1029/2022ms002984

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. https://doi.org/10.1256/qj.04.03

Xie, C., Wang, J., Li, H., Wan, M., & Chen, S. (2020). Spatially multi-scale artificial neural network model for large eddy simulation of compressible isotropic turbulence. *AIP Advances*, *10*(1). https://doi.org/10.1063/1.5138681

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. https://doi.org/10.1038/s41467-020-17142-3

Zacharuk, M., Dolaptchiev, S. I., Achatz, U., & Timofeyev, I. (2018). Stochastic subgrid-scale parametrization for one-dimensional shallow-water dynamics using stochastic mode reduction. *Quarterly Journal of the Royal Meteorological Society*, *144*(715), 1975–1990. https://doi.org/10.1002/qj.3396

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376. https://doi.org/10.1029/2020gl088376

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. https://doi.org/10.1016/j.ocemod.2017.01.004

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

Zhu, Y., Zhang, R.-H., Moum, J. N., Wang, F., Li, X., & Li, D. (2022). Physics-informed deep-learning parameterization of ocean vertical mixing improves climate simulations. *National Science Review*, *9*(8), nwac044. https://doi.org/10.1093/nsr/nwac044

References From the Supporting Information

Lund, T. (2003). The use of explicit filters in large eddy simulation. *Computers & Mathematics with Applications*, *46*(4), 603–616. https://doi.org/10.1016/s0898-1221(03)90019-8

Orszag, S. A. (1971). On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of the Atmospheric Sciences*, *28*(6), 1074. https://doi.org/10.1175/1520-0469(1971)028<1074:oteoai>2.0.co;2