Surface to Seafloor: A Generative AI Framework for Decoding the Ocean Interior State

Andre N. Souza[a] , Simone Silvestri[a] , Katherine Deck[b] , Tobias Bischoff[c] , Raffaele Ferrari[a] , and Glenn R. Flierl[a]

[a] *Massachusetts Institute of Technology, Cambridge, MA, USA,*
[b] *California Institute of Technology, Pasadena, CA, USA,*
[c] *Aeolus Labs, San Francisco, CA, USA*

Corresponding author: Andre Nogueira de Souza, andrenogueirasouza@gmail.com

ABSTRACT: Understanding subsurface ocean dynamics is essential for quantifying oceanic heat and mass transport, but direct observations at depth remain sparse due to logistical and technological constraints. In contrast, satellite missions provide rich surface datasets—such as sea surface height, temperature, and salinity—that offer indirect but potentially powerful constraints on the ocean interior. Here, we present a probabilistic framework based on score-based diffusion models to reconstruct three-dimensional subsurface velocity and buoyancy fields, including the energetic ocean eddy field, from surface observations. Using a 15-level primitive equation simulation of an idealized double-gyre system, we evaluate the skill of the model in inferring the mean circulation and the mesoscale variability at depth under varying levels of surface information. We find that the generative model successfully recovers key dynamical structures and provides physically meaningful uncertainty estimates, with predictive skill diminishing systematically as the surface resolution decreases or the inference depth increases. These results demonstrate the potential of generative approaches for ocean state estimation and uncertainty quantification, particularly in regimes where traditional deterministic methods are underconstrained or ill-posed.

1. Introduction

Inferring subsurface ocean dynamics from surface observations remains a fundamental challenge in physical oceanography. The ocean interior plays a crucial role in regulating Earth's climate by transporting heat, carbon, and other tracers, yet direct measurements at depth are sparse due to logistical and technological limitations. In contrast, satellite missions have provided decades of continuous, high-resolution observations of sea surface height (SSH) and temperature and large-scale salinity observations, offering a vastly richer dataset than what is available below the surface (Liu and Hu 2024). Theoretical and numerical studies suggest that surface anomalies—such as sea surface height—strongly correlate with subsurface structures at mesoscale and submesoscale scales, meaning surface data could provide meaningful constraints on the ocean interior (Hurlburt 1986; Cooper and Haines 1996). However, existing methods for reconstructing subsurface states from surface observations are often limited by strong dynamical assumptions or high computational costs.

This study focuses on reconstructing interior velocity and temperature fields, with a particular emphasis on their correlations, which are essential for estimating mass and heat transport throughout the ocean. Mesoscale eddies, which account for 20–50% of total heat transport (Su et al. 2018; Saenko et al. 2018), are particularly difficult to infer from surface observations alone. A deterministic reconstruction, which attempts to match a single "true" state, may fail to capture these dynamically significant features. Instead, we take a probabilistic approach, where a generative model reconstructs a distribution of plausible subsurface states statistically consistent with available surface data. This methodology ensures that eddy statistics, such as their variance, spatial correlations, and contributions to heat transport, are preserved without erasing dynamically important details.

Several physically motivated methods have been developed to infer subsurface ocean states from surface data, each with distinct advantages and limitations. The Surface Quasi-Geostrophic (SQG) framework has been widely used for subsurface reconstruction. Still, it is constrained to scales above 150 km (Held et al. 1995; Lapeyre and Klein 2006; Wang et al. 2013; Liu et al. 2019) due to the inherent assumptions in QG theory and the lack of detailed information about the vertical stratification. Recent satellite missions such as the Surface Water and Ocean Topography (SWOT) Mission (Fu et al. 2024) offer much higher resolution (down to 1–10 km), limiting the applicability

of traditional large-scale geostrophic approximations. Theoretical extensions, such as the SQG-mEOF-R method, improve upon standard SQG by accurately reconstructing both surface- and subsurface-intensified eddies, particularly in regions with complex vertical stratification, making it more effective across oceanic regimes such as the Kuroshio Extension and the Peru-Chile upwelling system, (Yan et al. 2020), but still build on assumptions applicable only to larger scale ocean structures.

Theoretical limitations can be overcome in the presence of data-abundance. In particular, machine learning offers a data-driven alternative to augment or completely replace physically constrained methods for ocean state estimation. Its applications range from model emulation (Dheeshjith et al. 2024) to data assimilation (Champenois and Sapsis 2024), with increasing success in leveraging satellite observations. Linear models have demonstrated the statistical predictability of large-scale ocean currents, such as the Atlantic Meridional Overturning Circulation (AMOC), using satellite-derived data (Cromwell et al. 2007; DelSole and Nedza 2022). More recently, feed-forward neural networks (FNNs) and convolutional neural networks (CNNs) have been applied to reconstruct meridional overturning circulations from ocean bottom pressure, zonal wind stress, and sea surface properties (Solodoch et al. 2023a; Meng et al. 2024). Additionally, studies incorporating satellite and Argo float data have demonstrated skill in inferring velocity at depth (Chapman and Charantonis 2017), while CNNs trained on quasi-geostrophic (QG) simulations have successfully mapped surface stream function anomalies to subsurface flow fields in idealized settings (Bolton and Zanna 2018). Complementary efforts using QG theory (George et al. 2021) and data assimilation techniques (Martin et al. 2023) have refined our understanding of large-scale ocean dynamics. Meanwhile, operationally constrained models integrate observational data with numerical simulations, leading to more accurate reconstructions of subsurface ocean conditions (Solodoch et al. 2023b).

While these studies have demonstrated that satellite-derived surface information can infer the ocean interior, most have relied on supervised learning, where a model is trained to minimize residuals against a single "true" state. However, a given surface ocean state does not deterministically correspond to a unique deep ocean state. A more physically sound approach requires probabilistic modeling, which allows for generating an ensemble of plausible subsurface states constrained by surface observations, while preserving the uncertainties and variability inherent in ocean dynamics.

Importantly, probabilistic modeling allows for uncertainty quantification (UQ) (Forget et al. 2015; Souza et al. 2020): not only can samples from the inferred distribution be explored, but these samples also provide estimates of plausible subsurface fields when surface information is insufficient to constrain the output fully. Generative modeling offers a flexible approach to sampling from such distributions, with score-based diffusion models emerging as a particularly effective tool for high-dimensional statistical inference. Generative-adversarial models (GANs) have also been applied to predict ocean temperature profiles from sea surface temperature (Meng et al. 2021; Zhang et al. 2023), but these works did not leverage the potential for uncertainty quantification. More recently, Martin et al. (2025) have leveraged score-based generative methods in the context of ocean surface data assimilation but again, they did not focus on uncertainty quantification. Beyond conditional sampling, score-based generative models have demonstrated utility in diverse applications, including calculating response functions (Giorgini et al. 2024, 2025b), estimating the dimensionality of data manifolds (Stanczuk et al. 2023), stabilizing the dynamical evolution neural network models (Pedersen et al. 2025), or devising stochastic analogs to chaotic equations (Giorgini et al. 2025a).

This study develops a probabilistic generative approach to infer the ocean interior state conditional on observable surface variables, such as sea surface height. By framing subsurface inference as a conditional sampling problem, we aim to quantify the variability and uncertainty of ocean interior states while leveraging the wealth of information encoded in surface observations. Rather than simply predicting a single deterministic state of ocean currents from the surface, our objective is to determine to what extent observations of the ocean surface constrain the ocean interior state while ensuring that the whole subsurface ocean state, including correlations amongst different variables, are properly captured. As we will show, a well-trained model can even predict the extent to which it is wrong in its predictions.

With these aims in mind, the rest of the paper is organized as follows. In Section 2, we provide the physical context for the study, introduce the numerical simulation that we use as a testbed for the generative method, and describe the generative modeling framework. Results are presented in Section 3, where we apply the generative model to the ocean simulation and assess its predictive skill in reconstructing subsurface velocities and temperature alongside their variability as a function of sea surface height. We also repeat the exercise with spatially filtered sea surface heights to

understand how less informative surface conditions alter predictions. Section 4 concludes and discusses further extensions of this work.

2. Numerical Simulations and Generative AI Model

In the following subsections, we describe the details of the ocean model output, the generative AI model, and training. In essence, we will use an ocean simulation to generate a time series of the ocean state, consisting of the sea surface height (SSH), velocities, and temperature at various vertical levels. This dataset then forms the backbone for training a neural network to determine plausible velocities and temperatures given a sea surface height.

a. Ocean Simulation

As with any idealized simulation, we focus on particularly salient features that generalize to more complex scenarios. Thus, the role of the simulation is not to capture realistic features of the ocean but rather to serve as a testing ground for whether or not statistical inference of the free surface to the ocean interior can be performed. With this in mind, we choose a baroclinic double gyre, similar to (Cox and Bryan 1984), as a representative ocean simulation for the methodology. The simulation generates circulation patterns consistent with those observed in the subtropical and subpolar gyres. *We will not run the ocean model to statistical equilibrium* so that we can study both statistically stationary and non-stationary features, (Hasselmann 1976).

Using Oceananigans.jl (Ramadhan et al. 2020; Silvestri et al. 2024a; Wagner et al. 2025a; Silvestri et al. 2024b; Wagner et al. 2025b), we solve the Boussinesq equations under the hydrostatic approximation. The dynamical core uses a nonlinear free surface and $z^{\star}$ coordinates. The prognostic variables are the horizontal velocities, U and V, the SSH η, and the temperature T. A linear equation of state relates the buoyancy of seawater B to the temperature, (Roquet et al. 2015). The figures will refer to the Oceananigans simulation output as "OcS" for brevity.

The idealized ocean domain extends from 15 to 75 degrees north in latitude and from 0 to 60 degrees in longitude with walls on all sides. The domain is 1800 m deep with a flat bottom. A zonal wind stress is applied at the surface, independent of longitude, with a cosine profile peaking eastward (positive) in the middle of the domain at a value of $0.1\,\mathrm{Nm^{-2}}$ and at a minimum and westward (i.e. negative) at the northern and southern walls. The surface temperature is restored

6

to a linear profile in latitude ranging from 30 C° on the southern boundary at 15°N, to 0 C° at the northern boundary at 75°N. A quadratic bottom drag is imposed at the seafloor and convective adjustment is used for vertical mixing. The (dimensionless) quadratic drag coefficient is 10^{-3} and we use a horizontal diffusivity and viscosity of 2.5×10^2 m^{2}s^{-1}, a background vertical diffusivity of 10^{-5} m^{2}s^{-1}, and background vertical viscosity of 10^{-2} m^{2}s^{-1}.

The domain is discretized into $256 \times 256 \times 15$ grid points, leading to a 25-kilometer resolution in latitude. The longitudonal resolution depends on the given latitude, and is calculated by multiplying the latitudonal resolution with the cosine of the latitude, which ranges from factors of 0.96 to 0.26. The vertical grid spacing varies from 50 meters at the surface to 180 meters at depth. The simulation is initialized with zero velocity, a flat sea surface, and uniform buoyancy and runs for ~ 425 years. The ocean circulation spins up in the first ~ 100 years, which are discarded in the following analysis. Once spun up, the solution is characterized by a double gyre circulation with convection along the northern boundary and a rich mesoscale eddy field. However, the ocean depths have not yet reached statistical equilibrium and continues to evolve.

We show the resulting fields and their evolution in time in Figure 1. From left to right, each column shows the SSH η, the zonal U, meridional V, vertical velocity W, and temperature T at a 400 meter depth. Each row is a snapshot at years 20, 200, and 300. There is a cyclonic gyre in the northern half of the domain and an anti-cyclonic gyre in the lower half. The convection in the north half of the domain cools fluid parcels that sink and fill the abyss with cold water. The resulting simulation exhibits a non-trivial depth-dependent structure, deviating from equivalent barotropic behavior and includes a substantial ageostrophic component at all vertical levels.

The generative AI task of this manuscript is to generate plausible predictions for the ocean state (U, V, W, T) at all depths, given the first column, the SSH η. We use the instantaneous SSH to infer plausible subsurface fields. We use this simulation as a proxy for performing ocean inference given satellite information and a model for the ocean interior. We only focus on the SSH since the imposed forcing constrains the present simulation's surface temperature field. In a more realistic setup, we would use more information.
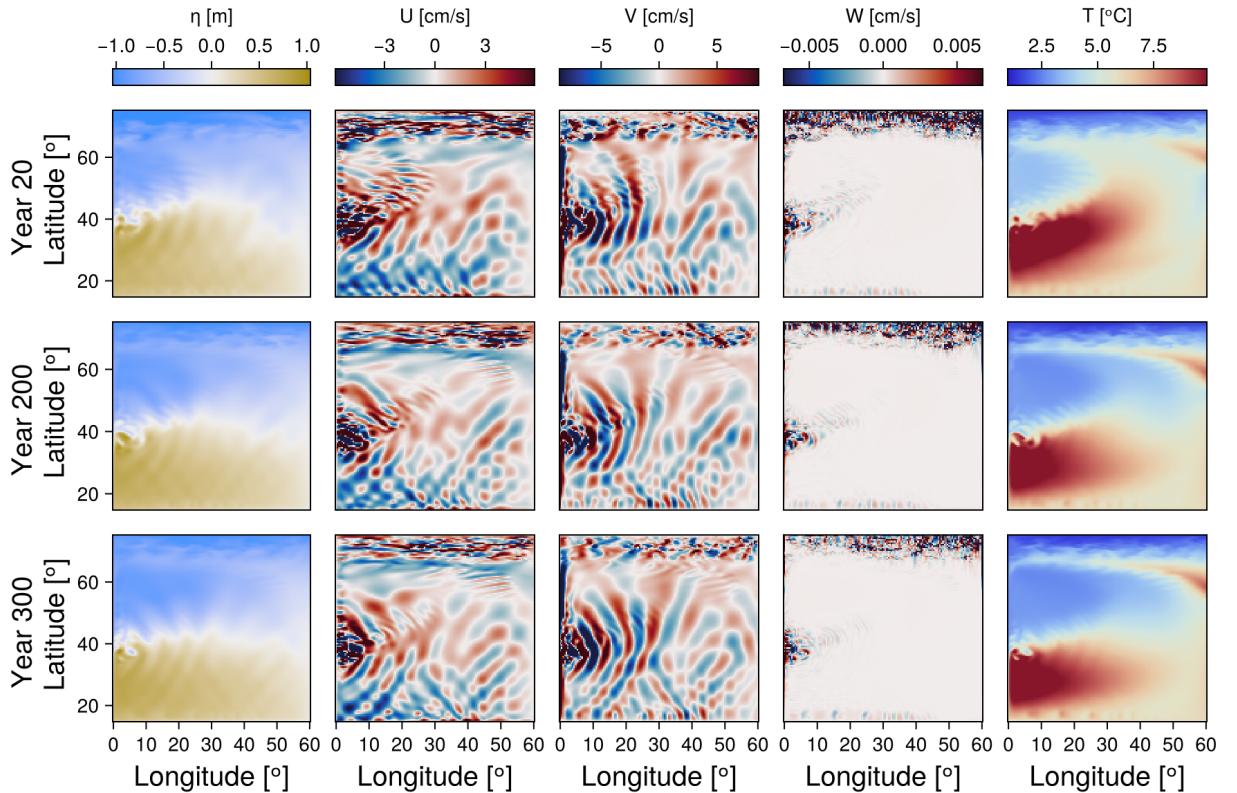
FIG. 1. Snapshots of the ocean state from a double-gyre simulation. Each column is a different ocean state, and each row is a fixed moment in time. The first row consists of year 20, the middle row year 200, and the last row year 300. The columns are the sea surface height, the zonal velocity U, the meridional velocity V, the vertical velocity w, and the temperature T. The velocities and temperature are evaluated at a 400-meter depth.

Our simulation is not in statistical equilibrium, as can be observed from Figure 2, which shows the L^1 norm (the average absolute value) of all variables at a few selected depths, e.g.

$$\|V\|_1 = \frac{1}{60°} \frac{1}{60°} \int_{15°}^{75°} \int_{0}^{60°} |V(\lambda, \varphi, z, t)| \, \mathrm{d}\lambda \mathrm{d}\phi. \tag{1}$$

We use this measure as it is sensitive to localized fluctuations, such as the root-mean-squared value. Including metric terms does not change the overall trend and is unnecessary for the present case, but we will include them later when calculating physically meaningful quantities rather than statistical measures of non-stationarity. The initial spinup period (blue) is discarded. Thus, all the

data in the work uses the time periods associated with the middle portion (red) and the end portion (orange) of the simulation. Closer to the surface (first column), it appears that a quasi-statistical equilibrium has been reached, whereas at depth (last column), statistical equilibrium has not been reached, especially for temperature. This quasi-equilibrium nature of the ocean is a feature we wished to capture in our dataset when investigating the limits of statistical inference predicated on past data. We comment that the sea surface height's L^1 norm is also not in equilibrium. Stated differently, our "data distribution" (red) will be different than our "validation distribution" (orange) and will contribute as an extra source of error.

b. Score-Based Diffusion Model

Our work leverages recent advancements in score-based generative modeling, particularly building upon foundational contributions, (Song and Ermon 2019, 2020; Ho et al. 2020; Karras et al. 2022). These methods allow for efficient sampling of arbitrary probability distributions, including conditional distributions.

In the present work, we draw samples from subsurface fields, U, V, W, T, given the sea surface height η, where the underlying probability distribution comes from the primitive equations. Mathematically, this is represented as a conditional probability distribution, $\mathcal{P}$,

$$\text{conditional distribution} = \mathcal{P}(\text{subsurface fields}|\text{surface fields}). \tag{2}$$

The conditional probability distribution outputs four fields U, V, W, T at every grid point in three-dimensional space, conditioned on the SSH η. Concretely, samples from this distribution produce $4 \times 256 \times 256 \times 15$ numbers conditioned on 256×256 numbers using the full resolution of the simulation. This high-dimensional conditional probability distribution is learned from the data set.

If the conditional information is perfectly informative and there is a functional relationship, $\mathcal{F}$, between SSH and predicted fields and predicted fields. In this case, the conditional distribution collapses to a delta function,

$$\mathcal{P}(\text{subsurface fields}|\text{surface fields}) = \delta(\text{subsurface fields} - \mathcal{F}[\text{subsurface fields}]). \tag{3}$$

9
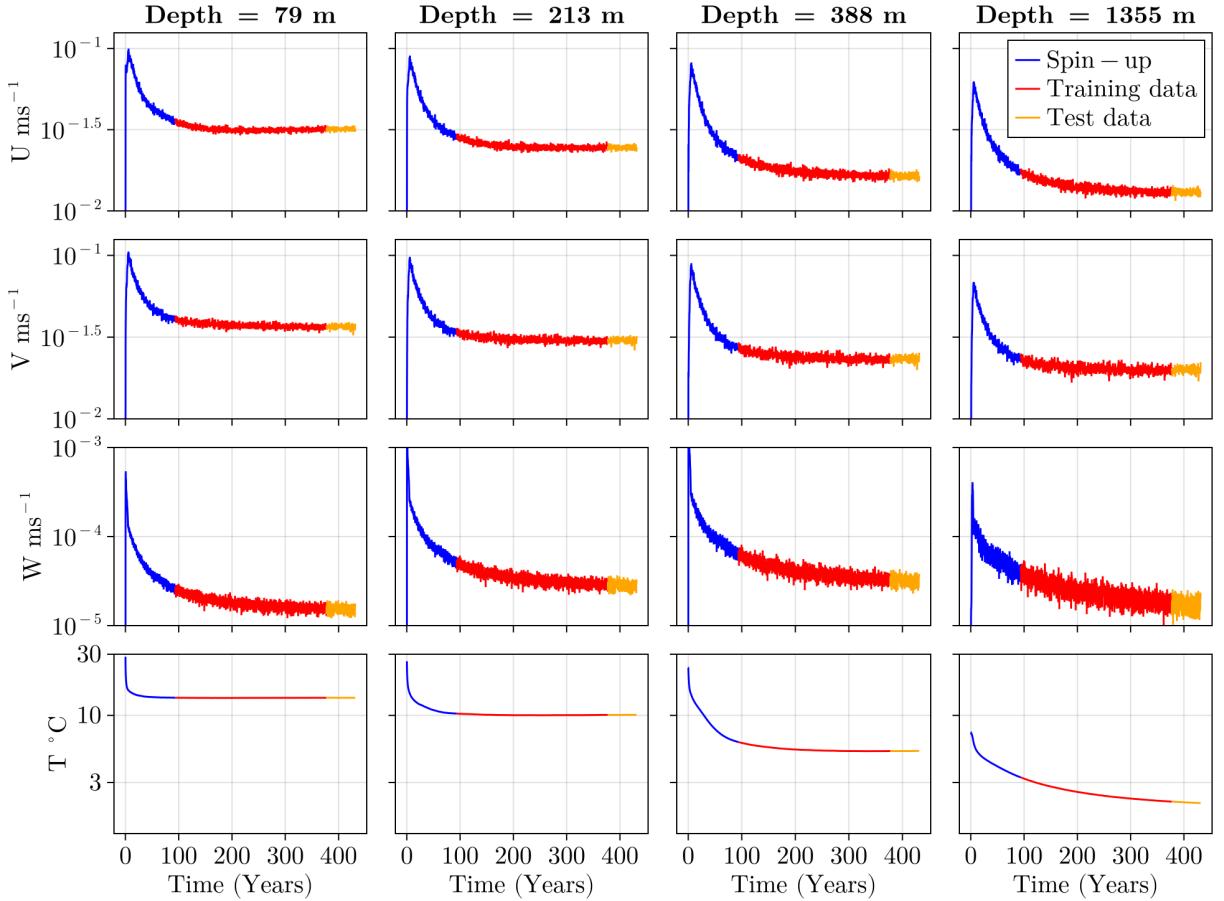
FIG. 2. Evolution of L^1 norms in time. We show the average value of the absolute value of various fields (see Equation 1) as a function of time in years on a logarithmic scale. The surface fields are in quasi-statistical equilibrium, whereas the fields near the bottom of the domain are still converging, especially for the temperature field. The different colors represent the discarded spin-up data (blue), training data (red), and test data (orange) for training a score-based diffusion model.

If the provided information is "not useful" or maximally uninformative, then the conditional probability distribution, $\mathcal{P}$, should reduce down to the marginal distribution of the data, e.g., a random sample with statistics similar those the generative model has been trained on independently of the surface information,

$$\mathcal{P}(\text{subsurface fields}|\text{surface fields}) = \mathcal{P}_M(\text{subsurface fields}), \tag{4}$$

10

where $\mathcal{P}_M$ is the marginal distribution of the subsurface fields. Most inference tasks lie between these two extremes, Equations 3 and 4. We comment that we use the word "marginal" as is commonly used in the statistics literature with regards to distributions.

Traditional machine learning methods find a functional form $\mathcal{F}$ between input and outputs. This methodology works best when such a functional form exists. In the absence of an approximate functional relationship, traditional methods that minimize a root-mean square (RMS) error with the aim of finding a functional form can only infer the average value of the conditional probability in Equation 2. (This happens for the same reason that a "mean" statistic minimizes the RMS error of a random variable.) We surmise that this observation partially explains why fields drawn from a conditional distribution better represent the statistical properties of fluid flows compared with those learned via a deterministic functional relationship. Details will not be "washed away" from averaging.

The key deep-learning methodologies to sample the conditional distribution include reverse diffusion and conditional denoising score-matching (Anderson 1982; Hyvärinen and Dayan 2005; Vincent 2011; Batzolis et al. 2021). Architecturally, we adopt the U-Net framework (Ronneberger et al. 2015), a well-established model utilized in generative modeling (Deck and Bischoff 2023; Bischoff and Deck 2024).

The simulated SSH data is coarse-grained from its original resolution of $256 \times 256 \times 15$ to $128 \times 128 \times 15$ by averaging neighboring horizontal points to expedite the training of the generative model. This lower-resolution dataset loses few features from the original but dramatically expedites the training process since the entire dataset can be loaded into computer memory. We take monthly snapshots over the simulation period to ensure that our data is sufficiently decorrelated.

A coarse-grained sea surface height is the conditional information provided to the model, which serves as the primary input for generating samples. We examine various SSH coarse-graining levels to determine the subsurface fields' predictability. Specifically, we average neighboring points by factors of two until only a single value for the SSH remains as input for the deep-learning model. That is to say, we progress from a 50 kilometer resolution to a 6,400 kilometer resolution in latitude. This procedure results in 8 candidate SSH "inputs", at any fixed moment in time due to the different resolutions. Each state, U, V, W, T, η, is separately normalized at each of the 15 levels of depth so that each field has the spatio-temporal mean removed and is divided by twice the spatio-temporal

11

standard deviation. This rescaling keeps the fields roughly $O(1)$ for training purposes. Thus, 8 numbers are used separately to rescale U, V, W, T at each of the 15 vertical level, and 2 numbers are used to rescale η, leading to a total of $8 \times 15 + 2$ scaling factors.

We train a single neural network that simultaneously outputs the entire 3D field for U, V, W, T and takes in all coarse-grained inputs. We enable the latter step by randomly choosing among the 8 different coarse-grained SSH. We use a 2D UNet with 61 input channels, 60 output channels (15×4 2D output fields and 1 2D input field), and $O(10^7)$ parameters. The resulting network has a $O(10^2)$ megabytes memory footprint, which we compare to the $O(10^2)$ gigabytes used to train the generative model. In this sense, we can think of the generative model as a statistical compression technique which, in the present case, reduces the storage by a factor of 10^3. The Adam optimizer (Kingma and Ba 2014) is employed for training the network.

3. Results

Given the ocean simulation data (OcS) from Section 2a and the neural network architecture from Section 2b, we perform statistical inference by drawing 100 sample fields using the reverse diffusion process, given SSH. We test the enural network by feeding SSH from the last 50 years of the double-gyre simulation which were not used during training. The predictions of the subsurface variables are tested for the various coarse-grained SSH input data.

The generative models draw many sample realizations from one SSH, yet we only have one ground-truth simulation to compare against for a given SSH. As such, we will look two ways of comparing the distributional output to the "ground truth". The first is to compare the ensemble mean of the generated samples to the ground truth since an "ensemble mean" is relatively easy to compute and serves as a good "average" guess. This choice is most useful when the conditional information is sufficiently informative, as in the limit described by Equation 3. The second is to simultaneously compare the ocean simulation output to all resulting ensemble members, which yields a probability distribution. We can then compare this probability distribution to one where we pick one of our generated samples (for example, the first) and compare it to all the others (for example, 2-100). We then check to see if the two probability distributions are consistent. This latter choice is a purely "data-driven" way to determine how accurate the resulting predictions of the model are. A well-trained model (which requires sufficient data and an appropriate choice of

architecture) makes meaningful estimates of the range of plausible fields for a given SSH, and thus the two resulting distributions should be similar. We will return to this point later.

a. Baseline Performance of Generative Model

We start by examining how well the SSH constrains the three components of velocity and temperature at a 400 meter depth. We show the result of a trained neural network model for generating conditional samples in Figure 3. The sea surface height (first column) and corresponding simulated variables at a 400-meter depth (second column) are taken from 50 years in the future from the end of the training set. Each row corresponds to a different ocean state component, (U, V, W, T), as labeled. Using the generative model, we draw 100 samples from the conditional distribution with this sea surface height as the conditional information. We then take the ensemble average over these samples (third column). We show 2 of the 100 samples in the fourth and fifth columns to illustrate the "generative" nature of the prediction from the neural network.
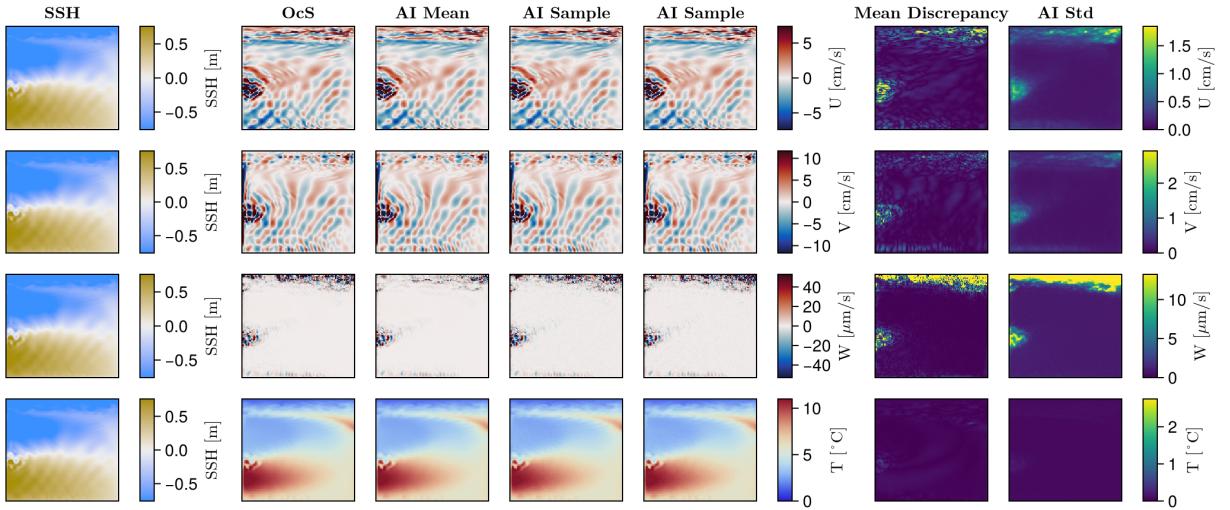
FIG. 3. Statistical inference of the ocean interior state from SSH. Given the SSH input field (left column), we reconstruct the ocean state: the velocity component fields U, V, W, and temperature T (second column). All fields are generated simultaneously. We show the mean over 100 samples predicted by the neural network (third column) and two representative samples (fourth and fifth column). The absolute difference between the AI mean and the OsC ground truth is displayed in the second-to-last column alongside the AI standard deviation in the last column.

We see an overall agreement between the ground truth (second column) and the expected value of the model predictions (third column). This result means that a substantial ageostrophic component is also captured. The ensemble mean of the generated samples looks similar to the ground truth and the representative generated samples since there is little diversity in the generated samples. The conditional distribution is nearly a delta function, suggesting that the sea surface height is very informative of the ocean state and that a functional relationship can be found between the sea surface height and sub-surface currents. In other words, the present inference task is leaning towards Equation 3.

The sixth column shows the absolute value of the difference between the OcS and the AI mean, and the seventh column shows the ensemble standard deviation of the AI method. We represent predictive uncertainty by using the generated fields' standard deviation at each grid point. The largest standard deviation is associated with the northern convecting region, especially for vertical velocity. We surmise that the sea surface height alone is insufficient to constrain the pointwise vertical velocity magnitude in these regions. The second largest source of uncertainty is the eddying gyre region, which contributes to the uncertainty of the velocity components. The mean discrepancy column and the AI Std column have similar spatial structures, suggesting that the method can identify regions in which it is less capable of prediction.

b. Ocean State Predictions and Uncertainties with Degraded Surface Height Input

Next, we check the ability to predict the ocean interior upon losing sea surface height information. We do this by applying a filter to the sea surface height where we average neighboring points and provide this filtered free surface as an input to the generative model. We use the same neural network model for this inference task as it was simultaneously trained on all coarse-grained fields. We show the resulting statistical inferences for the zonal velocity in Figure 4 and quantitative errors as a function of depth and coarse-graining in Figure 6.

We first focus on Figure 4. In the first column, we show the sea surface height at a 200 kilometer resolution in latitude (32×32 number of points) in the top, with a 800 kilometer resolution (8×8 points) in the second row and coarse grain by a factor of two in each subsequent row to 1,600 kilometers and 3,200 kilometer resolutions in latitude, respectively. Although the network predicts all the velocity components and buoyancy at each vertical level, we restrict the visualization to
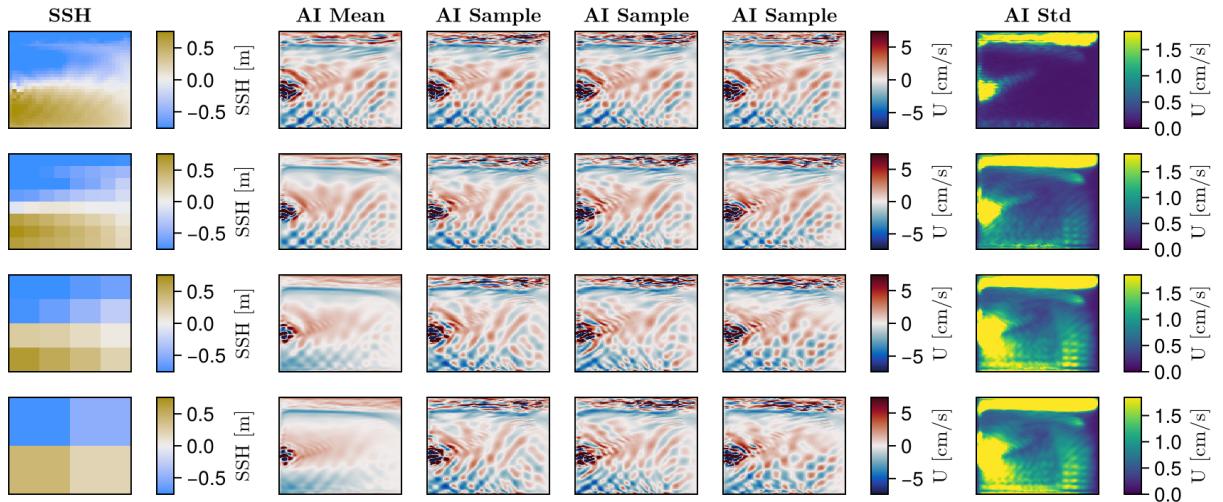
FIG. 4. Statistical inference of zonal velocity given the different resolutions of the free surface. Given the sea surface height at several coarse-grained resolutions (first column), we show the output of the generative AI model ensemble in the subsequent columns. We show the average value over 100 samples predicted by the neural network (second column) and representative samples (third, fourth and fifth columns). The pointwise standard deviation of the generative AI ensemble is in the last column. The latitudinal resolutions of SSH are, from top to bottom, 200 kilometers, 800 kilometers, 1,600 kilometers, and 3,200 kilometers.

the zonal velocity at a 400 meter depth. The outputs can be directly compared to the first row of Figure 3. The generative model produces 100 samples for each coarse-graining level. The second column shows the ensemble average zonal velocity, and columns three through five show randomly selected samples. We see by inspection that the ensemble average departs progressively the samples as we proceed downwards through the rows and that the samples become increasingly diverse. In addition, we show the point-wise standard deviation of the 100-member ensembles in the last column to quantify the diversity of the generated samples.

These observations suggest that an SSH coarse-grained beyond $O(1,000)$ kilometers is no longer informative since the generated samples become so diverse as to reproduce statistics of the data distribution, a "worst case" scenario for the generative method. As we coarse-grain the free surface (lose information), the inference task moves from something informative (similar to Equation 3) towards the uninformative (similar to Equation 4) marginal distribution of the training data.
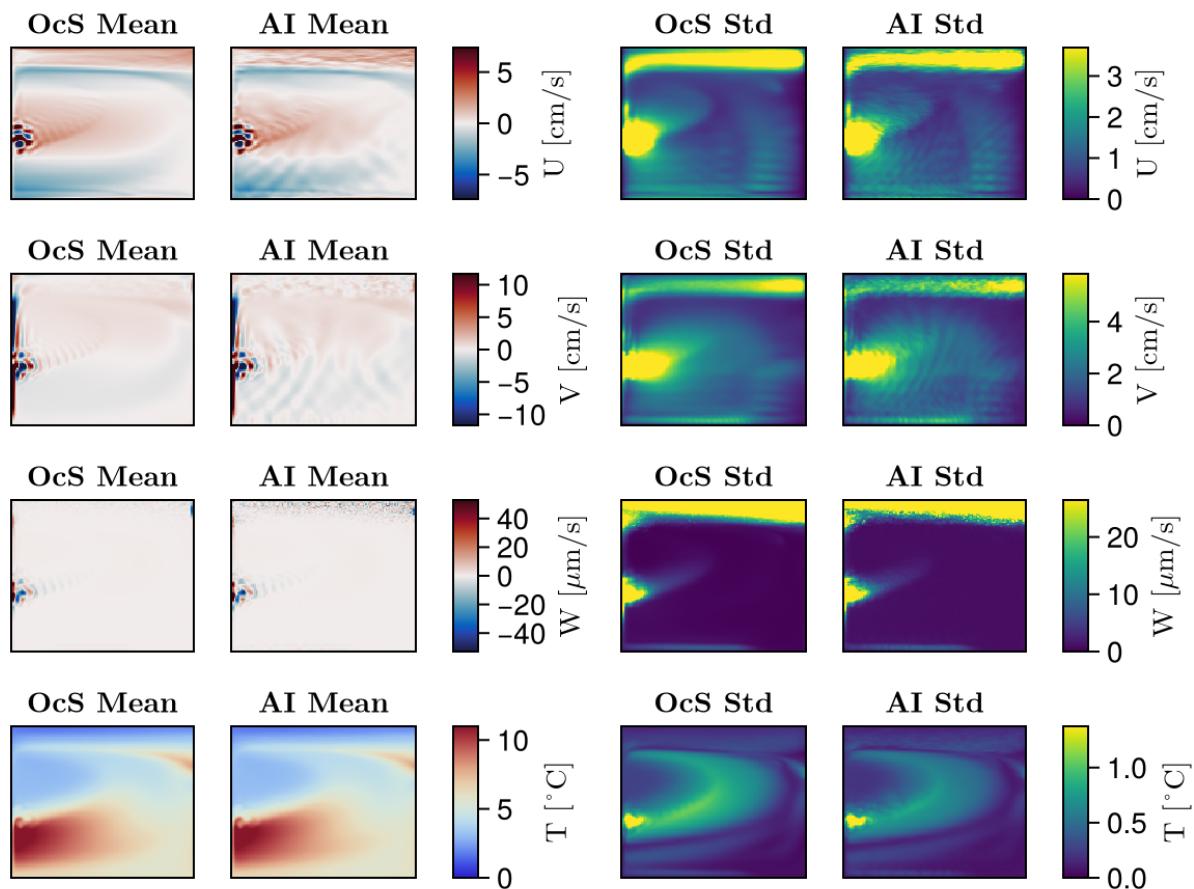
FIG. 5. Comparison of the AI ensemble mean to the OsC temporal mean. Using fully coarse-grained SSH as inpute to the network, we generate 100 samples and calculate point-wise statistics. We compare these ensemble statistics to the temporal analog of the ocean simulation over the training period.

We compare the sampled distribution using the most uninformative (i.e. fully coarse-grained over the entire domain) SSH as the conditional information to the marginal distribution of the data over the training period in Figure 5. This corresponds to the maximally uninformative case and the goal is to show that uninformative data reduces to the drawing samples similar to the training data. To do so we compare the ensemble statistics of the generative model to the temporal statistics of the training data. For brevity we use a fixed depth. The rows correspond to the different fields U, V, W, T at a 400 meter depth. The OcS mean/std is a temporal mean/std over the training data set and the AI mean/std is generated from a 100 member ensemble using only the domain average

16

free-surface height as input, i.e. we remove spatial structure from the input field. We see that there are still wave-like oscillations present in the AI predictions which we attribute to insufficient averaging from the limited generated samples since using less ensemble members exacerbated the oscillations, but overall there seems to be good agreement between the two calculations.
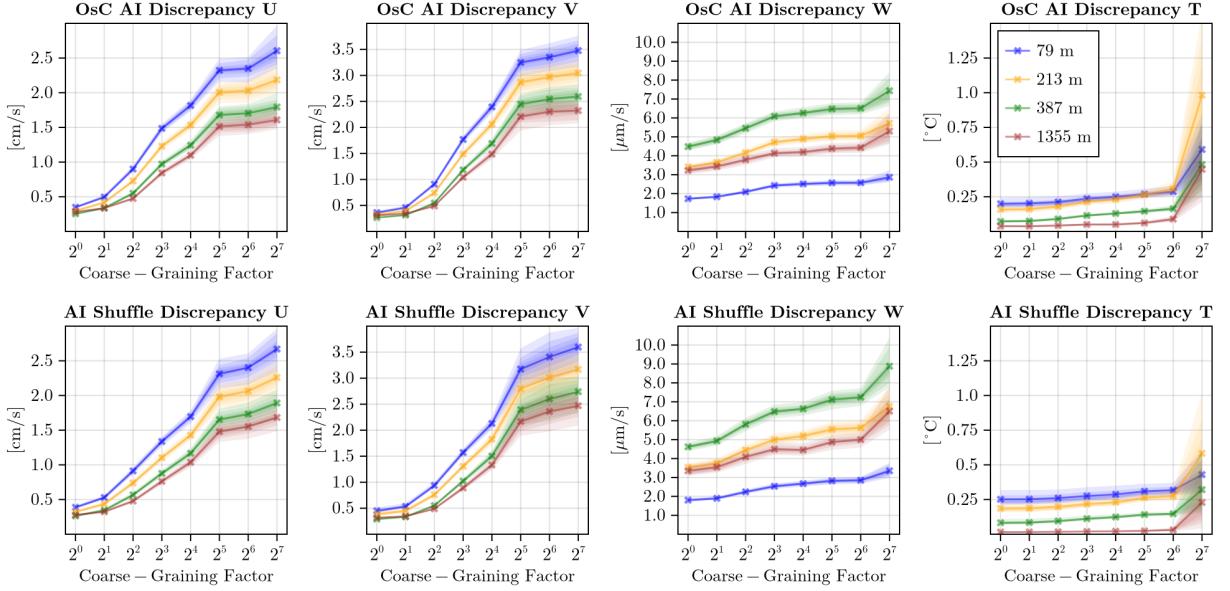
FIG. 6. Discrepancy as a function of depth and coarse-graining for a fixed SSH. We show the distribution calculated by the difference in the neural network prediction 100 member ensemble and the ocean simulation with the L^1 norm in the first row. In the second row we show a purely AI calculation of error given by taking differences between ensemble members and using the L^1 norm. The columns represent the different fields, U, V, W, T. Within a panel the different colors represent different depths, and the horizontal axis corresponds to different coarse-graining factors of SSH with 2^0 representing the full resolution (50 kilometers) and 2^7 representing an average value of SSH (6,400 kilometer resolution).

In Figure 6, we quantify the discrepancy between the ensemble average of the AI prediction and the ocean simulation in predicting the various fields as a function of resolution and depth. We take the difference between the OcS simulation and each ensemble member individually. We use the L^1 norm (average absolute error) to illustrate the discrepancy and the resulting distribution with different bands whose contours represent the $0.6, 0.7, 0.8$, and 0.9 quantiles, from darker to lighter. The points are the ensemble average of the resulting distribution. We denote this as the "OsC AI

17

Discrepancy". We also perform an analogous computation purely from the AI ensemble, which we call the "AI Shuffle Discrepancy" in the second row. Here, we plot the distribution given by the difference between different ensemble members. Specifically, we take the L^1 norm between ensemble members 1 and 2, 2 and 3, 3 and 4, and so forth to ensemble members 100 and 1, yielding 100 samples of an L^1 norm, which we then take to be a distribution to compare against the distribution given by the "OsC AI Discrepancy". A similar calculation is to take ensemble member 1 as a "fake" ground truth, and take the difference between the other ensemble members 2 through 100. We choose the "shuffled" version to reduce variation due to a "unlucky" choice of ground truth ensemble member.

We see that (for all fields), as we decrease the resolution of the sea surface height, the discrepancy between generated fields and the OsC increases. The average discrepancy of the U and V velocity field components decreases as a function of depth since their magnitudes decrease with depth. The temperature field T has the opposite trend, with the largest average discrepancies coming from the surface, perhaps due to the larger overall temperature values near the ocean's surface. The vertical velocity exhibits a less clear trend, with the largest average discrepancy coming from a 400-meter depth and the lowest coming from the near-surface value. A potential explanation lies in the fact that $w = 0$ at the surface and the seafloor, so the largest value should come from somewhere at mid-depth.

The trends of the first row are quantitatively reflected in the second row. The "Shuffle Discrepancy" is similar to the difference between the simulation output and the AI ensemble. This observation suggests that the AI method can provide meaningful uncertainty estimates in so far as the data it was trained on is similar in distribution to what is seen in the future. For example, we see a difference between the 1355-meter depth "OsC AI" temperature discrepancy and the corresponding "AI Shuffle Discrepancy" where the AI Shuffle Discrepancy underpredicts the associated error. We attribute this to the lack of equilibrium of the ocean simulation depths, as reflected in Figure 2.

c. Heat Flux Predictions and Uncertainties with Degraded Surface Height Input

Thus far, we have only examined statistics within a given field but have not looked at correlations across fields. We address this by calculating the vertical and meridional eddy heat flux. We do not

show the full heat flux because its value depends on arbitrary temperature units. Also, we use the zonal averages of the fluxes for illustrative purposes. Mathematically, we use

$$\bar{\cdot} \overset{\text{def}}{=} \frac{1}{60°} \int_0^{60°} \cdot \, d\lambda, \tag{5}$$

$$\text{vertical eddy heat flux} = \cos(\varphi) \left(\overline{wT} - \overline{w}\,\overline{T} \right), \tag{6}$$

$$\text{meridional eddy heat flux} = \cos(\varphi) \left(\overline{vT} - \overline{v}\,\overline{T} \right). \tag{7}$$
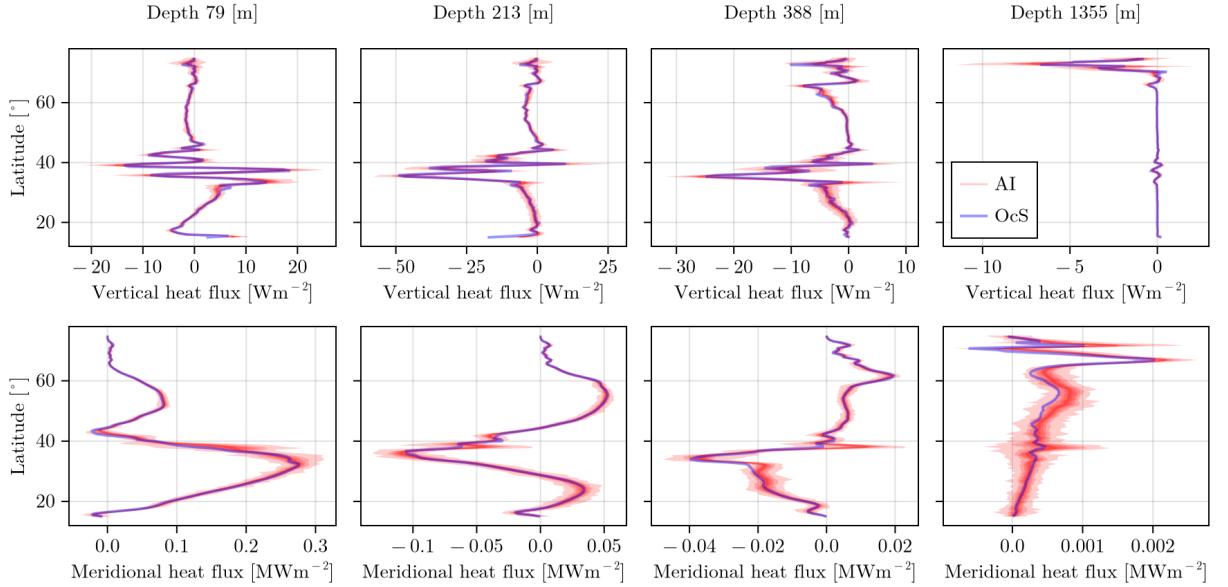
FIG. 7. Zonal average of vertical and horizontal eddy heat flux for a fixed SSH at various depths. We show the zonal average of the vertical (top row) and meridional eddy heat flux (bottom row) over the domain for each latitude at a fixed moment time and various depths in blue. The neural network results using the full-resolution sea surface height for input are in red, with the solid red line representing the ensemble average heat flux. The ribbons represent the 0.6, 0.7, 0.8, and 0.9 quantiles, from darker to lighter, as computed with the 100-member ensemble. Each column shows a different depth, and the relative uncertainty (red ribbon region) increases as we go deeper into the domain.

We show the heat fluxes as a function of latitude at various depths (in each column) in Figure 8 for the highest resolution SSH. The "ground truth" results are the blue lines, and the generative AI
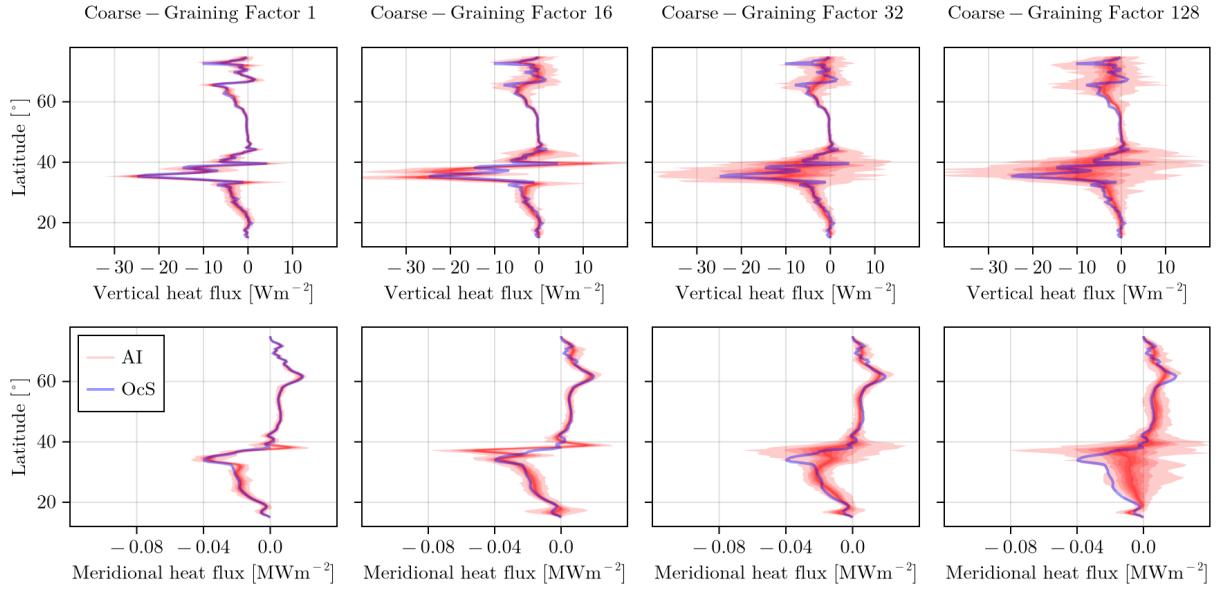
FIG. 8. Zonal average of vertical and horizontal eddy heat flux for a fixed SSH at a 400 meter depth. We show the zonal average of the vertical (top row) and meridional heat flux (bottom row) over the first fifteen degrees of the domain for each latitude at a fixed moment time and a depth of 400 meters in blue. The neural network results are in red, with the solid red line representing the ensemble average heat flux. The ribbons represent the $0.6, 0.7, 0.8,$ and 0.9 quantiles, from darker to lighter. Each column uses different levels of coarse-grained sea surface height. As the amount of coarse-graining increases, the confidence bounds (as represented by the red region) increase and reduce to the marginal distribution of the data distribution). Information about the location of the meander is lost when going from a factor of 8 reduction to a factor of 16 in reduction of resolution. See the first column and third row of Figure 4 for the corresponding SSH.

results with 100 samples are shown as ribbon plots that capture the $0.6, 0.7, 0.8, 0.9$ quantiles from darker to lighter shadings of red. The top row is the vertical heat flux, and the bottom row is the meridional heat flux, both as a function of latitude. Generally, we see that the generative model predicts the heat flux well for all cases, with the ribbons increasing in relative size with depth: the meridional heat flux at the lowest shown depth is not (relatively) as well constrained.

We also calculate the eddy heat fluxes as a function of SSH resolution at a 400 meter depth in Figure 8. Similar to the previous plot, the top row is the vertical heat flux, the bottom row is the meridional heat flux, the "ground truth" results are the blue lines, and the generative AI results with

100 samples are shown as a ribbon with $0.6, 0.7, 0.8, 0.9$ quantiles from darker to lighter shadings of red. The heat flux prediction grows in uncertainty as we coarse-grain the sea surface height. The loss of information here is understood upon examining the first column of Figure 4. As the sea surface height (SSH) field is increasingly coarse-grained, we lose precise information about the location of the jet separation point — the region where the western boundary current detaches from the wall and begins to meander eastward. This separation point is critical in organizing mesoscale eddy activity and lateral heat transport. When spatial detail in the SSH is removed, the model can no longer resolve where the jet begins to meander, leading to increased uncertainty in the predicted eddy heat fluxes, particularly in the downstream region where the current interacts with ambient stratification.

d. Statistics Combining Various Depths

Understanding how well generative models predict subsurface ocean dynamics at different depths is crucial for assessing their viability in state estimation and climate modeling. Next we assess the generative model's ability to aggregate statistics across different vertical levels. Specifically, to evaluate the model's performance, we compare the predictions of the generative model against the ocean simulation for correlations across depths, the zonally and vertically integrated meridional heat flux, and the meridional overturning stream function in depth coordinates. We do this for the same 50-year future SSH used in previous sections.

Figure 9 quantifies the vertical correlation of the subsurface fields under various inference settings. Here we use the correlation coefficient C. This quantity for the meridional velocities at two different depths V_1 and V_2 is calculated as

$$C(V_1, V_2) = \frac{\mathbb{E}\left[(V_1 - \mathbb{E}V_1)(V_2 - \mathbb{E}V_2)\right]}{\sqrt{\mathbb{E}\left[(V_1 - \mathbb{E}V_1)^2\right]} \sqrt{\mathbb{E}\left[(V_2 - \mathbb{E}V_2)^2\right]}} \tag{8}$$

where the expected value is over the latitude-longitude points (i.e., a horizontal average without the metric terms), at a fixed depth, and for a fixed ensemble member (when appropriate). A similar calculation is performed for other quantities. For ensemble-based comparisons, results are shown as ribbon plots, with shading corresponding to the 0.6, 0.7, 0.8, and 0.9 quantiles (from dark to light).

21
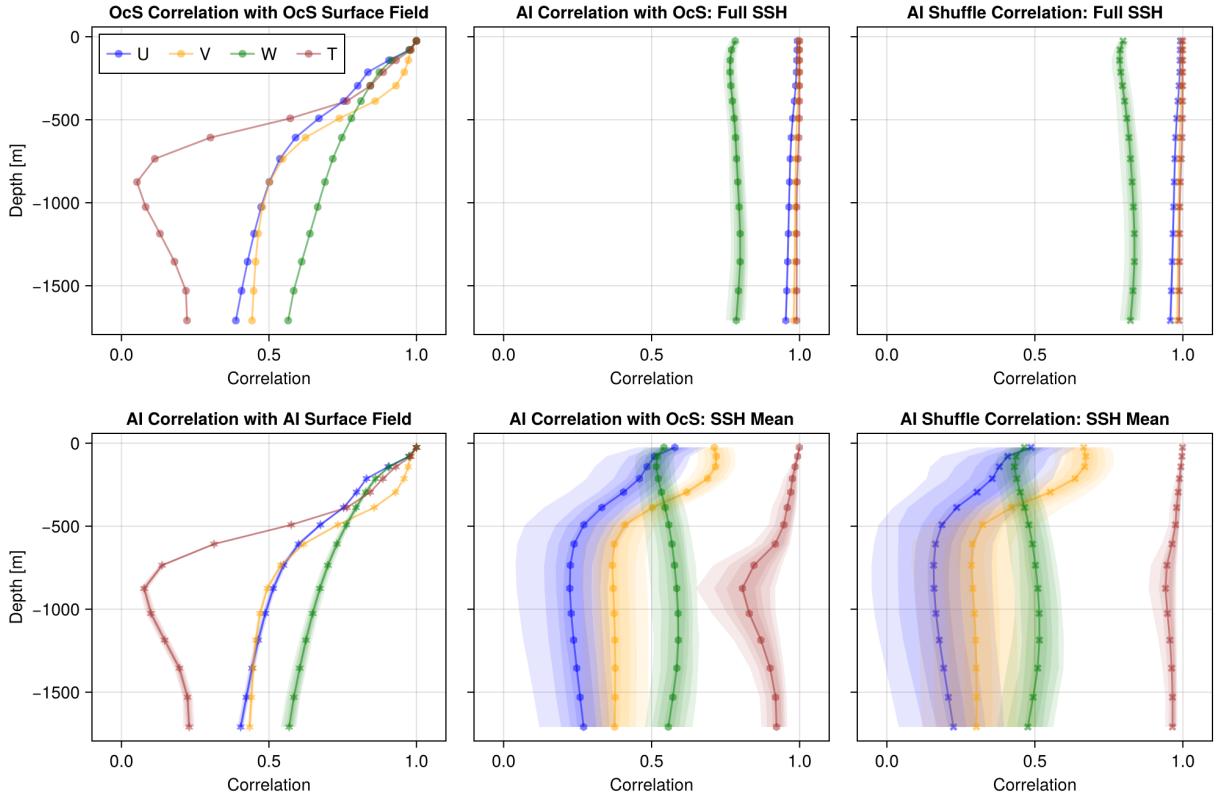
FIG. 9. Correlation coefficients. Left column: correlations with the surface field in OcS (top) and the generative AI outputs (bottom). Middle column: cross-correlations between AI-predicted fields and OcS profiles, using full-resolution SSH (top) and domain-averaged SSH (bottom) as conditioning inputs. Right column: ensemble-based correlations computed purely from the AI model for full-resolution SSH (top) and domain-averaged SSH (bottom).

The left column of Figure 9 shows the correlation between each field and its surface value, computed from the OcS simulation (top) and the generative AI model (bottom). These panels confirm that the generative model reproduces the vertical coherence present in the simulation. Furthermore, we note that the simulation deviates from equivalent barotropic behavior due to the changing correlation coefficient. The middle column compares AI-generated fields to the OcS profiles across depth, conditioned on either full spatial-resolution SSH (top) or mean (i.e. no spatial structure) SSH (bottom). The stronger correlations in the top row indicate that the fine-scale structure in SSH provides a meaningful constraint on subsurface variability. The existence

Zonal and Vertically Integrated Meridional Heat Flux
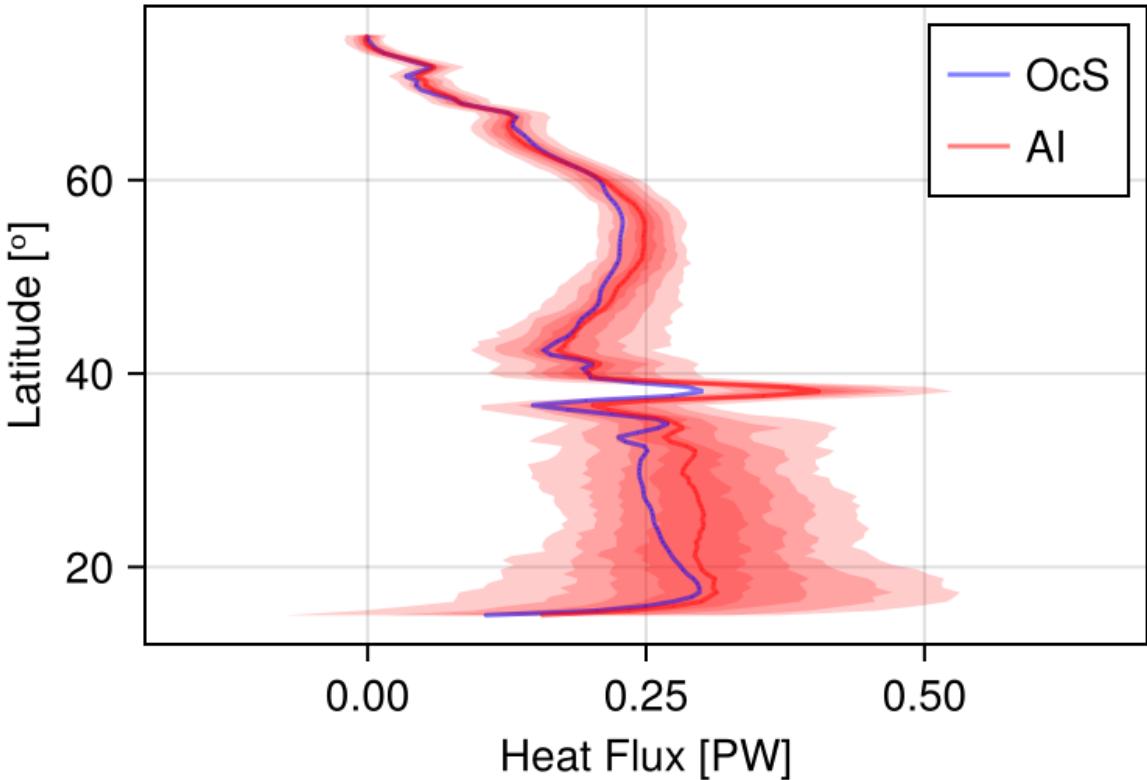
FIG. 10. Zonal and vertically averaged heat transport. The instantaneous transport as a function of latitude is estimated from the model simulation (blue line) and with the AI model using only sea surface height information (red line). The pink shading represents the $0.6, 0.7, 0.8$, and 0.9 quantiles, from darker to lighter.

of *some* level of correlation in the lower row reflects the similarity between past ocean states and future ocean states as given by the simulation. After all the fields are not "arbitrary" but rather related to the past flow states of the model, e.g. regions in the past with no flow will remain with no flow. The right column presents ensemble-based correlations among AI samples conditioned on identical SSH inputs (top: full-resolution; bottom: no-resolution), serving as a "shuffle test" analogous to the discrepancy analysis in Figure 6. These correlations largely mirror the patterns seen in the middle column, indicating internal consistency in the generative model's outputs. A

23

notable exception is seen in the temperature field at depth, where correlations drop significantly—a result we attribute to the lack of statistical equilibrium in the abyssal ocean.

We show in Figure 10 the generative AI's ability to reproduce the zonally and vertically integrated meridional heat flux,

$$\text{Meridional Heat Flux} = \rho c_p R \cos(\varphi) \int_{-1800\,m}^{0\,m} \int_{0°}^{60°} V(\lambda, \varphi, z, t) T(\lambda, \varphi, z, t) \, d\lambda dz. \tag{9}$$

where R is the radius of Earth, c_p is the heat capacity of water, and ρ_0 is the reference density. The red line (ensemble mean of the generative AI prediction) and the resulting (from dark to light) ribbons represent the $0.6, 0.7, 0.8$, and 0.9 quantiles ribbons correctly encapsulate the "ground truth" ocean simulation prediction. This quantity is evaluated 50 years in the future, similar to the previous sections. The uncertainty is quite large, yet the ensemble mean prediction is similar to ground truth. The uncertainty can likely be reduced by improving the model by better accounting for the vertical structure. Furthermore, the generative method need not be non-divergent, which may affect the resulting statistics.

Lastly, we calculate the meridional overturning stream function,

$$\psi(\varphi, z, t) \stackrel{\text{def}}{=} R \cos(\varphi) \int_{z}^{0} \int_{0°}^{60°} V(\lambda, \varphi, z', t) d\lambda dz' \tag{10}$$

Figure 11 shows the resulting overturning stream function calculation. The top left figure is the ocean simulation output, the middle top figure is the mean of the 100 generated samples, and the top right is the difference. The bottom left and bottom middle panels are arbitrarily chosen samples, and the AI uncertainty (bottom right) is the standard deviation of the 100 generated samples.

We see that the overall structure in the upper 400 meters of the domain is well captured, with errors increasing with depth. Furthermore, we see a larger discrepancy at lower latitudes and mid-depth. The largest standard deviation of the AI ensemble is at low latitudes and at depth. The lack of statistical equilibrium contributes to the error here. The largest error comes from the pool of cold water slowly finding its way along the seafloor, from northern latitudes towards the lower latitudes. This process takes time, and was not shown to the neural network during training. The samples of the generative AI method appear noisier than physically plausible, suggesting that a 3D
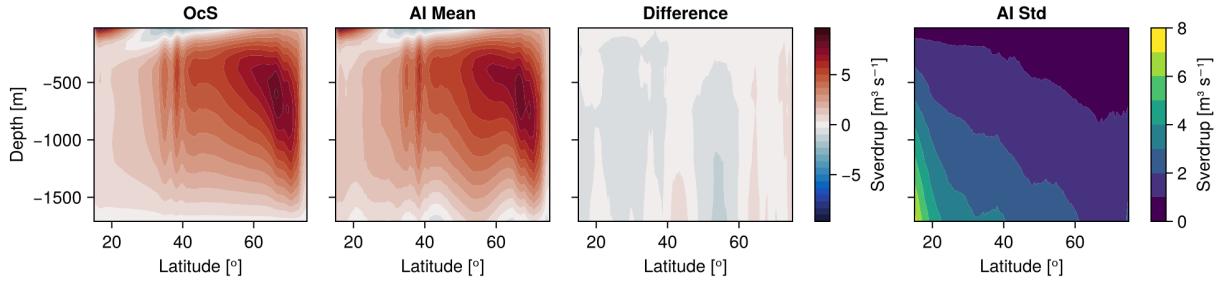
Fɪɢ. 11. The instantaneous overturning stream function in depth coordinates. Here, we show the overturning stream function at each vertical level z and latitude for a fixed moment in time. The ocean simulation is compared to the AI mean and samples. The difference between the ocean simulation and the AI mean is shown in the top right. The pointwise standard deviation of the 100 generated samples is in the bottom right.

UNet would be a better architecture for the present case. Regardless, the resulting variability is overestimated.

4. Conclusions

This study has demonstrated the potential of generative modeling as a statistical framework for inferring the interior state of the ocean from surface observations. We explored the ability of diffusion-based generative models to reconstruct subsurface velocity and buoyancy field statistics using only information about the free surface in a statistically non-equilibrated system, offering a probabilistic approach to state estimation. Our key findings can be summarized as follows:

1. We have demonstrated, in an idealized setting, that ocean state estimation can be formulated as a conditional generative modeling problem, where score-based diffusion models sample from the distribution of subsurface velocity and buoyancy fields given surface data. This approach captures inherent uncertainties in subsurface reconstructions, providing plausible subsurface fields consistent with the given information.

2. The generative AI method accurately reconstructs key ocean circulation features—including mesoscale eddies, zonal jets, eddy-heat flux, and overturning circulations—across different depths, dynamical regimes, and surface data resolutions. Notably, the generative approach

25

provided meaningful uncertainty estimates, particularly in eddying and convective regions, and correctly estimated its own predictive inaccuracy.

3. While the model performs well in regions where surface observations strongly constrain subsurface dynamics, its predictive skill diminishes at greater depths and in areas where the ocean state is out-of-sample (because the ocean simulation is drifting over time) from ocean states similar to what has been seen previously.

Our results demonstrate that generative models might capture the ocean interior's statistical variability, even when conditioned on coarse-resolution surface data. However, decreasing surface observations' spatial resolution or increasing inference depth diminishes predictive skill. In the limiting case, when the surface signal is insufficient to constrain the interior, the generative model reverts to sampling from the marginal distribution of the training data. As shown in Figure 6, the most pronounced deterioration in performance occurs when the surface resolution is degraded from 100km (coarsening factor 2^2) to 200km (2^3), particularly for the horizontal velocity components where the eddy length scale is on the order of 100km. For temperature, predictive accuracy declines when large-scale gradient information is lost. These findings suggest that improvements in satellite resolution—from current AVISO-class missions (200km) to SWOT (50km)—could significantly enhance our ability to infer subsurface ocean dynamics (Ballarotta et al. 2019; Fu et al. 2024). Incorporating additional dynamical priors, such as quasi-geostrophic constraints or hybrid physics-informed machine learning architectures, may further improve the fidelity and robustness of generative reconstructions.

Future work will extend this approach to more complex ocean simulations that include seasonal cycles, bathymetry, and finer-scale turbulence, providing additional constraints on the generative model. Furthermore, integrating observational datasets with multiple modalities such as sea surface salinity, Argo float profiles, sea surface temperature, and velocity measurements from moorings will enable direct evaluation of the model's performance in real-world ocean conditions, (Martin et al. 2025). Measurement uncertainty must also be accounted for. One avenue for doing so would be to draw from a distribution of consistent free-surface heights, effectively sampling from $\mathcal{P}(\text{subsurface}|\text{surface})\mathcal{P}_s(\text{surface})$ where $\mathcal{P}_s(\text{surface})$ incorporates uncertainty in the surface field. Incorporating theoretical predictions as additional "input" for the generative model will also be

explored. This way, we can use a theory as a "physics-informed" guess, further refined through the generative method and a numerical ocean simulation with higher fidelity physics.

The methods developed in this study contribute to the broader effort of integrating machine learning with physical modeling in climate science. By leveraging advances in generative AI and ocean state estimation, this work provides a foundation for a new approach to data assimilation, parameterization, and statistical inference in ocean modeling. Such methodologies can also be applied to atmospheric geophysical models as well as Earth system models to provide a probabilistic view of multi-scale dynamics.

Data availability statement. Scripts for reproducing the figures and the neural network training are located at `https://github.com/sandreza/DoubleGyreInference`. The Oceananigans script for producing simulations is located at `https://github.com/CliMA/Oceananigans.jl/blob/as/z-star-coordinate/double_gyre_5.jl`. We developed a Jax code for implementing the deep learning method `https://github.com/sandreza/JaxDiffusion`.

References

Anderson, B. D., 1982: Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, **12 (3)**, 313–326.

Ballarotta, M., and Coauthors, 2019: On the resolutions of ocean altimetry maps. *Ocean Science*, **15 (4)**, 1091–1109, https://doi.org/10.5194/os-15-1091-2019.

Batzolis, G., J. Stanczuk, C.-B. Schönlieb, and C. Etmann, 2021: Conditional image generation with score-based diffusion models. *arXiv preprint*, **2111.13606**.

Bischoff, T., and K. Deck, 2024: Unpaired downscaling of fluid flows with diffusion bridges. *Artificial Intelligence for the Earth Systems*, **3 (2)**, e230 039.

Bolton, T., and L. Zanna, 2018: Applications of deep learning to ocean data inference and subgrid parameterization. 376 - 399 pp., https://doi.org/10.1029/2018MS001472.

Champenois, B., and T. Sapsis, 2024: Machine learning framework for the real-time reconstruction of regional 4d ocean temperature fields from historical reanalysis data and real-time satellite and buoy surface measurements. *Physica D: Nonlinear Phenomena*, **459**, 134 026, https://doi.org/https://doi.org/10.1016/j.physd.2023.134026.

Chapman, C., and A. A. Charantonis, 2017: Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geoscience and Remote Sensing Letters*, **14 (5)**, 617–620.

Cooper, M., and K. Haines, 1996: Altimetric assimilation with water property conservation. *Journal of Geophysical Research: Oceans*, **101 (C1)**, 1059–1077, https://doi.org/https://doi.org/10.1029/95JC02902, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/95JC02902.

Cox, M. D., and K. Bryan, 1984: A numerical model of the ventilated thermocline. *Journal of Physical Oceanography*, **14 (4)**, 674 – 687, https://doi.org/10.1175/1520-0485(1984)014⟨0674:ANMOTV⟩2.0.CO;2.

Cromwell, D., A. Shaw, P. Challenor, R. Houseago-Stokes, and R. Tokmakian, 2007: Towards measuring the meridional overturning circulation from space. *Ocean Science*, **3 (2)**, 223–228.

Deck, K., and T. Bischoff, 2023: Easing color shifts in score-based diffusion models. *arXiv preprint arXiv:2306.15832*.

DelSole, T., and D. Nedza, 2022: Reconstructing the atlantic overturning circulation using linear machine learning techniques. *Atmosphere-Ocean*, **60 (5)**, 541–553.

Dheeshjith, S., A. Subel, A. Adcroft, J. Busecke, C. Fernandez-Granda, S. Gupta, and L. Zanna, 2024: Samudra: An ai global ocean emulator for climate. URL https://arxiv.org/abs/2412.03795, 2412.03795.

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch, 2015: Ecco version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, **8 (10)**, 3071–3104, https://doi.org/10.5194/gmd-8-3071-2015.

Fu, L.-L., and Coauthors, 2024: The surface water and ocean topography mission: A breakthrough in radar remote sensing of the ocean and land surface water. *Geophysical Research Letters*, **51 (4)**, e2023GL107 652, https://doi.org/https://doi.org/10.1029/2023GL107652, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL107652.

George, T. M., G. E. Manucharyan, and A. F. Thompson, 2021: Deep learning to infer eddy heat fluxes from sea surface height patterns of mesoscale turbulence. *Nature Communications*, **12 (1)**, 800, https://doi.org/10.1038/s41467-020-20779-9.

Giorgini, L. T., T. Bischoff, and A. N. Souza, 2025a: Kgmm: A k-means clustering approach to gaussian mixture modeling for score function estimation. URL https://arxiv.org/abs/2503.18054, 2503.18054.

Giorgini, L. T., K. Deck, T. Bischoff, and A. Souza, 2024: Response theory via generative score modeling. *Phys. Rev. Lett.*, **133**, 267 302, https://doi.org/10.1103/PhysRevLett.133.267302.

Giorgini, L. T., F. Falasca, and A. N. Souza, 2025b: Predicting forced responses of probability distributions via the fluctuation-dissipation theorem and generative modeling. URL https://arxiv.org/abs/2504.13333, 2504.13333.

Hasselmann, K., 1976: Stochastic climate models part i. theory. *Tellus*, **28 (6)**, 473–485, https://doi.org/https://doi.org/10.1111/j.2153-3490.1976.tb00696.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2153-3490.1976.tb00696.x.

Held, I. M., R. T. Pierrehumbert, S. T. Garner, and K. L. Swanson, 1995: Surface quasi-geostrophic dynamics. *Journal of Fluid Mechanics*, **282**, 1–20, https://doi.org/10.1017/S0022112095000012.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

Hurlburt, H. E., 1986: Dynamic transfer of simulated altimeter data into subsurface information by a numerical ocean model. *Journal of Geophysical Research: Oceans*, **91 (C2)**, 2372–2400, https://doi.org/https://doi.org/10.1029/JC091iC02p02372, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JC091iC02p02372.

Hyvärinen, A., and P. Dayan, 2005: Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6 (4)**.

Karras, T., M. Aittala, S. Laine, and T. Aila, 2022: Elucidating the design space of diffusion-based generative models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS '22.

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

Lapeyre, G., and P. Klein, 2006: Dynamics of the upper oceanic layers in terms of surface quasigeostrophy theory. *Journal of Physical Oceanography*, **36 (2)**, 165 – 176, https://doi.org/10.1175/JPO2840.1.

Liu, L., H. Xue, and H. Sasaki, 2019: Reconstructing the ocean interior from high-resolution sea surface information. *Journal of Physical Oceanography*, **49 (12)**, 3245 – 3262, https://doi.org/10.1175/JPO-D-19-0118.1.

Liu, S., and S. Hu, 2024: Seeing through the sea with satellites: Reconstructing ocean subsurface temperature and salinity with satellite observations. *Earth System Science Data Discussions*, **2024**, 1–29, https://doi.org/10.5194/essd-2024-334.

Martin, S. A., G. E. Manucharyan, and P. Klein, 2023: Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. *Journal of Advances in Modeling Earth Systems*, **15 (5)**, e2022MS003 589, https://doi.org/https://doi.org/10.1029/2022MS003589, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003589.

Martin, S. A., G. E. Manucharyan, and P. Klein, 2025: Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. *Journal of Advances in Modeling Earth Systems (JAMES)*, 8688.

Meng, S., A. L. Stewart, and G. Manucharyan, 2024: Circumpolar transport and overturning strength inferred from satellite observables using deep learning in an eddying southern ocean channel model. *Journal of Advances in Modeling Earth Systems*, **16 (9)**, e2024MS004 262.

Meng, Y., E. Rigall, X. Chen, F. Gao, J. Dong, and S. Chen, 2021: Physics-guided generative adversarial networks for sea subsurface temperature prediction. *IEEE transactions on neural networks and learning systems*, **34 (7)**, 3357–3370.

Pedersen, C., L. Zanna, and J. Bruna, 2025: Thermalizer: Stable autoregressive neural emulation of spatiotemporal chaos. URL https://arxiv.org/abs/2503.18731, 2503.18731.

Ramadhan, A., and Coauthors, 2020: Oceananigans.jl: Fast and friendly geophysical fluid dynamics on gpus. *Journal of Open Source Software*, **5 (53)**, 2018, https://doi.org/10.21105/joss.02018.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 234–241.

Roquet, F., G. Madec, T. McDougall, and P. Barker, 2015: Accurate polynomial expressions for the density and specific volume of seawater using the teos-10 standard. *Ocean Modelling*, https://doi.org/10.1016/j.ocemod.2015.04.002.

Saenko, O. A., D. Yang, and J. M. Gregory, 2018: Impact of mesoscale eddy transfer on heat uptake in an eddy-parameterizing ocean model. *Journal of Climate*, **31 (20)**, 8589 – 8606, https://doi.org/10.1175/JCLI-D-18-0186.1.

Silvestri, S., G. L. Wagner, J.-M. Campin, N. C. Constantinou, C. N. Hill, A. Souza, and R. Ferrari, 2024a: A new weno-based momentum advection scheme for simulations of ocean mesoscale turbulence. *Journal of Advances in Modeling Earth Systems*, **16 (7)**, e2023MS004 130, https://doi.org/https://doi.org/10.1029/2023MS004130, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023MS004130.

Silvestri, S., and Coauthors, 2024b: A gpu-based ocean dynamical core for routine mesoscale-resolving climate simulations. *Authorea, Inc.*, https://doi.org/10.22541/essoar.171708158.82342448/v1.

Solodoch, A., A. L. Stewart, A. McC. Hogg, and G. E. Manucharyan, 2023a: Machine learning-derived inference of the meridional overturning circulation from satellite-observable variables in an ocean state estimate. *Journal of Advances in Modeling Earth Systems*, **15 (4)**, e2022MS003 370.

Solodoch, A., A. L. Stewart, A. McC. Hogg, and G. E. Manucharyan, 2023b: Machine learning-derived inference of the meridional overturning circulation from satellite-observable variables in an ocean state estimate. *Journal of Advances in Modeling Earth Sys-*

tems, **15 (4)**, e2022MS003 370, https://doi.org/https://doi.org/10.1029/2022MS003370, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003370.

Song, Y., and S. Ermon, 2019: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, **32**.

Song, Y., and S. Ermon, 2020: Improved techniques for training score-based generative models. *Advances in neural information processing systems*, **33**, 12 438–12 448.

Souza, A. N., and Coauthors, 2020: Uncertainty quantification of ocean parameterizations: Application to the k-profile-parameterization for penetrative convection. *Journal of Advances in Modeling Earth Systems*, **12 (12)**, e2020MS002 108, https://doi.org/https://doi.org/10.1029/2020MS002108, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002108.

Stanczuk, J., G. Batzolis, T. Deveney, and C.-B. Schönlieb, 2023: Your diffusion model secretly knows the dimension of the data manifold. 2212.12611.

Su, Z., J. Wang, P. Klein, A. F. Thompson, and D. Menemenlis, 2018: Ocean submesoscales as a key component of the global heat budget. *Nature Communications*, **9 (1)**, 775, https://doi.org/10.1038/s41467-018-02983-w.

Vincent, P., 2011: A connection between score matching and denoising autoencoders. *Neural computation*, **23 (7)**, 1661–1674.

Wagner, G. L., and Coauthors, 2025a: Formulation and calibration of catke, a one-equation parameterization for microscale ocean mixing. *Journal of Advances in Modeling Earth Systems*, **17 (4)**, e2024MS004 522, https://doi.org/https://doi.org/10.1029/2024MS004522, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2024MS004522.

Wagner, G. L., and Coauthors, 2025b: High-level, high-resolution ocean modeling at all scales with oceananigans. URL https://arxiv.org/abs/2502.14148, arXiv:2502.14148 [physics.ao-ph], https://doi.org/10.48550/arXiv.2502.14148, 2502.14148.

Wang, J., G. R. Flierl, J. H. LaCasce, J. L. McClean, and A. Mahadevan, 2013: Reconstructing the ocean's interior from surface data. *Journal of Physical Oceanography*, **43 (8)**, 1611 – 1626, https://doi.org/10.1175/JPO-D-12-0204.1.

Yan, H., H. Wang, R. Zhang, J. Chen, S. Bao, and G. Wang, 2020: A dynamical-statistical approach to retrieve the ocean interior structure from surface data: Sqg-meof-r. *Journal of Geophysical Research: Oceans*, **125 (2)**, e2019JC015 840, https://doi.org/https://doi.org/10. 1029/2019JC015840, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019JC015840.

Zhang, J., P. Ning, X. Zhang, X. Wang, and A. Zhang, 2023: Deriving sea subsurface temperature fields from satellite remote sensing data using a generative adversarial network model. *Earth and Space Science*, **10 (4)**, e2022EA002 804.