



## RESEARCH ARTICLE

10.1029/2023MS003915

# On the Choice of Training Data for Machine Learning of Geostrophic Mesoscale Turbulence

F. E. Yan<sup>1</sup> , J. Mak<sup>1,2,3</sup> , and Y. Wang<sup>1,2</sup> 

<sup>1</sup>Department of Ocean Science, Hong Kong University of Science and Technology, Hong Kong, Hong Kong, <sup>2</sup>Center for Ocean Research in Hong Kong and Macau, Hong Kong University of Science and Technology, Hong Kong, Hong Kong, <sup>3</sup>National Oceanography Centre, Southampton, UK

**Key Points:**

- Investigated the dependence of convolution neural networks on the choice of training data for geostrophic turbulence
- Models are trained on eddy fluxes with rotational component filtered out by means of an eddy force function
- Resulting models as accurate but less sensitive to small-scale features than models trained on divergence of eddy fluxes

**Correspondence to:**

F. E. Yan and J. Mak,  
feyan@connect.ust.hk;  
julian.c.l.mak@googlemail.com

**Citation:**

Yan, F. E., Mak, J., & Wang, Y. (2024). On the choice of training data for machine learning of geostrophic mesoscale turbulence. *Journal of Advances in Modeling Earth Systems*, 16, e2023MS003915. <https://doi.org/10.1029/2023MS003915>

Received 2 JULY 2023

Accepted 20 DEC 2023

**Author Contributions:**

**Conceptualization:** J. Mak  
**Data curation:** F. E. Yan, J. Mak  
**Formal analysis:** F. E. Yan, J. Mak  
**Funding acquisition:** J. Mak  
**Investigation:** F. E. Yan, J. Mak, Y. Wang  
**Methodology:** J. Mak, Y. Wang  
**Project administration:** J. Mak  
**Resources:** J. Mak  
**Software:** F. E. Yan, J. Mak  
**Supervision:** J. Mak  
**Validation:** F. E. Yan, J. Mak  
**Visualization:** F. E. Yan  
**Writing – original draft:** F. E. Yan, J. Mak, Y. Wang  
**Writing – review & editing:** F. E. Yan, J. Mak, Y. Wang

**Abstract** Data plays a central role in data-driven methods, but is not often the subject of focus in investigations of machine learning algorithms as applied to Earth System Modeling related problems. Here we consider the problem of eddy-mean interaction in rotating stratified turbulence in the presence of lateral boundaries, where it is known that rotational components of the eddy flux plays no direct role in the sub-grid forcing onto the mean state variables, and its presence is expected to affect the performance of the trained machine learning models. While an often utilized choice in the literature is to train a model from the divergence of the eddy fluxes, here we provide theoretical arguments and numerical evidence that learning from the eddy fluxes with the rotational component appropriately filtered out, achieved in this work by means of an object called the eddy force function, results in models with comparable or better skill, but substantially reduced sensitivity to the presence of small-scale features. We argue that while the choice of data choice and/or quality may not be critical if we simply want a model to have predictive skill, it is highly desirable and perhaps even necessary if we want to leverage data-driven methods to aid in discovering unknown or hidden physical processes within the data itself.

**Plain Language Summary** Data-driven methods are increasingly being utilized in various problems relating to the numerical modeling of the Earth system. While there are many investigations focusing on the machine learning algorithms or the problems themselves, there have been relative few investigations into the impact of data choice or quality, given the central role of data. We consider here the impact of the choice of data for a particular problem relevant to ocean modeling, that of eddy-mean interaction, where it is known that the training data generically contains a component that plays no role in the eddy-mean interaction, and its presence in the training phase is expected to degrade the model performance. We provide arguments and evidence that one choice is preferable over a more standard choice utilized in related research. While the choice of data choice and/or quality may not be critical if we simply want a data-driven model to be skillful, we argue it is highly desirable, possibly even a necessity, if we want to leverage data-driven methods as a means to aid in discovery of unknown or hidden physical processes within the data itself.

## 1. Introduction

Data-driven methods and machine learning algorithms are increasingly being utilized in problems relating to Earth system and/or climate modeling, and there is no doubt such methods have a strong potential in greatly enhancing model skill and/or reducing computation cost in Earth System Modeling. Some examples include modeling of dynamical processes in the atmosphere (e.g., Brenowitz & Bretherton, 2019; Connolly et al., 2023; Mooers et al., 2021; Sun et al., 2023; Yuval & O’Gorman, 2020), climate modeling (e.g., Besombes et al., 2021; Sonnewald & Lguensat, 2021), sea ice prediction (e.g., Andersson et al., 2021; Bolibar et al., 2020), identification problems in oceanography (e.g., Jones et al., 2019; Sonnewald et al., 2019, 2023; Thomas et al., 2021), and our primary focus here, ocean mesoscale turbulence parameterizations (e.g., Bolton & Zanna, 2019; Guillaumin & Zanna, 2021; Zanna & Bolton, 2021). We refer the reader to the works of Reichstein et al. (2019), Irrgang et al. (2021), Sonnewald et al. (2021), and Camps-Valls et al. (2023) for a more comprehensive review.

One criticism of some data-driven methods and machine learning algorithms is the “black-box” nature of the resulting models. In general, for a problem with input  $x$  and target  $y$ , a focus of data-driven methods is to find some mapping  $f$  such that  $f(x) = y$ , where  $f$  could be deterministic or probabilistic depending on the deployed algorithm. It is often not clear how or why the resulting  $f$  was returned by the algorithm, or what  $f$  is in fact doing in terms of

© 2024 The Authors. Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

known transformations or operations to map  $x$  into  $y$ . This lack of interpretability for  $f$  brings into question several important issues with the use of data-driven methods. The first is whether the models are skillful in the predictions for the “right” reasons, or at least not the “wrong” ones? For example, if  $x$  and  $y$  are related by some known physical process, does  $f$  in fact represent that physical process, or failing that, is the resulting  $f$  at least physically valid, such as not violating conservation laws? Without appropriate constraints on the model, it is perfectly plausible that resulting models can behave erratically, and given the nonlinear and convoluted nature of the model itself, the models can generate subtly wrong results that might be close to impossible to check. The second relates to further utilities of the methods themselves: is it possible to use such methods to aid process discovery from the data itself? A lack of interpretability would suggest a negative answer to that question. With that in mind, there has been an increasing focus on physically constrained and/or interpretable/explainable models (e.g., Barnes et al., 2022; Beucler et al., 2021; Brenowitz et al., 2020; Clare et al., 2022; Guan et al., 2023; Kashinath et al., 2021; Lopez-Gomez et al., 2022; Sonnewald & Lguensat, 2021; Yuval et al., 2021; Zanna & Bolton, 2021; Zhang & Lin, 2018). While the tools and algorithms do exist, this is a fundamentally harder problem, since the training step ultimately becomes one of constrained optimization.

While the algorithms and nature of the resulting model  $f$  (e.g., linear vs. nonlinear, generative vs. discriminative, model complexity) are important details, at the very base level we are really dealing with the problem of *data regression*. We would thus expect *data choice* and/or *data quality* to critically affect the training, the performance or the useful information that could be encoded by the model, but are issues that have not received much investigation. If we simply want models that have skill in whatever metric we think is relevant (e.g., low mismatch values in the predictions compared to a chosen reference), then the issue of data quality and/or content may not be critical, since we are simply looking for some optimal fit. If, on the other hand, we are interested in the harder problem of optimal fit with constraints, such as having a model that is constrained by physical conservation laws, or using data-driven methods for process discovery from data, then one might expect the choice and quality of data exposed to the model to be important. Furthermore, certain data may be more accessible for the machine learning algorithms to extract/predict features from, which has practical consequences for the optimization procedure at the model training and prediction step.

To demonstrate that not all choices of data are equal, we consider in this work the problem of eddy-mean interaction in rotating stratified turbulence in the presence of boundaries, of relevance to ocean modeling and parameterization of geostrophic mesoscale eddies. We assert that the parameterization problem is affected by presence of what are known as the rotational component of eddy fluxes (e.g., Fox-Kemper et al., 2003; Maddison et al., 2015; J. C. Marshall & Shutts, 1981) in the training data. We provide some theoretical arguments and evidence on why learning from the eddy fluxes with the rotational component removed to various degrees is preferable to the divergence of the eddy fluxes, the latter having been considered in some existing works (Bolton & Zanna, 2019; Zanna & Bolton, 2021). We will largely leverage the experimental procedure of Bolton and Zanna (2019), albeit with important differences to be detailed. While the present investigation is largely empirical and relies on input of external knowledge that is somewhat specific to the problem considered, the current work serves to open a discussion into data choice and/or quality, as well as probing the available information content in data in the general case, possibly in a more systematic and objective fashion than one performed here.

The technical problem statement relating to rotational fluxes and its impact on data quality for data-driven methods is outlined in Section 2. In Section 3 we outline our experimental procedure, numerical model used and data-driven method. Section 4 summarizes the impact of data choice on the skill of the trained models, and additionally explores sensitivity to small-scale features in data, and to the amount of data exposed to model. We close in §5 and provide outlooks, focusing particularly on further experiments to probe the information content of data being for use in data-driven methods of relevance to the present eddy-mean interaction problem.

## 2. Rotational Fluxes and the Eddy Force Function

### 2.1. Formulation

We consider turbulent motion under the influence of strong rotation and stratification. Specifically, we consider the Quasi-Geostrophic (QG) limit (e.g., Vallis, 2006), which is a widely used and applicable limit for oceanic mesoscale dynamics where the motion is geostrophic at leading order. If we consider the standard Reynolds decomposition (e.g., L. Sun et al., 2021; Sun et al., 2023) with

$$A = \bar{A} + A', \quad \overline{A+B} = \bar{A} + \bar{B}, \quad \overline{A'} = 0, \quad (1)$$

where the overbar denotes a mean (with the projection operator assumed to commute with all relevant derivatives), and a prime denotes a deviation from the mean, the mean QG Potential Vorticity (PV) equation takes the form

$$\frac{\partial \bar{q}}{\partial t} + \nabla \cdot (\bar{\mathbf{u}} \bar{q}) = -\nabla \cdot \overline{\mathbf{u}' q'} + \bar{Q}. \quad (2)$$

Here,  $t$  denotes time,  $\nabla$  denotes the horizontal gradient operator, so that the PV  $q$  is defined as

$$q = \nabla^2 \psi + \beta y + \frac{\partial}{\partial z} \frac{f_0}{N_0^2} \frac{\partial b}{\partial z}, \quad (3)$$

where  $\psi$  is the streamfunction,  $f = f_0 + \beta y$  is the Coriolis frequency (background value and leading order meridional variation),  $N_0$  is the constant background buoyancy frequency related to the imposed background stratification,  $b = f_0 \partial \psi / \partial z$  is the buoyancy,  $\mathbf{u} = (-\partial \psi / \partial y, \partial \psi / \partial x)$  is the non-divergent geostrophic velocity, and  $Q$  encapsulates all forcing and dissipation.

Studies of eddy-mean interaction often seek to understand the inter-dependence of the nonlinear eddy flux terms on the right hand side of Equation 2 and the mean state variables. A particular goal with eddy parameterization is to relate the eddy flux term  $\overline{\mathbf{u}' q'}$  with some large-scale mean state, normally as

$$\overline{\mathbf{u}' q'} \sim f(\bar{q}, \dots; \kappa, \dots), \quad (4)$$

where  $f$  is some mapping between mean state variables (such as  $\bar{q}$ ) and associated parameters (such as  $\kappa$ ) to the eddy fluxes. Once such a relation exists, we take a divergence, from which we obtain the eddy forcing on the mean state variables. A notable example would be PV diffusion (e.g., Green, 1970; J. C. Marshall, 1981; Rhines & Young, 1982), where we directly postulate for the form of  $f$  as

$$\overline{\mathbf{u}' q'} = -\kappa \nabla \bar{q} \quad \Rightarrow \quad -\nabla \cdot \overline{\mathbf{u}' q'} = \nabla \cdot (\kappa \nabla \bar{q}). \quad (5)$$

We emphasize the ordering of the operations here: we obtain a functional relation between the mean and eddy fluxes first, then we take a divergence to obtain the eddy forcing (cf. Fickian diffusion closures).

## 2.2. The Issue of Rotational Fluxes

The form as given in Equation 4 suggests that data-driven approaches would be useful by either directly regressing/learning for an empirical mapping  $f$  or, when a prescribed mapping  $f$  such as Equation 5 is given, to learn for parameters such as  $\kappa$ . Note, however, that a two-dimensional vector field such as  $\overline{\mathbf{u}' q'}$  can be generically written as

$$\overline{\mathbf{u}' q'} = \nabla \tilde{\Psi} + \hat{\mathbf{e}}_z \times \nabla \tilde{\Phi} + \tilde{\mathbf{H}} \quad (6)$$

via a Helmholtz-type decomposition, where  $\hat{\mathbf{e}}_z$  is the unit vector pointing in the vertical,  $\tilde{\Psi}$  and  $\tilde{\Phi}$  are scalar potentials encoding a divergent (vanishing under a curl) and a rotational (vanishing under a divergence) component respectively, and  $\tilde{\mathbf{H}}$  is a vector potential encoding a harmonic component (vanishing under either a curl and divergence). Since the eddy forcing on the mean state in Equation 2 appears as a divergence, the rotational (and harmonic) eddy fluxes play no role in the eddy forcing, and questions arise as to whether the presence of such rotational fluxes is going to be detrimental to the regression/learning by data-driven methods. Similar issues arise for example, in a diagnostic problem for the PV diffusivity  $\kappa$ , where rotational fluxes are known to severely contaminate the calculation (e.g., Mak et al., 2016; L. Sun et al., 2021). More generally, the eddy forcing arises from a divergence of the Eliassen–Palm flux tensor, with various eddy fluxes as the tensor

components (e.g., Maddison & Marshall, 2013; Young, 2012), and this problem of gauge freedom is generic for problems relating to eddy-mean interaction.

One way around the issue of rotational fluxes would be to perform a Helmholtz decomposition as above, and perform learning/regression/diagnoses using only the divergent term  $\nabla\tilde{\Psi}$ . This approach is however complicated by the issue of gauge freedom in the presence of boundaries (e.g., Fox-Kemper et al., 2003; Maddison et al., 2015; Mak et al., 2016). Since there is generically no inherited natural boundary condition for arbitrary choices of vector fields (although there may be ones that are physically relevant depending on the problem), the divergent term  $\nabla\tilde{\Psi}$  is defined only up to an arbitrary rotational gauge.

Another possibility might be to utilize the divergence of the eddy flux directly (e.g.,  $\nabla\cdot\overline{\mathbf{u}'q'}$ ). This is somewhat the approach taken in the works of Bolton and Zanna (2019) and Zanna and Bolton (2021) for example, who consider applying data-driven methods to learn about sub-grid momentum forcing in an idealized ocean model. While they report positive results from data-driven methods in their work, there are some points that are worth revisiting, particularly regarding learning from the divergence of the eddy flux. One issue is the spatial resolution of data itself: the eddy flux data is characterized by significant small-scale variability, and now we want its divergence, which further amplifies the relative variance at smaller-scales. Questions arise whether such a choice is unnecessarily taxing on the machine learning algorithms, which is now trying to find a mapping between very small-scale data and large-scale mean-state data for the parameterization problem. Following on from this point is the issue of model sensitivity to small-scale fluctuations in the training data, which could arise from numerical model resolution or the choice of averaging window. If a model is sensitive to data variation, one might question its robustness, and whether the associated degree of uncertainty is acceptable for practical deployment of model. A final point is more subtle and more speculative, to do with *commutativity*, that is, ordering of operations. Eddy parameterizations are usually formulated as in Equation 4: we learn a  $f(\dots) = \overline{\mathbf{u}'q'}$ , from which we take a divergence of the learned  $f$  to get the eddy forcing. If we are learning from  $\nabla\cdot\overline{\mathbf{u}'q'}$ , then the ordering is different, because we are seeking a mapping  $\hat{f}$  such that  $\nabla\cdot\overline{\mathbf{u}'q'} = \hat{f}(\dots)$ , where we would hope that  $\hat{f} = \nabla\cdot f$ . There is however no reason to expect such an equality, since the resulting mappings  $f$  or  $\hat{f}$  obtained from machine learning algorithms are generically nonlinear.

If we are simply interested in a model that is skillful, then these aforementioned points may not actually matter. If, on the other hand, we are interested in learning about the underlying physics via data-driven methods, then it is not clear whether the aforementioned properties (or the lack thereof) become fundamental limitations in the applicability of the procedure.

### 2.3. The Eddy Force Function

If we consider learning from data at the eddy flux level, then we probably want to filter the rotational component in some way. Generically, a simple Helmholtz decomposition as in Equation 6 would be one possibility, subject to the caveat that we have a freedom to choose a boundary condition. However, within a simply connected QG system as is considered in this work, there is a choice of employing an object called the *eddy force function* (Maddison et al., 2015; D. P. Marshall & Pillar, 2011). The eddy force function  $\Psi_{\text{eff}}^q$  associated with the eddy PV flux (denoted by the super-script  $q$ ) is obtained by solving the Poisson equation

$$\nabla\cdot\overline{\mathbf{u}'q'} = -\nabla^2\Psi_{\text{eff}}^q \quad (7)$$

subject to homogeneous Dirichlet boundary conditions  $\Psi_{\text{eff}}^q = 0$ , where the boundary condition is inherited from the zero normal mean geostrophic flow condition (Maddison et al., 2015). The eddy force function is related to the Helmholtz decomposition in Equation 6 where  $\tilde{\Psi}$  is replaced by  $-\Psi_{\text{eff}}^q$ , with Equation 7 obtained from taking a divergence of Equation 6, and is thus one way of synthesizing the divergent component of the eddy flux. Within a simply connected QG system,  $\Psi_{\text{eff}}^q$  can be shown to be optimal in the sense that  $\nabla\Psi_{\text{eff}}^q$  is as small as possible in the  $L^2$  norm (see Equation 10 for the definition of the  $L^2$  norm, as well as Appendix A of Maddison et al., 2015).

Via the linearity assumption of the eddy force function and boundary condition inheritance (Maddison et al., 2015), we can define an eddy force function for the components that contribute toward the definition of eddy PV flux. For example, from the definition of PV given in Equation 3, we can define an eddy relative vorticity and a buoyancy force function as solutions to

$$\nabla \cdot \overline{\mathbf{u}'\zeta'} = -\nabla^2 \Psi_{\text{eff}}^{\zeta}, \quad \nabla \cdot \overline{\mathbf{u}'b'} = -\nabla^2 \Psi_{\text{eff}}^b, \quad (8)$$

subject also to homogeneous Dirichlet boundary conditions, where  $\zeta = \nabla^2 \psi$  is the relative vorticity. The eddy relative vorticity and eddy buoyancy fluxes are related to the Reynolds stress (via the Taylor identity, e.g., Maddison & Marshall, 2013) and form stress respectively.

Physically, the eddy force function is a quantity that encapsulates momentum tendencies associated with eddy forcings (Maddison et al., 2015; D. P. Marshall & Pillar, 2011). The eddy force functions have been previously demonstrated to be a useful quantity for diagnoses problems (e.g., Mak et al., 2016), and we might expect that it would be a useful quantity for data-driven methods applied to eddy parameterization of rotating stratified turbulence. The gradient of the eddy force function  $-\nabla \Psi_{\text{eff}}^q$  removes a portion of the rotational fluxes, suggesting that  $-\nabla \Psi_{\text{eff}}^q$  would serve as a better choice of data compared to training on the full eddy flux  $\overline{\mathbf{u}'q'}$ , which contains rotational components. Additionally, given parameterizations are more naturally formulated as a relation between the eddy fluxes and the mean state (cf. Equation 4), learning from  $-\nabla \Psi_{\text{eff}}^q$  avoids the possible issue with commutativity mentioned above.

Given the useful properties of the eddy force function, for the present work, we principally focus on the eddy force function, although we provide some sample results in calculations that employs a standard Helmholtz decomposition in Appendix A.

### 3. Model Details

For a problem  $y = f(x)$ , the focus here is principally on the skill of the models  $f$ , trained on various target data  $y$  for the same inputs  $x$ , where skill is to be measured by mismatches between  $y_{\text{data}}$  and  $y_{\text{predict}} = f(x_{\text{data}})$ . We detail here a set of experiments to test and explore the following hypotheses:

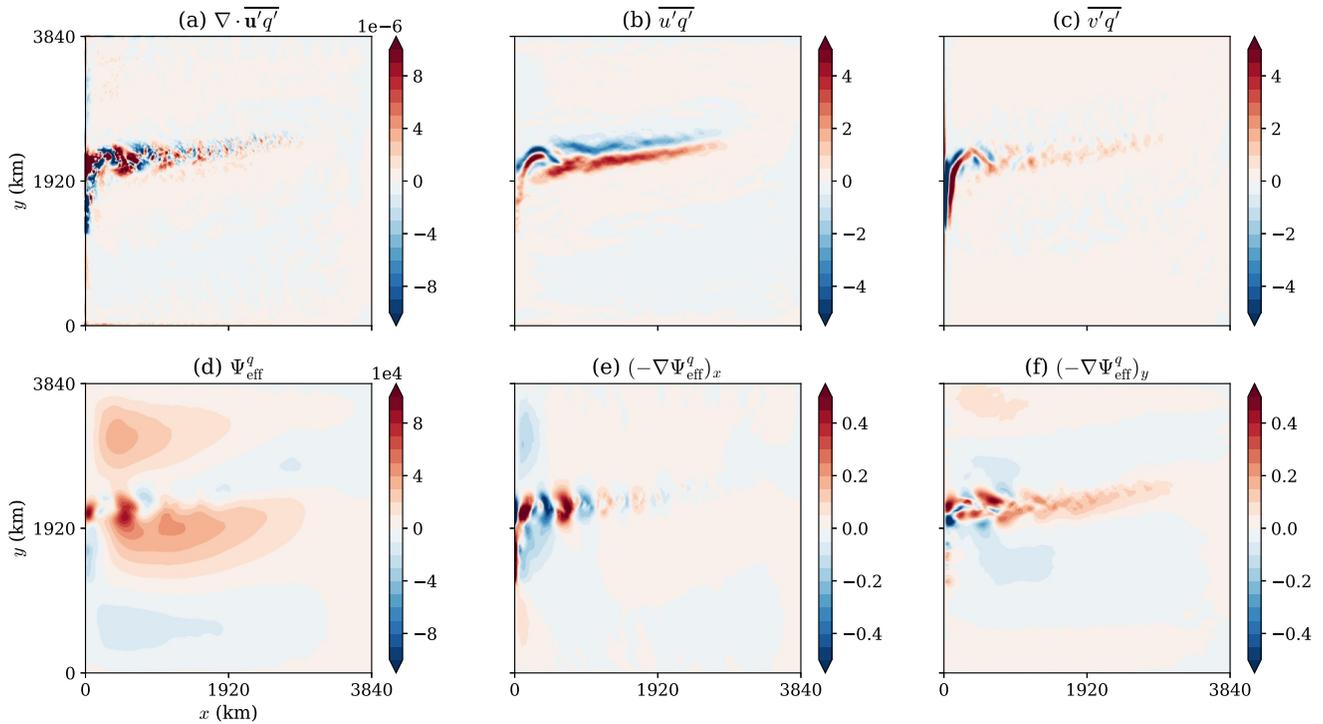
1. Models trained on the filtered eddy flux  $-\nabla \Psi_{\text{eff}}^q$  would be more skillful than ones trained on the full eddy flux  $\overline{\mathbf{u}'q'}$ ,
2. Models trained on the filtered eddy flux  $-\nabla \Psi_{\text{eff}}^q$  would possibly be comparable in skill to ones trained on the divergence of the eddy flux  $\nabla \cdot \overline{\mathbf{u}'q'}$ , but the latter models might be more sensitive to small-scale features in the training data.

The experimental approach will largely mirror that of Bolton and Zanna (2019). However, one important fundamental difference of our work is the choice of average, which impacts the definition of eddies from Equation 1. Where Bolton and Zanna (2019) take a low-pass spatial filter as the projection operator (with  $\overline{A'} \neq 0$ ), we employ a time-average, which has the property that  $\overline{A'} = 0$ , in line with properties of a Reynolds operator. Our eddy forcing is then in the more familiar form of a nonlinear eddy flux (e.g.,  $\nabla \cdot \overline{\mathbf{u}'q'}$ ), rather than as a difference between the spatially averaged quantities (e.g.,  $\mathbf{S} = \overline{\mathbf{u}} \cdot \nabla \overline{q} - \overline{\mathbf{u}} \cdot \nabla q$ , cf. Equation 8 of Bolton & Zanna, 2019). The existing definition of the eddy force function  $\Psi_{\text{eff}}^q$  assumes a Reynolds average (Maddison et al., 2015), and while there are likely extensions and relaxation of assumptions possible, we do not pursue this avenue.

#### 3.1. Numerical Ocean Model Setup

The physical setup we consider is essentially the same three-layer QG square double gyre configuration as Bolton and Zanna (2019) (cf. Berloff, 2005; Karabasov et al., 2009; D. P. Marshall et al., 2012; Mak et al., 2016), but solved with a pseudo-spectral method instead of the finite difference CABARET scheme of Karabasov et al. (2009). The numerical model (qgm2) generating the data presented in this work utilizes the parameters detailed in Mak et al. (2016), with the stratification parameters chosen such that the first and second Rossby deformation radii are 32.2 and 18.9 km, with a horizontal grid spacing of  $\Delta x = \Delta y = 7.5$  km (which is 512 by 512 in horizontal grid points), a horizontal viscosity value of  $\nu = 50 \text{ m}^2 \text{ s}^{-1}$ , and a time-step of  $\Delta t = 30$  min. A wind forcing with peak wind stress of  $\tau_0 = 0.8 \text{ N m}^{-2}$  is used (correcting a typo in Table 1 of Mak et al., 2016). The model is spun up from rest for 20,000 days, and a further integration period of 5,000 days after this spin up is performed for computing time-averages.

The accumulated time-averages of the eddy fluxes are used to compute the eddy force function  $\Psi_{\text{eff}}$  via solving the Poisson equation in Equation 7 with homogeneous Dirichlet boundary conditions per layer, although we only

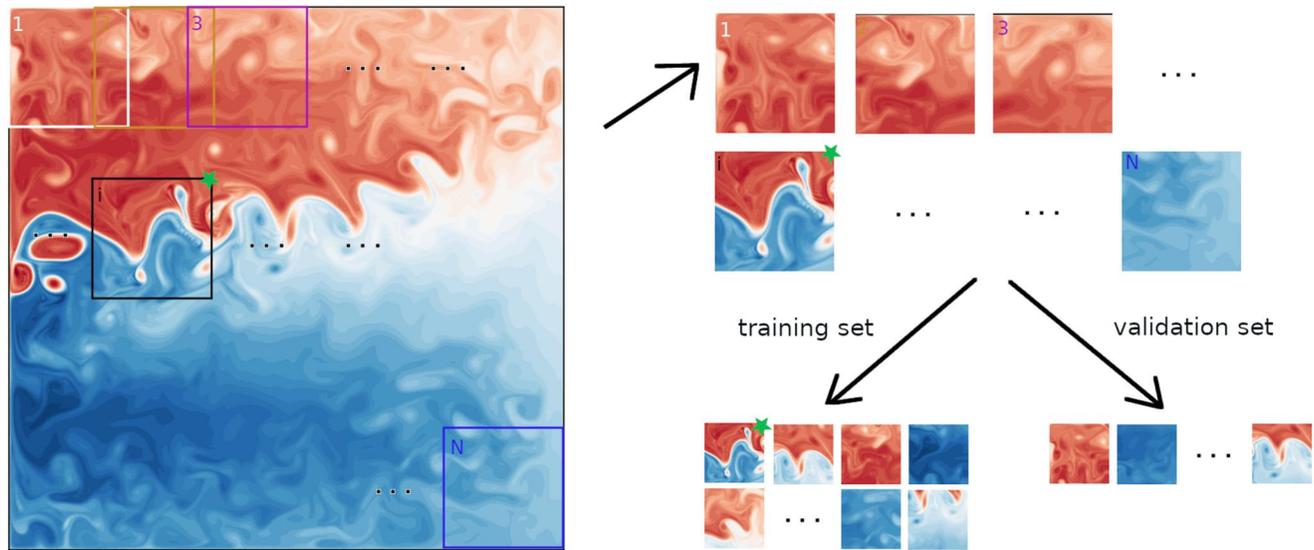


**Figure 1.** (a) The divergence of the eddy PV flux (units of  $s^{-2}$ ), calculated from the diagnosed time-averaged (b) zonal and (c) meridional component of the PV fluxes (units of  $m^2 s^{-2}$ ). (d) The associated eddy force function  $\Psi_{\text{eff}}^q$  (units of  $m^2 s^{-2}$ ) calculated from the data shown in panel (a), and the (e) zonal and (f) meridional component of  $-\nabla\Psi_{\text{eff}}^q$  (units of  $m s^{-2}$ ), the associated eddy PV fluxes with a rotational component removed. Note the different choices of colorbar limits between the data range in panels (b, c, e, f).

present the analysis for the top surface layer in this manuscript. We leverage the FEniCS software (Alnæs et al., 2014; Logg & Wells, 2010; Logg et al., 2012) following the previous works of Maddison et al. (2015) and Mak et al. (2016) to solve the Poisson equation, making use of the high level abstraction, automatic code generation capabilities and the numerous inbuilt solvers that are particularly suited to elliptic equations. The data from each grid point of the numerical model are the nodal values on a regular structured triangular mesh, with a projection onto a piecewise linear basis (CG1). All derivative operations are performed on the finite element mesh, and the nodal values of the relevant fields are restructured into arrays for feeding into the machine learning algorithms.

Figure 1 shows some sample data in the surface layer. The two horizontal components of the time-averaged eddy PV fluxes in panels (b, c) are the data sets returned by the pseudo-spectral model, which is sampled onto a finite element mesh as a vector object. The resulting object's divergence can then be computed in FEniCS, and the result is given in panel (a). As expected, the divergence of the eddy PV flux has more smaller-scale fluctuations and is less smooth than the eddy PV fluxes. Solving the relevant Poisson equation in FEniCS, the PV eddy force function  $\Psi_{\text{eff}}^q$  is shown in panel (d). From Maddison et al. (2015), the gradient of the eddy force function  $\nabla\Psi_{\text{eff}}^q$  has a physical interpretation when considered together with the time-mean streamfunction  $\bar{\psi}$  (not shown, but see Maddison et al., 2015), interpreted as whether eddies are accelerating the mean-flow (if  $\nabla\Psi_{\text{eff}}^q \cdot \nabla\bar{\psi} > 0$ , interpreted as an input of energy into the mean by eddies) or decelerating the mean flow (if  $\nabla\Psi_{\text{eff}}^q \cdot \nabla\bar{\psi} < 0$ , interpreted as an extraction of energy from the mean by eddies). Here, the eddy force function can be shown to correspond to the regimes where the eddies are slowing down the mean-flow via baroclinic instability when the Western Boundary Current first separates (the first positive-negative pattern emanating from the western boundary, which is anti-correlated with  $\nabla\bar{\psi}$ ), while the next dipole pattern (the first negative-positive patterns, which is correlated with  $\nabla\bar{\psi}$ ) is an eddy forcing of the mean-flow corresponding to an eddy driven regime (cf. Waterman & Hoskins, 2013; Waterman & Jayne, 2011).

From this  $\Psi_{\text{eff}}^q$ , the horizontal components of the gradient gives an eddy PV flux with a portion of the rotational component removed, and are shown in panels (e, f). While not obvious at first sight, the divergence of the full



**Figure 2.** Model training strategy demonstrated here with a snapshot of the instantaneous PV from model output. The domain is partitioned into small square regions of size 40 by 40 pixels, overlapping in the  $x$  and  $y$  direction by 6 pixels, resulting in 6,400 entries of input and output data. Each pair of input and output data is assigned with equal probability to be in the training set and validation set at the 80:20 ratio, from which a trained model results. An ensemble of models with 20 members is created, and are tested according to the procedure detailed in text.

eddy PV flux (panels b, c) and the divergence of the filtered eddy PV flux (panels e, f) are both equal to  $\nabla \cdot \overline{\mathbf{u}'q'}$  (panel a) up to numerical solver errors (here at least four orders of magnitude smaller than the data). In this instance, note also that the filtered eddy flux has qualitatively different spatial patterns to the full eddy flux, and that the filtered eddy flux is around an order of magnitude smaller than the full eddy fluxes. The behavior is consistent with observations that the rotational eddy fluxes can be large (e.g., Griesel et al., 2009), and suggests that the presence of rotational fluxes can be expected to have a significant impact on the model training.

### 3.2. Model Training Procedure

Following Bolton and Zanna (2019) we employ Convolutional Neural Networks (CNNs; e.g., Section 9, Goodfellow et al., 2016) to map between the inputs and targets. In line with the intended investigation, the choice of parameters for training the CNNs are kept fixed and chosen as in Bolton and Zanna (2019), and the main quantity we vary is the choice of training data. The mappings that are returned as a CNN are denoted:

- $f_{\text{div}}^q(\dots)$ , a mapping between mean state variables to be specified and the divergence of the eddy PV flux  $\nabla \cdot \overline{\mathbf{u}'q'}$ ,
- $f_{\text{full}}^q(\dots)$ , a mapping between mean state variables to be specified and the full eddy PV flux  $\overline{\mathbf{u}'q'}$ ,
- $f_{\text{eff}}^q(\dots)$ , a mapping between mean state variables to be specified and the gradient of the PV eddy force function  $-\nabla\Psi_{\text{eff}}^q$ .

Note that  $f_{\text{div}}^q(\dots)$  predicts a scalar field, while the  $f_{\text{full/eff}}^q(\dots)$  returns a vector field. A possible choice could be to train a model from the eddy force function, and from the trained model's predicted eddy force function compute its Laplacian to obtain the divergence of the eddy flux. As mentioned above, this is an extremely difficult test for model skill since gradient operations amplify mismatches, and we comment on related results and observations in the conclusions section.

To obtain these mappings in the present time-averaged case, we follow the schematic given in Figure 2, partially inspired by the approach of Bolton and Zanna (2019). The model domain is partitioned into small overlapping boxes. The input and target data associated with each of these boxes are paired up, and the pairs are each assigned an integer number and randomly shuffled (i.e., sampling from a uniform probability distribution function) depending on a choice of a random seed, and subsequently assigned to the training set (for training up the model) and validation (for tuning the hyperparameters in order to minimize a specified loss function) at the 80:20 ratio. In the 512 by 512 pixel domain, we take the small boxes to be 40 by 40 pixels, with a stride of six, resulting in a

collection of  $80^2 = 6,400$  images of the domain. For statistical significance, an ensemble of 20 such models were trained, each ensemble member only differing in the choice of the random seed, and the same sets of random seeds are used for all ensembles. The CNNs are built using the PyTorch platform (Paszke et al., 2019), where the CNN architecture consists of three hidden convolutional layers with square kernels (of size 8, 4 and 4 respectively), with a two-dimension max pooling layer with square kernel of size 2, and a fully connected linear activation layer as the output. The CNNs are trained with a batch size of 64, using the Adam optimizer (Kingma & Ba, 2015) with a mean squared error loss function. An early stopping criterion, where the training is stopped if the validation loss does not decrease by  $10^{-6}$  after 20 epochs, is used to monitor the loss function during the training to avoid overfitting. For simplification, we use a constant learning rate of  $10^{-4}$  during training. Models that predict scalars are typically trained for around 200 epochs, taking around three mins with a NVIDIA Tesla T4 GPU in Google Colab; the number of epochs and training time is doubled for models that predict vector quantities with two components.

In this work, the skill of the model is its ability to be able to predict the global field. We note that while it is customary to withhold a portion of data that the model has not been exposed to until the testing stage, because of our choice in partitioning the domain into overlapping boxes, in some sense the model will have “seen” the whole domain if the percentage of total data exceeds a certain threshold (around 30% of total data). For this work we make the simplest choice of exposing the model to *all* the data in a set of control calculations in Sections 4.1 and 4.2, but consider the dependence of our conclusions to *decreasing* percentage of data exposed to the model during the training phase in Section 4.4. While it is certainly true our choice leads to the criticism that we may be testing the model’s skill in overfitting in Section 4, our results in Section 4.4 provides some evidence that this is in fact not the case, and the conclusions we draw are robust.

#### 4. Model Skill

We first evaluate the predictive skill of the various models to the choice of target data. The skill of the models are judged by its ability to reduce mismatches of the divergence of the eddy PV flux, via repeated predictions of smaller patches over the whole domain (here taken with a stride of 2 pixels), with averages taken as necessary. Note that while  $f_{\text{div}}^q(\dots)$  already predicts the divergence of the eddy PV flux, we will take a divergence of the outcome of  $f_{\text{full/eff}}^q(\dots)$  to give the predicted divergence of the eddy PV flux. The normalized mismatch between data and prediction will be judged as

$$\epsilon_{L^2}^q(F_{(\cdot)}^q) = \frac{\|\nabla \cdot \overline{\mathbf{u}'q'} - F_{(\cdot)}^q(\dots)\|_{L^2}^2}{\|\nabla \cdot \mathbf{u}'q'\|_{L^2}^2}, \quad (9)$$

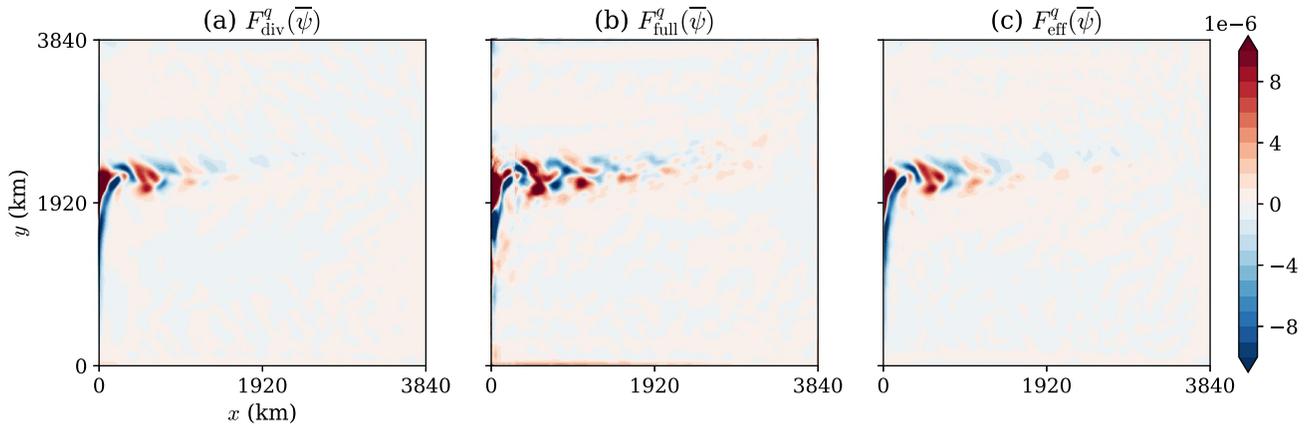
where  $F_{(\cdot)}^q$  denotes the divergence of the eddy PV flux predicted from the models  $f_{(\cdot)}^q(\dots)$ , and the  $L^2$  norm is defined as

$$\|g\|_{L^2}^2 = \int_A g^2 \, dA \quad (10)$$

for some scalar field  $g$ , where the integration is done over the whole computation domain. Each ensemble member makes a set of predictions with an associated mismatch, and the associated averages and standard deviations over the ensemble were computed to judge model skill.

We note that the test for skill chosen here is inherently harder and biased *against* the models trained on the eddy PV fluxes (filtered or otherwise), since an extra divergence operation is required. The above choice to compare the divergence of the eddy PV flux was taken noting that we want a quantity that is comparable across the three sets of models, and there is a theoretical issue in comparing quantities at the eddy PV flux level since that requires integrating the relevant predictions, but is then subject to a choice of boundary condition.

One could argue whether it is the  $L^2$  mismatches we are ultimately interested in, since we may be interested in the patterns rather than the exact locations of the predicted quantity for example. While a Wasserstein metric (e.g., Villani, 2008) would serve that purpose, a simpler and more readily computed quantity is the Sobolev semi-norms (e.g., Thiffeault, 2012) given by



**Figure 3.** Prediction of the divergence of eddy PV flux (units of  $\text{m}^2 \text{s}^{-2}$ ) from one of the ensemble member of models. (a)  $F_{\text{div}}^q(\bar{\psi})$ , (b)  $F_{\text{full}}^q(\bar{\psi})$ , (c)  $F_{\text{eff}}^q(\bar{\psi})$ . The target reference data shown in Figure 1a.

$$\|g\|_{\dot{H}^p}^2 = \int_A |(-\nabla^2)^p g|^2 dA = \sum_{k^2+l^2 \neq 0} (k^2 + l^2)^p |\hat{g}_{k,l}|^2, \quad (11)$$

where  $\hat{g}_{k,l}$  are the Fourier coefficients of  $g$ ,  $(k, l)$  are the respective wavenumbers, and the link between integral and sum follows from Parseval's theorem (e.g., if  $p = 0$  then it is the  $L^2$  norm above when the  $k = l = 0$  mode is included). Sobolev semi-norms with negative  $p$  will weigh the lower wavenumbers (i.e., the larger-scale patterns) more, and in this instance a lower value of the normalized mismatch

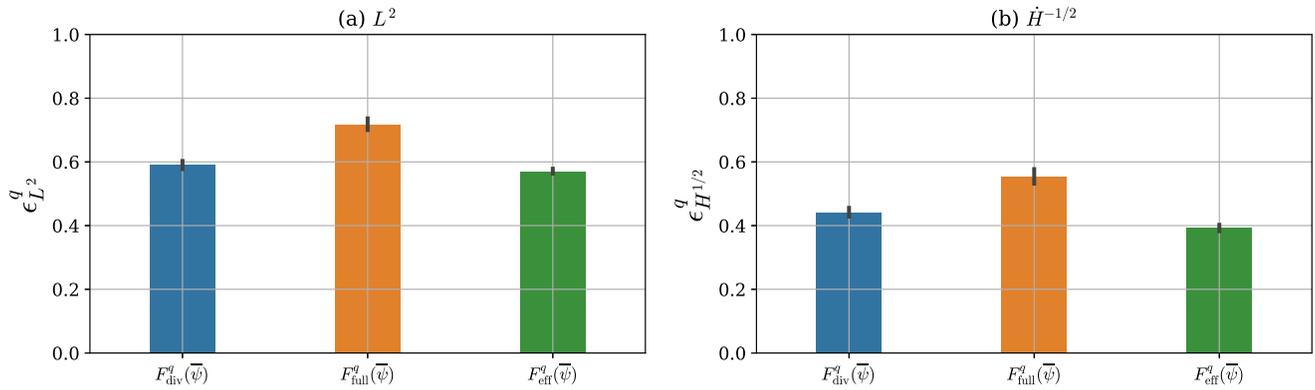
$$\epsilon_{\dot{H}^p}^q(F_{(\cdot)}^q) = \frac{\|\nabla \cdot \bar{\mathbf{u}}' q' - F_{(\cdot)}^q(\dots)\|_{\dot{H}^p}^2}{\|\nabla \cdot \bar{\mathbf{u}}' q'\|_{\dot{H}^p}^2} \quad (12)$$

indicates that the mismatches at the large-scales are smaller. Since we are dealing with finite approximations so that  $k^2 + l^2 < \infty$ , we can perform the computation, even if the formal definition for the  $\dot{H}^p$  semi-norms is generally for fields with zero mean and on a periodic domain and such that the infinite sum converges. For the work here we will focus on the case of  $p = -1/2$ , sometimes referred to as the mix-norm (e.g., Thiffeault, 2012); conclusions below are qualitatively the same if  $p = -1$  or  $p = -2$  were chosen (not shown).

#### 4.1. Models Trained on Eddy PV Fluxes

We first focus on models trained on the data based on the eddy PV flux with the time-mean streamfunction  $\bar{\psi}$  as the input. Figure 3 shows the predicted divergence of the eddy PV flux  $F_{\text{div}/\text{full}/\text{eff}}^q(\bar{\psi})$  as an output from one of the model ensemble members. Compared to the target given in Figure 1a, the predictions are more smooth with fewer small-scale features, arising from a combination of the fact that CNNs were used, and that our prediction step leads to some averaging of the overlapping regions. Visually, the predictions  $F_{\text{div}}^q(\bar{\psi})$  and  $F_{\text{eff}}^q(\bar{\psi})$  are almost indistinguishable, the latter having a slightly stronger signal downstream of the Western Boundary Current. On the other hand, the prediction  $F_{\text{full}}^q(\bar{\psi})$  shows more fluctuations than the other two cases. The larger amount of small-scale features in  $F_{\text{full}}^q(\bar{\psi})$  likely arises because the model is predicting the eddy PV flux first and then having its numerical divergence computed, so any small fluctuations that arise from the prediction is amplified. In that regard, the fact that the prediction  $F_{\text{eff}}^q(\bar{\psi})$  is so visually similar to  $F_{\text{div}}^q(\bar{\psi})$  is rather remarkable.

Figure 4 shows the more quantitative measure of computing the  $L^2$  and the  $\dot{H}^{-1/2}$  mismatches given in Equations 10 and 11 respectively. The results show that the models trained on the filtered eddy PV flux  $-\nabla \Psi_{\text{eff}}^q$  outperforms the models trained on the full eddy PV flux  $\bar{\mathbf{u}}' q'$ , and have a comparable or even better performance compared to the models trained on the divergence of the eddy PV flux  $\nabla \cdot \bar{\mathbf{u}}' q'$ . The differences in skill are visually obvious between the models trained on the full eddy flux  $\bar{\mathbf{u}}' q'$  and the filtered eddy flux  $-\nabla \Psi_{\text{eff}}^q$ . The  $\dot{H}^{-1/2}$  mismatch is smaller in the models trained from the filtered eddy flux  $-\nabla \Psi_{\text{eff}}^q$  compared to the divergence of



**Figure 4.** Ensemble average and quartiles of the normalized (a)  $L^2$  and (b)  $\dot{H}^{-1/2}$  mismatches, given by Equations 9 and 12 respectively, for the models predicting the divergence of the eddy PV flux (Figure 1a). Blue denotes models trained on the divergence of the eddy fluxes, orange denotes models trained on the full eddy fluxes, and green denotes models trained on the filtered eddy fluxes.

the eddy flux  $\nabla \cdot \overline{\mathbf{u}'\mathbf{q}'}$ , but is too close to call in the  $L^2$  mismatch (e.g., we do not have  $p < 0.05$  using the Student's  $t$ -test (Student, 1908) under the null hypothesis that the means of  $F_{\text{div}}^q(\overline{\psi})$  and  $F_{\text{eff}}^q(\overline{\psi})$  are the same).

The results here lend support to our expectation that the presence of rotational fluxes contaminate and degrade the accuracy of a trained model, and the eddy force function provides one means of filtering such rotational fluxes that leads to comparable model performance to the more standard choice of training from the divergence of the eddy fluxes. The result is all the more remarkable when we note that tests based on the models' ability in reproducing the divergence of the eddy flux is intrinsically harder and biased against models trained on the filtered flux, since an additional divergence operation that is expected to amplify errors is required.

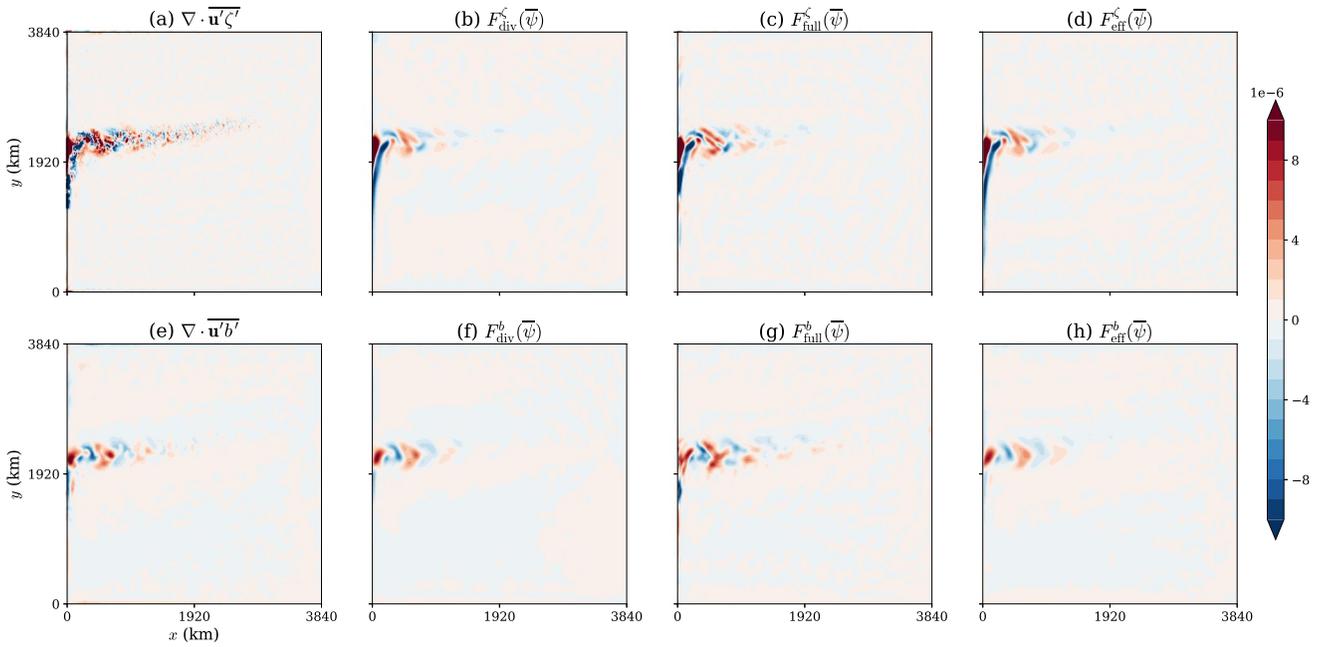
#### 4.2. Other Choice of Targets and Inputs

Following the notation outline above, Figure 5 shows the target data  $\nabla \cdot \overline{\mathbf{u}'\boldsymbol{\zeta}'}$  and  $\nabla \cdot \overline{\mathbf{u}'\mathbf{b}'}$ , and the analogous predictions of the divergence of the fluxes denoted by  $F_{\text{div/full/eff}}^{\zeta/b}(\overline{\psi})$  from one of the ensemble members.

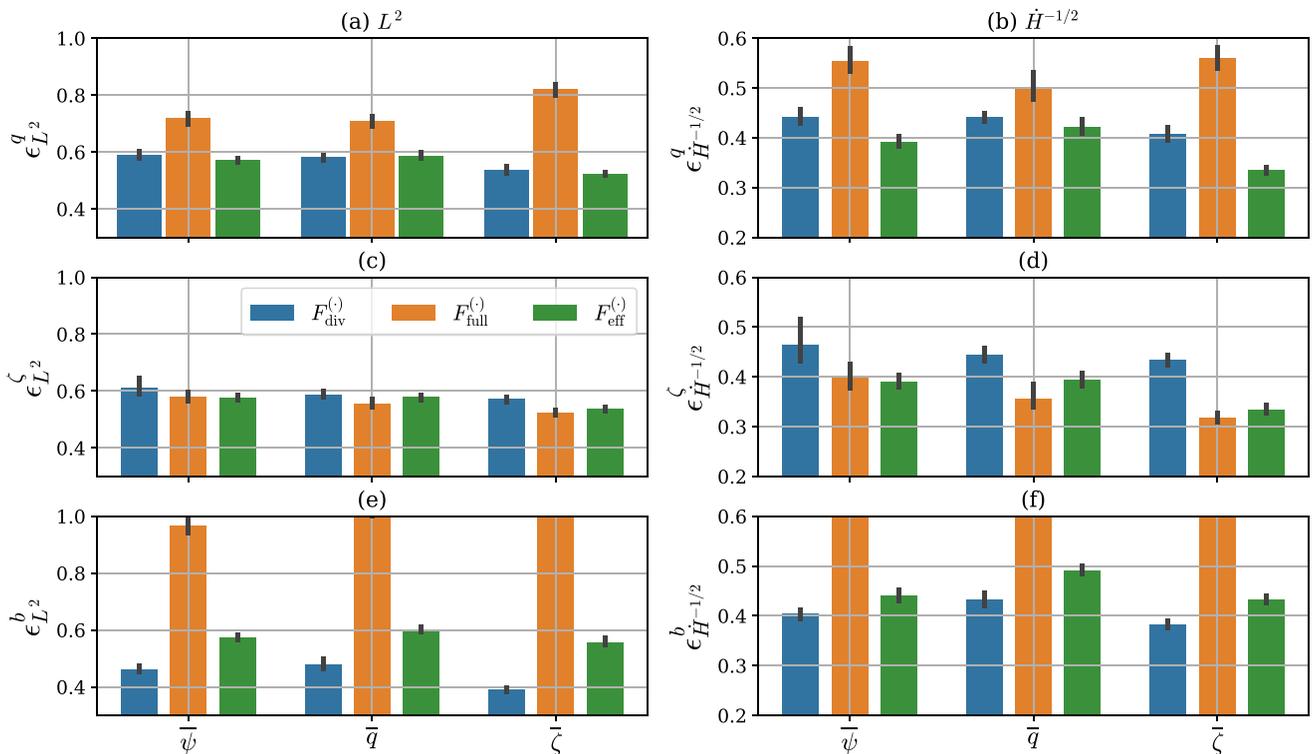
The predictions are again more smooth than the diagnosed target data, particularly noticeable for the prediction of the divergence of the eddy relative vorticity flux in Figures 5b–5d. For the eddy buoyancy flux case, the diagnosed target data is already relatively smooth. We note that, visually,  $F_{\text{full}}^b(\overline{\psi})$  in Figure 5g seems to possess extra features particularly in the downstream region, while  $F_{\text{eff}}^b(\overline{\psi})$  and  $F_{\text{div}}^b(\overline{\psi})$  in Figures 5f and 5h seems to be capturing the patterns in the target data well, with some visual hints that the prediction from  $F_{\text{div}}^b(\overline{\psi})$  has slightly sharper features.

For a more quantitative measure, we show in Figure 6 the  $L^2$  and  $\dot{H}^{-1/2}$  mismatches in  $F_{\text{div/full/eff}}^{q/\zeta/b}(\overline{\psi}/\overline{q}/\overline{\zeta})$ , totaling the  $3^3 = 27$  possible combinations. The conclusions over all these possible choices are largely what was drawn from before but with minor differences. The models trained on the filtered eddy fluxes outperform those trained on the full eddy fluxes (except for the case of eddy relative vorticity fluxes), and are comparable or better than models trained on the divergence of the flux (except in the case of the eddy buoyancy fluxes).

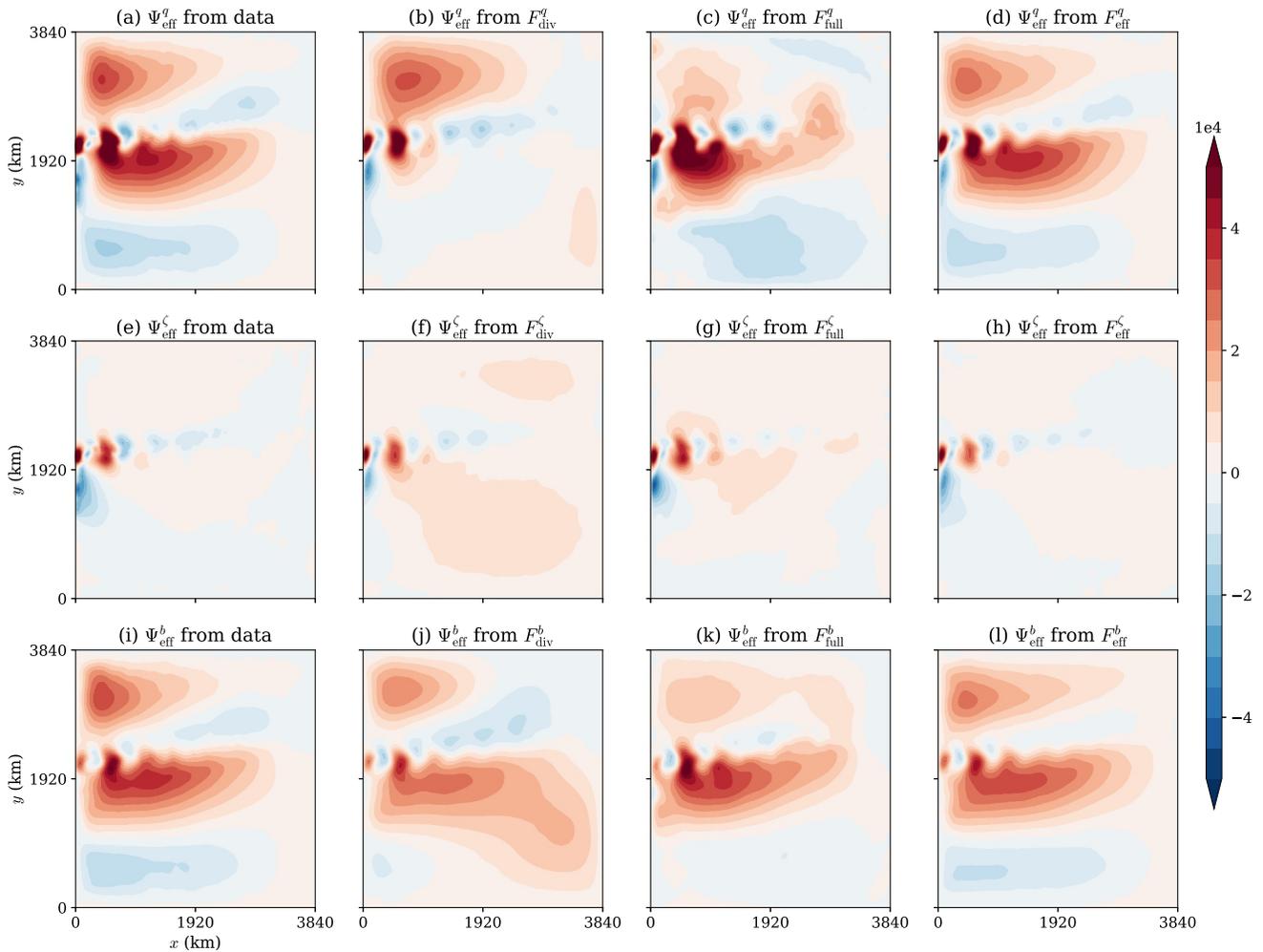
Noting that eddy PV fluxes have contributions from the eddy buoyancy as well as eddy relative vorticity fluxes, it is curious that models trained on the filtered eddy fluxes compared with models trained on the divergence of the flux appear to perform worse for the eddy buoyancy flux case (bottom row of Figure 6). However, the performance is reasonable in the eddy relative vorticity flux case (middle row of Figure 6), such that the resulting skill in the eddy PV flux case (top row of Figure 6) still remains comparable, and possibly slightly better in the  $\dot{H}^{-1/2}$  mismatch, indicating better matching in terms of large-scale patterns. One possible explanation for the degradation in performance for eddy buoyancy fluxes is that  $\nabla \cdot \overline{\mathbf{u}'\mathbf{b}'}$  is already relatively smooth and larger-scale (Figure 5e), which might be favorable for direct use as training data. On the other hand, the eddy relative vorticity fluxes are inherently smaller-scale (Figure 5a), and the presence of small-scale fluctuation might be unfavorable for direct use as training data, but does not affect models trained on the filtered fluxes as such since



**Figure 5.** Target data and predictions associated with (top row, a–d) eddy relative vorticity flux (related to the Reynolds stress, units of  $\text{m s}^{-2}$ ) and (bottom row, e–h) eddy buoyancy flux (related to the form stress, units also of  $\text{m s}^{-2}$  taking into account of the extra factors). Showing (a, e) the divergence of the time-averaged eddy relative vorticity and buoyancy flux, and a sample (b, f)  $F_{\text{div}}^{z/b}(\bar{\psi})$ , (c, g)  $F_{\text{full}}^{z/b}(\bar{\psi})$ , (d, h)  $F_{\text{eff}}^{z/b}(\bar{\psi})$  from one of the ensemble members.



**Figure 6.** Ensemble average and quartiles of the normalized (a, c, e)  $L^2$  and (b, d, f)  $\dot{H}^{-1/2}$  mismatch, given by Equations 9 and 12 respectively, for the models predicting the divergence of the eddy PV flux (a, b), relative vorticity (cf. momentum) flux (c, d), and buoyancy flux (e, f), over various choices of inputs. Blue denotes models trained on the divergence of the eddy fluxes, orange denotes models trained on the full eddy fluxes, and green denotes models trained on the filtered eddy fluxes. Top row is identical to Figure 4. The mismatches in  $F_{\text{full}}^b(\bar{q})$  and  $F_{\text{full}}^b(\bar{\zeta})$  are out of range in panels (e and f), with values around 1.0 and 1.5 respectively.



**Figure 7.** (a, e, i) Target eddy force functions  $\Psi_{\text{eff}}^{q/c/b}$ , and eddy force functions associated with prediction from (b, f, j) divergence of the eddy fluxes, (c, g, k) full eddy fluxes, and (d, h, l) filtered eddy fluxes, respectively for the PV, relative vorticity and buoyancy variable from one of the ensemble members. All data shown here are in units of  $\text{m}^2 \text{s}^{-2}$ .

the training data is by definition more smooth. The performance of models based on the full eddy relative vorticity fluxes is somewhat surprising, but may be to do with the smaller component of the rotational fluxes. On examining the decomposition into divergent and rotational parts via the eddy force function (cf. Figures 1b, 1c, 1e, and 1f, not shown), it is found that the divergent component is smaller by about a factor of 2 in the eddy relative vorticity flux, but a factor of 10 in the eddy buoyancy and PV flux. The results seem to suggest that the main benefits of filtering dynamically inert rotational fluxes would be in the eddy buoyancy and PV where the rotational component is large.

For completeness, we show in Figure 7 the analogous eddy force functions associated with the predictions from the trained models from one of the ensemble members (although observations detailed here are robust upon examining the outputs from other members); note the analogous mismatches would be closely related to the  $\dot{H}^{-2}$  semi-norm as defined in Equation 11, but with a difference in the choice of boundary conditions. The predictions from models trained on the filtered eddy fluxes (panels d, h, i) have patterns that are largely aligned with the diagnosed eddy force functions from the data (panels a, e, i) up to minor discrepancies (e.g., downstream patterns in panel d compared to panel a, and panel l compared to panel i). The predictions from models trained on the full eddy fluxes (panels c, g, k) show similar patterns although with somewhat more mismatches, particularly in the PV and buoyancy eddy force functions. By contrast, the predictions from the divergence of the eddy fluxes (panels b, f, j) show large-scale disagreements in all three variables, the mismatches being visually the gravest in

the PV and buoyancy variables. Given that the eddy force function has an interpretation that  $\nabla\Psi_{\text{eff}} \cdot \nabla\overline{\psi}$  encodes the sign of energy exchange between the mean and eddy component (Maddison et al., 2015), the finding here suggests the predictions from models trained on the divergence of the eddy fluxes are very likely representing erroneous energy transfers, particularly for processes associated with eddy buoyancy fluxes.

### 4.3. Model Skill and Sensitivity to Noise

The above observations bring into question of whether the models are sensitive to small-scale features, such as that arising from the numerical model (e.g., spatial resolution) and/or amount of averaging (e.g., time window in a time-average, number of ensemble members in an ensemble average). Sensitivity to small-scale features would be suggestive of large uncertainties in the models, with implications on their possible use in extracting information from data for example. To explore the sensitivity of skill to noise in the data, we consider a set of experiments where we add noise  $\eta(x, y)$  to the data at the *training* stage, and judge the models' performance on its ability in predicting the target data without noise. To make sure we are comparing models in a consistent manner, we add an appropriately scaled Gaussian distributed noise  $\eta(x, y)$  to the eddy fluxes  $(\overline{u'q'}, \overline{u'\zeta'}, \overline{u'b'})$ , from which we compute the divergence of the eddy flux as well as the eddy force function from the noisy data, and train the models using the procedure outlined above. In that sense the whole set of models are exposed to the *same* choice of noise, since 1 unit of noise at the divergence level is not necessarily the same as 1 unit of noise at the streamfunction level. The noise level here is measured in units of the standard deviation of the eddy flux data. The hypothesis is that the models trained on the filtered eddy fluxes are more robust than those trained on the divergence of the eddy fluxes, and able to maintain model skill with increased levels of noise.

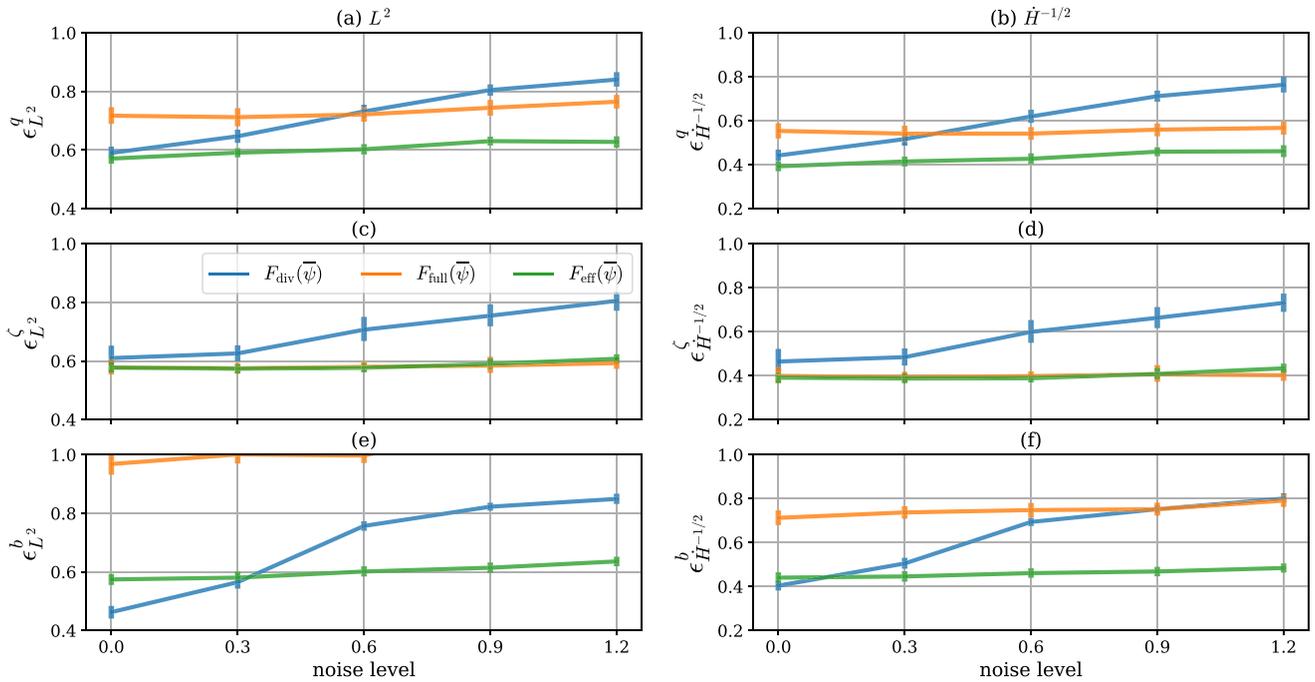
Note the stochastic noise  $\eta(x, y)$  is formally non-differentiable in space, so that its divergence is not well-defined. In terms of numerical implementation, however, the random numbers sampled from the appropriately scaled Gaussian distribution are the nodal values of the finite element mesh used in FEniCS, and there is a projection onto a linear basis, so that a derivative operation on the projected  $\eta(x, y)$  is allowed within FEniCS, though the operation may be numerically sensitive. An approach we considered is spatially filtering the noise field. We consider solving for some  $\tilde{\eta}(x, y)$  satisfying

$$(1 - L^2\nabla^2)^2\tilde{\eta} = \eta \quad (13)$$

with no-flux boundary conditions, and add the resulting  $\tilde{\eta}(x, y)$  to the training data; note that the “noise level” here refers to the magnitude of  $\eta(x, y)$ , and that  $\max|\tilde{\eta}(x, y)| < \max|\eta(x, y)|$  by construction. The resulting  $\tilde{\eta}$  is by construction differentiable at least once so that a divergence is well-defined. For the operator  $(1 - L^2\nabla^2)^2$ , the associated Green's function has a characteristic length-scale  $L$  that can be interpreted as a filtering length-scale where the radial spectral power density decreases significantly after  $L$  (closely related to the Matérn autocovariance, e.g., Lindgren et al., 2018; Whittle, 1963).

The  $L^2$  and  $\dot{H}^{-1/2}$  mismatches of  $F_{\text{div}/\text{full}/\text{eff}}^{\eta/\zeta/b}(\overline{\psi})$  to the data as a function of noise level for the ensemble of models is shown in Figure 8, and consistently we find that the models trained up on the eddy force function outperform the models trained upon the divergence of the eddy flux. The former shows a relative insensitivity to noise level, while the latter shows a rapid degradation in skill with noise level. It would seem that the use of eddy force function data alleviates the sensitivity to small-fluctuations in data, at least in the present measure and approach.

The reduced sensitivity in the models trained on the filtered fluxes to noise might have been anticipated, since the eddy force function is a result of an elliptic solve of a Poisson equation, which leads to a smoothing of the data (via an inverse Laplacian operator). We would however argue that the relative insensitivity to noise level is somewhat surprising, since there is no guarantee the presence of even reduced fluctuations at the streamfunction level would stay small after spatial derivatives operations, since we are using the divergence of the eddy flux as the target for the measure of skill. While one could also argue that the present test is inherently a hard test for models trained upon the divergence of the eddy flux, we argue the conclusions are robust regardless of whether the noise is added at the flux, divergence of flux or streamfunction level. In fact, the use of the divergence of a flux as training data is likely the cause for sensitivity to noise: a inherently small-scale field is sensitive to the presence of small-scale features, so is likely to lead to issues sensitivity to such features.



**Figure 8.** Ensemble average and quartiles of the normalized (a, c, e)  $L^2$  and (b, d, f)  $\hat{H}^{-1/2}$  mismatch given by Equations 9 and 12 respectively, as a function of noise level for the models predicting the divergence of the eddy PV flux (a, b), relative vorticity (cf. momentum) flux (c, d), and buoyancy flux (e, f), using the time-mean streamfunction  $\bar{\psi}$  as the input. Blue denotes models trained on the divergence of the eddy fluxes, orange denotes models trained on the full eddy fluxes, and green denotes models trained on the filtered eddy fluxes. Model skill is out of range for  $F_{full}^b(\bar{\psi})$  in panel (e).

The conclusions in the above are qualitatively robust for different choices of the filtering length-scale  $L$ : with reduced  $L$ , the degradation of skill in models trained on the divergence of the eddy fluxes is more rapid with noise level, but the skill of models trained on the filtered eddy fluxes is still relatively insensitive to noise level, and consistently more skillful than models trained on the divergence of the eddy fluxes. The conclusions are also robust for different choices of inputs ( $\bar{\zeta}$  and  $\bar{q}$ ), and with sample calculations employing other choices of smoothing, coarse-graining (e.g., Aluie, 2019) or filtering (e.g., Grooms et al., 2021) of the noise field  $\eta(x, y)$ .

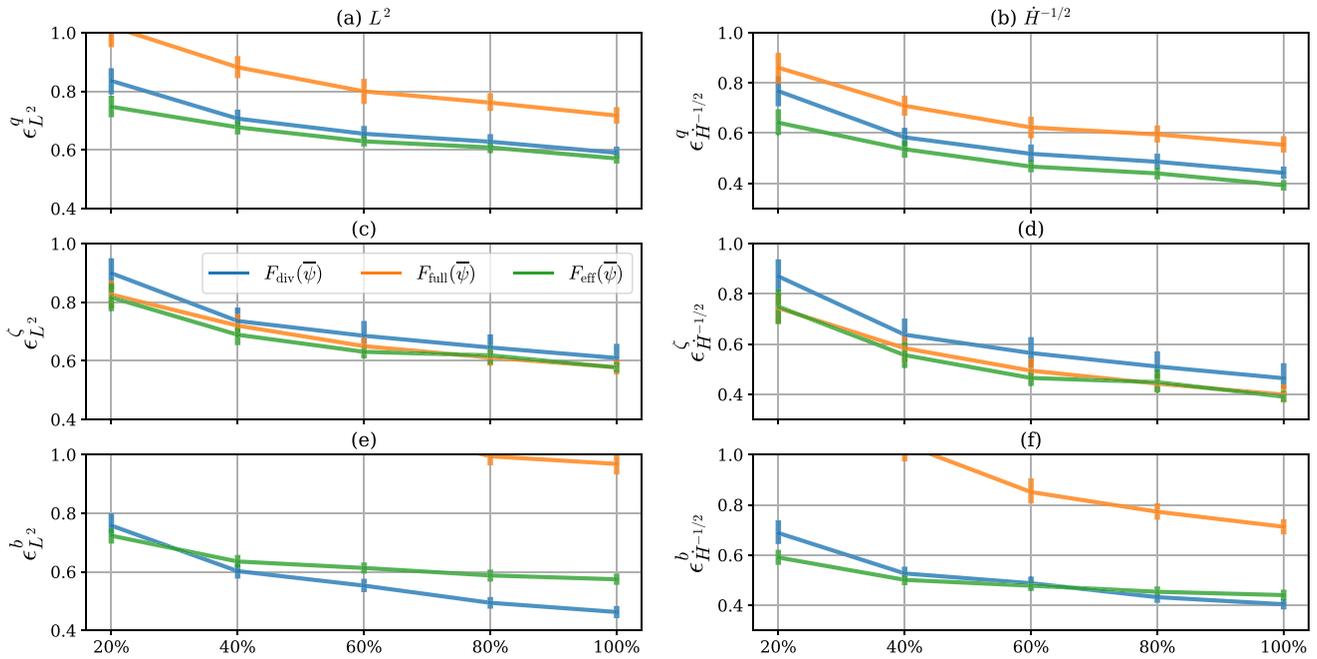
#### 4.4. Model Skill and Dependence on Data Amount

The reported results so far are from models trained upon a 100% of the data (6,400 images) in an 80:20 ratio of training to validation data, from which the skill is computed for the model's ability to reduce the global mismatches throughout the domain. Figure 9 shows the skill of the models as a function of data percentage used as part of the training, with the time-mean streamfunction as the input, keeping the same 80:20 ratio of training to validation data in all cases.

As expected, the skill of models decrease with the percentage of training data provided. However, the conclusions drawn from Sections 4.1 and 4.2 continue to hold. The conclusions in Section 4.3 regarding reduced sensitivity hold in the models trained on the eddy fluxes (be it filtered or otherwise) in sample calculations at reduced percentage of training data (not shown). Thus we have evidence in support that the conclusions made thus far are robust even in the more standard regime where some amount of training data is withheld for testing purposes.

### 5. Conclusions and Outlooks

Data-driven methods are increasingly being employed in problems of Earth System Modeling. Such methods can in principle be leveraged to not only improve our modeling efforts, and also deepen our underlying understanding of the problems. Most works in the literature thus far has focused on demonstrating the computational efficacy and predictive power of the machine-learning methods and algorithms. Here we take a complimentary line of investigation in considering the choice and quality of data itself being fed to the algorithms, for a case where we



**Figure 9.** Ensemble average and quartiles of the normalized (a, c, e)  $L^2$  and (b, d, f)  $\dot{H}^{-1/2}$  mismatch given by Equations 9 and 12 respectively, as a function of percentage of training data exposed to model during training phase (100% = 6,400 frames), for the models predicting the divergence of the eddy PV flux (a, b), relative vorticity (cf. momentum) flux (c, d), and buoyancy flux (e, f), using the time-mean streamfunction  $\bar{\psi}$  as the input. Blue denotes models trained on the divergence of the eddy fluxes, orange denotes models trained on the full eddy fluxes, and green denotes models trained on the filtered eddy fluxes. Model skill is out of range for  $F_{\text{full}}^b(\bar{\psi})$  in panel (e and f).

have some theoretical understanding to inform our choices. While one could argue this is not entirely necessary if we just want something that has skill as measured by the relevant metric(s) for the problem, we argue it is incredibly useful and if not necessary if we want to leverage data-driven methods to learn about the underlying physical problems, and/or to go beyond “black-box” models. Furthermore, the choice of data can in principle improve the training and/or the performance of the data-driven models themselves, so there is a need for such an investigation into data quality and information content.

For this work we focused on the problem of eddy-mean interaction in rotating stratified turbulence in the presence of boundaries, relevant to the modeling and parameterization of ocean dynamics. In such systems it is known that the large-scale mean affects and is affected by the small-scale eddy fluxes, and while we might want to leverage data-driven methods to learn about this relationship, it is known that in the presence of boundaries the eddy feedback onto the mean is invariant up to a rotational gauge (e.g., Fox-Kemper et al., 2003; Eden et al., 2007; J. C. Marshall & Shutts, 1981). The rotational component can be quite large (e.g., Griesel et al., 2009, and also Figure 1 here), and its presence might be expected to negatively impact the training and eventual performance of trained models. One possible way round is to train models based on the divergence of the eddy fluxes (e.g., Bolton & Zanna, 2019; Zanna & Bolton, 2021). Here we propose that data that filters out rotational component of the eddy fluxes be used instead. The approach outlined here we argue to have the advantage in that the resulting field is inherently larger-scale, which would help with model training and sensitivity, and be theoretically more appropriate to use if we want to learn about the underlying physics of the problem, because we do not expect commutativity (i.e., given the nonlinearity, learning from the divergence is not guaranteed to be the same as the divergence of the learned result).

The experimental approach here largely follows that of Bolton and Zanna (2019), where we diagnose the relevant data from a quasi-geostrophic double gyre model to feed to the machine learning algorithm, and compare the trained models' performance in their global predictions. For filtering the eddy flux we employ the eddy force function (e.g., Maddison et al., 2015; Mak et al., 2016; D. P. Marshall & Pillar, 2011), which in the present simply connected quasi-geostrophic system is optimal in removing the rotational fluxes (see Appendix of Maddison et al., 2015). We made the choice here to measure a model's skill in its ability to reproduce the divergence of the

eddy fluxes over an ensemble of models with 20 members and over a variety of inputs. We find that the models trained on the eddy force function are (a) more skillful than those trained on the full eddy flux (except for the relative vorticity eddy fluxes), (b) at least comparable (and on occasion better) in skill than models trained on the divergence of the eddy fluxes (except for the buoyancy eddy fluxes), and (c) the trained models are less sensitive to small-scale fluctuations in the training data. The conclusions appear to be robust up to the amount of training data provided (see Figure 9). Furthermore, there is evidence that the conclusions are also robust as long as a rotational component is filtered out in some sensible way, for example, by the use of a standard Helmholtz decomposition, up to the caveat that we have to choose a boundary condition (see Figure A1).

The first finding is perhaps not unexpected. The latter two findings we argue are not entirely obvious, given divergence operations acting at various steps. For example, sample calculations where a model is trained on the eddy force function directly (and then taking a Laplacian to obtain a prediction of the divergence of eddy flux) leads to larger mismatches, which we attribute to the fact that any mismatches in the predicted eddy force function is significantly amplified by the two derivative operations. With that in mind, the fact that models trained on the filtered flux reported here leads to models with comparable or better skill and less sensitivity to small-scale features in the data are non-trivial results.

Exceptions to the above conclusions are that models trained on the divergence of the eddy buoyancy flux are more skillful (bottom row of Figure 6), and models trained on the eddy relative vorticity flux appear comparable whether the rotational component is filtered out or not (middle row of Figure 6). The former might be rationalized in that the eddy buoyancy flux is already relatively smooth and somewhat larger-scale, so that training on its divergence is not such an issue; however, we also note that the buoyancy eddy force functions associated with the predictions of models trained on the divergence of the eddy buoyancy flux seem to perform the worse (bottom of Figure 7), implying erroneous predictions of eddy energy pathways. The latter observation is possibly to do with the fact that in the eddy relative vorticity flux, the rotational component is comparable in size to the divergent component, as opposed to the rotational component being a factor of 10 smaller in the eddy PV and buoyancy flux (see Figures 1b, 1c, 1d, and 1e for eddy PV flux), and the effect of filtering is somewhat marginal. One saving grace is that, in the quasi-geostrophic system, the potential vorticity (with contributions from relative vorticity and buoyancy) is the master variable, and that while models trained up on the relative vorticity or buoyancy fluxes perform better separately, the models trained up on the filtered eddy flux have skill in the PV eddy flux, where PV is the master variable in the quasi-geostrophic system.

One thing we caution here is drawing a one-to-one comparison of the present work with that of Bolton and Zanna (2019) and Zanna and Bolton (2021). While it is true those works utilize a similar model, experimental procedure and data to this work, the choice of averages are different. Their work utilizes a spatial average, and the eddy flux data there is defined as the difference between the filtered divergence and the divergence of the filtered field. Here we utilize a *time* average, which is in line with the definition of the eddy force function in Maddison et al. (2015) via a Reynolds average. While we have not attempted a similar investigation in the case of spatial averaging or the more general case beyond the quasi-geostrophic setting, the eddy force function presumably can be similarly defined going back to the original definition in D. P. Marshall and Pillar (2011), even if optimality cannot be shown as in Maddison et al. (2015). Failing that, a Helmholtz decomposition might suffice, again on the caveat that boundary conditions need to be chosen. This part is beyond the scope of the present work and left as a future investigation.

Because of the choice of time average, we have limited data in time, and one could wonder whether our conclusions are simply to do with the limited data availability. This is unlikely the case: we also carried out an analogous investigation with rolling time averages as well as *ensemble* averages (not shown), and the conclusions drawn from those results are essentially identical to those here. This is perhaps not surprising noting that the rolling time averages for a long enough window and the ensemble averages shown no strong deviations from each other in the present system, but we note this is likely only true for a sufficiently simple system with no strong evidence of internal modes of variability.

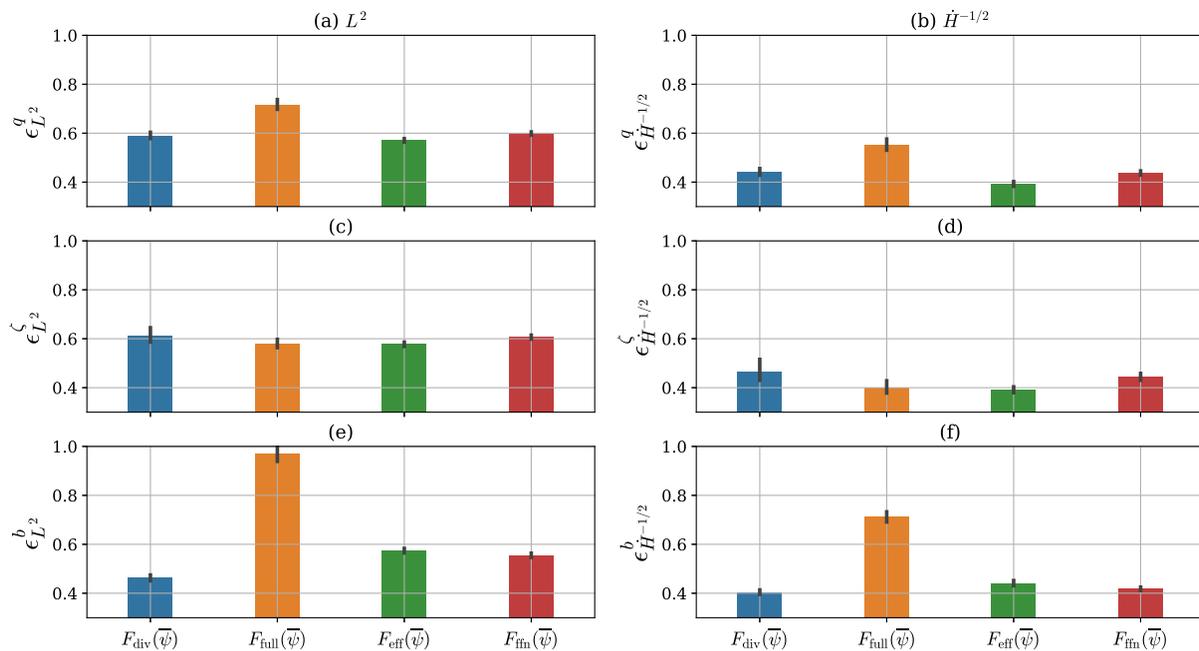
The main intention of the present work is to demonstrate that not all data choices are equal for machine learning in the Earth System Modeling setting. For the case of rotating stratified turbulence, the eddy force function is a potentially useful quantity if we aim to leverage data-drive methods for model skill or for learning about the underlying physics of the problem, given the various theoretical expectations highlighted in this work. Other choices such as a standard Helmholtz decomposition could be used to solve for the divergent component may be

useful, although the eddy force could still be used for physical interpretation. We note that while skill in reproducing eddy forcing is one target, we have not examined here on the ability of the model to reproduce the mean state, and the present procedure might be termed an “offline” approach. Learning “online” (e.g., Frezat et al., 2022) may be more appropriate for parameterization purposes to improve on the mean response, and it would be of interest to see whether filtering of the eddy flux as discussed here would confer any benefits to model learning. We note that the computation of the Helmholtz decomposition should be relatively quick in a periodic domain via transformations into Fourier spectral space.

The present work also highlights questions relating to information content of data. While quantifying absolute data information content is likely quite difficult, it should be at least possible to compute a relative measure, even if empirically. One might ask an analogous question of the input data. The work of Bolton and Zanna (2019) suggests that training with data from regions with higher eddy kinetic energy leads to better model performance in terms of accuracy for example, suggestive of higher information content in said region. Within the present experimental framework, we could consider training based on a biased sampling that favor regions with higher eddy energy content, with the hypothesis that the latter case leads to models with higher accuracy from a statistical point of view. Further, we could investigate the case of multiple inputs, where we hypothesize that eddy energy and a mean state variable as inputs might lead to improved performance compared to say two mean state variables: in the current quasi-geostrophic setting, the mean state variables are functionally related to each other, possibly leading to redundant information, while the eddy energy might be dependent on the mean state, but captures eddy statistics instead and providing complementary information. This investigation is ongoing and will be reported elsewhere in due course.

### Appendix A: Training From Data Beyond the Eddy Force Function

Here provide sample results from an analogous investigation into using data from a standard Helmholtz decomposition, as in Equation 6. In this case we have to make a choice on the imposed boundary condition, and we take this to be a homogeneous Neumann condition, which corresponds to zero normal eddy flux. This is in



**Figure A1.** Ensemble average and quartiles of the normalized (a, c, e)  $L^2$  and (b, d, f)  $H^{-1/2}$  mismatch given by Equations 9 and 12 respectively, for the models predicting the divergence of the eddy PV flux (a, b), relative vorticity (cf. momentum) flux (c, d), and buoyancy flux (e, f), using the time-mean streamfunction  $\bar{\psi}$  as the input. Blue denotes models trained on the divergence of the eddy fluxes, orange denotes models trained on the full eddy fluxes, green denotes models trained on the filtered eddy fluxes by means of an eddy force function (cf. Figure 6), and red denotes the models trained on the filtered eddy flux by means of a standard Helmholtz decomposition in Equation 6 using a homogeneous Neumann condition (i.e., zero normal eddy flux). The figure is to be directly compared with Figure 4, but noting the use of a different y-axis limit.

contrast to the homogeneous Dirichlet condition used for the eddy force function, which corresponds to the zero normal *mean* geostrophic flow condition (related to the momentum tendency; Maddison et al., 2015).

Figure A1 shows the results analogous to Figure 4 (but note the change in the y-axis limits) for varying model data but for a fixed input of the time-mean streamfunction. It can be seen that while the performance of models trained on the filtered fluxes by means of eddy force functions are generally more skillful, the conclusions drawn in this work appear to also hold for the case where filtering is performed using a Helmholtz decomposition. The results would suggest that some degree of rotational flux removal is desirable. Whether this continues to hold in the non-simply connected quasi-geostrophic setting or the more general primitive equation setting remains to be investigated.

## Data Availability Statement

This work utilizes FEniCS (2019.1.0) that is available as a Python package. The qgm2 source code, sample model data and scripts used for generating the plots in this article from the processed data are available through the repository at Yan and Mak (2023).

## Acknowledgments

This research was funded by both RGC General Research Fund 16304021 and the Center for Ocean Research in Hong Kong and Macau, a joint research center between the Qingdao National Laboratory for Marine Science and Technology and Hong Kong University of Science and Technology. We thank James Maddison and Liying Yeow for various scientific and technical comments in relation to the present investigation, the former for providing the qgm2 code for use in the present work, and the referees for providing insightful comments that enhanced the scientific and presentation aspect of the manuscript.

## References

- Alnaes, M. S., Logg, A., Ølgaard, K. B., Rognes, M. E., & Wells, G. N. (2014). Unified form language: A domain-specific language for weak formulations of partial differential equations. *ACM Transactions on Mathematical Software*, 40, 9:1–9:37.
- Aluie, H. (2019). Convolutions on the sphere: Commutation with differential operators. *GEM: International Journal of Geometry*, 10, 1–31. <https://doi.org/10.1007/s13137-019-0123-9>
- Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., et al. (2021). Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1), 5124. <https://doi.org/10.1038/s41467-021-25257-4>
- Barnes, E. A., Barnes, R. J., Martin, Z. K., & Rader, J. K. (2022). This looks like that there: Interpretable neural networks for image tasks when location matters. *Artificial Intelligence for the Earth Systems*, 1(3), e220001. <https://doi.org/10.1175/AIES-D-22-0001.1>
- Berloff, P. (2005). On dynamically consistent eddy fluxes. *Dynamics of Atmospheres and Oceans*, 38(3–4), 123–146. <https://doi.org/10.1016/j.dynatmoce.2004.11.003>
- Besombes, C., Pannekoucke, O., Lapeyre, C., Sanderson, B., & Thual, O. (2021). Producing realistic climate data with generative adversarial networks. *Nonlinear Processes in Geophysics*, 28(3), 347–370. <https://doi.org/10.5194/npg-28-347-2021>
- Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9), 098302. <https://doi.org/10.1103/PhysRevLett.126.098302>
- Bolibar, J., Rabatel, A., Gouttevin, I., Galiez, C., Condom, T., & Sauquet, E. (2020). Deep learning applied to glacier evolution modelling. *The Cryosphere*, 14(2), 565–584. <https://doi.org/10.5194/tc-14-565-2020>
- Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and sub-grid parameterisation. *Journal of Advances in Modeling Earth Systems*, 11(1), 376–399. <https://doi.org/10.1029/2018MS001472>
- Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, 77(12), 4357–4375. <https://doi.org/10.1175/JAS-D-20-0082.1>
- Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, 11(8), 2728–2744. <https://doi.org/10.1029/2019MS001711>
- Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., et al. (2023). Discovering causal relations and equations from data.
- Clare, M. C. A., Sonnewald, M., Lguensat, R., Deshayes, J., & Balaji, V. (2022). Explainable Artificial Intelligence for Bayesian neural networks: Toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, 14(11), e2022MS003162. <https://doi.org/10.1029/2022MS003162>
- Connolly, C., Barnes, E. A., Hassanzadeh, P., & Pritchard, M. (2023). Using neural networks to learn the Jet Stream forced response from natural variability. *Artificial Intelligence for the Earth Systems*, 2, e220094. <https://doi.org/10.1175/AIES-D-22-0094.1>
- Eden, C., Greatbatch, R. J., & Olbers, D. (2007). Interpreting eddy fluxes. *Journal of Physical Oceanography*, 37(5), 1282–1296. <https://doi.org/10.1175/jpo3050.1>
- Fox-Kemper, B., Ferrari, R., & Pedlosky, J. (2003). On the indeterminacy of rotational and divergent eddy fluxes. *Journal of Physical Oceanography*, 33(2), 478–483. [https://doi.org/10.1175/1520-0485\(2003\)033<0478:otiora>2.0.co;2](https://doi.org/10.1175/1520-0485(2003)033<0478:otiora>2.0.co;2)
- Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, 14(11), e2022MS003124. <https://doi.org/10.1029/2022MS003124>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from <http://www.deeplearningbook.org>
- Green, J. S. A. (1970). Transfer properties of the large-scale eddies and the general circulation of the atmosphere. *Quarterly Journal of the Royal Meteorological Society*, 96(408), 157–185. <https://doi.org/10.1002/qj.49709640802>
- Griesel, A., Gille, S. T., Sprintall, J., McClean, J. L., & Maltrud, M. E. (2009). Assessing eddy heat flux and its parameterization: A wavenumber perspective from a 1/10° ocean simulation. *Ocean Modelling*, 29(4), 248–260. <https://doi.org/10.1016/j.ocemod.2009.05.004>
- Grooms, I., Loose, N., Abernathy, R., Steinberg, J. M., Bachman, S. D., Marques, G., et al. (2021). Diffusion-based smoothers for spatial filtering of gridded geophysical data. *Journal of Advances in Modeling Earth Systems*, 13(9), e2021MS002552. <https://doi.org/10.1029/2021MS002552>
- Guan, Y., Subel, A., Chattopadhyay, A., & Hassanzadeh, P. (2023). Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate LES. *Physica D*, 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>

- Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, 3(8), 667–674. <https://doi.org/10.1038/s42256-021-00374-3>
- Jones, D. C., Holt, H. J., Meijers, A. J. S., & Shuckburgh, E. (2019). Unsupervised clustering of Southern Ocean Argo float temperature profiles. *Journal of Geophysical Research: Oceans*, 124(1), 390–402. <https://doi.org/10.1029/2018JC014629>
- Karabasov, S. A., Berloff, P. S., & Golovizin, V. M. (2009). CABARET in the ocean gyres. *Ocean Modelling*, 30(2–3), 155–168. <https://doi.org/10.1016/j.ocemod.2009.06.009>
- Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmailzadeh, S., et al. (2021). Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200093. <https://doi.org/10.1098/rsta.2020.0093>
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *Iclr (poster)*.
- Lindgren, F., Rue, H., & Lindström, J. (2018). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B (Methodology)*, 73(4), 423–498. <https://doi.org/10.1111/j.1467-9868.2011.00777.x>
- Logg, A., Mardal, K. A., & Wells, G. N. (2012). *Automated solution of differential equations by the finite element method*. Springer.
- Logg, A., & Wells, G. N. (2010). DOLFIN: Automated finite element computing. *ACM Transactions on Mathematical Software*, 37(2), 20:1–20:28. <https://doi.org/10.1145/1731022.1731030>
- Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, 14(8), e2022MS003105. <https://doi.org/10.1029/2022MS003105>
- Maddison, J. R., & Marshall, D. P. (2013). The Eliassen–Palm flux tensor. *Journal of Fluid Mechanics*, 729, 69–102. <https://doi.org/10.1017/jfm.2013.259>
- Maddison, J. R., Marshall, D. P., & Shipton, J. (2015). On the dynamical influence of ocean eddy potential vorticity fluxes. *Ocean Modelling*, 92, 169–182. <https://doi.org/10.1016/j.ocemod.2015.06.003>
- Mak, J., Maddison, J. R., & Marshall, D. P. (2016). A new gauge-invariant method for diagnosing eddy diffusivities. *Ocean Modelling*, 104, 252–268. <https://doi.org/10.1016/j.ocemod.2016.06.006>
- Marshall, D. P., Maddison, J. R., & Berloff, P. S. (2012). A framework for parameterizing eddy potential vorticity fluxes. *Journal of Physical Oceanography*, 42(4), 539–557. <https://doi.org/10.1175/JPO-D-11-048.1>
- Marshall, D. P., & Pillar, H. R. (2011). Momentum balance of the wind-driven and meridional overturning circulation. *Journal of Physical Oceanography*, 41(5), 960–978. <https://doi.org/10.1175/2011jpo4528.1>
- Marshall, J. C. (1981). On the parameterization of geostrophic eddies in the ocean. *Journal of Physical Oceanography*, 11(2), 257–271. [https://doi.org/10.1175/1520-0485\(1981\)011<0257:otpoge>2.0.co;2](https://doi.org/10.1175/1520-0485(1981)011<0257:otpoge>2.0.co;2)
- Marshall, J. C., & Shutts, G. J. (1981). A note on rotational and divergent eddy fluxes. *Journal of Physical Oceanography*, 11(12), 1677–1680. [https://doi.org/10.1175/1520-0485\(1981\)011\(1677:ANORAD\)2.0.CO;2](https://doi.org/10.1175/1520-0485(1981)011(1677:ANORAD)2.0.CO;2)
- Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentile, P. (2021). Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *Journal of Advances in Modeling Earth Systems*, 13(5), e2020MS002385. <https://doi.org/10.1002/2020MS002385>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Rhines, P. B., & Young, W. R. (1982). Homogenization of potential vorticity in planetary gyres. *Journal of Fluid Mechanics*, 122(-1), 347–367. <https://doi.org/10.1017/s0022112082002250>
- Sonnewald, M., & Lguensat, R. (2021). Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, 13(8), e2021MS002496. <https://doi.org/10.1029/2021MS002496>
- Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., & Balaji, V. (2021). Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, 16(7), 073008. <https://doi.org/10.1088/1748-9326/ac0eb0>
- Sonnewald, M., Reeves, K. A., & Lguensat, R. (2023). A Southern Ocean supergyre as a unifying dynamical framework identified by physics-informed machine learning. *Communications Earth & Environment*, 4(1), 153. <https://doi.org/10.1038/s43247-023-00793-7>
- Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5), 784–794. <https://doi.org/10.1029/2018EA000519>
- Student (1908). The probably error of a mean. *Biometrika*, 6, 1–25. <https://doi.org/10.2307/2331554>
- Sun, L., Haigh, M., Shevchenko, I., Berloff, P., & Kamenkovich, I. (2021). On non-uniqueness of the mesoscale eddy diffusivity. *Journal of Fluid Mechanics*, 920, A32. <https://doi.org/10.1017/jfm.2021.472>
- Sun, Y. Q., Hassanzadeh, P., Alexander, M. J., & Kruse, C. G. (2023). Quantifying 3D gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations. *Journal of Advances in Modeling Earth Systems*, 15(5), e2022MS003585. <https://doi.org/10.1029/2022MS003585>
- Thiffeault, J.-L. (2012). Using multiscale norms to quantify mixing and transport. *Nonlinearity*, 84(2), R1–R44. <https://doi.org/10.1088/0951-7715/25/2/R1>
- Thomas, S. D. A., Jones, D. C., Faul, A., Mackie, E., & Pauthenet, E. (2021). Defining Southern Ocean fronts using unsupervised classification. *Ocean Science*, 17(6), 1545–1562. <https://doi.org/10.5194/os-17-1545-2021>
- Vallis, G. K. (2006). *Atmospheric and oceanic fluid dynamics*. Cambridge University Press.
- Villani, C. (2008). *Optimal transport: Old and new*. Springer.
- Waterman, S., & Hoskins, B. J. (2013). Eddy shape, orientation, propagation, and mean flow feedback in western boundary current jets. *Journal of Physical Oceanography*, 43(8), 1666–1690. <https://doi.org/10.1175/JPO-D-12-0152.1>
- Waterman, S., & Jayne, S. R. (2011). Eddy-mean flow interactions in the along-stream development of western boundary current jet: An idealized model study. *Journal of Physical Oceanography*, 41(4), 682–707. <https://doi.org/10.1175/2010JPO4477.1>
- Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40, 974–994.
- Yan, F. E., & Mak, J. (2023). Data collection for machine learning using eddy force function data [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.8072817>

- Young, W. R. (2012). An exact thickness-weighted average formulation of the Boussinesq equations. *Journal of Physical Oceanography*, 42(5), 692–707. <https://doi.org/10.1175/JPO-D-11-0102.1>
- 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>
- Yuval, J., O’Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, 48(6), e2020GL091363. <https://doi.org/10.1029/2020GL091363>
- Zanna, L., & Bolton, T. (2021). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17), e2020GL088376. <https://doi.org/10.1029/2020GL088376>
- Zhang, S., & Lin, G. (2018). Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217), 20180305. <https://doi.org/10.1098/rspa.2018.0305>