JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003370

Key Points:

- The Meridional Overturning Circulation (MOC) is a key ocean circulation pattern, which is under-observed
- We use an ocean state estimate to test the feasibility of inferring the MOC based solely on satellite observations
- A unified machine learning methodology shows high skill in MOC reconstruction as a function of latitude and basin

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Solodoch,
Aviv.Solodoch@mail.huji.ac.il

Citation:

Solodoch, A., Stewart, A. L., McC. Hogg, A., & Manucharyan, G. E. (2023). Machine learning-derived inference of the Meridional Overturning Circulation from satellite-observable variables in an ocean state estimate. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003370. https://doi.org/10.1029/2022MS003370

Received 23 AUG 2022
Accepted 3 APR 2023

Machine Learning-Derived Inference of the Meridional Overturning Circulation From Satellite-Observable Variables in an Ocean State Estimate

Aviv Solodoch[1], Andrew L. Stewart[1], Andrew McC. Hogg[2,3], and Georgy E. Manucharyan[4]

[1]Department of Atmospheric and Oceanic Sciences, University of California in Los Angeles, Los Angeles, CA, USA, [2]Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia, [3]ARC Centre of Excellence for Climate Extremes, Sydney, NSW, Australia, [4]School of Oceanography, University of Washington, Seattle, WA, USA

Abstract The oceanic Meridional Overturning Circulation (MOC) plays a key role in the climate system, and monitoring its evolution is a scientific priority. Monitoring arrays have been established at several latitudes in the Atlantic Ocean, but other latitudes and oceans remain unmonitored for logistical reasons. This study explores the possibility of inferring the MOC from globally-available satellite measurements via machine learning (ML) techniques, using the ECCOV4 state estimate as a test bed. The methodological advantages of the present approach include the use purely of available satellite data, its applicability to multiple basins within a single ML framework, and the ML model simplicity (a feed-forward fully connected neural network (NN) with small number of neurons). The ML model exhibits high skill in MOC reconstruction in the Atlantic, Indo-Pacific, and Southern Oceans. The approach achieves a higher skill in predicting the model Southern Ocean abyssal MOC than has previously been achieved via a dynamically-based approach. The skill of the model is quantified as a function of latitude in each ocean basin, and of the time scale of MOC variability. We find that ocean bottom pressure generally has the highest reconstruction skill potential, followed by zonal wind stress. We additionally test which combinations of variables are optimal. Furthermore, ML interpretability techniques are used to show that high reconstruction skill in the Southern Ocean is mainly due to (NN processing of) bottom pressure variability at a few prominent bathymetric ridges. Finally, the potential for reconstructing MOC strength estimates from real satellite measurements is discussed.

Plain Language Summary The Meridional Overturning Circulation (MOC) plays a key role in the exchange of heat and chemical constituents between oceans globally, and between the atmosphere and the deep ocean. However, it is currently directly monitored only at a handful of different latitudes in the Atlantic Ocean. Using an ocean simulation constrained by available measurements, we examine the feasibility of monitoring the MOC indirectly using measurements made by satellites. We train a so-called "neural network" to learn relations between the MOC and satellite-observable ocean properties, such as pressure at the sea floor. These relations are used to produce reconstructions of the simulated MOC derived from the satellite-observable ocean properties alone, and to test which satellite-observable ocean properties and regions may be most impactful in MOC reconstruction. We find that this approach yields high skill in reconstructing the MOC over much of the ocean, suggesting that this approach could be transferred to indirectly monitor the MOC in nature.

1. Introduction

The Meridional Overturning Circulation (MOC) is a critical component of the climate system, playing key roles in heat and material transport, deep ocean ventilation, and global water mass distribution and stratification (Talley, 2013). MOC deep waters form in several regions of the subpolar North Atlantic and of the Antarctic margins, from which they are exported globally in complicated patterns, carrying to their destinations the signatures of their air-sea interactions prior to ventilation. Monitoring and deducing the global MOC transport is thus crucial for understanding of ocean circulation and climate variability. Traditionally, the MOC was estimated using geostrophy and inverse methods from ship-based hydrographic measurements (Wunsch, 1996). However, the sparseness of ship hydrography in time and space limits the recoverable information, and can lead to severe aliasing, for example, of temporal variability (Frajka-Williams et al., 2019). In the last two decades, dedicated in situ monitoring arrays have been put in place at several latitudes, although only in the Atlantic Ocean. Several arrays span the width of the basin: the RAPID-MOCHA array at 26.5°N (Cunningham et al., 2007; Kanzow

et al., 2007; McCarthy et al., 2015), the OSNAP array near 56°N (Lozier et al., 2017), and the SAMBA array at 34.5°S (Ansorge et al., 2014). Two additional Atlantic arrays were recently extended to full basin width: the NOAC array at 47°N and the TSAA array at 11°S (McCarthy et al., 2020). While the Atlantic is thus presently observed at multiple latitudes, the entire Indo-Pacific and Southern Oceans are not covered by MOC monitoring arrays.

Closing this gap in monitoring of the MOC would require scaling up the few continuously maintained cross-basin in-situ MOC-sampling arrays to a significantly larger number of latitudes (and other basins), which presents a serious logistical challenge. An alternative is to explore indirect approaches to monitoring the MOC using physical ocean properties that have been continuously monitored by satellite observations for several decades. These properties include sea surface height (SSH), Ocean bottom pressure (OBP), surface wind stress, sea-surface temperature (SST), and sea-surface salinity (SSS). Indeed, the present in-situ monitoring arrays depend on satellite-measured surface wind stress to derive the Ekman transport component of the MOC, as well as satellite-measured SSH to derive geostrophic surface velocity in some regions (McCarthy et al., 2020).

There are previously-established theoretical reasons to assume that satellite observable variables can be used to infer at least some of the MOC's variability. For example, upper ocean meridional Ekman transport may be estimated from the zonal wind stress (ZWS), while bottom pressure differences predict the abyssal geostrophic transport component, although complications arise due to sloping bottom relief (Bingham & Hughes, 2008; McCarthy et al., 2015) as well as satellite resolution and noise. Stewart and Hogg (2017) have suggested theoretically and demonstrated within an idealized numerical model that in the Southern Ocean bottom form stress (i.e., bottom pressure multiplied by bathymetric slope), and its associated AABW fluctuations, are dynamically related to SSH fluctuations due to the deep penetration of the Antarctic Circumpolar Current near significant topographic features. Sea surface temperature has also been shown to be correlated with Atlantic MOC (AMOC) strength over multi-decadal and centennial time scales via an advective mechanism (Knight et al., 2005; Rahmstorf et al., 2015). Seasonal-lagged correlations between SST distribution and AMOC strength were reported by Duchez et al. (2016), and again suggested to occur due to advection, although Alexander-Turner et al. (2018) have shown the correlation and its associated spatial patterns can be far from stationary on time scales shorter than 30 years.

Previous studies have used satellite-observed SSH and wind stress to reconstruct AMOC in conjunction with partial in-situ observations, for example, using Florida Strait flow measurement only (i.e., in contrast with use of basin-wide monitoring arrays). Willis (2010) combined satellite SSH measurements and in-situ Argo floats data to estimate AMOC and its variability by integration of geostrophic velocity. The method worked well when tested within a numerical model (root mean square error (RMSE) ∼1 Sv) near 41°N, but not elsewhere (RMSE ∼5 Sv). Frajka-Williams (2015) has shown that AMOC at 26° can be reconstructed with high skill (with >90% interannual variance explained) from satellite SSH and wind stress observations in combination with Florida Strait in-situ transport measurements. The SSH to basin-interior AMOC relation was derived via linear regression. Sanchez-Franks et al. (2021) have derived the basin-interior AMOC component via geostrophy from satellite SSH, wind stress data, and from the observed climatological stratification structure. Their AMOC reconstructions explained ≈70% of RAPID-MOCHA array AMOC variability at inter-annual timescales. Additional relevant work is reviewed in Jackson et al. (2022).

Ocean bottom pressure (OBP), observed via the GRACE mission (Tapley et al., 2004), has also been used to reconstruct the MOC in several studies. Based on OBP at the western boundary alone (within a numerical model), Bingham and Hughes (2008) have reconstructed 90% of interannual AMOC variability at 42°N, via gesotrophy. Landerer et al. (2015) used solely satellite-measured OBP to calculate AMOC variability at 26.5° via gesotrophy. On near-annual time scales, the reconstructed AMOC had a correlation of 0.7 with the RAPID-MOCHA array. Mazloff and Boening (2016) demonstrated that a linear regression of OBP spatial differences can explain over 85% of variability of Antarctic Bottom Water northward transport across a relatively flat section in the South Pacific (183.7–209.7°E, 36.4°S) within the Southern Ocean State Estimate (Mazloff et al., 2010). Stewart et al. (2021) have shown that ZWS determines a large variability fraction (e.g., ≈65% at 6-month time scales) of Antarctic Bottom Water export in the Southern Ocean within the ECCO state estimate (Forget et al., 2015), and that the relevant signal is also present in bottom form stress.

Previous work thus examined MOC reconstruction from satellite observations within specific depth cells and regions, and via geostrophy and/or linear regression. In this work we use a ML methodology to perform a

near-global examination of the potential for multiple satellite-observable variables to be used for MOC reconstruction. To this end we use the ECCO state estimate as a test bench, treating its overturning circulation, as well as the distribution of its satellite-observable variables as "truth." We use a neural network (NN) approach to gauge the maximal MOC-relevant information that may be inherent in these satellite observables, rather than as a suggested practical reconstruction method at this point. The methodology, reconstructing abyssal or mid-depth MOC cells strength latitude by latitude and separately at each of three main ocean basins (the Southern Ocean, Indo-Pacific Oceans, and Atlantic Oceans), is introduced in Section 2. In Section 3 we evaluate the skill of the methodology and the conditions under which it performs well within four test cases: AMOC at 26°N, the Indo-Pacific MOC at 30°S, and the Southern Ocean mid-depth and abyssal MOC cells at 55°S and 60°S, respectively. We further examine the potential utility of different satellite-observable variables, as well as the length of training period required with the present approach. We then apply and test the skill of this approach for MOC reconstruction near-globally within ECCO in Section 4. A discussion of the dynamical reasons underlying regional reconstruction skill differences, and of prospects for the application of the technique is presented in Section 5, followed by a summary in Section 6.

2. Methods

2.1. Numerical Model

Our analyses are conducted on the output data of the ECCOV4r3 (ECCO henceforth here) state estimate (Forget et al., 2015), based on the MITgcm ocean model (Adcroft et al., 2004; Marshall et al., 1997) at 1° resolution with 50 vertical levels, and including a sea-ice module (Losch et al., 2010). The state estimate framework is optimal in the sense that it minimizes the deviation of the model fields from observations (in an integral sense) while maintaining self-consistent model dynamics. That is, no artificial nudging terms are employed in the model forward run. We use output fields given as monthly averages, which are available over the period 1992–2015.

2.2. Meridional Overturning Circulation

In this study we address the mid-depth residual MOC cell in the Atlantic Ocean (AMOC), the residual abyssal MOC cell in the Indo-Pacific Ocean (henceforth IPMOC), and the residual mid-depth and abyssal MOC cells in the Southern Ocean (henceforth SOMOC). At any particular latitude and basin (e.g., the entire Southern Ocean zonal extent at 60°S, or the Atlantic Basin at 26°N), the residual monthly-mean MOC streamfunction Ψ as a function of time t is computed as:

$$\Psi(\text{basin}, y, \sigma_2, t) = -\int_{x_1(\text{basin},y)}^{x_2(\text{basin},y)} \int_{\eta_b(x,y)}^{\eta_{\sigma_2}(x,y,t)} v_r(x, y, \sigma_2, t)\, \mathrm{d}x\mathrm{d}z. \tag{1}$$

Here x is a west-east coordinate, y is latitude, "basin" denotes the ocean basin of interest, for example, Southern Ocean, σ_2 is potential density anomaly referenced to 2 km depth, x_1 and x_2 are the zonal boundaries of the basin at the particular latitude (defined in Appendix A), η_b is the bottom depth, η_{σ_2} is isopycnal depth, $v_r = v_m + v_e$ is residual velocity, v_m is monthly-mean velocity, and v_e is the parameterized eddy-induced monthly-mean "bolus" velocity (Gent & Mcwilliams, 1990). Time-mean ECCO residual overturning circulation streamfunctions in the Atlantic, Indo-Pacific, and Southern Oceans are shown in Figure 1. The MOC in the figure is remapped to depth coordinates by the method described in Stewart et al. (2021).

For the present goal of reconstructing the strength of the MOC, we define the MOC strength (ψ, hereafter also MOC time series) as the monthly-mean streamfunction (Ψ) evaluated at the density σ_2^0 at which the time-mean MOC magnitude is maximal for the same latitude, MOC "cell," density range, and ocean basin. To be precise, σ_2^0 is defined by the following equation:

$$\left|\overline{\Psi\left(\text{basin}, y, \sigma_2^0(\text{basin}, y), t\right)}\right| = \max_{\sigma_2}\left\{\left|\overline{\Psi(\text{basin}, y, \sigma_2, t)}\right|\right\}. \tag{2}$$

Here the overline indicates a time mean over 1992–2015. The maximum over density levels is taken over a subset of the full density range, corresponding to the time-mean density range of the overturning cell of interest (mid-depth or abyssal cell) at the same latitude. The MOC time series are thus defined by

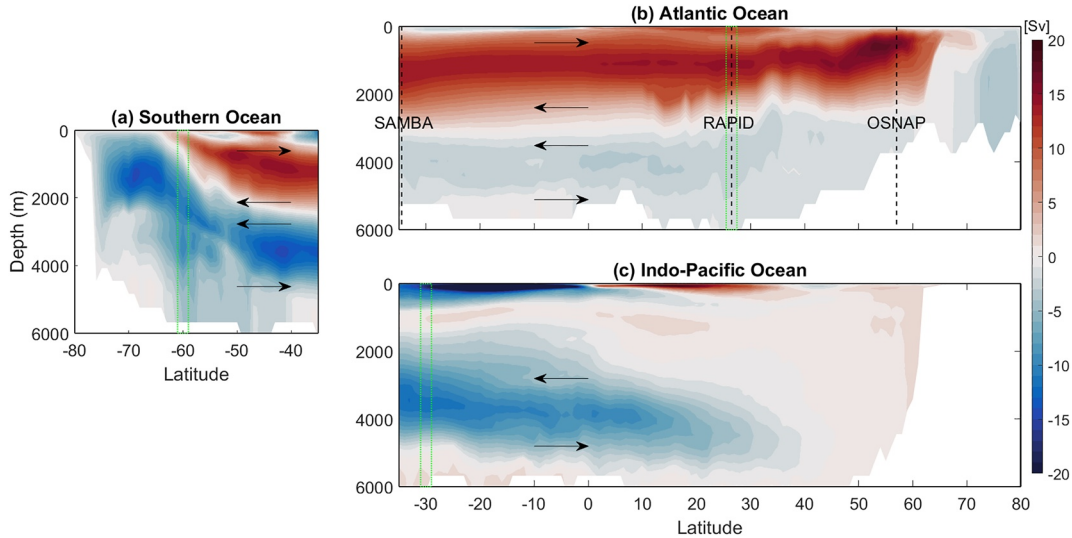

Figure 1. Residual meridional overturning circulation (MOC) in ECCOV4r3 (a) Southern Ocean, (b) Atlantic Ocean, and (c) Indo-Pacific Oceans. Arrows indicate the residual MOC direction. Dotted green rectangles show zonal strips that we study in more detail than other latitudes in this paper. Dashed black lines in panel (b) mark the locations of the OSNAP (≈57°N), RAPID-MOCHA (26.5°N), and SAMBA (34.5°S) arrays.

$$\psi(\text{basin}, y, t) = s\Psi\left(\text{basin}, y, \sigma_2^0(\text{basin}, y), t\right), \tag{3}$$

where s is the sign of $\overline{\Psi\left(\text{basin}, y, \sigma_2^0(\text{basin}, y), t\right)}$, that is, chosen to make times with higher than average MOC transport magnitude correspond to positive ψ anomalies.

We also experimented with another often-used MOC time series definition (Frajka-Williams et al., 2019; Li et al., 2021), identical to the above except that the density level $\sigma_2^0(\text{basin}, y, t)$ varies from month to month, and coincides with the monthly-mean maximum of the streamfunction,

$$\left|\Psi\left(\text{basin}, y, \sigma_2^0(\text{basin}, y, t), t\right)\right| = \max_{\sigma_2}\{|\Psi(\text{basin}, y, \sigma_2, t)|\}. \tag{4}$$

In a small subset of cases that we examined, the methods described below led to quantitatively similar reconstruction skills for this alternative MOC time series definition.

2.3. Machine Learning Methodology

The ML methodology here produces an estimate of the MOC strength (defined in 2.2), based on one or more satellite-observable model variables within a zonal strip centered on the MOC latitude of interest ("covariates"). The meridional extent of the zonal strip in the reported results is 3° for the AMOC and SOMOC mid-depth MOC cell, and 5° for the abyssal IPMOC and SOMOC cells. However, the results are insensitive to varying the widths from 3 to 7° (Section 5.3). The longitude range per latitude ("masks") included for covariates is limited by the MOC basin of interest. The basin masks are defined in Appendix A.

The model (ECCO) variables used as covariates in the NNs are subsets of the following list: sea-surface height (SSH), SST, SSS, ZWS, and OBP. These variables have been chosen because they are also measured by satellites, and hence are informative for assessment of real-world applicability of our approach. The variable SSH (designated with the same acronym in ECCO output fields) is a "dynamic" SSH variable, which adds sea-ice and snow loading to the SSH under sea ice. Equivalently, by hydrostatic balance it corresponds to SSH in leads between sea ice and hence is best suited for comparison with altimetry (Wang et al., 2017), or for emulating altimetry. For ZWS we prescribe the basin-zonal-mean (within the basin of interest), since it predicts the total meridional Ekman flux, and since it has a high correlation with abyssal export in the Southern Ocean (Stewart et al., 2021).

As a pre-processing step, the MOC time series (ψ, Section 2.2) and the covariates are each de-trended and their seasonal cycles are removed, that is, we are trying to reconstruct non-secular anomalies from the climatological

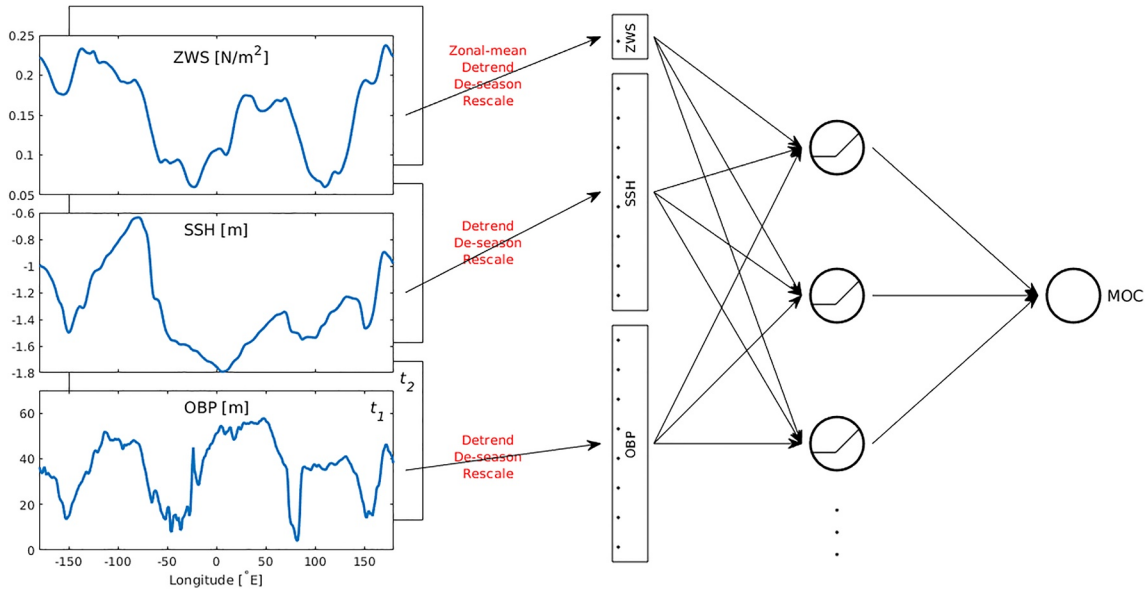

Figure 2. Neural network (NN) schematic architecture. The Meridional Overturning Circulation at a single latitudinal strip is reconstructed using satellite-observable variables at the same latitude. The left side shows an example of input data ("covariates"): zonal wind stress (marked ZWS), sea surface height (SSH), and Ocean bottom pressure (OBP). An example data zonal distribution at a single time-sample (monthly-average) is shown. Additional blank squares in the background symbolize the other time samples. Pre-processing (red text) includes input variable detrending and scaling (standardization). Wind stress, if used, is first replaced by its zonal mean. For other covariates, full zonal information is used in the NN. The input variables are then stacked into one long 1D array (vertical array in the figure). The multiple dots (single dot) within the stacked variables symbolize the zonal dependence (lack of zonal dependence due to zonal averaging). The stacked variables are fed to a feed-forward fully connected NN (three neurons with ReLu activation functions are shown as an example). A single hidden layer is found to be sufficient for this application, as discussed in Section 2.

series. All covariates are then standardized (means are removed, followed by division by the temporal standard deviation at each spatial location), since NN training is generally more effective with standardization. To study MOC reconstruction skill at various time scales of variability, we apply a temporal low pass filter (LPF) to the ECCO MOC time series and to its reconstruction by the NN trained on the unfiltered MOC (hereafter, LPF after training). Alternatively, we include the option of applying a LPF to the covariates and the MOC time series prior to NN pre-processing and training (hereafter, LPF before training). This filtering is performed via matlab's "lowpass" function, which constructs a delay-compensated filter with 60 dB stopband attenuation. The filtered series maintain monthly time intervals between their samples.

We train a statistical model that predicts a single output (the MOC strength) from one or more inputs (the covariates). Figure 2 illustrates the model architecture for our focal case of a NN, described in more detail below. The NNs are implemented using the matlab NNs toolbox. The code required to reproduce our findings is publically available (see data availability statement). We also briefly report on results with linear regression, but focus on results with nonlinear NNs, which have a greater generalization potential than linear regression. Regularized (see below) linear regression is considered as well, as a special case of NNs with linear activation function.

As indicated by Figure 2, our NN architecture comprises a fully-connected feed-forward NN, with only one hidden layer. The feed-forward NN determines its output at each time sample (month) from the input variables at the same time only. The NN's linear weights are constant in time and "learned" based on all time samples in the training subset (see below). The neurons use either a ReLu activation function (tanh activation produced similar results) or a linear activation function. We have examined using several hidden layers, but found no gain in the performance metrics relative to one layer with the present methodology. The chosen training method is Levenberg-Marquardt back-propagation (Hagan & Menhaj, 1994), as implemented in the matlab "trainlm" function. The Levenberg-Marquardt method requires choice of initial value (μ_i) and rates of decrease (μ_{dec}) and increase (μ_{inc}) of the parameter μ, which controls the switch over from gradient descent to Newton method-like behavior (the latter being more accurate near a minimum). We experimented with different values but found no significant improvements beyond the matlab functions's default values: $\mu_i = 0.001$ (or $\mu_i = 0.005$ with Bayesian training, see below), $\mu_{dec} = 0.1$, and $\mu_{inc} = 10$.

Originally we limited over-training via an early stopping criterion on a validation subset, as well as by adding a regularization L_2 penalty, with a relative penalty weight that we set to 75% based on experimentation. We subsequently switched to a more complex NN learning algorithm, known as Bayesian interpolation (Foresee & Hagan, 1997; MacKay, 1992). This approach is more computationally intensive, but has the advantage of automatically finding the Bayesian "optimal" value of the NN regularization parameter given the observations (input data). The Bayesian framework also obviates the need for a validation data set and early stopping criterion, that is, it has the advantage of increasing the training data set size without increasing the overfitting tendency. The input data is divided into "training" and "testing" datasets that comprise the first 70% and last 30% of the model time series, respectively (except in Section 3.3). The NN is trained only via the "training" data set, so that the potential for reconstructing MOC from using a trained NN can be examined from the goodness of fit of MOC reconstruction in the "testing" data set. A nonzero mean-square-error (MSE) training goal is applied and found useful in improving convergence with the linear activation function. To limit over-training and aid convergence, the MSE threshold is defined as 1% of the temporal variance of the MOC time series.

2.4. Skill Metrics

To quantify the skill of the MOC reconstructions, we examine several metrics to compare the NN reconstruction $\psi_r(t)$ (t being model time) with the model (ECCO) time-series $\psi_m(t)$. Values presented for these metrics in the manuscript relate strictly to the "testing" portion of the time series. The metrics are linear correlation (r), the coefficient of determination (R^2), and the root (RMSE). R^2 measures the percentage of variability in ψ_m that is statistically accountable by ψ_r. For completeness, the metrics are defined as follows (overline denotes a time-mean over the testing period, prime denotes temporal deviation from the mean, and N_t is the number of time samples):

$$r^2 = \frac{\sum_t (\psi_m(t)\psi_r(t))}{\sqrt{\sum_t (\psi_m(t))^2}\sqrt{\sum_t (\psi_r(t))^2}}, \tag{5}$$

$$R^2 = 1 - \frac{\sum_t (\psi_m(t) - \psi_r(t))^2}{\sum_t \left(\psi_m(t) - \overline{\psi_m}\right)^2}, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N_t}\sum_t (\psi_m(t) - \psi_r(t))^2}. \tag{7}$$

The skill achieved by the NN may vary depending on the initial network weights, which are chosen randomly prior to training. We therefore report the median skill metrics over 20 NN realizations, in some cases complemented by additional statistical measures such as the range or maximal skill over realizations. The realizations differ only in the seed value of the random number generator used to initialize the NN training.

We also compare the power spectral density (PSD) levels of the MOC time series and its NN reconstruction, as well as their spectral coherence. The PSD and coherence are calculated using the Welch method with 10 chunks (each detrended and Hann-windowed) and 50% overlap (Percival & Walden 1993). The statistical significance of the coherence is estimated by a Monte Carlo approach, following Thompson (1979).

2.5. NN Interpretability

We employ several methods to determine the impact of different input locations (longitudes) on NN reconstruction skill. A brute-force approach to acquire this information is to retrain the NN with various subsets of the input data locations. This approach is computationally intensive due to the need to retrain the NN over large numbers of configurations. We therefore use two previously-developed methods of interpreting NNs (McGovern et al., 2019; Toms et al., 2020), the method of permutations and Layerwise Relevance Propagation, to explore this question without retraining the NN.

In the method of permutations (Breiman, 2001; McGovern et al., 2019), a modified version of the "testing" portion of the input data is passed through the NN. This modified input data is identical to the original input data,

except that for one input channel (channel denoting here a specific ECCO variable, e.g., SSH, at a specific longitude) the time samples are randomly reshuffled (permuted). The prediction of the MOC strength from this modified input data typically exhibits a reduction in skill, which is used to quantify the importance of the permuted input channel (i.e., a specific longitude) to the NN skill. Note that rather than permuting at a single longitude, we permute the time series of all the cells up to 7° longitude away from the longitude of interest.

Layerwise Relevance Propagation (Bach et al., 2015; Montavon et al., 2017; Toms et al., 2020) propagates backward (from output to input) across the NN the relative impact of each neuron on the output. This allows the relative impact ("relevance") of each input element (longitude) in determining the NN output to be quantified. The computation of relevance $\mathcal{R}_i$ of input element (longitude) i for a specific time sample is particularly simple for the one-layer NNs employed here:

$$A_j = \mathrm{ReLu}\left[\sum_i W_{l=1,i,j} X_i + b_{l=1,j}\right] \tag{8}$$

$$\mathcal{R}_{l=1,i} = \sum_j \frac{A_i W_{l=2,i,j}}{\sum_i A_i W_{l=2,i,j}} Y, \tag{9}$$

$$\mathcal{R}_i = \mathcal{R}_{l=0,i} = \sum_j \frac{W^2_{l=1,i,j}}{\sum_i W^2_{l=1,i,j}} \mathcal{R}_{l=1,j}. \tag{10}$$

Here the time sample input data is denoted by X and its NN output is denoted by Y. The output of neuron j is denoted by A_j. We define that the input, hidden, and output layers have layer numbers $l = 0$, 1, and 2 respectively. Weight matrices from layer $l-1$ element i to layer l element j are denoted by $W_{l,i,j}$, and the associated bias arrays are $b_{l,j}$. The relevance of layer 1, $\mathcal{R}_{l=1,i}$ is an intermediate quantity used to compute the relevance of the input elements, $\mathcal{R}_i = \mathcal{R}_{l=0,i}$. LRP has multiple variants, and here we use the approach outlined in Bach et al. (2015). Other variants we tried produced qualitatively similar results.

3. Evaluating the Methodology: Reconstructing MOC at Select Latitudes

In this section we center our analysis on one latitude of interest per basin and MOC cell. We use these sections as test cases to illustrate the impact of some of the key choices made in our methodology, and under what conditions the methodology yields skillful predictions. We choose the 26°N AMOC latitude, due to the presence of the RAPID-MOCHA array and for comparison with several recent AMOC reconstructions at this latitude (Section 1). We choose the 30°S IPMOC latitude since it is in the region where the IPMOC is most intense (IPMOC gradually decays north of the equator, as seen in Figure S1 in Supporting Information S1), and for comparison with the subtropical latitude we chose for AMOC. We choose 60°S for the SOMOC abyssal cell since this latitude band passes through Drake Passage and hence is dynamically distinct from a zonally-bounded basin (e.g., AMOC). We choose a slightly different latitude, 55°S, for the SOMOC mid-depth cell since the mid-depth cell strength is much weaker at or south of 60°S, and since our reconstruction skill diminishes sharply for the mid-depth cell near and southward of 60°S as shown in Section 4.1.

We present results using a hidden layer with number of neurons between 1 and 5. Larger layers did not produce any significant advantage in our tests. Unless specified otherwise, reported skill values are median scores over 20 NN realizations, and presented time series correspond to the realization with the median score. We defer the presentation of results with linear activation functions or with linear regression to Section 4, where these are compared across a range of latitudes and in different basins.

3.1. Timeseries and Time Scale Dependence

A NN-reconstructed AMOC time series at 26°N is compared with the "real" (ECCO) time series in Figure 3a. The NN input data provided is comprised of OBP and zonal mean ZWS. The number of neurons is three. The AMOC NN reconstruction has a correlation of ∼0.98 with the model AMOC time series during the testing period (Figure 3a). We note that the reconstruction approximately captures several anomalous AMOC lows in the testing period, despite their values lying outside the range of AMOC values in the training data. IPMOC 30°S

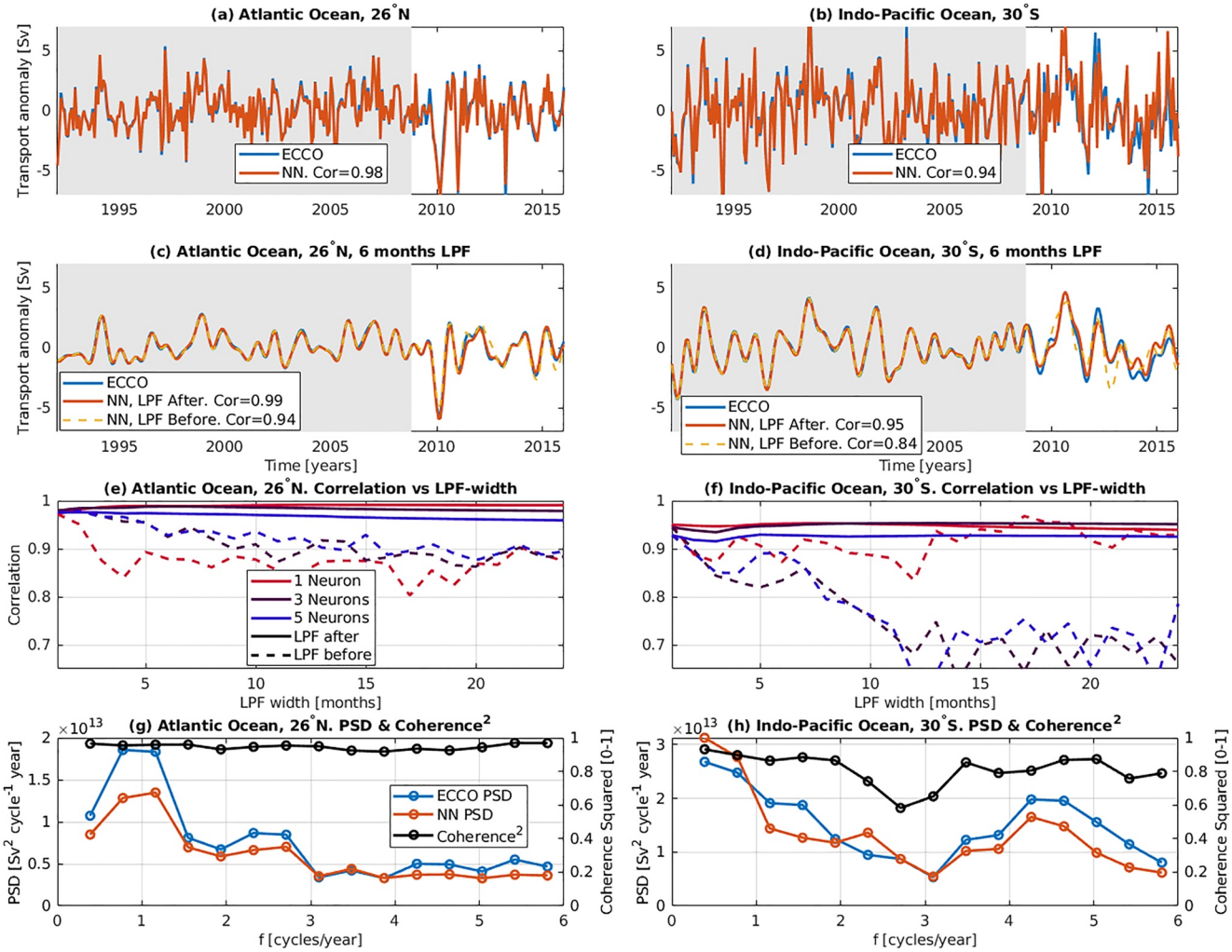

Figure 3. (a, c) ECCO AMOC time series at 26°N and reconstruction, from a neural network (NN) trained with input of bottom pressure and zonal-mean zonal wind-stress, with one hidden layer of three neurons. (b, d) ECCO IPMOC abyssal Meridional Overturning Circulation (MOC) at 30°S and reconstruction, from a NN trained with input of bottom pressure only, with one hidden layer of three neurons. Training and testing data partitioning is shown by the light gray and white backgrounds, respectively. The testing-period correlation between the model and reconstruction MOC time series is shown in each panel title. In (c and d), short time scales are filtered out from the reconstruction shown in panel (a or b) using a 6-month low-pass filter (red line). Alternatively, the NN is trained based on low passed filtered input data and filtered MOC time series (dashed orange line). The filtered ECCO MOC time series is shown in blue. Time series in panels (a–d) are those with median reconstruction skill over 20 NN realizations. In panels (e and f) the reconstruction (correlation) skill is shown for different choices of low pass filter (LPF) width applied to the full reconstruction (solid lines) or to the covariates and MOC series before NN training (dashed lines), for a NN hidden layer of 1, 3, or 5 neurons. For all LPF widths we plot the median over 20 realizations of the trained NN. All shown correlation values are statistically significant at the $p = 0.05$ level. In (g and h) we show the power spectral density for the NN series from panels (a and b), respectively, as well as (with values on the right axis) the coherence between these series and the respective ECCO MOC time series.

reconstruction is similarly skillfull (correlation ~0.94, panel b) although based on OBP alone. For the Southern Ocean 60°S abyssal MOC (Figure 4a) and 55°S mid-depth MOC (Figure 4b) cells reconstruction, as for the IPMOC 30°S time series, only OBP is used as an input variable, as we find that adding ZWS or other variables does not increase reconstruction skill there (Section 3.3). The NN reconstruction correlation with three neurons is 0.98 for both Southern Ocean MOC cells at the latitudes shown.

From a climatic viewpoint, it is also of interest to examine the reconstructed MOC fidelity on longer time scales. We begin by examining a 6-month LPF, applied before or after training (see Section 2.3) in Figures 3 and 4 panels c, d. In all cases the reconstruction skill does not change relative to the unfiltered case when applying the LPF after training, but does degrade slightly when the LPF is applied before training (correlation values given in the legend). This trend continues when considering longer time scales, at least up to 24 months (Figures 3 and 4 panels e, f). Significantly longer time scales can not be examined because the testing period is just 7 years long.

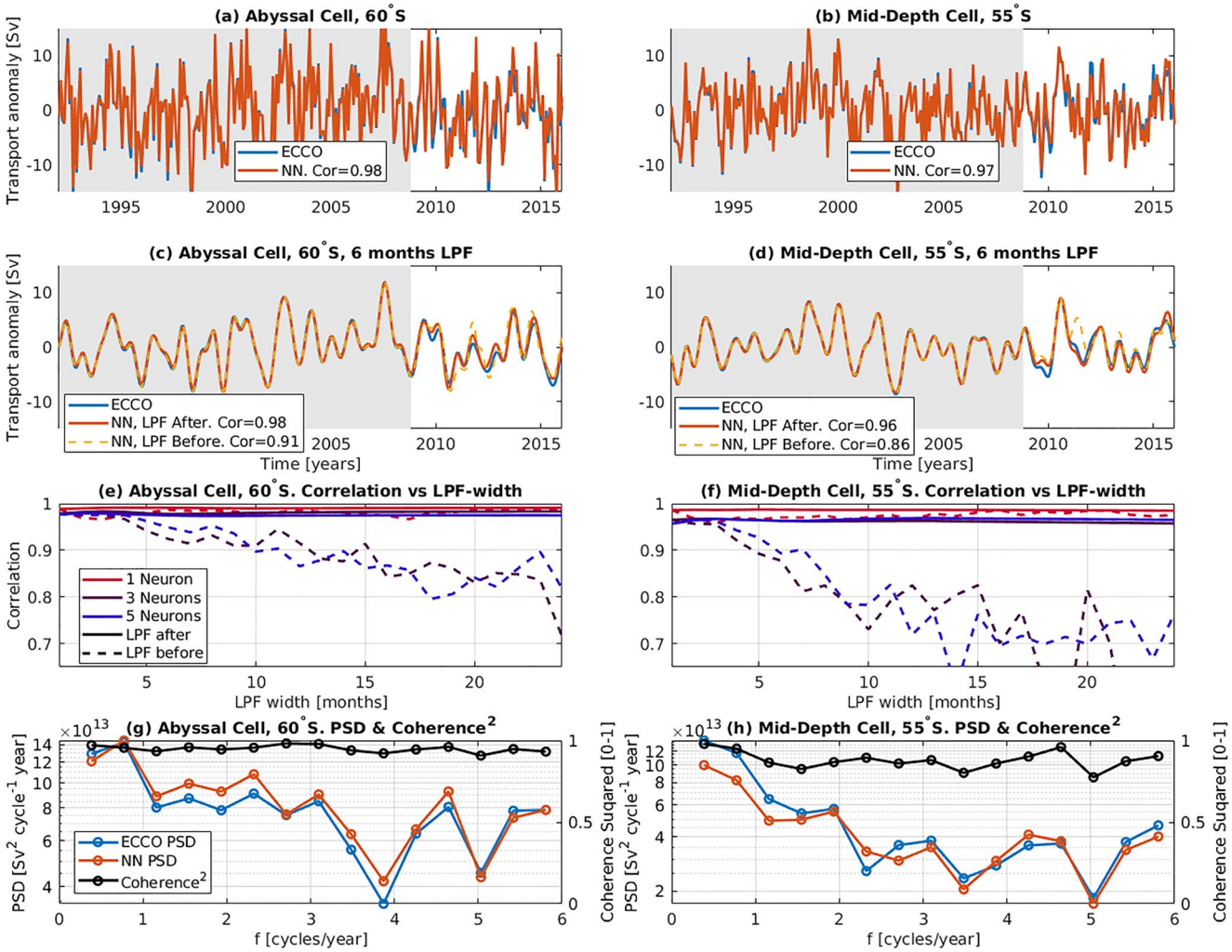

Figure 4. As in Figure 3 except that panels (a, c, e, g) relate to the SOMOC abyssal cell at 60°S, and panels (b, d, f, h) relate to the SOMOC mid-depth cell at 55°S (both based on Ocean bottom pressure alone).

The fact that skillfull reconstruction of long time scales is aided by training with short time scales (i.e., the higher skill of LPF after rather than before training) suggests that short and long time scales (up to 2 years) of variability have the same underlying dynamics. Panels e, f also show the dependence of NN skill on the number of neurons (up to five) in the hidden layer of the network. This dependence is generally quite weak (provided that the LPF is not applied before training). The high skill achieved with even a single neuron suggests a relatively simple underlying dynamics (which should be at least partially related to geostrophy and to Ekman flux (Bingham & Hughes, 2008; Kanzow et al., 2008; McCarthy et al., 2015)). The lack of significant improvement with multiple versus single neurons may be attributed to the small number of temporal samples used to train the network not being sufficient to constrain a large number of parameters.

At interannual time scales, the present AMOC skill is similar to the maximal skill achieved by the dynamically-based method of Sanchez-Franks et al. (2021), but here the skill additionally remains high over a wide range of time scales. The method of Frajka-Williams (2015) achieved even higher skill based on linear regression of altimetric SSH (also using wind stress and the in situ observed Florida Strait velocity), but did not separate the data into training and testing partitions. However, these real-world MOC reconstruction efforts are also likely more challenging than the model MOC reconstruction due to factors such as measurement noise (Section 5). The abyssal Southern Ocean NN reconstruction accounts for (i.e., $R^2 \geq$) 95% of the 60°S MOC variability. This is ≈25% points higher than the MOC variability explained at similar latitudes using the theoretical relationship between the abyssal MOC and ZWS proposed by Stewart et al. (2021).

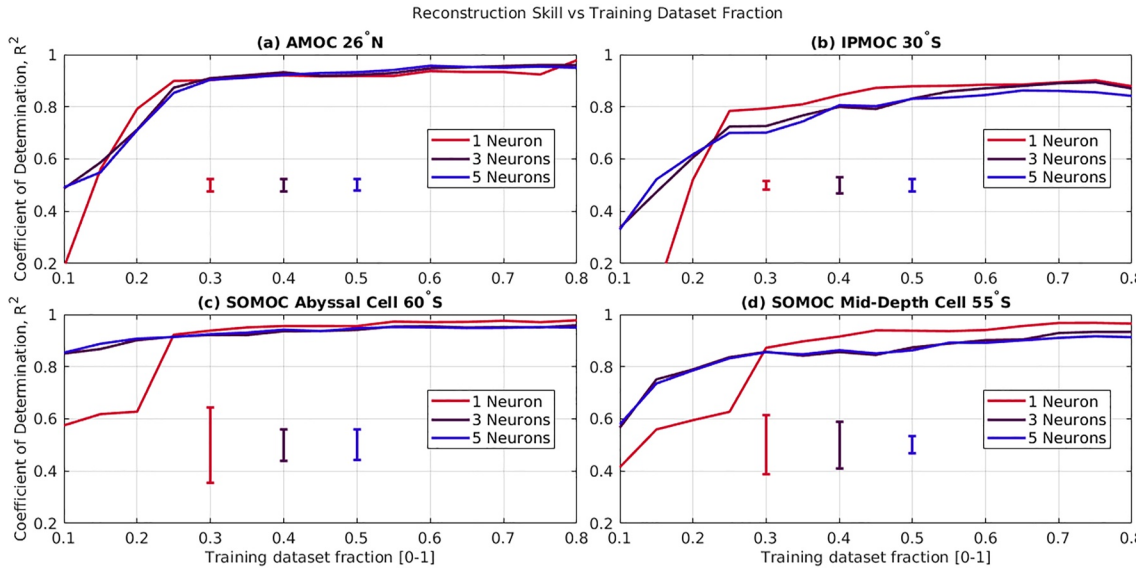

Figure 5. Meridional Overturning Circulation (MOC) reconstruction R^2 skill score versus the fraction of the time samples (i.e., fraction of 24 years) used for the training set. Each presented data point is the median score for 20 neural network realizations. Error bars show the (median value over all data set fraction values of) realization-standard deviation. Results with one, three, or five neurons are shown (color coded). Results shown per panel are: (a) AMOC at 26°N reconstructed using Ocean bottom pressure (OBP) and zonal wind stress, (b) IPMOC at 30°S reconstructed using OBP, (c) Southern Ocean abyssal MOC cell at 60°S reconstructed using OBP, and (d) Southern Ocean mid-depth MOC cell at 55°S reconstructed using OBP.

We show further support for the reconstruction skill at various time scales by examining spectral properties of the reconstruction (Figures 3 and 4 panels g, h). In this case we train the NN using just 40% (rather than 70%) of the data set, in order to increase the range of frequencies over which the reconstruction spectrum can be calculated (in the testing data set). The PSD of the MOC time series and its NN reconstruction, as well as their spectral coherence, are calculated (Section 2.4). The longest period in the PSD is 26 months. At each frequency, the difference between the reconstruction and the ECCO MOC is an order of magnitude smaller than the spectral 95% confidence interval (not shown). The mean and standard deviation of squared spectral coherence across the presented frequency range in each case are as follows: SOMOC abyssal cell $C^2 = 0.95 \pm 0.02$, SOMOC mid-depth cell $C^2 = 0.88 \pm 0.06$, AMOC $C^2 = 0.92 \pm 0.03$, and IPMOC $C^2 = 0.81 \pm 0.1$,

3.2. Reconstruction Skill Versus Training Set Size

Since satellite observations and ocean state estimates have relatively short time spans of a few decades, it is imperative to examine the dependence of reconstruction skill on the temporal duration of the data set. Therefore, we retrained the NN cases shown in Section 3.1 using training sets with sizes between 10% and 80% of the total 24 years-long data set (Figure 5). The covariates used in each case are the same as in the previous subsection.

We find that high skill can be attained with relatively short training periods: the NNs attain R^2 values of (depending on basin and number of neurons) 0.35–0.85 for a training set size of just 10%, that is, 2.5 years. The reconstruction skill grows most steeply until a training set size of ≈30% to reach values of $R^2 \approx 0.7$–0.9, after which the skill grows little (by 0.1 or less) with further training. The high performance at the smallest training data set sizes demands several neurons, with one neuron becoming competitive for training datasets exceeding 20% of the time series.

3.3. Optimal Satellite-Observable Input Variables

To determine which combinations of satellite-observable variables considered here (OBP, SSH, ZWS, SST, and SSS) result in the most skillful prediction of the MOC, in Figure 6 we show their R^2 skill scores for each basin and latitude considered in the previous subsection. The median skills over 20 realizations are shown in green, and realization-maximum skill is in black. Of all single-variable cases OBP has the most skill, except for AMOC 26°N, where ZWS has nearly equal skill to OBP. SSH has low skill in the AMOC and IPMOC cases, but substantial skill in the Southern Ocean, if smaller than that of OBP.

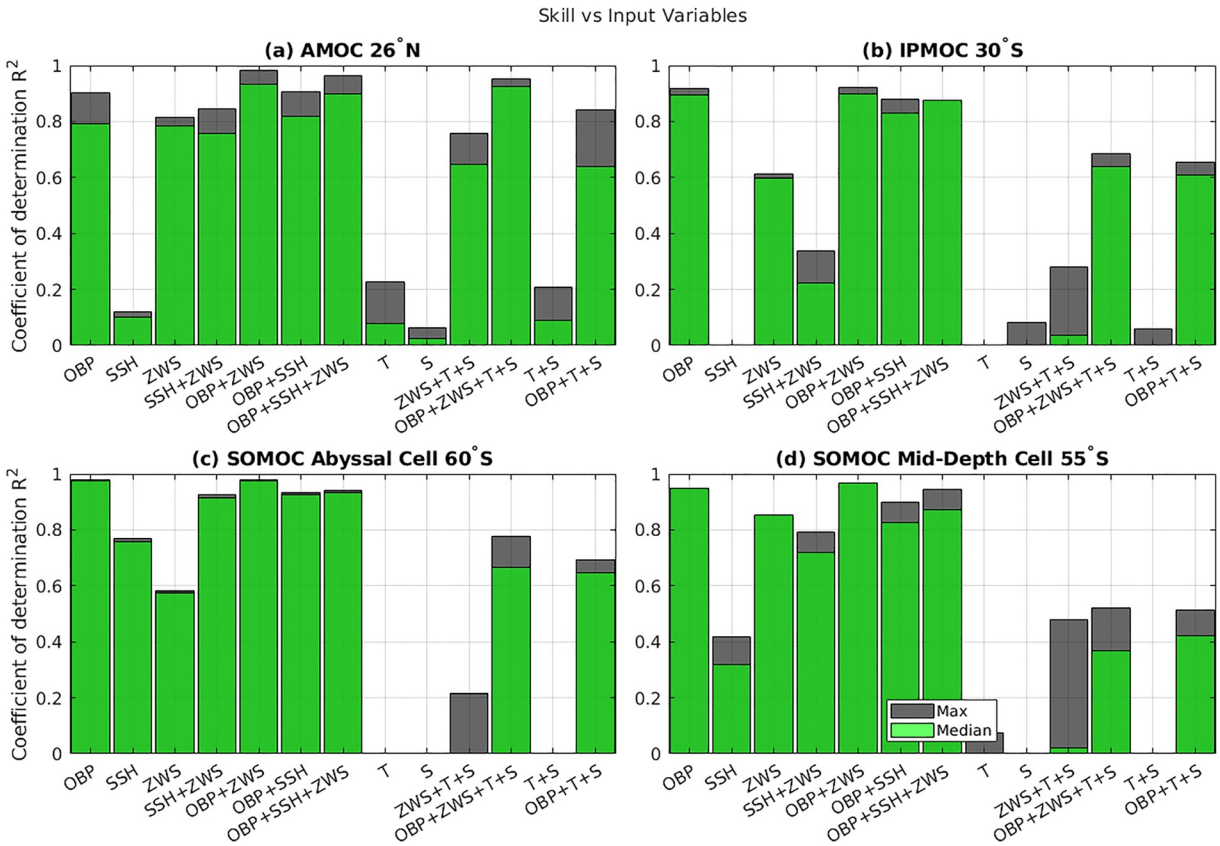

Figure 6. Reconstruction skill (R^2) of various combinations of input data (satellite-observable variables) for (a) AMOC 26°N, (b) Indo-Pacific Meridional Overturning Circulation (MOC) 30°S, (c) Southern Ocean abyssal MOC cell 60°S, and (d) Southern Ocean mid-depth MOC cell 55°S. The median skill between 20 neural network realizations is shown in green, and the maximum in black. Used variables and thier acronyms in the figure: sea-surface temperature (T) and salinity (S), Ocean bottom pressure (OBP), sea surface height (SSH), and zonal-mean zonal wind stress (ZWS).

The significance of local ZWS for MOC variability in the Southern Ocean is expected from the considerations of the zonal momentum balance (Stewart & Hogg, 2017; Stewart et al., 2021). Zonal wind stress as a sole predictor yields $R^2 \approx 0.58$, which is approximately the value of correlation squared found between the MOC and wind stress based on their dynamical relation in Stewart et al. (2021) on monthly time scales. It is also expected that ZWS would be skillful for AMOC based on simple Ekman flux considerations, AMOC observations (Frajka-Williams et al., 2019), and inverse modeling attribution studies (Kostov et al., 2021; Pillar et al., 2016). Smith and Heimbach (2019) have shown that almost 80% of AMOC variability at 34.5°S is due to local wind forcing. In contrast, SSH alone yields a very low skill in predicting AMOC at 26°N. It may be expected that OBP should be more directly related to AMOC, as it more directly quantifies pressure gradients at the depth of the meridional flow that comprises AMOC. We note that previous studies that found higher AMOC reconstruction skills with SSH have utilized SSH in conjunction with additional variables, for example, Florida Straits transport data (Frajka-Williams, 2015) at 26°N, or Argo data near 41°N (Willis, 2010).

Adding ZWS to OBP improves skill in most cases, presumably since a component of ZWS forces ageostrophic circulation which its flux is not well addressed using OBP. And presumably for the same reason adding ZWS to SSH improves skill in the SOMOC abyssal cell, where both variables have substantial skill already. That is not so in the other cells, where SSH has low or negative R^2 values and hence adding it to ZWS complicates the NN needlessly and degrades the skill. Additionally, adding SSH to OBP or adding SSH to OBP and ZWS does not increase skill relative to the latter's skill, that is, there isn't enough MOC variability information embedded in SSH that isn't already included in OBP, or that can be utilized in the present framework.

Reconstructions based on SST and salinity (SSS) yield very low R^2 values ($\leq$0.1) in all analyzed cases. The addition of SST and/or SSS (except for marginal increase from $R^2 = 0.1$ with SSH alone to $R^2 = 0.2$ with SSH + SST)

degrades the reconstructions based on SSH, OBP and/or ZWS. The degradation is possibly due to over-training of the NN associated with the larger number of NN weights, that is, degrees of freedom (DOF). For AMOC, SSS, and SST are expected to become more important in the subpolar North Atlantic (Pillar et al., 2016). However, the NN reconstruction skill at 55°N is similar to 26°N if based on SST, and increases only modestly to $R^2 = 0.25$ if based on SSS. The combination at 55°N of SST, SSS, and ZWS is more skilled ($R^2 = 0.78$) than ZWS alone there ($R^2 = 0.64$), but less skilled than OBP or SSH combined with ZWS ($R^2 \approx 0.9$, Figure 10). Adding SST or SSS to other variable combinations does not increase their skill at 55°. The low skill found with SST and SSS in the present NN architecture input is on par with previous studies: AMOC variability has been shown to be related to surface thermohaline variables or to surface thermohaline fluxes primarily on basin-scale and decadal or longer time scales (Duchez et al., 2016; Knight et al., 2005; Pillar et al., 2016; Rahmstorf et al., 2015; Smith & Heimbach, 2019). The basin-scale dependence cannot be captured by the present local framework, nor is the decadal time scale resolved in the (detrended) 24-year duration of the state estimate used here. Regarding the lack of skill using SST or SSS for SOMOC reconstruction, while Stewart and Hogg (2017) have shown that SST is correlated with MOC variability in an idealized model configuration with a single bathymetric ridge, it may be more relevant for reconstruction of local rather than zonal-mean SOMOC variability. Satellite observable variables related to sea ice are considered in Section 4.1 for the Antarctic margins latitudes.

We have also examined the possibility of augmenting OBP, SSH, and ZWS by their values at one or several lagged times. For example, to reconstruct the MOC at a particular month we might use the zonal OBP distribution at the same month, as well as in the previous 2 months. We examined additions of up to 8 consecutive prior months to the input, but found that this augmentation led to reconstruction skills that were similar to or lower than those obtained using only contemporaneous input information. This is perhaps to be expected due to the small number of training samples available, and due to the spatially-local specification of the covariates relative to the reconstructed MOC. We might expect temporal lags become important if covariate latitudinal dependence were included in the NN.

3.4. Interpretation of MOC-Covariate Relations

In this section we examine the dependence of the NN skill on the zonal structure of the input covariates, to understand which (if any) longitudinal locations are associated more strongly with MOC variability. Such knowledge is of use practically in designing NN architecture, and theoretically to explain why covariates are associated with MOC variability.

The first method we use in order to address the above question is zonal subsampling of the input data, that is, using a subset of the ECCO longitudinal grid points in a particular basin and latitude band. This also reduces the NN DOF (the number of weights and biases), which can limit overtraining; note that the input consists of only 288 time samples, which without zonal subsampling is less than the number of DOF even with a single neuron. We use a number N_d of zonally equal-spaced points, with the first and last points being the western and eastern boundaries in the case of a zonally-bounded basin. We find (Figure 7a) that the AMOC reconstruction skill at 26°N is virtually independent of the number of interior basin points, as long as the western boundary point is included. There is a moderate decrease to a correlation of 0.9 if the western boundary point is the only one included. This finding is consistent with the results of Bingham and Hughes (2008) at 42°N (Section 1). We verify that the interannual skill is similar to that of Bingham and Hughes (2008) at 42°N with the present methodology as well, although the different models and methodologies preclude an exact comparison.

For zonal subsampling of input data in the Southern Ocean at 60°, a starting longitude specification for the N_d equally spaced points is required since there are no zonal boundaries. We repeat the NN calculation for all possible starting point (or "phase") choices spaced by 5° longitude (from −180°E). We find that the maximal reconstruction skill (correlation) over all possible phases does not decrease appreciably until $N_d \leq 10$, and remains above 0.9 for $N_d \geq 5$. The phase-minimal skill is close to the phase-maximal skill until it drops sharply under $N_d = 8$, while the realization-spread (≈0.05) is much smaller than the phase spread. This means that the actual zonal locations used for the covariates, rather than only their number, are significant determining factors of the NN skill for $N_d \leq 8$. The dependence on phase for a single observational point ($N_d = 1$, with maximal correlation ≈0.7, panel b) has a distinct zonal structure. The skill is maximized for observations near significant bathymetric ridges (see red line in Figure 7d and Appendix A). The widest and highest peak is in the region of Drake Passage and the Scotia Arc, near −50°E. Additional peaks occur at the Pacific Antarctic Ridge crossover at −150°E, and at the eastern Enderby basin slope near 60°E.

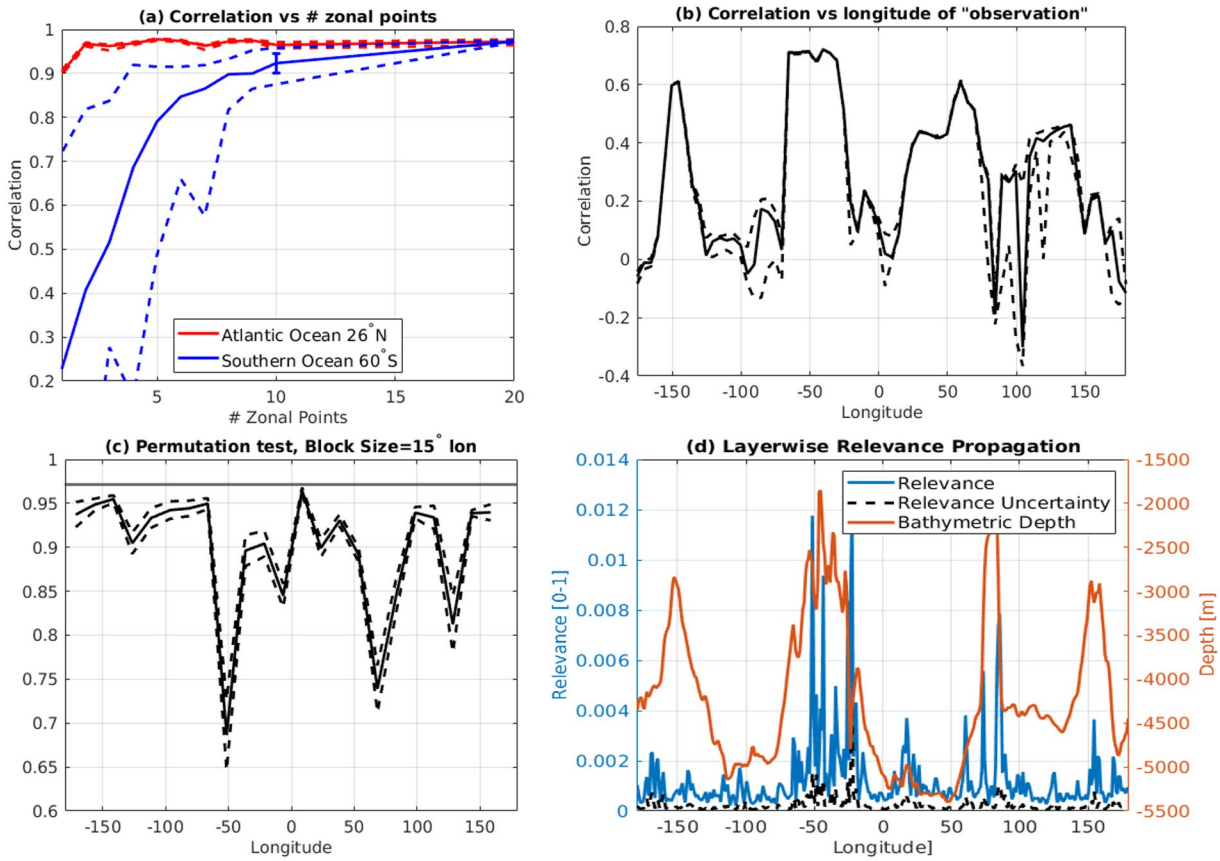

Figure 7. (a) Neural network (NN) skill versus number of zonal input longitudes. AMOC 26°N (SOMOC 60°S) reconstruction is shown in red (blue). The red (AMOC) solid and dashed lines show reconstruction skill realization-median and interquartile range (IQR), respectively. The blue (SOMOC 60°S) solid line shows the median reconstruction skill for all realizations and longitude "phases" (see main text). The dashed blue lines show the phase minimum and maximum of realization-median skills at the same section. (b) Reconstruction skill of SOMOC at 60°S using a single longitudinal point as a function of the longitudinal position of that point (i.e., the phase). (c) Permutation test for the relative importance of different longitudinal positions to the NN skill in predicting the SOMOC at 60°S. A permutation block size of 15° longitude is used. (d) Relative "relevance" (based on Layerwise Relevance Propagation) of different longitudinal positions to the NN prediction of the SOMOC at 60°S. The solid and dashed lines in panels (b–d) show the median and interquartile range, respectively. In all panels, 20 NN realizations (per phase, where relevant) are used. Bathymetric depths are shown in the red line in panel d.

We additionally use two ML interpretablity techniques: permutation and LRP (Section 2.5) to examine the relative importance of data from different locations when using full longitudinal input data (not limited to one location as above). The permutation test shows that the largest drop in skill occurs when scrambling covariate data from the Drake Passage, or from the eastern Enderby basin slope (panel c). Finally, LRP (panel d) also finds the largest "relevance" is located in the Drake Passage region, with three narrow peaks near gaps in and at the eastern slope of the Scotia Arc. LRP also identifies smaller, narrower peaks of longitudinal relevance at the eastern Enderby basin slope, as well as peaks on both flanks of the Kerguelen Plateau, and at the Pacific Antarctic Ridge near 154 and 169°E.

In summary the various interpretability techniques are in agreement in that the covariates' impact on the NN skill is maximal near deep ridge systems, and they also identify some of the same specific locations of largest impact/relevance. The importance of OBP distribution around sub-sea ridge systems can be understood as a consequence of the balance between bottom form stress anomalies (which are linear in OBP) and of anomalies in the Coriolis force acting on meridional transport anomalies (Stewart et al., 2021; Stewart & Hogg, 2017).

4. Application to the Global MOC

In this section we expand our reconstructions of the MOC to the full range of latitudes occupied by the MOC in the Southern, Atlantic, and Indo-Pacific Oceans. The configuration of the NN is as described in Section 3.1,

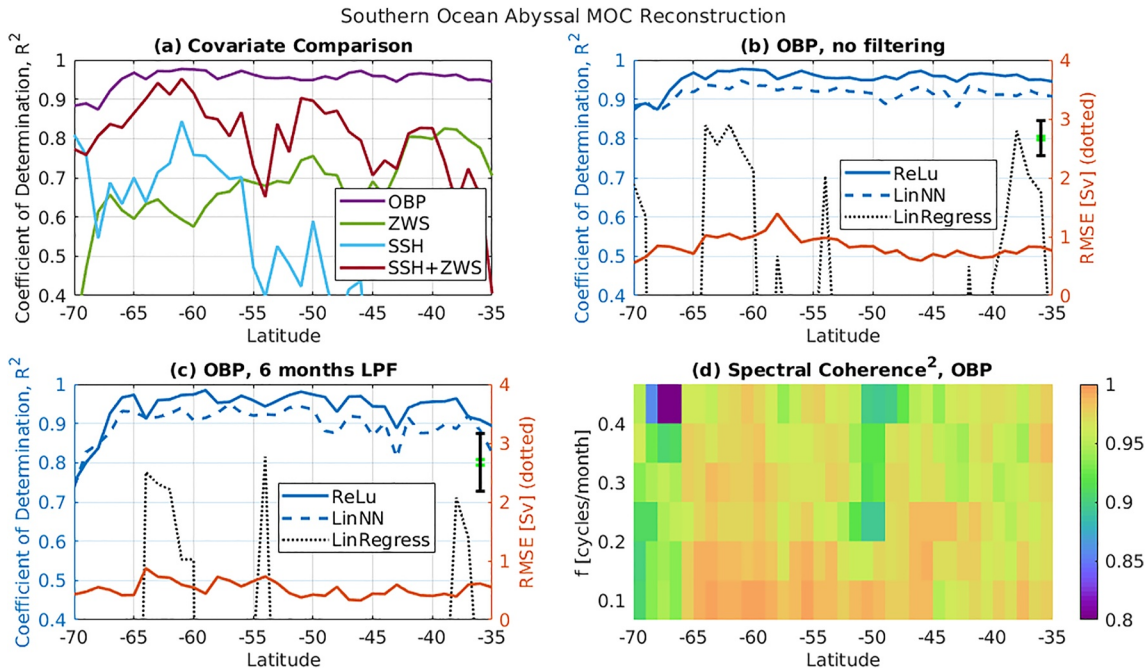

Figure 8. (a) Neural network (NN) Reconstruction skill (R^2) versus latitude for the Southern Ocean Meridional Overturning Circulation (MOC) abyssal cell, based on input of different covariates, as given in the legend: ocean bottom pressure (OBP), sea surface height (SSH), (zonal-mean of) zonal wind stress (ZWS), or a combination of SSH and ZWS. (b) The solid blue line shows results from a ReLu-based NN (as in panel a) while the dashed blue line shows results from a linear activation function -based NN, and the dashed black line shows results for linear regression. The covariate used is OBP (as in the purple panel (a) line). Green (black) error bars show $\pm S$ for ReLu (linear) NNs, where S is the latitude-median of the NN skill standard deviation across realizations. Panel (c) is identical to panel b except that a 6-month low pass filter is applied to the reconstruction (and to the ECCO time series) before estimating its skill. Panel (d) shows the squared spectral coherence between the OBP ReLu-based NN reconstruction (from panels a and b) and the ECCO MOC time series. Other symbols and skill metrics are as described in this figure.

unless stated otherwise. Note that in the figures of this section we also examine the NN skill obtained using linear, rather than ReLu, activation functions, and also with linear regression. However, discussion of these alternative approaches is mostly postponed to Section 5. We present results using a hidden layer comprised of a single neuron. Results with more neurons, which in most cases are similar or of lower skill, are given in the Supporting Information S1 (figure S1). Reported skill values are median scores over 20 randomly-initialized NN realizations.

4.1. Southern Ocean Abyssal MOC Reconstruction Skill Versus Latitude

Using the same NN configuration found to be optimal at 60°S (Section 3.3), that is, with input data composed of OBP alone, the abyssal MOC cell strength reconstruction skill with just one neuron is higher than $R^2 = 0.94$ everywhere north of 65°S (Figure 8a). As was found at 60°S (Figure 3), performance with more neurons is similar to or worse than with a single neuron (Figure S1 in Supporting Information S1). The RMSE varies between 0.55 and 1.4 Sv, with a peak that occurs near 58°S (Figure 8b). This peak in RMSE is associated with a relatively small decrease in R^2, since the RMSE peak is partially compensated by a local peak in the temporal variability of the MOC strength (see Supporting Information S1).

Reconstruction using ZWS results in skill $R^2 \approx 0.7$, a similar value to that found in Stewart et al. (2021) based on a dynamical framework. SSH also has skill $R^2 \approx 0.7$ south of 55°S, but very low skill further north. Stewart and Hogg (2017) have suggested that abyssal water export variability along bathymetric obstacles should also modulate the route that the deep-reaching ACC takes around the same obstacles, and hence modulate the SSH distribution as well. This mechanism may produce a part of the SSH signature learned by the NN. North of 55°S, SOMOC-related Deep Western Boundary Currents flow northward in paths adjacent to land, and hence their variability cannot affect the ACC in the same manner, perhaps explaining the lower skill of SSH in these latitudes. However, a combination of SSH and ZWS results in a substantial increase in skill to a latitudinal mean value of $R^2 \approx 0.8$, with a substantial dip in skill only near 35°S. In contrast, using OBP and ZWS together (not shown) results in only a marginal increase in skill (mean R^2 increase of ~0.01) relative to OBP alone.

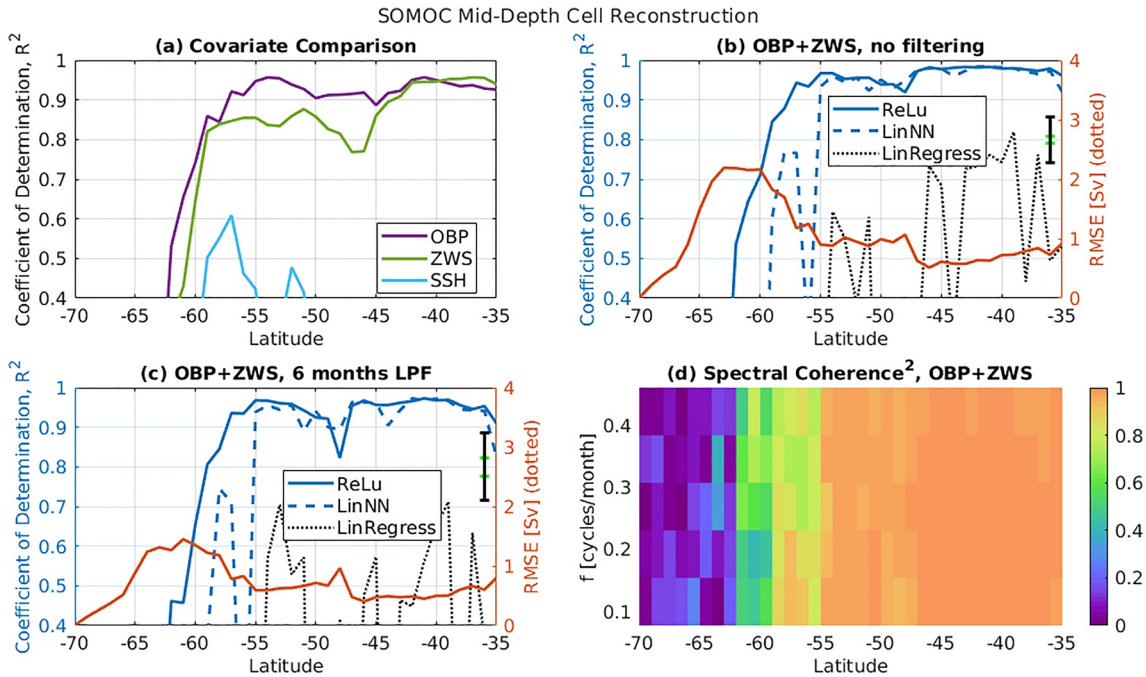

Figure 9. As in Figure 8, but applied to the Southern Ocean mid-depth Meridional Overturning Circulation cell. Covariate choices differ from Figure 8—see panel (a) legend and panels (b–d) titles).

The reconstruction skill using OBP at 6-monthly or longer time scales is explored in Figure 8c. At most Southern Ocean latitudes, the NNs skill changes little with application of a 6-month LPF, as was shown for 60°S above (Figure 3). However, skill does decrease slightly to $R^2 \approx 0.9$ in several latitude bands north of 45°S. Zonal land barriers appear near this latitude in the Indo-Pacific sector, which perhaps complicates or at least changes the MOC-OBP relation over long time scales. The spectral coherence of the NN reconstruction and ECCO MOC time series is shown as a function of latitude and frequency in Figure 8d. It is calculated in the same manner to Section 3.1 except that the training data set fraction is not changed from 70% to 40%, and hence the lowest frequency calculated is higher here, that is, 15 months. The coherence squared has near uniform (and high) values at all latitudes and frequencies, that is, a mean value of 0.98 and standard deviation of 0.2. All shown coherence values are statistically significant relative to a 0.01 p-value (Section 2.4).

South of 65°S the OBP reconstruction skill gradually decreases to $R^2 = 0.88$ at 70°S in the unfiltered case, and more so in the LPF filtered case. We tested using other covariates or their combinations in this latitude band, including OBP together with SSH and/or ZWS, zonally averaged or not, and (motivated by the complex shelf bathymetry) meridional wind stress. We performed an additional test to examine if using explicit sea-ice thickness information (which can also be derived from satellite data, albeit still with significant uncertainty) (Kurtz & Markus, 2012; Kwok, 2010; Petty et al., 2020) at these latitudes can counter the skill reduction around Antarctica. To do so we added as covariates to OBP the ECCO variables "SIheff" and "SIhsnow," that is, sea-ice thickness and snow pack thickness, respectively, scaled by grid-cell fractional sea-ice area. However, no improvement in skill so resulted from any combination of the above-described input variables (not shown) and OBP appears to be an optimal input variable for NN MOC reconstruction with the present methodology at these latitudes as well. The decrease in skill there may signify more complicated OBP-MOC relations at the southernmost latitudes due to shallower waters, and additional processes at the continental slope region, for example, topographic waves which may be aliased in the monthly data.

4.2. Southern Ocean Mid-Depth MOC Reconstruction Skill Versus Latitude

The SOMOC mid-depth cell is reconstructed with high skill using OBP alone: mean R^2 north of 58° is 0.93 (Figure 9a). Further south, skill degrades sharply, possibly due to the weakness of the mid-depth MOC cell there (e.g., weaker than 1 Sv south of 65°S, Figure 1 and Figure S2 in Supporting Information S1). Zonal wind stress attains similar if slightly lower skill values, in line with much of the variability on these time scales being wind-forced. However,

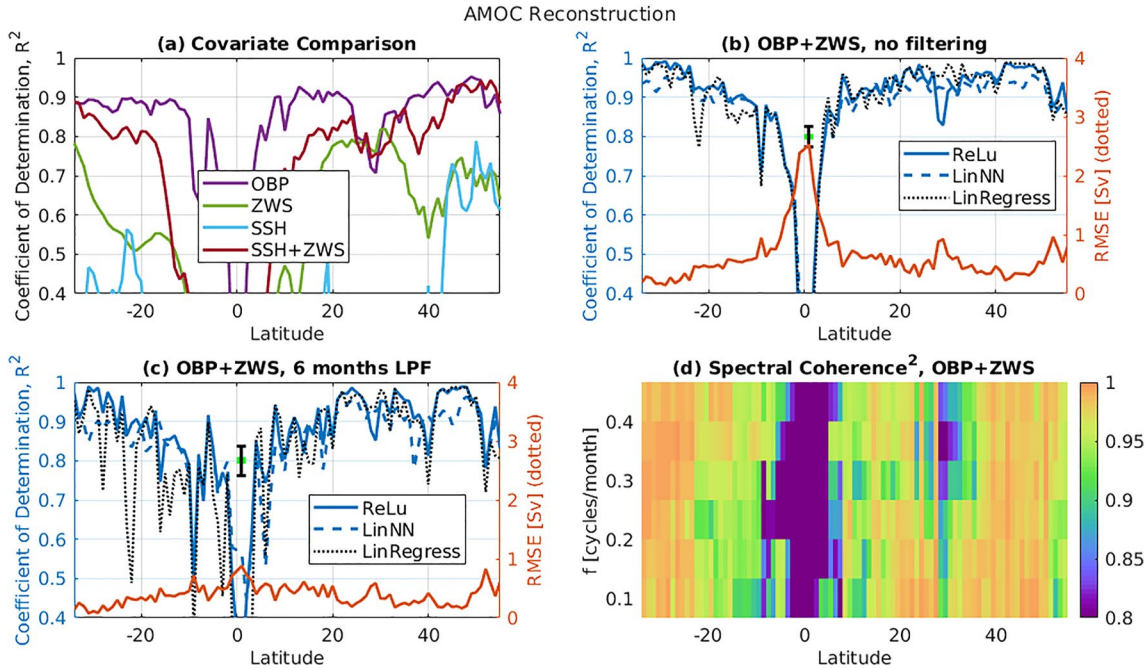

Figure 10. As in Figure 8, applied to AMOC. Covariate choices differ from Figure 8—see panel a legend and panels (b–d) titles).

combining OBP and ZWS does increase reconstruction skill in most latitudes (Figure 9b). SSH has very low reconstruction skill in most latitudes, but achieves a local peak between 55 and 60°S, similarly to the abyssal cell, albeit with somewhat lower R^2 (≈0.5 vs. 0.7). The high skill of OBP similarly to the abyssal cell, and the co-location of SSH-skillful latitudes (in latitudes where both cells are strong), suggests that the mid-depth reconstruction skill is due to the bottom form stress variability due to barotropic response to wind stress anomalies, and the associated SSH adjustment, as suggested by Stewart et al. (2021) in the context of the abyssal cell. Skill of longer than 6-month time scale reconstruction is comparable to the unfiltered case (Figure 9c). The mean coherence value north of 56°S is 0.92 (Figure 9d). All shown coherence values are statistically significant relative to a 0.01 p-value (Section 2.4).

4.3. AMOC Reconstruction Skill Versus Latitude

As for the SOMOC, the AMOC reconstruction skill is high ($R^2 \approx 0.9$) at most latitudes when using OBP alone with a single neuron (Figure 10a). Using more than one neuron (figure S1 in Supporting Information S1) produces similar or lower skill near 30°N (Figure 3). The skill drops equatorward to $R^2 \approx 0.7$ at ±4° latitude, and from there precipitously drops to near zero values near the equator. The skill drop toward the equator is likely a result of the breakdown of geostrophic balance there, that is, the diagnostic relationship between OBP and current velocity. Another, less severe, drop in skill ($R^2 = 0.6$) occurs at 7–9°S, that is, where the Deep Western Boundary Current locally breaks up into eddies (Dengler et al., 2004), again possibly associated with a breakdown between the diagnostic relation between (monthly-mean) pressure and transport (see discussion in Section 5). Skill also drops to $R^2 = 0.7$ near 30°N, possibly due to the shallow bathymetric depths at which the western boundary current occurs at these latitudes (Willis, 2010).

Zonal wind stress on its own attains its peak reconstruction skill $R^2 \approx 0.8$ between 20 and 32°N, that is, around the RAPID array latitude (26.5°N). Its skill drops northwards to a local minimum near 40°N (perhaps related with the boundary between the subtropical and subpolar gyres), before rising again to $R^2 \approx 0.7$ in the subpolar region. Indeed, a greater influence of wind stress on subtropical than subpolar AMOC variability has been demonstrated on interannual time scales by Kostov et al. (2021). Zonal wind stress skill is generally lower in the Southern Hemisphere, for reasons that remain unclear, although it increases at the southernmost latitudes. The large variation in ZWS skill, if interpreted as related to the actual role of ZWS in AMOC forcing, means that caution should be taken in applying results from AMOC variability attribution studies from one latitude to other, even nearby latitudes.

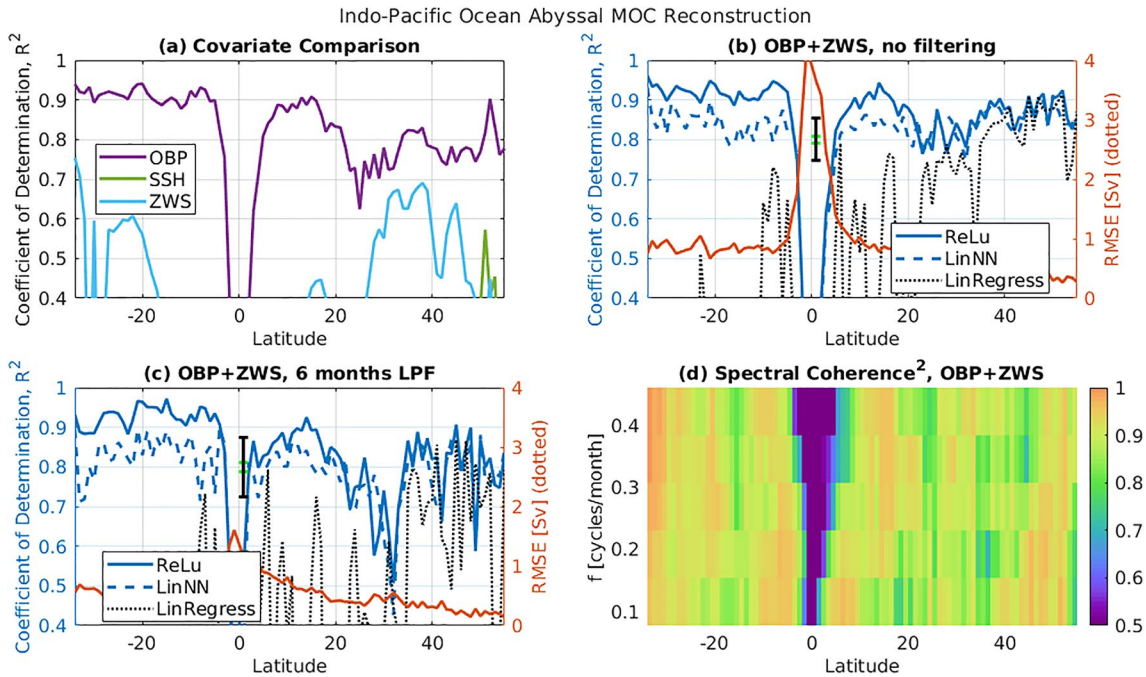

Figure 11. As in Figure 8, applied to the Indo-Pacific Ocean abyssal Meridional Overturning Circulation cell. Covariate choices differ from Figure 8—see panel a legend and panels (b–d) titles).

SSH has little reconstruction skill on its own except north of 43°N, where it attains $R^2 \approx 0.6$. That may be since stratification is on average lower in subpolar regions while the Coriolis parameter is higher, which tends to result in a more barotropic circulation (Isachsen et al., 2003; Nøst & Isachsen, 2003; Salmon, 1998). Using a combination of SSH and ZWS as joint covariates results in a significant improvement, with skill higher than $R^2 = 0.8$ in most latitudes. While this combination is still slightly less skillfull than using OBP alone, there may be advantages to using SSH rather than OBP in real conditions, for example, because satellite observations of SSH generally have higher resolution than those of OBP.

Using a combination of OBP and ZWS generally increases the reconstruction skill further (Figure 10b). Poleward of ±15°N, the latitudinal-mean skill is $R^2 \approx 0.95$. Mean RMSE in the same region is ∼0.4 Sv. The skill is quite similar with a 6-month LPF applied to the reconstruction (after training), that is, $R^2 \approx 0.93$ poleward of ±15° (Figure 10c). The dip in skill near 9°S (described above) is more severe with LPF. An additional dip occurs in the Atlantic Basin near 37–40°N. Similarly to the 8°S dip, there is a change of western boundary current state in this region, that is, the Gulf Stream separation starts near 37°N in the model, and becomes aligned more along than across the NN-input zonal strip by 40°N, after which the North Atlantic Current turns north again. In the same latitude band Leroux et al. (2018) find that AMOC has an anomalously high fraction of "chaotic," that is, intrinsic (vs. forced) variability. Additionally, Jamet et al. (2020) find there anomalously high forced variability near decadal and longer time scales. Decadal variability cannot be learned properly by the NNs in the present work since the ECCO4 output is limited to 24 years. The mean coherence value poleward of 5° latitude is 0.98 (Figure 9d), and its standard deviation across these latitudes is 0.025.

4.4. Indo-Pacific MOC Reconstruction Skill Versus Latitude

In this subsection we analyze a reconstruction of the abyssal MOC cell in the combined Indo-Pacific oceans (IPMOC). The reconstruction skill based on OBP varies little with latitude south of 20°N, attaining values around $R^2 = 0.9$ except for a large drop near the equator similarly to the AMOC case (Figure 11a). RMSE is everywhere smaller than 1 Sv (panel b). Further north, the skill based on OBP drops to around $R^2 = 0.75$, perhaps related to the decrease in the IPMOC MOC cell strength (Figure 1 and Figure S2 in Supporting Information S1). Reconstruction based on SSH has poor skill: its latitudinal-mean value being negative ($R^2 = -0.05$), that is, its RMSE is larger than the MOC strength standard deviation. That is perhaps unsurprising: variability of basin scale flow, for example, wind-driven gyres which are surface-intensified, and of its associated SSH, may be largely decoupled from the abyssal cell.

Reconstruction based on ZWS attains R^2 values larger than 0.4 only south of 30°S and between 30 and 50°N, and nowhere higher than 0.75, that is, much lower than OBP. Including ZWS and OBP as joint covariates does not increase skill south of 20°N (unlike the situation with AMOC, see previous subsection), but does increase skill (R^2) by ∼0.1 north of 20°N. We also tested using wind stress curl (WSC) as a (single or joint with OBP) covariate in a subset of latitudes, and found that this yielded even poorer skill. The relatively low skill with ZWS or WSC is perhaps surprising given either variable has been shown to exert strong influence on Eulerian IPMOC variability (Han, 2021; Tandon et al., 2020). The resolution may be that a large fraction of the wind-induced Eulerian MOC variability is likely compensated by the eddy-rectified component of the isopycnal MOC (which is the quantity considered in the present work).

The reconstruction skill using OBP and ZWS changes little on time scales longer than 6 months (Figure 11c) south of 20°N. Further north, where the unfiltered skill is lower, the R^2 skill drops on average by 0.08, and in some latitudes reaching values as low as $R^2 = 0.5$. However, the spectral coherence (panel d) does not show a clear degradation with increasing time scale toward the longest scales considered (15 months).

5. Discussion

5.1. Spatial Patterns in Reconstruction Skill

The MOC reconstruction skill is quite high ($R^2 \geq 0.8$) in most of the considered latitudes and basins. That is especially so for AMOC in the extratropics and for the abyssal cell across most of the Southern Ocean, with typical values of $R^2 \sim 0.95$. However, skill is somewhat lower in several regions. While some of the regional differences might be due to non-optimal training or NN architectures, differences in training efficiency should ultimately be related to dynamic or geometric reasons as well, as we discuss here.

Regions of reduced skill appear in most cases to be related to bathymetry (Figure A1) or bathymetry-related transitions. In the Southern Ocean lower skill occurs near the Antarctic continent, south of 65°S. Drops in skill also appear in the Indo-Pacific basin near 25°N, near the northern boundary of the deep Indian Ocean, and where also the Hawaiian Ridge in the Pacific Ocean is nearly zonally aligned. In these shallower regions, nonlinear and agesotrophic effects, and relatively larger dissipation may all introduce challenges to MOC reconstruction, since they result in deviations from the diagnostic capability of circulation from pressure gradients alone (i.e., geostrophy). The zonal-boundaries also introduce near-zonal currents in some cases, which increase covariate variabillty ("noise") without contributing directly to the (zonally-averaged) MOC. The latter issue may be addressed by not suppressing the latitudinal structure of covariates in reconstructing the MOC strength at a given latitude, but such extension remains to be explored in future work. Skill drops (as a function of latitude) are typically more severe in the LPF case, that is, in reconstructing longer variability time scales. That may reflect an increase (for longer time scales) of the impact of the suppressed covariates' latitudinal dependence. Additionally, the limited training time (≈17 years) may be insufficient for training the longer time scales.

In the equatorial regions of both Atlantic and Indo-Pacific basins, the NN skill is practically zero, likely due to the breakup of geostrophic balance. Additionally, two locations associated with western boundary current separation (37–40°N) and Deep Western Boundary Current breakup into eddies (9°S) are associated with lower AMOC reconstruction skill. The disintegration of the boundary current and increase of variability due to larger amplitude meanders and vortices also means that geostrophic balance may be somewhat compromised when considering monthly-mean quantities, as used for NN input here. For example, if the shed eddies (which may still contribute to MOC transport) travel at a moderate 10 cm s^{-1} velocity near the energetic separation area, over 1 month they will cover a distance of 250 km, which is comparable to or larger than typical mesoscale eddy diameters. Hence the pressure field associated with such a propagating vortex would be greatly reduced in amplitude within an Eulerian monthly average, and would also become elongated in its propagation direction. We note that while the observed ≈100 km radius eddies near 9°S (Dengler et al., 2004) can not be spatially resolved explicitly in ECCO, the change in boundary current character there may still manifest by higher variability and meandering, and by the parameterized eddy transport.

5.2. Linearity of MOC-Covariates Relations

MOC reconstruction with more than 1 neuron generally did not result in better reconstructions in our tests, as we show for up to five neurons in Figures 3–5, 8–11 and Figure S1 in Supporting Information S1. Tests with more than 5 neurons and up to 30 neurons of subsets of the cases discussed (not shown) did not result in

improvements either. The simplicity of the "optimal" NNs prompted us to consider linear activation functions as well (Figures 8–11). Perhaps the main advantage of linear versus nonlinear reconstruction (if their skill scores are similar) is in terms of greater simplicity of the linear reconstruction, although as argued by Lipton (2018), for complex multiparameter models linear relations are not necessarily more interpretable than nonlinear ones. We find, however, that for all considered MOC cells (AMOC, IPMOC, SOMOC) nonlinear NNs have similar or higher reconstruction skill across most latitudes when using monthly data. This result may be due to more successful reconstruction of rectified sub-monthly variability, or due to better noise reduction using ReLu.

Although in some of the examined cases linear NNs are competitive with nonlinear NNs, simple linear regression generally produces significantly lower MOC reconstruction skill in most of our tests (Figures 8–11). Thus the regularization imposed on the linear weights (Section 2.3) and training methodology are crucial. That is the especially the case for the Southern Ocean MOC cells and for the IPMOC (Figures 4, 8, and 11), where linear regression has near zero coefficient of determination skill metric in most latitudes. However, for AMOC linear regression reconstruction skill is generally only slightly lower than with NNs (although linear regression is not as successful for longer time scales, i.e., after LPF application). The relative success of (un-regularized) linear regression for AMOC reconstruction is in line with previous studies (Section 1). A possible reason for its failure for the abyssal and Southern Ocean MOC cases is that the driving dynamics are then more localized (which may be better imitated with regularized learning) due to anomalous bathymetric features (Stewart et al., 2021), as suggested for the SOMOC in Section 3.4.

5.3. Reconstruction of Real-World MOC Variability

We have demonstrated that the ECCO MOC can be reconstructed with high skill from knowledge of the considered proxy variables. Additionally, reconstruction root mean-squared-errors (RMSE) are generally similar to or lower than uncertainties associated with present monitoring arrays, where they exist (Section 4): the reconstructed AMOC RMSE in the subtropics (∼0.5 Sv) is similar to the uncertainty value estimated in the RAPID-MOCHA array, which is 1.5 Sv for 10-day averages, and 0.9 Sv for annual averages. The reconstructed RMSE near 55°N (∼2 Sv), is similar to the uncertainty associated with OSNAP array monthly estimates, which is ∼4 Sv (Lozier et al., 2019). Despite the success of the present approach as tested within the ECCO state estimate, we expect that the trained NN, as-is, would perform poorly on real world data due to inherent biases in the test bed, that is, the ocean model used here. Although ECCO is in a sense an optimal estimate of ocean circulation, non-zero state biases necessarily remain (Forget et al., 2015).

One model issue is the relatively low (non eddy-resolving) spatial resolution, that is, 1°. Therefore, eddy transport processes are parameterized rather than resolved. An encouraging finding is that while ECCO parameterized transport contribution to time-mean MOC is large in the Southern Ocean, its contribution to time-variability is relatively small (Stewart et al., 2021). Despite that, and although the effective resolution of satellite products is often not "eddy-resolving" (Pujol et al., 2016; Tapley et al., 2004), ideally an eddy-resolving or at least eddy-permitting model resolution should be used and its output appropriately averaged and sampled to represent satellite-observations. We therefore consider the presented results as a plausibility demonstration that the considered proxy variables in the real ocean contain the necessary information for skillful global MOC reconstruction, while the potential to (machine- or otherwise) learn the specific real-world reconstruction operator based on satellite observables remains to be investigated.

In addition to spatial resolution, a second inherent model limitation is the use of the Boussinesq approximation, which conserves fluid volume instead of mass (Greatbatch, 1994; Griffies & Greatbatch, 2012), and hence may cause errors in deep ocean pressure (Ponte, 1999) and, by geostrophy, in deep circulation as well. Other sources of model error include the hydrostatic assumption, atmospheric forcing errors, and parametric errors in the formulation of the equation of state and of dissipative operators. Each of these terms can result, within numerical model tests, in SSH and OBP errors of similar magnitude to the Boussinesq approximation (Greatbatch et al., 2001; Losch et al., 2004). The biases introduced by these approximations in MOC and in its reconstruction from satellite observable variables remain to be investigated in future work.

Another point of difficulty in a hypothetical real-world application of the present method is partial knowledge of the proxy variables. Partial knowledge occurs due to measurement and processing errors, and due to sparser or coarser spatial sampling than the model in some cases (e.g., ∼300 km in GRACE). While these issues are

not addressed in full here, it is encouraging that coarse-grained model data was used successfully here: (a) we pre-averaged NN input data over several degrees latitude N_{dl} before training, effectively mimicking lower effective resolution of satellite observations. In cases that were tested (including AMOC at 26° and SOMOC at 60°), the results were not sensitive to the value of N_{dl} between 3 and 7° latitude. (b) Training NNs based on (more sparse) subsets of the full proxy variables did not seriously deteriorate reconstruction skills (Section 3.4). (c) We also tested zonally smoothing the covariates. A subset of the results tested with a range of different (Gaussian) smoothing radius choices (1, 2, 3, 4, 5° longitude) were not sensitive to the actual smoothing radius value (not shown). (d) For AMOC and SOMOC we found that a combination of SSH and ZWS can lead to high skill, although generally lower than OBP. It is therefore possible that any measurement noise or error (which we did not consider here) in, for example, OBP, can be partially compensated for by inclusion of SSH and ZWS as joint covariates.

We note that the present approach, if successfully implemented based on "real world" input, may be considered as a complementary approach to "state estimates" of ocean circulation (Errico, 1997) such as ECCO (Forget et al., 2015), but focused on the MOC specifically, and using a different model type (NN) and optimization procedure. While state estimates have several advantages including an exact physical basis for their forward model, the NN approach is computationally much lighter (especially after training) and optimizes the model specifically for the quantity of interest (MOC).

5.4. Reconstruction of Secular Variations

The holy grail in MOC reconstruction is arguably to be able to reconstruct not only variability around a long time mean, but also temporal trends or secular variations. Here we only dealt with the former, and in fact detrended the used ECCO variables. Recently Worthington et al. (2021) have trained an empirical model of AMOC strength at 26°N based on western and eastern density values at several depths, and used it to estimate MOC values in preceding decades. Due to their dynamical significance, it is possible that boundary density profiles can generalize well to conditions outside of the training set, and may be used to augment the present approach as boundary density values are more readily monitored in situ than basin interior values. Alternatively, it may be desirable to use measurements from the fleet of profiling autonomous Argo floats, at least for the mid-depth MOC cells (the majority of Argo floats generally profile down to 2 km depth, although deeper Argo floats are deployed in greater numbers in recent years). The use of observations at time-varying locations would require substantial changes to the present ML architecture, but arguably for inter-annual and longer time scales it may be sufficient to use smoothed Eulerian annual temperature and salinity distributions from the float observations. In addition, we suggest transfer learning (Weiss et al., 2016) may be used to train a NN over multiple climate states within one or several numerical models. Training over multiple models, each suffering from different biases, can also be hoped to improve the fidelity of the model(s) trained NN when applied to real ocean proxy data.

To address decadal and longer variability time scales, the NN should not only be based on a longer training period, but should also likely allow for delayed and non-local (in latitude) connections between observables and MOC. Pillar et al. (2016) has shown that interannual AMOC anomalies at 26°N are strongly affected by the surface heat flux distribution over the whole North Atlantic as far back as 15 years prior to the anomaly year. AMOC at 34.5°S has been shown (Smith & Heimbach, 2019) to be significantly affected by surface heat and momentum fluxes even outside of and remote from the Atlantic Ocean for interannual time scales. We expect the same may be true for the IPMOC. Thus, while time-delays were of no consequence in the present framework (Section 3.3), they can potentially increase reconstruction skill if combined with non-local affects.

6. Summary and Conclusions

Jackson et al. (2022), in a review of AMOC observations and variability, have stressed the need for employing a variety of sustainable long-term monitoring capabilities. That need is even greater in the Southern, Pacific, and Indian Oceans, where no continuous monitoring systems are or have been in operation. A variety of MOC observing system based on different methodologies is advisable, since all present methods have different associated uncertainties and biases. For example, significant uncertainties were recently brought to light regarding in-situ monitoring arrays' reference level assumptions (Frajka-Williams et al., 2018). The need for cost-effective methods is exacerbated by potentially long monitoring times which may be required to identify secular MOC

changes. For example, Baehr et al. (2008) have shown, using a numerical model test, that detecting a ∼50% secular decrease in 26° AMOC (for IPCC A1B emissions scenario) requires 60–90 years of monitoring array data.

Inference of the MOC based on satellite-observed variables is an attractive alternative to in-situ monitoring array, due to the continuous coverage in space and time and since satellite sensing is expected to continue to be funded irrespective of MOC programs. Several decades of satellite observations records already exist and may be utilized for historical reconstruction as well (Frajka-Williams et al., 2019). Previous studies have each examined MOC reconstruction skill in specific regions, and based on specific satellite data products (Section 1). Here we examine near-global MOC reconstruction skill from satellite-observable variables within a numerical model (the ECCO state estimate). We do not limit ourselves to one variable, and examine which combinations are optimal. We apply a data-based reconstruction method, that is, we use NNs to find the "best" empirical models.

We find that MOC strength time series are reconstructed with high skill in the present testing methodology for the four considered branches: the abyssal and mid-depth MOC cells in the Southern Ocean (SOMOC), the abyssal MOC Indo-Pacific cell (IPMOC), and the mid-depth MOC cell in the Atlantic (AMOC). The coefficient of determination (R^2), roughly interpretable as the fraction of time series variance accounted for by the reconstruction, is higher than 0.8 in most latitudes for all four considred MOC cells. The reconstruction is particularly successful for the SOMOC and (aside from the equatorial region) for the AMOC, where R^2 values around 0.95 are achieved based on OBP for abyssal SOMOC, and based on OBP in conjunction with ZWS for AMOC and for the mid-depth SOMOC cell. Overall we find that OBP generally has the highest reconstruction skill potential, followed by ZWS, and find in which regions their combination adds skill. SSH and ZWS are found to be a (slightly less) skillful pair as well for AMOC, particularly in the subpolar North Atlantic, and for the SOMOC abyssal cell. The identification of the optimal observables in the present context may help guide future studies on MOC reconstruction.

To provide context for the modes of variability captured by the reconstruction, we have also examined the reconstruction skill for (filtered) longer periods of variability (Sections 3.1 and 4). For all of the MOC basins and cells considered, reconstruction skill and squared coherence decreased very little if at all with increase in variability time scale up to the longest periods tested (≈2 years). This indicates that the pertinent dynamics are only weakly time-scale dependent (at the range of scales considered). Consistently, we find that several years of training data already produces moderate reconstruction skill, and that skill is near-maximized after about a decade of training (Section 3.2). The required training period in this test is thus shorter than the present record length of real satellite observations of OBP, SSH, and ZWS, or of the RAPID-MOCHA array time series.

To quantify the relative importance of different regions to the MOC reconstruction we tested zonal subsampling of the input fields (Section 3.4). For the AMOC at 26°N, we found similar results to those of Bingham and Hughes (2008) at 42°N, that is, the western boundary OBP time series on its own leads to almost the same skill as the entire zonal distribution. In the Southern Ocean, reconstruction with as little as five zonal points can lead to almost no loss in skill relative to using the full zonal structure (correlation of 0.9 vs. 0.99), depending on which zonal points are retained. The importance of the retained points is explored via ML interpretability techniques (Section 2.5). We find that the prominent ridge systems, for example, the Scotia Arc and Pacific Antarctic Ridge, play an out-sized role in determining the output of the high-skill NNs. This is consistent with a dynamic balance between bottom form stress anomalies and meridional transport anomalies (Stewart & Hogg, 2017), and hence supports further leveraging of this dynamic understanding in SOMOC theory, observation, and monitoring.

The present work demonstrates, within a numerical model, a high potential for skillful data-driven global MOC reconstruction based on satellite-observable variables. Extension of the present test to decadal and longer variabilty time scales, where surface heat and freshwater fluxes can be expected to be most important, is desirable and would require a longer time series than presently available from ECCO4 (24 years). Such an extension would likely require using non-local and time-delayed covariates information (Jamet et al., 2020; Larson et al., 2020; Pillar et al., 2016; Smith & Heimbach, 2019).

To apply model-learned relations to real observations would require additional considerations. Testing the degree of generalizability of the trained NN across numerical models can be a step forward in application to the real ocean. It is likely that training using an eddy-resolving model would be required to better represent real satellite observations. We suggest that training the NN across different numerical models via transfer learning may

produce a NN suitable for real ocean MOC reconstruction, as the learned features would ideally be the basic physics (primitive equations etc) rather than model-specific artifacts.

Appendix A: Basin Masks

For defining the MOC in each basin we integrated transport (Equation 1) between zonal limits dependent on the basin and latitude. The definitions of each basin are shown in Figure A1b. Atlantic or Indo-Pacific variable masks are defined between 34°S and 55°N. The Atlantic mask includes the entire Atlantic Ocean between Africa or Europe and the Americas. The Mediterranean is excluded from the Atlantic Sea mask, which amounts to "blocking" the one grid point wide Straits of Gibraltar. The Indo-Pacific Oceans mask is defined as all the ocean points in the 34°S−55°N latitude range, outside of the Atlantic Ocean and the Mediterranean Sea. The covariates used in reconstructing each MOC series were distributed between the same zonal limits (Section 2.3). Panel a shows the ECCO bathymetry over the same basins.

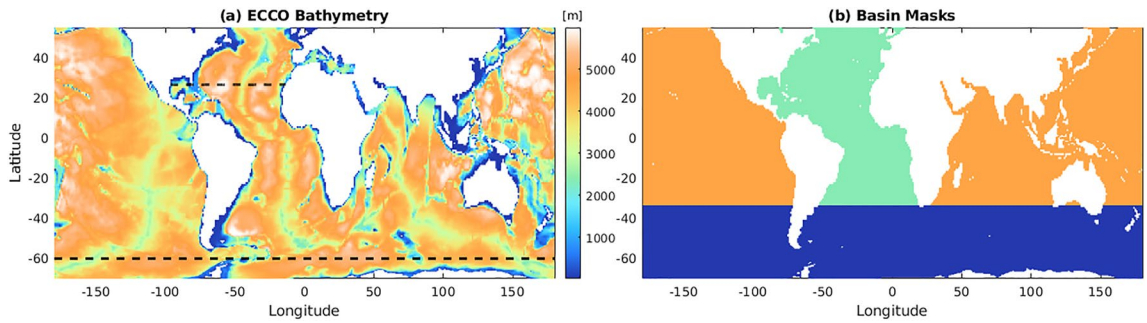

Figure A1. (a) ECCO bathymetry for the analyzed regions. Two sections which are analyzed in detail in Section 3 are shown in dashed lines. (b) Basin Masks for the input data to the neural networks. The Atlantic mask is in green, the Indo-Pacific mask is in orange, and the Southern Ocean mask is in blue. Land mass is in white. The Mediterranean Sea, although included in ECCO, is painted white here since it is not included the Atlantic mask, that is, our Atlantic mask ends outside of the Mediterranean Sea.

Data Availability Statement

The ECCOV4r3 state estimate data used in this study is publicly available from the ECCO website (Fukumori et al., 2017). The MATLAB code that was used to train and analyze the neural networks in this study is freely available for download at the Zenodo website (https://zenodo.org/) under the following https://doi.org/10.5281/zenodo.7016247 (Solodoch & Stewart, 2022). The MATLAB scripts additionally make use of several freely available MATLAB toolboxes: gcmfaces toolboxes (Forget, 2018), the Climate Data Toolbox (Greene et al., 2019), and the cmocean colormaps package (Thyng et al., 2016).

Acknowledgments
The authors thank Graeme MacGilchrist and two anonymous reviewers for very helpful feedback which greatly imporved this manuscript. The authors also thank Jemma Jeffree for very helpful discussions on MOC machine learning. AS and ALS were partially supported by the National Aeronautics and Space Administration ROSES Physical Oceanography program under Grant 80NSSC19K1192, and by the National Science Foundation (NSF) under Grant OPP-2023244. GEM was supported by the NASA OSTST Grant 20-OSTST20-0004.

References

Adcroft, A., Campin, J.-M., Hill, C., & Marshall, J. (2004). Implementation of an atmosphere–ocean general circulation model on the expanded spherical cube. *Monthly Weather Review*, *132*(12), 2845–2863. https://doi.org/10.1175/mwr2823.1

Alexander-Turner, R., Ortega, P., & Robson, J. (2018). How robust are the surface temperature fingerprints of the Atlantic Overturning Meridional Circulation on monthly time scales? *Geophysical Research Letters*, *45*(8), 3559–3567. https://doi.org/10.1029/2017gl076759

Ansorge, I. J., Baringer, M., Campos, E. J., Dong, S., Fine, R., Garzoli, S. L., et al. (2014). Basin-wide oceanographic array bridges the south Atlantic. *Eos, Transactions American Geophysical Union*, *95*(6), 53–54. https://doi.org/10.1002/2014eo060001

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Baehr, J., Keller, K., & Marotzke, J. (2008). Detecting potential changes in the meridional overturning circulation at 26 N in the Atlantic. *Climatic Change*, *91*(1), 11–27. https://doi.org/10.1007/s10584-006-9153-z

Bingham, R. J., & Hughes, C. (2008). Determining north Atlantic meridional transport variability from pressure on the western boundary: A model investigation. *Journal of Geophysical Research*, *113*(C9), C09008. https://doi.org/10.1029/2007jc004679

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cunningham, S. A., Kanzow, T., Rayner, D., Baringer, M. O., Johns, W. E., Marotzke, J., et al. (2007). Temporal variability of the Atlantic meridional overturning circulation at 26.5 N. *Science*, *317*(5840), 935–938. https://doi.org/10.1126/science.1141304

Dengler, M., Schott, F., Eden, C., Brandt, P., Fischer, J., & Zantopp, R. J. (2004). Break-up of the Atlantic deep western boundary current into eddies at 8 S. *Nature*, *432*(7020), 1018–1020. https://doi.org/10.1038/nature03134

Duchez, A., Courtois, P., Harris, E., Josey, S. A., Kanzow, T., Marsh, R., et al. (2016). Potential for seasonal prediction of Atlantic sea surface temperatures using the RAPID array at 26N. *Climate Dynamics*, *46*(9), 3351–3370. https://doi.org/10.1007/s00382-015-2918-1

Errico, R. M. (1997). What is an adjoint model? *Bulletin of the American Meteorological Society*, *78*(11), 2577–2592. https://doi.org/10.1175/1520-0477(1997)078<2577:wiaam>2.0.co;2

Foresee, F. D., & Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. *Proceedings of international conference on neural networks*, *3*(icnn'97), 1930–1935.

Forget, G. (2018). gcmfaces. [Software]. Zenodo. https://doi.org/10.5281/zenodo.3606908

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

Frajka-Williams, E. (2015). Estimating the Atlantic overturning at 26 n using satellite altimetry and cable measurements. *Geophysical Research Letters*, *42*(9), 3458–3464. https://doi.org/10.1002/2015gl063220

Frajka-Williams, E., Ansorge, I. J., Baehr, J., Bryden, H. L., Chidichimo, M. P., Cunningham, S. A., et al. (2019). Atlantic meridional overturning circulation: Observed transport and variability. *Frontiers in Marine Science*, *260*. https://doi.org/10.3389/fmars.2019.00260

Frajka-Williams, E., Lankhorst, M., Koelling, J., & Send, U. (2018). Coherent circulation changes in the deep north Atlantic from 16 n and 26 n transport arrays. *Journal of Geophysical Research: Oceans*, *123*(5), 3427–3443. https://doi.org/10.1029/2018jc013949

Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., & Ponte, R. M. (2017). ECCO version 4 release 3. [Dataset]. ECCO. https://ecco-group.org/

Gent, P. R., & Mcwilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, *20*(1), 150–155. https://doi.org/10.1175/1520-0485(1990)020<0150:imiocm>2.0.co;2

Greatbatch, R. J. (1994). A note on the representation of steric sea level in models that conserve volume rather than mass. *Journal of Geophysical Research*, *99*(C6), 12767–12771. https://doi.org/10.1029/94jc00847

Greatbatch, R. J., Lu, Y., & Cai, Y. (2001). Relaxing the Boussinesq approximation in ocean circulation models. *Journal of Atmospheric and Oceanic Technology*, *18*(11), 1911–1923. https://doi.org/10.1175/1520-0426(2001)018<1911:rtbaio>2.0.co;2

Greene, C. A., Thirumalai, K., Kearney, K. A., Delgado, J. M., Schwanghart, W., Wolfenbarger, N. S., et al. (2019). The climate data toolbox for MATLAB. *Geochemistry, Geophysics, Geosystems*, *20*(7), 3774–3781. https://doi.org/10.1029/2019GC008392

Griffies, S. M., & Greatbatch, R. J. (2012). Physical processes that impact the evolution of global mean sea level in ocean climate models. *Ocean Modelling*, *51*, 37–72. https://doi.org/10.1016/j.ocemod.2012.04.003

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, *5*(6), 989–993. https://doi.org/10.1109/72.329697

Han, L. (2021). The sloshing and diapycnal meridional overturning circulations in the Indian Ocean. *Journal of Physical Oceanography*, *51*(3), 701–725. https://doi.org/10.1175/jpo-d-20-0211.1

Isachsen, P. E., LaCasce, J., Mauritzen, C., & Häkkinen, S. (2003). Wind-driven variability of the large-scale recirculating flow in the Nordic seas and Arctic ocean. *Journal of Physical Oceanography*, *33*(12), 2534–2550. https://doi.org/10.1175/1520-0485(2003)033<2534:wvotlr>2.0.co;2

Jackson, L. C., Biastoch, A., Buckley, M. W., Desbruyères, D. G., Frajka-Williams, E., Moat, B., & Robson, J. (2022). The evolution of the north Atlantic meridional overturning circulation since 1980. *Nature Reviews Earth & Environment*, *3*(4), 1–14. https://doi.org/10.1038/s43017-022-00263-2

Jamet, Q., Dewar, W. K., Wienders, N., Deremble, B., Close, S., & Penduff, T. (2020). Locally and remotely forced subtropical AMOC variability: A matter of time scales. *Journal of Climate*, *33*(12), 5155–5172. https://doi.org/10.1175/jcli-d-19-0844.1

Kanzow, T., Cunningham, S. A., Rayner, D., Hirschi, J. J.-M., Johns, W. E., Baringer, M. O., et al. (2007). Observed flow compensation associated with the MOC at 26.5 N in the Atlantic. *Science*, *317*(5840), 938–941. https://doi.org/10.1126/science.1141293

Kanzow, T., Hirschi, J. J.-M., Meinen, C., Rayner, D., Cunningham, S. A., Marotzke, J., et al. (2008). A prototype system for observing the Atlantic meridional overturning circulation–scientific basis, measurement and risk mitigation strategies, and first results. *Journal of Operational Oceanography*, *1*(1), 19–28. https://doi.org/10.1080/1755876x.2008.11020092

Knight, J. R., Allan, R. J., Folland, C. K., Vellinga, M., & Mann, M. E. (2005). A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophysical Research Letters*, *32*(20), L20708. https://doi.org/10.1029/2005gl024233

Kostov, Y., Johnson, H. L., Marshall, D. P., Heimbach, P., Forget, G., Holliday, N. P., et al. (2021). Distinct sources of interannual subtropical and subpolar Atlantic overturning variability. *Nature Geoscience*, *14*(7), 491–495. https://doi.org/10.1038/s41561-021-00759-4

Kurtz, N. T., & Markus, T. (2012). Satellite observations of Antarctic sea ice thickness and volume. *Journal of Geophysical Research*, *117*(C8), C08025. https://doi.org/10.1029/2012jc008141

Kwok, R. (2010). Satellite remote sensing of sea-ice thickness and kinematics: A review. *Journal of Glaciology*, *56*(200), 1129–1140. https://doi.org/10.3189/002214311796406167

Landerer, F. W., Wiese, D. N., Bentel, K., Boening, C., & Watkins, M. M. (2015). North Atlantic meridional overturning circulation variations from Grace Ocean bottom pressure anomalies. *Geophysical Research Letters*, *42*(19), 8114–8121. https://doi.org/10.1002/2015gl065730

Larson, S. M., Buckley, M. W., & Clement, A. C. (2020). Extracting the buoyancy-driven Atlantic meridional overturning circulation. *Journal of Climate*, *33*(11), 4697–4714. https://doi.org/10.1175/jcli-d-19-0590.1

Leroux, S., Penduff, T., Bessières, L., Molines, J.-M., Brankart, J.-M., Sérazin, G., et al. (2018). Intrinsic and atmospherically forced variability of the AMOC: Insights from a large-ensemble ocean hindcast. *Journal of Climate*, *31*(3), 1183–1203. https://doi.org/10.1175/jcli-d-17-0168.1

Li, F., Lozier, M. S., Bacon, S., Bower, A., Cunningham, S., de Jong, M., et al. (2021). Subpolar North Atlantic western boundary density anomalies and the meridional overturning circulation. *Nature Communications*, *12*(1), 1–9. https://doi.org/10.1038/s41467-021-23350-2

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Losch, M., Adcroft, A., & Campin, J.-M. (2004). How sensitive are coarse general circulation models to fundamental approximations in the equations of motion? *Journal of Physical Oceanography*, *34*(1), 306–319. https://doi.org/10.1175/1520-0485(2004)034<0306:hsacgc>2.0.co;2

Losch, M., Menemenlis, D., Campin, J.-M., Heimbach, P., & Hill, C. (2010). On the formulation of sea-ice models. Part 1: Effects of different solver implementations and parameterizations. *Ocean Modelling*, *33*(1–2), 129–144. https://doi.org/10.1016/j.ocemod.2009.12.008

Lozier, M. S., Bacon, S., Bower, A. S., Cunningham, S. A., De Jong, M. F., De Steur, L., et al. (2017). Overturning in the subpolar North Atlantic program: A new international ocean observing system. *Bulletin of the American Meteorological Society*, *98*(4), 737–752. https://doi.org/10.1175/bams-d-16-0057.1

Lozier, M. S., Li, F., Bacon, S., Bahr, F., Bower, A. S., Cunningham, S., et al. (2019). A sea change in our view of overturning in the subpolar North Atlantic. *Science*, *363*(6426), 516–521. https://doi.org/10.1126/science.aau6592

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447. https://doi.org/10.1162/neco.1992.4.3.415
Marshall, J., Adcroft, A., Hill, C., Perelman, L., & Heisey, C. (1997). A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *Journal of Geophysical Research*, *102*(C3), 5753–5766. https://doi.org/10.1029/96jc02775
Mazloff, M. R., & Boening, C. (2016). Rapid variability of Antarctic bottom water transport into the Pacific ocean inferred from GRACE. *Geophysical Research Letters*, *43*(8), 3822–3829. https://doi.org/10.1002/2016gl068474
Mazloff, M. R., Heimbach, P., & Wunsch, C. (2010). An eddy-permitting Southern Ocean state estimate. *Journal of Physical Oceanography*, *40*(5), 880–899. https://doi.org/10.1175/2009jpo4236.1
McCarthy, G. D., Brown, P. J., Flagg, C. N., Goni, G., Houpert, L., Hughes, C. W., et al. (2020). Sustainable observations of the AMOC: Methodology and technology. *Reviews of Geophysics*, *58*(1), e2019RG000654. https://doi.org/10.1029/2019rg000654
McCarthy, G. D., Smeed, D. A., Johns, W. E., Frajka-Williams, E., Moat, B. I., Rayner, D., et al. (2015). Measuring the Atlantic meridional overturning circulation at 26 N. *Progress in Oceanography*, *130*, 91–111. https://doi.org/10.1016/j.pocean.2014.10.006
McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1
Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008
Nøst, O. A., & Isachsen, P. E. (2003). The large-scale time-mean ocean circulation in the Nordic Seas and Arctic Ocean estimated from simplified dynamics. *Journal of Marine Research*, *61*(2), 175–210. https://doi.org/10.1357/002224003322005069
Percival, D. B., & Walden, A. T. (1993). *Spectral analysis for physical applications*. Cambridge University Press.
Petty, A. A., Kurtz, N. T., Kwok, R., Markus, T., & Neumann, T. A. (2020). Winter Arctic sea ice thickness from ICESat-2 freeboards. *Journal of Geophysical Research: Oceans*, *125*(5), e2019JC015764. https://doi.org/10.1029/2019jc015764
Pillar, H. R., Heimbach, P., Johnson, H. L., & Marshall, D. P. (2016). Dynamical attribution of recent variability in Atlantic overturning. *Journal of Climate*, *29*(9), 3339–3352. https://doi.org/10.1175/jcli-d-15-0727.1
Ponte, R. M. (1999). A preliminary model study of the large-scale seasonal cycle in bottom pressure over the global ocean. *Journal of Geophysical Research*, *104*(C1), 1289–1300. https://doi.org/10.1029/1998jc900028
Pujol, M.-I., Faugère, Y., Taburet, G., Dupuy, S., Pelloquin, C., Ablain, M., & Picot, N. (2016). DUACS DT2014: The new multi-mission altimeter data set reprocessed over 20 years. *Ocean Science*, *12*(5), 1067–1090. https://doi.org/10.5194/os-12-1067-2016
Rahmstorf, S., Box, J. E., Feulner, G., Mann, M. E., Robinson, A., Rutherford, S., & Schaffernicht, E. J. (2015). Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nature Climate Change*, *5*(5), 475–480. https://doi.org/10.1038/nclimate2554
Salmon, R. (1998). *Lectures on geophysical fluid dynamics*. Oxford University Press.
Sanchez-Franks, A., Frajka-Williams, E., Moat, B. I., & Smeed, D. A. (2021). A dynamically based method for estimating the Atlantic meridional overturning circulation at 26°N from satellite altimetry. *Ocean Science*, *17*(5), 1321–1340. https://doi.org/10.5194/os-17-1321-2021
Smith, T., & Heimbach, P. (2019). Atmospheric origins of variability in the South Atlantic meridional overturning circulation. *Journal of Climate*, *32*(5), 1483–1500. https://doi.org/10.1175/jcli-d-18-0311.1
Solodoch, A., & Stewart, A. L. (2022). Machine learning reconstruction of the meridional overturning circulation in the ECCOv4r3 Ocean State estimate. [Software]. Zenodo. https://doi.org/10.5281/zenodo.7016247
Stewart, A. L., Chi, X., Solodoch, A., & Hogg, A. M. (2021). High-frequency fluctuations in Antarctic bottom water transport driven by Southern Ocean winds. *Geophysical Research Letters*, *48*(17), e2021GL094569. https://doi.org/10.1029/2021gl094569
Stewart, A. L., & Hogg, A. M. (2017). Reshaping the Antarctic circumpolar current via Antarctic bottom water export. *Journal of Physical Oceanography*, *47*(10), 2577–2601. https://doi.org/10.1175/jpo-d-17-0007.1
Talley, L. D. (2013). Closure of the global overturning circulation through the Indian, Pacific, and Southern Oceans: Schematics and transports. *Oceanography*, *26*(1), 80–97. https://doi.org/10.5670/oceanog.2013.07
Tandon, N. F., Saenko, O. A., Cane, M. A., & Kushner, P. J. (2020). Interannual variability of the global meridional overturning circulation dominated by Pacific variability. *Journal of Physical Oceanography*, *50*(3), 559–574. https://doi.org/10.1175/jpo-d-19-0129.1
Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). Grace measurements of mass variability in the Earth system. *Science*, *305*(5683), 503–505. https://doi.org/10.1126/science.1099192
Thompson, R. O. (1979). Coherence significance levels. *Journal of the Atmospheric Sciences*, *36*(10), 2020–2021. https://doi.org/10.1175/1520-0469(1979)036<2020:csl>2.0.co;2
Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M., & DiMarco, S. F. (2016). True colors of oceanography: Guidelines for effective and accurate colormap selection. *Oceanography*, *29*(3), 9–13. https://doi.org/10.5670/oceanog.2016.66
Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019ms002002
Wang, O., Fukumori, I., & Fenty, I. (2017). An overview of ECCO version 4 (release 3). Retrieved from https://ecco.jpl.nasa.gov/drive/files/Version4/Release3/doc/v4r3_overview.pdf
Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 1–40. https://doi.org/10.1186/s40537-016-0043-6
Willis, J. K. (2010). Can in situ floats and satellite altimeters detect long-term changes in Atlantic Ocean overturning? *Geophysical Research Letters*, *37*(6), L06602. https://doi.org/10.1029/2010gl042372
Worthington, E. L., Moat, B. I., Smeed, D. A., Mecking, J. V., Marsh, R., & McCarthy, G. D. (2021). A 30-year reconstruction of the Atlantic meridional overturning circulation shows no decline. *Ocean Science*, *17*(1), 285–299. https://doi.org/10.5194/os-17-285-2021
Wunsch, C. (1996). *The ocean circulation inverse problem*. Cambridge University Press.