

Conditional updates of neural network weights for increased out of training performance

Jan Saynisch-Wagner¹ and Saran Rajendran Sari¹

¹GFZ Helmholtz Centre for Geosciences, Potsdam, Germany

Correspondence: Jan Saynisch-Wagner (saynisch@gfz.de)

Abstract. This study proposes a method to enhance neural network performance when training data and application data are not very similar, e.g., out of distribution problems, as well as pattern and regime shifts. In contrast to previous approaches to out of distribution problems, which alter the inputs and outputs of an operator, our approach alters the operator itself. Here, we especially aim to alter the weights and biases of a neural network to mitigate the out of distribution problem. The method consists of three main steps: 1) ~~Retrain the~~ Fine-tune a trained neural network towards reasonable subsets of the training data set and note down the resulting weight anomalies. 2) Choose reasonable predictors and derive a regression between the predictors and the noted weight anomalies. 3) Extrapolate the weights, and thereby the neural network, to the application data. We show and discuss this method in three nonlinear use cases from the climate sciences, which include successful temporal, spatial ~~and cross-domain~~, and cross-regime extrapolations of neural networks.

1 Introduction

In physics, especially in nonlinear geosciences and climate sciences, the poor performance of neural networks (NN) when applied outside their training distribution or their trained dynamics poses a very strong limitation to their general applicability (Irrgang et al., 2021; Landsberg and Barnes, 2025).

In these fields, physical relations such as laws, dependencies, or sensitivities are commonly derived (or learned) under ~~well observed~~ well-observed conditions and are then applied to ~~less-observed~~ less-observed conditions to gain knowledge about the latter. For example, results from lab or numerical model experiments are regularly applied to ~~real-world~~ real-world problems or observations (e.g., Mehta et al., 2025); knowledge from our Earth and our Solar System ~~are~~ is transferred to other planets and other star systems (e.g., Kvorka et al., 2026); learned relations that are derived today are transferred to the distant past, or to the future (e.g., Eyring et al., 2016; Wang et al., 2024; Koutsodendris et al., 2014). Usually, well-documented problems are transferred, e.g., across country borders (Sebastianelli et al., 2024; Kotz et al., 2021), across spatial scales (Ortega-Cisneros et al., 2025), from the ~~well-observed~~ well-observed upper ocean to the deep ocean (Llovel et al., 2014; Lee and Gentemann,

2018), or from the lower latitudes to the poles (Irrgang et al., 2020); or a mixture of it all (Jung et al., 2024; van der Deure et al., 2025).

From a machine-learning perspective, this means that available training data are typically limited in their spatial and temporal extent, as observations or model simulations only cover specific time periods or regions. For example, Earth observing satellites often leave the higher latitudes unobserved depending on their inclination ~~;~~ ~~oceanographic~~ ~~ARGO~~ (e.g., Li et al., 2025); oceanographic Argo floats measure only the upper 2000 m of the marine water column (e.g., Wong et al., 2020); paleo records such as ice cores may exist for the past but only in regions where there is still ice today (e.g., Bereiter et al., 2015). In general, ~~the majority of measurements since the onset of satellite measurements, the majority of Earth observations~~ cover only the last few decades~~;~~ no measurements exist for the future, and comparatively few measurements exist for the distant past. Furthermore, the environmental conditions that exist on Earth are very heterogeneous (e.g., Peel et al., 2007), and most processes, for instance, the climate, are non-static in nature (e.g., Schannwell et al., 2024). The combination of spatiotemporally-localized data distributions in combination with differing conditions between possible inference regions/times/environments and the application regions/times/environments makes it very difficult for NN to operate (e.g., Rasp et al., 2018; Hernanz et al., 2022; Hsieh, 2023).

In some cases, individual standardization or rescaling of the variables from the application data set is a valid option to mitigate the out of distribution (OOD) problem. However, in many real use cases, the standardization or rescaling removes essential information from the application data. This is true for nonlinear problems where standardizing or rescaling the input will lead to wrong outputs~~as by,~~ as by the definition of non-linearity: $f(a+b) \neq f(a) + f(b)$ and $f(n*a) \neq n*f(a)$. Especially in the climate sciences, the conditions for contemporary ~~AI~~ artificial intelligence (AI) are challenging and raise the demand for new approaches (e.g., Eyring et al., 2024; Fang et al., 2025)(e.g., Eyring et al., 2024; Beucler et al., 2024; Fang et al., 2025; El Ghawi et al., 2025). For example, removing offsets and trends in variables such as sea level and temperature would discard information essential for predicting climate change, climate impacts, and tipping. Likewise, the amplitudes of extreme events should not be rescaled to fit ~~that those~~ of the training data. In addition, standardization and rescaling ~~requires~~ require a certain amount of data points, which are not always available in the application data. Often, the learned relations have to be applied to one single (possibly extreme) event, to a few ice cores, to one other planet, moon, or star system, and so on, i.e., they pose ~~so-called~~ so-called single-shot or few-shot problems in AI terminology. Consequently, in such cases ~~a standardization or rescaling, application-distribution-based standardization, rescaling, re-mapping, or basis transformation~~ (e.g., Beucler et al., 2024) is not possible. Finally, even without any distribution shift or with normalization, the NN can still be confronted with unseen combinations of input values, which can pose a serious OOD problem.

Here, we propose a method to reduce this problem for NN by exploiting the NN's own sensitivity towards samples of the training data to predict NN weights and biases, which are intended to perform better in the OOD application realm (cf. Sec. 2.1, Method).

~~We believe that~~ However, this paper does not discuss if and when NN are suitable choices for nonlinear OOD problems. Furthermore, we will not discuss (in depth) the many other methods that might be able to extrapolate the inputs or outputs of some NN to OOD regions. The reason is that many tasks do require NN, and these other approaches cannot be used, e.g.,

reasoning and decision making, (conditional) generative AI, big data, and large classification problems such as image feature detection. These are applications where the extrapolation or averaging (cf. Hsieh, 2023) of NN output will not be meaningful. To conclude, this study aims at problems where NN are used for various reasons, e.g., ease of use and implementation, existing workflows, compute limitations, invertibility, lack of alternatives, etc., and these NN should be used on OOD data. Still, for completeness, we provide a comparison of our experiment 1 results to several standard extrapolation methods in Appendix A, together with a brief discussion.

We think that AI applications in many fields of physics and other fields of science, such as medicine, biology, economic sciences, or social sciences, can benefit from the proposed method. Nonetheless, we demonstrate and evaluate the proposed methodology in three classes of experiments that are mainly motivated by climate sciences and Earth sciences (cf. Sec. 3, Experiment Design and Data). One experiment, where NN weights are predicted through time, one experiment where NN weights are predicted through space, and one experiment where NN weights are predicted across domain-boundaries regime boundaries. Within similar training and application data sets and domains, the chosen experiments do not represent any challenges to NN training (nor to basic mathematics and physics). The challenges we demonstrate and mitigate here arise only by training the NN in times and places that have very different conditions than the times and places the model has to operate on during application. Consequently, these experiments should be understood entirely as symbols or place-holders for more sophisticated AI application tools.

2 Experiment Design and Data Method

2.1 Experiment 1: Tipping of the Atlantic Meridional Overturning Circulation

We base experiment 1 on van Westen et al. (2024), a paper that shows that critical weakening of the Atlantic Meridional Overturning Circulation's (AMOC) strength can occur in complex climate models, too. This so-called AMOC tipping is triggered by forcing the northern Atlantic Ocean of the respective climate model with increased freshwater flux (e.g., related to the rapid melting of the Greenland Ice Sheet). These so-called hosing experiments are a common tool in climate sciences (e.g., Rahmstorf, 1996; Saynisch et al., 2016). The paper's supplementary material and the data we use here is located in the following repository:-

As input to the NN, we use 2d velocity fields representing vertical slices through the Atlantic Ocean, i.e., the zonal velocity maps (VVEL). These model based horizontal velocity maps are taken always at the same location and are given as annual values. The VVEL data can be downloaded from the above given repository from the folder: . The files cover each 50 years of data, the first one is called and so forth.-

The target output of the NN is the strength of the AMOC given in Sverdrup ($1\text{ Sv} = 10^6 \text{ m}^3/\text{s}$), i.e., a single value per input velocity map. The AMOC strength is defined as the total meridional volume transport at 26°N over the upper 1000 m (cf. eq. 3 in van Westen et al. (2024)). The respective time series is given in-

The challenge in experiment 1 is that the training bases only on data before the AMOC tipping point, occurring around year 1800 of the model data, cf. Fig. 1. However, the NN has to calculate the correct AMOC strength also during tipping and even after tipping, a serious OOD problem.

2.1 Experiment 2: Estimation of Sea Water Density

Experiment 2 bases on the highly nonlinear Equation Of State for sea water (EOS). For every desired oceanic location, the NN input is salinity (S), temperature (T), pressure (P), latitude (ϕ), longitude (λ) and depth (z) and the target output is the local in-situ density (ρ) at this particular location. Consequently, the NN task is to convert 6 numbers into one. For data we use the gridded observations of the World Ocean Atlas (Locarnini et al., 2024; Reagan et al., 2024) which are available under . We use the 5°-decadal averaged annual fields of S and T for every available latitude (36 levels), longitude (72 levels), and depth (102 levels) of the provided 3d ocean grid. These values represent long term averages and no temporal dimension is present in this experiment. We then calculate P and ρ by using the routines `swPressure( $z, \phi$ )` and `swRho( $S, T, P$ )` from the `oce` R-package (Kelley and Richards, 2025) which follow the Thermodynamic Equation Of Seawater – 2010 (TEOS-10) standards (IOC et al., 2010).

Experiment 2 is motivated by the diminishing of oceanic monitoring capabilities with depth. Let’s assume that we have found an interesting relation between our targets and the ARGO float data. Unfortunately, the ARGO floats cover only the upper 2000 m of the water column. However, we want to exploit our findings also below 2000 m. In a practical application case this relation could estimate krill, CO_2 or nutrients. Here, we restrict ourselves to the estimation of sea water density.

In contrast to experiment 1 where knowledge has to be transferred over time (actually, also over to different post-tipping dynamics), the challenge of experiment 2 is to transfer learned knowledge spatially. The training of the NN is restricted to a spatial subset of the world oceans. However, the trained NN has to redo the calculation for the entire ocean where OOD. The proposed weight-prediction method for NN follows the pseudo-code given in Tab. 1. At its core, the approach applies existing extrapolation techniques to the weights of a NN instead of applying them to the input and output values can occur. A further difficulty is that the EOS follows a complicated, highly nonlinear empirical polynomial (cf. Roquet et al., 2015). of NN as other approaches suggest (e.g., Hsieh, 2023; Beucler et al., 2024).

2.1 Experiment 3: Uncertainty Estimation of Global Wind Velocity Reanalyses

The third experiment focuses on cross-domain extrapolation of learned uncertainties in global wind velocity reanalyses. The setup follows the concept proposed by Irrgang et al. (2020), who demonstrated that spatiotemporal uncertainty patterns learned at a single oceanic location can be generalized to the global ocean domain. Here, we extend this approach by transferring the learned uncertainty information from an oceanic reference point to continental regions. This setup represents an OOD scenario, since wind dynamics over land are governed by different boundary layer processes, surface roughness, and convective regimes compared to those over the ocean.

In contrast to Irrgang et al. (2020), who used three atmospheric reanalysis products as model input, we restrict the input data to two widely used reanalyses: ERA5 (Hersbach et al., 2023) from the European Centre for Medium-Range Weather

Forecasts (ECMWF) and CFSv2 (Saha et al., 2014) from the National Centers for Environmental Prediction (NCEP). Data set preparation, NN model architecture, and NN model training follow the same strategy as in the reference study to ensure comparability. This configuration tests whether the learned sensitivities in wind velocity uncertainty over marine environments can meaningfully generalize to terrestrial conditions, thereby illustrating the broader potential of the spatial transferability of the proposed method.

3 Method

The proposed weight-prediction method for NN follows the pseudo-code given in Tab. 1. At its core, the Our approach consists of 3 parts. After the training of a suitable NN ~~to for~~ the problem at hand with a training data set (Tab. 1, step 1), one has to do the following. A: For each weight and bias of this hereafter termed "parent model", collect the ~~weight deviations~~ weight deviations arising from retraining the parent model with single data points or subsets of the training data set (Tab. 1, step 2). B: For each weight and bias of ~~this the~~ parent model, establish a dependency between said weight deviations and suitable predictors, e.g., a linear regression (Tab. 1, step 3). C: For each weight and bias of ~~this the~~ parent model, extrapolate the established dependency towards target-application data points and thus generate new ~~target-data-tailored models~~ application data tailored NN, the "child models" (Tab. 1, step 4). From now on, we summarize a group of child models originating from the same trained parent model (cf. ParentModel₀ in Tab. 1, step 1) under the singular "child model". All models in one child model differ only due to the fact that their weights are predicted onto different target-application data points (cf. ~~x_t and x_a~~ and ChildModel_{t,g} in Tab. 1, step 4), until this step ~~they base~~, they are based on exactly the same data. Consequently, if we use the plural from now on, we refer to different child model groups, e.i.e., groups originating from different ParentModel₀. ~~g., each originating from a different parent model.~~

Note, in Tab. 1 we show a total-valued form of the method. Naturally, handling only the weight anomalies or increments ΔW_{ik} is equally possible.

Table 1: Weight-Prediction Approach (pseudo-code)

1. Train:

- Train a suitable NN on the training data set $\mathbb{I}$, consisting of I data points x_i
- Store $ParentModel_0$

2. Forgetful online learning:

- Loop over a reasonable subset of training data points x_i , e.g., for all $x_i \in \mathbb{I}$:
 - (a) ~~Finetune~~ Fine-tune $ParentModel_0$ on a single training data point x_i , regularized towards $\mathbb{I}$. For example, by compiling the x_i -focused training data set $\mathbb{I}_i$ ~~consisting of n copies of~~, consisting of n copies of a single data point x_i and ~~n~~ n random other training data points $x_{j \neq i} \in \mathbb{I}$.
 - Result: $ParentModel_i$
 - (b) Store $ParentModel_i$'s K weights (including biases) of all layers, nodes, and connections: W_{ik}
 - (c) Reload $ParentModel_0$
 - (d) Repeat from (a)

3. Weight-Regression:

- Choose a suitable regression algorithm, e.g., linear regression $\text{lm}()$
- Choose (and prepare) a respective set of predictors/regressors $\mathbb{R}$, which correspond to the training data points x_i , used in step 2 above ~~and the target~~, and the application data point(s) ~~x_T~~ x_a , used in step 4 below: R_i & ~~$R_T \in \mathbb{R}$~~ $R_a \in \mathbb{R}$, e.g., time steps t_i & ~~t_T~~ t_a
- Loop over a suitable subset of the $ParentModel_0$'s K weights and biases, e.g., for all $k \in \mathbb{K}$:
 - (a) Fit chosen regression model, e.g., $RegressionModel_k = \text{lm}(W_{ik}, \mathbf{R}_T \mathbf{R}_i)$
 - (b) Store $RegressionModel_k$
 - (c) Repeat from (a)

4. Weight-Prediction:

- Loop over a reasonable subset of ~~target data points~~ application data points x_a , e.g., for all ~~$x_T \in \mathbb{T}$~~ $x_a \in \mathbb{A}$:
 - (a) Predict all required weights for ~~target data point~~ application data point x_a : ~~W_{Tka}~~ W_{Tka} = $RegressionModel_k(\mathbf{R}_T \mathbf{R}_a)$
 - (b) Store $ChildModel_{T_a}$
 - (c) Repeat from (a)

5. Actual Application:

- Apply child model to ~~target~~ application data point(s): ~~$ChildModel_{T_a}(x_T)$~~ $ChildModel_{T_a}(x_a)$

The reasoning why the above approach should work is that fine-tuning the $ParentModel_0$ further on a training data point x_i can result in a new model ($ParentModel_i$) whose weights and biases differ a) only slightly and b) in a meaningful way from $ParentModel_0$. Ideally, all $ParentModel_i$ form a point cloud around $ParentModel_0$ in NN parameter space, which encodes the respective sensitivities between this specific $ParentModel_0$'s weights and the input data. For this to be the case, the deviations between $ParentModel_0$ and the related $ParentModel_i$ have to be constrained in a way that no entirely new and therefore unrelated $ParentModel_i$ are created in the process. In other words, all $ParentModel_i$ should be closely yet meaningfully distributed around $ParentModel_0$. To achieve this, it should be theoretically sufficient to restrict the fine-tuning of $ParentModel_0$ (cf. Tab. 1, step 2a) towards each x_i to a few update steps (epochs). However, we found that a regularized fine-tuning of $ParentModel_0$ towards individual training data points x_i works better and allows for more flexible fine-tuning approaches. To regularize the fine-tuning step, we add a representative fraction of the training data to the x_i fine-tuning batch. Naturally, the impact of such a regularization and the impact of the fine-tuning towards a single x_i in generating a new $ParentModel_i$ has to be weighted reasonably. Tab. 1, step 2a gives an example of how to achieve a suitable regularization easily in practice, e.g., we take n random samples from the training data combined with n duplicates of x_i to ensure equal weighting in the fine-tuning batch. Consequently, weighting in the given example is 0.5 for each impact, i.e., equal impact from the entire training set ($\mathbb{I}$) and from a single data point x_i .

Note, in Tab. 1, we show a total-valued form of the method. Naturally, handling only the weight anomalies or increments ΔW_{ik} is equally possible.

The exact execution of each ~~of the mentioned steps can vary~~. However, some given step of the core approach (Tab. 1) can vary substantially and should be adapted to the problem at hand and to domain knowledge if available. For example, two main choices have to be made during the approach ~~in any case~~. First, choices for suitable predictors for the weight-regression (Tab. 1, step 3) and the subsequent weight-prediction (step 4) ~~have to be made~~. It is out of ~~the~~ scope of this paper to discuss these choices here in detail. Ideally ~~the choices~~, ~~the choice~~ should be heavily influenced by prior knowledge of the problem at hand ~~and how it may~~. Especially, how the problem might behave outside the training data range. In principle, the predictors for the weights can be anything, e.g., i) time, ii) (subsets of) the original NN input, iii) features derived from the original input, or iv) additional information not used by the ~~parent~~ NN model. A careful selection and pruning of the predictors is advised. In our examples, we demonstrate the use of ii) in experiment 2, and iii) in ~~experiment~~ experiments 1 and 3. Note that in ~~experiments~~ experiment 2, the inputs of depth, latitude, and longitude are not necessary for the NN to calculate the correct sea water density as long as temperature, salinity, and pressure are given. However, they are promising weight-predictors. As we did not want to limit the ~~parent~~ NN by withholding information, we unified the inputs among ~~parent~~ NN training and weight-regression in experiment 2.

Second, a choice ~~of for the~~ weight-prediction method has to be made. Of these, many suitable exist, e.g., auto-regressive integrated moving average (ARIMA), ~~XGboost~~, eXtreme Gradient Boosting (XGBoost), Gaussian Processes, linear models (LM), generalized linear models (GLM, Nelder and Wedderburn, 1972), generalized additive model (GAM, Hastie and Tibshirani, 1986), support vector machines (SVM, Cortes and Vapnik, 1995), random forest (Ho, 1995), NN itself (cf. Chauhan et al., 2024), and many more. Again, it is not part of this paper to review ~~regressions~~ regression methods and their rightful

application. Nonetheless, we provide a comparison of these methods acting on our toy problems in Appendix A. In experiments 1 and 2, we ~~choose~~chose LM for their ease of use and very fast computation. The latter is especially needed as we do regression and prediction for every weight and bias of every single node of the NN. However, if the NN is very large, low-rank adaptation (LoRA, Hu et al., 2021) or similar decompositions should be considered. We would not generally recommend ~~to use other purely data-driven~~using other purely data-driven methods like RF and NN as they ~~can~~can suffer from poor OOD performance, ~~too.~~Still as already stated (e.g., Rasp et al., 2018; Hernanz et al., 2022; Hsieh, 2023). Still, in some cases, we found improvements in our experiments by these methods. For instance, in experiment 3, we demonstrate the use of a NN for the weight-regression and weight-prediction tasks. In general, if you know a regression method that works well for your problem, domain, or data (cf. Figs. A1 - A3), then this method could be a good candidate for extrapolating the weights of a NN, too (if the method is fast enough, that is).

3 Experiment Design and Data

3.1 Experiment 1: Tipping of the Atlantic Meridional Overturning Circulation

We base experiment 1 on van Westen et al. (2024), a paper that shows that critical weakening of the Atlantic Meridional Overturning Circulation’s (AMOC) strength can occur in complex climate models, too. This so-called AMOC tipping is triggered by forcing the northern Atlantic Ocean of the respective climate model with increased freshwater flux (e.g., related to the rapid melting of the Greenland Ice Sheet). The paper’s supplementary material and the data we use here is located in the following repository: https://github.com/RenevanWesten/SA-AMOC-Collapse/tree/SA-AMOC-Collapse_v1.0

These so-called hosing experiments are a common tool in climate sciences (e.g., Rahmstorf, 1996; Saynisch et al., 2016).

As input to the NN, we use 2d velocity fields representing vertical slices through the Atlantic Ocean, i.e., the zonal velocity maps (VVEL). These model-based horizontal velocity maps are always taken at the same location and are given as annual values. The VVEL data can be downloaded from the above given repository from the folder: `/Data/CESM/Data/AMOC_section_26N/`. The files cover each 50 years of data; the first one is called `CESM_year_0001-0050.nc` (and so forth).

The target output of the NN is the strength of the AMOC given in Sverdrup ($1 \text{ Sv} = 10^6 \text{ m}^3/\text{s}$), i.e., a single value per input velocity map. The AMOC strength is defined as the total meridional volume transport at 26°N over the upper 1000 m (cf. eq. 3 in van Westen et al. (2024)). The respective AMOC strength time series is given in `/Data/CESM/Ocean/AMOC_transport_depth_0-1000m.nc`

The challenge in experiment 1 is that the training is based only on data before the AMOC tipping point, occurring around year 1800 of the model data, cf. Fig. 1. However, the NN has to calculate the correct AMOC strength also during tipping and even after tipping, a serious OOD problem.

3.2 Experiment 2: Estimation of Sea Water Density

Experiment 2 is based on the highly nonlinear Equation Of State for sea water (EOS). For every desired oceanic location, the NN input is salinity (S), temperature (T), pressure (P), latitude (ϕ), longitude (λ), and depth (z). The target output is the local in-situ density (ρ) at this particular location. Consequently, the NN task is to convert 6 numbers into one. For data, we use the gridded observations of the World Ocean Atlas (Locarnini et al., 2024; Reagan et al., 2024), which are available under <https://www.ncei.noaa.gov/access/world-ocean-atlas-2023/>. We use the 5° decadal averaged annual fields of S and T for every available latitude (36 levels), longitude (72 levels), and depth (102 levels) of the provided 3d ocean grid. These values represent long-term averages, and no temporal dimension is present in this experiment. We then calculate P and ρ by using the routines `swPressure(z,  $\phi$ )` and `swRho(S, T, P)` from the `oce` R-package (Kelley and Richards, 2025), which follow the Thermodynamic Equation Of SeaWater - 2010 (TEOS-10) standards (IOC et al., 2010).

Experiment 2 is motivated by the diminishing of oceanic monitoring capabilities with depth. Let's assume that we have found an interesting relation between our targets and the Argo float data as input. Unfortunately, the Argo floats cover only the upper 2000 m of the water column (Wong et al., 2020). However, we want to exploit our findings also below 2000 m. In a practical application case, this relation could estimate krill, CO₂, or nutrients. Here, we restrict ourselves to the estimation of sea water density.

In contrast to experiment 1 where knowledge has to be transferred over time (actually, also over to different post-tipping dynamics), the challenge of experiment 2 is to transfer learned knowledge spatially. The training of the NN is restricted to a spatial (upper) subset of the world oceans. However, the trained NN has to redo the calculation for the entire ocean where OOD input and OOD output values occur. A further difficulty is that the EOS follows a complicated, highly nonlinear empirical polynomial (cf. Roquet et al., 2015).

3.3 Experiment 3: Uncertainty Estimation of Global Wind Velocity Reanalyses

The third experiment focuses on cross-regime extrapolation of learned uncertainties in global wind velocity reanalyses. The setup follows the concept proposed by Irrgang et al. (2020), who demonstrated that spatiotemporal uncertainty patterns learned at a single oceanic location can be generalized to the global ocean domain. Here, we extend this approach by transferring the learned uncertainty information from a single oceanic reference point to continental regions, too. Since wind dynamics over land are governed by different boundary-layer processes, surface roughness, and convective regimes compared to those over the ocean, this setup represents a severe OOD scenario.

In contrast to Irrgang et al. (2020), who used three atmospheric reanalysis products as model input, we restrict the input data to two widely used reanalyses: ERA5 (Hersbach et al., 2023) from the European Centre for Medium-Range Weather Forecasts (ECMWF) and CFSv2 (Saha et al., 2014) from the National Centers for Environmental Prediction (NCEP). Data set preparation, NN model architecture, and NN model training follow the same strategy as in the reference study to ensure comparability. This configuration tests whether the learned sensitivities in wind velocity uncertainty over marine environments can meaningfully generalize to terrestrial conditions, too.

4 Results and Discussion

4.1 Experiment 1: Tipping of the Atlantic Meridional Overturning Circulation

Table 2: Notable deviations from the examples given in Tab. 1:

Skipped step 2c (reload $ParentModel_0$)

EOF of the $ParentModel_0$ inputs are used as weight predictors

To predict AMOC strength from velocity maps, we use a standard shallow convolutional NN (CNN), with 2 convolution-
245 all/max pooling layers followed by two dense layers. The last layer has linear activation ~~while~~, and all other layers have the
Exponential Linear Unit (ELU) activation. For the ~~online learning fine-tuning regularization~~ (Tab. 1, step 2a), ~~n~~ is set to 200.

For the regression of the NN weights, the predictors are based on the decomposition into empirical orthogonal functions
(EOF) of the ~~training data itself~~ NN's training data inputs. The principal components of the 4 leading EOF, i.e., 4 numbers per
time step (respectively per velocity field), are used for the regression of the NN weights. Outside the training data, exactly
250 the same EOF-basis as during training is used to decompose the ~~respective~~ velocity maps. Another possible set of predictors
could be mean, maximum, minimum, and standard deviation of each individual velocity map, i.e., again 4 numbers per time
step, or even time itself if a suitable ~~low-dimensional~~ low-dimensional bifurcation model were to be employed for the NN
weight-prediction (cf. Rahmstorf, 1996). Note, however, that the sensitivity of the weights is not dictated by the domain
dynamics alone but also depends on the involved activation functions (and not only the activation function from the node under
255 consideration). For example, if a complex nonlinear change of model output is required by the application domain dynamics,
it can, in principle, be realized by a less complex (e.g., linear) weight change when nonlinear activation functions are involved.
Furthermore, the required weight changes can be distributed over several layers, which then will entangle all contributing
activation functions.

As there is a clear order of the data in experiment 1, i.e., the temporal order, we skip the "reload ~~$ParentModel_0$~~ $ParentModel_0$ "
260 step of the approach (cf. Tab. 1, step 2c). This way, the online learning is not forgetful, and the weights will evolve ~~naturally~~
~~along-time~~ continuously along the time axis as the AMOC system does itself. This improves the results in this experiment
as it reduces noise in the ~~weight sensitivities~~ (generation of the weight sensitivities W_{ik}) ~~generated during step 2.~~. After
establishing weight-predictor relations for every weight and bias of the CNN, ~~the child model was including the CNN layers,~~
the child model is predicted with sub-models (cf. $ChildModel_{t_g}$) for every single time step of the problem by using the same
265 weight-predictor relations. ~~As a consequence, each of these sub-models was then applied only to one dedicated data point.~~ Each
of these sub-models is then applied to one single data point, i.e., time step, only.

Figure 1 shows the time series of the target value (black line), the output of the parent model (red line), and the output of the
same model after the application of the weight-prediction approach (orange line), i.e., the child model. The training data starts
at year 1, and the end of the training data (EOT) is shown by a vertical dashed line, here at 800 years. After the tipping of the
270 AMOC sets in at around 1800 years, the parent model has clearly problems reproducing the black line. Its output is strongly
biased ~~in the direction of the~~ towards the output it had learned during training. The child model surpasses the parent model in
its performance during and after tipping.

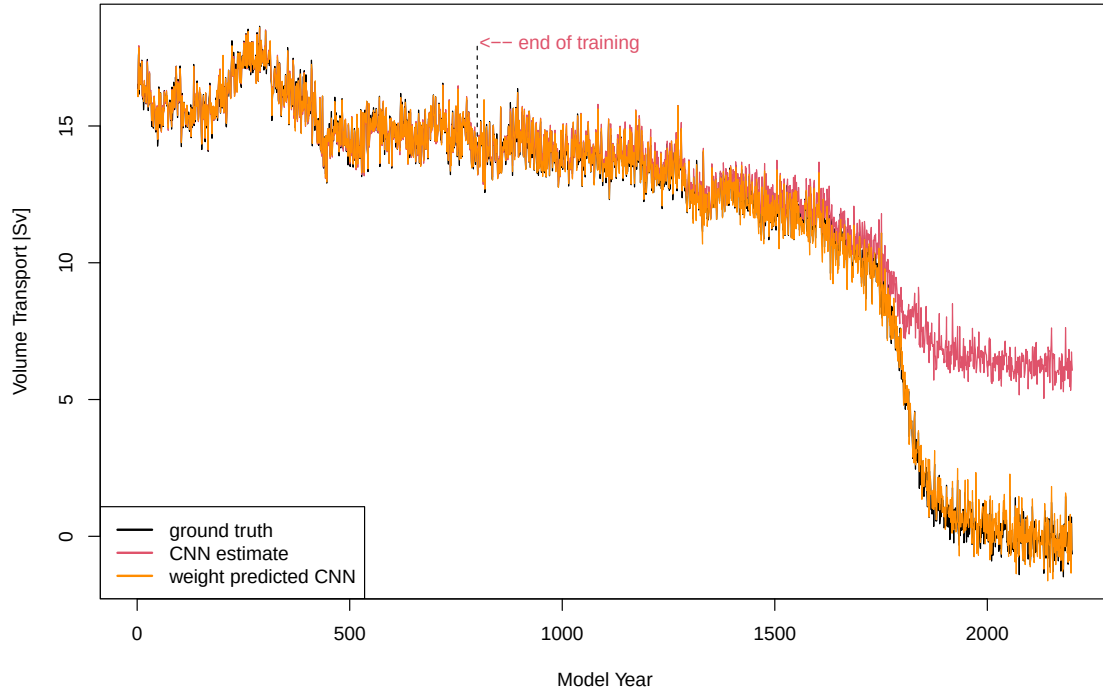


Figure 1. AMOC strength and an example of good performance of the weight-prediction method. Black Line: AMOC overturning decay and tipping, i.e., the target ground truth values. Red Line: Output of parent model trained until the dashed end of training (EOT) line. Orange Line: Output of weight-predicted child model. ~~Forgetful-online~~ Online learning and subsequent fitting of EOF-to-weight relations ~~use-used~~ only data before the dashed EOT line.

Please note ~~, that since ParentModel₀ was that since~~ ParentModel₀ is not reset during step ~~2-2c~~ of Tab. 1, at the end of step 2 we get a NN model that has seen the training data more often than ~~ParentModel₀. Still~~ ParentModel₀. Still, the performance of this model (not shown) is the same as that of ~~ParentModel₀~~ ParentModel₀, and the improvements plotted in Fig. 1 must be attributed entirely to steps 3, 4, and 5 of our approach.

Since in general, the training of NN can have largely varying results, we ~~repeat~~ repeated the experiment several hundred times with randomly varying EOT times, while the beginning of the training window ~~remains~~ remained fixed at year 1. As a consequence, the amount of data each NN ~~was-is~~ trained on varies as well as how close the EOT comes to the actual tipping event. Furthermore, two weight-regression methods are tested. The results are aggregated in Fig. 2.

The figure shows root mean squared errors (RMSE) of parent models (yellow boxes), child models (green boxes), and their pairwise differences (blue boxes). For a successful weight-prediction, the child models should show lower RMSE than the parent models, and the pairwise differences should show as positive values as possible. The variations summarized by each box ~~bases-are based~~ are based on the rather random generation of parent models in combination with the random selection of an EOT between years 800 and 1800. In each panel, three horizontal subdivisions are shown where the RMSE are calculated for 3

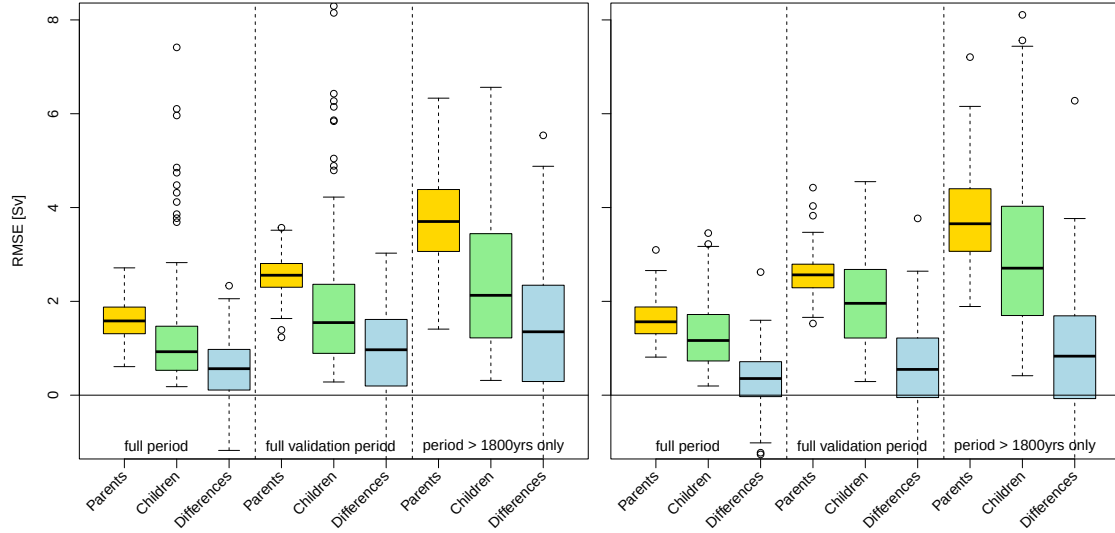


Figure 2. Impact of the choice of the weight-prediction method on the predicted-child model's performance. Polynomials of order 1 (left panel) and order 2 (right panel) where-were fitted to the leading 4 principal components of the input variable (i.e., the meridional velocity maps). Yellow Boxes: RMSE of unperturbed-parent models. Green Boxes: RMSE of weight-predicted-child models. Blue Boxes: Pairwise RMSE differences (i.e., unperturbed "parent" model minus corresponding weight-predicted "child" model). Positive values represent improvements due to the presented weight-prediction approach. Note that "full period" and "period >1800" is the same for every simulation of the ensemble. The term "full validation period" refers to all data point-points after each simulation's individual EOT (which is randomly chosen for every simulation before training). Each box bases-is based on 300 parents/children/pairs.

different time windows: the full 2200yrs of data including the training data (left), the time beyond each NN's individual EOT, i.e., everything except the training data (middle), and a fixed time window covering years 1800-2200, i.e., the tipping event and after (right).

The two panels of Fig. 2 are both based on a linear model for the weight-regression. The standard R function `lm()` was-is used here. The difference between the two panels is the order of polynomials built from the predictors before weight-fitting. While the left panel uses polynomials of order 1, the right panel uses polynomials of order up to 2. As mentioned, the weight-regression bases-on EOF-decomposition-is based on the EOF decomposition of the 2d meridional velocity fields that form the input for the NN itself. The predictors are always the 4 principal components associated with the 4 leading EOF as derived in the training time window.

One can see that in all cases the (green) child model boxes show lower RMSE than the (yellow) parent model boxes, leading to positive pairwise RMSE differences depicted as blue boxes. In addition, the performance gain of the child models compared to their parents increases when the models face conditions that are increasingly different from their training data. The average RMSE of both-parent and child -boxes become larger from left to right in each of the two panels. From left to right, the RMSE in the subdivisions increasingly focus-focuses on the OOD performance of the NN. However, the RMSE increase stronger

300 in the parent models than in the child models, resulting in larger pairwise improvements when the tipping is approached and surpassed.

Since the regressions of polynomials of order 1 and ~~order-2~~ give comparable results ~~with no clear preference so far~~ (with a slight advantage of the order 1 results), this could be interpreted as a sign of robustness of the approach towards the regression method. However, polynomials of even ~~higher-order~~ higher-order (as far as tested) gave worse results (not shown). Here, the
305 higher order terms get random/noise-based slopes/gradients assigned to them during the regression (Tab. 1, step 3) as the ~~resulting~~ higher-order terms have only a negligible influence within the training data ~~(where the regression is conducted)~~. Nonetheless, these terms can still lead to significant yet wrong influences during and after tipping in step 4 of Tab. ~~1-1~~, the weight prediction. More research on how to stabilize the child generation is needed. A signal-to-noise ~~ration~~ ratio guided pruning of insignificant slopes returned by the regression step could help with this problem.

310 In general, as can be seen in Fig. 2, the approach's success can vary considerably, resulting ~~even in worse results of in even worse results from~~ some child models. The same mechanism described above for higher polynomials ~~probably is is probably~~ the reason for the poor results in some of the polynomials of order 1 and ~~order-2~~ cases, too. On average, the results are positive as the ensemble mean of the child models shows a clear improvement over the ensemble mean of the parent models. More research is needed to stabilize the results ~~—~~ either by identifying parent models unfit for the approach, or if possible,
315 by subjecting child models to forms of validation before their application. Until then, ~~an-the~~ ensemble approach should be considered.

In the following, we will discuss some noteworthy examples from the ensemble simulations aggregated in Fig. 2. Apart from the many good results that look more or less similar to Fig. 1, it can happen that the regression shows too little sensitivity to the predictors and the parent, and child model outputs are very similar.

320 How ~~does it come it comes~~ to this poor performance is rather unclear at the moment and needs more investigation. The most likely explanation is that the respective parent model sits either in a deep narrow minimum or a very flat part of the loss hyperplane, and due to this, the proposed ~~forgetful~~ online learning resulted only in minimal variations of the weights (ΔW_{ik}). As the particular parent model is not sensitive to the online training step (Tab. 1, step 2), the weight-regression and the subsequent weight-prediction inherit this limitation. Still, the small induced weight variations are not pure noise, and the
325 weight-prediction generates a ~~better performing better performing~~ child model, even if it is very similar to its parent (not shown).

Furthermore, in some rare cases, deterioration can happen when a child model is applied to data similar to the training conditions, as shown in Fig. 3. The child model (orange line) bends too early towards the post-tipping output and then overshoots. As a consequence, in the ~~500-year~~ 500-year period following the EOT (at year 1200) the parent model (red line) slightly
330 outperforms the child model (orange line). When calculating the RMSE for the entire post-EOT time, the child model still outperforms its parent model.

Figure 4 ~~bases is based~~ on the same ~~data results~~ as Fig. 2, but here the RMSE are sorted by EOT. In addition, the figure shows only the RMSE for the tipping and the past tipping period. While in the view of Fig. 2, the polynomial of order 1 and order 2 based regressions did look very similar, here in Fig. 4, some large differences become evident. If trained only very far

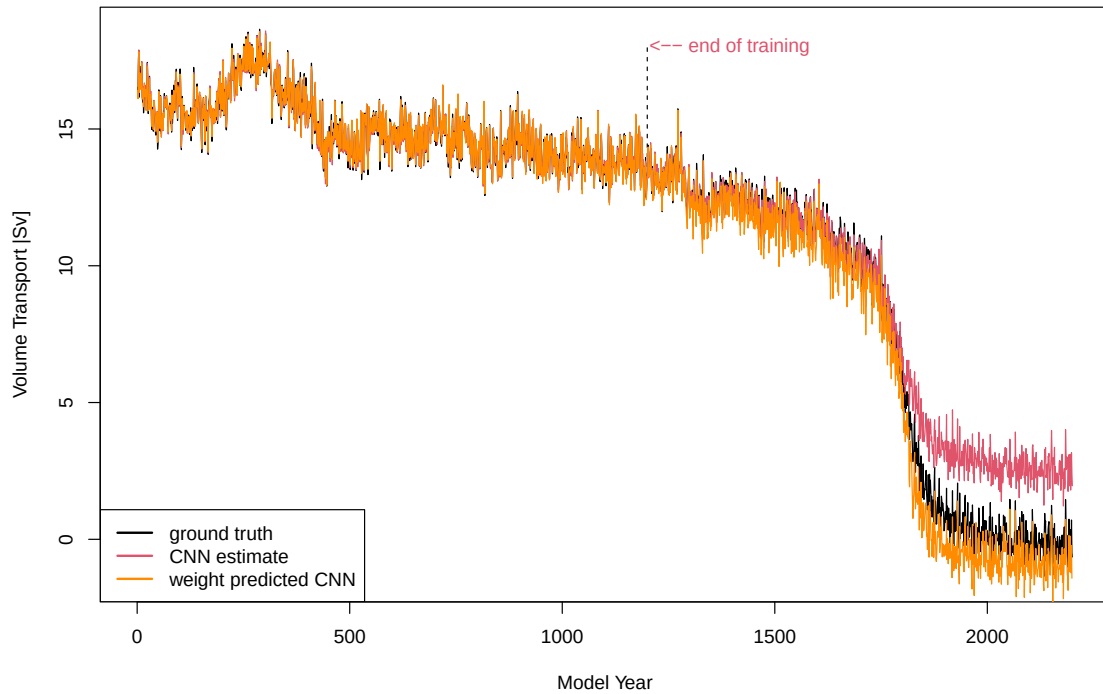


Figure 3. As Fig. 1, but with a sub-optimal example of child model performance. Still, the predicted child model (orange) **outperform** outperforms the parent model (red) beyond the EOT threshold in the RMSE sense.

335 from the tipping point, the regression based on ~~higher-order~~ higher-order polynomials leads to wrong results. This strengthens the argument for the fail-mechanism already described, i.e., that ~~higher-order~~ higher-order terms get noise-based gradients assigned to them as they are not varying significantly during training and regression. When they then do start to vary during tipping, they will give large yet wrong contributions to the weight prediction, and this way will deteriorate the child model and its output.

340 In general, the skill of all models (parents and children) improves with larger EOT. This is not only due to more training data per se, but rather that the training data then also includes information located closer to the tipping point. Consequently, the models that are trained up-until tipping perform best, i.e., all yellow boxes go down from left to right in each panel of Fig. 4. On average, even as children and parents show increasingly similar performances with EOT closer to the tipping point, ~~on-average the~~ the on average better RMSE of the child models are never reached by the parent models.

345 A note on temporal data: In experiment 1, we shuffle the training data before applying the training/validation split. In some problems, this is not possible, either due to the choice of architecture (as LSTM) or not advised due to correlations within time (e.g., Schnaubelt, 2019; Bergmeir et al., 2018). If data shuffling is not an option but a test/validation split is still necessary, then the training data is pushed even further away from the tipping point by the presence of the validation window. As a result,

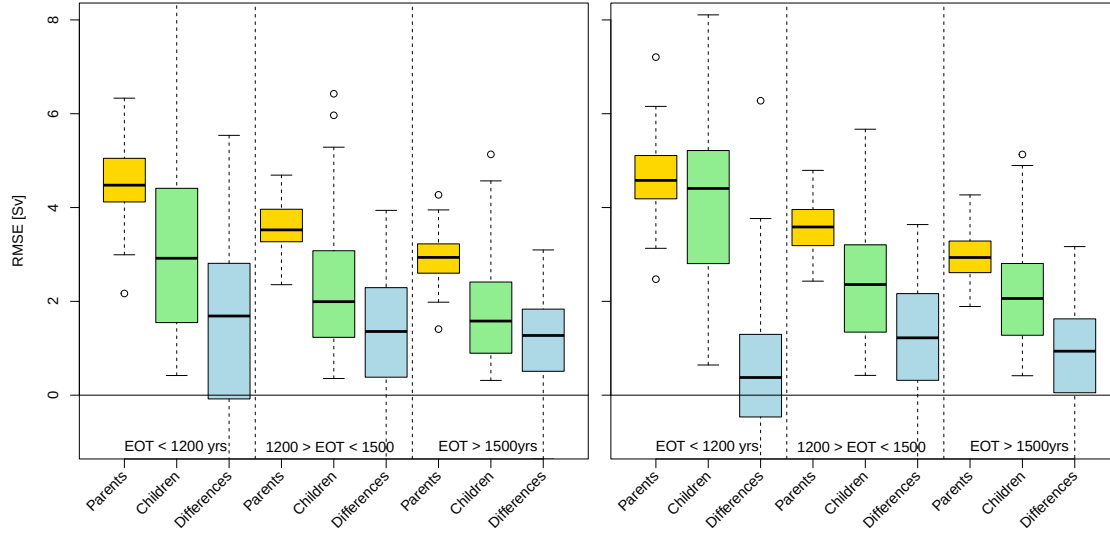


Figure 4. Impact of EOT on the RMSE of a fixed target-application range (1800yrs,2000yrs]. Polynomials of order 1 (left panel) and order 2 (right panel) where were fitted to the leading 4 principal components of the input variable (i.e., the meridional velocity maps). Yellow Boxes: RMSE of unperturbed-parent models. Green Boxes: RMSE of weight-predicted-child models. Blue Boxes: Pairwise RMSE differences (i.e., unperturbed "parent" model minus corresponding weight-predicted "child" model). Positive values represent improvements due to the presented weight-prediction approach. Each box bases-is based on 100 parents/children/pairs.

the parent models will show an even worse performance. Consequently, our weight-prediction approach will become more
 350 advantageous as it also uses the validation data located closer to the tipping point for the weight-regression.

4.2 Experiment 2: Estimation of Sea Water Density

Table 3: Notable deviations from the examples given in Tab. 1:

Quantile approach to sample only a fraction of the training data set $\mathbb{I}$ in step 2a
No regularization towards $\mathbb{I}$ is applied during the fine-tuning step 2a (epochs are limited to 4)
NN inputs (S, T, P, ϕ , λ , and z) are used as weight predictors as well

The NN of choice here is a simple two-layer-two-layer feed-forward NN with ELU activation functions in the first layer and linear activation in the final layer. The regression of the NN weights uses the same set of variables as the NN input, i.e., S, T,
 355 P, ϕ , λ , and z (cf Sec. 3.2). For the forgetful online learning (Tab. 1, step 2a), we finetune-fine-tune the parent model only on individual x_i without the regularization towards the entire training data set $\mathbb{I}$. The parent-to-child model differences (ΔW_{ik}) were limited by limiting the fine-tuning iterations (here: epochs=4), only.

In experiment 2, the training data set is very large, and thus retraining the parent-model-parent model on every training data point is prohibitively expensive. A random sub-sampling of the training set could lead to undersampling with respect to
 360 some of the predictors if the random draw is not sufficiently large. To be efficient and at the same time avoid undersampling the

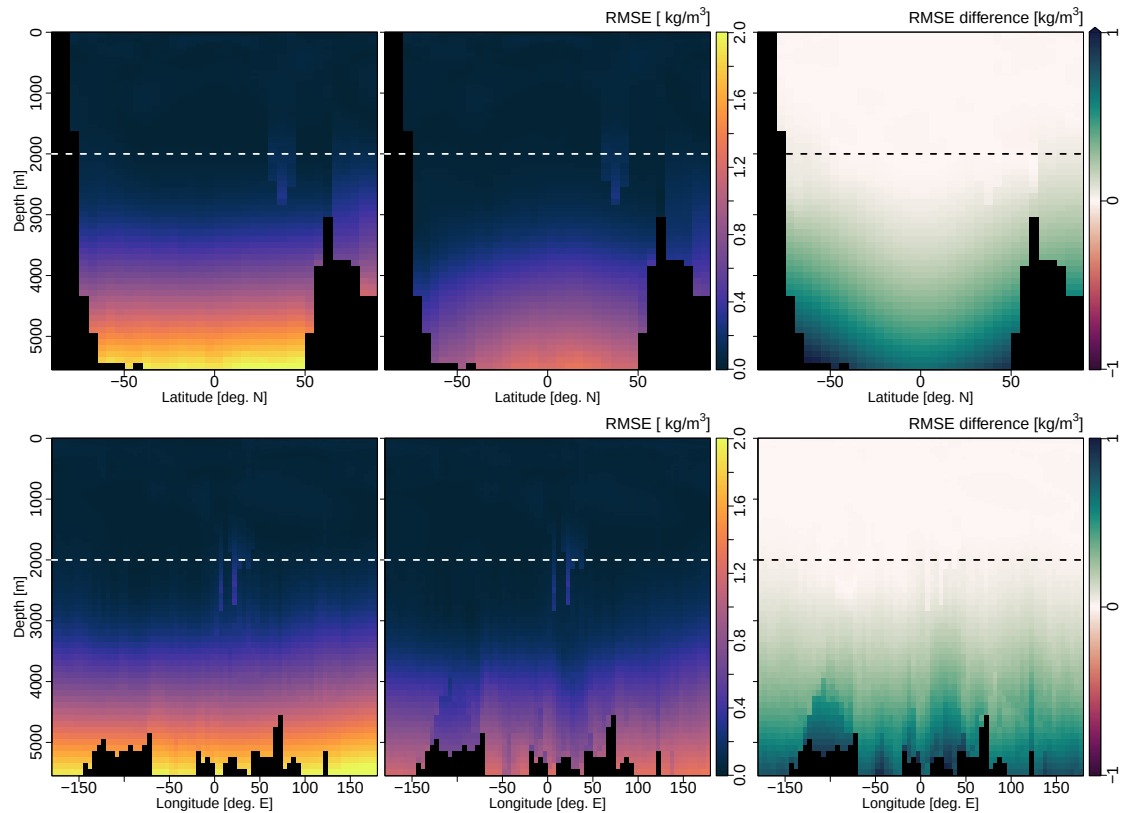


Figure 5. ~~Depth-dependent zonally-averaged~~Depth-dependent zonally-averaged* (upper panels) and meridionally averaged* (lower panels) RMSE of ~~NN-based~~NN-based sea water density estimation after training above the dashed line only. Left panels: Performance of the parent model. Middle panels: Performance of weight-predicted child model. Right panels: RMSE differences (parent minus child, the green color represents improvements by the child model). *: In giving every global output location equal weight, these averages represent a data-centered view and do not follow oceanographic standards.

predictors, we sort the training data into quantiles (e.g., 300) for every predictor. Then we draw several x_i from every parameter-specific ~~quantile-interval-for-finetuning~~quantile interval for fine-tuning the parent model (Tab. 1, step 2). If it is suspected that special functions (or interactions) of predictors should play a role in the weight-regression, then sampling quantiles of the respective function-values (or interactions) is advised (e.g., one has not only to guarantee sufficient sampling of the predictors
 365 x and y , but also of $\log(x)$, y^2 , and $x*y$ etc.).

As in experiment 1, individual $ChildModel_{t_q}$ are generated for every single ~~target data point x_t , here~~application data point x_a , here, every grid point of the 3d ocean grid.

Figure 5 shows the impact of the weight-prediction approach on a spatial OOD problem, where the parent model is only trained above the dashed line at 2000 m.

370 As a result, the weight-prediction approach improves the performance of the NN sea water density calculation in the deep ocean considerably. At depth, the improvements amount to 2 kg/m^3 in the global average, which is substantial by oceanographic standards. Below the EOT of 2000 m, the child model improves the RMSE of the parent model by 50-100%. Regions where the parent model already performs well show less or no improvements. As already discussed in experiment 1, even slightly worse RMSE can be found if the approach is applied to the training data. In contrast to experiment 1, the ~~result~~results of experiment 375 2 appear to be very stable (not shown).

4.3 Experiment 3: Uncertainty Estimation of Global Wind Velocity Reanalyses

Table 4: Notable deviations from the examples given in Tab. 1:

A single NN is used instead of several RegressionModel_k in steps 3a & 4a
Only the final layer of ParentModel_0 is changed to generate the ParentModel_i and ChildModel_a
EOF of the ParentModel_0 inputs are used as weight predictors

As outlined in Sec. 3.3, this experiment assesses the proposed weight-prediction framework in the context of global wind velocity uncertainty estimation. The focus is on evaluating whether knowledge about uncertainty characteristics learned at 380 one oceanic location can be extrapolated to terrestrial regions with different atmospheric dynamics. This task represents a demanding OOD test case for the proposed method.

Unlike the previous experiments, the child model generation (cf. Tab. 1, steps 3 and 4) is conducted by a NN, too. This choice of a NN-based child model generation underscores the significance of the underlying weight-prediction methodology over ~~regression-model-architectures~~the regression model choice. The child model generation is implemented as a NN consisting 385 of two fully connected hidden layers with 32 nodes each and parameter-specific output heads. Instead of predicting all network weights and biases of the child model, we restrict the weight-regression and prediction to the final fully connected layer, while keeping the remaining parameters identical to those of the parent model. For this to work correctly, the not-to-be-predicted weights and biases have to be set as `trainable=False` already during the online-learning step (Tab. 1, step 2). The use of multiple output heads, together with a scaled mean-squared-error (MSE) loss, mitigates scaling disparities across the predicted 390 weights and biases. All layers use ELU activation functions, and the Adam optimizer (learning rate = 0.001) ~~are~~is used.

The training and evaluation periods span 2012–2017, using six-hourly data from two reanalysis products (ERA5 and CFSv2, cf. Sec. 3.3). Data from 2012–2016 are used for training, January–June 2017 for validation, and July–December 2017 for testing. All results reported below correspond to this testing period. From each of the two input data sets, the principal components of the four leading EOF, yielding eight predictor values in total per time step, are used as weight-regressors. For the 395 regularization of the forgetful online learning (Tab. 1, step 2a), ~~n~~n is set to 200. The parent model is trained at a single oceanic grid point located at 180° E , 60° S , while the child model predictions are evaluated over the entire land mass to assess the method’s ability to generalize beyond the training ~~domain~~regime. As in the other experiments, individual ChildModel_{t_a} are generated for every single ~~target data point~~target data point x_t application data point x_a , that is, every grid point of the 2d globe.

Figure 6 summarizes the spatial distribution of the RMSE in wind speed uncertainty prediction. Panel (a) shows the RMSE 400 obtained ~~with the weight predicted by the child~~model trained only at the oceanic reference point (black dot) and panel (b)

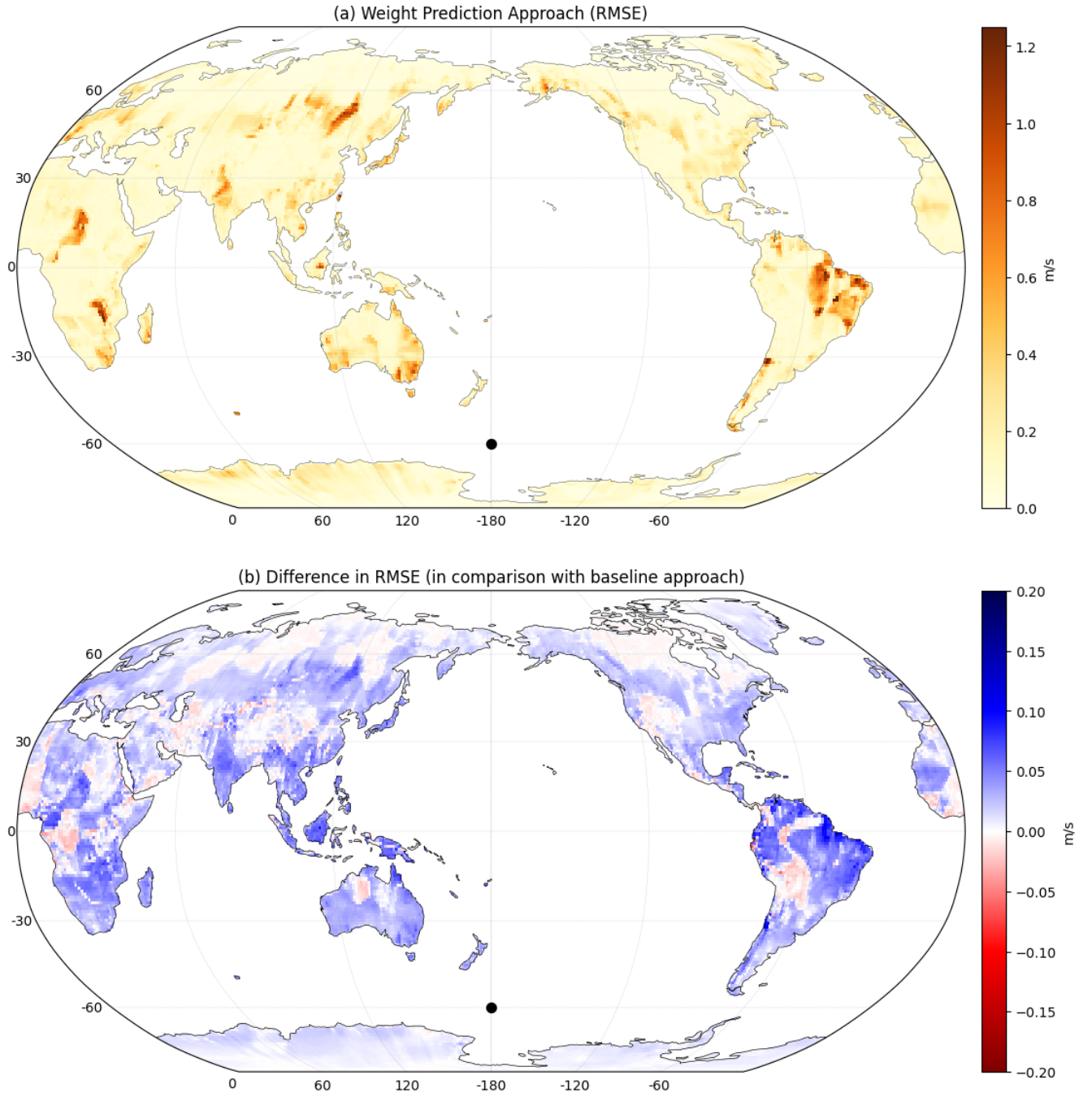


Figure 6. Spatial distribution of RMSE in the prediction of wind-speed uncertainty using the proposed weight-prediction framework. (a) RMSE of the ~~weight-predicted-neural-network-child model~~ trained at the reference oceanic location (180° E, 60° S), represented by ~~the~~ black dot. (b) Difference in RMSE between the parent model and the ~~proposed method, where blue-child model~~. Blue regions indicate ~~improvement~~improvements.

illustrates the parent minus child model RMSE differences. Blue areas in Fig. 6(b) mark regions where the proposed method outperforms the parent model, i.e., where RMSE ~~is~~are reduced by the weight-prediction method.

From these results, it is evident that the weight-prediction approach substantially enhances the wind velocity uncertainty estimation over continental regions, particularly where the parent model struggles the most. Notable improvements are observed
405 over the southeast coast of Australia, eastern Brazil, the East African Highlands, the Indonesian archipelago, and southeastern Russia. In contrast, regions where the parent model already performs well show no major improvement, as expected. This applies ~~for~~to most of the oceanic region as well (not shown in the figure). This behavior ~~,~~suggests that the proposed method provides the greatest benefit under conditions of strong distributional shift, where the parent model’s learned representations are least transferable (cf. ~~experiment~~experiments 1 and 2).

410 Quantitatively, the terrestrial global mean RMSE of the uncertainty estimates decreases from 0.1236 m/s to 0.1104 m/s, corresponding to an average improvement of approximately 10.7%. Repeated experiments indicate that this value varies between -2% and 14%, with a mean improvement of 5.25%, reflecting again the stochastic variability inherent in NN training. Despite this variability, the consistent tendency across repeated experiments confirms that the proposed method systematically improves the predictive performance of the parent model, particularly in regions where generalization is most challenging.

415 That a NN can be used to improve the OOD problem of another NN seems contradictory at ~~best~~first. Overall, the terrestrial input is equally OOD to the $ParentModel_0$ as well as to the child-generating NN. One could ask why the child-generating NN is not limited by OOD, and why then not one NN alone can handle the problem. A full answer to these questions remains open to further research. So far, we ~~only can~~can only argue that ~~we as user know that in experiment 3 the~~, as users, know if an OOD task we gave to ~~the~~a NN remains generally the same but has to be adapted to perform outside the training realm. Consequently,
420 splitting an OOD-experiment into problem-solving and problem-adaptation represents a form of structural implementation of prior knowledge. This understanding, the required task separation, and the subsequent training of both NN, a single NN cannot (yet) develop itself. Furthermore, as already discussed, a nonlinear regime shift in a NN’s in- and output data could be less nonlinear in the weight-space of the same NN when nonlinear activation functions are involved.

5 Summary and Final Remarks

425 In this study, a general method is presented that aims at improving neural network (NN) performance outside of their training data distribution (OOD). The method is proposed for problems where NN are used for one of the various reasons and re-normalization or standardization of the inputs or outputs is not possible due to limited data in few-shot problems or due to governing nonlinear relations, e.g., in climate sciences. Additionally, OOD is meant in a wider sense here. Even when all input variables have the same distribution within and out of training, a trained NN can operate on uncharted territory due to unseen
430 value-combinations and interactions of the input variables.

After the training of a NN, which we call "parent model", on a specific problem, the method requires the following three steps: First, the method collects the series of weight anomalies that arise by online-relearning or fine-tuning ~~within the~~of the parent model towards subsets of the training data. Second, a relation (e.g., a linear regression) is to be established between these

weight anomalies and suitable predictors (e.g., the input variables of the training data set itself). Third, the established relation
435 is then used to generate new weights and thus a new NN that ~~correspond~~corresponds to an application data set by employing
the same predictors but now with values that correspond to the application data set (e.g., the application data itself). In ~~analogue~~
analogy to the parent model, we call this new model the "child model". A child model can encompass a group of NN as a new
NN can be generated for every single application data point ~~or location~~(time step or location). However, sub-setting is possible
in all steps of the approach. For example, only one new NN could be generated per cluster of the application data set. Likewise,
440 the weight sensitivities could be derived from quantiles or clusters of training data only.

In short, the approach is generating NN fine-tuned on subsets of the training data and then is extrapolating along these NN to
data-regions outside of the training data distribution. This way, the network's learned tasks can be transferred, e.g., to different
geographic locations, future climates, or across (physical) regimes. Note, reasonable choices for the mentioned subsets should
depend on the self-similarity of the training data, respectively, the application data has. For example, if the training data shows
445 notable clusters, fine-tuning on these clusters should be sufficient. Furthermore, if the application data is rather homogeneous,
one does not need to calculate a single child model for every application data point.

The method is demonstrated on 3 very different examples from Earth sciences, ~~;~~ one temporal problem, one spatial prob-
lem, and one spatiotemporal problem. These examples and their execution demonstrate further the possible flexibility ~~in-the~~
of the core approach in using linear, nonlinear, and NN-based regression methods; full and partial NN weight prediction;
450 and different realizations of the online training step ~~;~~ ~~Despite the fact that all given examples are from Earth sciences, the~~
~~authors strongly believe that the demonstrated approach is useful in other fields, as astrophysics, biology, social sciences, or~~
~~technical applications, too. For example, in medicine it could be used to transfer learned skills and knowledge to target groups~~
~~underrepresented or unrepresented in the training data~~and its regularization.

In summary, our experiments demonstrate that the proposed weight-prediction framework enhances the extrapolation capa-
455 bility of NN under strong temporal, spatial, and ~~domain~~regime shifts. The method proves most effective where traditionally
trained and therefore static models exhibit limited skill, underscoring its potential to extend NN capabilities to previously
unseen regimes and conditions. As a result, ~~the NN~~NN can become adaptive and ~~dynamical~~more dynamic operators.

Despite all these promising results, there are still some caveats, as choices have to be made that can heavily impact the
method's performance. The choice of suitable predictors, e.g., if one does not choose the input variables themselves, ~~have~~has
460 to be based on prior knowledge. This should pose no major problem for the experienced researcher. The same can be stated
about the choice of whether ~~only specific parts of a NN or~~ all weights and biases ~~should be subject to the~~of a NN or only
specific subsets should be subjected to weight-regression and subsequent weight-prediction.

In contrast to these rather straightforward choices, the choice of the weight-regression and prediction method cannot be
based on prior domain knowledge alone. The reason is ~~;~~ that the weight sensitivities the proposed approach has to model
465 are entwined with the network architecture and the employed activation functions ~~as well~~. As a first guide, linear or slightly
nonlinear regression methods should work fine for most cases.

The biggest drawback identified so far is that the weight-prediction method has inherited the notorious stochastic nature of
NN training. From several parent NN (with identical architecture and ~~hyper-parameters~~hyperparameters) that are trained on

the same data, not all ~~do produce better performing~~ produce better performing offspring. To this issue, the most follow-up
470 research will be dedicated. Although we already discussed possible reasons and solutions within the manuscript, we will have
to understand the problem better to be able to identify and select parent models suitable for weight-regression or, in the long
run ~~even foster suitable parent model generation through tailored~~, even foster the generation of more suitable parent models
through adapted training procedures. Additionally, we see ~~room to improve~~ much room for improvement in the generation of
the child models by predictor pruning, ~~weight-prediction~~ weight-regression validation, regression-ensembles, transformers, or
475 other means.

As a conclusion, most steps of the approach may be called more or less experimental and ~~surely~~ can and will be improved in
follow-up research. However, already now the method has to be considered very promising and is at least in our experiments
robust in the ensemble-average sense.

As a further outlook, we see this method also as a promising tool for NN applications without OOD. During our many tests,
480 we got the impression that the method can improve the performance of undertrained, overfitted, or by design under-complex
models as well.

Despite the fact that all given examples are from Earth sciences, the authors strongly believe that the demonstrated approach
is useful in other fields, such as astrophysics, biology, social sciences, or technical applications, too. For example, in medicine,
the approach could be used to transfer learned skills and knowledge to patient groups underrepresented or unrepresented in the
485 training data.

Acknowledgments

JSW receives funding from the German Federal Ministry of Education and Research (BMFTR) through the project PalMod.
SRS receives funding from the Helmholtz International Berlin Research School in Data Science (HEIBRiDS) and the Helmholtz
Foundation.

490 ~~Authors contribution statement~~ Author Contribution Statement

JSW contributed with methodology development, experiment design, implementation, and interpretation. JSW wrote the first
draft of the manuscript, ~~did finance~~ financed the project, and did project supervision. SRS contributed with experiment design,
method development, implementation, and refining the manuscript.

6 Data Availability

495 The study was conducted entirely with open data. The AMOC data was taken from the supporting material of van Westen et al. (2024)
available at https://github.com/RenevanWesten/SA-AMOC-Collapse/tree/SA-AMOC-Collapse_v1.0. The World Ocean Atlas
data (Locarnini et al., 2024; Reagan et al., 2024) is available under <https://www.ncei.noaa.gov/access/world-ocean-atlas-2023/>.
The ECMWF-ERA5 climate data (Hersbach et al., 2023) is available from the Copernicus Climate Change Service (C3S)

Climate Data Store (CDS) at <https://cds.climate.copernicus.eu/>. The NCEP-CFSv2 data (Saha et al., 2014) is available from the NOAA Climate Prediction Center at <https://cfs.ncep.noaa.gov/>. The study was conducted entirely with open software, namely Keras, Tensorflow, R, and Python. Exemplary source code for the weight prediction approach is available here: <https://doi.org/xx.xxxx/xxxx.xxxxxxx>

Competing interests

The authors declare ~~no-competing-interests~~ there are no conflicts of interest for this manuscript.

Appendix A

To provide a baseline, we present and compare a range of standard linear and nonlinear methods used to extrapolate the targets of experiments 1 and 2, i.e., AMOC strength and sea water density, without any use of NN. The methods (Gaussian process, linear model, random forest, and support vector machine) were fitted to the same predictors as the weight prediction method, i.e., 4 leading EOF (experiment 1) and S , T , P , ϕ , λ , and z (experiment 2). Likewise, the respective OOD projections are based on the same predictors that our weight prediction method uses. Note that none of these methods outperform the child models in the OOD application cases. Most of the methods do not outperform the parent NN ($ParentModel_0$) either. However, even if a method exists that outperforms both the parent and the child NN in our specific use cases, our approach remains useful in all cases where NN are used for one of the various reasons, e.g., ease of use and implementation, existing NN workflows, compute limitations, invertibility requirements, or simple lack of alternatives (e.g., decision making, generative AI, image classification).

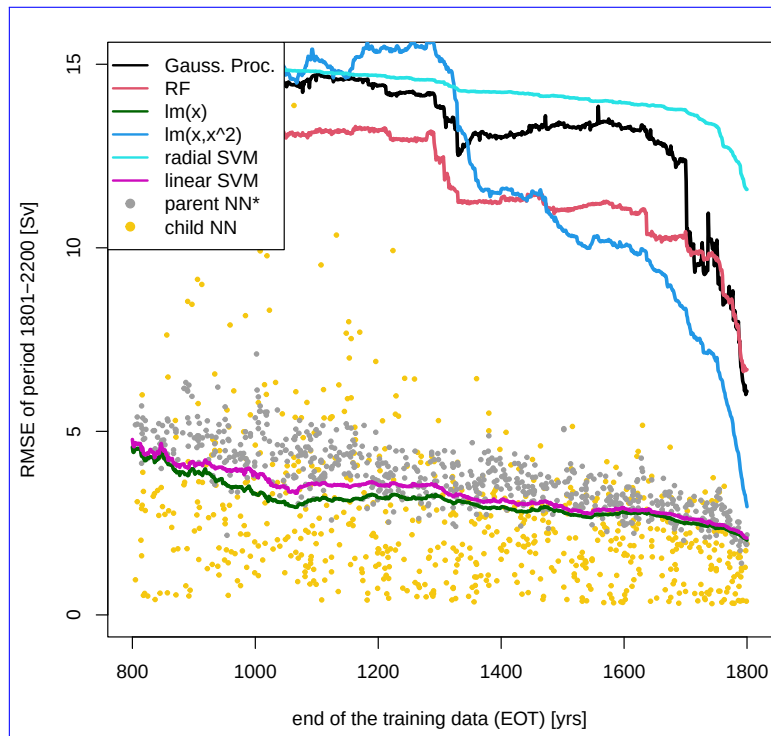


Figure A1. Performance comparison of parent and child NN to non-NN extrapolation methods in experiment 1. Impact of EOT on the RMSE of a fixed application range (1800yrs,2000yrs). Lines show non-NN approaches: Gaussian processes (linear predictors), random forest (linear predictors), linear regression (linear predictors), linear regression (linear and squared predictors), linear and radial kernel SVM (linear predictors) are fitted to the leading 4 principal components of the meridional velocity maps. Dots show the parent NN and linear fit based child NN (cf. Fig. 4). *Note that the parent NN base their AMOC strength estimates on the full velocity fields only, and not on the principal components, as the other methods do, i.e., the NN have to learn a suitable representation themselves while the other methods benefit from prior knowledge.

References

- Bereiter, B., Eggleston, S., Schmitt, J., Nehrbass-Ahles, C., Stocker, T. F., Fischer, H., Kipfstuhl, S., and Chappellaz, J. (2015). Revision of the EPICA Dome C CO₂ record from 800 to 600 kyr before present. *Geophysical Research Letters*, 42(2):542–549.
- Bergmeir, C., Hyndman, R., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, 120:70–83.
- Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O’Gorman, P. A., Neelin, J. D., Lutsko, N. J., and Pritchard, M. (2024). Climate-invariant machine learning. *Science Advances*, 10(6):eadj7250.
- Chauhan, V. K., Zhou, J., Lu, P., Molaei, S., and Clifton, D. A. (2024). A brief review of hypernetworks in deep learning. *Artificial Intelligence Review*, 57(9):1–29.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

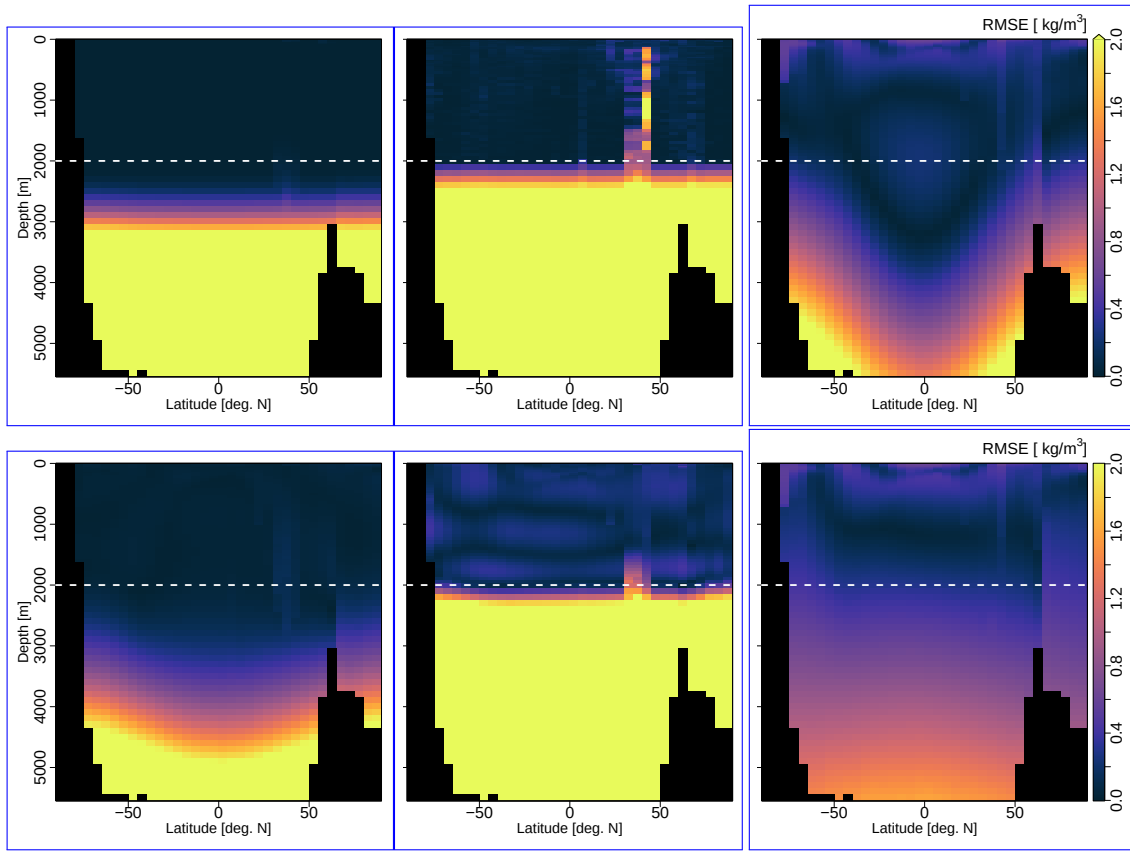


Figure A2. Non-NN model performance of experiment 2 (cf. Fig. 5, upper panel). Depth-dependent zonally-averaged* RMSE of non-NN sea water density estimation after training on ground truth data above the dashed line only. Upper row: Gaussian processes (linear predictors), random forest (linear predictors), linear regression (linear predictors). Lower row: Linear regression (squared predictors), radial kernel SVM (linear predictors), linear kernel SVM (linear predictors). Predictors in all cases are S , T , P , ϕ , λ , and z (cf. Sec. 3.2). *: In giving every global output location equal weight, these averages represent a data-centered view and do not follow oceanographic standards.

El Ghawi, R., Winkler, A., Reimers, C., Schall, A., Gensheimer, J., and Kraft, B. (2025). Imitation or Identification: Limitations of Deep Learning in Extrapolating to Future Climate-Carbon Cycle Change. *Machine Learning: Earth*, 1.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

530 Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucier, T., Bocquet, M., Bretherton, C. S., Christensen, H. M., Dagon, K., Gagne, D. J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M. C., Meehl, G. A., Molina, M. J., Monteleoni, C., Mueller, J., Pritchard, M. S., Rolnick, D., Runge, J., Stier, P., Watt-Meyer, O., Weigel, K., Yu, R., and Zanna, L. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9):916–928.

Fang, S., Wang, Z., Kurths, J., and Fan, J. (2025). Tipping Points and Cascading Transitions: Methods, Principles, and Evidences.

535 arXiv:2511.01168 [physics.ao-ph].

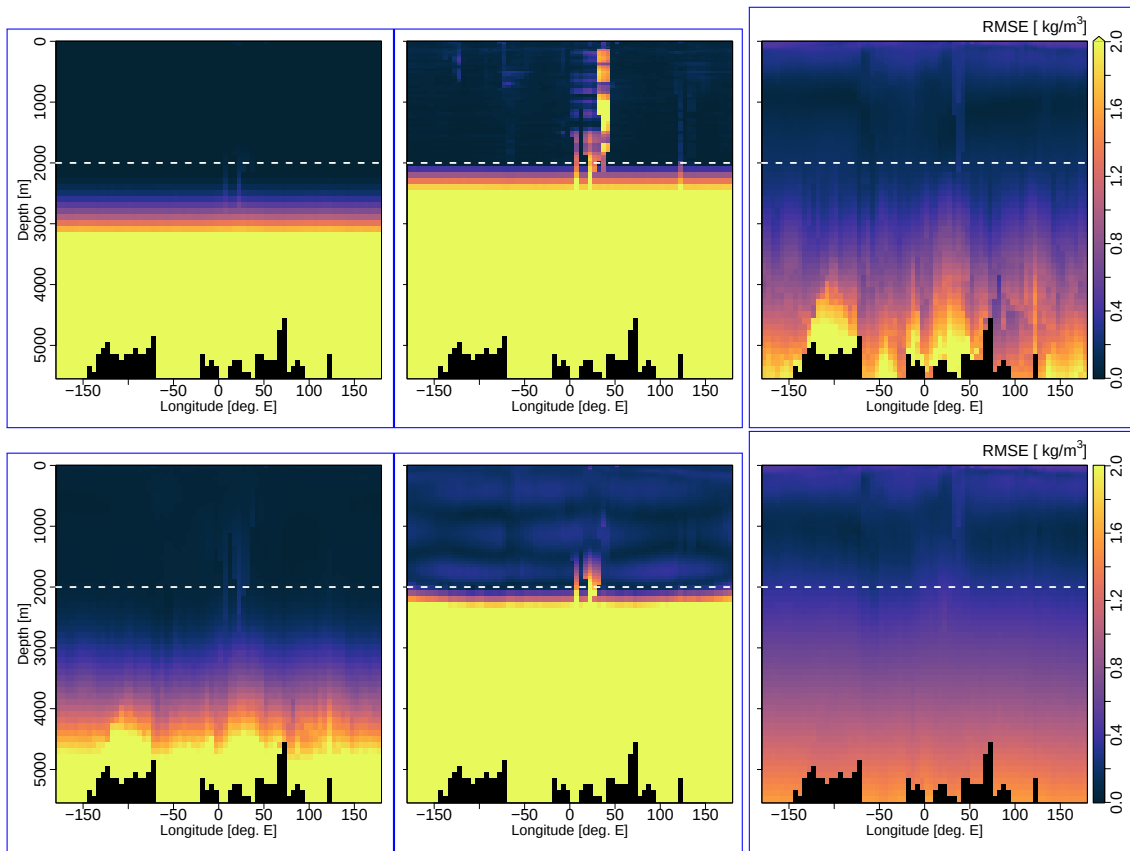


Figure A3. Non-NN model performance of experiment 2 (cf. Fig. 5, lower panel). Depth-dependent meridionally-averaged* RMSE of non-NN sea water density estimation after training on ground truth data above the dashed line only. Upper row: Gaussian processes (linear predictors), random forest (linear predictors), linear regression (linear predictors). Lower row: Linear regression (squared predictors), radial kernel SVM (linear predictors), linear kernel SVM (linear predictors). Predictors in all cases are S , T , P , ϕ , λ , and z (cf Sec. 3.2). *: In giving every global output location equal weight, these averages represent a data-centered view and do not follow oceanographic standards.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297–310.

Hernanz, A., García-Valero, J. A., Domínguez, M., and Rodríguez-Camino, E. (2022). A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment. *Atmospheric Science Letters*, 23(6):e1087.

- 540 Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2023). ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI:10.24381/cds.adbb2d47 (Accessed on 08-06-2025).
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

- 545 [Hsieh, W. W. \(2023\). Improving Predictions by Nonlinear Regression Models from Outlying Input Data. *Journal of Environmental Informatics*, 41\(2\).](#)
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.
- IOC, SCOR, and IAPSO (2010). The international thermodynamic equation of seawater – 2010: Calculation and use of thermodynamic
- 550 properties. Technical Report 56, UNESCO. Intergovernmental Oceanographic Commission, Manuals and Guides No. 56, UNESCO, 196 pp.
- Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, 3:667–674.
- Irrgang, C., Saynisch-Wagner, J., and Thomas, M. (2020). Machine Learning-Based Prediction of Spatiotemporal Uncertainties in Global
- 555 Wind Velocity Reanalyses. *Journal of Advances in Modeling Earth Systems*, 12(5):e2019MS001876.
- Jung, H., Saynisch-Wagner, J., and Schulz, S. (2024). Can eXplainable AI Offer a New Perspective for Groundwater Recharge Estimation?—Global-Scale Modeling Using Neural Network. *Water Resources Research*, 60(4):e2023WR036360.
- Kelley, D. and Richards, C. (2025). *oce: Analysis of Oceanographic Data*. R package version 1.8-4.
- Kotz, M., Wenz, L., Stechemesser, A., Kalkuhl, M., and Levermann, A. (2021). Day-to-day temperature variability reduces economic growth.
- 560 *Nature Climate Change*, 11:319–325.
- Koutsodendris, A., Pross, J., and Zahn, R. (2014). Exceptional Agulhas leakage prolonged interglacial warmth during MIS 11c in Europe. *Paleoceanography*, 29:1062–1071.
- Kvorka, J., Čadek, O., Šachl, L., and Velínský, J. (2026). Convective flow in Ganymede’s subsurface ocean: Implications for the induced magnetic field and topography. *Icarus*, 444:116807.
- 565 Landsberg, J. B. and Barnes, E. A. (2025). Forecasting the Future with Yesterday’s Climate: Temperature Bias in AI Weather and Climate Models. arXiv:2509.22359 [physics.ao-ph].
- Lee, T. and Gentemann, C. (2018). Satellite SST and SSS observations and their roles to constrain ocean models. In Chassignet, E., Pascual, A., Tintoré, J., and Verron, J., editors, *New Frontiers in Operational Oceanography*, chapter 11, pages 271–288. GODAE OceanView. doi:10.17125/gov2018.ch11.
- 570 [Li, X.-M., Chang, L., Zhou, J., Sun, J., Lu, X., Han, X., Wang, J., Lu, H., Li, X., Wang, N., Qiu, Y., and Wu, Y. \(2025\). Haishao-1 satellite: Low-inclination orbit spaceborne synthetic aperture radar. *The Innovation*, 6\(9\). doi: 10.1016/j.xinn.2025.100949.](#)
- Llovel, W., Willis, J. K., Landerer, F. W., and Fukumori, I. (2014). Deep-ocean contribution to sea level and energy budget not detectable over the past decade. *Nature Climate Change*, 4(11):1031–1035.
- Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Reagan, J. R., Boyer, T. P., Seidov, D., Wang, Z., Garcia, H. E., Bouchard, C., Cross,
- 575 S. L., Paver, C. R., and Dukhovskoy, D. (2024). World Ocean Atlas 2023, Volume 1: Temperature. Technical Report 89, A. Mishonov Technical Editor, NOAA Atlas NESDIS.
- Mehta, A., Parsons, D., Lukas Holch, T., Berge, D., and Weidlich, M. (2025). Convolution and Graph-based Deep Learning Approaches for Gamma/Hadron Separation in Imaging Atmospheric Cherenkov Telescopes. *Proceedings of Science*, ICRC2025(752):1–8.
- [Nelder, J. A. and Wedderburn, R. W. M. \(1972\). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A \(General\)*, 135\(3\):370–384.](#)
- 580 Ortega-Cisneros, K., Fierros-Arcos, D., Lindmark, M., Novaglio, C., Woodworth-Jefcoats, P., Eddy, T. D., Coll, M., Fulton, E., Oliveros-Ramos, R., Reum, J., Shin, Y.-J., Bulman, C., Capitani, L., Datta, S., Murphy, K., Rogers, A., Shannon, L., Whitehouse, G. A., Adekoya,

- E., Dias, B. S., Fuster-Alonso, A., Hansen, C., Husson, B., McGregor, V., Morell, A., Morzaria Luna, H.-N., Ouled-Cheikh, J., Ruzicka, J., Steenbeek, J., Stollberg, I., Subramaniam, R. C., Tulloch, V., Bryndum-Buchholz, A., Harrison, C. S., Heneghan, R., Maury, O.,
585 Pozo Buil, M., Schewe, J., Tittensor, D. P., Townsend, H., and Blanchard, J. L. (2025). An Integrated Global-To-Regional Scale Workflow for Simulating Climate Change Impacts on Marine Ecosystems. *Earth's Future*, 13(2):e2024EF004826.
- [Peel, M. C., Finlayson, B. L., and McMahon, T. A. \(2007\). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, 11\(5\):1633–1644.](#)
- [Qin, R., Zhang, G., and Tang, Y. \(2025\). On the transferability of semantic segmentation for very-high-resolution remote sensing data of multi-city environments. *Photogrammetric Engineering & Remote Sensing*, 91\(8\):517–528.](#)
590
- Rahmstorf, S. (1996). On the freshwater forcing and transport of the Atlantic thermohaline circulation. *Climate Dynamics*, 12(12):799–811.
- [Rasp, S., Pritchard, M. S., and Gentile, P. \(2018\). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115\(39\):9684–9689.](#)
- Reagan, J. R., Seidov, D., Wang, Z., Dukhovskoy, D., Boyer, T. P., Locarnini, R. A., Baranova, O. K., Mishonov, A. V., Garcia, H. E.,
595 Bouchard, C., Cross, S. L., and Paver, C. R. (2024). World Ocean Atlas 2023, Volume 2: Salinity. Technical Report 90, A. Mishonov Technical Editor, NOAA Atlas NESDIS.
- Roquet, F., Madec, G., McDougall, T. J., and Barker, P. M. (2015). Accurate polynomial expressions for the density and specific volume of seawater using the TEOS-10 standard. *Ocean Modelling*, 90:29–43.
- Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang,
600 R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, 27(6):2185–2208.
- Saynisch, J., Peterett, J., Irrgang, C., Kuvshinov, A., and Thomas, M. (2016). Impact of climate variability on the tidal oceanic magnetic signal — a model-based sensitivity study. *Journal of Geophysical Research: Oceans*, 121(8):5931–5941.
- [Schannwell, C., Mikolajewicz, U., Kapsch, M.-L., and Ziemer, F. \(2024\). A mechanism for reconciling the synchronisation of Heinrich events and Dansgaard-Oeschger cycles. *Nature Communications*, 15\(1\):2961.](#)
605
- Schnaubelt, M. (2019). A comparison of machine learning model validation schemes for non-stationary time series data. Technical Report 11/2019, FAU Discussion Papers in Economics.
- Sebastianelli, A., Spiller, D., Carmo, R., Wheeler, J., Nowakowski, A., Jacobson, L. V., Kim, D., Barlevi, H., Cordero, Z. E. R., Colón-González, F. J., Lowe, R., Ullo, S. L., and Schneider, R. (2024). A reproducible ensemble machine learning approach to forecast dengue
610 outbreaks. *Scientific Reports*, 14(1):3807.
- van der Deure, T., Nogués-Bravo, D., Njotto, L. L., and Stensgaard, A.-S. (2025). ~~Climate change favors african malaria vector mosquitoes~~[Climate Change Favors African Malaria Vector Mosquitoes](#). *Global Change Biology*, 31(11):e70610. e70610 GCB-25-1290.R2.
- van Westen, R. M., Kliphuis, M., and Dijkstra, H. A. (2024). Physics-based early warning signal shows that AMOC is on tipping course.
615 *Science Advances*, 10(6):eadk1189.
- Wang, J. Y., Nikolaou, N., an der Heiden, M., and Irrgang, C. (2024). High-resolution modeling and projection of heat-related mortality in Germany under climate change. *European Journal of Public Health*, 34:ckae144.025.
- [Wong, A. P. S., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., Gilson, J., Johnson, G. C., Martini, K., Murphy, D. J., Scanderbeg, M., Bhaskar, T. V. S. U., Buck, J. J. H., Merceur, F., Carval, T., Maze, G., Cabanes, C., André, X., Poffa, N., Yashayaev, I., Barker, P. M., Guinehut, S., Belbéoch, M., Ignaszewski, M., Baringer, M. O., Schmid, C., Lyman, J. M., McTaggart, K. E., Purkey, S. G.,](#)
620

Zilberman, N., Alkire, M. B., Swift, D., Owens, W. B., Jayne, S. R., Hersh, C., Robbins, P., West-Mack, D., Bahr, F., Yoshida, S., Sutton, P. J. H., Cancouët, R., Coatanoan, C., Dobbler, D., Juan, A. G., Gourrion, J., Kolodziejczyk, N., Bernard, V., Bourlès, B., Claustre, H., D'Ortenzio, F., Le Reste, S., Le Traon, P.-Y., Rannou, J.-P., Saout-Grit, C., Speich, S., Thierry, V., Verbrugge, N., Angel-Benavides, I. M., Klein, B., Notarstefano, G., Poulain, P.-M., Vélez-Belchí, P., Suga, T., Ando, K., Iwasaka, N., Kobayashi, T., Masuda, S., Oka, E., Sato, K., Nakamura, T., Sato, K., Takatsuki, Y., Yoshida, T., Cowley, R., Lovell, J. L., Oke, P. R., van Wijk, E. M., Carse, F., Donnelly, M., Gould, W. J., Gowers, K., King, B. A., Loch, S. G., Mowat, M., Turton, J., Rama Rao, E. P., Ravichandran, M., Freeland, H. J., Gaboury, I., Gilbert, D., Greenan, B. J. W., Ouellet, M., Ross, T., Tran, A., Dong, M., Liu, Z., Xu, J., Kang, K., Jo, H., Kim, S.-D., and Park, H.-M. (2020). Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. *Frontiers in Marine Science*, Volume 7 - 2020.