https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

Correction of Air-Sea Heat Fluxes in the NEMO Ocean General Circulation Model Using Neural Networks

Andrea Storto[1,2], Sergey Frolov[3], Laura Slivinski[3], Chunxue Yang[1,2]

[1]National Research Council of Italy (CNR), Institute of Marine Sciences (ISMAR), Rome, Italy.
[2]National Research Center for High Performance Computing, Big Data and Quantum Computing (ICSC), Italy
[3]National Oceanic and Atmospheric Administration (NOAA), Physical Sciences Laboratory (PSL), Boulder, CO, United States

Correspondence to: Andrea Storto (andrea.storto@cnr.it)

Abstract. The atmospheric forcing and the heat exchanges between the ocean and the atmosphere represent one of the major sources of uncertainty for numerical ocean reconstructions and predictions. Air-sea heat fluxes may suffer from inaccuracies in meteorological fields, sea surface variables, and bulk formulations, which have a strongly non-linear dependence on the ocean state. Here, state-dependent errors of the heat fluxes are learned by artificial neural networks (ANN) from a dataset of heat flux correction terms, derived in turn from previous sea surface temperature nudging experiments. The pre-trained model predictors include stationary fields, atmospheric forcing data, ocean state, and stratification indices. Variable importance scores emphasize the dependence of the air-sea heat flux errors on the wind forcing. The pre-trained model of heat flux correction is then used to adaptively correct the fluxes online, in a series of global ocean experiments performed with the NEMO (Nucleus for European Modelling of the Ocean) ocean general circulation model, augmented with ANN inference capabilities. Results indicate the positive impact of the correction procedure, beyond the training period, e.g., in independent observation-poor and -rich periods, leading to the same dynamical and subsurface signature as in nudging experiments. Prediction experiments also indicate the method's potential for operational forecast applications. The method may also be adopted in coupled long-term reanalyses, long-range predictions, and projections.

Short summary. Inaccuracies in air-sea heat fluxes severely downgrade the accuracy of ocean numerical simulations. Here, we use artificial neural networks to correct the air-sea heat fluxes as a function of oceanic and atmospheric state predictors. The correction successfully improves surface and subsurface ocean temperatures beyond the training period and in prediction experiments.

1 Introduction

The ocean and the atmosphere interact by exchanging momentum, heat, and freshwater. These interactions drive the ocean circulation and ventilation (e.g., Marzocchi et al., 2021), its energy and water budgets, which are crucial, in turn, to understanding the ocean's role in Earth's climate and its variability over a wide range of spatial and temporal scales (e.g., Roberts et al., 2016; Small et al., 2019). Unfortunately, direct measurements of these fluxes are only available in limited buoy

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

locations, making their global and precise estimate a challenging problem (Cronin et al., 2019). Typically, air-sea fluxes are estimated using bulk flux parameterizations, which rely on near-surface meteorological variables, obtained from numerical weather prediction systems or atmospheric reanalyses (e.g., Yu, 2019). Bulk formulations are strongly non-linear, and there are significant uncertainties in these parameterization-based flux estimates; when averaged over ocean basins, heat fluxes may result in considerable imbalances (see, e.g., Kato et al., 2013; Storto et al., 2016; Valdivieso et al., 2017). Inaccuracies in ocean model vertical mixing and solar radiation penetration schemes may further amplify the sea surface errors (e.g., Deppenmeier et al., 2020, Jia et al., 2021), inducing in turn additional errors in the fluxes (e.g., Richards et al., 2009).

For both retrospective ocean simulations (e.g., OMIP, Ocean Model Intercomparison Project, Griffies et al., 2016), long-term reanalyses (Storto et al., 2021) and coupled model simulations (e.g., CMIP, Coupled Model Intercomparison Project, Small et al., 2019), inaccuracies in the air-sea heat flux calculation affect ocean heat redistribution (convection, stratification, and large-scale circulation), potentially compromising climate change signals (Storto et al., 2016; Carton et al., 2018). Errors in the air-sea heat fluxes thus remain among the most critical sources of uncertainty for many numerical ocean applications, including climate monitoring (e.g., Hakuba et al., 2024) and operational forecasting (e.g., Lewis et al., 2019; Lin et al., 2023; Ohishi et al., 2024).

Attempts to empirically correct errors in the fluxes have generally developed along two directions: i) bias-correction methodologies applied directly to ocean variables, i.e. correcting the effects of the air-sea heat flux systematic errors, see e.g. Balmaseda et al. (2007); ii) calibrating atmospheric reanalyses through comparison with observed climatology (Large and Yeager, 2009; Brodeau et al., 2010; Tsujino et al., 2018). Both strategies have their merits and weaknesses; bias-correcting ocean variables requires an adequate and dense ocean observing network, namely relying on the Argo float network limited to the last ~15 years, and cannot be used for attributing ocean model errors to specific processes; on the other hand, calibrating atmospheric reanalyses can mitigate errors in the atmospheric forcing, but not in the bulk formula approximations, therefore is only partially able to improve air-sea heat fluxes. Stochastic approaches can also, to some limited extent, improve the estimation of air-sea heat fluxes through rectification of the mean ocean state (Agarwal et al., 2023; Storto and Yang, 2023).

In this work, we use a state-of-the-science ocean general circulation model to showcase the idea of finding a predictor-correction empirical relationship, formulated in terms of neural networks, to correct the non-solar component of the air-sea heat fluxes. As neural networks have been proven to be universal approximators of any function (Hornik et al., 1989), they represent an obvious and flexible choice to model non-linear relationships between the atmospheric and oceanic states and the heat flux errors. Indeed, previous works (Bonavita and Laloyaux, 2020; Chen et al., 2022) have shown their ability to infer systematic errors in atmospheric models. The use of data assimilation increments was also demonstrated to be a robust strategy to learn such errors, with both theoretical (e.g., Mitchell and Carrassi, 2015) and practical (Farchi et al., 2021) arguments. The relationship is learned offline from present-day ocean model simulations that exploit the availability of space-borne sea surface temperature to estimate a corrective heat flux term. The correction is then tested online in ocean model simulations, for periods beyond the learned (training) one.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

The article's structure is as follows: after this Introduction, Section 2 describes the modeling system, the neural network setup, the relevant datasets, and the experimental setup. Section 3 summarizes the results of the reconstruction of the corrective heat flux terms and the online correction experiments, while Section 4 discusses and concludes.

2 Materials and Methods

2.1 The NEMO model and the nudging scheme

In this work, we use the NEMO ocean model (version 4.0.7, Madec et al., 2017) including the sea-ice dynamic and thermodynamic model SI[3]. NEMO is implemented at about 1/3°-1° of horizontal resolution, with 75 vertical depth levels and partial steps (Barnier et al., 2006). The surface boundary conditions are calculated through the CORE bulk formulas (Large and Yeager, 2009) implemented in the AEROBULK package (Brodeau et al., 2016), using meteorological variables extracted from the ECMWF ERA5 atmospheric reanalysis (Hersbach et al., 2020). The discharge from land is provided by the JMA JRA-55-do reanalysis (Tsujino et al., 2018). The model setup includes i) a 3-band RGB scheme for the net shortwave radiation, with extinction coefficients that depend on a monthly climatology of chlorophyll; ii) the TKE scheme for the vertical mixing; iii) a Laplacian operator and a bi-Laplacian operator for tracers and momentum, respectively. We use the same ocean model setting and parameters as the CIGAR reanalysis (Storto and Yang, 2024).

In the NEMO model, the air-sea heat flux can be optionally corrected with a nudging scheme (see e.g., Storto et al., 2016b). In practice, the net heat flux is decomposed into a penetrative (solar) component and a non-penetrative (non-solar) component. The non-solar component, which includes latent, sensible, and net longwave heat flux, can be corrected as:

$$Q'_{ns} = Q_{ns} + Q_{rp} = Q_{ns} + \kappa\,(SST_o - SST) \quad (1)$$

where the misfit between the observed (SST_o) and modeled (SST) sea surface temperature, multiplied by the nudging coefficient (or strength) κ, represents the corrective flux Q_{rp} added to the uncorrected non-solar flux. SST nudging is still a popular assimilation methodology for many climate-scale applications, where the use of gap-filled SST data ensures temporal consistency of the simulated ocean state compared to the direct assimilation of SST measurements (see, e.g., Yang et al., 2017). A 2000-2020 experiment (referred to as REF) with nudging to the SST data from the UKMO HadISST dataset (Rayner et al., 2003) was conducted, with a nudging coefficient equal to 100 W m^{-2} K^{-1}, which roughly corresponds to a 20-day relaxation time scale for a 50 m deep mixed layer. Note that nudging coefficients may be related to error characteristics and set up in a statistically optimal way (e.g., Zou et al., 1992; Vidard et al., 2003), although here, for the sake of simplicity, the nudging coefficient κ is spatially and temporally constant. Additionally, preliminary experiments considered using alternative SST datasets, for instance, the NOAA DOISST v2.1 (Huang et al., 2021), but those using HadISST provided the best results, and are the only ones considered in the remainder of the article.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

2.2 Artificial Neural Networks

The artificial neural network (ANN) employs a feed-forward architecture to infer the corrective flux Q_{rp} using several predictors, whereas the gridded predictors are unrolled but contain information about the geographical location, the so-called column neural networks as in Bonavita and Laloyaux (2020). Consequently, Q_{rp} will not depend any longer on SST observations but on several predictors, representative of the atmospheric and oceanic states. We grouped the predictors in several categories, listed in Table 1, to represent different sources of errors: i) stationary errors (location and day of the month); ii) surface temperature and its diurnal cycle; iii) heat flux components; iv) atmospheric wind forcing; v) surface salinity and the freshwater components; vi) ocean stratification and its diurnal cycle. Within the ANN training, the input variables are taken as daily means from the REF experiment (with SST nudging enabled), except the variables referring to diurnal amplitudes (defined as the maximum value minus the minimum value, at hourly frequency, within each day). The output fields used for training the ANNs are the Q_{rp} fields from the REF experiment, taken as the average between the same day as the predictors and the following day, to be nominally valid at 24:00 (end of each daily window). No sea-ice predictors are used, as it is expected that over sea-ice-covered areas the heat flux corrections vanish, due to the use of the sea-ice mask in the construction of the Q_{rp} fields in the REF experiment; this is because sea surface temperature data beneath sea ice are extrapolated from sea ice concentration data and are less reliable (Rayner et al., 2003).

We tested the impact of the correction frequency and training dataset's timescale in preliminary experiments with the NEMO model and the online correction of Q_{rp} with the pre-trained ANN, where we aimed to assess the impact of high temporal frequency in the inference step, ranging from monthly to daily sets of predictors-corrections, and investigate the impact of the frequency of the inference step in NEMO from daily to 3-hourly. The results are reported in Figure 1, in terms of global sea surface temperature RMSE, during the independent verification period 2019-2020. We increasingly improve the performances of the ANN-based inference in NEMO, closely approaching the REF experiment with SST nudging, by increasing both the temporal frequency of the predictor-correction datasets and the temporal frequency of the inference step. The best results are obtained for daily sets of predictors-corrections and 3-hourly inference step frequency. Note that we cannot increase it further, because 3 hours is the frequency of the surface boundary condition calculation in our configuration of NEMO.

After a preliminary comparison of different model architectures (not shown), the best-scoring neural network model includes 3 hidden layers (5 total), 256 neurons (considering an input size of 24 features and an output size of 1), and uses the rectified linear unit (ReLU) activation function in all layers but the last one. All input and output variables were normalized by their global mean and standard deviation. During the training, we used daily means, subsampled every 5 days during the period 2003-2017; while, at the same temporal frequency, the years 2001, 2002, 2017, and 2018 were used for validation within the ANN training, and 2019-2020 as independent test datasets.

Table 1 reports the list of predictors, grouped into categories, together with their impact in terms of Variable Importance Scores (VIS). VIS for the predictors is calculated through the permutation-based method of Fisher et al. (2019), through the *vip* R package (Greenwell and Boehmke, 2020), applied on the entire pre-trained model, or pointwise for each model grid-point (see

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

Table 1's caption for details). The explainability results for the entire pre-trained model suggest a large impact from static data, wind forcing, and temperature; a significant impact from the heat flux components, and a relatively smaller impact from salinity, freshwater fluxes, and ocean stratification.

There may exist, however, non-exclusive attributions of the errors to the predictors, as important correlations between parameters exist. For instance, VIS for temperature may partly indicate errors in climatological flux (due to the climatological state of the sea surface) or air-sea heat flux (e.g., the upward longwave heat flux); the wind forcing may also explain systematic errors in, e.g., latent heat flux; and so on for other correlated fields. Due to the strong non-linearity of air-sea interactions (and bulk formulas), these correlations are not reducible, and we take the practical approach to diagnose their impact as it comes from the VIS metrics.

Figure 2 shows the most impacting predictors as a function of longitude and latitude (both individual predictors and categories). This indicates that in most of the global ocean, the most important predictor is associated with wind forcing (either wind speed or stress). Interestingly, mesoscale active areas (e.g., western boundary current regions and the Antarctic Circumpolar Current, ACC) exhibit turbulent heat fluxes (latent and sensible heat) as the dominant predictor, consistently with the large influence of ocean mesoscale dynamics in air-sea exchanges therein (see, e.g., Frolov et al., 2021). In many coastal regions, the most important predictor is associated with freshwater fluxes. Only a few grid points exhibit another dominant predictor.

Figure 3 shows the individual impact of each predictor (in %), disclosing interesting spatial patterns, closely related to physical and dynamical processes. For instance, the mixed layer depth appears important near the Equator, likely related to ENSO variability; precipitation's impact is relevant in correspondence to the ITCZ (Inter-tropical convergence zone) likely due to its possible misplacement, and around the maritime continent. Eastern boundary upwelling systems are impacted by the solar heat flux, and the diurnal and seasonal variability (namely, the SST diurnal amplitude and the day of the year, respectively). The salinity flux is relevant over marginal ice zones, in both polar regions, associated with ice-ocean freshwater and heat exchanges; river runoff impacts the flux errors in the proximity of the shorelines. Sensible heat flux dominates in areas of strong mesoscale activity.

2.3 Experimental setup

Several experiments were run with the NEMO ocean model equipped with new functionalities, to store in a rolling array the predictors at the desired temporal frequency (see also section 2.2 and Figure 1). We use an in-house Fortran90 library (see the Code availability section) for online inference from the pre-trained model. The prediction step is natively implemented in Fortran90 to avoid the need for external software interfaces. The inferred corrective flux is then added to the uncorrected (bulk formula-derived) non-solar heat flux component every 3 hours.

The experiments with the ANN-based heat flux correction, presented hereafter, are named NNC (Neural Network-based Correction) and cover four different scenarios: i) validation in the training phase (self-consistency), i.e. during the period 2002-2018; ii) validation in the test phase (independent verification), i.e. during the period 2019-2020, after the training period; iii) validation in earlier periods, where no dense SST data were available (1961-1979), aiming to test the impact of the new method

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

for retrospective simulations and reanalyses, without any memory in the ocean state initialization; iv) validation in prediction experiments, namely 7-day forecasts initialized every 10 days in 2021 and 2022 from the data assimilation-enabled CIGAR reanalysis (Storto and Yang, 2024) and forced at the sea surface by the ECMWF operational forecasts replacing the ERA5 reanalyses used in the scenarios i), ii) and iii). These setups allow us to provide a full assessment of the methodology for different applications (long- or short-term simulations, historical reanalysis, and operational oceanography).

Further to NNC, we show results from REF (standard SST nudging enabled), CTRL (no corrections), and CLIMC (climatological corrections). The latter corrects the air-sea heat fluxes with a monthly climatology of corrections derived from the REF experiment, representing a linear benchmark for the methodology used in the NNC experiment.

3 Results

3.1 Contemporary simulations

The reconstruction of corrective fluxes with the pre-trained model is shown in Figure 4, which indicates the close correspondence between the SST nudging-derived and the neural network inferred fields, during the full period 2001-2020. Large corrections occur in mesoscale active areas (with large but not exclusive role of turbulent heat fluxes, see Figures 2 and 3), the North Atlantic subpolar gyre (with significant role of freshwater-related predictors, see Figure 3), in the Tropical and Southern Oceans. Signs are in general reversed in the Northern and Southern Hemispheres during the winter and summer seasons (namely, the non-solar heat fluxes are underestimated in wintertime and over-estimated in summertime, because of generally cold and warm biases of sea surface temperature, respectively). The seasonality of the corrections in deep convection areas suggests also systematic misrepresentation of convective processes therein, with much too deep mixed layer in the North Atlantic oceans, and more complex patterns in the Southern Ocean and ACC region.

The application of the correction leads to satisfying bias correction during the independent verification period 2019-2020, as shown in Figure 5. Large negative biases in the Gulf Stream, Kuroshio Extension, and central Tropical Pacific, plus locally in the Southern Ocean, present in CTRL are equally mitigated in REF and NNC, and likewise for warm biases in the eastern regions of Tropical basins, in the Indian Ocean, and locally elsewhere. Over the mid-latitudes, SST biases approach zero, while elsewhere the remaining biases that the SST data ingestion was not able to mitigate in the REF experiment are reproduced also in the NNC experiment. The global mean absolute error (MAE) over 2019-2020 decreases from 0.37°C in CTRL to 0.20°C and 0.19°C in NNC and REF, respectively, while CLIMC exhibits a MAE of 0.23°C. Differences between NNC and REF experiments are very small and limited only to polar areas (north of 60°N and south of 60°S), where the NNC corrections are small by construction.

The effects of the correction are also well reproduced in the ocean stratification, shown in Figure 6 in terms of mixed layer depth differences in March and September 2020 compared to the CTRL experiment. Either the SST assimilation or the neural network-based heat flux corrections induce an identical shift in the deep convection areas; in the Southern Ocean, during September 2020, a westward shift is visible in the Pacific sector; other local adjustments are visible in both the Atlantic and

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

Indian sectors of the ACC region. Adjustments are also visible in the Atlantic subpolar gyre, where enhanced convection appears in the Iceland basin and Irminger Sea, equally present in both REF and NNC experiments, along with attenuated mixing south of the Labrador Sea.

The global ocean heat content (OHC) anomaly interannual variations are visible in Figure 7 and show that NNC and REF lead to the same linear trends, and seasonal and interannual variations. Neglecting air-sea heat flux corrections in CTRL produces underestimated global ocean warming (0.15 W m^{-2}), which is identically corrected in NNC and REF (0.41 and 0.43 W m^{-2}, respectively). Using climatological corrections only partly mitigates the warming under-estimation (0.33 W m^{-2}), resulting in an intermediate solution. The correlation of OHC anomalies with respect to independent datasets such as the CIGAR reanalysis is also equally improved (from 0.48 in CTRL to 0.92 in NNC and REF). This suggests that the subsurface signature of the correction method is identical to the original nudging experiment.

Similarly, the global overturning circulation (Figure 8) shows again the same behavior for the REF and NNC experiments, indicating that also the dynamical signature of our approach provides the same results as in the assimilation experiment REF. The assimilation of the SST observations in REF reduces the North-South Hemisphere contrast of the overturning circulation (panel b in Figure 8), which is equally found in NNC.

Finally, the impact is evaluated against fully independent data, namely in-situ profiles extracted from the UKMO EN4 dataset (Good et al., 2013), during the period 2019-2020. This is shown in Figure 9 (left panels) where the RMSE of CTRL is shown, together with the differences of RMSE between REF or NNC minus CTRL. Negative (positive) values indicate an improvement (deterioration) borne by the correction method. The figure indicates the comparable impact of the SST nudging and the neural network correction on reducing the errors in the subtropical and mid-latitude regions, with the sub-surface Tropics less impacted by the corrections.

3.2 Retrospective simulations

Retrospective simulations were conducted to evaluate the potential of the method for long-term historical simulations, e.g., for OMIP- and CMIP-like exercises, and in multi-decadal reanalyses where the paucity of observation data in early periods limits the impact of conventional data assimilation and cannot take advantage of space-borne satellite measurements of SST. To this end, the same set of experiments presented earlier is performed for the period 1961-1979, initialized by the same initial conditions in 1961 taken from previous simulations.

We show the impact of NNC in terms of RMSE decrease versus the CTRL experiment in Figure 9, compared also, as an independent reference, to the CIGAR reanalysis (Storto and Yang, 2024) that assimilates all in-situ surface and sub-surface observations, and includes a deep-ocean large-scale bias-correction scheme. Improvements are present everywhere in NNC, except in the high-latitude 100-300 m depth layer, although smaller than CIGAR, especially in the Northern Hemisphere. The total average improvement (RMSE decrease) in the top 300 m of depth, compared to CTRL, is 22% for CIGAR and 7% for NNC, meaning that about one-third of the improvement borne by assimilating the full oceanic observing network and applying conventional bias-correction is achieved with the neural network-based correction. The improvement is remarkable at all

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

latitudes, also in the sub-surface Tropical region where the correction over the more recent years 2019-2020 failed to provide significant improvement (middle-left panel in Figure 9).

3.3 Forecast experiments

Forecast experiments are set up with the same model configuration but different initialization and forcing as detailed in Section 2.3. The correction is then applied online within the forecasts, as a proof-of-concept for operational purposes. Unlike the nudging scheme that depends on observational data and cannot be used in forecasts, the ANN-based correction depends only on the oceanic and atmospheric states, thus it can be adopted in operational forecasting systems.

Sea surface temperature errors (verified against mapped satellite data from DOISST v2.1, Huang et al., 2021) as a function of forecast lead time (Figure 10) indicate that NNC provides improvements comparable to nudging – shown as a benchmark –, except in the North Hemisphere Extra-Tropics, likely because of the intense mesoscale variability. The climatological corrections (CLIMC) fail to improve the CTRL experiment, as they cannot adapt to the variations of the atmospheric forcing in the forecast experiments. Compared to CTRL and considering the errors given by the climatology (dashed lines in the panels of Figure 10), the NNC scheme extends the horizon of useful forecasts by about 1 day in all regions. The impact of the method increases with the forecast lead time, suggesting that the approach might be fruitfully applied in long-range forecasting systems (sub-seasonal and beyond), although it should be demonstrated that coupled feedbacks in the case of Earth System models do not compromise the algorithm.

Similar results are found in the verification against in-situ profiles for the upper ocean (sea surface to 50 m of depth), shown in Figure 11. The top 50 m exhibit significant improvement in the Southern extra-Tropics and the Tropics, with an improvement borne by NNC increasing with forecast lead time. In the Northern extra-tropics, the ANN correction leads to negligible improvements.

4 Summary and Discussion

In this work, we propose an algorithm to correct air-sea heat fluxes by letting a neural network pre-trained model learn the relationships between ocean and atmospheric state predictors and heat flux corrective terms, estimated from a previous experiment that adopted sea surface temperature nudging to estimate and apply such terms. The predictors include several oceanic and atmospheric variables representative of the heat, freshwater, momentum fluxes, ocean temperature and salinity, and stratification. A feed-forward column neural network architecture is adopted, and the NEMO ocean general circulation model is augmented with online inference capability to allow collecting predictors and inferring corrections of the air-sea heat fluxes, based on the pre-trained model. Variable Importance Scores indicate the large impact that wind forcing has on the errors in most parts of the global ocean, with other variables dominating locally, e.g. turbulent fluxes in mesoscale active areas and freshwater fluxes near the coasts.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024


Geoscientific Model Development Discussions

The online use of the correction in the experiments indicates that the approach successfully reproduces the surface, sub-surface, and dynamical signature of the SST correction, even beyond the training data period. Next, the approach is demonstrated in early periods (1960s and 1970s) where surface temperature data are sparse, to mimic a long-term simulation or reanalysis application. In this context, the methodology provides a significant improvement in subsurface temperature errors, roughly equal to one-third of the improvement in a corresponding reanalysis system where all available observations are directly assimilated.

We also showcase the method in short-range prediction experiments, where observations cannot be used to correct the forecast step; the methodology is proven to significantly reduce surface and subsurface temperature errors, at a negligible extra computational cost and without the use of any observational information, increasing the SST predictability of about 1 day at all latitudes. Subsurface errors are also mitigated everywhere except in the Northern Extra-tropics.

While ANNs cannot provide improvements compared to the data assimilative experiments that are learning from, their use is appealing for several different applications that cannot rely on the observational input: simulations and projections, and multi-decadal reanalyses spanning early periods with scarce observations, as demonstrated in this article. Coupled model experiments may benefit from the method as well, although possible drawbacks from non-linear coupled feedback may arise, which need to be assessed in detail. Additionally, in the future, the algorithm could include corrections also to freshwater and momentum fluxes, subject to long and reliable datasets of e.g. sea surface salinity and currents to first estimate their corrections, whose availability is limited now.

The method represents the first attempt to leverage data assimilation correction increments, in this case from SST nudging, to learn systematic errors in ocean models attributable to inaccuracies in air-sea flux computations. It is also expected that higher-resolution implementations than that presented here may further benefit from the ANN compared to climatological corrections, due to their higher spatial and temporal variability. While providing good results in hindcast mode compared to the control experiment, climatological corrections fail in predictive experiments without proper retuning and re-computation through computationally expensive re-forecast experiments. This in turn suggests the possibility of extending the approach for calibrating forecasts without the need for long re-forecasts.

Further extension of the approach will consider full column increments for three-dimensional corrections, not only associated with heat fluxes but also vertical physics and model parametrizations; while this has been proven successful in atmospheric (e.g., Chen et al., 2022) and sea-ice (Gregory et al., 2023) applications, ocean implementations are more challenging due to scarce observing networks in the ocean interior, potentially hampering the use of analysis increments at depth, which is an active area of investigations at the moment.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

Code availability

The NEMO model is available through the official website https://www.nemo-ocean.eu; version 4.0.7, used in this study, can be downloaded at http://forge.ipsl.jussieu.fr/nemo/changeset/15814/NEMO/releases/r4.0/r4.0-HEAD?old_path=%2F&format=zip.

In-house modifications to the NEMO model code as used in the experiments presented here – including the module for ANN-based corrections, plus other modifications – are available as a git repository at https://baltig.cnr.it/nemo_ismar-rm/nemo_4.0.7/-/tree/3.0?ref_type=tags

There are several additional modifications than just the ANN-correction routine, which can be found in the *dnnqcorr* module. The ANN correction routine can be however isolated taking only the modules dnnqcorr.F90, and adding the call to dnn_qcorr in sbcmod.

The library for online inference in Fortran90 used in our NEMO experiments is available as a git repository at https://baltig.cnr.it/andrea.storto/nnt4nemo/-/tree/main/F90_Inference

It includes the ANNIF module for reading pre-trained ANN in NetCDF format, plus inference routines and their tangent-linear and adjoint versions.

The frozen version of both source codes, together with the scripts and the data to analyze and plot the results presented in the figures, are available as a dataset at Zenodo as https://doi.org/10.5281/zenodo.13380698.

Data availability

Atmospheric fields from ECMWF to force the ocean have been taken from the Climate Data Store (CDS, https://cds.climate.copernicus.eu) archive (ERA5) and the operational archive (operational forecasts, see www.ecmwf.int). SST data are available from the U.K. Met Office Hadley Centre (https://www.metoffice.gov.uk/hadobs/hadisst). For verification purposes, we used SST analyses from the NOAA DOISSTv2 dataset (https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html) and in-situ profiles from the U.K. Met Office EN4 dataset (https://www.metoffice.gov.uk/hadobs/en4/download-en4-2-2.html).

Author contribution

AS and SF have designed the methodology; AS has coded the methodology and run the experiments; LS and CY have provided guidance on the model developments and assessment results. AS drafted the initial version of the manuscript; all coauthors have discussed the results and revised the paper.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

Financial Support

This research is supported by the Research Program CN00000013 "National Centre for HPC, Big Data and Quantum Computing" Directorial Decree (grant no. 1031 of 17 June 2022) from the resources of the PNRR MUR – M4C2 – Investment 1.4 – "National Centers" Directorial Decree (grant no. 3138 of 16 December 2021). Financial support from NOAA/PSL for a scientific visit of A. Storto in Boulder (CO, USA) during June/July 2023 is also acknowledged.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

References

Agarwal, N., Small, R. J., Bryan, F. O., Grooms, I., & Pegion, P. J. (2023). Impact of stochastic ocean density corrections on air-sea flux variability. Geophysical Research Letters, 50, e2023GL104248. https://doi.org/10.1029/2023GL104248

Balmaseda, M.A., Dee, D., Vidard, A. and Anderson, D.L.T. (2007), A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans. Q.J.R. Meteorol. Soc., 133: 167-179. https://doi.org/10.1002/qj.12

Barnier, B., Madec, G., Penduff, T. et al. Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy-permitting resolution. Ocean Dynamics 56, 543–567 (2006). https://doi.org/10.1007/s10236-006-0082-1

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems, 12(12), e2020MS002232. https://doi.org/10.1029/2020MS002232

Brodeau, L., Barnier, B., Treguier, A.-M., Penduff, T., & Gulev, S. (2010). An ERA40-based atmospheric forcing for global ocean circulation models. Ocean Modelling, 31(3), 88–104. https://doi.org/10.1016/j.ocemod.2009.10.005

Brodeau, L., B. Barnier, S. Gulev, and C. Woods, 2016: Climatologically significant effects of some approximations in the bulk parameterizations of turbulent air-sea fluxes. J. Phys. Oceanogr., 47 (1), 5–28, 10.1175/JPO-D-16-0169.1

Carton, J. A., Chepurin, G. A., Chen, L., & Grodsky, S. A. (2018). Improved global net surface heat flux. Journal of Geophysical Research: Oceans, 123, 3144-3163. https://doi.org/10.1002/2017JC013137

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., & Tulich, S. (2022). Correcting systematic and state-dependent errors in the NOAA FV3-GFS using neural networks. Journal of Advances in Modeling Earth Systems, 14, e2022MS003309. https://doi.org/10.1029/2022MS003309

Cronin MF, Gentemann CL, Edson J, Ueki I, Bourassa M, Brown S, Clayson CA, Fairall CW, Farrar JT, Gille ST, Gulev S, Josey SA, Kato S, Katsumata M, Kent E, Krug M, Minnett PJ, Parfitt R, Pinker RT, Stackhouse PW Jr, Swart S, Tomita H, Vandemark D, Weller RA, Yoneyama K, Yu L and Zhang D (2019) Air-Sea Fluxes With a Focus on Heat and Momentum. Front. Mar. Sci. 6:430. doi: 10.3389/fmars.2019.00430

Deppenmeier, AL., Haarsma, R.J., LeSager, P. et al. The effect of vertical ocean mixing on the tropical Atlantic in a coupled global climate model. Clim Dyn 54, 5089–5109 (2020). https://doi.org/10.1007/s00382-020-05270-x

Farchi, A., Laloyaux, P., Bonavita, M. & Bocquet, M. (2021) Using machine learning to correct model error in data assimilation and forecast applications. Q J R Meteorol Soc, 147(739), 3067–3084. Available from: https://doi.org/10.1002/qj.4116

Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J Mach Learn Res. 2019;20(177):1–81. pmid:34335110

Frolov, S., C. A. Reynolds, M. Alexander, M. Flatau, N. P. Barton, P. Hogan, and C. Rowley, 2021: Coupled Ocean–Atmosphere Covariances in Global Ensemble Simulations: Impact of an Eddy-Resolving Ocean. Mon. Wea. Rev., 149, 1193–1209, https://doi.org/10.1175/MWR-D-20-0352.1.

Greenwell BM , Boehmke BC , Gray B. Variable importance plots - an introduction to the vip package . The R Journal, 2020 ;12 (1):343

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Gregory, W., Bushuk, M., Adcroft, A., Zhang, Y., & Zanna, L. (2023). Deep learning of systematic sea ice model errors from data assimilation increments. Journal of Advances in Modeling Earth Systems, 15, e2023MS003757. https://doi.org/10.1029/2023MS003757

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., et al.: OMIP contribution to CMIP6: experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project, Geosci. Model Dev., 9, 3231–3296, https://doi.org/10.5194/gmd-9-3231-2016, 2016.

Hakuba, M.Z., Fourest, S., Boyer, T. et al. Trends and Variability in Earth's Energy Imbalance and Ocean Heat Uptake Since 2005. Surv Geophys (2024). https://doi.org/10.1007/s10712-024-09849-5

Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. Q J R Meteorol Soc. 2020; 146: 1999–2049. https://doi.org/10.1002/qj.3803

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5), 359-366, https://doi.org/10.1016/0893-6080(89)90020-8.

Huang, B., C. Liu, V. Banzon, E. Freeman, G. Graham, B. Hankins, T. Smith, and H.-M. Zhang, 2021: Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1, Journal of Climate, 34, 2923-2939. doi: 10.1175/JCLI-D-20-0166.1

Jia, Y., Richards, K. J., & Annamalai, H. (2021). The impact of vertical resolution in reducing biases in sea surface temperature in a tropical Pacific Ocean model. Ocean Modelling, 157, 101722. https://doi.org/10.1016/j.ocemod.2020.101722

Kato, S., N. G. Loeb, F. G. Rose, D. R. Doelling, D. A. Rutan, T. E. Caldwell, L. Yu, and R. A. Weller (2013), Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances, J. Clim., 26, 2719–2740.

Large, W.G., Yeager, S.G. The global climatology of an interannually varying air–sea flux data set. Clim Dyn 33, 341–364 (2009). https://doi.org/10.1007/s00382-008-0441-3

Lewis, H. W., Siddorn, J., Castillo Sanchez, J. M., Petch, J., Edwards, J. M., and Smyth, T.: Evaluating the impact of atmospheric forcing and air–sea coupling on near-coastal regional ocean prediction, Ocean Sci., 15, 761–778, https://doi.org/10.5194/os-15-761-2019, 2019.

Lin, X., Massonnet, F., Fichefet, T., and Vancoppenolle, M.: Impact of atmospheric forcing uncertainties on Arctic and Antarctic sea ice simulations in CMIP6 OMIP models, The Cryosphere, 17, 1935–1965, https://doi.org/10.5194/tc-17-1935-2023, 2023.

Madec, G. & The NEMO System Team. NEMO Ocean Engine. Note Du Pole De Modélisation (Institut Pierre-Simon Laplace, Paris, France, 2017). https://doi.org/10.5281/zenodo.3248739

Marzocchi, A., Nurser, A. J. G., Clément, L., and McDonagh, E. L.: Surface atmospheric forcing as the driver of long-term pathways and timescales of ocean ventilation, Ocean Sci., 17, 935–952, https://doi.org/10.5194/os-17-935-2021, 2021.

Mitchell, L. and Carrassi, A. (2015) Accounting for model error due to unresolved scales within ensemble Kalman filtering. Quarterly Journal of the Royal Meteorological Society, 141(689), 1417–1428. https://doi.org/10.1002/qj.2451.

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Ohishi, S., Miyoshi, T. & Kachi, M. (2024) Impact of atmospheric forcing on SST biases in the LETKF-based ocean research analysis (LORA). Ocean Modelling, 189, 102357. Available from: https://doi.org/10.1016/j.ocemod.2024.102357

Rayner, N. A.; Parker, D. E.; Horton, E. B.; Folland, C. K.; Alexander, L. V.; Rowell, D. P.; Kent, E. C.; Kaplan, A. (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century J. Geophys. Res.Vol. 108, No. D14, 4407 10.1029/2002JD002670

Richards, K. J., S. Xie, and T. Miyama, 2009: Vertical Mixing in the Ocean and Its Impact on the Coupled Ocean–Atmosphere System in the Eastern Tropical Pacific. J. Climate, 22, 3703–3719, https://doi.org/10.1175/2009JCLI2702.1.

Roberts, M. J., H. T. Hewitt, P. Hyder, D. Ferreira, S. A. Josey, M. Mizielinski, and A. Shelly (2016), Impact of ocean resolution on coupled air-sea fluxes and large-scale climate, Geophys. Res. Lett., 43, 10,430–10,438, doi:10.1002/2016GL070559.

Small, R. J., F. O. Bryan, S. P. Bishop, and R. A. Tomas, 2019: Air–Sea Turbulent Heat Fluxes in Climate Models and Observational Analyses: What Drives Their Variability?. J. Climate, 32, 2397–2421, https://doi.org/10.1175/JCLI-D-18-0576.1.

Storto, A., C. Yang, and S. Masina (2016), Sensitivity of global ocean heat content from reanalyses to the atmospheric reanalysis forcing: A comparative study, Geophys. Res. Lett., 43, 5261–5270, doi:10.1002/2016GL068605.

Storto A and Yang C (2023) Stochastic schemes for the perturbation of the atmospheric boundary conditions in ocean general circulation models. Front. Mar. Sci. 10:1155803. doi: 10.3389/fmars.2023.1155803

Storto, A., Balmaseda, M. A., de Boisseson, E., Giese, B. S., Masina, S., & Yang, C. (2021). The 20th century global warming signature on the ocean at global and basin scales as depicted from historical reanalyses. International Journal of Climatology, 41(13), 5977–5997. https://doi.org/10.1002/joc.7163

Storto, A., Masina, S. and Navarra, A. (2016b), Evaluation of the CMCC eddy-permitting global ocean physical reanalysis system (C-GLORS, 1982–2012) and its assimilation components. Q.J.R. Meteorol. Soc., 142: 738-758. https://doi.org/10.1002/qj.2673

Storto, A., Yang, C. Acceleration of the ocean warming from 1961 to 2022 unveiled by large-ensemble reanalyses. Nat Commun 15, 545 (2024). https://doi.org/10.1038/s41467-024-44749-7

Tsujino et al., 2018: JRA-55 based surface dataset for driving ocean-sea-ice models (JRA55-do), Ocean Modelling, 130(1), pp 79-139. https://doi.org/10.1016/j.ocemod.2018.07.002

Valdivieso, M., Haines, K., Balmaseda, M., Chang, Y.-S., Drevillon, M., Ferry, N., et al. (2017). An assessment of air–sea heat fluxes from ocean and coupled reanalyses. Clim. Dyn. 49, 983–1008. doi: 10.1007/s00382-015-2843-3

Vidard PA, Le Dimet F-X, Piacentini A (2003) Determination of optimal nudging coefficients. Tellus A 55(1):1–15

Yang, C., Masina, S. and Storto, A. (2017), Historical ocean reanalyses (1900–2010) using different data assimilation strategies. Q.J.R. Meteorol. Soc., 143: 479-493. https://doi.org/10.1002/qj.2936

Yu, L. (2019). Global air–sea fluxes of heat, fresh water, and momentum: energy budget closure and unanswered questions. Annu. Rev. Mar. Sci. 11, 227–248. doi: 10.1146/annurev-marine-010816-060704

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

420 Zou X, Navon IM, Ledimet FX (1992) An optimal nudging data assimilation scheme using parameter estimation. Q J R Meteorol Soc 118(508):1163–1186

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

Category	Predictors	Total VIS (%)	Gridpoint-averaged Local VIS (%)
Stationary	Lon, Lat, Time	30	7
Temperature	SST, OHC, SST_da	22	12
Salinity	SSS, OSC	4	1
Heat flux	Qlat, Qsen, Qlw, Qsw, Qemp	11	18
Freshwater flux	Precip, Runoff, Salt flux,	6	16
Wind forcing	Stress modulo, Wind speed, SSH	26	44
MLD	MLD, MLD_da	1	2

Table 1. List of predictors, grouped by categories, with their aggregated Variable Importance Score (VIS), given as percent impact, both as the impact on the pre-trained model (Total VIS), and averaged over the global domain from the pointwise application (Gridpoint-averaged Local VIS). MLD: mixed layer depth; OHC: ocean heat content; SSS: sea surface salinity; OSC: ocean salt content; SSH: sea surface height; the suffix *_da* refers to diurnal amplitude. The Total VIS refers to the VIS over the full columnar ANN model, while the local VIS is calculated for each gridpoint by fixing the longitude-latitude pair to the corresponding gridpoint separately. The different VIS results respond to different questions, i.e. either the global impact of each predictor on the final ANN, or the spatial average of local variable importance. Diagnosing local VIS allows investigating maps of variable impact (see Figures 2 and 3).

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

435

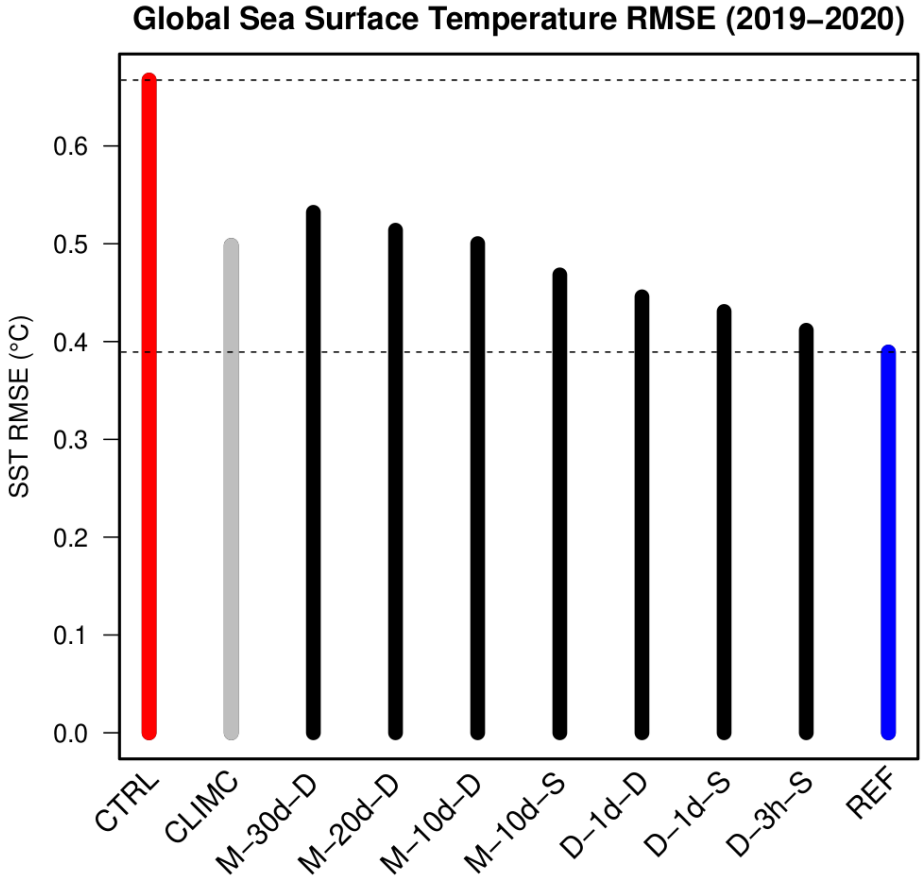
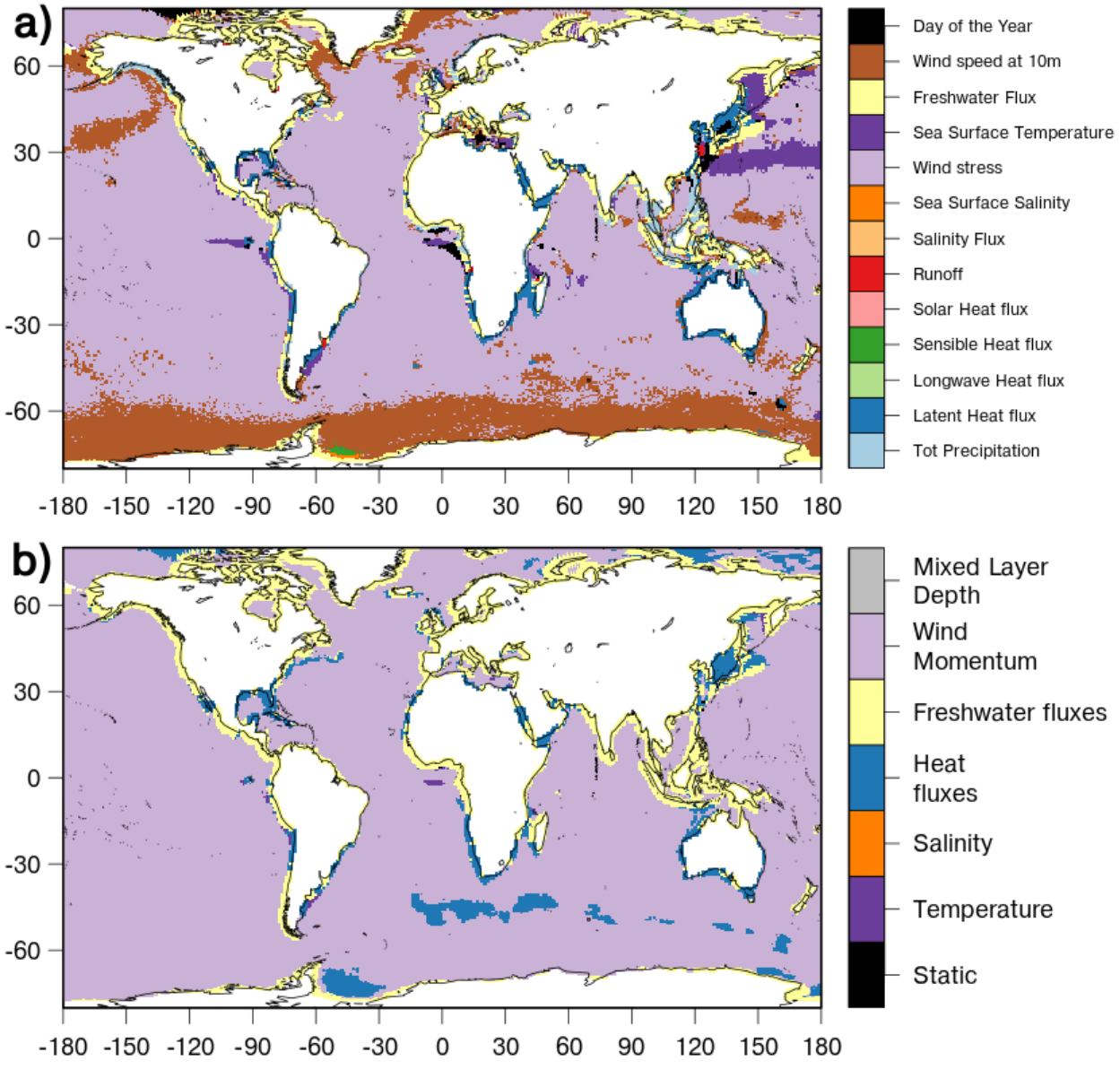


Figure 1. Sea surface temperature globally averaged RMSE for preliminary experiments, over the independent verification period 2019-2020. M-* experiments and D-* experiments refer to the use of monthly versus daily averaged nudging increments in the ANN
440 **training; the second string in the experiment name (30d, 20d, …, 3h) refers to the length of the predictor rolling archive; the last letter refers to the frequency of the update in the online experiments ("D" as daily, "S" as sub-daily, namely every 3 hours).**

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions



445

Figure 2. Dominant predictors identified by Variable Importance Scores (by individual predictor, panel a, and by predictor categories, panel b), from the optimal pre-trained model described in the text. The predictors' list is as in Table 1, but for the sake of clarity only those predictors with at least one dominant gridpoint are considered in panel a).

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions

450

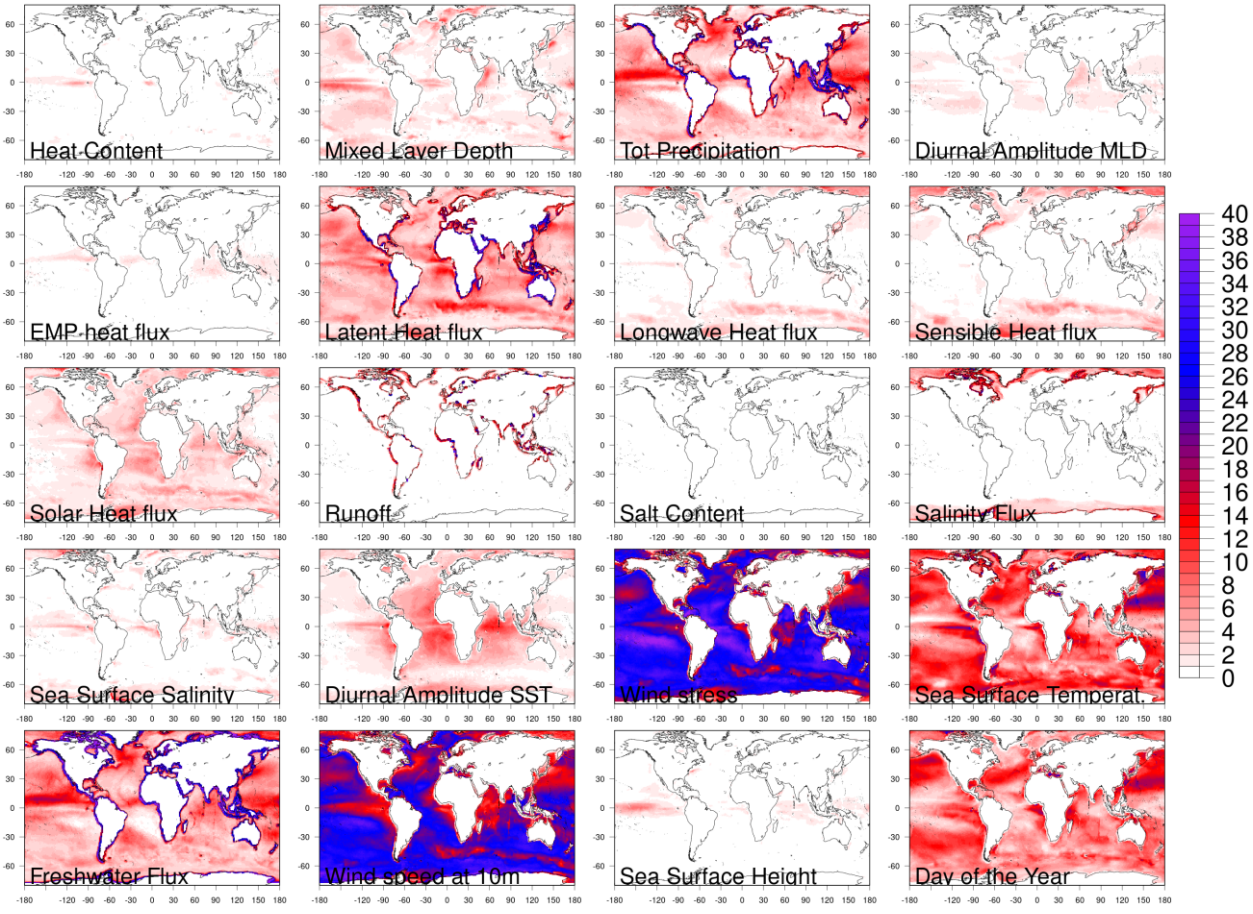
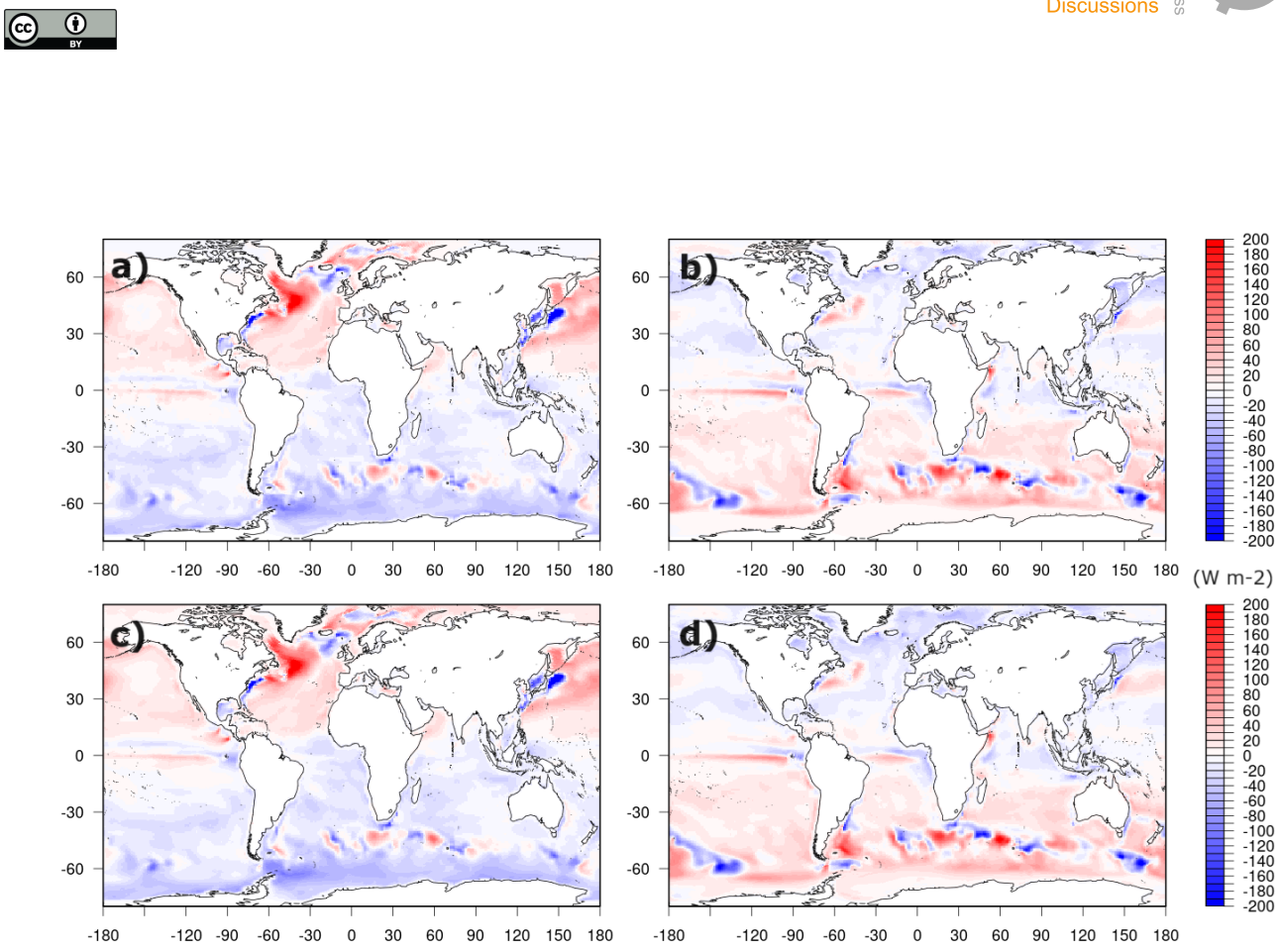


Figure 3. Variable Importance Score (VIS, in % values) for each predictor used in the neural network pre-trained model, as a function of grid point (namely, fixing the values of longitude and latitude). VIS maps are used to locally attribute different sources of air-sea heat flux errors to the predictors.

455

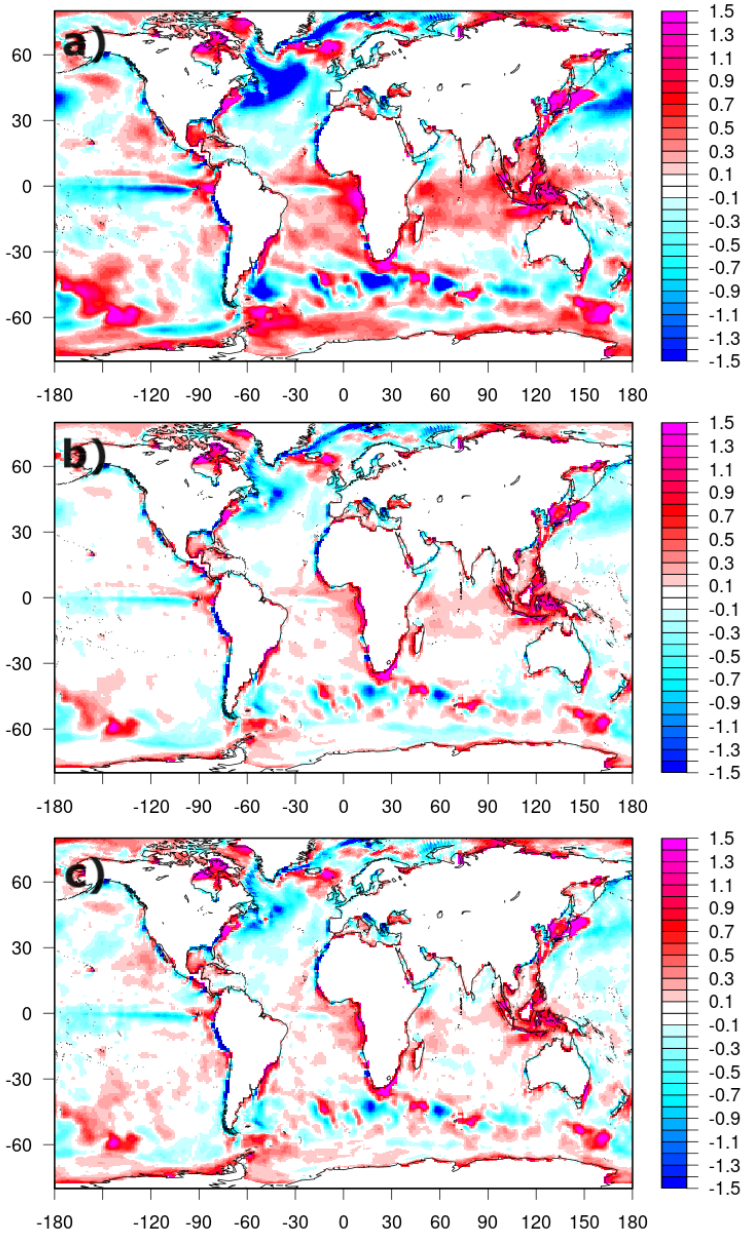
https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024



460 **Figure 4. Reconstructed heat flux correction fields versus the original ones from the REF (a, b) and NNC (c, d) experiments, for JFM (a, c) and JJA (b, d), during the 2002-2020 period.**

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions



465

Figure 5. SST Bias over the independent period 2019-2020 against the SST observations (from UKMO HadISST), for the three experiments CTRL (panel a), REF (panel b), and NNC (panel c).

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development
Discussions
Open Access
EGU

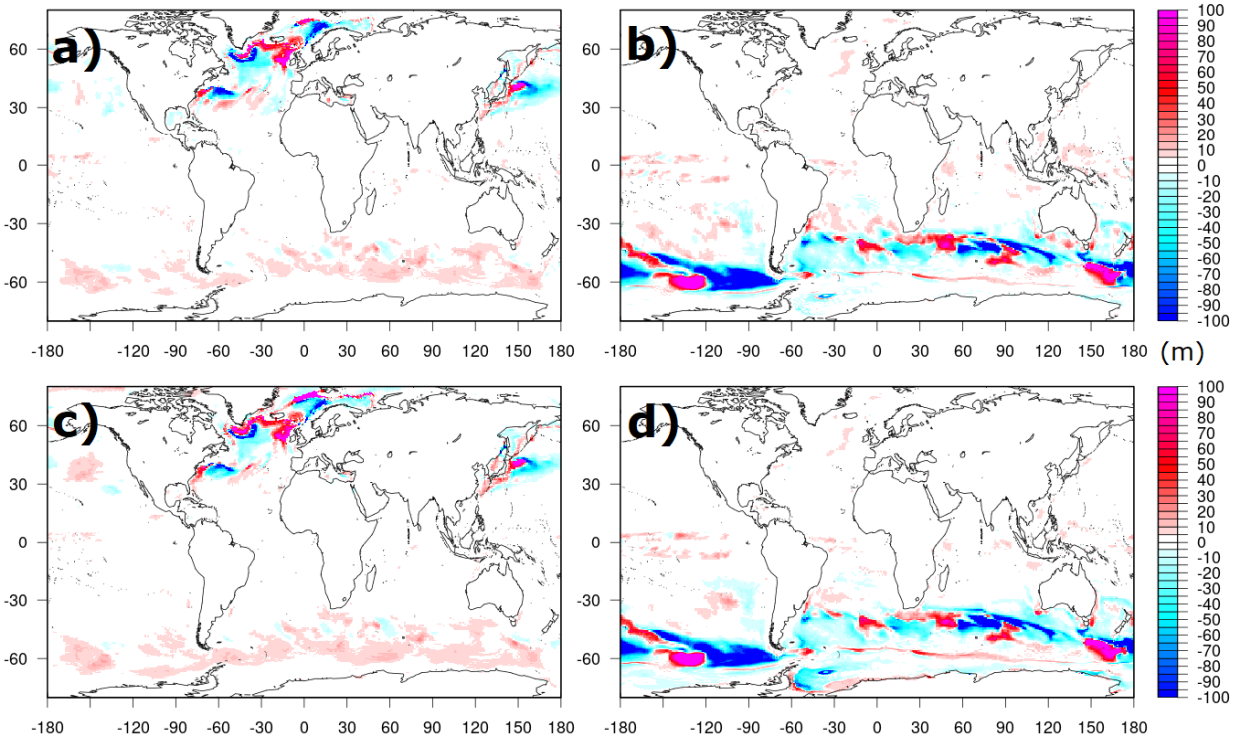


Figure 6. Mixed layer depth differences with respect to the CTRL experiment during March 2020 (panels a, c) and September 2020 (panels b, d), for experiments REF (panels a, b) and NNC (panels c, d).

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development
Discussions
Open Access
EGU

475

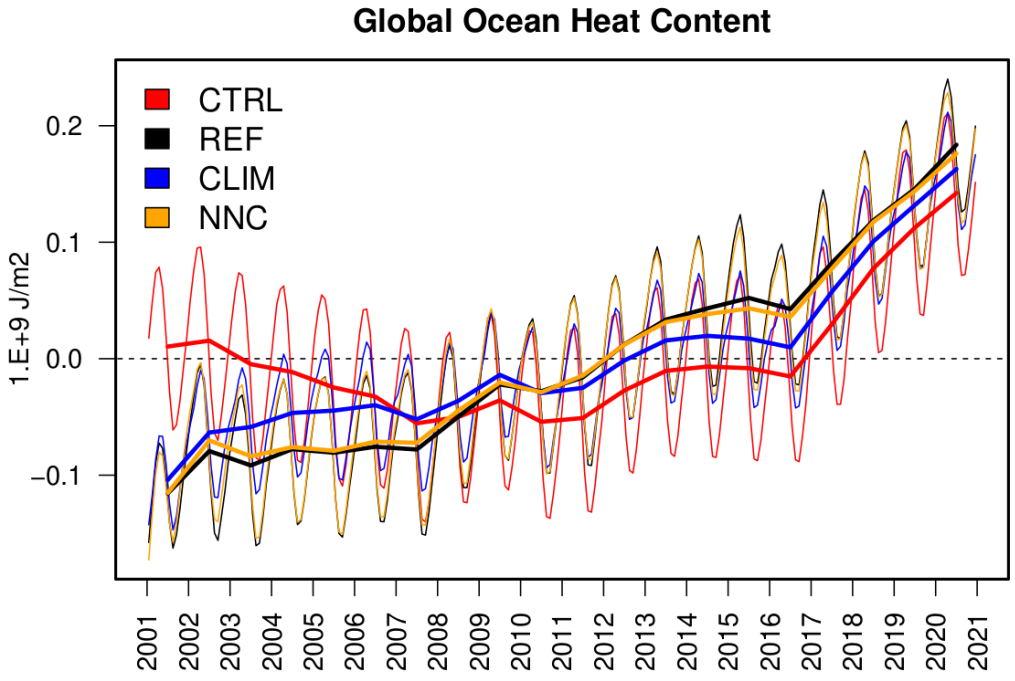


Figure 7. Global ocean heat content anomaly vertically integrated over the period 2001-2020 for the four experiments presented in the text, as monthly (thin lines) and yearly (thick lines) means.

480

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

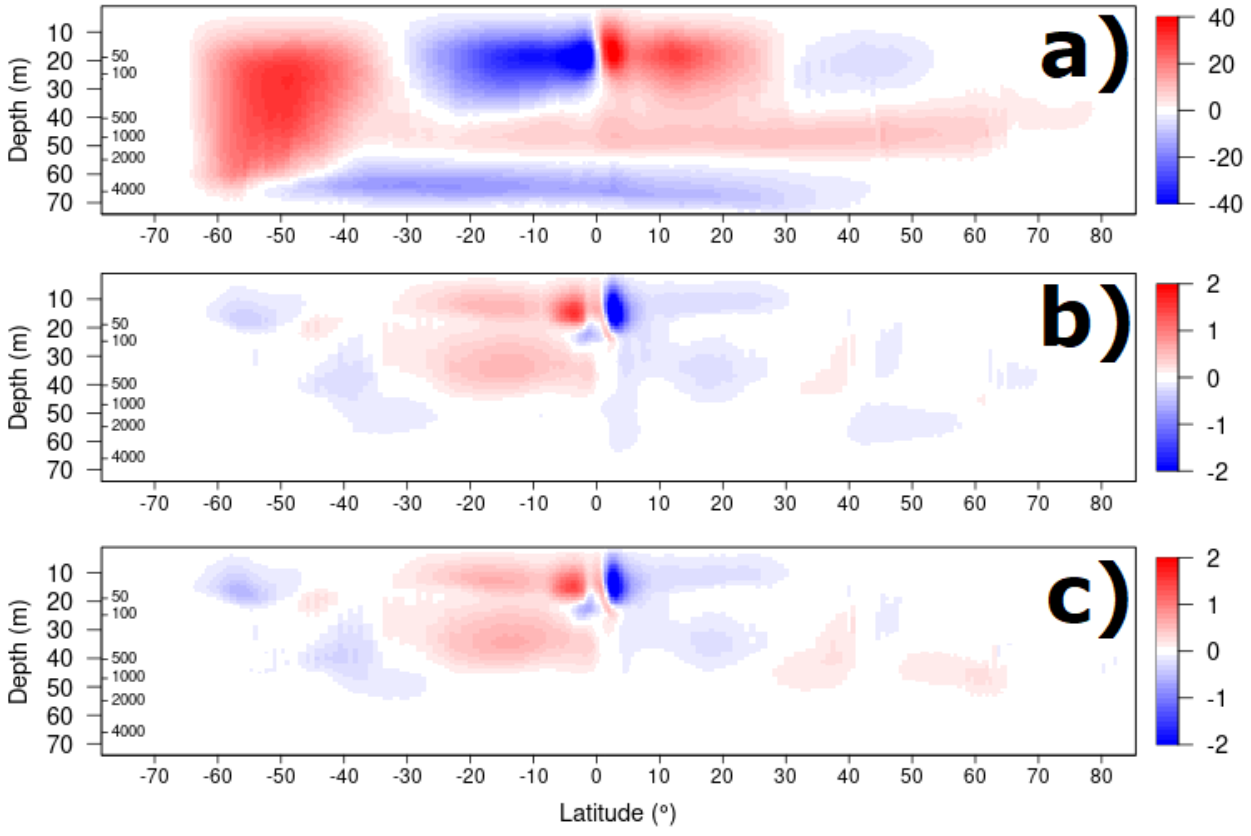
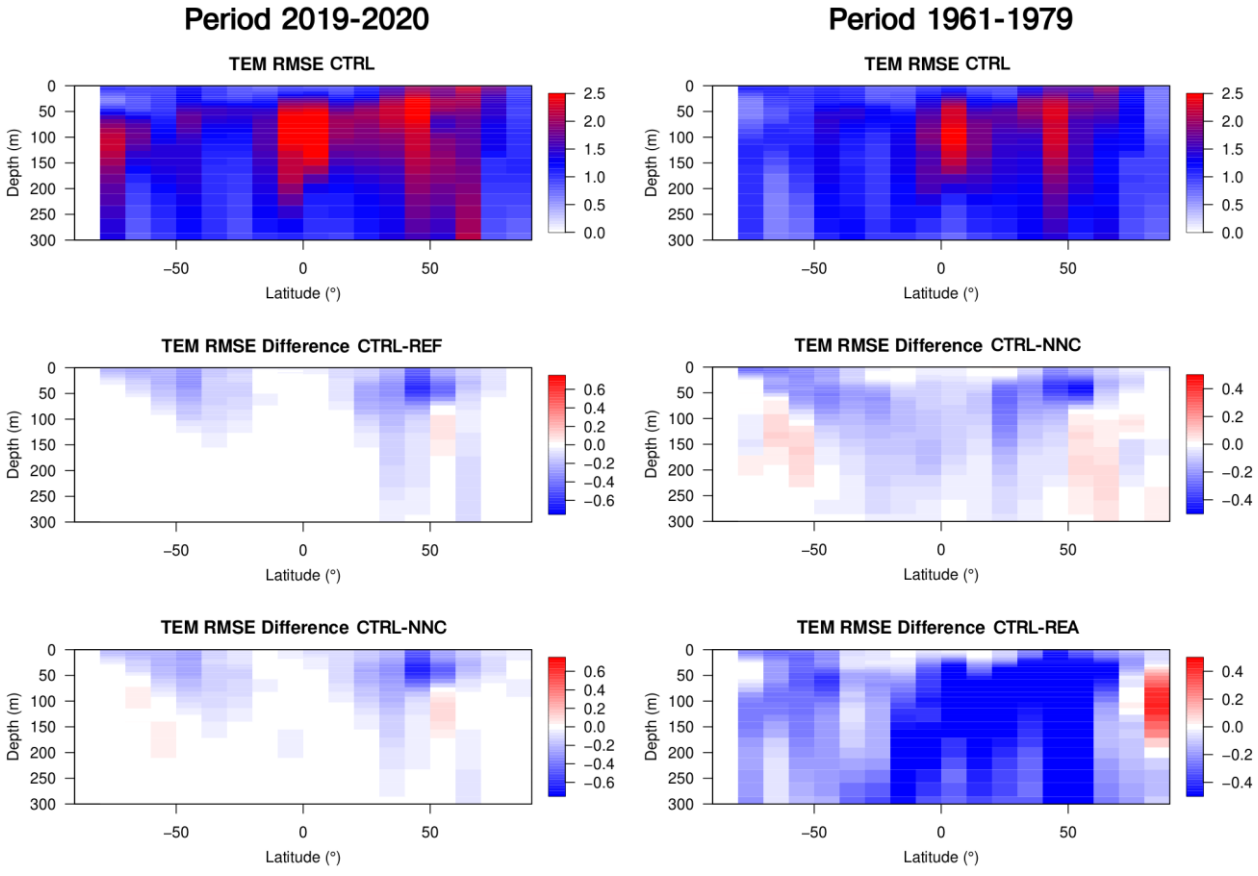


Figure 8. Reconstructed global overturning circulation (in Sverdrups, with 1 Sv = 1E+6 m3 s-1) for the CTRL (a), and as a difference between CTRL and REF (b), or NNC (c) experiments.

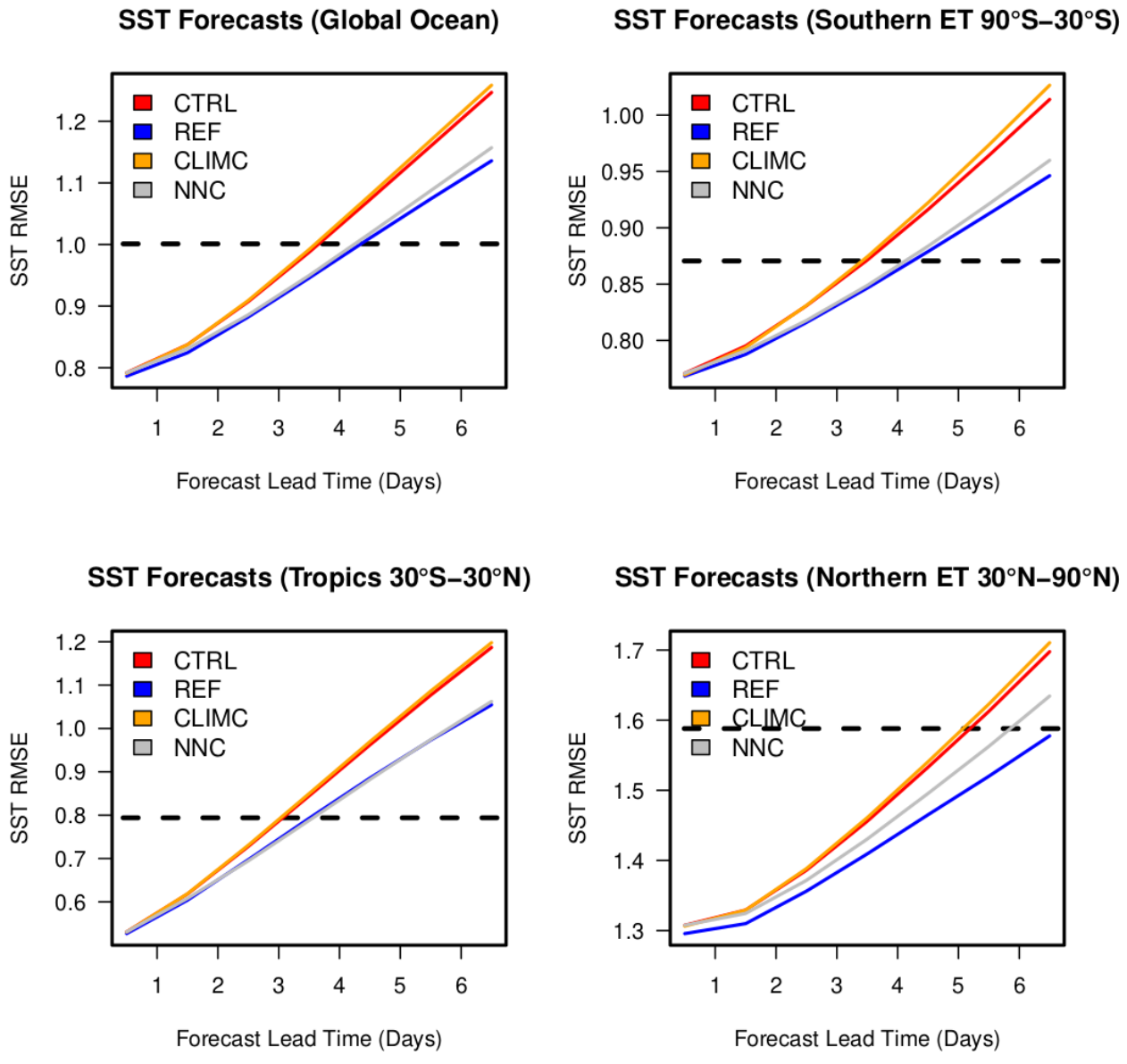
https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024



490 **Figure 9. Temperature RMSE as a function of latitude and depth for the CTRL experiments (top panels) and for the period 2019-2020 (left) and 1961-1979 (right), and differences between CTRL and REF or NNC (left) and NNC or REA (right) for their respective periods. REA is the CIGAR reanalysis.**

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

Geoscientific Model Development Discussions



495

Figure 10. Forecast skill score metrics (RMSE), for sea surface temperature at different latitudinal bands, as a function of forecast lead time, for the experiments presented in the text. The dashed line corresponds to the RMSE of climatology, i.e. for values of RMSE greater than the climatology the forecasts are not useful. Note that the REF experiment is shown as a benchmark, but its
500 **setup cannot be used in operational experiments, as it relies on future observations.**

https://doi.org/10.5194/gmd-2024-185
Preprint. Discussion started: 21 October 2024

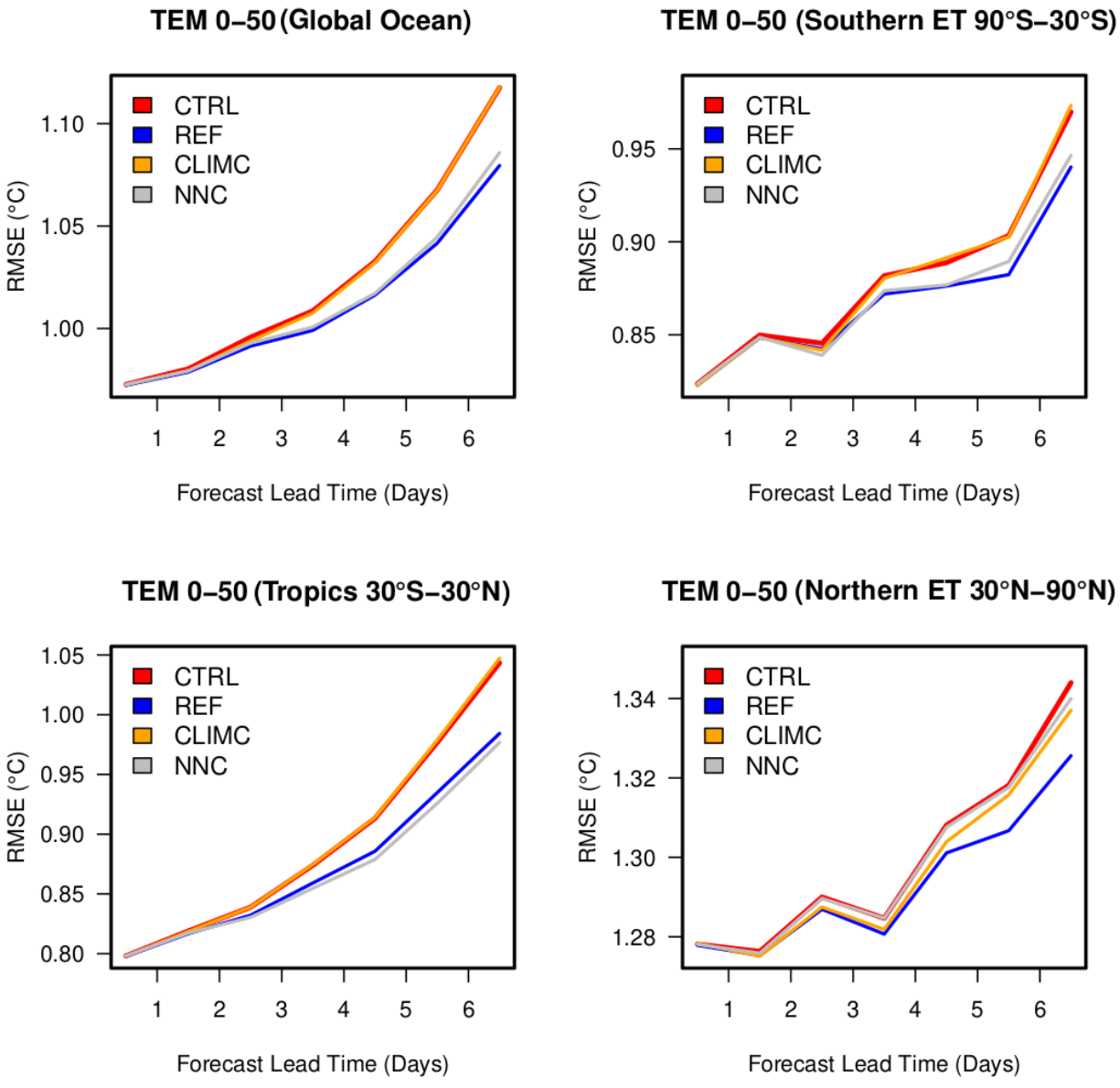


Figure 11. As Figure 10 but for the verification against in-situ profiles in the top 50 m of depth.