Reconstruction of Surface Kinematics from Sea Surface Height Using Neural Networks

Qiyu Xiao[1], Dhruv Balwada[2], C. Spencer Jones[3], Mario Herrero-González[4], K. Shafer Smith[1], Ryan Abernathey[2]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[3]Texas A&M University, College Station, TX, USA
[4]École Nationale Supérieure de Techniques Avancées, Brittany, France

Key Points:

- Neural networks reasonably reconstruct surface vorticity, strain and divergence, from sea surface height.
- Neural networks naturally filter wave divergence, leaving only the desired divergence associated with fronts.
- Transfer learning shows promise when task-specific data is limited but data from reasonably close simulations is available.

Corresponding author: K. S. Smith, `kss3@nyu.edu`

Abstract

The Surface Water and Ocean Topography (SWOT) satellite is expected to observe the sea surface height (SSH) down to scales of $\sim 10-15$ kilometers. While SWOT will reveal submesoscale SSH patterns that have never before been observed on global scales, how to extract the corresponding velocity fields and underlying dynamics from this data presents a new challenge. At these soon-to-be-observed scales, geostrophic balance is not sufficiently accurate, and the SSH will contain strong signals from inertial gravity waves — two problems that make estimating surface velocities non-trivial. Here we show that a data-driven approach can be used to estimate the surface flow, particularly the kinematic signatures of smaller scales flows, from SSH observations, and that it performs significantly better than directly using the geostrophic relationship. We use a Convolution Neural Network (CNN) trained on submesoscale-permitting high-resolution simulations to test the possibility of reconstructing surface vorticity, strain, and divergence from snapshots of SSH. By evaluating success using pointwise accuracy and vorticity-strain joint distributions, we show that the CNN works well when inertial gravity wave amplitudes are weak. When the wave amplitudes are strong, the model may produce distorted results; however, an appropriate choice of loss function can help filter waves from the divergence field, making divergence a surprisingly reliable field to reconstruct in this case. We also show that when applying the CNN model to realistic simulations, pretraining a CNN model with simpler simulation data improves the performance and convergence, indicating a possible path forward for estimating real flow statistics with limited observations.

Plain Language Summary

Satellite measurements of SSH have for the past few decades provided weekly global estimates of upper ocean currents at scales larger than approximately 100 km. The new Surface Water and Ocean Topography satellite promises to improve the resolution of these SSH observations. However, these new observations will introduce a new challenge, since a simple physics-based diagnostic relationship does not exist between the SSH and upper ocean currents for the finer scales ($O(10)$ km) that will now be visible. Here we show that a neural network can be used to estimate the surface flow from SSH observations. In particular, our trained neural networks are able to use SSH to predict the surface kinematic variables: vorticity, strain, and divergence, which are particularly sensitive to the smaller scale flows. We also find that appropriate choice of the loss function can help filter unwanted waves signals from the divergence. Finally, we show that when applying the neural network to realistic simulations, pretraining a model with simpler simulation data improves the performance and convergence, indicating a possible path forward for estimating real flow statistics with limited observations.

1 Introduction

Since the mid-1990s oceanography has been revolutionized by the use of satellite nadir altimetry to provide global observations of sea surface height (SSH) (Munk, 2002). Products such as AVISO (Ducet et al., 2000) interpolate this one-dimensional track data to gridded form, with an effective lateral spatial resolution of order 100 km and a temporal resolution of a few weeks. At these scales, non-equatorial motions are accurately described by geostrophic balance, allowing for regular global estimates of upper ocean currents, from the basin scale down to larger mesoscale eddies and meanders, without the need for an assimilating model. The recently-launched Surface Water and Ocean Topography (SWOT) satellite is expected to significantly improve the effective spatial resolution to approximately 15 km (Fu et al., 2012; Chelton et al., 2019) through the use of radar interferometry to provide two-dimensional swaths of SSH measurements. The smaller scales that will be observed will likely include at least the larger end of the submesoscale regime, where geostrophy is not a good approximation, obviating its use as a diagnostic relationship for estimating currents at the new scales to be resolved by SWOT.

The nongeostrophic nature of these "near-submesoscale" flows is due to the impact on SSH at these scales of both ageostrophic features, like fronts, and inertia-gravity waves (IGWs), including internal tides. The waves present an exceptionally vexing challenge, as SWOT's 21-day repeat cycle during its main operational phase will prevent the use of averaging over inertial times to remove IGW signals. Yet, despite that IGWs comprise a significant fraction of vertical kinetic energy, they do not contribute much to tracer transport (e.g. Balwada et al., 2018; Uchida et al., 2019). By contrast, the remaining nongeostrophic near-submesoscale motions contribute significantly to the vertical transport of tracers between the ocean's surface and interior, as seen in both observations (Omand et al., 2015; Siegelman et al., 2020; Balwada et al., 2016) and modeling studies (Balwada et al., 2021; Bachman & Klocker, 2020).

Estimating this near-submesoscale transport-active velocity field is a major challenge for the interpretation and use of SWOT data. To do so one must solve two difficult problems. First, one must find a method to filter IGW signals from the data, and since the repeat cycle period is an order of magnitude longer than the inertial time of roughly one day, the method must work on individual snapshots of SSH. This unfortunately obviates the use of methods such as Eulerian spectral filtering (Torres et al., 2018, 2022) and Lagrangian filtering (Jones et al., 2022), since each requires high temporal resolution. Second, one needs a model through which to infer the nongeostrophic flow from the filtered SSH signal. While a number of papers have demonstrated success in recovering ageostrophic flows from submesoscale-permitting numerical simulations using the eSQG analytical model (e.g. J. Wang et al., 2013; Qiu et al., 2016, 2020, and others), the method still requires data to first be low-pass filtered to remove IGW signals.

The present paper seeks to sidestep these issues, forgoing a full reconstruction of the velocity field in favor of an approach that reconstructs dynamically-relevant flow statistics. Balwada et al. (2021) found that the joint probability densities (JPDFs) of surface vorticity, strain magnitude (referred to henceforth simply as 'strain'), and divergence are highly informative; these are given by

$$\zeta = v_x - u_y, \quad \sigma = \sqrt{(u_x - v_y)^2 + (v_x + u_y)^2}, \quad \text{and} \quad \delta = u_x + v_y, \tag{1}$$

where u and v are the zonal and meridional components of the surface velocity. As also noted by Shcherbina et al. (2013), the shapes and properties of these JPDFs are a statistical way to characterize the presence, magnitude and spatial scale of front-like flow structures, which are associated with sub-surface vertical transport. JPDFs can easily be calculated from individual snapshots of the surface velocity field to infer the magnitude and lateral scales of convergent frontal flows. We show here that IGWs have a distinct signature on these JPDFs, and that it may be possible to remove the wave signal, even without temporal data.

We wish to estimate the JPDFs from the sea surface height directly, and we choose a machine learning model for this task. By training the machine learning model on output from two different numerical simulations, we show that a neural network can be used to learn the surface vorticity, strain and divergence statistics directly from raw SSH. Moreover, due to a surprising kinematical fact about IGWs discussed in section 5, the method is especially useful for reconstructing the wave-filtered divergence field.

Specifically, we train a convolution neural network (CNN) to estimate the surface kinematics directly from simulated SSH data provided by two submesoscale-permitting general circulation models: the global LLC4320 simulation (Rocha et al., 2016) and a Southern-Ocean-like channel model (Balwada et al., 2018). The former, forced by 6-hourly winds and 16 tide modes, has a well-developed realistic wave field, providing a difficult but important challenge for the method. In addition, for some questions, we also consider a synthetic wave model that approximates the SSH due to a linear superposition of intertia gravity waves.

The paper is organized as follows. In section 2 we discuss the channel model and LLC4320 simulations, and the fields from each used as datasets in this study. Section 3 introduces the neural network architecture used for the reconstruction problem. Section 4 introduces the use and significance of joint distributions of surface vorticity, strain and divergence, as a tool for revealing flow structure and tracer transport, and demonstrate their reconstruction from the neural network model. In section 5 we show that when internal waves are present in the surface fields, the neural network is unable to reconstruct the wave-divergence field. This surprising fact is discussed in detail, and speculative explanations are provided. Section 6 investigates how well neural networks trained on one model can be used to predict the surface kinematic fields for another. Finally, caveats, additional points, and implications, along with a concluding summary, are given in section 7.

2 Simulation data and their statistics

To train our machine learning models, we use output from two submesoscale-permitting general circulation model simulations: the idealized channel model used in Balwada et al. (2018), and a subset of the LLC4320 simulation (Rocha et al., 2016) located near the Agulhas in the Southern Ocean. The former has minimal wave activity, while the latter has a well-developed wave field, driven by high-frequency winds and tidal forcing. For a part of the investigation, we also use output from a synthetic wave model.

The key metrics through which we analyze the models and their reconstructed statistics are the joint probability density functions (JPDF) of surface vorticity, strain and divergence. The JPDFs of these kinematical quantities allow one to identify flow signatures of submesoscale vortices and fronts, as well as their lateral scales (Balwada et al., 2021). In addition, we show below that internal waves have a distinct signature, allowing them to be identified clearly in the JPDFs, even from single snapshots of the flow.

2.1 Submesoscale-permitting channel simulation data

The channel model output is taken from a submesoscale-permitting MITgcm simulation, intended as an idealized analogue of the Southern Ocean, with a horizontal grid-spacing of 1 km and an internal deformation radius of around 40 km, set in a 2000 km $\times$ 2000 km domain with a topographic ridge in the center (see Balwada et al., 2018, for details). It is forced by time-independent surface wind and surface temperature relaxation; consequently this simulation produces a strong eddy field, and a relatively weak field of inertia gravity waves. Figure 1 shows snapshots of the surface SSH and vorticity fields, and denotes the parts of the domain used for training and testing the CNN.

The 1 km resolution simulation was the highest-resolution case in a set that included 5 km and 20 km resolution simulations as well. As the lateral resolution increased, the vorticity-strain JPDFs of the surface flow share the same qualitative shape, but the ranges of vorticity and strain increase, and the JPDF becomes increasingly cyclonically skewed, with a clustering of points just above the line with slope 1 (Figure 2). The latter is indicative of convergent fronts, which have cyclonic vorticity, with $|\zeta| \approx \sigma$ — this is especially apparent in the 1 km simulation (lower-left JPDF in Figure 2). The ± 1 slope lines moreover serve to distinguish between strain-dominated and cyclone-dominated points. The probability contours also serve as a proxy for spatial scale — lower probability points towards the high vorticity and strain parts of the JPDF tend to be smaller in scale, while points near the origin tend to represent the largest features in the flow.

Though not crucial to the present story, we note that Balwada et al. (2021) also demonstrated that the kinematic JPDFs of the surface flow reveal information about vertical transport. When conditioned on surface vorticity and strain, it was found that large negative values of the *sub-surface divergence* (i.e. convergent regions) are strongly correlated with the frontal regions of the vorticity-strain JPDF noted above. Moreover, vertical transport by submesoscale fronts was found to increase by an order of magnitude as resolution was increased, and to extend below the mixed layer (see section 2.c of Balwada et al. (2021) for details). Because of this relationship, surface vorticity-strain JPDFs inferred from SSH may provide a means to estimate submesoscale transport between the ocean surface and interior directly from SWOT.

To investigate the non-geostrophic nature of the submesoscale features in the high-resolution flows, we compare JPDFs of vorticity and strain computed from geostrophic estimates of the velocities for the same two simulations (right-most panels in bottom two rows of Figure 2). In the 5 km resolution simulation, where submesoscales are barely permitted, the geostrophic result looks qualitatively similar to the true JPDF, but underestimates the extreme values and captures less of the cyclone-anticyclone asymmetry. For the submesoscale-rich 1 km simulation, the geostrophic estimate not only fails to capture the asymmetry, it also *overestimates* anticyclonic strain and vorticity, and differs more qualitatively from the true JPDF, appearing somewhat diffused. This is a reflection of the highly inaccurate finer-scale structure that emerges in from taking derivatives of the raw SSH field used in the geostrophic estimate. It also suggests that dynamics have become much more complicated at 1 km resolution, with non-geostrophic features like strong, fast fronts, submesoscale cyclones, and some wave activity more strongly affecting the SSH.

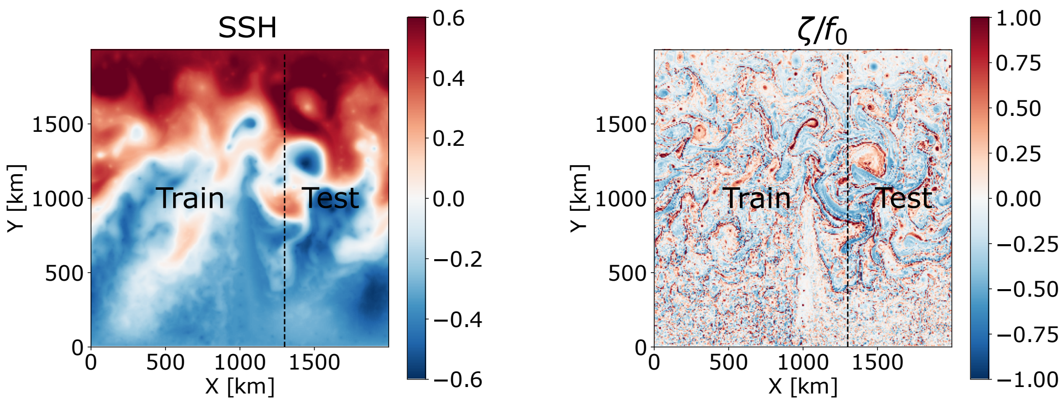

Figure 1. Snapshots of SSH (left) and normalized vorticity ζ/f (right) from a snapshot of the channel simulation of Balwada et al. (2018). The training and testing regions are marked.

2.2 The Agulhas region of the LLC4320 simulation

The second set of model output is taken from the high-resolution global LLC4320 simulation. This is a latitude–longitude–polar cap MITgcm (Marshall et al., 1997) simulation forced by surface fluxes from the European Centre for Medium-range Weather Forecasting (ECMWF) atmospheric operational model analysis for years 2011-2012. The simulation has a nominal lateral grid resolution of 1/48°, and is forced by 6-hourly winds and the 16 most significant tidal components (Rocha et al., 2016). As a result, in addition to resolving mesoscale and near-submesoscale currents, the model also exhibits strong internal tides and IGW signals that are not present in the channel simulation.

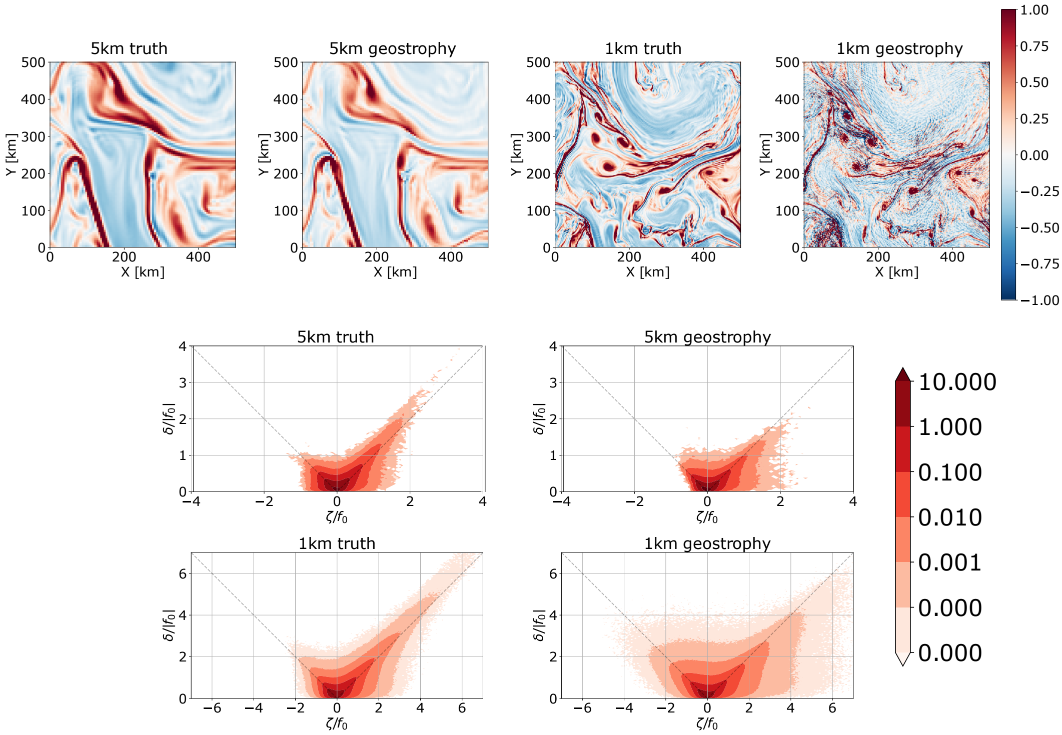

Figure 2. Top row: Normalized vorticity, ζ/f, from a 500 km square subregion of the 5 km and 1 km simulations analyzed in Balwada et al. (2021) and Balwada et al. (2018), computed directly from their velocity fields, as well as from geostrophic estimates of velocity (see panel titles for identification). Middle row: vorticity-strain JPDFs from the 5 km simulations, computed from velocity field (left) and from geostrophic estimate (right). Bottom row: same as the middle row, but for the 1 km simulation.

We focus on three local regions in the Agulhas region, with latitudes between 35° and 47° south and longitudes 4−21° west, 12−28° east, 28−45° east, respectively, as marked in Figure 3. Out of the total simulation time spanning from September 2011 to October 2012, we focus on data from March 2012, when the mixed layers in the three regions are at their deepest, and September 2012, when the mixed layers are shallowest; these two months are thus termed 'summer' and 'winter', respectively.

Many of the same qualitative patterns seen in the surface vorticity-strain JPDFs for the channel simulation are found in observational data (Shcherbina et al., 2013; Berta et al., 2020) as well as in the winter-time data for the three target regions, and summer-time data for region 2, of the LLC4320 simulation (top row and middle column of Figure 4; see also JPDFs computed by Rocha et al. (2016)). However, new features not seen in the channel simulation arise in regions 1 and 3 of the summer LLC4320 data (bottom two rows of Figure 4). These new features, characterized by clusters of points with high strain, high divergence and low vorticity, are consistent, we argue below, with the stronger surface IGW activity expected in the presence of shallow summertime mixed layers.

2.3 Wave signatures in surface kinematic JPDFs

These JPDF signatures for IGWs can be most easily understood by computing kinematic fields for a single plane inertia-gravity wave in constant stratification. Writing the pressure field for wavenumber (k, l, m) as $p = \Re\, \hat{p} \exp\left[i(kx + ly + mz - \omega t)\right]$ and us-

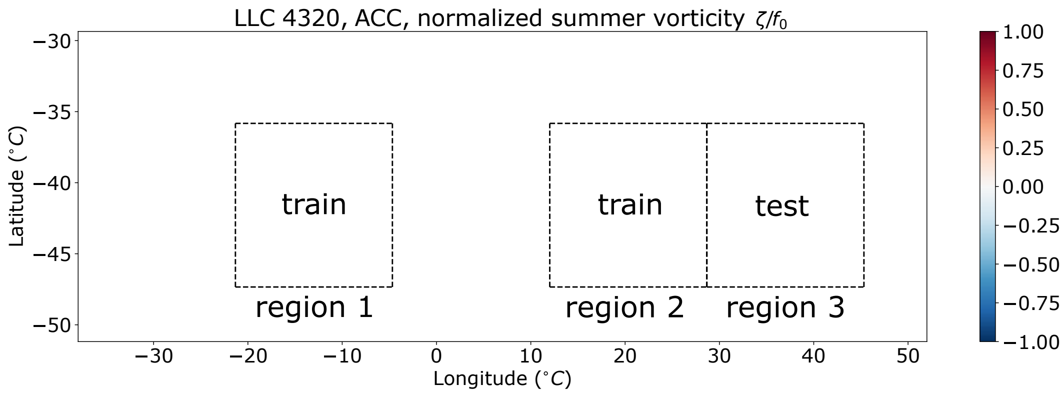

Figure 3. A snapshot of normalized summer vorticity ζ/f in the target regions of the LLC4320 simulation, with training and testing regions as marked.

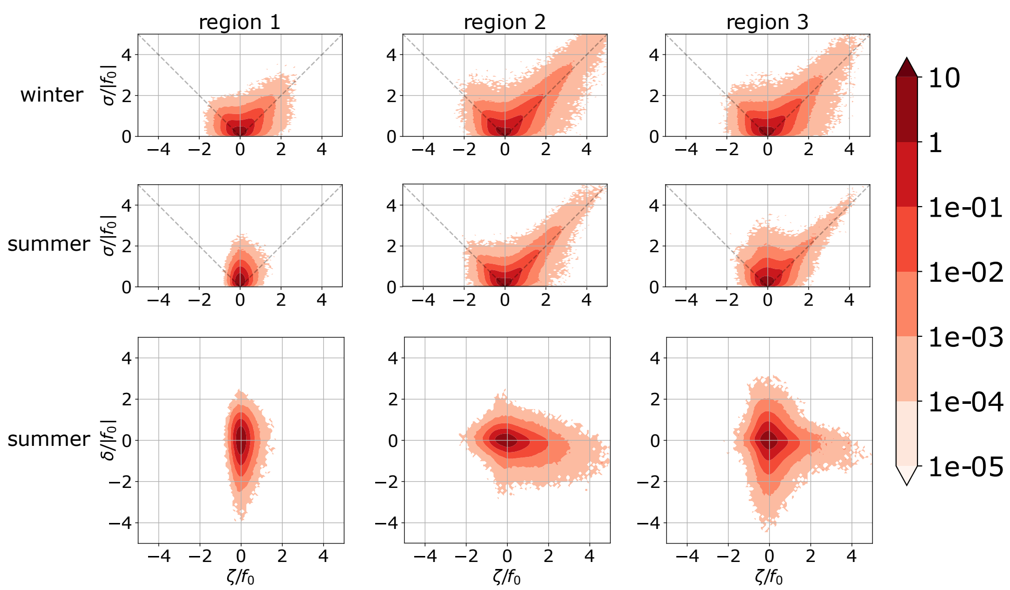

Figure 4. Winter vorticity-strain JPDFs (top row), summer vorticity-strain JPDFs (middle row) and summer vorticity-divergence JPDFs (bottom row) for the three local regions of the LLC4320 simulation marked in Figure 3.

ing the hydrostatic IGW dispersion relationship $\omega^2 = f^2 + N^2(k^2 + l^2)/m^2$, the horizontal velocity amplitudes are

$$\hat{u} = \frac{k\omega + ilf}{\omega^2 - f^2}\,\hat{p} \quad \text{and} \quad \hat{v} = \frac{l\omega - ikf}{\omega^2 - f^2}\,\hat{p},$$

where N is the buoyancy frequency, and f the Coriolis parameter. From the wave velocity, and taking $\hat{p}$ to be real, the vorticity and divergence are

$$\zeta = \frac{fm^2}{N^2}\hat{p}\cos\left(kx + ly + mz - \omega t\right) \quad \text{and} \quad \delta = -\frac{\omega m^2}{N^2}\hat{p}\sin\left(kx + ly + mz - \omega t\right) \tag{2}$$

and the strain turns out to be just

$$\sigma = \sqrt{\zeta^2 + \delta^2}. \tag{3}$$

The ratio of vorticity to divergence thus scales as $O(|\zeta/\delta|) \sim |f/\omega|$. Because ω grows large relative to f as the horizontal wavenumber increases, at smaller scales divergence increasingly dominates vorticity, and then strain is approximated by divergence instead of vorticity.

We test this simple argument by computing the JPDFs for a synthetic internal wave model (Early et al., 2021). This Matlab-based package generates linear internal waves following the Garrett-Munk spectrum (Munk, 1981) by numerically solving the linearized Boussinesq equations for a user-defined domain, with a specified background stratification and resolution. Here we use the mean stratification and resolution from the channel simulation to compute its kinematic surface fields, and vorticity-strain and divergence-vorticity JPDFs; snapshots of SSH, vorticity, and the JPDFs are shown in Figure 5. The resulting JPDFs behave as predicted, and moreover bear resemblance the summertime JPDFs for region 1 of the summer LLC4320 data (Figure 4). The JPDFs for region 3 of the summer LLC4320 data seem to indicate a superposition of submesoscale and IGW structures, especially so in the vorticity-divergence JPDF (bottom row of Figure 4), where the wave-dominated and front-dominated signatures are almost orthogonal to each other.

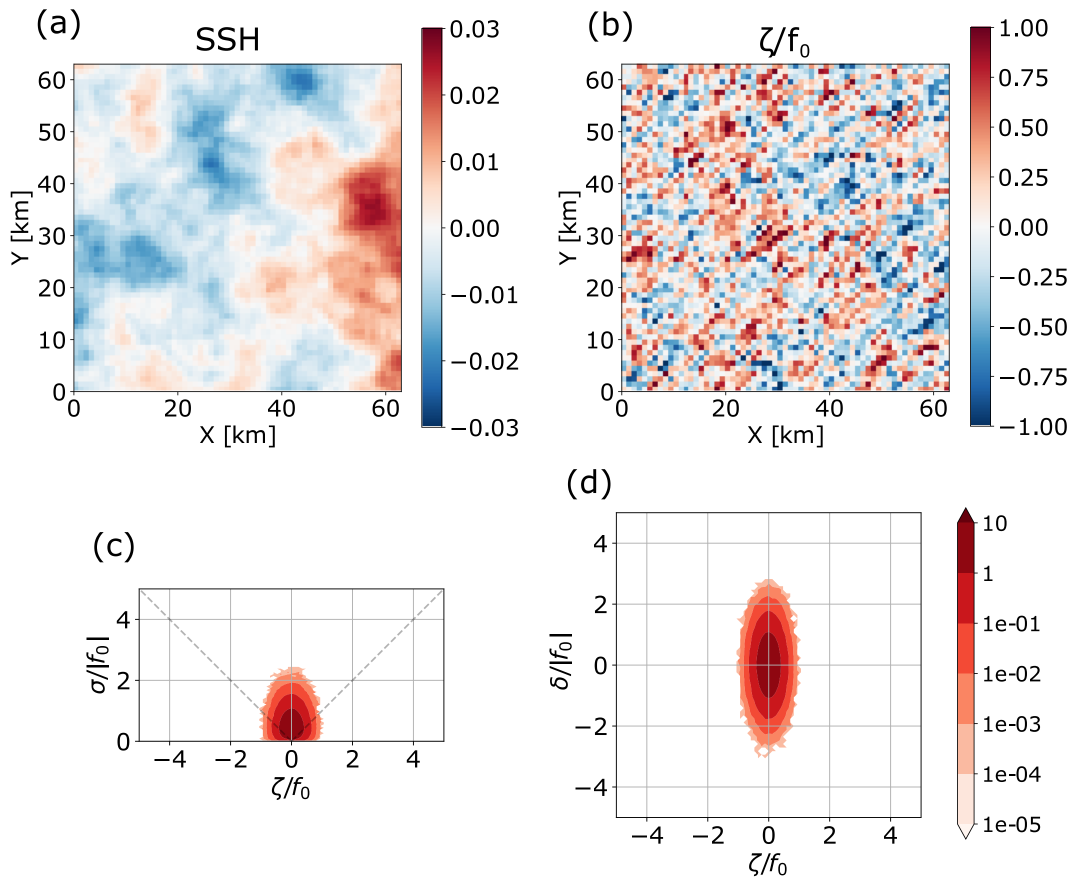

Figure 5. Snapshots of (a) SSH and (b) vorticity normalized by f (strain and divergence show similar structure, and so are not shown) from the synthetic internal wave model. Vorticity-strain (c) and vorticity-divergence (d) JPDFs from the same data.

In summary, statistics of surface vorticity, divergence and strain are robust indicators of surface flow features, and geostrophy does a poor job at reconstructing these from SSH fields at higher resolution (or smaller spatial scales). In the next section, we introduce the machine learning architecture used, and in the following sections show that this framework can be used to more accurately reconstruct these surface kinematic variables.

3 Deep Learning Model

Neural networks, among other machine learning models, have gained a lot of attention in the atmosphere-ocean science community over the past few years and have shown better performance relative to traditional approaches for many tasks (Bolton & Zanna, 2019; Manucharyan et al., 2021; Sinha & Abernathey, 2021; George et al., 2021). Briefly speaking, a neural network consists of several hidden layers that transform its input into the final output. Each hidden layer is a combination of multiple linear matrix multiplications or additions and a simple nonlinear element-wise function such as a sigmoid. The elements of these matrices are tuned during the training of the model using gradient descent. The number of operations in each layer (usually called the 'width' of a layer) and the number of layers in the whole network (usually called the 'depth' of a network) determine the capability or flexibility of a neural network.

The theoretical basis for neural networks is the Universal Approximation Theorem (Hornik et al., 1989): given an arbitrarily wide or deep network, there exists a set of matrices, such that any continuous function can be approximated by the neural network as closely as desired. However, the Universal Approximation Theorem doesn't provide a construction recipe for the target neural network. In practice, due to limitations on computing resources and the amount of data, the architecture of the neural network is no less critical than the width or depth for efficiently building a useful model.

Here we use a Convolution Neural Network (CNN) (LeCun & Bengio, 1995), which is known for its ability to capture spatial patterns in 2D physical data. When passing the data within a layer, the CNN uses a set of 'convolutional filters' (a 3×3 matrix for example) to do convolution with each local patch of the input before feeding the result to a point-wise nonlinear function to generate the output. Abstractly, this can be represented

$$Y_j^{(k)} = \gamma^{(k)} \left(\beta_j + \sum_{i=1}^{c^{(k-1)}} F_{ij} * Y_i^{(k-1)} \right) \tag{4}$$

where $Y_j^{(k)}$ is the jth channel at layer k, and $\gamma^{(k)}$ is a nonlinear function that could be composite of activations, normalizations and poolings. The parameter β_j is a scalar bias term, $c^{(k-1)}$ is the number of channels in layer $(k-1)$, and F_{ij} is a filter matrix that transform $Y_i^{(k-1)}$ to another feature space through the 2D cross-correlation '$*$'. During training, these filter matrices from each layer are believed to converge to representations in abstract feature space that are crucial for generating predictions.

In this work, we use a specific type of CNN called a 'Unet' (Ronneberger et al., 2015), the structure of which is shown schematically in Figure 6. The Unet has two parts: the 'encoder' condenses the variable resolution and expands the number of feature maps to extract information from the input, while the 'decoder' does the opposite, using the information extracted to construct the output. Unet tries to overcome the loss of information in previous CNN models by delivering input in the encoding layers not only through the feature mapping pathway but also directly to the decoding layers. Each layer has two sets of convolution filters of dimension 3×3 as well as batch normalization and Scaled Exponential Linear Units as activation functions.

Throughout this work, we train Unets on simulated SSH data, and test their ability to reconstruct surface vorticity, strain and divergence, given only SSH data under different scenarios. Though a Unet is flexible in the dimensions of its input, we chop our training data into non-overlapping sections of 64 × 64 grid points each. This is a trade-

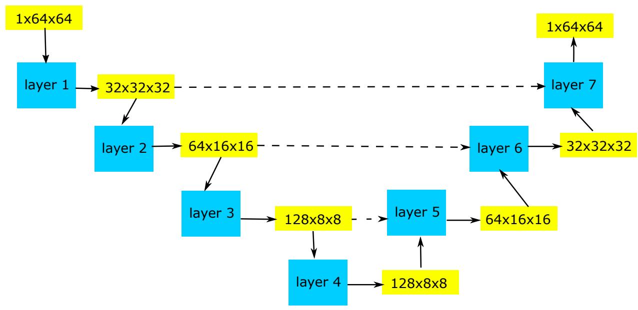

Figure 6. The structure of the Unet CNN used in this work. Blue boxes represent convolution layers, and yellow boxes represent input, intermediate and final outputs. The sets of three numbers refer to channels, height, and width. Solid lines indicate delivery of data to the next layer. Dashed lines indicate delivery to the layer not directly following.

off in the sense that while we use a smaller size of the input, we have a larger collection of samples. But at the same time, the model needs to be exposed to mesoscale features during training. We found a 64×64 box suitable for these purposes. On the other hand, when we test the performance of our model we take a slightly different approach. We still chop our target region into 64×64 local regions as input, but now these local regions overlap with each other, with a stride of 5. The reason for doing this is that when building the output using non-overlapping data, the points closer to the boundary of the input would get less information available for its reconstruction compared to points at the center, and this largely impairs the capability of the model. Samples of the training set are randomly shuffled, preventing the neural net from learning temporal information.

Note that we omit the difficulty of transforming swath data to grid data, assuming SSH is given naturally on the grid without loss of information. In theory, neural networks applied here can be extended to use swath data as input (Manucharyan et al., 2021; Fablet & Chapron, 2022).

For loss functions, we use mean squared error for most of the work and mean absolute error for models used in Figure A1. In the past few years, innovative loss functions such as adversarial loss (Ledig et al., 2017; Zhang et al., 2019) and perceptual loss (Johnson et al., 2016) have trended in the computer vision community and helped build state of art image processing models. However, the main focus of those studies is to improve model performance against the perceptual feeling of humans, and the mathematical foundation of these new techniques is not fully explored. While we believe that the application of a task-specific loss function is important to the application of a machine learning model, the discussion of that is out of the scope of this work and awaits future investigation.

Besides the configuration above, we use Adam (Kingma & Ba, 2014) as the optimizer with a learning rate 0.0001, a batch size of 32 and 100 epochs, unless specified otherwise. Additional details about can be found in the sample code provided in our Github repository.

4 Learning surface kinematics with a neural network model

4.1 Channel simulation

We train Unet CNNs with the output of the channel model SSH and velocity fields in the top grid cell (with $z = -0.5$ m) to construct surface vorticity, strain and divergence separately. We perform the training and testing on regions of the 1 km simulation (marked in Figure 1). Temporally, we use 80 days of 6-hourly snapshot data for training, and the following 10 days are used for testing. After chopping, there are about 40,000 samples of 64×64 tiles for training. In Figure 7 we show the true vorticity and strain and the reconstructed result in the downstream testing region, and also compare to the reconstruction using the geostrophic balance. The Unet has successfully captured most features on both large and small scales. In comparison, the vorticity and strain computed from geostrophic balance deviate much more from the truth. Visually this deviation is most severe in submesoscale vortices and filaments, though also visible in larger-scale features. This can be explained by the fact that small-scale features usually have larger Ro and under this scenario the geostrophic relation no longer dominates in the asymptotic expansion in orders of Ro, even given that this is a simulation with relatively weak waves.

The discrepancy is even more obvious in the point-wise performance of the reconstruction, measured by its prediction skill

$$\text{skill} = 1 - \left[\frac{\overline{(\text{truth} - \text{prediction})^2}}{\overline{\text{truth}^2}}\right]^{\frac{1}{2}}$$

and correlation between the true target and the reconstructed result (Table 1). The Unet reconstruction yields high correlation as well as decent prediction skill, surpassing that of the geostrophic estimation.

Table 1. Correlations and prediction skills of machine learning and geostrophic results against the 'truth' from the channel simulation.

Variable	ζ_{Unet}	σ_{Unet}	δ_{Unet}	ζ_{geo}	σ_{geo}
Correlation	0.93	0.91	0.80	0.73	0.75
Skill	0.65	0.71	0.41	0.2	0.31

Greater insight into the performance of the reconstruction methods can be gauged by considering the true, reconstructed, and geostrophic vorticity-strain JPDFs for the channel model (Figure 8, top row). Overall it can be seen that the neural network result captures the basic structure of the JPDF, especially the small scales asymmetric frontal part. By contrast, the geostrophic result shows excessive symmetry between cyclonic and anticyclonic features, and smaller extreme values, as also seen in the previous section.

The neural network is also able to capture properties of the distribution of surface divergence conditioned on the vorticity and strain (Figure 8, bottom row). Here we can see that the Unet result reproduces the separation between downwelling and upwelling regions of the JPDF, as well as the magnitude of divergence. This holds promise for estimating vertical transport from snapshots of SWOT-measured SSH.

In conclusion, we see that while the machine learning solution captures the relationship between the SSH and surface kinematic variables, while the geostrophic relation provides an unsatisfactory reconstruction for the high-resolution simulation.

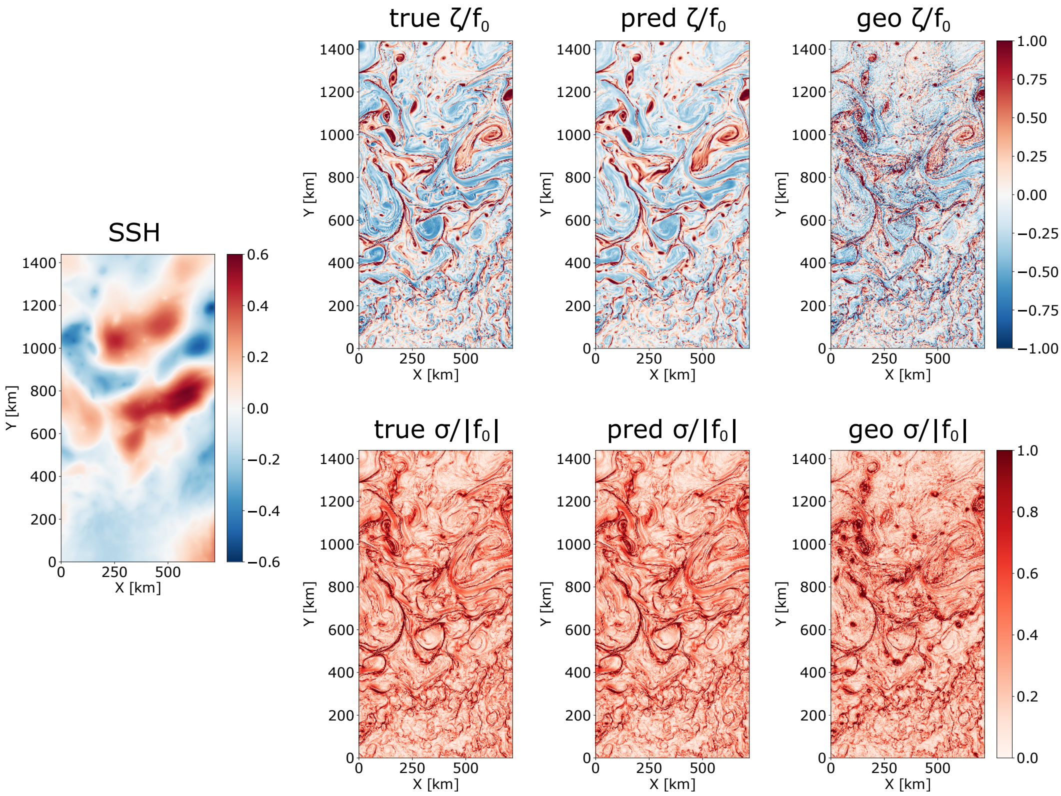

Figure 7. The SSH field in the test region of the channel simulation (left); true normalized vorticity, predicted vorticity and vorticity from geostrophic relation (top right three panels); true normalized strain, predicted strain and strain from geostrophic relation (bottom right three panels).

4.2 LLC4320 simulation

After finding success with the channel simulation, here we test the ability of a Unet neural network to reconstruct surface kinematic quantities for the more complex LLC4320 simulation. As denoted in Figure 3, we train the Unet with data from Regions 1 and 2 and test preditions in Region 3. Specifically, we use 30 days of 4-hourly snapshot data in either winter or summer for Regions 1 and 2 — giving a total of about 50,000 samples for training — and test predictions for Region 3 in the same seasons. The vorticity field in Region 3 shows a combination of wavy and turbulent sub-regions that are roughly located in the southeast and northwest parts of the spatial domain (Figure 9). While the frontal features, at both meso- and submesoscale in either season, are captured well in the northwest part of the region, the properties in the wavy sub-region in the southeast are farther from the truth.

From Table 2, we see that, compared to the channel simulation, the point-wise correlation and skill metrics have significantly dropped for the Unet reconstructions of the kinematic fields, especially for summer, when IGWs are stronger. We also experimented with using a neural network model trained with one season of the LLC4320 simulation to reconstruct vorticity in another season, and found that the result is indistinguishable from reconstruction when using a model that is trained with the same season as the test input (not shown).

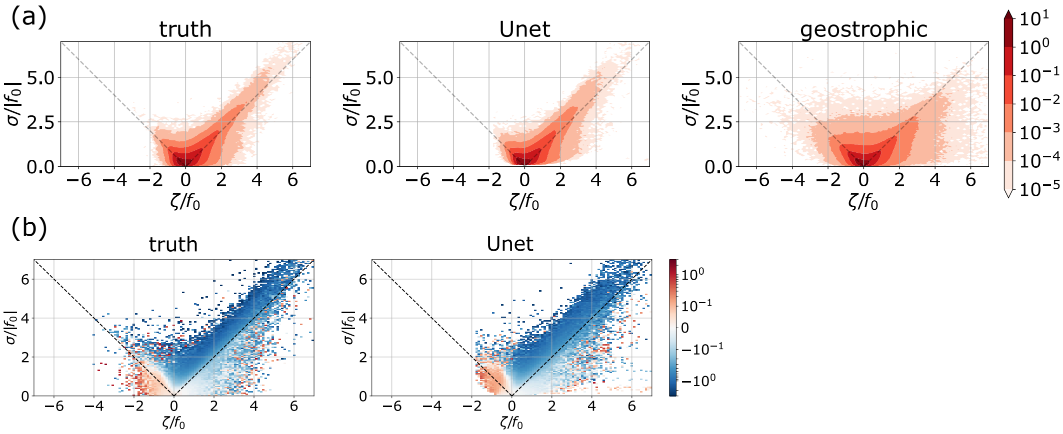

Figure 8. Vorticity-strain JPDF for the channel model truth (upper left), Unet reconstruction (upper middle), and geostrophic estimates (upper right); mean divergence conditioned on vorticity and strain for the true channel simulation data (lower left) and for the Unet reconstruction (lower right).

Table 2. Correlations and prediction skills for the kinematic fields reconstructed using the Unet model against those computed from the true LLC4320 simulation.

Variable	ζ_{winter}	σ_{winter}	δ_{winter}	ζ_{summer}	σ_{summer}	δ_{summer}
Correlation	0.9	0.81	0.5	0.84	0.63	0.5
Skill	0.57	0.67	0.15	0.46	0.55	0.15

In Figure 10 we show the vorticity-strain JPDF for Region 3 in winter and summer. Because of the extra complexity introduced by the strengthening of inertia gravity waves, in neither season could the machine learning model produce a result as good as that for the channel simulation. For winter, though suffering more from missing extreme values, the shape of the JPDF is still consistent with the truth.

The JPDF for summer is more severely distorted. The predicted joint distribution doesn't fall into either the wave-dominated or turbulence-dominated regime we have seen above. The marginal distribution of vorticity is roughly reproduced, but the distribution of strain becomes more concentrated at small values. The small-scale large vorticity values (likely from the southeast part of the Region 3 domain) are replaced by smoothed small values, most obvious in the summer (the same is true for strain, not shown). This suggests that the Unet isn't able to properly reconstruct IGW vorticity and strain. It remains a question if this is because the model wasn't able to distinguish the wave signal from the SSH, or because it couldn't find a way to transform the wave signal it sees in SSH to vorticity and strain.

The Unet's reconstruction of divergence behaves particular poorly when measured in terms of correlation and skill. This is because, relative to strain and vorticity, divergence is dominated by wave signals. Despite this dramatic drop in both metrics, and a prediction skill as low as 0.15, Figure 11 suggests that the models give a prediction that preserves fronts and filaments in different scales, while much of IGW signal is reduced. The particularly poor ability of the neural net to capture IGW signals in divergence — and the potential advantages of this weakness — are discussed in the next section.

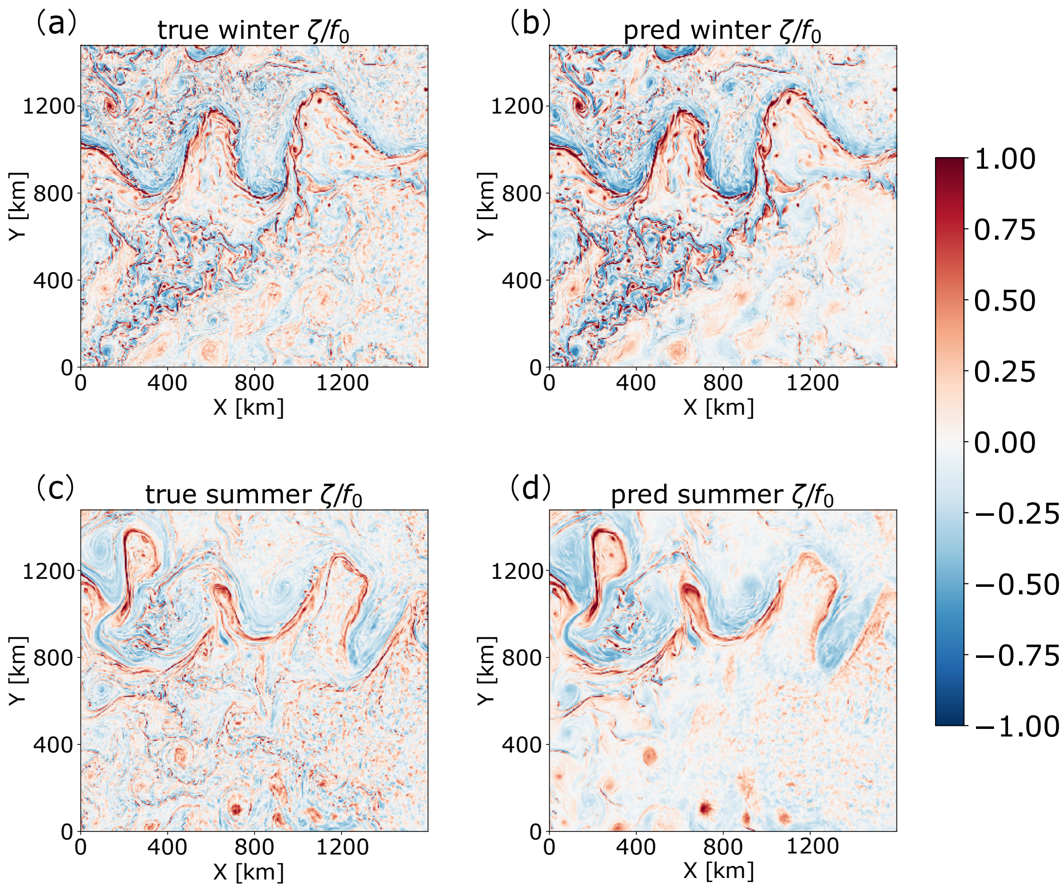

Figure 9. LLC4320 Region 3 true winter vorticity (a) and reconstructed winter vorticity (b);Region 3 true summer vorticity (c) and reconstructed summer vorticity (d).

5 Neural networks may automatically filter IGW divergence

Here we show that the divergence associated with IGW cannot be estimated using only SSH. This is because the same SSH anomaly can produce equal and opposite signed IGW surface divergence depending on the sign of the frequency, thus the relationship between the surface divergence and SSH is not one-to-one and partly random.

5.1 Expected values of wave and balanced divergence

If we assume that the flow can be separated as a linear combination of a balanced part (denoted by subscript 'bal') and a wave part (denoted by subscript 'wave'), then using a mean squared error as loss function results in a neural network that predicts,

$$\begin{aligned} f_\theta(\eta_{\mathrm{bal}} + \eta_{\mathrm{wave}}) &= E[\delta_{\mathrm{bal}} + \delta_{\mathrm{wave}} | \eta_{\mathrm{bal}} + \eta_{\mathrm{wave}}] \\ &= E[\delta_{\mathrm{bal}} | \eta_{\mathrm{bal}} + \eta_{\mathrm{wave}}] + E[\delta_{\mathrm{wave}} | \eta_{\mathrm{bal}} + \eta_{\mathrm{wave}}], \end{aligned} \tag{5}$$

where f_θ is the neural network function and E denotes the expectation of a distribution.

Considering the plane-wave polarization relations discussed in section 2.3, we see that the surface pressure p (and thus η_{wave} through hydrostatic balance $\eta_{\mathrm{wave}} = p_{\mathrm{wave}}|_{z=0}/\rho_0 g$) and the surface divergence, are related through a ratio $\omega m^2/N^2$. The frequency ω can take both positive and negative values, which impacts the direction of wave propagation. However, if no temporal information is available or incorporated into the loss function,

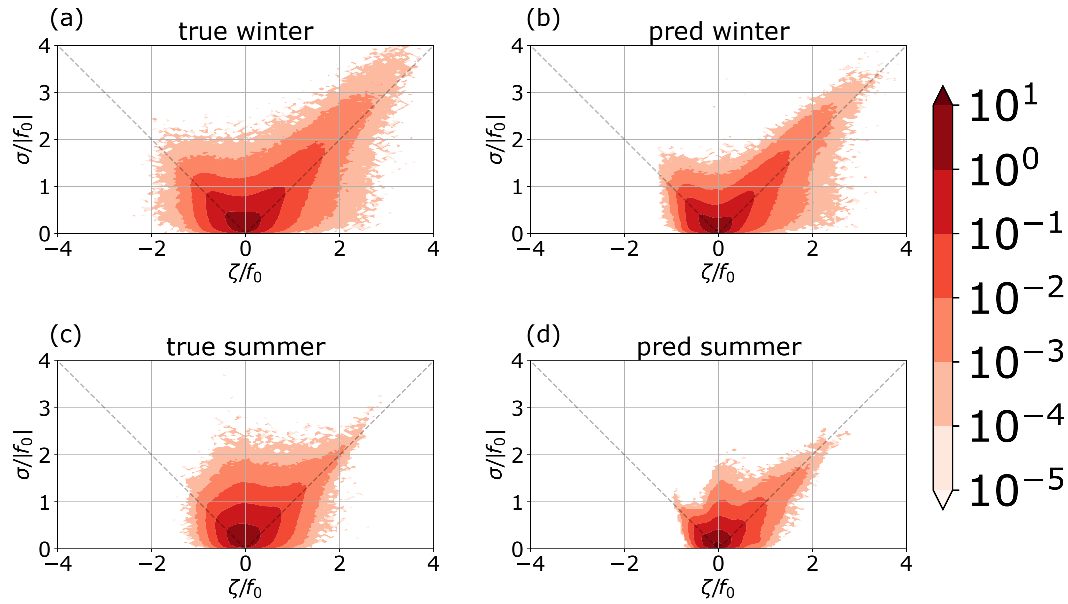

Figure 10. LLC4320 Region 3 true winter vorticity-strain JPDF (a) and Unet predicted winter vorticity-strain JPDF (b); Region 3 true summer vorticity-strain JPDF (c) and Unet predicted summer vorticity-strain JPDF (d).

the conditional distribution of surface wave divergence is symmetric about zero and

$$E[\delta_{\text{wave}}|\eta_{\text{wave}}] = 0. \tag{6}$$

This suggests that given a divergence field with both wave and balanced parts, a neural network will automatically filter out the wave divergence.

When balanced flow $\mathbf{u}_{\text{bal}}$ is taken into consideration, Doppler shifting can happen. Assuming $\mathbf{u}_{\text{bal}}$ is relatively slowly varying in both space and time, then the intrinsic frequency ω is replaced by $\Omega = \omega + \mathbf{u}_{\text{bal}} \cdot \mathbf{k}$ in the phase of wave divergence (2). However, the change in frequency due to Doppler shift doesn't affect the intrinsic frequency ω in the factor $\frac{\omega m^2}{N^2}$. Thus following the same argument, if one is able to separate the sea surface height generated by waves from that due to the balanced flow, we find

$$E[\delta_{\text{wave}}|\eta_{\text{wave}}, \eta_{\text{bal}}] = 0. \tag{7}$$

[Here the comma between η_{wave} and η_{bal} means that we observe each of them at the same time but separately.]

Through the law of total expectation, when observing the superposition of sea surface height from both IGW and balanced parts instead of these two separately, we still have

$$\begin{aligned} E[\delta_{\text{wave}}|\eta_{\text{wave}} + \eta_{\text{bal}}] &= E[E[\delta_{\text{wave}}|\eta_{\text{wave}}, \eta_{\text{bal}}]|\eta_{\text{bal}} + \eta_{\text{wave}}]] \\ &= E[0|\eta_{\text{bal}} + \eta_{\text{wave}}] = 0, \end{aligned} \tag{8}$$

and thus

$$\begin{aligned} f_\theta(\eta_{\text{bal}} + \eta_{\text{wave}}) &= E[\delta_{\text{bal}} + \delta_{\text{wave}}|\eta_{\text{bal}} + \eta_{wave}] \\ &= E[\delta_{\text{bal}}|\eta_{\text{bal}} + \eta_{\text{wave}}]. \end{aligned} \tag{9}$$

The model converges to only output the divergence from the balanced part.

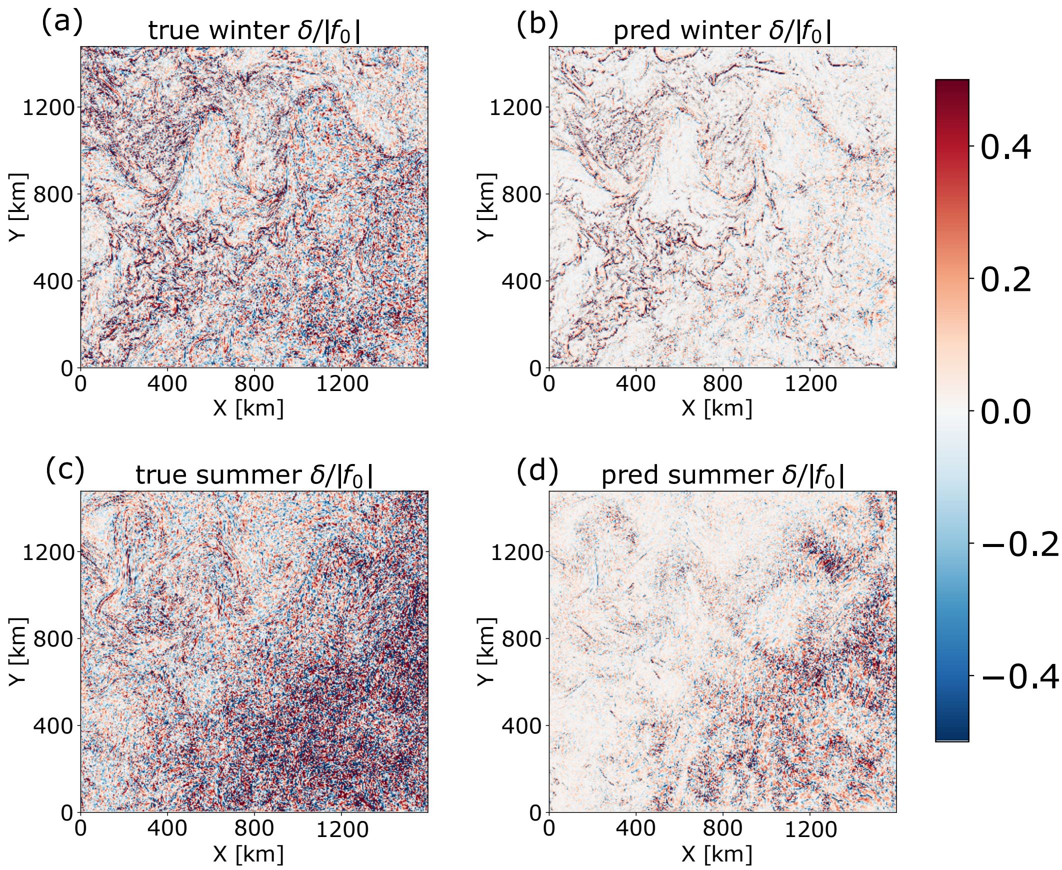

Figure 11. LLC4320 Region 3 true winter divergence (a) and Unet reconstructed winter divergence (b); Region 3 true summer divergence (c) and Unet reconstructed summer divergence (d).

This argument is inspired by Lehtinen et al. (2018), where the authors creatively use only noisy images as both inputs and targets to train an image denoiser. The idea backing this method is that as long as the 'corrupted' data has the same conditional expectation as the 'clean' data, the model will converge to the ideal set of configurations even just fed with corrupted data, at the cost of needing more training data and more iterations of training before convergence.

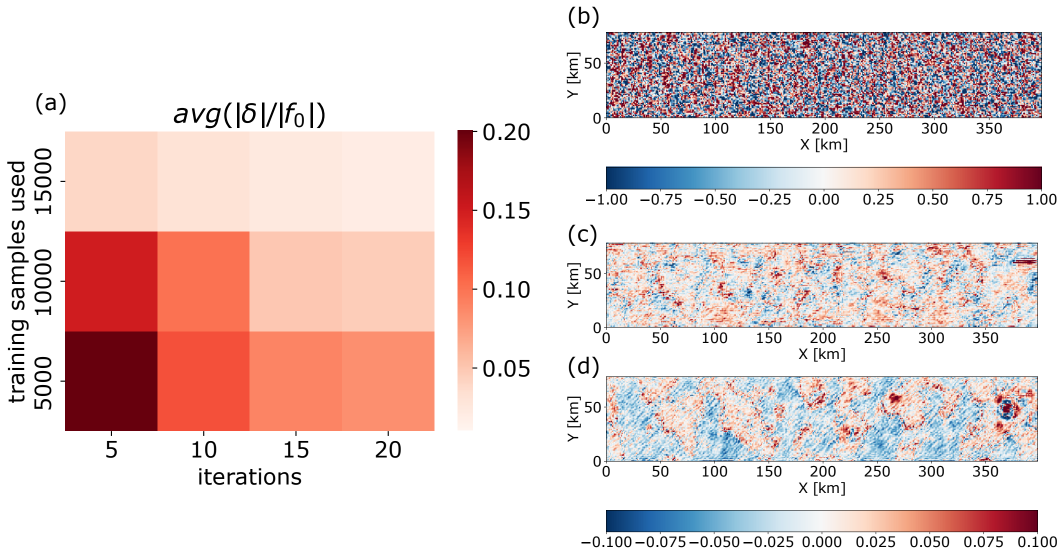

Figure 12. (a) Mean of absolute values of wave divergence prediction using different amount of training iterations and training samples for the synthetic wave model. (b) Sample of target wave divergence. (c) Unet predicted wave divergence after 20 iterations using 5,000 training samples. (d) Same as (c) except using 15,000 training samples.

5.2 Testing divergence reconstruction with synthetic wave data

To empirically justify (6), we trained a neural network using the synthetic wave data to generate around 18,000 training samples, and then predict wave divergence from wave SSH (Figure 12). We can see, as expected, that as more training data and more training iterations are provided, the model converges towards a field of zeros (Figure 12a). We also see from the Unet predictions (Figure 12c,d) that no clear pattern is learned. [Note that filtering lower wavelength waves takes longer as the number of their relative samples per snapshot is lower].

Unfortunately, this is not a property broadly shared by other kinematic quantities like vorticity and strain. For example, based on the polarization relationships (see section 2.3), the wave pressure and the wave vorticity are related by a factor of $-fm^2/N^2$ and a phase of $\pi/2$. Thus for a single-plane wave, the wave SSH can uniquely determine the wave vorticity. When multiple waves exist, the expectation of wave vorticity conditioned on wave sea surface height depends on the distribution of vertical wavenumber m from the training data and thus the GM spectra (Munk, 1981; Levine, 2002).

When trained with more data and more iterations, the IGW vorticity converges to a limit that is neither zero nor the true target value (Figure 13). When waves are weak, this will add a small distortion to the reconstruction of the balanced vorticity. For a strong wave scenario, we may need to develop more advanced loss functions to either better reconstruct the wave vorticity or remove it more precisely.

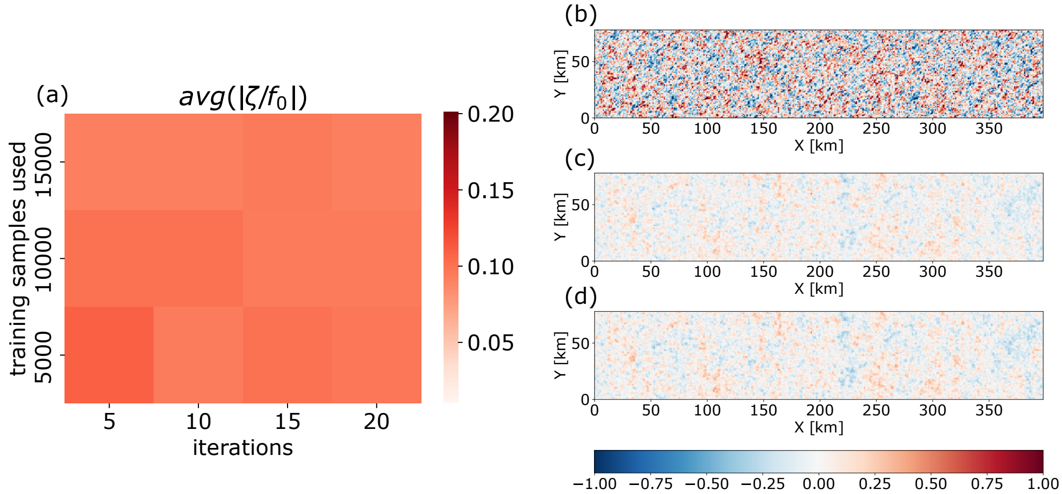

Figure 13. Same as Figure 12 but for synthetic wave model vorticity.

5.3 Testing divergence reconstruction using Lagrangian filtered velocities

Filtering inertia-gravity waves from the simulated flow is a key aim of this paper. Implicit in that goal is the idea of a well-defined balanced flow that can be cleaved away from the wave part. In fact, this is a notoriously difficult and unsolved problem, though progress has been made on practical methods to do so. Here we use the Lagrangian-filtered flow computed in Jones et al. (2022) as an approximation of the balanced flow, and train the CNN to extract it from the raw LLC data. The Lagrangian filtered data available to us includes daily snapshots within the region bounded by longitudes 15° west − 29° east and latitudes 26−52° south, spanning from September to October 2011, which provides about 35,000 samples for training in total. Unfortunately this excludes the summer month that exhibits the strongest wave activity.

We train two neural network models using raw LLC4320 SSH fields to predict either the raw divergence or the Lagrangian filtered divergence. The divergence in the former should converge to $E[\delta_{\text{bal}} + \delta_{\text{wave}}|\eta_{\text{bal}} + \eta_{\text{wave}}]$ and the latter should converge to $E[\delta_{\text{bal}}|\eta_{\text{bal}} + \eta_{\text{wave}}]$, but the two should be similar based on the discussion above.

Figure 14 suggests that at least visually the predictions from the two models are quite similar. It should be remarked that the Lagrangian filtering does a good job at removing IGWs, as can be seen by comparing true Lagrangian filtered divergence to true raw divergence, but still preserves many small-scale features. In contrast, we see that the predictions from both the neural networks result in divergence fields that have diminished smaller-scale structure than even the Lagrangian filtered divergence field. This aspect will be investigated more in future studies, but might indicate that smaller scale features have less of a unique connection to the SSH field.

It is worth mentioning that this conditional expectation that the model converges to doesn't really rely on the strength of the wave part, but rather on the interaction between the wave and balanced parts. This could be seen in the convergence of the model trained on the raw data towards the model trained on Lagrangian data (Figure 14). However, the amount of training data needed for the model to converge **is** dependent on the strength of wave-like motions in the chosen region. As the signal-to-noise ratio gets smaller, we require more data to recover the signal.

To conclude, if we only want to extract information about the balanced flow from a SSH input that contains both balanced and wave signatures, using a neural network and reconstructing the divergence may be a reliable option. This is because the neural network using conventional loss functions will converge towards giving wave-free output due to the isotropic-in-time behavior of the wave divergence.

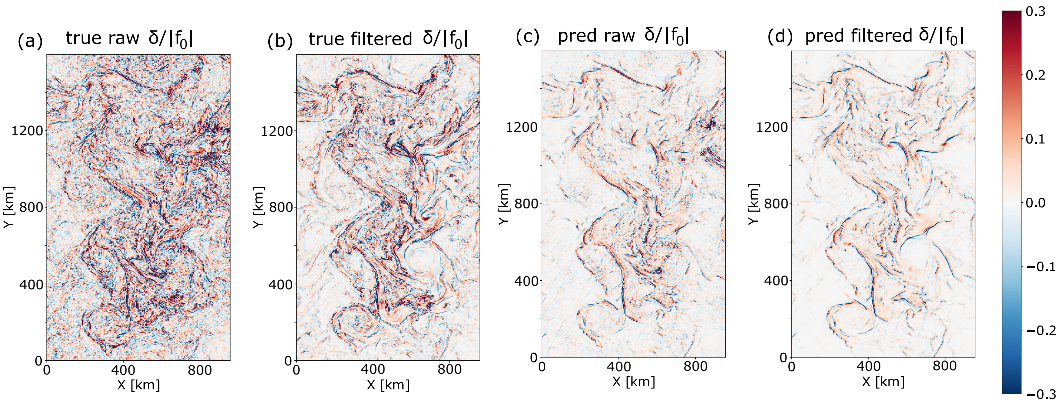

Figure 14. (a) True raw divergence from the region of the LLC4320 simulation analyzed by Jones et al. (2022), and (b) the Lagrangian filtered divergence from the same region. (c) Unet predicted divergence trained on true divergence, and (d) Unet predicted divergence trained on the Lagrangian filtered divergence.

6 Learning from limited data: Transfer Learning

While training with simulation data, we can in theory continuously boost the performance by adding more complexity to the machine learning model and supplementing extra simulation data during training, if computing resources are not a limitation. However when working with real world observations, reliable observational data for training is always scarce and likely never enough to train a model from scratch. One paradigm to overcome this challenge is to train a model with some closely linked dataset for which large-amount of data is available, and then fine-tune the model with task-specific data. This procedure is referred to as "transfer learning," and the expectation is that the 'knowledge' learned previously could be transferred and thus compensate for the missing task-specific data. The intuition behind this is that universal representations could be learned even when a model is trained with non-task-related data. The first few layers of the model often learn to recognize lines and shapes in the input regardless of the task, and these features can be reused when we try to apply the model to more specific datasets. Though the theoretical understanding of transfer learning is still a topic of ongoing research, the adoption of this methodology has led to prominent results in practice (Y. Wang et al., 2020).

With SWOT-derived SSH data, we won't have simultaneous high-resolution in-situ observations of the corresponding velocity field, and thus no "truth" with which to train a neural network model. In analogy to this problem, in this section we test whether transfer learning from the channel model could help a neural network reconstruct the surface kinematic variables from SWOT-like SSH data from the LLC4320 simulation.

Specifically, here we pretrain a Unet with channel model simulation data using 40,000 samples. During the training stage using the LLC4320 simulation data (which, again, consists of 30 days of 4-hourly snapshot data from Regions 1 and 2, for either summer or winter), all the weights from the pretrained model are allowed to be tuned. For comparison, we also train a second model with randomly initialized weights using the LLC4320 simulation dataset, with the same randomly chosen subsets from the LLC4320 winter dataset. We denote these neural network models as either 'CS' for channel simulation pretrained, or 'scratch' for the model with randomly initialized weights, appended by the number of LLC4320 winter samples used to tune or train the model. For example, 'scratch-20000' means the model is initialized from scratch (randomly initialized) and trained with 20,000 samples from the LLC4320 dataset.

First, we test the performance of these models when the number of training samples is cut to 10,000 or 20,000 from the total 53,000 samples used in earlier sections. Figure 15 shows a subregion of LLC4320 Region 3 winter vorticity, along with reconstructed vorticity fields from the randomly initialized model (scratch-10000), and from the channel simulation pretrained model (CS-10000). Both models were trained for the same number of iterations. We can see that though the two show similar structure, the latter performs better in recovering the details and amplitude of the structures. A more comprehensive comparison of prediction skills from models with different setups is summarized in Figure 16 (correlations share the same trend). We can see that when less data is available, the model pretrained with channel simulation data can offer both better performance and faster convergence. This suggests that the model can reuse some of the features learned from channel simulation data to help reconstruct LLC simulation surface dynamics.

Note also that while the channel simulation pretrained model consistently performs better than the randomly initialized model, the gap is narrowing when more training samples are provided. In Figure 16 we show how many extra training samples are needed to supply to the randomly initialized model to make its performance match the channel-simulation pretrained model. We see that as more training samples are used, the superiority of the pretrained model (measured in the number of extra samples supplied to the scratch model to gain equal performance) fades out, and finally the difference between these models is negligible.

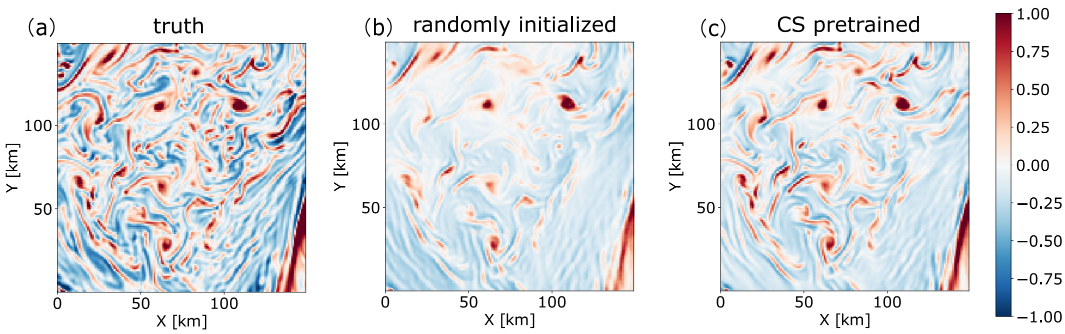

Figure 15. (a) The true normalized vorticity, ζ/f, from a subregion of LLC4320 Region 3 in winter; (b) Unet-predicted normalized vorticity from the scratch model using 10,000 samples and 60 iterations of training; (c) Same as (b) but with the channel simulation pretrained model.

These results raise the questions: what has been transferred or reused from the pretrained model? When training samples are plentiful, do pretrained weights in the model make any difference from the randomly initialized ones? To address these, we use the centered kernel alignment (CKA) (Kornblith et al., 2019; Nguyen et al., 2020) to measure the similarity between layers from different models. This empirical metric first computes the principal components of the correlation matrix between the outputs from a layers of a model when given a large amount of inputs, and then compare the similarity be-

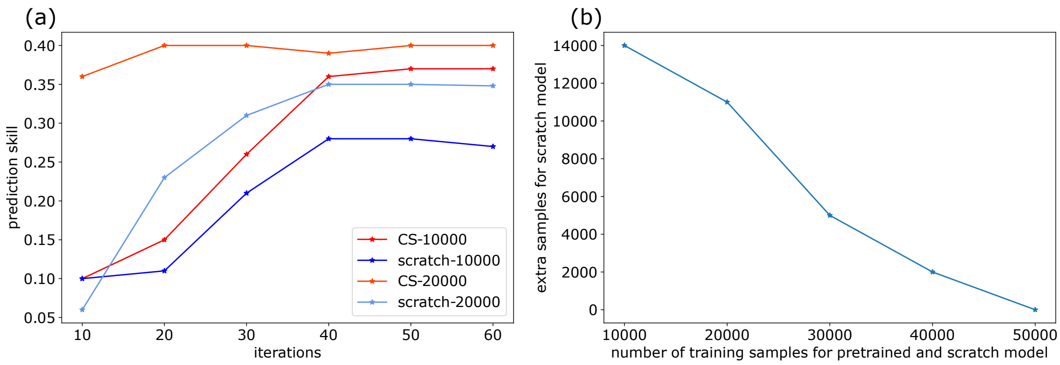

Figure 16. (a) Prediction skill measured for pretrained models and model trained from scratch, using either 10,000 or 20,000 samples of LLC4320 data. (b) Extra training examples needed to boost the performance of scratch model to match channel simulation pretrained model when they are given different number of LLC4320 training samples.

tween principal components from layers of two different models when given the same inputs. The values 1 suggests identical and 0 means orthogonal.

In the upper panel of Figure 17 we show the CKA between the pretrained models with and without tuning using the LLC4320 data. We can see high similarity along the diagonal regardless of the amount of LLC4320 data used, indicating the changes that happen during tuning are mostly small modifications of the original feature space. In the lower panel of Figure 17 we show the CKA between pretrained models and randomly initialized models. The high similarity along the diagonal of the first three layers suggests that similar features are learned by the first few layers, regardless of the starting state of the model. But this similarity doesn't last through the full model, in particular the last two layers. This suggests that even though both models extract information from the input in similar ways, they are taking different approaches in utilizing it to reconstruct the output; even though when measured in correlation and prediction skill, their results show negligible differences.

Results from the CKA analysis in Figure 17 have two important implications. First, it suggests that feature-reuse does happen and is most significant in the first few layers. On the other hand, the pretrained weights set the basis for modification during tuning and this could be a restriction when the training data is largely available and the data for pretraining is very different from the data for training.

When applied to real observation data, the pretrained simulation data should follow similar dynamics and boundary conditions as closely as possible, and it may be worth adding extra layers at the end or just randomly initializing the last few layers of the model. Another implication is the fact that while giving a similar performance, two neural networks with different initial weights have vastly different intermediate results. This poses the difficulty of trying to extract the physical knowledge learned by the machine learning model, if there is any. While the physical law governing the data should be unique, the approximations derived by machine learning models are not and may be very different from one trained model to another.

7 Discussion and Conclusion

In this study, we explored the possibility of using a neural network to reconstruct surface kinematic variables — vorticity, strain and divergence — from snapshots of SSH.

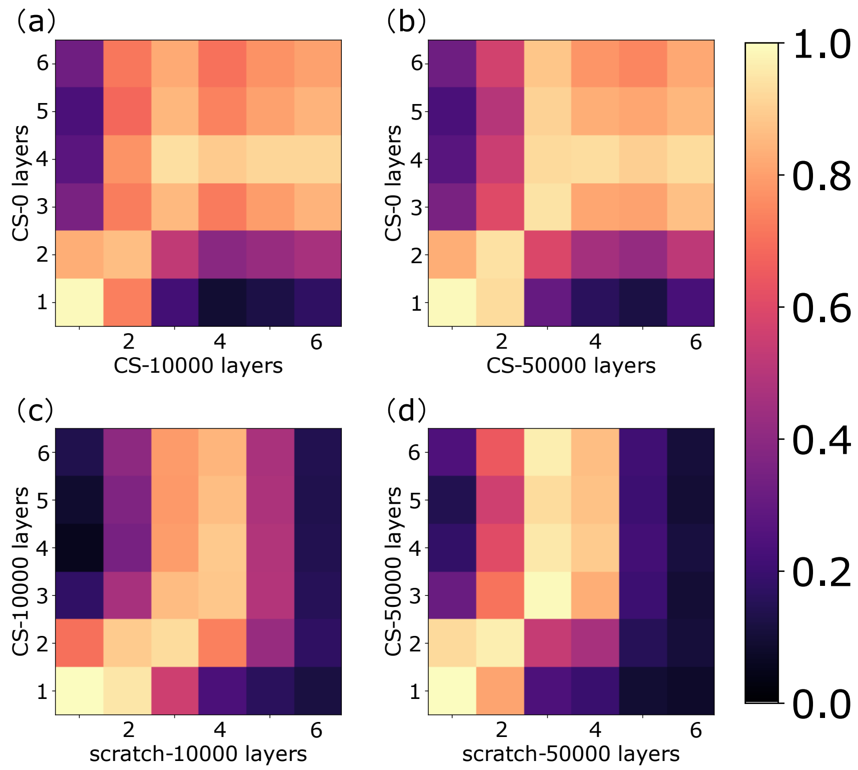

Figure 17. (a) CKA between the channel simulation pretrained models with and without tuned with 10,000 samples of LLC data; (b) same as left but with 50,000 samples of LLC data; (d) CKA between the channel simulation pretrained model and randomly initialized model, with 10,000 samples of LLC data; (d) same as left but with 50,000 samples.

This work was motivated by the anticipated challenges that will emerge once the data from the SWOT satellite becomes available. SWOT will present an unprecedented 2D view of SSH at scales smaller than ever seen before, but this will also raise a number of questions about how to best utilize and interpret these observations (Chelton et al., 2019). These include questions about how to reconstruct surface flows at scales where geostrophy may not be appropriate, and when the SSH perturbations may be strongly influenced by the presence of IGWs. We use neural networks because we currently lack dynamics-based methods like geostrophy. The neural network model works more like traditional analog forecasting methods based on pattern recognition (Balaji, 2021). They unfortunately come with the cost of being less interpretable.

Here we used a particular type of convolution neural network called Unet, which has previously shown to be very successful at different 2D prediction tasks. However, we believe that the success of applying neural networks to our task is not limited to this model. Other CNN-based models should have similar capabilities, and there may be other neural networks with architectures more suited to this task. Also, we used pointwise mean squared error and mean absolute error as loss functions during training, as they are simple to understand conceptually and their properties are well-known. In the future, more complex and task specific loss functions can be devised (Ebert-Uphoff et al., 2021). Since a neural network may never be able to converge to a zero error, due to incomplete knowledge of the hidden states, we also focus on the overall pattern reconstruction rather than only on point-wise errors to evaluate the success and predictions properties of our model.

To do this, we used vorticity-strain JPDFs (Balwada et al., 2021), which help us assess statistically if the predictions appropriately capture the structures present in the flow.

For training our models, we used data from three sources, an idealized channel model with weak IGWs, a region of a realistic high-resolution global simulation (LLC4320) with seasonally varying IGW amplitudes, and a synthetically generated field of IGWs. We are interested in how neural network performs in situations with different strengths of IGWs, since though both the IGW and balanced part get enhanced with finer resolution and expected to be part of the SSH observations gathered by SWOT, their kinematic properties are very different. The IGWs don't contribute much to the passive tracer transport, and may be less relevant for research applications corresponding to transport. It is thus important to understand if the neural network can preserve and predict both signals, or whether it imposes different distortions to them.

When the Unet is trained on the channel simulation, in which IGWs are weak, we find that the reconstruction of surface kinematics is superior to a naive application of geostrophic balance. Not only are point-wise correlation and prediction skills high, but both vorticity-strain joint distributions and conditional divergence distributions, are close to the truth. A similar result is found for the LLC4320 during the winter, when IGWs are relatively weak. However, when training is done on LLC4320 summer, when IGWs are strong, the quality of prediction is decreased.

The quality of these predictions can be understood by considering the loss functions we use. When optimization is done using the mean squared error or mean absolute error, the neural network should converge to the conditional expectation or the conditional median conditioned to the input, respectively. At least for the waves, it can be shown that these conditional metrics for the vorticity and strain conditioned on the SSH snapshots are not necessarily equal to the true target values, but depend on the wavenumber distribution embedded in the training data. For the balanced or frontal part of the flow, no such simple reasoning can be done, but empirically, given the success of the prediction when the waves are weak, it seems that the conditional metrics do converge towards the true surface kinematic variables.

The situation for prediction of the wave divergence is particularly interesting since its conditional expectation and median converge to 0. This implies a neural network predicting the conditional expectation of divergence associated with waves will have a natural tendency to filter them out. We confirmed this result by not only using an idealized synthetic field of IGWs, but also by comparing a model trained on LLC4320 raw data against a version where the waves were greatly filtered out before training. It remains to be examined whether this insight can be leveraged to filter waves from other kinematic variables by using specialized loss functions. This is a promising area for future study.

Overall, in future exploration, we should pay more attention to choosing a more task-specific loss function before turning to more complicated neural networks. While the latter decides how well the final model will be able to generalize, the former determines what the model converges to and is closely related to the underlying physical properties of the problem.

Finally, we also showed that a model pretrained on a simpler simulation can be tuned to work for a more complex model with a smaller amount of data, with the hope that a similar technique can be used to pretrain a model with realistic simulation data and tuned with observational data. This technique is referred to as transfer learning. However, more work needs to be done determine the minimal number of observational data that will be needed to carry out this procedure, and what realistic models will be most suited to perform the pretraining to work with actual SSH observations. It would be ideal if the in-situ data collected at the SWOT "adopt a crossover" sites, which are regions that will be heavily monitored during the first 3 months of the SWOT mission, could be used train machine learning models to recover the flow properties from SSH.

In summary, we show that a neural network can serve as a potential tool to reconstruct surface dynamics from snapshot SSH data. This study was a proof of concept, revealing a few different avenues that should be further investigated before such work can be used for operational purposes.

Appendix A Comparison between mean squared error and mean absolute error as loss functions

When considering the vorticity-strain JPDFs, we noticed that the JPDF of the predicted results is usually less spread out than the true JPDF (e.g. Figure 8 or Figure 10). This happens because at smaller scales, which are usually associated with the outer contours of the JPDF, the flow deviates more strongly from geostrophy. Thus, it is less likely that a one-to-one relationship exists between the SSH and the surface flow; many different flow structures are possible for the same SSH structure. In this case, the machine learning model offers a statistical estimate of the surface kinematic variable conditioned on the SSH, and this statistical estimate depends on the loss function we use. In section 5, we used this property to our advantage, and filtered out the IGW divergence. Here we show that changing the the loss function from mean squared error (MSE) to mean absolute error (MAE) changes the details of the predicted kinematic variables, and thus impacts the JPDF of the predicted variables. In particular, when using the mean absolute error a clear cut off in $\zeta/f_0 = -1$ appears (Figure A1), which is absent when using mean squared error.

We speculate that this sharp cut-off, when using MAE, may be associated with the fact that $\zeta/f_0 \leq -1$ is also the criterion for barotropic, centrifugal and inertial instabilities (Hoskins, 1974; Thomas et al., 2013). The relatively larger scale flow tries to push the $\zeta/f_0 \leq -1$, and the instability mechanism tries to restore the value to be $\zeta/f_0 \geq -1$, potentially resulting in a significant amount of variability centered near this threshold. Since $\zeta/f_0 \leq -1$ is likely to happen at small scales, it has a less deterministic dependence on SSH. So, for a similar SSH structure, the flow can form a wide range of ζ/f_0 values, and this distribution is likely a long tail distribution, peaking around -1 and extending to smaller negative values (≤ -1) that appear intermittently and are wiped out by the instabilities. When we use MSE, the machine learning model converges to the conditional expectation of vorticitygiven a SSH pattern. For long tail distributions, the expectations can be diverse and distinct from the peak value. However, when we use MAE, the model converges towards the conditional median instead. In this case, the results become less variant and cluster around the peak value of -1. This likely leads to the sharper cut-off in the vorticity prediction.

Thus, we conclude that predictions of surface kinematic variables from the model trained using the MSE looked more natural than ones from MAE, which is why we use MSE is this study. However, even the MSE based estimates are just statistical estimates from the training data and can be far from the truth. Since part of the variability is due to the missing information in the input to the model trained only using SSG, this cut-off disappears when we have more variables such as surface temperature in the model input (not shown).

Appendix B Data and Code Availability Statement

The Python notebooks and code samples required to train the models and recreate the figures can be found at `https://github.com/qyxiao/CNN-for-SSH-reconstruction`. The channel simulation and LLC4320 data can be accessed using the Pangeo (`https://pangeo.io/`) data catalog at `https://catalog.pangeo.io/browse/master/ocean/channel/`

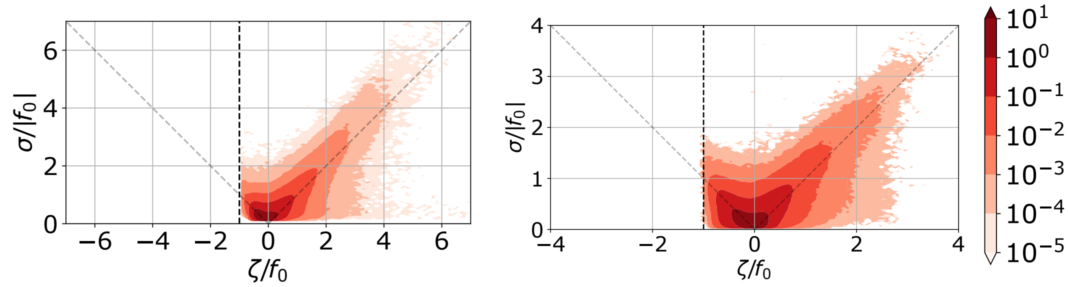

Figure A1. Vorticity-strain joint distributions of (left) reconstructed channel simulation vorticity and (right) reconstructed LLC4320 winter vorticity when using mean absolute error as a loss function to train the Unet. The dashed vertical line corresponds to $\zeta/f_0 = -1$, which seems to emerge as a hard cutoff when using the mean absolute errors as the loss function.

`channel_ridge_resolutions_01km/` and `https://catalog.pangeo.io/browse/master/ocean/LLC4320/` respectively. The Lagrangian filtered LLC4320 data can be accessed from `https://doi.org/10.5281/zenodo.6561068`. The synthetic IGW is generated with Matlab package GLOceanKit (`https://github.com/Energy-Pathways-Group/GLOceanKit`).

Acknowledgments

QX, CSJ, KSS, and RPA acknowledge funding from NASA's SWOT Science Team program, grant number 80NSSC20K1142. DB received M2LInES research funding by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program.

References

Bachman, S. D., & Klocker, A. (2020). Interaction of jets and submesoscale dynamics leads to rapid ocean ventilation. *Journal of Physical Oceanography*, *50*(10), 2873–2883.

Balaji, V. (2021). Climbing down Charney's ladder: Machine learning and the post-Dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200085.

Balwada, D., LaCasce, J. H., & Speer, K. G. (2016). Scale-dependent distribution of kinetic energy from surface drifters in the Gulf of Mexico. *Geophysical Research Letters*, *43*(20), 10–856.

Balwada, D., Smith, K. S., & Abernathey, R. (2018). Submesoscale vertical velocities enhance tracer subduction in an idealized Antarctic Circumpolar Current. *Geophysical Research Letters*, *45*(18), 9790–9802.

Balwada, D., Xiao, Q., Smith, S., Abernathey, R., & Gray, A. R. (2021). Vertical fluxes conditioned on vorticity and strain reveal submesoscale ventilation. *Journal of Physical Oceanography*, *51*(9), 2883–2901.

Berta, M., Griffa, A., Haza, A., Horstmann, J., Huntley, H., Ibrahim, R., ... Poje, A. (2020). Submesoscale kinematic properties in summer and winter surface flows in the Northern Gulf of Mexico. *Journal of Geophysical Research: Oceans*, *125*(10), e2020JC016085.

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.

Chelton, D. B., Schlax, M. G., Samelson, R. M., Farrar, J. T., Molemaker, M. J., McWilliams, J. C., & Gula, J. (2019). Prospects for future satellite estimation of small-scale variability of ocean surface velocity and vorticity. *Progress in Oceanography*, *173*, 256–350.

Ducet, N., Le Traon, P.-Y., & Reverdin, G. (2000). Global high-resolution mapping of ocean circulation from TOPEX/Poseidon and ERS-1 and ERS-2. *Journal of Geophysical Research: Oceans*, *105*(C8), 19477–19498.

Early, J. J., Lelong, M. P., & Sundermeyer, M. A. (2021). A generalized wave-vortex decomposition for rotating Boussinesq flows with arbitrary stratification. *Journal of Fluid Mechanics*, *912*.

Ebert-Uphoff, I., Lagerquist, R., Hilburn, K., Lee, Y., Haynes, K., Stock, J., ... Stewart, J. Q. (2021). CIRA guide to custom loss functions for neural networks in environmental sciences–Version 1. *arXiv:2106.09757*.

Fablet, R., & Chapron, B. (2022). Multimodal learning-based inversion models for the space-time reconstruction of satellite-derived geophysical fields. *arXiv:2203.10640*.

Fu, L.-L., Alsdorf, D., Morrow, R., Rodriguez, E., & Mognard, N. (2012). *SWOT: The Surface Water and Ocean Topography Mission: wide-swath altimetric elevation on Earth* (Tech. Rep.). Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration.

George, T. M., Manucharyan, G. E., & Thompson, A. F. (2021). Deep learning to infer eddy heat fluxes from sea surface height patterns of mesoscale turbulence. *Nature Communications*, *12*(1), 1–11.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hoskins, B. (1974). The role of potential vorticity in symmetric stability and instability. *Quarterly Journal of the Royal Meteorological Society*, *100*(425), 480–482.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711).

Jones, C. S., Xiao, Q., Abernathey, R., & Smith, K. S. (2022). Separating balanced and unbalanced flow at the surface of the Agulhas region using Lagrangian filtering. *EarthArXiv*. doi: 10.31223/X5D352

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... others (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4681–4690).

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. *arXiv:1803.04189*.

Levine, M. D. (2002). A modification of the Garrett-Munk internal wave spectrum. *Journal of physical oceanography*, *32*(11), 3166–3181.

Manucharyan, G. E., Siegelman, L., & Klein, P. (2021). A deep learning approach to spatio-temporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. *Journal of Advances in Modeling Earth Systems*, *13*(1), e2019MS001965.

Marshall, J., Hill, C., Perelman, L., & Adcroft, A. (1997). Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling. *Journal of Geophysical*

Research: Oceans, *102*(C3), 5733–5752.
Munk, W. (1981). Internal waves and small-scale processes. *Evolution of physical oceanography*, 264–291.
Munk, W. (2002). *Testimony before the U.S. Commission on Ocean Policy.* `http://govinfo.library.unt.edu/oceancommission/meetings/apr18_19_02/munk_statement.pdf`.
Nguyen, T., Raghu, M., & Kornblith, S. (2020). Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv:2010.15327*.
Omand, M. M., D'Asaro, E. A., Lee, C. M., Perry, M. J., Briggs, N., Cetinić, I., & Mahadevan, A. (2015). Eddy-driven subduction exports particulate organic carbon from the spring bloom. *Science*, *348*(6231), 222–225.
Qiu, B., Chen, S., Klein, P., Torres, H., Wang, J., Fu, L.-L., & Menemenlis, D. (2020). Reconstructing upper-ocean vertical velocity field from sea surface height in the presence of unbalanced motion. *Journal of Physical Oceanography*, *50*(1), 55–79.
Qiu, B., Chen, S., Klein, P., Ubelmann, C., Fu, L.-L., & Sasaki, H. (2016). Reconstructability of three-dimensional upper-ocean circulation from SWOT sea surface height measurements. *Journal of Physical Oceanography*, *46*(3), 947–963.
Rocha, C. B., Gille, S. T., Chereskin, T. K., & Menemenlis, D. (2016). Seasonality of submesoscale dynamics in the Kuroshio Extension. *Geophysical Research Letters*, *43*(21), 11–304.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).
Shcherbina, A. Y., D'Asaro, E. A., Lee, C. M., Klymak, J. M., Molemaker, M. J., & McWilliams, J. C. (2013). Statistics of vertical vorticity, divergence, and strain in a developed submesoscale turbulence field. *Geophysical Research Letters*, *40*(17), 4706–4711.
Siegelman, L., Klein, P., Rivière, P., Thompson, A. F., Torres, H. S., Flexas, M., & Menemenlis, D. (2020). Enhanced upward heat transport at deep submesoscale ocean fronts. *Nature Geoscience*, *13*(1), 50–55.
Sinha, A., & Abernathey, R. (2021). Estimating ocean surface currents from satellite observable quantities with machine learning. *Frontiers in Marine Science*, *8*. doi: 10.3389/fmars.2021.672477
Thomas, L. N., Taylor, J. R., Ferrari, R., & Joyce, T. M. (2013). Symmetric instability in the Gulf Stream. *Deep Sea Research Part II: Topical Studies in Oceanography*, *91*, 96–110.
Torres, H. S., Klein, P., D'Asaro, E., Wang, J., Thompson, A. F., Siegelman, L., . . . Perkovic-Martin, D. (2022). Separating energetic internal gravity waves and small-scale frontal dynamics. *Geophysical Research Letters*, *49*(6), e2021GL096249.
Torres, H. S., Klein, P., Menemenlis, D., Qiu, B., Su, Z., Wang, J., . . . Fu, L.-L. (2018). Partitioning ocean motions into balanced motions and internal gravity waves: A modeling study in anticipation of future space missions. *Journal of Geophysical Research: Oceans*, *123*(11), 8084–8105.
Uchida, T., Balwada, D., Abernathey, R., McKinley, G., Smith, S., & Levy, M. (2019). The contribution of submesoscale over mesoscale eddy iron transport in the open Southern Ocean. *Journal of Advances in Modeling Earth Systems*, *11*(12), 3934–3958.
Wang, J., Flierl, G. R., LaCasce, J. H., McClean, J. L., & Mahadevan, A. (2013). Reconstructing the ocean's interior from surface data. *Journal of Physical Oceanography*, *43*(8), 1611–1626.
Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few exam-

ples: A survey on few-shot learning. *ACM computing surveys (CSUR)*, *53*(3), 1–34.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354–7363).