Deep Learning of Systematic Ocean Model Errors in a Coupled GCM from Data Assimilation Increments

Tarun Verma[1], F. Lu[2], A. Adcroft[2], L. Zanna[3], A. Gnanadesikan[1]

[1]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ 08542, USA
[3]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

Key Points:

- Neural networks can predict spatiotemporal patterns of data assimilation increments using only local state variables.
- The skill of neural networks exceeds that of the state-independent climatological benchmark in the upper ocean.
- The predictability of upper ocean increments from fluxes and vertical gradients highlights biases in ocean mixed layer representation.

Corresponding author: Tarun Verma, `tarunverma.geos@gmail.com`

Abstract

We present a novel, data-driven approach to predict systematic model errors in the ocean component of a coupled general circulation model leveraging deep learning and data assimilation. We examine the skill of the proposed scheme in learning systematic model errors, including their spatial patterns, variance, scales, and test its sensitivity to different predictors and neural network architecture. The scheme utilizes local state variables such as ocean temperature, salinity, velocities, and surface fluxes to predict corrections to temperature tendency for the upper 1000 meters in the ocean on daily timescales. The performance is evaluated on the withheld test dataset and compared against the empirical climatological temperature corrections that are geographically dependent. The performance is depth-dependent, with significant improvements over the benchmark in the upper 20 meters in the ocean. It degrades rapidly with depth but remains comparable to the climatology benchmark. Neural networks can capture up to 40−50% of the daily variance in temperature increments in the upper 20 meters relative to the benchmark's 20%. The improvements are associated with networks predicting finer spatiotemporal scales than the benchmark. They are expected to perform better in reducing surface ocean mixed layer bias than previously used techniques. Despite being column-local without geographical inputs, networks can sufficiently reproduce spatial patterns on daily and longer timescales. The patterns consist of corrections to regional dynamical features such as western boundary currents, equatorial undercurrents, bathymetry-related corrections in the Southern Ocean, and warm surface increments over subtropical and midlatitude belts.

Plain Language Summary

The ocean is a complex system, and we use ocean general circulation models to study it. However, these models are imperfect and have errors in representing the subgrid-scale processes. We present a new way to correct these errors using deep learning and data assimilation. This method uses information about the ocean thermodynamic state, such as temperature, velocity, and surface fluxes, to predict and correct errors. We found that the new method performs better in the upper 20 meters of the ocean and captures a significant fraction of daily corrections to the temperature equation. This new method can help us reduce bias in the upper ocean mixed layer.

1 Introduction

Climate models, when used for climate predictions and projections, often exhibit systematic differences from the real world, wherein 'systematic' implies that the discrepancies are persistent rather than random over time. These systematic discrepancies are often called model drift or model bias, and they can manifest in forms of both fast model dynamics and physics, as well as slow climatological equilibrium. The sea surface temperature (SST) bias pattern is an example of a systematic error that is persistent across different generations and configurations of climate models (Farneti et al., 2022). SST bias is particularly detrimental, as it affects processes across the climate system. For example, it impacts climate sensitivity via SST-Cloud feedback (Hyder et al., 2018) , tropical cyclone density via surface heat fluxes and vertical shear in the tropical atmosphere (Vecchi et al., 2014), North American precipitation by altering large-scale atmospheric flow (Johnson et al., 2020), and the arctic amplification (Wu et al., 2023).

Much like the inaccurate parameter values in Lorenz63 or incomplete representation of sub-grid scale term in two-scale Lorenz96 models that lead to systematic errors in the evolution of the respective systems (Chen et al., 2015; Arnold et al., 2013), climate models develop biases partially due to numerical errors from discretization and truncation, and parameterization-related errors such as inaccurate and missing subgrid-scale parameterizations.

Reducing model bias is a priority of various climate modeling and prediction centers across the world (Fox-Kemper et al., 2021). The fundamental way to reduce model bias would require improvements in the model structure. Besides correcting the model structure, numerous bias correction methods have been developed for climate model applications such as prediction and projection. Some examples include flux adjustment for coupled climate model simulations (Robert et al., 1997), diagnostic lead-time-dependent bias correction for prediction post-processsing (Kirtman et al., 2014; Nadiga et al., 2019), and prognostic bias correction in the form of tendency adjustment for weather and climate prediction (Lu et al., 2020; Chang et al., 2019), all of which are state-independent and climatological in nature, and typically correct some persistent spatiotemporal patterns associated with the bias without explicit dependence on any specific structural deficiency in the model. For example, the Ocean Tendency Adjustment (OTA) method from Lu et al. (2020) uses grid-dependent climatology of data assimilation (DA) increments to prognostically correct temperature and salinity tendencies in an ocean component of the NOAA Geophysical Fluid Dynamics Laboratory's Seamless System for Predictions and EArth system Research (GFDL's SPEAR) model. It has been successfully implemented in NOAA GFDL's SPEAR-ocean data assimilation (ODA) and experimental real-time seasonal prediction systems, significantly reducing climatological model drift and improving forecasts of ENSO, Arctic and Antarctic sea ice (Bushuk et al., 2021, 2022), atmospheric rivers (Tseng et al., 2021), extratropical baroclinic waves (G. Zhang et al., 2021), and extreme events (Jia et al., 2023, 2024).

Despite the success of OTA in reducing the ocean model bias, there are drawbacks to the OTA procedure. First, the OTA corrections, by construction, can only capture the seasonally varying climatology of the DA increments. Second, the corrections are fixed on the SPEAR model grid, which could limit the method's capability to generalize to other modeling frameworks. Last but not least, the climatological DA increments highly depend on the spatial and temporal coverage of the assimilated datasets, particularly Argo floats (Wong et al., 2020). Therefore, the OTA corrections may be subject to sampling errors over locations or periods less frequented by Argo floats. To mitigate the sparsity of subsurface ocean observations, a seasonal climatology of DA increments is computed to increase sampling sizes for each model location and average out random variations not related to systematic model bias. These random variations are a result of a) significant subgrid-scale variations that are present in the observations but are not resolved in the ocean model, b) deterministic chaos which may result in errors, even on resolved scales, due to initial condition sensitivity, c) representational uncertainty (mismatch between the model grid and the observational points), for example, an observed mixed layer of 12.5 meters would be represented as either too shallow or too deep if the vertical resolution of the model is five meters in the upper ocean, and d) impacts of systemic biases, such as depth of the mixed layer or location of a boundary current on variability. For example, a location with a systematically shallow mixed layer will exhibit a response that is too large to transient warming and cooling events on subseasonal time scales. The neural network approach presented here may partially capture the effects of the last two factors discussed.

The availability of efficient optimization algorithms and fast computation has recently spurred interest in using machine learning (ML) to improve existing subgrid-scale parameterizations and develop new data-driven parameterizations. The rationale behind this push is that many subgrid-scale processes are complex, nonlinear, and involve multi-scale interactions, and can not be adequately described by low dimensional empirical and analytical relationships as in traditional parameterizations. Therefore, a high dimensional nonlinear model such as neural networks (NNs) could provide benefits over the traditional approach. These methods require a large amount of data for training. They may also need some physically relevant quantities that may not be directly observed in the physical world, so much so that a higher fidelity, higher resolution numerical simulation is almost always used for training machine learning models instead of the actual observations. Rasp et al. (2018), Yuval and O'Gorman (2020) and Brenowitz and Bretherton (2018) are a few recent examples of studies parameterizing deep convection in the atmosphere using cloud-resolving model outputs. They all attempt to build nonlinear mappings from spatially coarsened state variables to sub-grid scale fluxes to develop data-driven parameterizations. (Bolton & Zanna, 2019) and Guillaumin and Zanna (2021) used a similar coarsening approach to parameterize sub-grid mesoscale momentum fluxes in the surface ocean, with the latter using the state-of-the-art high-resolution ($1/10^o$) climate simulation and predicting both deterministic and stochastic parts.

ML applications in weather and climate modeling can also utilize real-world observations directly or indirectly. For example, Holder and Gnanadesikan (2023) train Random Forest on satellite-derived observations to predict phytoplankton biomass in the ocean. ML weather forecast models (Pathak et al., 2022; Lam et al., 2023; Arcomano et al., 2020) are trained on reanalysis datasets, which are the data assimilation products that combine numerical models with real-world observations, e.g., ERA5 (Hersbach et al., 2020). Direct use of observational data is seen in models like MetNet-3, which employs weather station data for training and evaluation (Andrychowicz et al., 2023). Similar approaches have also been applied to the ocean, where the historical lack of subsurface and long-term data is a challenge for training. The introduction of Argo floats over the past 20 years has improved in situ observations down to 2000 meters. However, their spatial and temporal coverage remains insufficient to characterize the multi-scale variability in the ocean. An alternative approach is to use ocean reanalysis datasets for training, but they themselves are inherently uncertain due to limited observations. Further, the long timescale variability in the ocean (days to multidecadal) relative to the atmosphere (hours to days) would require longer data records for training. In short, the combination of data sparsity and short data records in the ocean makes it challenging to train ML models for oceanic applications without the help of dynamic general circulation models.

In this study, we choose an alternate approach that uses real-world observations with the help of dynamic models, specifically the DA corrections or increments, to learn the state-dependent ocean component bias in the SPEAR coupled climate model. Instead of learning sub-grid scale fluxes, we directly target the difference between the model and the observed state through cycled ocean DA experiments. The DA increments act as a proxy of the fast errors that eventually lead to model drift, and could be linked to deficiencies in model parameterizations (Rodwell & Palmer, 2007). In other words, we plan to build a ML-enhanced version of the OTA bias correction scheme that makes state-dependent predictions of the tendency adjustment terms. Similar approaches have been tested in the atmosphere (Watt-Meyer et al., 2021; Chapman & Berner, 2024) and sea ice components (Gregory et al., 2023, 2024). Watt-Meyer et al. (2021) nudged a low-resolution atmosphere model to an observational analysis and used the nudging tendencies to train state-dependent ML models that can predict corrective tendencies for atmosphere temperature, specific humidity and horizontal winds. Gregory et al. (2023) and Gregory et al. (2024) use the increments from a sea ice DA system to train convolutional neural networks (CNN) that can predict skillful sea ice concentration increments, and apply such CNNs to reduce sea ice bias in SPEAR coupled climate simulations.

The following section 2 details the data and neural network problem formulation and training, followed by a rationale for learning DA increments in section 3. We then quantitatively summarize the skill of neural networks on the withheld-test dataset in section 4. Then we summarize mean and daily patterns of predictions in section 5 followed by its temporal characteristics in section 6. The broader implications of the results are discussed in section 7. We finally summarize the findings and conclude in section 8.

2 Data and Methods

We aim to build a state-dependent model of systematic ocean DA increments for the upper thousand meters that can either correct model errors in the MOM6 ocean component of a free-running SPEAR coupled simulation or serve as a bias correction scheme for the seasonal to decadal prediction system within SPEAR. To achieve this, we are employing a neural network-based approach. This section details the datasets, supervised learning problem formulation, various design choices, training procedure, evaluation, and lessons learned. We use Python's PyTorch library to accomplish this.

2.1 Dataset

Data for developing the state-dependent model comes from simulations using NOAA GFDL's SPEAR model. SPEAR is the current modeling system at GFDL that enables a wide range of climate research and operations, including large ensemble simulations (Delworth et al., 2020), seasonal prediction (Lu et al., 2020) as part of the North America MultiModel Ensemble (NMME), subseasonal prediction (Xiang et al., 2022), as well as decadal (Yang et al., 2021) and sea ice prediction (Bushuk et al., 2021, 2022) through international inter-comparison programs. SPEAR consists of the AM4.0/LM4.0 atmosphere and land models (Zhao et al., 2018), and the MOM6/SIS2 ocean and sea ice models (Adcroft et al., 2019). SPEAR includes models of various resolutions that can be selected to best suit the needs and computational capacity of specific applications. In this study, we use the SPEAR-LO model, in which the atmosphere/land resolution is about $100km$ and the ocean/sea ice resolution is about 1° with tropical refinement to 1/3°.

SPEAR ocean data assimilation (SPEAR-ODA) system was developed to facilitate the experimental prediction efforts at GFDL. It provides both an experimental ocean analysis product and the oceanic initial conditions for SPEAR seasonal predictions. SPEAR-ODA uses the Ensemble Adjustment Kalman Filter (EAKF) algorithm and a daily assimilation window. For this study, we only assimilate gridded daily OISST and Argo data since other data sources such as XBT (eXpendable BathyThermographs) or tropical moorings have very uneven spatial or temporal coverage. Full details of SPEAR-ODA can be found in Lu et al. (2020), including description of the DA increments. The OTA bias correction scheme implemented in the operational real-time SPEAR seasonal predictions takes the seasonal cycle of the SPEAR-ODA increments and applies them prognostically in the coupled climate predictions. This capability is made possible by the unique design of the SPEAR-ODA analysis, where the ocean DA is performed in the coupled SPEAR model without any direct observational constraint of the atmosphere component. The success of OTA in coupled climate predictions points to the possibility that a ML-enhanced state-dependent version of OTA can also be applied to coupled climate model predictions and projections.

2.2 Problem Formulation

We use fully connected dense layer architecture-based neural networks to develop a low-dimensional, column-local, and nonlinear mapping from state variables to temperature increments. The term "column-local" implies that non-locality is explicitly considered in the vertical direction. The horizontal gradient terms implicitly include the non-locality in the horizontal direction. Learning model errors from data is challenging as there is always a risk of learning i) geographical patterns instead of the underlying physics of model errors and ii) propagated errors rather than localized subgrid scale errors. Both of these situations can adversely affect the neural network's ability to generalize across different locations, times, models, spatial resolutions, and timescales. In this work, we avoid using explicit geographical information and instead rely on instantaneous ocean state and boundary flux fields as inputs to address the first issue. The second issue is partially addressed by building column-local models and modeling the increments on fast

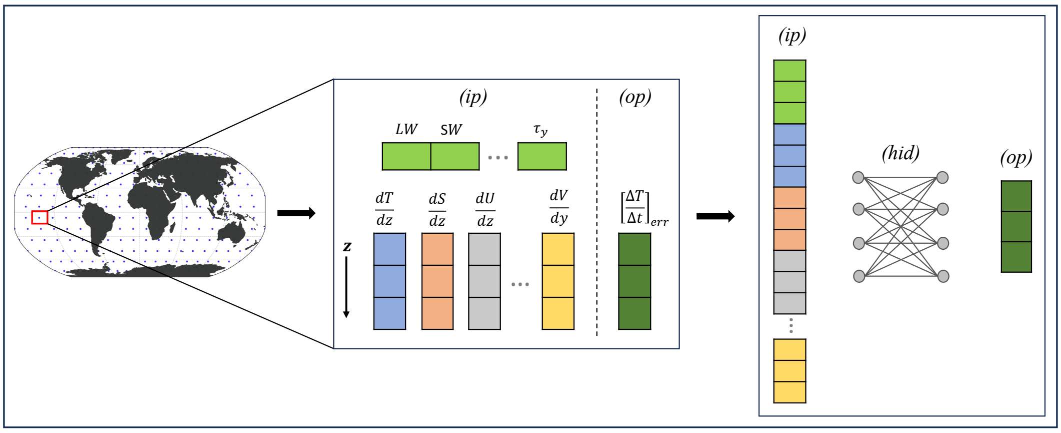
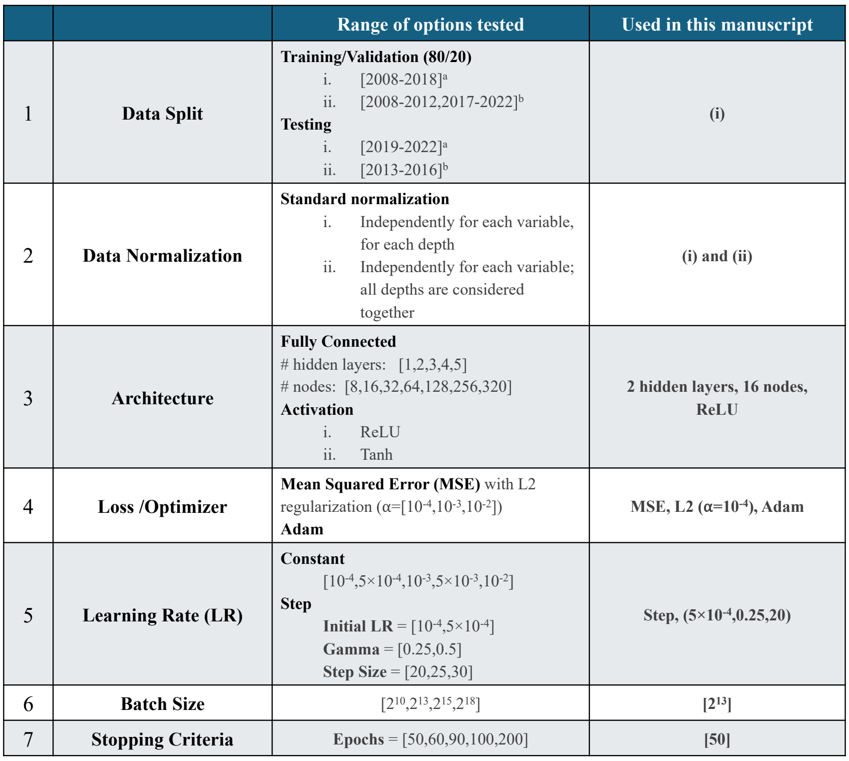


Figure 1. A schematic summarizing the supervised learning problem of modeling systematic ocean model errors in terms of column-local state variables. The dataset is curated by collecting data on 2^o subsampled ocean model grid across the globe and on the 3-day time frequency spanning 2008 to 2022. The vertical extent of the upper 1000 meters in the ocean is non-uniformly sampled in 51 levels. The feature and target pair consists of a combination of gradients of ocean temperature (T), salinity (S), zonal and meridional velocities (U and V) along with the radiative, heat and momentum fluxes as inputs (ip) and corresponding vertical profiles of temperature increments as outputs (op) of the neural network. Different features are stacked together to generate one long feature vector as an input to a fully connected dense layer architecture neural network. Every depth level is independently standard normalized (i.e., subtracting the sample mean and dividing by the sample standard deviation) for each of the input variables.

timescales of the DA cycles, thereby limiting error propagation across space, time, and processes. The column-local approach offers an additional advantage of reducing the network size, thus reducing the inference's computational cost. One caveat of this approach is that the DA increments contain corrections to both model and numerical errors and could also be corrupted by errors from other Earth system components in the coupled model. We utilize surface fluxes and gradients of ocean state variables (ocean stratification, vertical and horizontal velocity shears) as inputs to learn, to an extent, generalizable physical relationships and capture the subgrid-scale errors.

Figure 1 illustrates the machine learning workflow schematically. The workflow consists of curating vertical profiles of the gradients of state variables and surface fluxes as inputs and corresponding vertical profiles of temperature increments as outputs of the neural network from across the globe. The vertical profiles span from the sea surface down to 1000 meters deep in the ocean in about 51 discrete levels. We use the gradients of scalars such as ocean temperature (T), salinity (S), and zonal and meridional components of the velocity vector (U, V). The surface fluxes include net longwave and shortwave fluxes, latent and sensible heat fluxes, and momentum fluxes. The data is subsampled up to 2^o horizontal spacing and 3-day temporal frequency to reduce the computational cost of training and testing networks of different sizes and inputs. Each year, there are about 26.9 million daily samples on the native grid; however, the subsampling process reduces this by a factor of 20, significantly cutting down on training time. As the model output is on a staggered horizontal grid, we ensure that different variables are collocated in space, targeting local physical errors and not numerical artifacts. The feature vectors are stacked into a single vector before being fed into the fully connected neural network.

Table 1. Table summarizing different training choices and parameters tested, as well as one that is presented in this manuscript.



		Range of options tested	Used in this manuscript
1	**Data Split**	**Training/Validation (80/20)** i. [2008-2018][a] ii. [2008-2012,2017-2022][b] **Testing** i. [2019-2022][a] ii. [2013-2016][b]	**(i)**
2	**Data Normalization**	**Standard normalization** i. Independently for each variable, for each depth ii. Independently for each variable; all depths are considered together	**(i) and (ii)**
3	**Architecture**	**Fully Connected** # hidden layers: [1,2,3,4,5] # nodes: [8,16,32,64,128,256,320] **Activation** i. ReLU ii. Tanh	**2 hidden layers, 16 nodes, ReLU**
4	**Loss /Optimizer**	**Mean Squared Error (MSE)** with L2 regularization (α=[10^{-4},10^{-3},10^{-2}]) **Adam**	**MSE, L2 (α=10^{-4}), Adam**
5	**Learning Rate (LR)**	**Constant** [10^{-4},5×10^{-4},10^{-3},5×10^{-3},10^{-2}] **Step** **Initial LR** = [10^{-4},5×10^{-4}] **Gamma** = [0.25,0.5] **Step Size** = [20,25,30]	**Step, (5×10^{-4},0.25,20)**
6	**Batch Size**	[2^{10},2^{13},2^{15},2^{18}]	**[2^{13}]**
7	**Stopping Criteria**	**Epochs** = [50,60,90,100,200]	**[50]**

2.3 Training

The neural network training involves exploring various hyperparameters, such as learning rate, batch size, regularization rate, and additional factors, including the size of the neural network, methods for feature and label normalization, training and testing periods, and data subsampling. We will outline the training procedure, explain our rationale, and share lessons learned. An example of a canonical network, showcasing the choices and parameters used, is presented in Table 1.

We utilize a fully connected neural network architecture that employs a ReLU (Rectified Linear Unit) nonlinear activation function. For our loss function, we use mean squared error and select the Adam optimizer for training. The daily global data is subsampled up to 2^o horizontal grid and 3-day temporal frequency. The training period covers the years 2008 to 2018, during which 20% of the randomly shuffled grid points from around the globe are set aside for validation ($\sim$ 3 million samples), while the remaining 80% are used for training ($\sim$ 12 million samples). Additionally, the independent test period spans from 2019 to 2022 with total number of samples $\sim$ 5.5 million. Even though features and labels exhibit non-stationarity in time, the use of different training and testing periods does not affect the general results and conclusions presented in this study.

We trained different sizes of fully connected neural networks with a number of hidden layers ranging from one to five and a number of nodes ranging from 8 to 320. We find that for most combinations of input predictors, a neural network with two hidden layers with sixteen nodes in each layer is sufficient to outperform our benchmarks with little to no overfitting. We employ L2 (or ridge) regularization with a rate ranging between 1×10^{-2} to 1×10^{-4} across all our neural networks to further reduce any overfitting. Some example training and validation learning curves for networks using six sur-

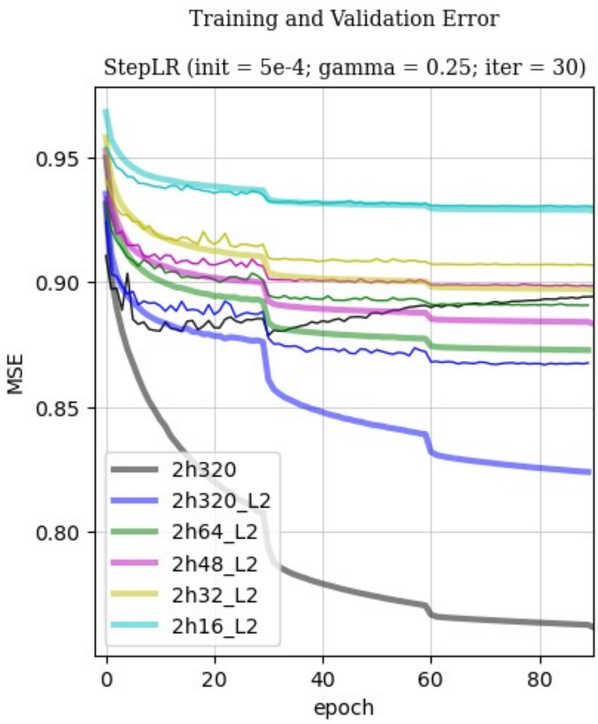


Figure 2. Training (bold) and validation (light) mean squared loss as a function of training epoch. Different colors represent different neural network sizes, specified as 2hN, where 2h denotes two hidden layers and N denotes the number of nodes in the two hidden layers.

face flux variables, four vertical gradient profiles of T,S,U and V and four horizontal gradient profiles of U and V are shown in Figure 2.

We examined the sensitivity to different batch sizes and ultimately selected a batch size of 2^{13}. This choice is influenced by two factors, i) the memory of the GPU node, and ii) the size of the input feature vector size which ranges between 51 and 618. We randomly shuffled the mini-batches across both space and time to ensure that each mini-batch is representative of profiles from different seasons and regions of the ocean. After selecting the network and batch sizes, we experimented with a range of constant learning rates from 1×10^{-4} to 1×10^{-2}. We observed that a relatively small learning rate leads to a stable solution; however, it results in slow convergence. Conversely, a larger learning rate accelerates the convergence rate, but the solution may become unstable. An unstable solution is indicated by predicted spatial patterns that fluctuate significantly between successive training epochs, particularly in sparse data regions such as subsurface and polar latitudes. This inter-epoch variability complicates the process of selecting an optimized network, making it somewhat arbitrary.

We found that using an adaptive learning rate effectively addresses issues related to slow convergence and the stability of the solution. Our approach involves starting with an initial learning rate of 5×10^{-4}, which is then reduced to a quarter of its value every 20 to 30 epochs, repeating this process 3 to 4 times. Each time we decrease the learning rate, we observe a notable reduction in both training and validation errors, although the magnitude of this reduction becomes smaller with each successive adjustment. Additionally, the learning curves become smoother following these reductions. The optimality and stability of the solution were evaluated based on two factors. The first factor was the plateauing of the validation error curve, while the second was the standard deviation of validation errors across successive training epochs. We tested both a single standard deviation metric for the entire validation dataset and a metric where the standard deviation was projected onto latitude-depth space. This approach allowed us to assess the stability of the solution across different depths and latitudinal zones. One drawback of this approach is that the network's weights and biases must be saved to disk after every epoch. However, this is a minor inconvenience for the relatively small networks used in this study. Early stopping is commonly used in prior research to prevent overfitting. However, we did not utilize this approach because our neural networks are relatively small and already incorporate regularization techniques. Instead, we followed a standard stopping criterion, which involves halting training after 50 to 60 epochs.

We use two different normalization approaches for the inputs and outputs of the network. The first method involves applying standard normalization (i.e., subtracting the sample mean and dividing by the sample standard deviation) independently to each variable and depth. The second approach, however, standard normalizes each variable while considering all depths together. The latter approach preserves the vertical structure of the oceanic variable, unlike the first approach. The first approach results in a slightly better performance and is presented here, with the overall metrics based on the second approach added to the supporting information. Additionally, we have either tried transforming temperature increments into fluxes through vertical integration or weighting them with layer thickness prior to the normalization step. We find similar performances in each of these cases and have decided to omit the comparisons for brevity.

We compared the performance of neural networks trained on datasets sampled near Argo locations and surfacing times (referred to as "training in Argo space") with those trained in the model grid space. This comparison is motivated by the significant influence that Argo measurements have on subsurface temperature increments at these specific locations and times. However, despite this motivation, the networks trained in Argo space struggled to learn large-scale and long-term patterns when tested in the model grid space. The challenges may stem from the reduced size of the training data and an increased occurrence of dynamical noise correction relative to systematic correction in the Argo space. For this reason, we present results only for the model grid training in this study.

The learning curves in Figure 2 initially trend downward before saturating at a specific non-zero value. We could interpret this non-zero residual MSE as caused by the unpredictable part of the DA increments- the part the network could not learn based on the given inputs. The residual MSE accounts for random increments due to unresolved dynamical variations and the unpredictable systematic part, either because of the neural network's lack of expressive power, not knowing the relevant predictors, or insufficient observational sampling. We will not dwell on the predictability issue here, as this is the topic for another study, and instead focus on learning and interpreting predicted DA increments.

3 Why model Data Assimilation Increments?

Figure 3 (a) shows the SST bias pattern in free-running coupled climate simulation using the GFDL's SPEAR-LO model. The assimilation of gridded sea surface temperature and the Argo data on daily timescales significantly reduces global mean SST bias (Figure 3 (b)), as expected in ocean analysis products compared to the free-running model. Refer to Lu et al. (2020) to see the spatial and vertical structure of the bias reduction on assimilating ocean observations. This reduction in bias in the SPEAR-ODA system results from a series of daily corrections sequentially applied to the model state, specifically to temperature and salinity fields. The 16-year average, spanning 2003 to 2018, of such daily temperature corrections or increments at the sea surface is shown in Fig-

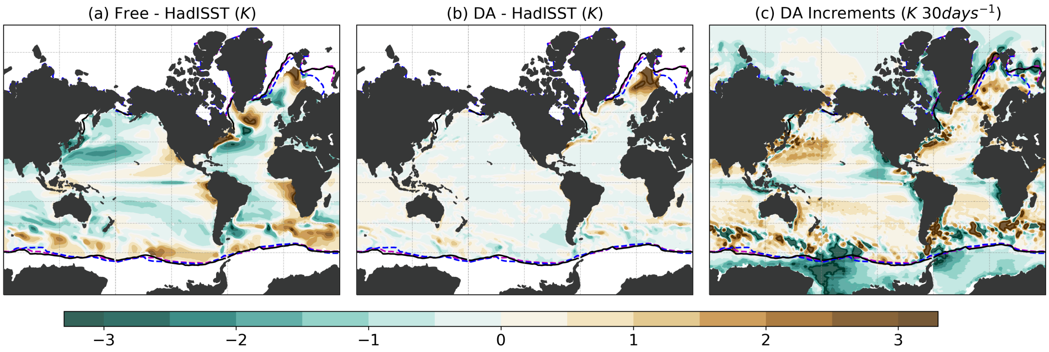


Figure 3. Sea surface temperature bias relative to Hadley centre global sea surface temperature (HadISST) record for 2003-2018 period for (a) free-running SPEAR simulation (Free), and (b) a SPEAR simulation with ocean data assimilation (DA). (c) 2003-2018 average sea surface temperature increments in the DA run. The colored contours in the polar regions represent the 2003-2018-mean 15 % sea ice concentration in HadISST (black), Free (blue), and DA (magenta) runs.

ure 3 (c). Regions with negative increments imply that the surface ocean in the model is biased warm on average for the short-term forecasts over the daily DA cycles, and positive increments mean that it is biased cold. The negative increments along the western coast of African and American continents correspond to the contemporary warm bias in climate models resulting from erroneous coastal upwelling and air-sea interactions. Notably, there is a prominent pattern of positive increments in the mid-latitude oceans along the western boundary current regions where large subgrid-scale variability exists due to processes that are not resolved in the coarse-resolution model like the one used in this study. In the Southern Ocean, alternating positive and negative increments extend zonally across all the longitudes. These corrections seem to be anchored to the ocean bathymetry, and alternating patterns imply that they are dependent on the local flow. Additionally, the Southern Ocean is also the region where coarse-resolution models can not explicitly represent subgrid-scale variability. Polar regions in both hemispheres are biased warm, perhaps an indication of low sea ice bias, resulting in negative increments except for the Greenland-Iceland-Norwegian seas. These mean increments are organized in large-scale patterns across the globe and reminiscent of the SST bias pattern shown in Figure 3(a). This high correspondence between the climatological bias and the mean increments based on fast error growth from the DA cycle indicates that the DA increments could be used to calculate spatially varying climatological correction tendencies. Lu et al. (2020) computed such climatological three-dimensional tendency fields for temperature and salinity and applied them prognostically in operational SPEAR seasonal predictions using the OTA procedure. OTA reduces the climatological drift in the ocean component of the coupled climate predictions, and improves the prediction skills across various processes such as ENSO. Additionally, Dee (2006) showed that bias in atmospheric GCMs can be corrected using the systematic components of the DA increments, which according to Rodwell and Palmer (2007), may also project onto subgrid-scale errors. This suggests that we can learn something about model errors from DA increments.

4 Offline Evaluation

The neural networks' overall performance is evaluated on a withheld test dataset spanning 2019 to 2022. Root mean squared error ($RMSE$) is the square root of the loss function that is minimized during the training. The coefficient of determination (R^2) is

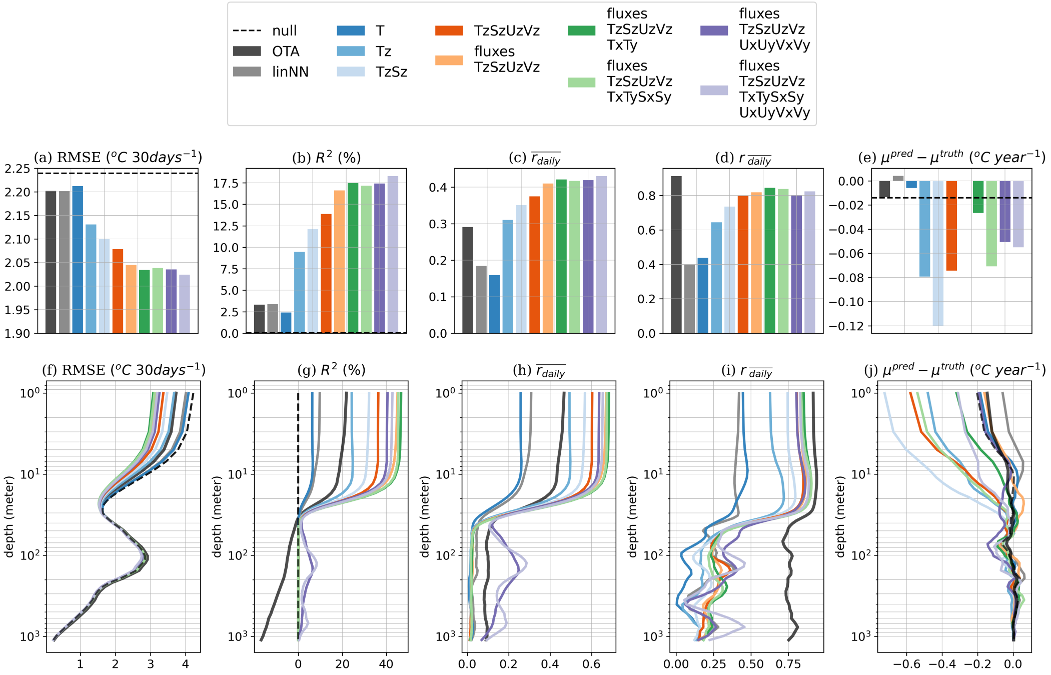


Figure 4. Overall (a-e) and depth-dependent (f-j) evaluation on the withheld test of the same size neural networks (2h16) using different input features (in colors) and comparison against the three benchmarks (in dashed line, black bar and grey bar). The performances are compared in terms of various metrics, including (a,f) root mean squared error, (b,g) coefficient of determination, (c,h) an average of daily pattern correlations between true and predicted fields, (d,i) the pattern correlation of the time-mean fields, and (e,j) differences between true and predicted means. The feature sets are indicated in the legend as concatenated strings. The letter 'z' indicates the vertical gradients of T (temperature), S (salinity), U (zonal velocity), and V (meridional velocity); similarly, the letters' x' and 'y' indicate zonal or meridional gradients, respectively. The term 'fluxes' indicates a collection of radiative, heat, and momentum fluxes at the ocean-atmosphere interface.

the fraction of variance (of the true labels) predicted by the network. Pattern correlations ($\overline{r_{daily}}$, $r_{\overline{daily}}$) characterize networks' ability to predict large-scale spatial patterns despite being column-local. We simultaneously care for solutions with lower $RMSE$, higher R^2, and higher pattern correlations. The three benchmarks used for comparison include a null model (which predicts the average of the training labels in the physical space or zero vector in a non-dimensional space), grid-dependent monthly climatology (or climatology benchmark), and a linear model (neural network without any nonlinear activation) in that order of complexity. The climatology benchmark, also referred to as OTA increments as described in Lu et al. (2020), is computed from the training dataset by linearly interpolating the monthly mean climatology onto the daily timescale at every grid point, and is also currently applied in the real-time SPEAR seasonal prediction system. The following subsections summarize networks' global space-time integrated performance, predictor dependencies, and depth variations in the physical space.

4.1 Global Metrics

Figures 4 (a-e) display the global space-time integrated metrics for the different neural networks. These networks have the same network architecture (two hidden layers with sixteen nodes) but different combinations of input predictors, as indicated by the concatenated string. The predictions are compared with two reference benchmarks - the climatology of the training increments (referred to as the climatology benchmark or OTA) and a linear neural network (linNN) without any hidden layer or non-linear activations. The dashed black line in Figures 4(a,b,e) corresponds to the null network that predicts the mean of the training labels in the physical space or, equivalently, zeros in the non-dimensional space. In non-dimensional space, it is expected to result in the mean squared error (MSE) of one, an R^2 of zero, and a zero bias. However, the non-dimensional MSE of greater than one (Supplementary Figure S1) indicates distributional shifts between the training and the test labels, associated with an increased variance in test labels relative to the training labels (not shown).

The RMSE loss decreases with the addition of new predictors (Figure 4(a)). The network with only temperature as input does not improve upon OTA or LinNN, while the use of vertical temperature gradient (T_z) makes a big difference. The RMSE is further decreased when adding additional input features, including vertical salinity gradient (S_z), vertical shear (U_z and V_z), and the fluxes. Beyond the local vertical gradients and fluxes, adding horizontal gradients provides less marginal improvement, which could benefit implementation since the vertical gradients and fluxes depend only on local variables. As the RMSE decreases on adding predictors, the R^2 increases, indicating that networks capture additional space-time variance (Figure 4(b)). Albeit small, the NNs show positive R^2 values with improvements ranging between 1-10% over the climatology benchmark. The relatively low R^2 values are due to subgrid-scale dynamical noise in temperature increments and significant representational errors associated with Argo profiles, which the network does not predict.

The two pattern correlation metrics based on the Pearson correlation statistics measure the similarity between the predicted and the actual three-dimensional fields of temperature increments. Figure 4(c) compares the mean of daily pattern correlations, revealing the degree of similarity between the predicted and the actual pattern on any given day. On average, most networks perform better than the climatology benchmark in predicting daily fields despite the relatively weak correlation (< 0.35) highlighting the importance of unpredictable noise. The second pattern correlation (Figure 4d) measures the similarity between time-averaged three-dimensional fields of actual and predicted increments. The high correlation (~ 0.9) for OTA is expected since the time-averaged increments for the training and test periods are sampled from the same underlying distribution. The correlations for various NNs reaches up to $\sim$ 0.8 compared to the upper limit from the climatology benchmark. The pattern correlation of 0.8 is noteworthy, given that the model is column-local and has no geographical inputs. All the NNs, as well as OTA have negative bias compared to the labels (Figure 4(e)). As mentioned earlier, the negative bias of OTA is likely caused by the shifting in the distribution of the increments between the training and testing periods due to low-frequency climate variability or changes in the Argo coverage. Such negative bias is amplified by the NNs, which are predicting increments with larger variance than OTA.

A linear neural network (linNN) is optimized using the stochastic gradient technique for comparison and provides a second benchmark. The linNN directly connects the input to the output layer, without intermediate hidden and non-linear activation layers. The input vector consists of six individual surface fluxes and four vertical gradients of T, S, U, and V. This is similar to a two-hidden layer neural network shown as a light orange bar in Figure 4 (a-e). All NNs, except the one using temperature profiles, perform better than the linNN. However, the linNN has lower MSE and higher R^2 than the climatology benchmark, indicating that part of the variance in the temperature increments is linearly predictable.

4.2 Depth Metrics

The depth-varying metrics (Figure 4(f-l)) help distinguish the performance of different neural networks over the different depths in the ocean. As evident from vertical R^2, the performance is coherent in the upper 20 meters in the ocean, with a sharp decline in R^2 below 20 meters, followed by an increase that peaks at 150 meters around the typical thermocline depth. The NNs with horizontal gradients show a $\sim 20\%$ increase in R^2 relative to the climatological benchmark in the top 20 meters and a $7-8\%$ increase at 150 meters, while other NNs show improvement mostly in the upper 20 meters only, indicating that the horizontal gradients are necessary for NN prediction skill around the thermocline. These subsurface improvements are concentrated around the equator in the thermocline layer, as discussed later in the section 5.4. At other depths, the amount of variance explained by neural networks is similar to the climatological benchmark. We can not see the aforementioned depth-dependence as clearly in the vertical profile of RMSE in the physical space as the shape of the standard deviation curve overwhelms it. The non-dimensional MSE (shown in supplementary Figure S1(f)), however, does show the depth-dependence more clearly. A close correspondence between vertical profiles of R^2 and $\overline{r_{daily}}$ also suggests that both these metrics are driven by predictions of fine-scale spatio-temporal variability that peaks at the surface and around 150 meters. On the other hand, the pattern correlation of predicted time-mean fields is always smaller than that of the climatological benchmark derived from the training dataset, which sets an upper limit on the predictions. In the following sections, we will choose 2 NNs for more detailed analysis, with NN1 (TzSzUzVz) including only column-local ocean variables, and NN2 (fluxesTzSzUzVzUxUyVxVy) additionally including surface fluxes and horizontal velocity gradients.

4.3 Spatial Metrics

Spatial maps evaluating the performance of NN1, NN2 and OTA for the upper 20 meters in the ocean are shown in Figure 5. While RMSE patterns for NNs and OTA look comparable (as RMSE is dominated by variability), the R^2 distinguishes the NNs from OTA, as NNs display widespread improvements of 20–30% over the low and mid latitudes except for small regions near the coasts and in the equatorial eastern Pacific. On the other hand, OTA displays large R^2 over the polar regions where NNs have low and even negative values, which suggests that NNs have difficulty learning increments over the poles. A possible explanation include lack (Argo) of observations to assimilate, and different dynamics including the impact of sea ice. The time correlation also displays similar patterns as R^2. The bias patterns, on the other hand, are proportional to the RMSE, thus the variance. The comparison with R^2 also suggests that some negative R^2 values are associated with NNs being unable to correctly predict the mean values, e.g., near coasts and the eastern equatorial Pacific.

5 Patterns of time-averaged and daily predictions

5.1 Average of the 2019-2022 test data

In this section, we compare the spatial maps of seasonal and annual averages of true and predicted fields over the three depth ranges, namely, 0-20 meters, 100-300 meters, and 700 to 1000 meters (Figure 6). We show DJF and JJA averages for the upper 20 meters and annual averages for two deeper layers. We compare predictions from two neural networks with the same number of hidden layers (2) and nodes (16), NN1 and NN2 as described in 4.2. The two networks differ in the input predictors used. NN1 is truly

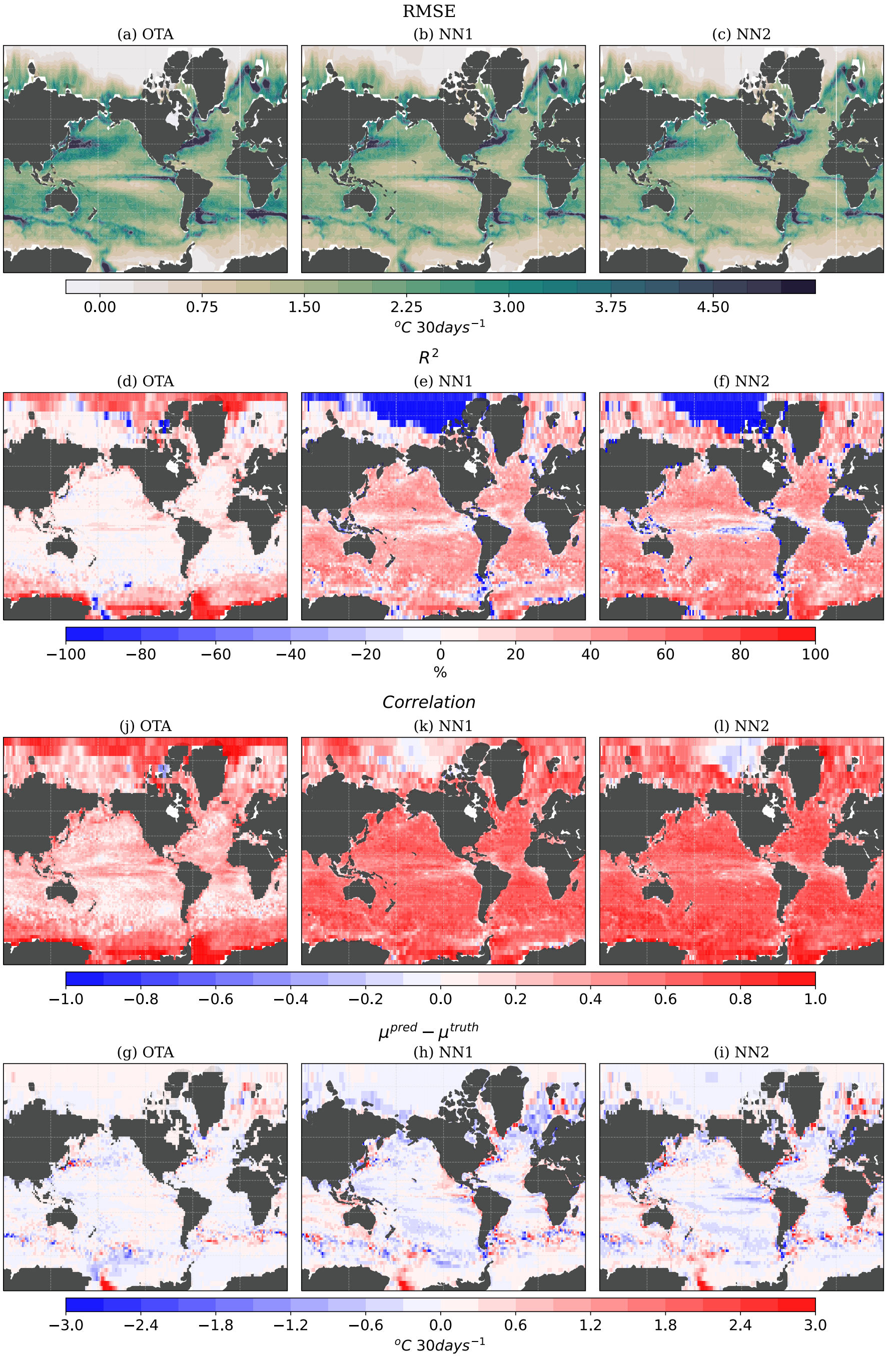


Figure 5. Spatial maps of offline metrics for OTA and NN predictions over the test dataset for 0 to 20 meters depth range. (a-c) RMSE, (d-f) R^2, (g-i) Pearson correlation coefficient, and (j-l) bias.

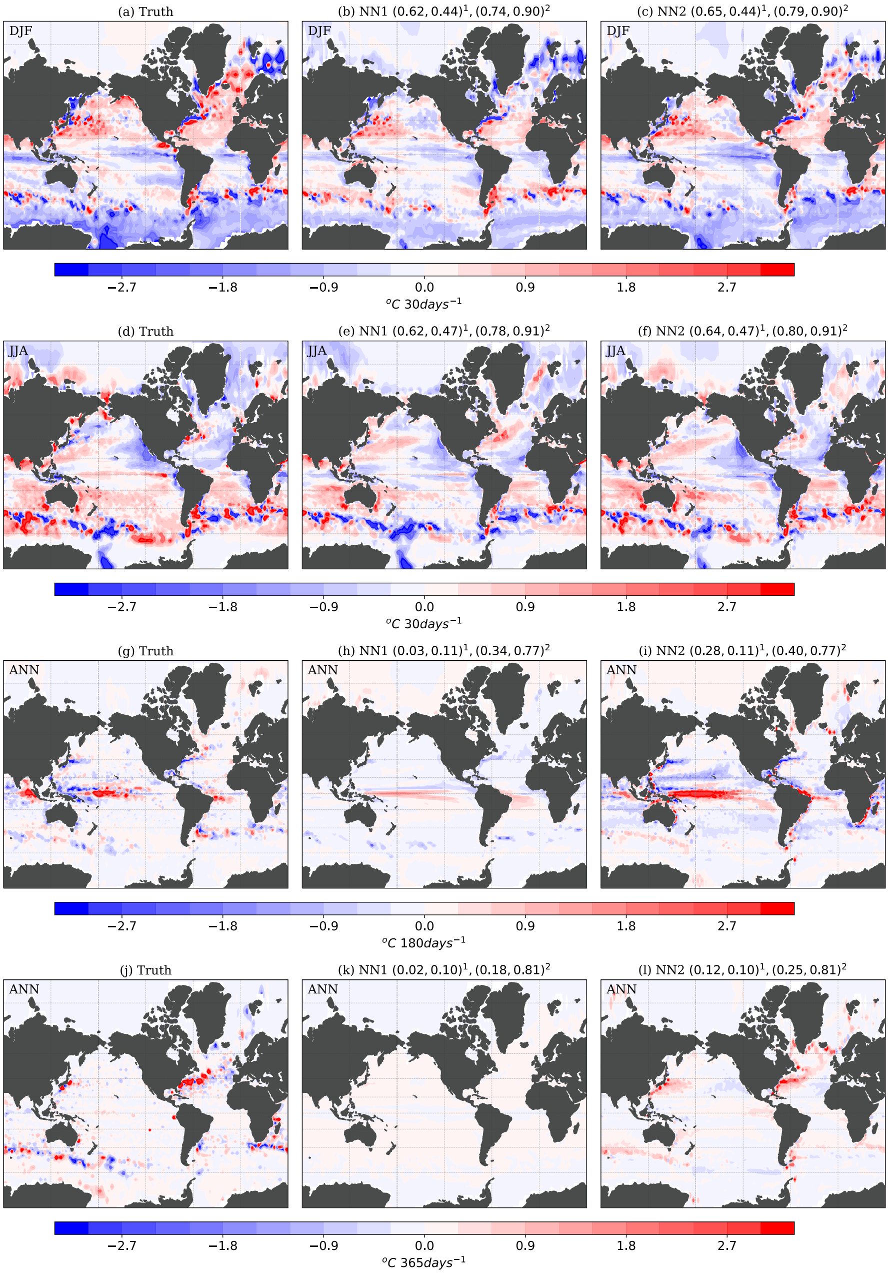

(a) Truth
(b) NN1 (0.62, 0.44)^1, (0.74, 0.90)^2
(c) NN2 (0.65, 0.44)^1, (0.79, 0.90)^2
DJF
DJF
DJF
−2.7
−1.8
−0.9
0.0
0.9
1.8
2.7
^oC 30$days^{-1}$
(d) Truth
(e) NN1 (0.62, 0.47)^1, (0.78, 0.91)^2
(f) NN2 (0.64, 0.47)^1, (0.80, 0.91)^2
JJA
JJA
JJA
−2.7
−1.8
−0.9
0.0
0.9
1.8
2.7
^oC 30$days^{-1}$
(g) Truth
(h) NN1 (0.03, 0.11)^1, (0.34, 0.77)^2
(i) NN2 (0.28, 0.11)^1, (0.40, 0.77)^2
ANN
ANN
ANN
−2.7
−1.8
−0.9
0.0
0.9
1.8
2.7
^oC 180$days^{-1}$
(j) Truth
(k) NN1 (0.02, 0.10)^1, (0.18, 0.81)^2
(l) NN2 (0.12, 0.10)^1, (0.25, 0.81)^2
ANN
ANN
ANN
−2.7
−1.8
−0.9
0.0
0.9
1.8
2.7
^oC 365$days^{-1}$

Figure 6. 2019-2022 seasonal and annual mean of true and predicted patterns for two different networks and the three depth ranges: (a-f) 0 to 20 meters, (g-i) 100 to 300 meters, and (j-l) 700 to 1000 meters. The seasonal or annual averaging periods are indicated at the top right corner of each map. The seasonal means are shown for the surface layer ((a-c) DJF and (d-f) JJA) and the annual means for the deeper layers (g-l). The two neural networks, NN1 and NN2, differ in input features. NN1 uses vertical gradients of T, S, U, and V, indicated by string, 'TzSzUzVz' in Figure 2; NN2 uses six flux variables, vertical gradients of T, S, U, and V, along with the horizontal gradients of U and V, as indicated by the string, 'fluxes-TzSzUzVz-UxUyVxVy' in Figure 2. The pattern correlation metrics between a) true and predicted fields and b) true and OTA fields (reference benchmark) are indicated in the two parentheses in the title of each plot. The first parenthesis indicated by superscript '1' is for the average of daily pattern correlations, and the second is for the pattern correlation of the mean fields.

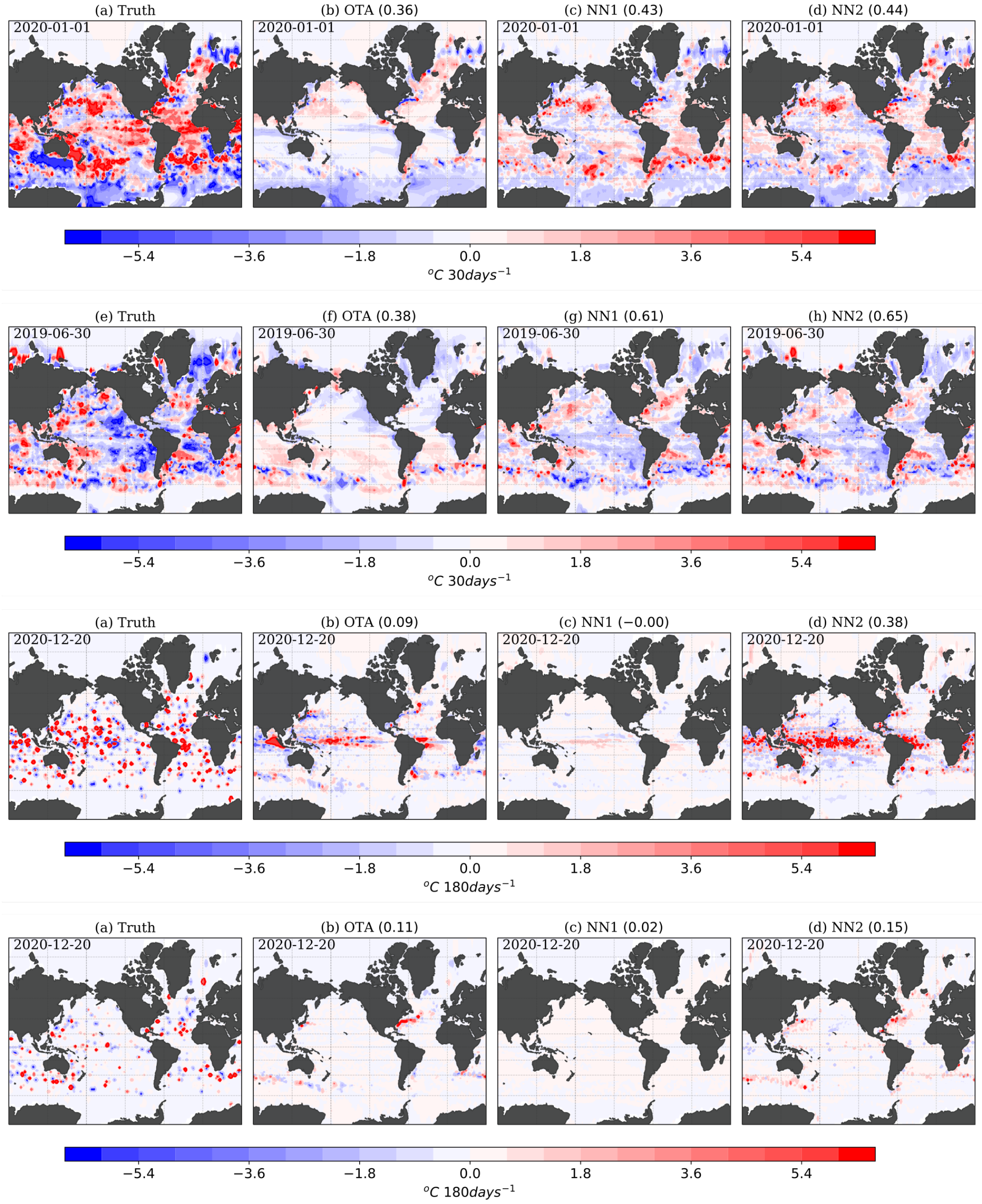


Figure 7. Daily snapshots of true, OTA and predicted increments for three depth ranges. Refer to Figure 6 for details.

column local in the ocean as it only depends on the vertical gradients of T, S, U, and V. Meanwhile, NN2 also uses six surface fluxes and four horizontal gradients of U and V, with 4419 and 7779 as a total number of parameters in NN1 and NN2, respectively. The horizontal gradients in NN2 make it implicitly nonlocal in horizontal directions. The pattern correlations between true and predicted fields on daily and climatological timescales are provided for further interpretation in the titles. We also compare daily snapshots of true, predicted, and from the climatology benchmark in Figure 7.

We can identify several spatial features in predicted fields reminiscent of the ocean's thermal and dynamic structures. Some examples include the Gulf Stream and the Kuroshio current in the northern hemisphere, alternate bands of positive and negative increments following the bathymetry in the Southern Ocean, basin-wide increments over the equatorial Pacific between 100 and 300 meters, and widespread positive increments over the winter hemisphere subtropical oceans in the upper 20 meters. These regions are essential for local and global climate and have significant implications for climate predictability. Reducing biases over these regions is, therefore, essential.

The two NNs compared in Figure 6 have similarly good fidelity in capturing the mean patterns in the upper 20 meters, with DJF (JJA) pattern correlations of 0.74 (0.78) and 0.79 (0.80). Both NNs capture the hemispheric signal that changes signs between the two seasons, most evident over the polar latitudes. Such large-scale seasonal changes in temperature increments are associated with model bias in simulating the seasonal cycle. The correlations between the 2019-2022 average and the OTA fields (second in each parentheses) are high across seasons and depths, as expected, since the OTA fields are the 2008-2018 average. This is also confirmation that the climatological DA increments do not change significantly, although the climate is not stationary and the observation network changes from year to year.

The seasonal cycle and associated bias are suppressed in the deeper layers, thus we focus on annual mean patterns for the two sub-surface layers.As we go deeper into the ocean, the pattern correlation decreases quickly between the NN predictions and truth. NN1 has difficulty learning increments in 700-1000 meters and predicts lower amplitude equatorial corrections in the 100-300 meters depth range. NN2, with the additional surface fluxes and horizontal gradients, performs considerably better in predicting the mean patterns over the equatorial Pacific at 100-300 meters and around the WBCs and the Southern Ocean at 700-1000 meters.

5.2 Daily snapshots

The daily snapshots are on the opposite end of the spectrum to time-mean patterns. While the former relates to an average correction tendency applied to the temperature equation in the physical space and projects on the mean bias, the latter corrects errors on fast timescales. We care about fast timescale corrections because the model errors are localized and happen on short timescales. Correcting for the long-term mean alone is like any other bias correction technique. Having fast timescale corrections in addition to slower ones is where this approach significantly differs and may provide an improvement over other bias correction techniques used in climate modeling.

Figures 7 (a-p) show patterns of actual and predicted daily temperature increments in three depth ranges and the measure of NN's skill using the pattern correlation metric as shown in the titles. January 1^{st} and June 30^{th} of 2020 are shown for 0-20 meters, while only one date of December 20^{th} is shown for 100-300 meters and 700-1000 meters. For comparison, the daily corrections from OTA, which are linearly interpolated from the monthly seasonal cycle, are also shown.

In the surface layer of 0-20 meters, the true DA increments (a,e) show corrections over most of the global ocean thanks to the daily global coverage of SST observational data, except for polar sea-ice covered oceans in the winter hemisphere. The OTA increments are subdued in magnitude because large daily DA increments are likely averaged out in the seasonal climatology. On the other hand, state-dependent NN predictions show larger corrections, albeit still smaller than the true DA increments. The pattern correlations of the NN predictions are also higher than the OTA for these 2 dates, indicating that the NNs are better at providing the necessary corrections to reduce model error growth over short forecast windows.

As we go deeper into the ocean, the actual daily increments (i,m) are only present when and where Argo observations exist, resulting in sparse and spotty distributions. The daily OTA corrections take on completely different patterns due to the climatological averaging. This demonstrates that with over a decade (2008-2018) of Argo observations, we have enough samples to retrieve large-scale climatological corrections that correspond to certain model deficiencies based on the spotty daily increments.

The NN state-dependent predictions show coherent large-scale patterns that resemble the OTA corrections more than the spotty daily increments. This is important since the proper bias corrections should not be determined by the availability of the observations like in the case of the daily increments. It is reassuring that the NNs are able to predict corrections for the subsurface ocean that do not look spotty like daily increments. One primary purpose of using ML in this study is to generalize the daily increments to work everywhere all the time, while providing additional state-dependent information. Furthermore, we do not expect the bias corrections to vary significantly at depth, where the natural variability of the ocean is on the timescale of months to decades. Between NN1 and NN2, the NN1 predictions are much smaller for both depth ranges, indicating the importance of horizontal velocity gradients as inputs. Most increments in the 100-300 meters depth range are concentrated in the tropical belt and near the WBCs. In the deeper layer of 700-1000 meters, the corrections are limited to WBCs and the Southern Ocean, similar to the OTA climatology.

Overall, NNs, particularly NN2, are able to predict daily increments that, while matching the OTA corrections climatologically, also provide additional state-dependent corrections at spatial and temporal scales beyond the OTA climatology. The temporal characteristics of the NN predictions compared to OTA will be discussed further in Section 6.

5.3 Zonal mean cross-section

Figure 8 compares the time-average zonal mean patterns of true and predicted fields from six neural networks with layer-wise pattern correlations summarized in Figure 8h. We find that the zonal mean patterns of temperature increments are primarily made up of localized features, such as WBCs, and do not entirely project on the zonal mean of state variables (shown in the Supplementary Figure S3). Despite that, the latitude-depth structure facilitates comparison between different networks, as described below. Generally, different NNs can capture zonal mean patterns in the upper ocean fairly well but perform differently in capturing deeper increments.

All the six NNs shown here have very similar skills in predicting the pattern in the upper 78 meters in the ocean, with pattern correlations between 0.8 and 0.9 (Figure 8h). The pattern consists of alternating positive increments over the subtropical ocean and negative increments over the equatorial and polar oceans, strongly influenced by the seasonal changes. The negative increments over the polar regions are associated with warm bias in both hemispheres in the summer months. The positive increments in the subtropical belt in both hemispheres are determined by winter mean patterns, indicating cooler mixed layer temperatures than observed. The negative increments over the equatorial region, on the other hand, are associated with the warm bias of the eastern equatorial

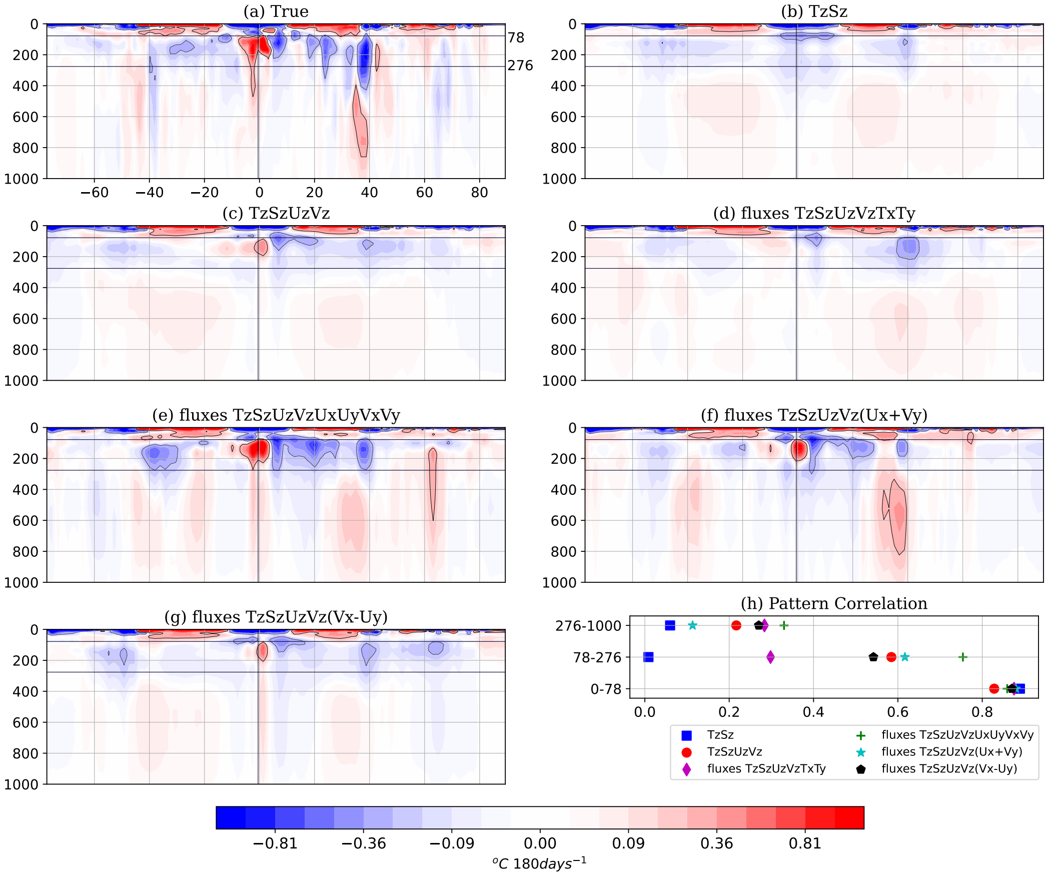


Figure 8. 2019-2022 mean zonal mean (a) true and (b-g) predicted patterns. The title of (b-g) indicates the combination of input features used by each neural network. (h) Layer-wise pattern correlation between true and NN predicted increments for three depth ranges: 0-78 meters, 78-276 meters, and 276-1000 meters, which are indicated by horizontal black lines in (a-g).

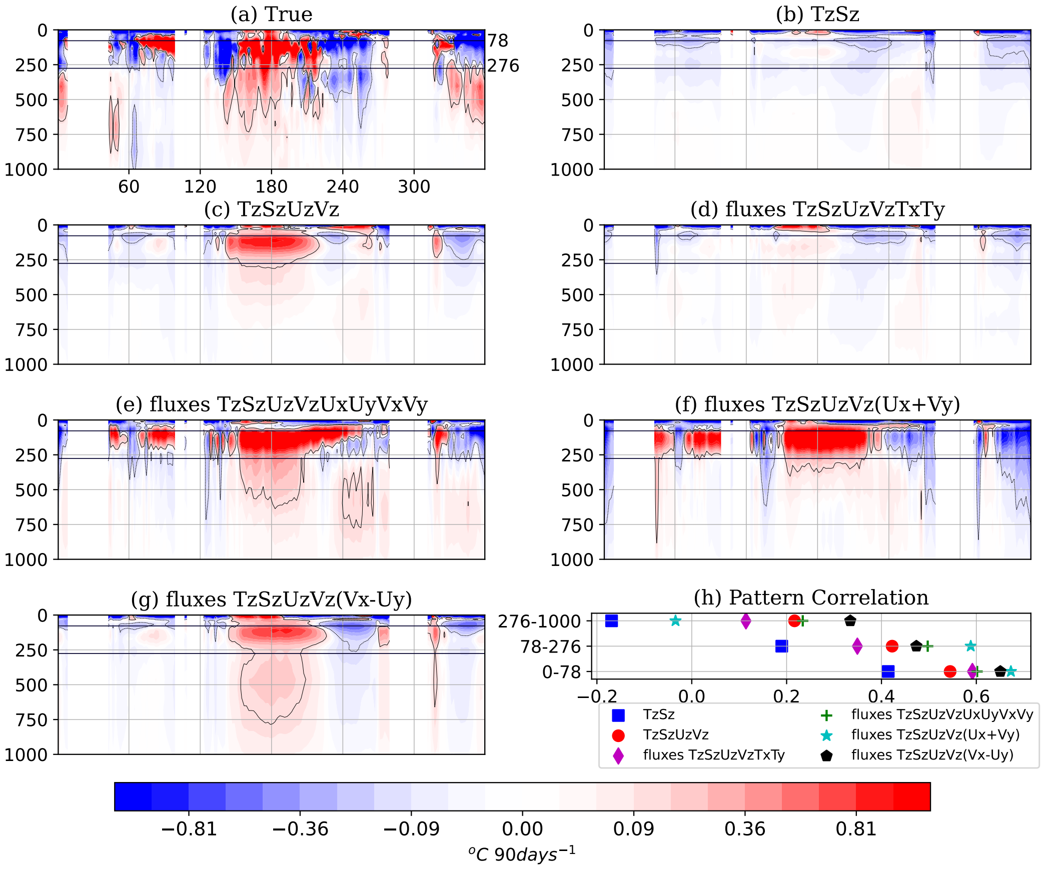


Figure 9. 2019-2022 mean vertical cross-section of (a) true and (b-g) predicted increments at equator for six different networks. (h) Pattern correlation between true and predicted increments in three depth ranges. Refer to Figure 8 for details.

basins in boreal winter months, as indicated by DJF means (Figure 6a-c), and may be tied to the coupled ocean-atmosphere interactions.

The dipole pattern seen in the 78-276 meter depth range between +-10 degree latitudes is associated with changes in the shape of the tropical thermocline. The positive increments at the equator are flanked by negative increments on either side, as seen in Figure 6 (g). Such a spatial pattern, primarily in the west to central equatorial Pacific, resembles the shape of the thermocline in the region, which is shallower at the equator and deeper on either side of it. Such a resemblance indicates corrections to the thermocline bias, associated temperature structure, and dynamical current systems like the eastward flowing equatorial undercurrent. The ability of different neural networks to reproduce this subsurface dipole pattern suggests that vertical and horizontal velocity gradients are required in addition to stratification for predicting positive increments at the equator and negative increments off the equator. We find that NN utilizing vertical gradients of T, S, U, and V (Figure 8 (c)) can predict the subsurface dipole over the equator. However, NN utilizing two additional horizontal temperature gradients and fluxes (Figure 8 (d)) can not, despite having a smaller overall RMSE than the former, as was shown in Figure 4 (a). This may be due to a trade-off in predicting surface versus subsurface increments between the two networks.

5.4 Equatorial cross-section

There are significant increments in the thermocline layer roughly 50-300 meters deep in the equatorial ocean (Figure 9), with positive increments in the eastern Indian Ocean, western Atlantic Ocean, and the central Pacific Ocean, and negative increments elsewhere. The central Pacific positive increments penetrate beneath the thermocline layer down to 1000 meters. The surface ocean has largely negative increments except over the central Pacific, where subsurface positive increments extend to the surface. Even though it is evident that these increments project strongly on the thermodynamical structure of the equatorial ocean, it is difficult to tease out the origin of these increments due to intricate coupling and feedback between different components and physical processes without targeted experiments.

The pattern correlation metrics provide information on the predictability of these patterns with a maximum value of 0.65 in the surface layer, which degrades with depth. The comparison of longitude-depth patterns, as predicted by NNs based on different inputs, once again reveals the importance of velocity shears in reproducing the mean pattern subsurface. The horizontal velocity shears, particularly Ux and Vy, add to the spatial variance as indicated by small spatial scales superimposed on the large-scale structure. The layerwise pattern correlation between actual and predicted mean fields in Figure 9 (h) shows that NN with horizontal divergence as one of its inputs performs the best in the top two layers. In contrast, the NN with the vertical component of the vorticity performs best in the deeper layer, 276-1000 meters. The NN with all four horizontal shears performs reasonably in all three layers.

The seasonal cycle of the zonal mean increments at the equator, as shown in Supplementary Figure S2, reveals significant seasonal dependence. The increments are positive during the boreal summer, fall, and late winter months, whereas they are negative during the spring and early summer months. The maxima in the negative increments in the spring season are reproduced even by the NN that only uses stratification as inputs. In contrast, velocity shears are necessary to capture positive increments in other months.

6 Temporal variability and timescale

Figures 10 (a-d) show maps of the standard deviation of actual, OTA, and NN-predicted increments averaged over the upper 20 meters in the ocean. The true increments have expectedly higher variance over the WBCs, equatorial Pacific, and the Southern Ocean. Even though NNs predict a fraction of the total variance of the actual increments (38% for NN1 and 42% for NN2), they are significantly higher than that of the OTA. It implies that NNs capture variances in the increments at timescales other than the seasonal cycle, also shown in the regionally averaged time series later (Figure 11).

Figure 10(e) shows the map of the Pearson correlation coefficient between the OTA and the true daily increments averaged between 0 and 20 meters, and Figures 10(f,g) show similar correlation maps for the two NNs but of the differences from the OTA. Over most of the global ocean between $60°S$ and $60°N$, the NN-predicted daily increments are more correlated with the true increments than OTA, which only contains seasonal-cycle variability. The zonal-mean correlations in Figure 10(h) confirm the improvement by the NNs.

The loss of correlation implies that NNs have difficulty capturing the seasonal variance over the high latitudes. Overall, there is 100% improvement in the median value globally. The zonal mean value of the correlation between actual increments and OTA, NN1, and NN2 are shown in Figure 10 (h). The two NNs perform better than the OTA over latitudes $\leq \pm 65°$, and worse otherwise. Moreover, NN2 performs better than NN1 at almost all latitudes.

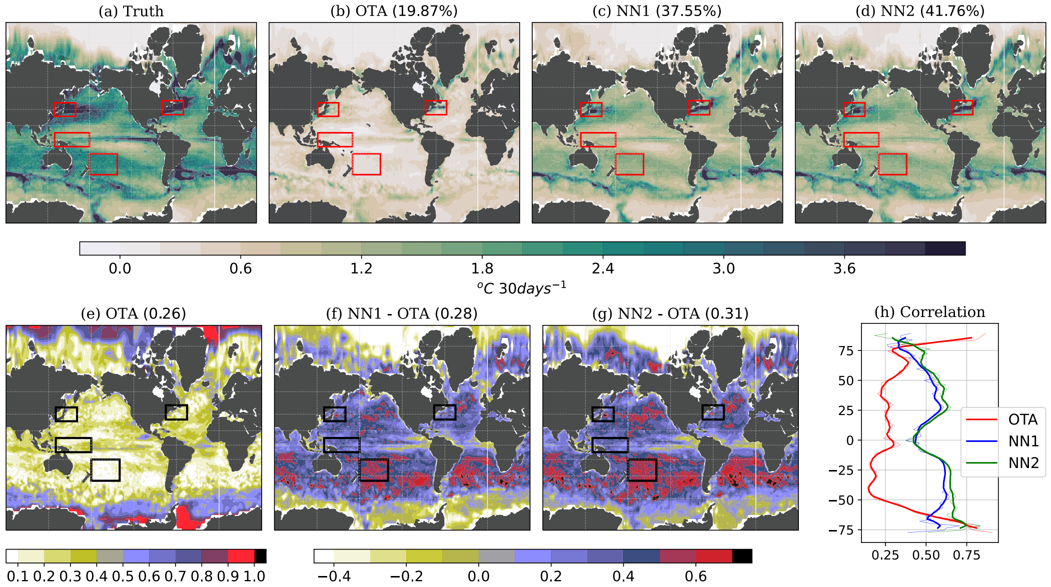


Figure 10. Spatial maps of (a) true,(b) OTA, and (c,d) predicted standard deviations in the upper 20 meters of the ocean for the 2019-2022 period. The numbers at the top (b-d) indicate the percentage of the spatial variance (of the true standard deviation in (a)) that is captured by OTA and the two neural networks: NN1 and NN2. (e) The correlation coefficient between the true and OTA time series spanning 2019 to 2022 at every grid point. The number at the top is the median of correlation coefficients globally. (f,g) Differences in maps of correlation coefficients of OTA and the two neural networks. The numbers indicate the median value of the differences. (h) The zonal mean of correlation coefficient maps for the OTA and the two networks.

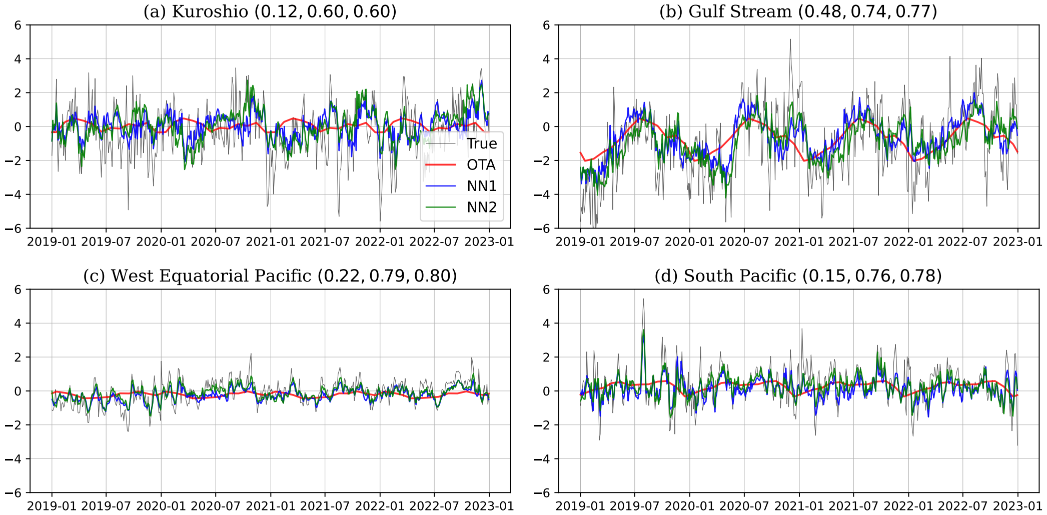


Figure 11. Regionally averaged true (grey), OTA (red), and predicted (NN1:blue; NN2:green) time series of upper 20 meters in the ocean for the four regions indicated by boxes in Figure 5. The sampling frequency is three days, and the period spans 2019 to 2022, equaling 488 time points. The three numbers in the parenthesis indicate the correlation coefficients between the true time series and OTA, NN1, and NN2 in that order.

The 2019-2022 daily time series of the average increments over the top 20 meters are plotted for four regions as marked in Figure 10, i.e. around Kuroshio extension, Gulf Stream, western Equatorial Pacific, and central south Subtropical Pacific. The correlation coefficients between the true daily increments and OTA/NN1/NN2 are shown in the title of each plot in that order. The true increments display the highest temporal variance, especially large spikes of positive or negative increments at synoptic timescale of a few days. These large increments over short periods are likely caused by synoptic variability that are not present in the atmosphere forcing, and have to been imposed by ODA. It is a desirable behavior that such large spikes are not learned and predicted by NNs, since they are not necessarily connected to intrinsic oceanic or coupled model biases. The NN-predicted increments do show larger variance at subseasonal, seasonal and interannual timescales that more closely follow the true increments, confirmed by the higher correlation coefficients of 0.6 to 0.8 depending on the regions. Among the four regions, the Gulf Stream region shows the most prominent seasonal cycle in the true increments, which leads to high correlation with OTA (0.48), while the other three regions do not show obvious seasonal cycle in the true increments, demonstrated also by the small magnitude of OTA increments throughout the year. The two NNs perform similarly for the top 20 meters, as indicated by previous

Figure 10 and 11 point to one of the key limitations of the climatological OTA scheme, which only retains climatological corrections that manifest as biases in the seasonal cycle. Although the seasonal cycle and the annual-mean climatology are important criteria for the fidelity of climate models, they certainly do not encompass all potential model biases. This limitation also provides the room for improvement with our ML-based OTA scheme, where the NNs could generalize the DA increments to predict corrections for a wide range of timescales, and the corrections are state-dependent to account for the nonlinear and non-stationary nature of the model biases.

7 Discussion

The time-averaged spatial and zonal-mean plots (Figures 6 and 8) show that NNs can learn biases associated with large-scale dynamical features in the ocean, namely, Antarctic Circumpolar Current (ACC) in the Southern Ocean, Equatorial Undercurrent (EUC) in the tropical Pacific, Kuroshio current and Gulf Stream in North Pacific and North Atlantic Oceans, and mixed layer depth, particularly over subtropical and mid-latitude open oceans. This section will briefly discuss the structure of these biases and potential sources of errors.

The true and predicted temperature increments in the Southern Ocean extend in a zonal direction and appear to follow the ocean bathymetry (Figure 6). These increments range from the surface down to a depth of 1,000 meters, indicating an equivalent barotropic structure similar to that of the Antarctic Circumpolar Current. The ACC consists of a westward-flowing current system strongly steered by ocean bathymetry. When these currents encounter undersea ridges, they create significant meanders, resulting in intense eddy activity downstream and standing meanders with pockets of warm (subtropical) and cold (polar) waters (Hughes, 2005). The zonally alternating positive and negative blobs of temperature increments seem to originate from systematic shifts in the locations and intensity of these meanders. It is important to note that in non-eddy-resolving models, the effect of eddy parameterization tends to smooth out the influence of meanders as internal interfaces are flattened. In contrast, eddy-resolving models, such as those in Hallberg and Gnanadesikan (2001), demonstrate that eddies can enhance bathymetry's influence, causing deep waters to mound over ridges.

In other words, errors in the representation of ocean flow-bathymetry interactions could be the underlying cause which could be related to errors in wind forcing, ocean stratification and smoothed bathymetry in coarse-resolution simulations like the one analyzed here (Hughes, 2005; Thompson & Naveira Garabato, 2014; X. Zhang et al., 2023). Additionally, since these regions are closely coupled with the overlying atmosphere, inaccuracies in atmospheric fluxes may also alter and introduce further errors in the ocean flow and the associated meanders (Vilela-Silva et al., 2024). We found that neural networks, based solely on thermal and salinity stratification, could not predict these alternating increments despite accurately identifying their locations. However, incorporating vertical and horizontal velocity shears allowed the neural networks to learn the zonally alternating features effectively.

Another significant correction is evident in the subsurface of the equatorial Pacific Ocean (Figures 6 (g,h,i), 8, and 9). This correction is connected to the meridional and zonal structure of the thermocline. Generally, the thermocline is shallower at the equator and progressively deepens toward the poles. Similarly, the negative increments at the surface at the equator spreads poleward and deeper along the thermocline, as seen in the zonal mean plots. Additionally, in the equatorial band below the thermocline, the equatorial undercurrent (EUC) appears colder than the observations, as indicated by positive increments in zonal-mean and equatorial cross-sections.

The eastward-flowing EUC originates from the meridional pressure gradient linked to the shape of the thermocline. The EUC plays a crucial role in the zonal redistribution of mass and heat across the tropical Pacific Ocean, significantly influencing the mean thermal structure and circulation, as well as the inter-annual variability associated with El Niño-Southern Oscillation (ENSO). A positive temperature correction in the EUC indicates that it is biased cold compared to observations. The stratification and currents in the equatorial Pacific Ocean and the atmospheric trade winds are intricately coupled (Coats & Karnauskas, 2018; Karnauskas et al., 2020; Verma et al., 2019). While the errors in EUC could have originated in any of the coupled processes, from an oceanic perspective, they could be linked to vertical mixing processes. Errors in vertical mixing can significantly impact the simulation of the equatorial thermocline and, consequently, the associated EUC, as shown in Pacanowski and Philander (1981). When comparing zonal-mean predictions, it becomes evident that vertical shears are critical for predicting the average positive corrections to the EUC, which seems consistent with the Pacanowski and Philander (1981). The mixing of heat into the EUC is influenced by diurnal mixing, which is modulated by variability in the surface winds (Schudlich & Price, 1992; Moum et al., 2022). This diurnal mixing also depends on how the mixing falls off at high Richardson numbers. While OTA is expected to capture the net effect of these processes, it does not account for variability caused by winds or meanders in the path of the EUC.

In zonal mean plots (Figure 8), particularly between 35-40°N, there is a notable negative correction at the surface and a positive correction at deeper depths. These corrections are linked to biases in the western boundary currents (WBCs) in the North Pacific and North Atlantic Oceans (Figure 6). Specifically, the Kuroshio current and the Gulf Stream are found to be too warm at the surface and too cold beneath compared to observational data. Climate models often exhibit inaccuracies in modeling the separation of these WBCs from the continental shelf (Schoonover et al., 2017), which can significantly impact oceanic and atmospheric conditions in their respective basins. These biases may arise from various sources, such as errors in the representation of bottom and lateral drag, meso- and submesoscale processes within the ocean, and interactions between the ocean and atmosphere. We find both ocean stratification and horizontal shears play a crucial role in predicting these corrections, wherein the latter may help define the boundaries and fronts associated with these currents.

Other interesting corrections learned by NNs include corrections in the surface mixed layer. As can be seen in Figure 4g, NN based solely on stratification (TzSz) outperforms (in terms of R^2) the state-independent climatology benchmark in the upper ocean, where both inputs are expected to be small. NNs must then be partially learning from biases in the mixed layer depth, which may have a distinct vertical signature in temperature increments near the base of the mixed layer. Comparing the spatial maps of R^2 for an NN based on stratification (not shown) with that of climatology benchmark predominantly shows improvement in subtropical and midlatitude bands similar to the ones highlighted in Figures 10 (f,g). We speculate that these biases may be linked to two factors: a) the parameterization of submesoscale processes in the ocean, which tend to restratify and shoal the mixed layer (Fox-Kemper et al., 2011) since data assimilation increments are produced with submesoscale parameterization disabled in this study, and b) the distinction between the ”mixed layer” in which vertical gradients are low and the ”mixing layer”(layer of active mixing) in which the gradients are essentially zero and dissipation is high (Giunta & Ward, 2022). By using vertical gradients as predictors, we could better characterize the mixing layer, which is vital for understanding short-term responses to heat fluxes.

Although we have shown that systematic corrections learned by NNs are associated with ocean dynamical features, we have not been able to attribute them to specific subgrid-scale physics, numerics, or atmospheric biases. This issue clearly hinders its adoption as model error parameterization in ocean models, wherein heat, salt, and momentum fluxes must be conservatively partitioned into different physical, dynamical, and numerical sources. We acknowledge that additional research is required, which is out of the scope of this manuscript. Despite the limitation, we expect that NNs are at least partially capturing some model errors and promote their case for testing and evaluation in online systems as a bias correction scheme and model error parameterization.

Future work may involve evaluating online skills and investigating issues related to the online implementation of such a scheme within the SPEAR system. A key concern is the stability of model integration; unphysical corrections and drifts associated with global imbalances may lead to instability in model integration. Other research directions could include quantifying the sensitivity of data assimilation increments to various subgrid-scale parameterizations and conducting specifically designed experiments to eliminate the impact of biases from other Earth system components, such as the atmosphere and sea ice.

Following the work of Rodwell and Palmer (2007), it may also be beneficial to save different subgrid-scale heat, salt, and momentum fluxes for use as additional predictors, which could aid in attribution. Further constraining the problem by limiting the physical space to surface mixed layer corrections or focusing on specific geographical regions may be helpful. From an algorithmic perspective, reducing the dimensionality of input and output profiles and enforcing physical constraints could enhance performance and generalization. All of these aspects are beyond the scope of this manuscript.

This problem formulation is column-local rather than three-dimensionally local, allowing predictors in the subsurface ocean to influence surface predictions and vice versa. Column-local models are not new and have been extensively used in many data-driven physical parameterizations, such as Yuval and O'Gorman (2020), and Laloyaux et al. (2022). A simple three-dimensional local model that relies only on local states may struggle to capture the complex space-time errors across the upper 1000 meters in the global ocean.

Other commonly used bias correction methods include flux adjustment, sea surface salinity restoration, and nudging toward observational products. While these methods help prevent long-term drifts in climate models, they have limitations. Unlike the flux adjustment and sea surface salinity restoration, the neural network-based approach evaluated here estimates systematic corrections at both the surface and the ocean's interior. While nudging to the climatology of a reanalysis product could correct some subsurface biases, our approach relies on in situ observations, potentially avoiding systematic biases in the reanalysis product. Lu et al. (2020) also demonstrated the benefits of using temperature increments for bias correction. This neural network method builds on their work by addressing fast-timescale systematic errors and the local state dependence of these errors.

One significant limitation is that the solutions do not always produce zero annual mean global averages for the upper thousand meters of the ocean despite being trained on near-zero averages. We speculate that non-zero averages are due to the NN's inability to capture all systematic behaviors across different scales and regions. This issue can lead to long-term drifts in simulated climate. A potential solution is to add a corrective term, but ensuring a bias-free model should be a priority for future research.

8 Summary and Conclusions

This study represents one of the earliest attempts at modeling systematic temperature increments using a full-complexity ocean general circulation model with neural networks. The end goals are to develop i) a state-dependent bias correction scheme for seasonal to decadal prediction systems and ii) an ocean model error parameterization for a free-running climate model within NOAA GFDL's SPEAR framework.

To achieve these goals, we employ relatively small, fully connected neural networks trained on data from the SPEAR-ODA system, which assimilates gridded OISST and Argo temperature and salinity profiles on daily timescales. The neural networks utilize a "column-local" state (which includes fluxes and vertical profiles) to predict vertical profiles of temperature tendency corrections for the upper 1,000 meters of the global ocean.

Specific goals are to determine what fraction of the space-time variance and to what extent the spatial patterns of temperature increments can be learned from the local state, its gradients, and surface fluxes. In this study, we evaluate the performance of neural networks on a withheld test dataset, often referred to as an "offline skill" in the existing literature, and compare it to a benchmark, state-independent climatology of temperature increments as outlined in Lu et al. (2020).

Our findings indicate that neural networks can learn systematic space-time variance and time-mean spatial patterns in the upper 1,000 meters of the global ocean despite being horizontally local. In terms of the global R^2 metric, the overall space-time variability is approximately 15–20% greater than that of the climatology benchmark (as shown in Figure 4 b). Moreover, nonlinear activation functions are crucial, as a linear network struggles to surpass the benchmark across many evaluated metrics.

Notably, the upper 20 meters of the ocean—typically part of the ocean surface mixed layer—exhibits the lowest root mean square error (RMSE), with an R^2 value reaching approximately 50%. Below 20 meters, R^2 values decline sharply, making prediction of subsurface variance more challenging. However, the minimum values remain non-negative, suggesting that performance is either better than or at least comparable to the climatology benchmark. One contributing factor to the low R^2 in the subsurface is the presence of small-scale dynamical noise in the daily temperature increments, which the chosen neural networks, based on coarse-resolution model state variables, are unable to predict.

Improvements in the upper 20 meters are uniformly observed across the global ocean, except in specific eastern equatorial and polar regions, where the R^2 values turn negative. Below 20 meters, there are localized areas—such as western boundary currents and equatorial regions—where R^2 values from neural network predictions are significantly above zero. These areas highlight the regions where subgrid-scale errors have a pronounced impact on large-scale ocean currents and where a neural network-based approach has the potential to enhance forecast skills. We presented these biases' characteristics, implications, and dynamics in the discussion section earlier.

The pattern correlations of mean fields reveal that neural networks cannot fully replicate time-mean patterns, particularly in the subsurface; the best-performing neural network achieves a maximum correlation of approximately 0.4, compared to about 0.8 for the climatology benchmark. However, neural networks outperform the climatology benchmark for daily timescale patterns and can reproduce spatial patterns in daily fields (two- or three-dimensional) with greater accuracy.

The performance of the neural networks is also influenced by the combination of input predictors, including stratification, vertical and horizontal velocity shears, and surface radiative, turbulent, and momentum fluxes. Analyzing performance changes by sequentially adding predictors to different neural networks provides qualitative insights into the relative importance of those predictors. Our analysis found that thermal and salinity stratification serves as better predictors of temperature increments than raw fields, resulting in lower test RMSE and higher R^2, especially in the upper 20 meters of the ocean. Including vertical and horizontal shears helps capture the space-time variance in the subsurface below 100 meters.

Overall, the improvements in depth- and location-dependent metrics demonstrate the advantages of using this data-driven approach to correct model errors compared to the previously employed climatological corrections by Lu et al. (2020). While our study showcases the potential benefits of this approach in an offline (diagnostic) context, further online (predictive) testing is needed to assess how it may reduce ocean model bias, affect the stability of model integration, and generalize across ocean models. Future research may also focus on strategies targeting specific subgrid-scale physics using data assimilation experiments and data or domain transformations to attribute corrections to various subgrid-scale processes better.

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ... Zhang, R. (2019). The gfdl global ocean and sea ice model om4.0:

Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167-3211. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001726 doi: https://doi.org/10.1029/2019MS001726

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., ... Kalchbrenner, N. (2023, 6). *Deep learning for day forecasts from sparse observations.* Retrieved from https://arxiv.org/abs/2306.06079

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020, 5). A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, *47*, e2020GL087776. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020GL087776 doi: 10.1029/2020GL087776

Arnold, H. M., Moroz, I. M., & Palmer, T. N. (2013, 5). Stochastic parametrizations and model uncertainty in the lorenz '96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*, 20110479. doi: 10.1098/rsta.2011.0479

Bolton, T., & Zanna, L. (2019, 1). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*, 376-399. doi: 10.1029/2018MS001472

Brenowitz, N. D., & Bretherton, C. S. (2018, 6). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*, 6289-6298. doi: 10.1029/2018GL078510

Bushuk, M., Winton, M., Haumann, F. A., Delworth, T., Lu, F., Zhang, Y., ... Zeng, F. (2021). Seasonal prediction and predictability of regional antarctic sea ice. *Journal of Climate*, *34*(15), 6207 - 6233. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/15/JCLI-D-20-0965.1.xml doi: https://doi.org/10.1175/JCLI-D-20-0965.1

Bushuk, M., Zhang, Y., Winton, M., Hurlin, B., Delworth, T., Lu, F., ... Zeng, F. (2022). Mechanisms of regional arctic sea ice predictability in two dynamical seasonal forecast systems. *Journal of Climate*, *35*(13), 4207 - 4231. Retrieved from https://journals.ametsoc.org/view/journals/clim/35/13/JCLI-D-21-0544.1.xml doi: https://doi.org/10.1175/JCLI-D-21-0544.1

Chang, Y., Schubert, S. D., Koster, R. D., Molod, A. M., & Wang, H. (2019, 1). Tendency bias correction in coupled and uncoupled global climate models with a focus on impacts over north america. *Journal of Climate*, *32*, 639-661. Retrieved from http://journals.ametsoc.org/doi/10.1175/JCLI-D-18-0598.1 doi: 10.1175/JCLI-D-18-0598.1

Chapman, W. E., & Berner, J. (2024, 4). Deterministic and stochastic tendency adjustments derived from data assimilation and nudging. *Quarterly Journal of the Royal Meteorological Society*, *150*, 1420-1446. doi: 10.1002/qj.4652

Chen, G., Kirtman, B. P., & Iskandarani, M. (2015, 10). An efficient perturbed parameter scheme in the lorenz system for quantifying model uncertainty. *Quarterly Journal of the Royal Meteorological Society*, *141*, 2552-2562. doi: 10.1002/qj.2541

Coats, S., & Karnauskas, K. (2018). A role for the equatorial undercurrent in the ocean dynamical thermostat. *Journal of Climate*, *31*(16), 6245–6261.

Dee, D. P. (2006, 1). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*, 3323-3343. doi: 10.1256/qj.05.137

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., ... others (2020). Spear: The next generation gfdl modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001895.

Farneti, R., Stiz, A., & Ssebandeke, J. B. (2022). Improvements and persistent biases in the southeast tropical atlantic in CMIP models. *npj Climate and Atmospheric Science*, *5*(1), 42. Retrieved from https://doi.org/10.1038/s41612

`-022-00264-4`

Fox-Kemper, B., Danabasoglu, G., Ferrari, R., Griffies, S., Hallberg, R., Holland, M., . . . Samuels, B. (2011). Parameterization of mixed layer eddies. iii: Implementation and impact in global ocean climate simulations. *Ocean Modelling*, *39*(1-2), 61–78.

Fox-Kemper, B., Hewitt, H. T., Xiao, C., Adalgeirsdottir, G., Drijfhout, S. S., Edwards, T. L., . . . Yu, Y. (2021, August). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. In V. Masson-Delmotte et al. (Eds.), (pp. 1211–1362). United Kingdom and New York, NY, USA: Cambridge University Press. Retrieved from `https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter_09.pdf` doi: 10.1017/9781009157896.011

Giunta, V., & Ward, B. (2022). Ocean mixed layer depth from dissipation. *Journal of Geophysical Research: Oceans*, *127*(4), e2021JC017904.

Gregory, W., Bushuk, M., Adcroft, A., Zhang, Y., & Zanna, L. (2023, 10). Deep learning of systematic sea ice model errors from data assimilation increments. *Journal of Advances in Modeling Earth Systems*, *15*. doi: 10.1029/2023MS003757

Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., & Zanna, L. (2024, 2). Machine learning for online sea ice bias correction within global ice-ocean simulations. *Geophysical Research Letters*, *51*. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL106776` doi: 10.1029/2023GL106776

Guillaumin, A. P., & Zanna, L. (2021, 9). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2021MS002534

Hallberg, R., & Gnanadesikan, A. (2001). An exploration of the role of transient eddies in determining the transport of a zonally reentrant current. *Journal of Physical Oceanography*, *31*(11), 3312–3330.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., . . . Thépaut, J. (2020, 7). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999-2049. Retrieved from `https://onlinelibrary.wiley.com/doi/10.1002/qj.3803` doi: 10.1002/qj.3803

Holder, C., & Gnanadesikan, A. (2023). How well do earth system models capture apparent relationships between phytoplankton biomass and environmental variables? *Global Biogeochemical Cycles*, *37*(7), e2023GB007701.

Hughes, C. W. (2005). Nonlinear vorticity balance of the antarctic circumpolar current. *Journal of geophysical research: Oceans*, *110*(C11).

Hyder, P., Edwards, J. M., Allan, R. P., Hewitt, H. T., Bracegirdle, T. J., Gregory, J. M., . . . Belcher, S. E. (2018, 9). Critical southern ocean climate model biases traced to atmospheric model cloud errors. *Nature Communications*, *9*, 3625. doi: 10.1038/s41467-018-05634-2

Jia, L., Delworth, T. L., Yang, X., Cooke, W., Johnson, N. C., McHugh, C., & Lu, F. (2023). Seasonal prediction of north american wintertime cold extremes in the gfdl spear forecast system. *Climate Dynamics*, *61*(3), 1769–1781.

Jia, L., Delworth, T. L., Yang, X., Cooke, W., Johnson, N. C., Zhang, L., . . . McHugh, C. (2024). Seasonal predictions of summer compound humid heat extremes in the southeastern united states driven by sea surface temperatures. *npj Climate and Atmospheric Science*, *7*(1), 180.

Johnson, N. C., Krishnamurthy, L., Wittenberg, A. T., Xiang, B., Vecchi, G. A., Kapnick, S. B., & Pascale, S. (2020, 3). The impact of sea surface temperature biases on north american precipitation in a high-resolution climate model. *Journal of Climate*, *33*, 2427-2447. Retrieved from http://journals.ametsoc.org/doi/10.1175/JCLI-D-19-0417.1https://journals.ametsoc.org/jcli/article/33/6/2427/347209/The-Impact-of-Sea-Surface-Temperature-Biases-on doi: 10.1175/JCLI-D-19-0417.1

Karnauskas, K. B., Jakoboski, J., Johnston, T. S., Owens, W. B., Rudnick, D. L., & Todd, R. E. (2020). The pacific equatorial undercurrent in three generations of global climate models and glider observations. *Journal of Geophysical Research: Oceans*, *125*(11), e2020JC016609.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., ... Wood, E. F. (2014, 4). The north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*, 585-601. Retrieved from http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00050.1 doi: 10.1175/BAMS-D-12-00050.1

Laloyaux, P., Kurth, T., Dueben, P. D., & Hall, D. (2022, 6). Deep learning to estimate model biases in an operational nwp assimilation system. *Journal of Advances in Modeling Earth Systems*, *14*. doi: 10.1029/2022MS003016

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... Battaglia, P. (2023, 11). Learning skillful medium-range global weather forecasting. *Science*. doi: 10.1126/science.adi2336

Lu, F., Harrison, M. J., Rosati, A., Delworth, T. L., Yang, X., Cooke, W. F., ... Adcroft, A. (2020). GFDL's SPEAR seasonal prediction system: Initialization and ocean tendency adjustment (OTA) for coupled model predictions. *Journal of Advances in Modeling Earth Systems*, *12*(12). Retrieved from https://doi.org/10.1029/2020MS002149

Moum, J. N., Natarov, A., Richards, K. J., Shroyer, E. L., & Smyth, W. D. (2022). Mixing in equatorial oceans. *Ocean Mixing*, 257–273.

Nadiga, B. T., Verma, T., Weijer, W., & Urban, N. M. (2019). Enhancing skill of initialized decadal predictions using a dynamic model of drift. *Geophysical Research Letters*, *46*(16), 9991–9999.

Pacanowski, R., & Philander, S. (1981). Parameterization of vertical mixing in numerical models of tropical oceans. *J. phys. Oceanogr*, *11*(11), 1443–1451.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022, 2). *Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators.* Retrieved from https://arxiv.org/abs/2202.11214

Rasp, S., Pritchard, M. S., & Gentine, P. (2018, 9). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*, 9684-9689. doi: 10.1073/pnas.1810286115

Robert, C. M., Gordon, C., & Cooper, C. (1997). The origin of flux adjustments in a coupled model. *Monthly Weather Review*, 909-926. doi: 10.1175/1520-0493(1997)125⟨0909:TOOFAI⟩2.0.CO;2

Rodwell, M. J., & Palmer, T. N. (2007, 1). Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society*, *133*, 129-146. doi: 10.1002/qj.23

Schoonover, J., Dewar, W. K., Wienders, N., & Deremble, B. (2017). Local sensitivities of the gulf stream separation. *Journal of Physical Oceanography*, *47*(2), 353–373.

Schudlich, R. R., & Price, J. F. (1992). Diurnal cycles of current, temperature, and turbulent dissipation in a model of the equatorial upper ocean. *Journal of Geophysical Research: Oceans*, *97*(C4), 5409–5422.

Thompson, A. F., & Naveira Garabato, A. C. (2014). Equilibration of the antarctic circumpolar current by standing meanders. *Journal of Physical Oceanography*, *44*(7), 1811–1828.

Tseng, K., Johnson, N. C., Kapnick, S. B., Delworth, T. L., Lu, F., Cooke, W., ... Jia, L. (2021, 9). Are multiseasonal forecasts of atmospheric rivers possible? *Geophysical Research Letters*, *48*, 1-12. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2021GL094000 doi: 10.1029/2021GL094000

Vecchi, G. A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A. T., ... Zhang, S. (2014, 11). On the seasonal forecasting of regional tropical cyclone activity. *Journal of Climate*, *27*, 7994-8016. doi: 10.1175/JCLI-D-14-00158.1

Verma, T., Saravanan, R., Chang, P., & Mahajan, S. (2019). Tropical pacific ocean dynamical response to short-term sulfate aerosol forcing. *Journal of Climate*, *32*(23), 8205–8221.

Vilela-Silva, F., Bindoff, N. L., Phillips, H. E., Rintoul, S. R., & Nikurashin, M. (2024). The impact of an antarctic circumpolar current meander on air-sea interaction and water subduction. *Journal of Geophysical Research: Oceans*, *129*(7), e2023JC020701.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021, 8). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*, 1-13. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2021GL092555 doi: 10.1029/2021GL092555

Wong, A. P. S., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., ... Park, H.-M. (2020, 9). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, *7*. doi: 10.3389/fmars.2020.00700

Wu, Y., Liang, Y., Kuo, Y., Lehner, F., Previdi, M., Polvani, L. M., ... Lan, C. (2023, 1). Exploiting smiles and the cmip5 archive to understand arctic climate change seasonality and uncertainty. *Geophysical Research Letters*, *50*. doi: 10.1029/2022GL100745

Xiang, B., Harris, L., Delworth, T. L., Wang, B., Chen, G., Chen, J.-H., ... Zhou, X. (2022). S2s prediction in gfdl spear: Mjo diversity and teleconnections. *Bulletin of the American Meteorological Society*, *103*(2), E463 - E484. Retrieved from https://journals.ametsoc.org/view/journals/bams/103/2/BAMS-D-21-0124.1.xml doi: https://doi.org/10.1175/BAMS-D-21-0124.1

Yang, X., Delworth, T. L., Zeng, F., Zhang, L., Cooke, W. F., Harrison, M. J., ... McColl, C. (2021). On the development of gfdl's decadal prediction system: Initialization approaches and retrospective forecast assessment. *Journal of Advances in Modeling Earth Systems*, *13*(11), e2021MS002529. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002529 (e2021MS002529 2021MS002529) doi: https://doi.org/10.1029/2021MS002529

Yuval, J., & O'Gorman, P. A. (2020, 7). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*, 3295. doi: 10.1038/s41467-020-17142-3

Zhang, G., Murakami, H., Cooke, W. F., Wang, Z., Jia, L., Lu, F., ... Zhang, L. (2021, 12). Seasonal predictability of baroclinic wave activity. *npj Climate and Atmospheric Science*, *4*, 50. Retrieved from https://www.nature.com/articles/s41612-021-00209-3 doi: 10.1038/s41612-021-00209-3

Zhang, X., Nikurashin, M., Peña-Molino, B., Rintoul, S. R., & Doddridge, E. (2023). A theory of standing meanders of the antarctic circumpolar current and their response to wind. *Journal of Physical Oceanography*, *53*(1), 235–251.

Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., ... Xiang, B. (2018). The gfdl global atmosphere and land model am4.0/lm4.0: 1. simulation characteristics with prescribed ssts. *Journal of Advances in Modeling Earth Systems*, *10*(3), 691-734. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001208 doi:

https://doi.org/10.1002/2017MS001208

Open Research Section

Python scripts to preprocess data and train neural networks can be obtained from https://zenodo.org/records/15191477. The subset of processed datasets, trained models, and Jupyter Notebooks for visualizations are also available at this location.

Acknowledgments
This research received support through Schmidt Sciences, LLC, and was intellectually guided by several members of the M^{2}LInES project. We thank Julien Le Sommer and Nathan Urban for helpful discussions regarding this work. The authors acknowledge that the research reported in this paper was primarily conducted using computational resources from the National Oceanic and Atmospheric Administration's Geophysical Fluid Dynamics Laboratory and from Princeton Research Computing at Princeton University.