UNIVERSITY OF CALIFORNIA

Santa Barbara

Assessing Historical Ocean Heat Content and Recent Global Mean Sea Level Rise Using Artificial Neural Networks

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Marine Science

by

Aaron C. Bagnell

Committee in charge:

Professor Timothy DeVries, Chair

Professor Qinghua Ding

Professor Paul Atzberger

September 2021

The dissertation of Aaron C. Bagnell is approved.

_____
Qinghua Ding

_____
Paul Atzberger

_____
Timothy DeVries, Committee Chair

August 2021

Assessing Historical Ocean Heat Content and Recent Global Mean Sea Level Rise Using Artificial Neural Networks

Aaron C. Bagnell

VITA OF AARON C. BAGNELL

August 2021

EDUCATION

Bachelor of Arts in Geophysics, University of California, Berkeley, May 2015

Doctor of Philosophy in Marine Science, University of California, Santa Barbara, August 2021

PROFESSIONAL EMPLOYMENT

2015-2016: Regents' Fellow, University of California, Santa Barbara

2016-2021: Graduate Researcher, Earth Research Institute, University of California, Santa Barbara

2017-2020: Teaching Assistant, Department of Geography, University of California, Santa Barbara

Summer 2017 and 2019: Summer Fellow, Earth Research Institute, University of California, Santa Barbara

PUBLICATIONS

"Correcting biases in historical bathythermograph data using artificial neural networks," *J. Atmospheric and Oceanic Technology* (2020).

"20th Century cooling of the deep ocean contributed to delayed acceleration of Earth's energy," *Nature Communications* (2021).

AWARDS

SAP + Esri Hackathon Finalist, Esri Annual Meeting, Palm Springs, 2018

Big Ocean Button Challenge Winner, XPRIZE Foundation and HeroX, 2018

FIELDS OF STUDY

Major Field: Ocean Data Science

Studies of Instrumental Biases in Ocean Temperature Measurements with Professor Timothy DeVries

Studies of Ocean Heat Content and Sea Level Rise with Professor Timothy DeVries

Studies of Stable Nitrate Isotopes with Drs. Patrick Rafter and Timothy DeVries

ABSTRACT

Assessing Historical Ocean Heat Content and Recent Global Mean Sea Level Rise Using Artificial Neural Networks

by

Aaron C. Bagnell

Global warming arises from an energy imbalance where increased radiative forcing from greenhouse gases traps additional heat in the Earth system. Over recent decades, more than 90% of this heat has gone into the ocean, causing thermal expansion that leads to sea level rise. Additionally, warming of the cryosphere has caused significant amounts of freshwater from melting ice to enter the ocean, representing another major contributor to rising sea levels. It is therefore essential to obtain accurate estimates of the historical changes in ocean heat content and sea level rise, as these are necessary to inform our understanding of future climate scenarios and some of the regional impacts from climate change. Improving these historical estimates, however, requires overcoming uncertainty related to instrumental and spatial sampling biases.

Prior to 2004 estimates of ocean heat content rely primarily on temperature measurements from mechanical and expendable bathythermograph (BT) instruments that were deployed on large scales by naval vessels and ships of opportunity. These BT temperature measurements are subject to well-documented biases, but even the best

calibration methods still exhibit residual biases when compared to high-quality temperature datasets. In Chapter I, we use a new approach to reduce biases in historical BT data. Our method consists of an ensemble of artificial neural networks that corrects biases with respect to depth, year, and water temperature in the top 10 meters. A global correction, as well as corrections optimized to specific BT probe types are presented for the top 1800 m. Even with comparable performances at reducing the instrumental biases, distinct patterns emerge across the calibration methods when they are extrapolated to BT data not included in our cross-instrument comparison. Multiple bias corrections should therefore be incorporated into studies of ocean heat content.

Sampling the deep ocean has remained a large technical challenge, leading to a sparse observational record below 2000 m. Due to methodological limitations at handling such sparse temperature data, many historical reconstructions of ocean heat content neglect this large volume of the ocean deeper than 2000 m. In Chapter II, we provide a global reconstruction of historical changes in full-depth ocean heat content based on interpolated subsurface temperature data using an autoregressive artificial neural network, providing estimates of total ocean warming for the period 1946-2019. We find that cooling of the deep ocean and a small heat gain in the upper ocean led to no robust trend in global ocean heat content from 1960-1990, implying a roughly balanced Earth energy budget within -0.16 to 0.06 W m^{-2} over most of the latter half of the 20th century. However, the past three decades have seen a rapid acceleration in ocean warming, with the entire ocean warming from top to bottom at a rate of 0.63±0.13 W m^{-2}. These results suggest a delayed onset of a positive Earth energy imbalance relative to previous estimates, although large uncertainties remain.

In Chapter III, we examine recent sea level rise due to thermal expansion of the oceans and the addition of freshwater from melting ice. Barystatic sea level rise caused by the addition of freshwater to the ocean from melting ice can in principle be recorded by a reduction in seawater salinity. However, instrumental biases and limited data coverage have thus far hindered efforts to infer barystatic sea level rise from ocean salinity measurements. Here, we develop an instrumental bias correction and adapt our autoregressive machine learning method to interpolate ocean salinity measurements to estimate salinity changes in the global ocean from 2001-2019. We find that the ocean mass rose by 13±3 Tt from 2001-2019, implying a barystatic sea level rise of 2.0±0.5 mm/yr. Combined with sea level rise of 1.3±0.1 mm/yr due to ocean thermal expansion, these results suggest that global mean sea level rose by 3.4±0.6 mm/yr from 2001-2019. These results provide an important validation of remote-sensing measurements of ocean mass changes and global sea level rise, as well as independent analyses of the global ice budget.

TABLE OF CONTENTS

I. Correcting Biases in Historical Bathythermograph Data

A. Introduction

The oceans play an enormous role in the Earth's energy budget, having gained roughly ten times as much heat over the past half century as all other parts of the Earth system combined (Church *et al.*, 2011). Reconstructing past changes in ocean heat content is therefore critical to understanding Earth's climate sensitivity and energy balance (Hansen *et al.*, 2005; Trenberth *et al.*, 2014, Meehl *et al.*, 2005). Reconstructions of ocean heat content (OHC) prior to roughly year 2005 are hampered by data sparsity and persistent instrumental biases. Recent studies have demonstrated that issues related to the choice of calibration applied to instrumental biases in the historical ocean temperature record contribute to significant uncertainty in OHC reconstructions (Lyman *et al.*, 2010; Boyer *et al.* 2016,

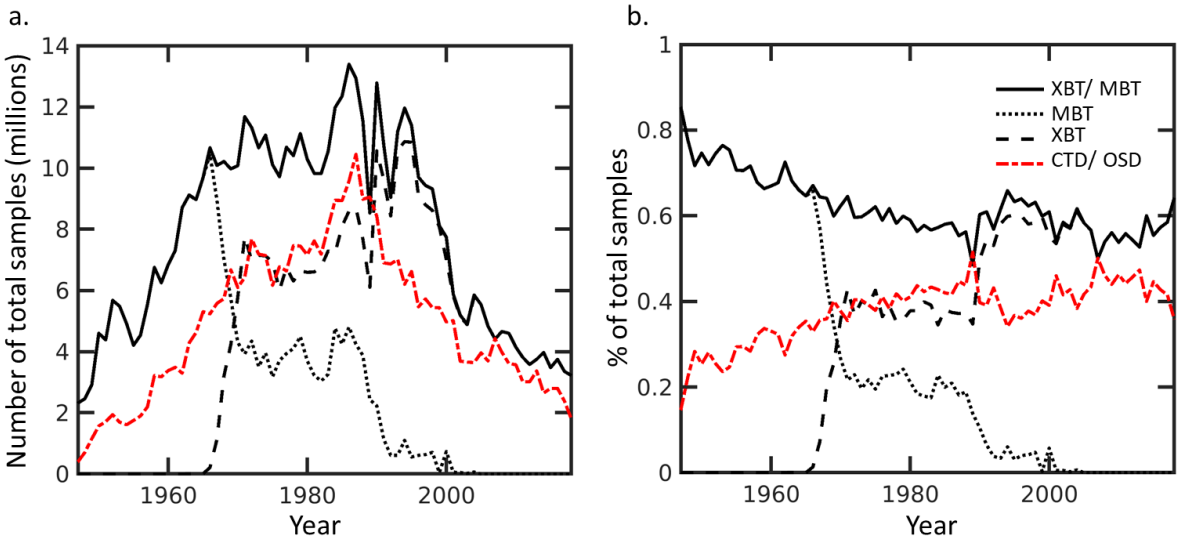


Fig. I.1 The relative annual sampling coverage of individual instrument casts for a. total discrete samples after linearly interpolating to 350 standard depths ($z_{standard}$) and b. percentage of total samples due to a particular temperature instrument.

Cheng *et al.*, 2016; Cheng *et al.*, 2018; Wang *et al.* 2018). It is crucial to correct these biases and further constrain the instrument calibration methods, as better constraints on past ocean temperature changes will impact projections of future warming.

The most widely-used instruments for measuring ocean temperature prior to the Argo era (2005-present, when autonomous profiling floats provide global data coverage) were the expendable bathythermograph (XBT) and its predecessor the mechanical bathythermograph (MBT). Excluding profiling floats, during the period 1945-present these instruments accounted for upwards of 60-70% of the raw temperature casts in a given calendar year (Fig. I.1). Additionally, because of their widespread use by the U.S. Navy and ships of opportunity, these instruments provide the community with vastly greater spatial sampling coverage (twice as much as other instruments on a 1-degree grid) in the historical ocean temperature record than would otherwise be afforded by scientific cruises alone. These instrument casts are therefore essential to numerous studies of historical ocean and climate trends, including assessments of ocean heat content (OHC) (Domingues *et al.*, 2008; Ishii and Kimoto, 2009; Levitus *et al.*, 2012; Cheng *et al.*, 2017).

The MBT probe was designed to reach nominal depths (60 m, 150 m, or 275 m depending on the model) and relied on being lowered on a winch at near free-fall speeds (Couper and LaFond, 1970). It contained instruments to measure temperature and pressure, which it continuously recorded on either a smoked or film-coated plate. This setup required that the probe be retrieved after every deployment, reducing the conditions under which it could be safely deployed. The XBT probe was designed to overcome this limitation by being expendable, with temperature recordings transmitted through a copper wire to a chart recorder on deck (Abraham *et al.*, 2013). After playing out the entire spool of wire the

connection severs, and the probe sinks to the bottom of the ocean. Several different probe types were designed over the long and continued period of use of the XBT, with most designs able to reach roughly 400-700 meters and a small subset able to reach depths up to 2000 m.

However, it is well understood that both XBT and MBT observations contain global systematic biases on the order of 0.1 ^{o}C, significantly impacting the estimation of OHC, which is highly sensitive to small temperature changes. Additionally, as first identified by Gouretski and Koltermann (2007), the biases of the XBT/ MBT instruments vary over time, leading to additional uncertainty about the rate of ocean warming on sub-decadal timescales. The community has therefore undertaken a concerted effort to examine and address the causes of these biases. However, uncertainties across an ensemble of prior bias corrections are of the same magnitude as natural intra-decadal variation (Lyman *et al.*, 2010; Cheng *et al.*, 2014), making these corrections one of the leading sources of uncertainty in estimates of OHC (Lyman *et al.*, 2010; Boyer *et al.*, 2016) and limiting the reliability of any single calibration method on shorter timescales.

Because the XBT instrument did not directly measure depth, but relied on a fall rate equation (FRE), this is an obvious source of systematic bias. After careful examination of a subset of probe deployments, Hanawa *et al.* (1995) proposed a modified FRE with new coefficients to correct for this depth bias. Against the warning of Hanawa *et al.* (1995) that this new FRE should not be introduced into original XBT data archives, the new equation was quickly adopted by the probe manufacturers, leading to XBT probes with a mixture of the old and new FRE being deployed in subsequent years (Abraham *et al.*, 2013). In addition, after applying the Hanawa *et al.* (1995) depth correction to historical XBT data,

Gouretski and Koltermann (2007) identified that a previously documented warm period in the XBT record could not be rectified with known climate patterns and was instead related to a residual time-variable bias in the XBT data itself, indicating that other sources of error exist. This finding triggered extensive interest in investigating the causes of this time-variable bias and generated numerous approaches to correct it (e.g. Wijffels *et al.*, 2008; Ishii and Kimoto, 2009; Gouretski and Reseghetti, 2010; Good, 2011; Gouretski, 2012; Hamon *et al.*, 2012; Cowley *et al.*, 2013, Cheng *et al.*, 2014).

Earlier studies corrected for the time-varying bias by using a FRE that varied based on the year of probe deployment (Wijffels *et al.*, 2008; Ishii and Kimoto, 2009). However, other studies indicated that the bias was also dependent on near-surface water temperature. Gouretski and Reseghetti (2010) found that the error could be further reduced by separating the bias into a depth dependent component and another they considered the "pure thermal" bias. Subsequently, Cowley *et al*. (2013) assembled thousands of side-by-side XBT and conductivity-temperature-depth (CTD) casts to look at cross-instrumental error, finding that there is a pure thermal bias that varies with both time and temperature but that is independent of depth, and also a depth error that varies not only with depth but also with time and perhaps with temperature.

The side-by-side comparisons of Cowley *et al*. (2013) relied on relatively ideal conditions where a CTD could be lowered from a stationary research vessel alongside an XBT probe deployment. This contrasts with normal operating conditions for the XBT, which was designed specifically so that it could be deployed from navy and merchant ships with minimal adjustments to the vessels' courses or speeds. Nevertheless, these findings were corroborated by Cheng *et al.* (2014) with a much larger global XBT dataset comprised

of probes deployed under more typical conditions. In addition, Cheng *et al.* (2014) sorted the XBT casts into groups of major probe types, confirming that differences in probe design not only led to different fall rates but also distinct time-varying bias histories.

The residual time-dependent biases that remain after the Hanawa *et al.* (1995) depth correction, or subsequent depth corrections that accounted for the effects of near-surface temperature on the probe fall rate (Cheng 2011) and the effect of probe design (Cheng *et al.*, 2014), are likely due to multiple sources of error such as variability in sea state, weather, and other deployment conditions, as well as technological developments that led to a shift in ship speeds, increasing deck heights, and a transition from analog to digital recorders. It has also been speculated that changes to probe manufacturing, such as the move by Sippican (one of two major XBT vendors) of operations from the USA to Mexico, could cause rapid shifts in the bias for certain XBT probe types (Wijffels *et al.*, 2008). These factors cannot be easily quantified for the global dataset without sufficient metadata. Unfortunately, in the global XBT dataset roughly half of the casts do not even contain the necessary metadata to assign them to a specific probe type. Because of this, developing a truly mechanistic model, one which accounts for errors using a purely physical explanation, remains difficult. Therefore, methods have been developed to account for only the most persistent biases, which evolve over multiyear time-scales.

Prior attempts to correct BT biases generally used one of two approaches. The first approach uses an empirical model, informed by the physics of the system, while also making necessary simplifying assumptions to estimate unknown model parameters. These approaches include methods that modify the original XBT probe FRE using time-variable parameters (Wijffels *et al.*, 2008; Ishii and Kimoto, 2009; Cowley *et al.*, 2013; Cheng *et al.*,

2014) instead of parameters that are constant with time (as with Hanawa *et al.*, 1995). The other approach uses statistical methods to remove biases, producing a single correction for the total bias, which is the approach taken by Levitus *et al.* (2009) and the current study. Arguments can be made for or against either approach, but both approaches can lead to significant reductions in the observed XBT/ MBT biases.

Here, we propose a new approach to correcting historical BT biases that sorts measurements from individual BT casts into categories based on the year of deployment, the temperature of the near-surface ocean, and the depth at which the measurement was taken, then uses an ensemble artificial neural network (EANN) to smooth and extrapolate the total BT bias to all times and locations for which BT data exist (Section B). This method is a statistical approach that does not attempt to separate components of the time-dependent bias, aside from considering the impact of different XBT probe types on the bias. While some of the underlying factors that contribute to the time-varying bias are likely independent of each other, our ability to disentangle these are limited by the completeness of available metadata. Therefore, for simplicity, our method considers the total bias to be an inseparable and nonlinear combination of the major sources of error identified in prior studies (Gouretski and Koltermann, 2007; Levitus *et al.*, 2009; Gouretski and Reseghetti, 2010; Cheng *et al.*, 2014). We apply our correction to both XBT and MBT datasets, with an option to either apply one global correction (what we call EANN-G) to each instrumental dataset or, for the XBT, to apply corrections to the individual probe types (EANN-P) (Section C). Using the metrics from Cheng *et al.* (2018) we demonstrate that we can favorably reduce the XBT/ MBT bias with respect to an independent validation dataset, with our calibration performing as well as or better than several widely-used existing methods (Section C). We close by

summarizing our main findings, and highlighting unresolved issues in correcting BT biases that should be addressed by the community (Section D).

B. Methods

B.1 Constructing a correction grid

We used individual casts of temperature data taken from the World Ocean Database (WOD) 2018 (Boyer *et al.*, 2018) (https://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html, accessed 9-30-19). These are provided as separate datasets for the different instrument types, which in our case are the two bathythermograph datasets (XBT and MBT) that will be corrected, and the Ocean Station (OSD) and Conductivity-Temperature-Depth (CTD) which we use as reference data (Fig. I.2, Step 1). Before quality control there are roughly 2.3 million XBT casts and 2.4 million MBT casts.

The XBT data we obtain is modified to have the original manufacturer FRE (MFR FRE hereafter) applied for the period after 1995. Without such a modification, there is a rapid transition in the bias simply due to instrument manufacturers adopting the Hanawa *et al.* (1995) FRE (H95 FRE hereafter) in the late 1990s (see Section A for further details). To modify individual casts from the 1990s onwards that have been flagged by the WOD as having the H95 FRE applied, we multiply their sample depths by the factor 0.9675, following Gouretski and Reseghetti (2010), which approximates the depths that would be given by using the original manufacturer FRE. Other studies such as Levitus *et al.* (2009) and Cheng *et al.* (2014) have instead opted to use the H95 FRE as a starting point, but some studies (Thadathil *et al.*, 2002; Gouretski and Resegheti, 2010) have indicated that the H95

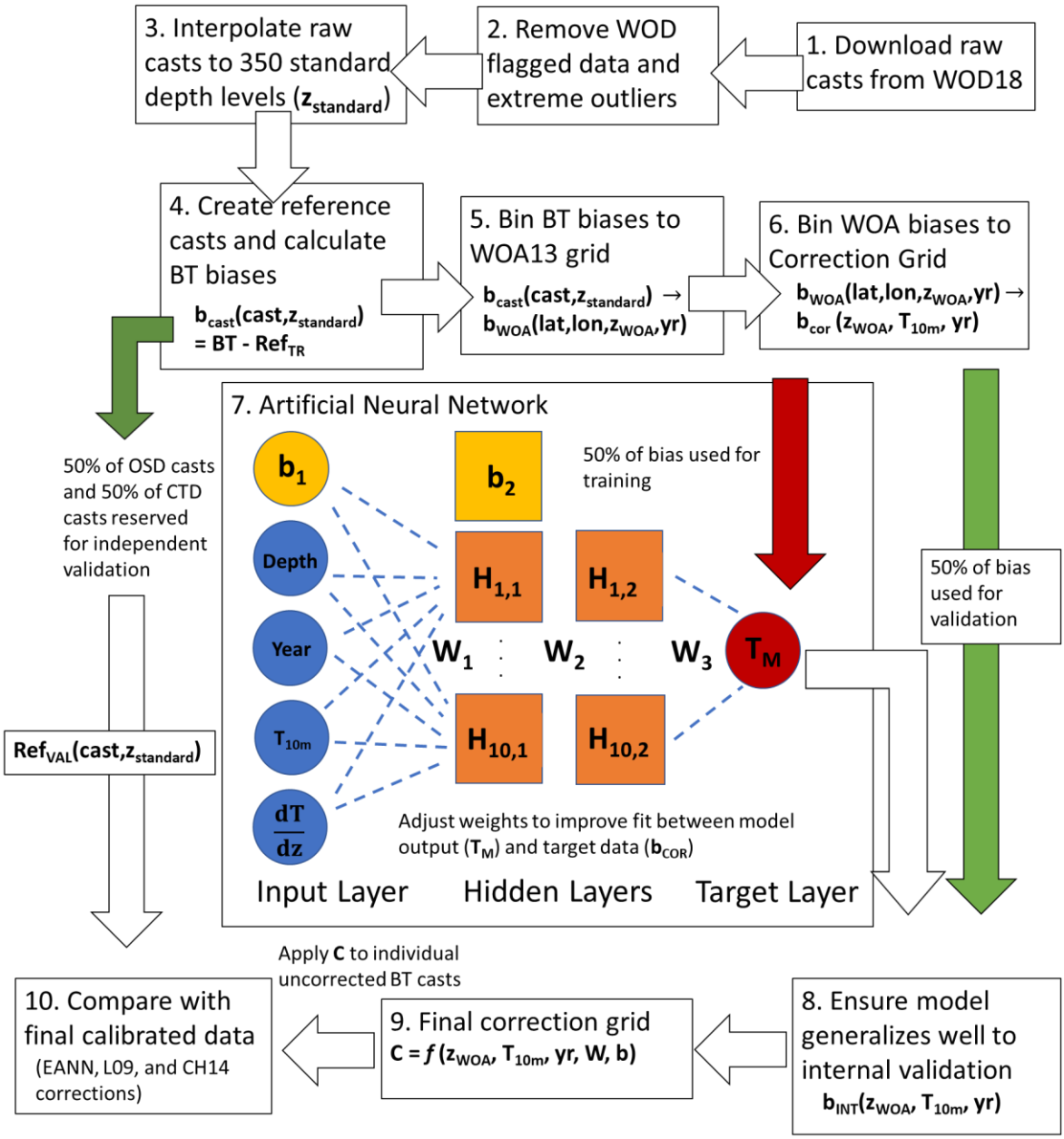


Fig. I.2 Schematic of this study's data flow and quality control, as well as the application of an ensemble of artificial neural networks (EANN) to calibrate bathythermograph data. Steps 6-9 are repeated 30 times to create an ensemble of bias corrections. See text for additional details on quality control, binning, and network architecture.

FRE may increase the time-dependent thermal bias relative to collocated CTD/ OSD data as compared to using the MFR FRE.

In addition, we exclude data in casts that have been flagged by the WOD for quality control issues with any flag other than zero, as well as unrealistic values that fall outside a -3 to 36 ^{o}C temperature range, as seawater does not normally fall outside this temperature range except for geographically isolated areas (e.g. hydrothermal vents) that are not relevant to our global study. Additionally, a cast must contain at least 4 temperature measurements as well as a measurement in the top 100 m in order to be included. After applying this quality control (Fig. I.2, Step 2), there are approximately 2.1 million XBT casts and 2.2 million MBT casts.

Following Cheng *et al*. (2014), each XBT probe is sorted into one of nine groups based on manufacturer, similarity in probe design, or terminal depth for probes of unknown type. These groups (approximate terminal depth in parentheses) are 1. T7/DB (760 m), 2. DX (760 m), 3. T4/T6 (450 m), 4. SX (450 m), 5. T10 (200 m), 6. T5 (1820 m), 7. TSK-T4/T6 (450 m), 8. TSK-T5 (1820 m), and 9. TSK-T7/DB (760 m). Probe groups 1, 3 5, and 6 are manufactured by Sippican, groups 7-9 are manufactured by TSK, and groups 2 and 4 are of unknown manufacturer/ type and are only differentiated by their greatest reported depth.

Cheng *et al*. (2014) noted that operators have generally considered XBT probes able to reach depths roughly 20% greater than their listed terminal depths with minimal loss in accuracy. However, due to the sparsity of data below these terminal depths and some apparently spurious observations, we chose to omit these data from the training of our model. In the end, these data will still receive a calibration.

Next, we interpolated individual casts for each instrument (XBT, MBT, OSD, and CTD) to the standard depth levels used by Cheng *et al*. (2018) (Fig. I.2 Step 3). These depth levels are spaced 1 m apart in the top 100 m, 5 m apart from 105-700 m, and 10 m apart from 710-2000 m, for a total of 350 depth levels. We then separately identified casts from the CTD and OSD datasets that were sampled within 1 degree of latitude and longitude and within the same 30-day window as a cast from either the XBT or MBT datasets (Fig. I.2 Step 4). Since multiple CTD and OSD casts could be associated with the same BT cast, the median of these "collocated casts" was computed separately for both the CTD and OSD datasets. Each respective BT cast could then have up to two reference casts. Keeping the two reference datasets separate, instead of taking the median of both CTD and OSD casts, was done so that our correction would not overly depend on a particular dataset, since the CTD and OSD data have different spatio-temporal sampling histories and vertical resolutions (Cheng and Zhu, 2014b).

After interpolation of the BT data to standard depth levels, we set aside 50% of the OSD and CTD data in order to use it for validating our calibration method. Due to the vastly different number of casts in each XBT probe category, removing a random 50% of data without consideration for probe type would leave some probe types with disproportionate amounts of training data versus validation data, so half of all reference CTD casts and half of all reference OSD casts are removed for each of the nine categories of XBT probes. The CTD/ OSD data that we set aside for validation is then concatenated and averaged. We refer to this dataset as $\mathrm{Ref_{VAL}}$ and use it only for validation of our calibration scheme, to ensure that our bias corrections extrapolate well to independent high-quality data (see Section C). One caveat to note is that some of the XBT/ MBT data were collected by ships of

opportunity from areas of the ocean far from any contemporaneous CTD/ OSD data, so the validation dataset Ref_{VAL} has a different spatio-temporal distribution than the full XBT dataset. This ultimately contributes additional uncertainty to the extrapolated bias corrections employed by various calibrations.

We use the remaining collocated casts Ref_{TR} to calculate the biases (b_{CAST}) in the individual XBT and MBT casts, defined as

$$b_{CAST}(\text{cast},z_{\text{standard}}) = \text{BT}(\text{cast},z_{\text{standard}}) - \text{Ref}_{\text{TR}}(\text{cast},z_{\text{standard}}). \qquad (1)$$

Because the cast level data is spatially biased towards regions, such as coasts, where significant repeat sampling occurred, we further bin the b_{CAST} data to a regular grid so that various geographic regions are more equally represented (Fig. I.2, Step 5). Our chosen grid, the World Ocean Atlas 2013 (WOA13) grid, has 1x1-degree resolution and 67 depth layers for 0-2000 m. When binning vertically, we use the depth layer whose value is closest to the observation's sampling depth (e.g. the first depth layer has a value of 0 m, the second of 5 m, and the third of 10 m, so all raw temperature values sampled between 0-2.5 m fall in the 0-m bin; between 2.5-7.5 m they fall in the 5-m bin). A point that lies exactly at the midpoint between depth intervals is binned to the shallower interval.

We opted to bin using the median of b_{CAST} instead of the mean, as it is more robust to noise caused by natural variability and instrumental errors. At sub-annual time-scales the time-varying biases appear to be dominated by changes in water temperature due to the seasonal cycle (Gouretski and Reseghetti, 2010) and not by other factors such as changes to probe design that occurred over multiple years (Cheng *et al*., 2014). Given this, we assume short-term changes to the temporal biases are purely a function of water temperature and are

not specific to a particular time or location. We can then bin b_{CAST} to the WOA13 grid annually based on the year of sampling, yielding b_{WOA}.

$b_{CAST}(\text{cast}, \text{z}_{\text{standard}})$ → median-binned to WOA13 grid → b_{WOA}(lon,lat,z_{WOA},yr). (2)

In the next step, biases on the WOA13 annual grid (b_{WOA}) are sorted into categories using the variables on which the XBT and MBT biases depend, namely year, depth, and temperature (Fig. I.2, Step 6). This forms the basis of our "correction grid". The dimensions of this grid are 52 years for the XBT and 65 years for the MBT data (at 1-year increments) by 67 depths (with depth increments coinciding with the WOA13) by 79 temperatures (from -3 to 36^{o} at 0.5 ^{o}C increments) for a total of 275,236 elements in the correction grid for the XBT and 344,045 elements for the MBT. For temperature binning, we use the 10-m temperature from the WOA climatology at the measurement location (lon, lat), since it can be used as a proxy for a spatial component of the bias, which varies with latitude and has been proposed to be dependent on near-surface ocean temperature (Kizu *et al.*, 2005a, Reverdin, 2009). Cheng *et al.* (2014) used the 0-100 m average temperature taken from the cast itself, but some of this near-surface data contains errors that cannot easily be accounted for by standard quality control procedures. Additionally, many casts do not contain measurements for the full 0-100 m depth interval, in which case a climatological value would need to be substituted.

Biases whose absolute values exceed 5 ^{o}C are omitted (following Cheng *et al.*, 2018), as these extremes are likely due to insufficient sampling coverage for that particular grid cell instead of a systematic bias. The binning of biases to the correction grid follows the same

median-binning procedure that we use to bin b_{CAST} to the WOA grid. After binning to the correction grid, there are 150 thousand bias data points for the XBT data, and 80 thousand for the MBT data. Step 6 (Fig. I.2) thus yields the value of the bathythermograph instrumental bias on the correction grid, which we refer to as b_{COR},

$$b_{WOA}(\mathrm{lon}, \mathrm{lat}, \mathrm{z_{WOA}}, \mathrm{yr}) \rightarrow \text{ median-binned to correction grid} \rightarrow b_{COR}(\mathrm{z_{WOA}}, \mathrm{T_{10m}}, \mathrm{yr}). \quad (3)$$

We produced separate correction grids for the XBT and MBT datasets. For the XBT data, we consider the years 1967-2018 in our correction, while for MBT we consider the years 1940-2004. While certain data exists in both datasets prior to these time intervals, they are insufficiently sampled, leading to large apparent biases with respect to the CTD/ OSD data, and therefore are not considered in this study. Nine additional correction grids are generated for probe-specific calibrations of the XBT data.

B.2 Creating an ensemble of artificial neural networks

The correction grids that result from the steps described above (Fig. I.2, Steps 1-6) are noisy and contain holes where there are no collocated data that satisfy the requirements of that grid cell. We employ an artificial neural network (ANN) to smooth out the result and fill in the gaps (Fig. I.2, Step 7), so the correction can be extrapolated to all of the data in the XBT and MBT instrumental datasets, including for testing on our internal validation sets (Fig. I.2, Step 8).

Our feedforward ANN is a machine learning approach that seeks to reduce cross-instrumental biases by minimizing the following "cost function"

$$\text{cost} = \sum_{z} \sum_{T_{10m}} \sum_{yr} (b_{COR}(z_{WOA}, T_{10m}, yr) - C(z_{WOA}, T_{10m}, yr))^2. \quad (4)$$

As such, the ANN is trained to produce a final correction (*C*) (Fig. I.2, Step 9) that replicates the bias (b_{COR}) as closely as possible, while extrapolating to areas without bias data.

Given that the fall rate of a probe may be partially dependent on water temperature (Thadathil *et al.*, 2002, Kizu *et al.*, 2005a, Reverdin, 2009, Cowley *et al.*, 2013, Cheng *et al.*, 2014), the vertical temperature structure likely has an impact on the resulting depth bias. For this reason, we also use the vertical temperature gradient derived from the annual WOA climatology as an additional input to our ANN (Fig. I.2, Step 7), which indirectly gives us spatial information about different water masses as well as their average vertical structure. Including the vertical temperature gradient improves the reconstructed bias in the shallow subsurface in our model.

We use a fully connected network that consists of two hidden layers, with 10 nodes each (Fig. I.2 Step 7). This architecture keeps the ratio of free parameters (151 total weights) versus training samples (~10,000-150,000 depending on probe type) well below 10%, thus reducing the chance of overfitting the training data. The value of each node is partially dependent on the transfer function used to propagate information from one layer to the next. Initially we opted for a network with only a single hidden layer and the hyperbolic tangent as the transfer function, but the use of this particular transfer function introduced artificial structure to the extrapolated correction that was inconsistent with the raw correction grid. Once we opted to go beyond a single hidden layer, the obvious candidate for the transfer

function was the rectified linear unit. This has become the default for deep networks because it has fewer problems with vanishing gradients (Glorot *et al.*, 2011).

For the back-propagation algorithm in our ANN, which iteratively updates the values of the weights, we chose the Levenberg-Marquardt algorithm (Marquardt, 1963) due to its improved performance at achieving a lower mean squared error between predicted and expected values for the targets, versus other common algorithms such as gradient descent (Hagan and Menhaj, 1994). There is a danger of over-fitting the model, which occurs when the neural network is over-trained on a dataset so that it cannot extrapolate well when presented with new data. This becomes more difficult to avoid with more nodes in a hidden layer or more layers in the network (Weigend *et al.*, 1990). To counteract this, we incorporate Bayesian regularization (MacKay, 1992; Foresee and Hagan, 1997) directly into the back-propagation algorithm to optimize the regularization procedure that prevents overfitting by penalizing large weights in the network.

Additionally, we employ early stopping (Prechelt, 1998) using our internal validation set, which stops training when performance extrapolating to the internal validation set begins to degrade.

As mentioned in Section B.1, we create an independent validation set by omitting half of the CTD/ OSD data before calculating the biases from the concatenated CTD/ OSD datasets that are then binned to the WOA and correction grids (Fig. I.2, Step 5-6). This independent validation dataset is not ever "seen" by the ANN or used to tune it in any way. We therefore also create an internal validation set (b_{INT}) by randomly dropping 50% of the data in the correction grid. The remaining 50% of the data is used for training. Utilizing random

validation sets helps ensure that the individual ANN generalizes well for the systematic biases that are independent of choice of reference dataset. For the ANN to be accepted, it must produce a correction (*C*) that reduces the sum of squared errors of the internal validation set (Fig. I.2, Step 8)

$$\sum_z \sum_{T_{10m}} \sum_{yr} (b_{INT} - C)^2 < \sum_z \sum_{T_{10m}} \sum_{yr} b_{INT}^2. \quad (5)$$

Steps 6-9 (Fig. I.2) are repeated until 30 validated ANNs are produced, and these 30 ANNs are combined to produce the EANN. About 10% of models fail to fulfill the validation criterion set by Eq. 5 and are discarded.

There are several advantages to using an ensemble of ANNs rather than a single ANN. Due to the random initialization of weights in the ANNs and differences in training sets across members, it is possible for many different networks to achieve similar performance on a validation set while extrapolating to areas with no data coverage differently. This randomization is a form of data subsampling similar to bootstrap aggregating (Breiman, 1996), which by averaging the solution across ensemble members affords better performance on the validation sets compared to an individual member. This ensemble averaging has been demonstrated to improve the robustness of the extrapolation in areas without data coverage (Hansen and Salamon, 1990; Lincoln and Skrzypek, 1990). The ensemble range also provides a measure of the uncertainty of our corrections.

Steps 6-9 are again repeated using data binned to correction grids for individual probe-specific bias corrections. The result is ten ensembles of corrections for the XBT and an additional one for the MBT. These ensembles consist of one global correction, which can be applied to all of an instrument's data, as well as corrections for the nine XBT probe types

that are applied individually to each category of probes. Data from depths greater than the terminal depths of each probe type (see Section B.1) are also corrected, as our EANN extrapolates the correction down to 2000 m.

Our correction grid organizes BT bias corrections into categories based on sample depth, year, and temperature in the top 10 m. Because our correction grid has the depth levels of the WOA13 grid, we linearly interpolate the correction grid to the 350 depth levels of our cast data. XBT/ MBT casts are corrected by identifying the grid cell in the interpolated correction grid that corresponds to each measurement in the XBT/MBT cast (based on standard depth level, year, and 10-m temperature), and applying the corresponding correction. The end result is 60 (30) different corrected datasets of individual XBT (MBT) casts, obtained by combining the two different correction schemes (global vs. probe type) with the 30 ensemble members of the EANN.

We compare the XBT/ MBT casts calibrated with the EANN ensemble to the independent validation set Ref_{VAL} (Fig. I.2, Step 10) as a final test of our method's ability to correct "never-before-seen" BT biases.

C. Results

In order to compare the performances of multiple existing XBT calibrations, Cheng *et al.* (2018) proposed a set of metrics that can be used to gauge the residual biases with regards to depth, probe type, year, and latitude, variables on which the bias has been demonstrated to depend (Gouretski and Koltermann, 2007; Levitus *et al.*, 2009; Gouretski and Reseghetti, 2010; Cheng *et al.*, 2014). We have used four of these metrics to assess our own method's

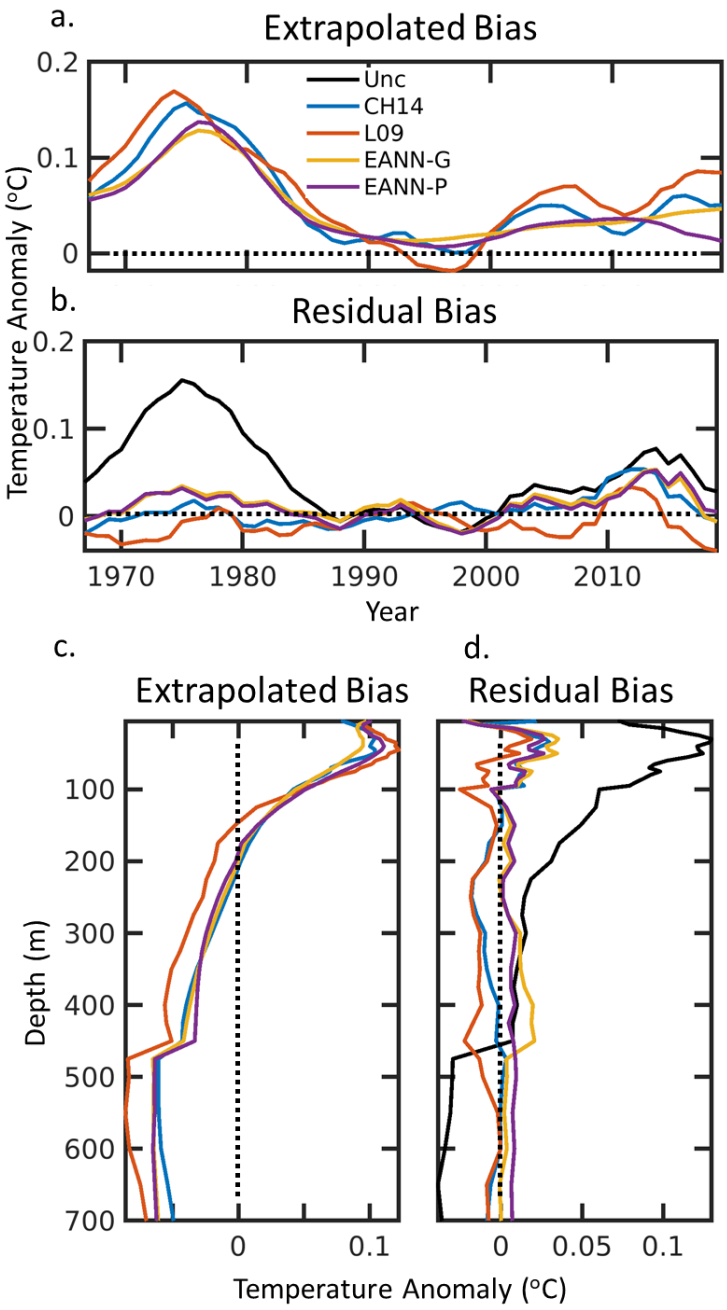


Fig. I.3 Global median extrapolated XBT bias (uncorrected XBT - corrected) for 0-700 m by year (a) and depth (c) for various XBT calibration methods (CH14, L09, and our EANN-G and EANN-P) after binning to the WOA13 grid. Also shown are residual biases with respect to our validation dataset (corrected XBT - Ref_{VAL}) for data corrected with the various XBT calibration methods, and the uncorrected XBT data (Unc), after binning to the WOA13 grid and taking the median by year (b) and depth (d).

performance when compared to the two top performing calibrations, L09 (Levitus *et al.*, 2009) and CH14 (Cheng *et al.*, 2014), as determined by Cheng *et al.* (2018)

These previous calibrations are dissimilar from this current study as well as from each other in their approach. L09 calculated the total bias as the difference between XBT data and a combined reference dataset of both OSD and CTD data after each dataset had been binned to a regular grid. Next, they took the global median of the bias for each depth level and year as their correction after smoothing with a 5-year moving average filter. CH14 applied independent depth and pure thermal bias corrections, while taking into

account the various probe types with respect to a reference set of CTD casts (later updated with a reference set of OSD, CTD, and PFL casts). Our method uses correction grids (Section B.1) that have been smoothed and filled using ensembles of artificial neural networks to reduce biases with respect to year, depth, water temperature in the top 10 m, and probe types.

We use four metrics adapted from Cheng *et al.* (2018), as well as a new metric of our own, to compare the residual biases of these disparate methods with respect to the same reference dataset of CTD/ OSD data (Ref_{VAL}) after the residual biases of the individual casts are binned to the WOA13 grid (Section C.1). Additionally, given the uncertainty across different ocean heat content estimates on basin scales (Wang *et al.*, 2018), we compare how these methods perform and extrapolate in the various ocean basins (Section C.2). We also consider differences in how these methods extrapolate to locations where there is no collocated reference data (Section C.3). Similarly, we consider both the residual biases and extrapolated biases for the MBT dataset using our calibration method as well as those of L09 and GR10 (Gouretski and Reseghetti, 2010) (section C.4). Our main analysis and figures use the global (EANN-G) and probe-specific (EANN-P) corrections to XBT data that uses the MFR FRE, but we also provide metrics for EANN methods that have been applied to XBT data that use the H95 FRE (see Tables I.1 and I.2).

C.1 Assessing global XBT bias correction

The World Ocean Database 2018 provides XBT data that have been pre-calibrated using some of the most popular XBT corrections including the CH14 and L09 corrections we consider here. We opted to use these pre-calibrated XBT data in our comparison, as we

could download, quality-control, and process them exactly in the same way as the uncorrected data (Section B.1). As with the uncorrected data, XBT data from the WOD 2018 corrected with the CH14 and L09 methods were interpolated to the 350 standard depth levels used in Cheng *et al.* (2018). The residual biases for the uncorrected XBT data, as well as the CH14, L09, and EANN corrections were then calculated by subtracting our reference validation set of CTD/ OSD data (Ref_{VAL}) from the interpolated XBT casts.

Next, these biases were binned to the WOA13 grid using the same procedure discussed in Section B.1. Calculating these residual biases on the individual casts and then binning them to a regular grid preserves information gained from the higher vertical resolution of the individual casts while also ensuring more frequently sampled regions, such as coastal areas, are not disproportionately represented in the metrics that follow. Additionally, this method

Metric	Calibration							
	Unc (MFR)	Unc (H95)	CH14 (H95)	L09 (H95)	EANN-G (H95)	EANN-P (H95)	EANN-G (MFR)	EANN-P (MFR)
Metric 1								
Global	0.048	0.049	0.017	0.017	0.011	**0.008**	0.016	0.017
Atlantic	0.040	0.043	0.020	0.022	0.018	0.017	0.020	**0.017**
Pacific	0.064	0.072	0.038	0.040	0.042	0.037	0.034	**0.031**
Indian	0.057	0.061	0.051	0.047	0.043	0.044	**0.038**	0.041
Metric 2								
Global	0.063±0.042	0.126±0.031	0.011±0.008	0.011±0.006	0.016±0.012	**0.010±0.009**	0.014±0.011	0.011±0.007
Atlantic	0.072±0.044	0.131±0.027	0.013±0.011	0.017±0.009	0.011±0.011	**0.009±0.007**	0.015±0.010	0.013±0.007
Pacific	0.062±0.041	0.132±0.035	0.013±0.010	0.009±0.006	0.017±0.009	0.010±0.008	0.012±0.009	**0.009±0.007**
Indian	0.043±0.023	0.094±0.039	0.030±0.020	0.032±0.017	**0.017±0.016**	0.024±0.016	0.032±0.025	0.033±0.024
Metric 4								
Global	0.083±0.054	0.119±0.058	0.033±0.029	0.039±0.031	0.031±0.027	**0.030±0.026**	0.037±0.032	0.036±0.032
Atlantic	0.097±0.073	0.138±0.083	0.053±0.050	0.060±0.053	0.054±0.051	**0.052±0.048**	0.054±0.053	0.053±0.052
Pacific	0.101±0.151	0.133±0.150	0.058±0.140	0.063±0.134	0.057±0.138	**0.055±0.133**	0.064±0.139	0.061±0.137
Indian	0.085±0.087	0.115±0.104	0.073±0.075	0.081±0.081	**0.070±0.079**	0.073±0.080	0.075±0.082	0.074±0.080
Metric 5								
Global	0.051±0.005	0.052±0.004	0.027±0.011	0.029±0.010	0.027±0.009	0.027±0.009	0.025±0.008	**0.024±0.008**
Atlantic	0.046±0.008	0.052±0.008	0.033±0.013	0.036±0.011	0.031±0.013	0.030±0.011	0.031±0.012	**0.028±0.008**
Pacific	0.083±0.021	0.085±0.021	0.061±0.027	0.062±0.031	0.060±0.024	0.052±0.018	0.056±0.025	**0.047±0.016**
Indian	0.114±0.081	0.109±0.065	0.111±0.086	0.102±0.063	**0.097±0.070**	0.097±0.072	0.102±0.083	0.102±0.084

Table I.1 Performance metrics of the Uncorrected XBT (Unc), CH14, L09, EANN-G, and EANN-P methods against a reference validation dataset REF_{VAL} (data that is not seen by the EANN calibrations). The choice of fall rate equation that each method uses, either the original manufacturer (MFR) or Hanawa *et al.* (1995) (H95) is also listed. Bolded values indicate best performance for that metric.

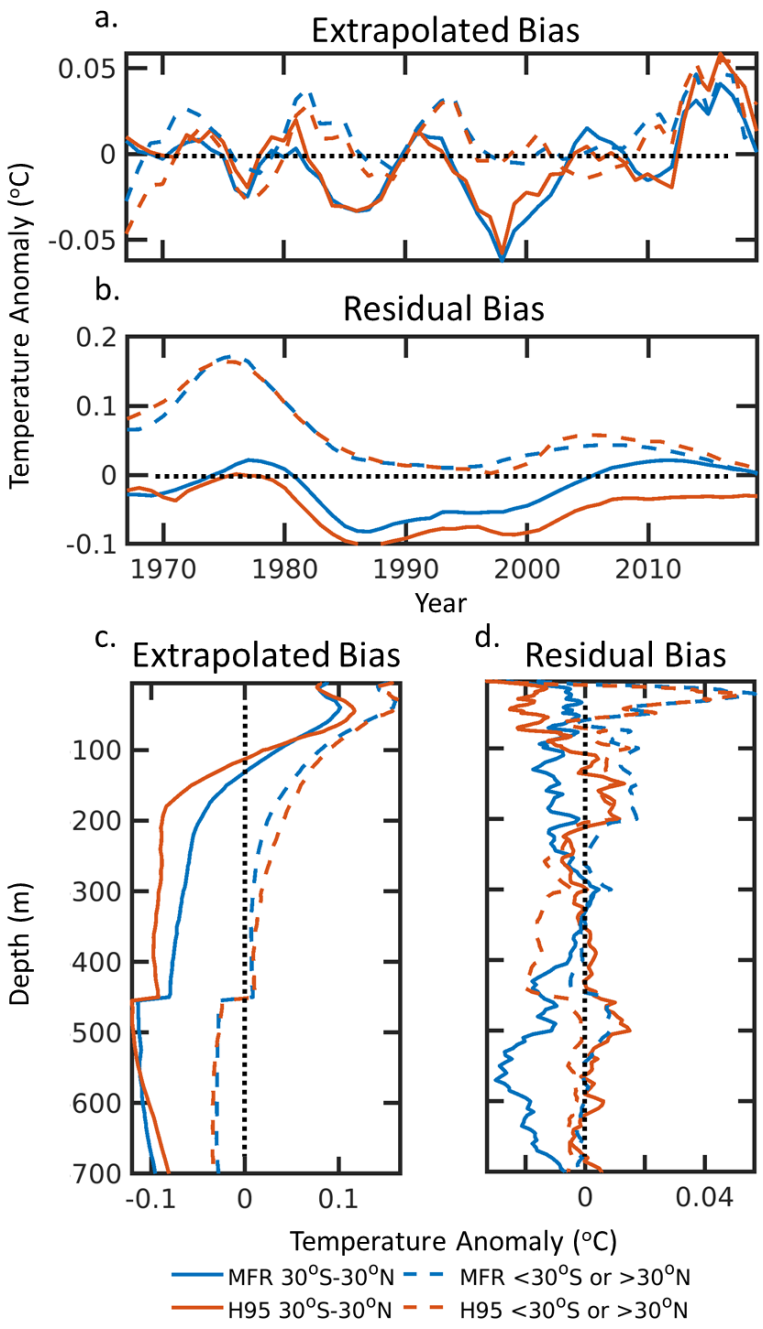
of comparison provides a reasonable compromise given that L09 biases were originally calculated on a regular grid, while the CH14 biases were calculated on interpolated casts at a much higher vertical resolution. Performance of the calibrations on a regular grid is the most relevant for studies of OHC, since studies of OHC rely on gridded temperature anomalies.

The five metrics that follow for assessing the original and residual XBT biases rely on taking a global median of the bias (represented as an overbar in the equations) on this regular grid (b_{WOA} in Eq. 3) to sort them into relevant bins that isolate components such as the temporal and depth biases.

The first metric we consider measures the reduction in the temporal bias, with a good method having minimal time variance in the residuals. As in Cheng *et al.* (2018), we too use a five-year moving filter when compositing our biases to annual temporal bins, however we opt to bin using the global median instead of the mean, as this has proven to be more robust. Thus, the first metric measures the standard deviation of the temporal component of the bias after aggregating the total XBT bias to annual bins using a five-year moving filter.

$$\text{Metric 1} = \sigma[\overline{b_{WOA}}(\text{yr}) \quad (6)$$

Fig. I.3a shows the global median extrapolated temporal bias (uncorrected XBT - corrected) of the various calibration schemes for 0-700 m. In our analysis, negative (positive) bias implies the uncorrected XBT data is too cold (warm) relative to the corrected data. All extrapolations from these different methods follow the same general temporal pattern with a few exceptions. CH14 and L09 show an earlier peak in the bias during the 1970s, occurring prior to 1975, while EANN-G and EANN-P peak after 1975. All methods show minimal bias in the late 1980s, with small biases persisting through the 1990s. The



Supplementary Fig. I.1 Median extrapolated XBT bias (uncorrected XBT - corrected) for 0-700 m by year (a) and depth (c) for the EANN-P XBT calibration method using either the MFR (blue) or H95 (red) fall rate equation and considering either the low latitudes (solid) or high latitudes (dashed). Also shown are residual biases with respect to the CTD casts (not binned to a regular grid) for data corrected with these same XBT calibration methods, taking the median by year (b) and depth (d).

CH14 and EANN methods are in general agreement throughout the 1990s, whereas the L09 method indicates that the bias becomes negative in the 1990s (Fig. I.3a). All methods agree that a positive bias re-appears in the 2000s, although the magnitude and temporal pattern of the bias differs between the various corrections by as much as > 0.05 °C. Both EANN-G and EANN-P exhibit a smooth leveling off in their extrapolated bias, whereas CH14 and L09 exhibit an oscillatory pattern (Fig. I.3a).

The choice of FRE also impacts the extrapolated bias of individual XBT probes (Supplementary Fig. I.1a). While at high latitudes (> 30°)

the difference is negligible, at low latitudes (30°S – 30°N) the extrapolated bias for the EANNP method using the H95 FRE is colder than the same method using the MFR FRE after 1970. Not only is the difference in the extrapolations due to the choice of FRE nonlinear, it also results in the extrapolations having the opposite sign after 2005 (Supplementary Fig. I.1a).

All four methods reduce the temporal bias against our independent validation dataset to within 0.05 °C for the entire period 1967-2018, as shown in Fig. I.3b. However, CH14 and the EANN methods appear to underestimate the positive bias in the XBT data for the period after 2010, when data becomes sparser. Prior to the minimum in the uncorrected XBT bias in 1987, the L09 method has a slight cold bias against the validation dataset, while the EANN methods have a slight warm bias (Fig. I.3b). The residual biases for all methods, including the uncorrected XBT bias, converge at the 1987 minimum and diverge again afterwards. The L09 residual bias transitions from positive in the early 1990s to negative in the early 2000s. The CH14 method is in broad agreement with the two EANN methods after 1990, except for a few years in the late 1990s where the CH14 method has a small warm bias and the EANN methods have a small cold bias. Overall, all methods perform quite similarly on Metric 1 (Table I.1), with a slight edge given to EANN-P or EANN-G, depending on the choice of FRE. Unlike the extrapolated biases, the residual time-dependent biases versus the available CTD data do not indicate a clear or consistent difference arising due to the choice of FRE (Supplementary Fig. I.1b).

The second metric we consider measures the residual depth bias, expressed as the average of the absolute XBT biases across depth bins from 1 to n (1 being the 0 m bin and *n*

= 41 being the 700 m bin). Once again, we bin using the global median since it is more robust to outliers.

$$\text{Metric 2} = \sum_{i=1}^{n} \left| \overline{b_{WOA}}(z_{WOA_i}) \right| / n \qquad (7)$$

Fig. I.3c shows profiles of the global median depth-dependent extrapolated biases of XBT data for the various calibrations for 0-700 m, demonstrating a positive XBT bias in the top 150 m and a negative bias below ~200 m for all calibration methods. EANN-P and L09 predict slightly larger extrapolated biases in the top 100 m, but the L09 extrapolated bias diminishes more rapidly, becoming negative at the shallowest depth of any of the methods (Fig. I.3c). All methods exhibit a transition in their extrapolated biases around 450 m, as this is the terminal depth of certain probe types and represents a change in the makeup of the probe data. The residual biases against our reference validation dataset (Fig. I.3d) demonstrate maximal disagreement among the methods at ~450 m depth, where the L09 method is biased cold and the EANN-G method is biased warm, while the probe-specific calibrations, CH14 and EANN-P, have smaller residual biases. Globally, all methods significantly reduce the depth-dependent bias, but EANN-P is slightly better than the other methods based on Metric 2 (Table I.1). Both the extrapolated and residual depth dependent biases reveal the impact of the choice of FRE, though the effect is most apparent for lower latitudes (Supplementary Fig. I.1c-d).

Although there is significantly less XBT data below 700 m (roughly 3 percent of that for 0-700 m), these data are nonetheless important to studies that consider warming in the deep ocean. While certain calibrations have previously been applied to deep XBT data, the performance of these methods has not been well investigated. We therefore briefly consider

the global median temporal and depth biases for 700-2000 m. The L09 method is not included here, as it was originally only applied to the 0-700 m depth interval (although in the WOD 2018, TSK-T5 probes have received a correction using the method of Kizu *et al.*, 2005, in addition to having the H95 FRE applied).

The temporal component of the extrapolated XBT bias below 700 m differs widely between the CH14 and EANN methods prior to the mid-1980s, when the methods disagree about both the sign and magnitude of the extrapolated bias (Fig. I.4a). This disagreement can mainly be attributed to the fact that the CH14 method employs a depth correction to the FRE, which adjusts the sampling depths of the XBT measurements and thus the overlap in gridded data with the uncorrected XBT data using the MFR FRE, whereas the EANN method only considers a thermal bias and makes no adjustment to the sampling depths. From the 1980s onwards, the CH14 extrapolated bias oscillates with both a similar period and phase to the extrapolated temporal bias in the top 700 m (Fig. I.3a), though the sign of the bias is reversed. This contrasts with the EANN methods, which exhibit a smoother time-dependence. The residual time-dependent bias (Fig. I.4b) for the uncorrected XBT data prior to 1990 likely isn't well resolved due to the sparsity of deep casts from this period, but it does indicate a larger temporal bias during the pre-1990s period, which is consistent with the timing and magnitude of this bias for depths shallower than 700 m (Fig. I.3b). After 1990 the temporal bias is quite minimal, indicating that deep probe types (such as the T5 and TSK-T5) may not suffer from a significant time-varying bias. For the period after 1990, it is not clear that any of the calibration methods have much of an effect or are even necessary. Prior to 1990, the CH14 method is more successful than the EANN methods at reducing the temporal bias.

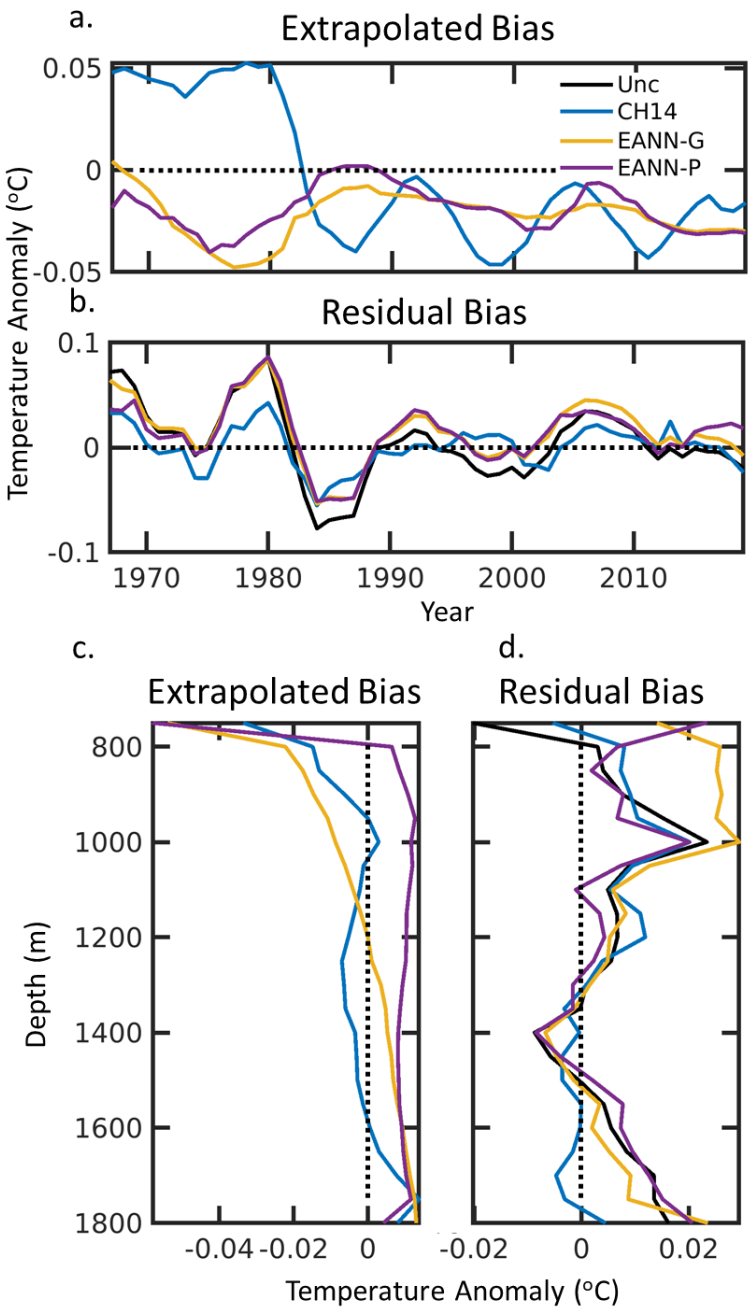


Fig. I.4 Same as Fig. I.3 but for the 700-1800 m depth interval.

Certain probe types have terminal depths at 760 m, producing a rapid transition in the extrapolated bias of the CH14 and EANN methods around this depth (Fig. I.4c). In the case of the EANN-P method this includes a reversal in the sign of the bias, though the magnitude remains stable below 800 m. This contrasts with the EANN-G method, which undergoes a smoother shift in the extrapolated bias with depth, and a change in sign at roughly 1200 m. In the case of the CH14 method, the extrapolated bias briefly reverses sign around 1000 m and again below 1600 m. The residual bias of the uncorrected XBT data below 700 m demonstrates that the deep probe groups do not exhibit significant bias at these depths (Fig. I.4d). The residual bias of the calibrated XBT data is not much improved from the uncorrected data, aside from the CH14

method for depths below 1500 m. The EANN-G method is a global correction and cannot handle the transition in the makeup of the probes around 760 m, leading to an increased residual bias compared to the uncorrected XBT data above 1000 m (Fig. I.4d). Based on the metrics (Table I.3), we find that the CH14 correction is most effective at reducing the bias for the 700-1800 m depth interval, however we emphasize that the uncorrected XBT data using the MFR FRE may already be sufficient, whereas the use of the H95 FRE greatly increases the bias with depth. We recommend using either the uncorrected XBT with the MFR FRE, or the CH14 or EANN-P corrections.

Metric	Calibration						
	Unc (MFR)	Unc (H95)	CH14 (H95)	EANN-G (H95)	EANN-P (H95)	EANN-G (MFR)	EANN-P (MFR)
Metric 1	0.036	0.047	**0.020**	0.030	0.029	0.031	0.030
Metric 2	0.008±0.006	0.083±0.019	**0.006±0.005**	0.009±0.007	0.009±0.005	0.011±0.010	0.008±0.007

Table I.3 Performance metrics of the Uncorrected XBT (Unc), CH14, EANN-G, and EANN-P methods against a reference validation dataset REF_{VAL} (data that is not seen by the EANN calibrations) for the 700-1800 m depth interval. The choice of fall rate equation that each method uses, either the original manufacturer (MFR) or Hanawa *et al.* (1995) (H95) is also listed. Bolded values indicate best performance for that metric.

Our third metric considers the depth-dependent bias separately for individual probe types, as depth profiles show distinct biases for different probes. Metric 3 is the same as Metric 2, but it is applied to the nine categories of XBT probes specified in Cheng *et al.* (2014) whose biases with respect to Ref_{VAL} have been separately binned to the WOA13 grid.

$$\text{Metric 3} = \sum_{i=1}^{n} \left| \overline{b_{WOA}}(z_i, \text{probe}) \right| / n \qquad (7)$$

The results indicate that global calibrations such as L09 and EANN-G perform similarly to probe-specific calibrations such as CH14 and EANN-P at reducing the bias of the most common XBT probe types (Fig. I.5 and Table I.2). While probe-specific calibrations are better overall (Table I.2), the global XBT bias is dominated by only a few probe types that

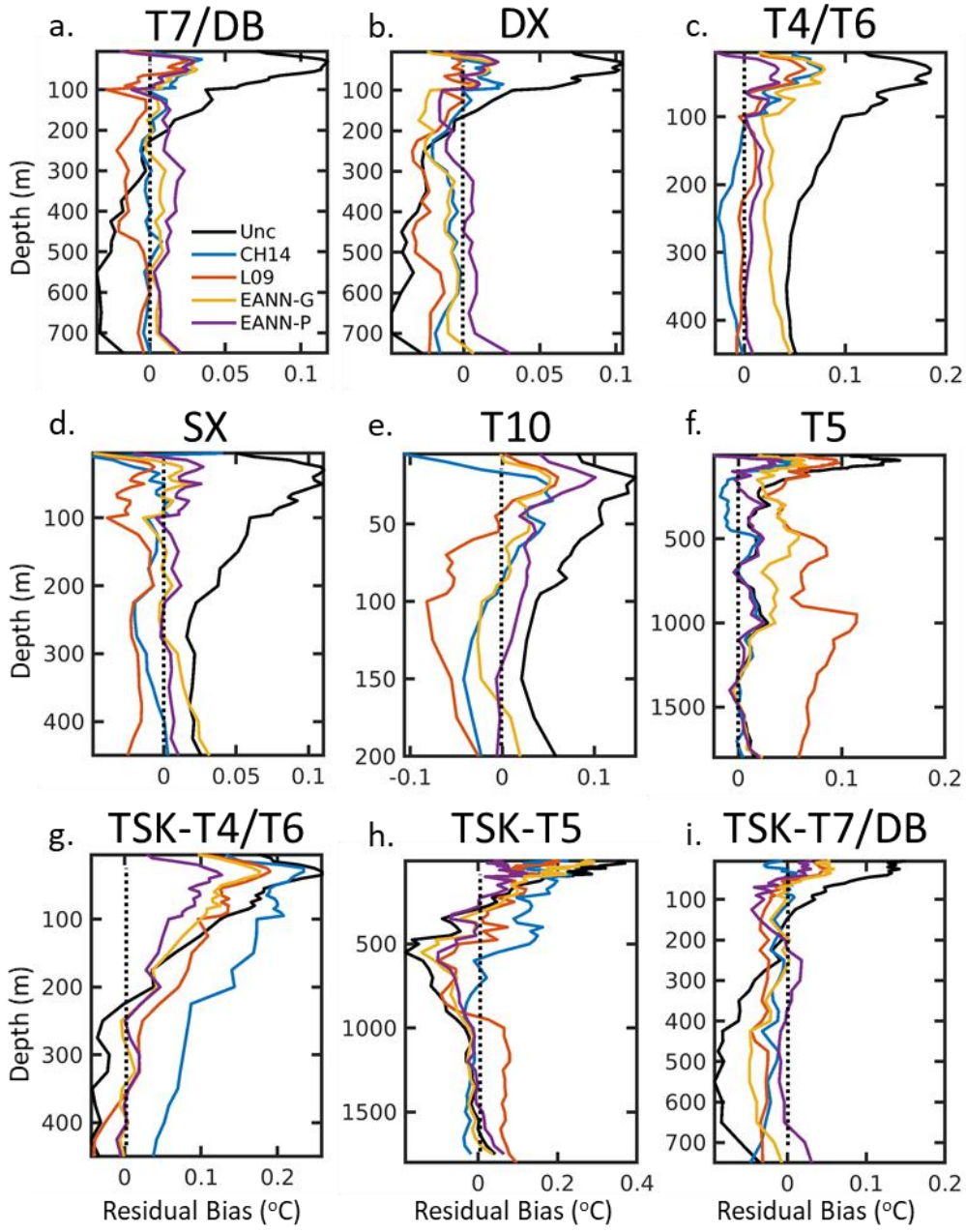


Fig. I.5 Residual XBT biases with respect to our validation dataset (corrected XBT - Ref_{VAL}) for various XBT calibration methods (CH14, L09, and our EANN-G and EANN-P) after separating the data into nine probe types, binning to the WOA13 grid, and taking the median by depth.

the global calibrations can adequately handle. The choice of FRE can also have a major impact on the residual bias of certain calibrations. As an example, the T5 and TSK-T5 probes have very little bias below 700 m, but applying the H95 FRE significantly increases

the depth dependent bias, especially for the L09 method, which does not perform a correction below 700 m aside from applying the method of Kizu *et al.* (2005) to the TSK-T5 probes. In addition, the most common probe types (Fig. I.5a.-d.) tend to have smaller biases for the uncorrected XBT data, which could be the result of having more collocated data to actually resolve the bias or due to mixing of probe types of different manufacturers for the unknown groups.

Due to insufficient metadata, nearly half of all XBT casts cannot be directly assigned to a manufacturer or probe type. Instead they are sorted into categories of shallow probes (SX) and deep probes (DX). Shallow probes are those with maximum reported depths less than 450 m, which could be probes that have reached their terminal depths (e.g. T4/ T6 from either Sippican or TSK and the Sippican T10) or any other probe type that was deployed without reaching terminal depth, which is likely in many coastal areas. Deep probes have maximum depths less than 930 m, including T7/ DB probes and some T5 probes.

Probe Group	Calibration							
	Unc (MFR)	Unc (H95)	CH14 (H95)	L09 (H95)	EANN-G (H95)	EANN-P (H95)	EANN-G (MFR)	EANN-P (MFR)
T7/DB	0.036±0.039	0.076±0.027	**0.007±0.010**	0.008±0.007	0.010±0.011	0.010±0.008	0.008±0.009	0.009±0.008
DX	0.036±0.031	0.067±0.030	0.008±0.007	0.010±0.011	0.008±0.006	**0.005±0.006**	0.008±0.011	0.006±0.006
T4/T6	0.060±0.051	0.096±0.035	0.017±0.023	0.010±0.019	0.024±0.024	**0.004±0.006**	0.022±0.019	0.008±0.011
SX	0.035±0.035	0.069±0.025	0.006±0.012	0.012±0.009	**0.005±0.010**	0.008±0.010	0.006±0.011	0.006±0.007
T10	0.031±0.035	0.047±0.035	0.012±0.023	0.016±0.025	0.010±0.023	0.038±0.037	**0.008±0.015**	0.014±0.026
T5	0.049±0.052	0.119±0.036	0.018±0.016	0.068±0.020	0.029±0.019	0.026±0.015	0.032±0.019	**0.014±0.011**
TSK-T4/T6	0.070±0.083	0.105±0.082	0.084±0.063	0.056±0.058	0.058±0.062	0.040±0.051	0.047±0.060	**0.032±0.041**
TSK-T5	0.134±0.119	0.184±0.137	0.098±0.072	0.084±0.044	0.117±0.100	0.054±0.047	0.102±0.091	**0.045±0.033**
TSK-T7	0.045±0.042	0.064±0.045	**0.009±0.011**	0.019±0.012	0.016±0.020	0.038±0.031	0.016±0.019	0.012±0.012

Table I.2 Results of applying Metric 3 to the Uncorrected XBT (Unc), CH14, L09, EANN-G, and EANN-P methods for each of the nine probe types considered in this study. The choice of fall rate equation that each method uses, either the original manufacturer (MFR) or Hanawa *et al.* (1995) (H95) is also listed. Bolded values indicate best performance for that probe type.

Manufacturer origin appears to have an impact on the depth bias of some probe types, which a global correction cannot address. For T4/ T6 probes (Fig. I.5c and Fig. I.5g.), the sign of the bias below 100 m depends on the manufacturer, and the resulting bias of the mixed probes of the SX group (Fig. I.5d.) falls somewhere in between. Global calibrations are also not suitable for T5 (Fig. I.5f) or TSK-T5 (Fig. I.5h) probes, as these apparently have different bias histories at intermediate depths. On the other hand, the use of probe-specific calibrations may run the risk of overfitting for the least common probe types. Considering the TSK-T4/T6 probes (Fig. I.5g), for instance, not many casts are known to exist, which could explain the poor performance of the CH14 method using our validation set (Table 2).

The fourth metric considers the reduction in the spatial component of the bias. An existing zonal bias may in fact be related to water temperature (Thadathil *et al.*, 2002, Kizu *et al.*, 2005, Reverdin, 2009, Cowley *et al.*, 2013, Cheng *et al.*, 2014) as well as the vertical temperature gradient (Gouretski and Reseghetti, 2010). Neither the CH14 nor the EANN calibrations directly correct for the zonal component of the XBT bias, but they do include factors to correct for a temperature-dependent bias. L09 corrects for total XBT bias as a function of both depth and year but does not consider a temperature-dependent bias. However, the L09 method uses the H95 FRE, which produces a spatial pattern in the bias correction relative to the MFR FRE that varies with water temperature. While the H95 depth correction factor attempted to be optimal for the global ocean, this appears to break down on regional scales due to in part to the time-varying pure thermal bias (Gouretski and Reseghetti, 2010).

We express the fourth metric as the average bias over *n* depth layers (*n* = 1 at 0 m to *n* = 41 at 700 m) and *m* latitudinal bins (*m* = 1 at 69.5° S to *m* = 140 at 69.5° N and progressing by 1° intervals) after binning using the global median.

$$\text{Metric 4} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left|\overline{b_{WOA}}(z_i, \text{lat}_j)\right| / (n \cdot m) \quad (8)$$

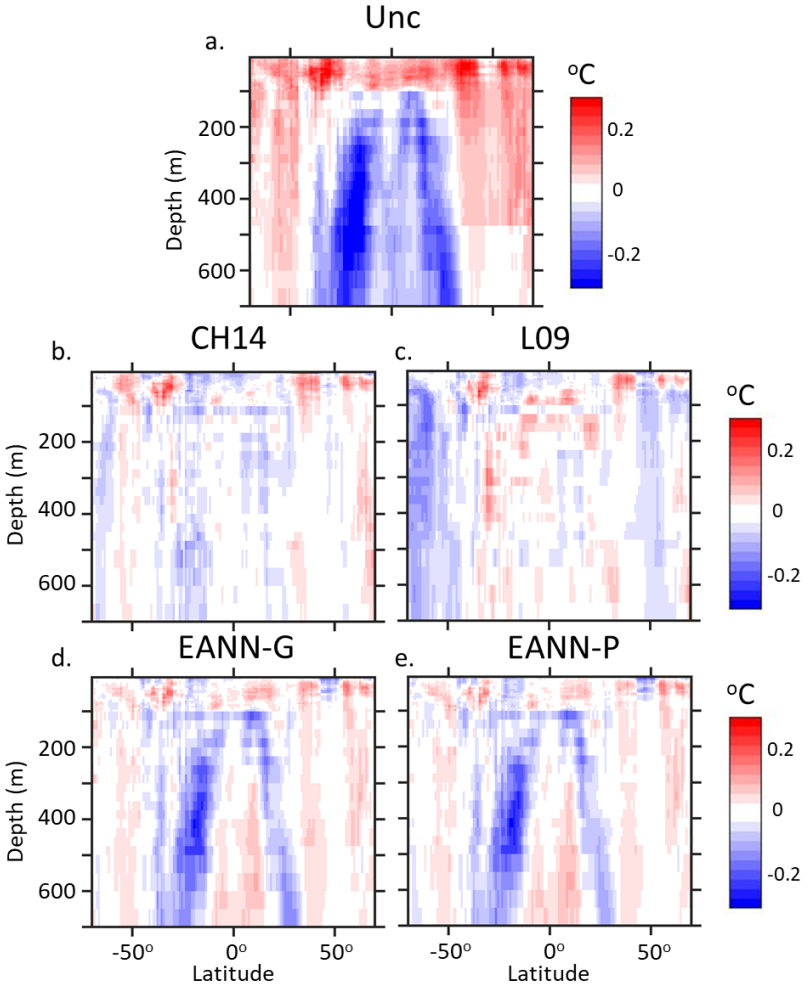
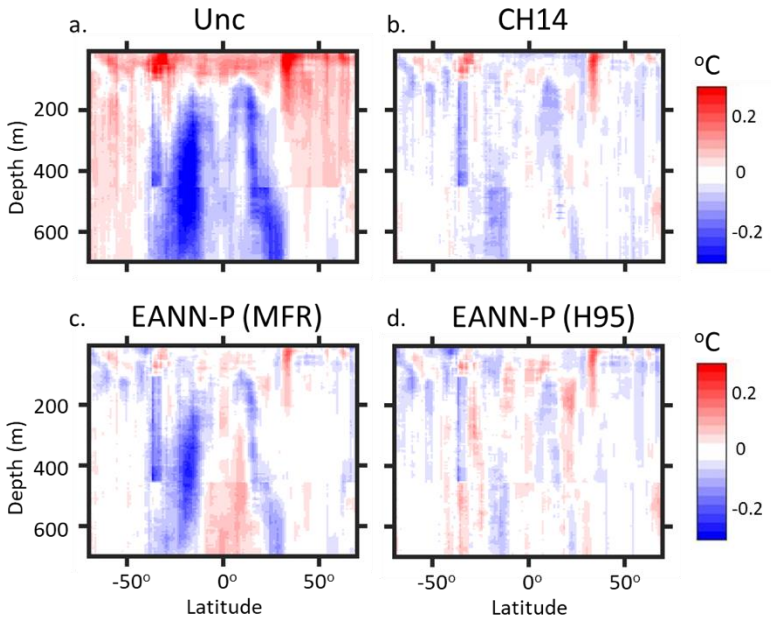


Fig. I.6 Residual XBT biases with respect to our validation dataset (corrected XBT - Ref_{VAL}) as a function of both depth and latitude for uncorrected XBT data (Unc) (a) , data corrected with the CH14 method (b), data corrected with the L09 method (c), data corrected with the EANN-G method (d), and data corrected with the EANN-P method (e).

The uncorrected zonal XBT bias exhibits a fairly symmetric pattern across the equator (Fig. I.6a), with positive biases below 100 m at high latitudes transitioning to negative biases at the low latitudes in a pattern that resembles the zonally averaged water temperature and vertical temperature gradient (Gouretski and Reseghetti, 2010, their Figure 12). Aside from a small overcorrection in the Southern Hemisphere

and in the tropics, the CH14 correction removes most of the original spatial pattern of the XBT bias (Fig. I.6b). At both high and low latitudes, the L09 method overcorrects, turning positive biases negative and vice versa (Fig. I.6c). Both the EANN-G (Fig. I.6d) and EANN-P (Fig. I.6e) calibrations reduce much of the original spatial bias in the XBT data. However, below 100 m the EANN methods under-correct the most prominent biases, most notably the strong negative biases in the low latitudes. Due to the lower vertical resolution of the OSD casts, including the interpolated OSD data in this comparison exaggerates the apparent spatial biases compared to the CTD data alone. However, reviewing the bias of the original XBT casts versus the CTD data at high vertical resolution prior to binning to the WOA grid shows that the overall pattern remains consistent (Supplementary Fig. I.2a-c). Additionally, these low-latitude negative biases can be eliminated by first applying the H95 FRE and subsequently the EANN correction (Supplementary Fig. I.2d), which based on Metric 4 does best at reducing the spatial bias on our combined reference dataset of both CTD and OSD data (Table I.1). This indicates that the H95 FRE is effective at reducing spatial biases in the XBT data, even though



Supplementary Fig. I.2 Median residual biases with respect to latitude and depth compared to individual CTD casts (not binned to a regular grid) for a. Uncorrected XBT casts, b. the CH14 method, and the EANN-P method using either c. the MFR or d. the H95 fall rate equation.

without additional corrections it worsens the overall XBT quality (Table I.1).

Our final metric is a combination of the first two metrics and considers both the depth and time dependent bias components. Conceivably a low score in Metric 2 could be achieved by having a positive bias and a negative bias of similar magnitudes in the same part of the water column but in different years of the time series. A similar logic could be applied to how a low value for Metric 1 might also be achieved. While such an occurrence for one metric would degrade the value of the other, the possibility of having uncorrected biases of different signs that can partially compensate for each other warrants this additional metric. After using the global median to bin with regards to both year (using a 5-year moving filter for the temporal binning) and depth, we consider the standard deviation of the temporal component at each depth bin and then average these values over all depth bins to obtain Metric 5,

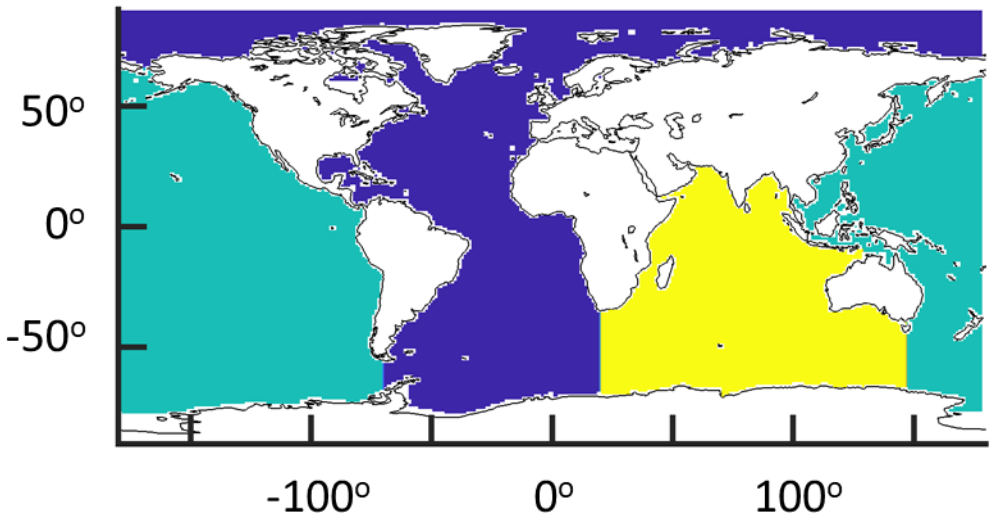
$$\text{Metric 5} = \sum_{i=1}^{n} \sigma[\overline{b_{WOA}}(\text{yr}, z_i)]/n \qquad (9)$$

The global median uncorrected XBT bias exhibits a nonlinear time evolution at different depth levels. It is always positive at depths shallower than 100 m, peaking in the mid-1970s, while below 100 m the bias shifts from positive in the 1970s to negative in the 1980s to more positive/ neutral again after 2000 (Fig. I.7a). The L09 method (Fig. I.7c) slightly overcorrects the original XBT bias in much of the top 600 m during the 1970-1990 period. CH14 (Fig. I.7b) and the EANN methods (Fig. I.7d-e) do not eliminate the shallow warm XBT biases from the mid-2000s onward but remove the majority of the bias prior to this. However, EANN-G (Fig. I.7d) under-corrects the bias prior to 1980 around the terminal

depth of some probe types at 450 m. Overall, the EANN-G, EANN-P, and CH14 calibrations all perform quite well based on Metric 5 (Table I.1).

C.2 XBT calibration performance in individual basins

The XBT calibrations differ in their performance on basin scales. In order to ensure that the performance of the XBT calibrations on Metrics 1, 2, 4, and 5 discussed in Section C.1 is not merely due to a canceling of errors across basins, we reproduced these metrics for the Atlantic, Pacific, and Indian Oceans. The geographic constraints we use for these basins are found in Supplementary Supplementary Fig. I.3. All metrics are calculated the same as in Section C.1 aside from Metric 4, which differs in the number of latitudinal bins for the Pacific (m = 135) and Indian (m = 95).



Supplementary Fig. I.3 Map of the basin extents used to do ocean basin scale analyses in the main text. The Pacific is in teal, the Atlantic blue, and the Indian yellow. The Arctic is grouped with the Atlantic and the Southern Ocean is grouped with other respective ocean basins due to the sparsity of data in these regions.

The Atlantic Basin exhibits a similar pattern in the uncorrected XBT bias with regards to depth and year as was seen globally, with alternating positive and negative biases over time

below 100 m (Fig. I.7f). However, the transition from positive to negative bias is double peaked for the global bias, with one negative peak in the late 1980s and a subsequent one in the 1990s, while for the Atlantic there is only one distinct negative peak that occurs later than the first peak in the global bias. Both CH14 and L09 overcorrect the initial positive bias to some extent before 1980 (Fig. I.7g-h), with the exception of L09 in the top 200 m after

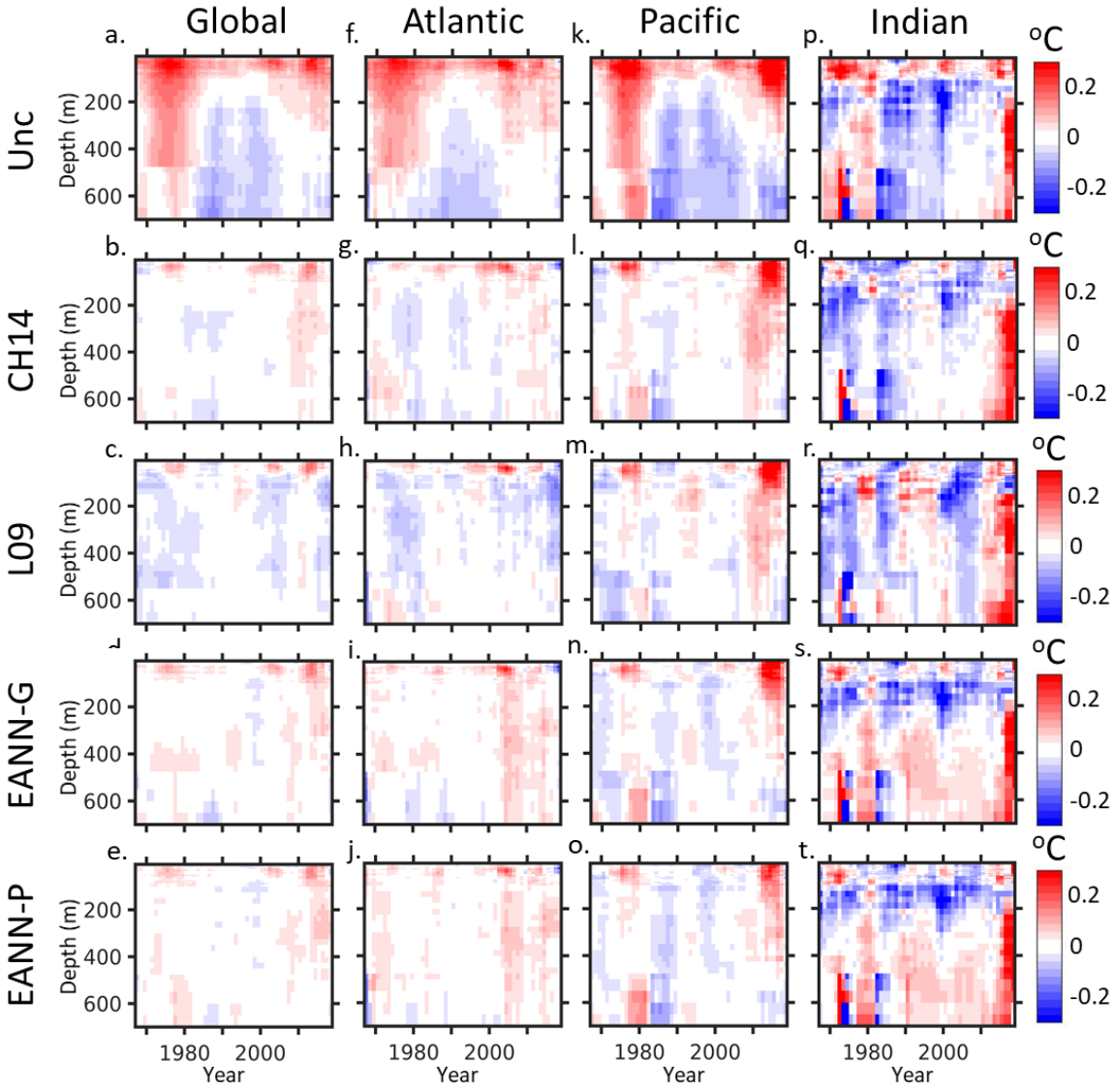


Fig. I.7 Residual XBT biases with respect to our validation dataset (corrected XBT - Ref_{VAL}) as a function of both depth and year for uncorrected XBT data (Unc), data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the EANN-P method, globally and for individual basins (Atlantic, Pacific, and Indian).

year 2000 where it overcorrects. The EANN methods (Fig. I.7i-j) reduce the bias more evenly across time and depth but under-correct significantly after 2010. CH14 also under-corrects to some extent after 2010, including extending the positive bias to greater depths like the EANN methods, whereas the L09 overcorrects during the same period. All methods perform similarly on Metric 5 in the Atlantic (Table I.1).

The shape of the uncorrected XBT bias in the Pacific (Fig. I.7k) dominates the global pattern (Fig. I.7a), but the pattern in the Pacific is more pronounced, as the offset in the timing of the negative bias in the top 450 m of the Atlantic partially compensates for that in the Pacific on global scales. All methods (Fig. I.7l-o) under-correct the positive biases in the 1970s and 2010s, but the EANN-G (Fig. I.7n) and EANN-P (Fig. I.7o) calibrations do a slightly better job at correcting the biases after 2010 below 200m. Again, based on the metrics, all methods perform quite similarly (Table I.1).

In the Indian Ocean (Fig. I.7p-t), the uncorrected XBT biases are larger and noisier than in the other basins because there is less data in this basin. The existing data does indicate that a pattern of positive biases before 1980 transitioning to negative biases in the 1980s and 1990s, that again transition to positive after 2000, is largely consistent across all of the ocean basins. Additionally, while the sparse data in the Indian makes it difficult to establish the performance of the different calibrations with any great certainty, it appears that all calibrations reduce the bias from the uncorrected XBT (Fig. I.7p). In the top 450 m, both CH14 and L09 (Fig. I.7q-r) appear to overcorrect the bias prior to 1980 and under-correct from the 1990s onward. The EANN (Fig. I.7s-t) methods leave a residual cold bias above ~ 300 m depth, and introduce a residual warm bias below 300 m. After 2010, there are large biases that no method is able to correct.

After reviewing the performance of the methods for individual ocean basins, all of the calibrations reduce the original XBT biases considerably. The differences in performance across the calibrations considered here are marginal, and we cannot distinguish a single best correction based on these metrics alone.

C.3 Spatial patterns of the extrapolated XBT bias for different methods

Each XBT calibration method reduces the biases present in the original uncorrected XBT data, but their different assumptions about the form of these biases and how to best correct for them lead to significant differences in how each correction generalizes to regions with no reference data to verify the calibration. The differences in how these calibrations extrapolate can contribute to uncertainty in estimates of ocean heat content on intra-decadal timescales and ocean basin scales. The ability of these products to extrapolate has previously been validated using independent datasets such as EN4 (Cheng *et al.*, 2018) or by withholding substantial amounts of CTD/ OSD, as was done in this study, but clearly reductions in the original XBT biases can be achieved by different methods with varying success on the global scale while leading to significant differences in the temporal and spatial patterns of the extrapolations.

On a global scale, the extrapolated biases of the CH14 and L09 methods over depth and year (Fig. I.8a-b) exhibit a similar pattern, but the L09 extrapolation is often of greater magnitude. One major distinction between CH14 and L09, and the EANN-G (Fig. I.8c) and EANN-P (Fig. I.8d) calibrations, is the presence of 2-3 distinct peaks of negative bias in the 1980s-2000s in the two former methods, versus a single peak of negative bias around 1990 in the latter two methods. The EANN-P method does infer a bias with a semblance of two

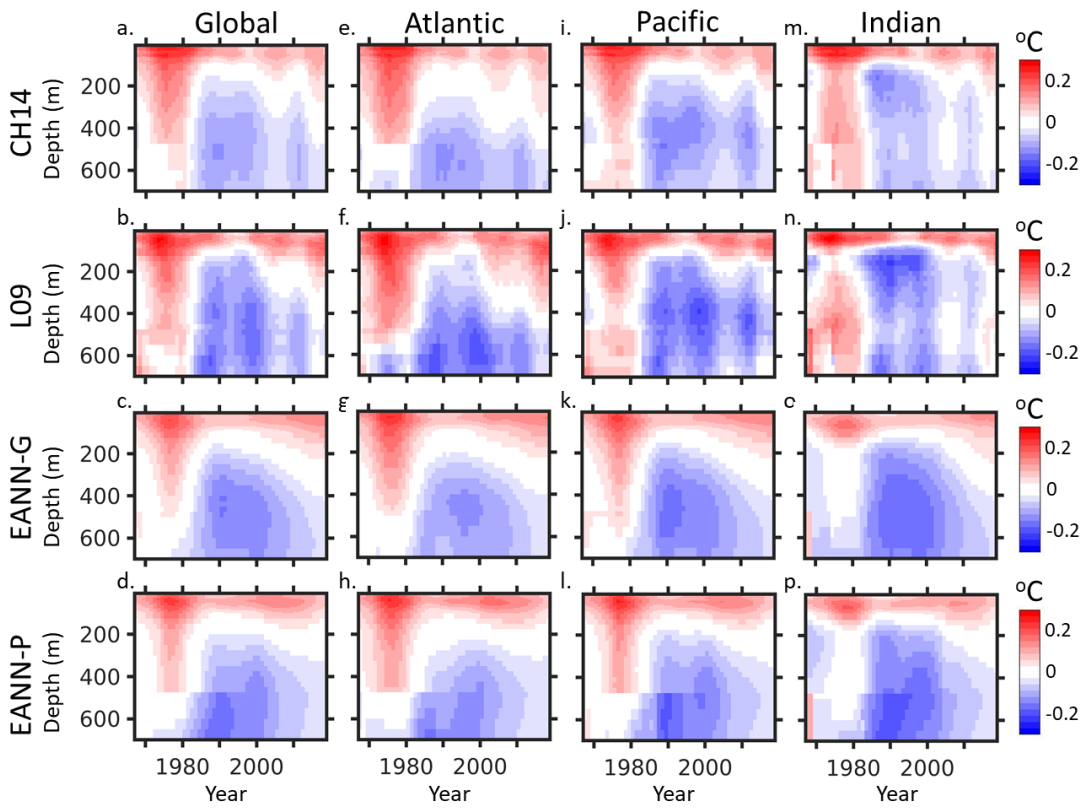


Fig. I.8 Median of the extrapolated XBT bias (uncorrected XBT - corrected) as a function of depth and year for data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the EANN-P method, globally and for individual basins (Atlantic, Pacific, and Indian).

negative peaks but remains smooth like the EANN-G after 2000. The EANN-P method is also distinguished from the EANN-G by a more distinct transition and stronger cold bias below 450 m, a depth that coincides with changes in the probe types.

The CH14 and L09 (Fig. I.8e-f) extrapolated biases in the Atlantic share a similar form that is mainly distinguished by a larger magnitude in the L09 extrapolation, especially for areas of negative bias, which is apparent at shallower depths in L09. EANN-G (Fig. I.8g) has a smoother extrapolated bias that is generally smaller than the other methods. The EANN-P (Fig. I.8h) method infers a larger bias but shallower positive bias above 200 m and after 2000 than the other methods.

Due to the amount of XBT data in the Pacific, this basin contributes an outsized amount to the shape of the global extrapolated bias, masking some of the distinctions arising in the Atlantic and Indian. Nonetheless, certain differences that were not as well resolved on the global scale become apparent when specifically considering the Pacific. For instance, in the 1960s and early 1970s there is a period where both the CH14 and L09 methods (Fig. I.8i-j) consider XBT data around 400 m to have a negative bias. The EANN methods (Fig. I.8k-l) differ from L09 and CH14 in their characterization of the bias below 450 m in the 1970s,

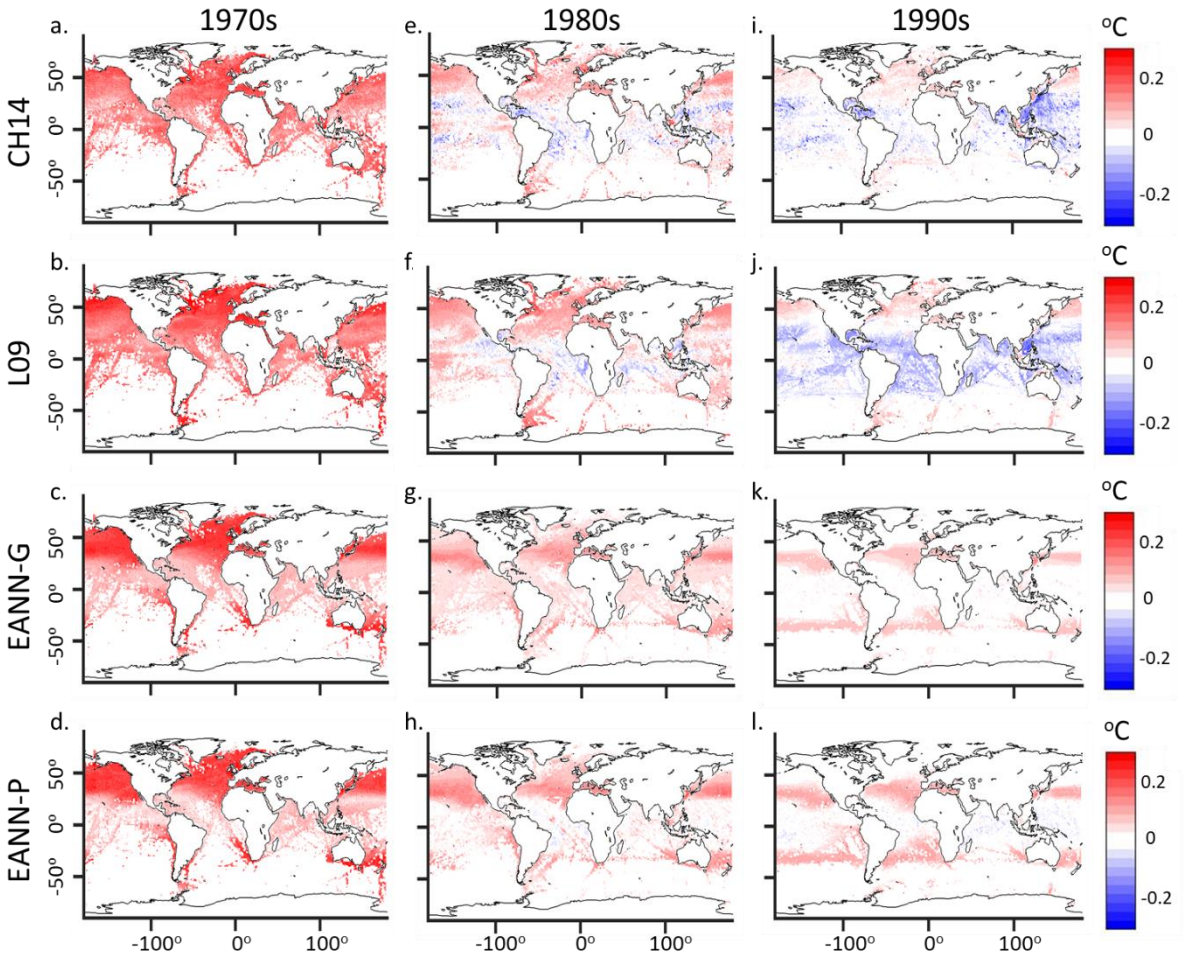


Fig. I.9 Decadal averages of the extrapolated XBT bias (uncorrected XBT - corrected) averaged over the top 700 m for data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the EANN-P method during the 1970s, 1980s, and 1990s.

which for the EANN methods is close to neutral while for L09 and CH14 the bias remains positive.

In the Indian, the extrapolated biases of the CH14 (Fig. I.8m) and L09 (Fig. I.8n) methods are highly distinct from those of the EANN methods (Fig. I.8o-p). While all methods consider some XBT biases to be negative in the 1960s, they disagree on the sign of the bias below 200 m in the 1970s, with CH14 and L09 indicating that the bias is positive, and the EANN methods indicating that it is neutral to negative. The L09 extrapolation also indicates that there are peaks of negative XBT biases at two different depths in the 1980s and 1990s, one around 200 m and one around 600 m (Fig. I.8n). The CH14 method shows a shallower peak in this negative bias (Fig. I.8m), while the EANN-G method spreads the negative bias more evenly across the water column (Fig. I.8o), and the EANN-P method concentrates this bias below 450 m (Fig. I.8p).

Maps of the extrapolated bias with the median taken over the top 700 meters for different decades reveal distinct spatial patterns in the biases over time (Fig. I.9). In the 1970s all methods overwhelmingly treat the uncorrected XBT data as warmer than the corrected, but while the CH14 method (Fig. I.9a) and L09 (Fig. I.9b) are more homogenous the EANN methods (Fig. I.9c-d) demonstrate an overall positive bias that is latitudinally dependent, with higher latitudes having larger biases than lower latitudes.

The four calibration methods have more distinct latitudinal patterns in the 1980s. While all methods indicate data in the high latitudes have a positive bias, data in tropics and subtropics differ in the sign of the bias across methods (Fig. I.9e-h). Both the CH14 (Fig. I.9e) and L09 (Fig. I.9f) have negative biases at low latitudes but these differ on basin

scales, with negative biases in the CH14 extrapolation being more extensive and farther west in their respective ocean basins than for the L09 extrapolated bias. The EANN-G method (Fig. I.9g) extrapolates positive biases at all latitudes, whereas the EANN-P method (Fig. I.9h) indicates a neutral to slightly negative bias in the tropics. For both EANN methods the positive bias peaks near 30^{o} N and S.

During the 1990s, these various methods diverge the most of any decade in the spatial patterns of their extrapolated biases. Although the CH14 (Fig. I.9i) and L09 (Fig. I.9h) methods share similarities, with high latitudes exhibiting positive biases and low latitudes having negative biases, the L09 method indicates larger magnitudes for the bias and a greater homogeneity to the negative biases. The EANN methods (Fig. I.9k-l) consider the XBT bias to be neutral at most latitudes, aside from bands of positive biases around 30^{o} N and S that mirror those from the 1980s.

C.4 Performance and extrapolation of the MBT calibrations

Systematic biases in the MBT data remain less well studied than the XBT, even though they are the dominant source of temperature data prior to 1967 (Fig. I.1). As with the XBT data, we consider the performance of several MBT calibration methods at removing depth- and time- dependent biases, as well as the form of the global extrapolated biases.

The global median extrapolated temporal biases are quite distinct for the different calibration schemes (Fig. I.10a), with the GR10 and L09 calibrations abruptly varying in the magnitude of their bias. This contrasts with the extrapolated bias inferred by the EANN-G method, which is temporally smooth (Fig. I.10a). The uncorrected time-dependent bias with respect to our validation dataset is also smooth (Fig. I.10b), peaking in 1954 around 0.1 ^{o}C

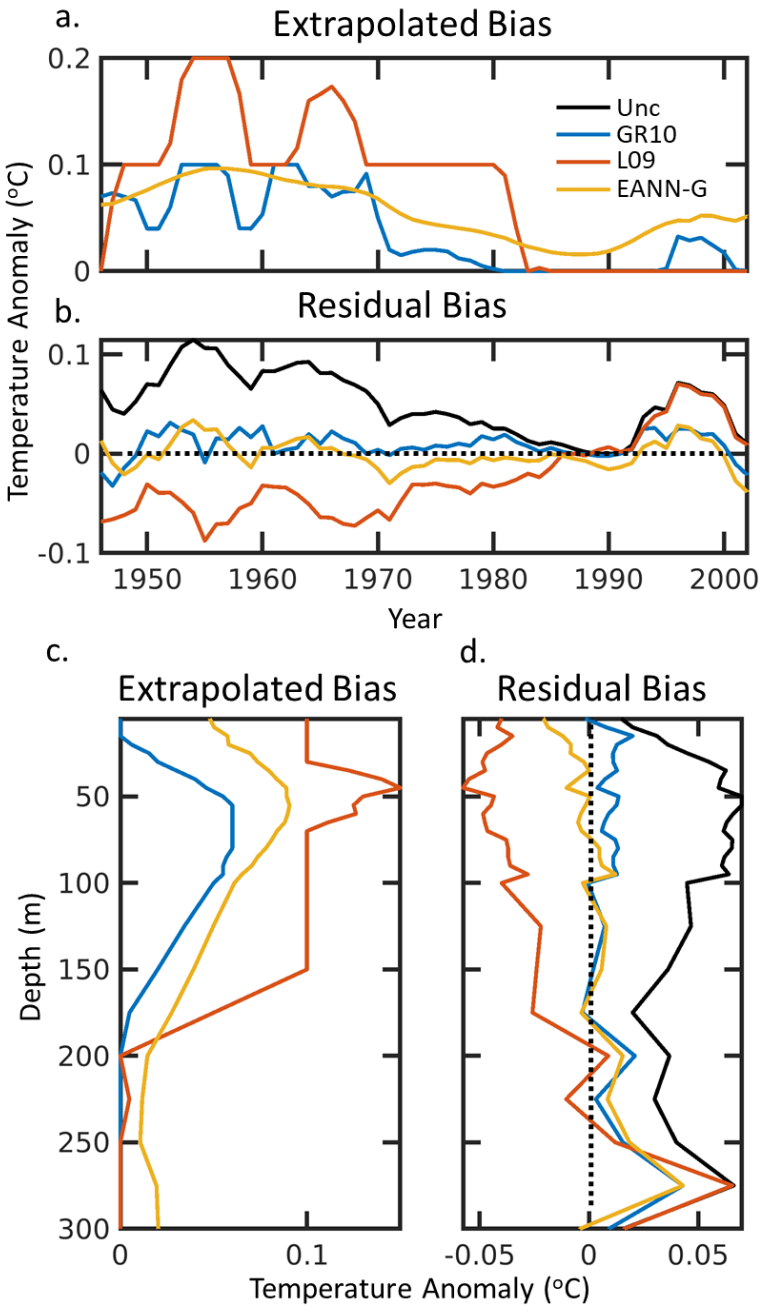


Fig. I.10 Global median extrapolated MBT bias (uncorrected MBT - corrected) for 0-300 m by year (a) and depth (c) for various MBT calibration methods (GR10, L09, and our EANN-G) after binning to the WOA13 grid. Also shown are residual biases with respect to our validation dataset (corrected MBT - Ref_{VAL}) for data corrected with the various MBT calibration methods, and the uncorrected MBT data (Unc), after binning to the WOA13 grid and taking the median by year (b) and depth (d).

and declining steadily so that by 1990 the remaining bias is almost zero. After 1990, the bias in the uncorrected MBT data increases again, but there is very little MBT data during this period. GR10 slightly under-corrects the original positive bias in the MBT data with respect to the validation dataset, whereas the EANN-G slightly overcorrects this positive bias (Fig. I.10b). However, both calibrations significantly reduce the original bias, outperforming the L09 method, which overcorrects for most of the period and offers no correction after 1994 (Fig. I.10b).

The median extrapolated biases with depth (Fig. I.10c), and the residual depth biases of the corrected data with respect

to our validation dataset (Fig. I.10d), tell a similar story. GR10 has a smaller extrapolated bias, and positive residual biases in the top 100 m, whereas the EANN-G method exhibits a larger extrapolated bias that also peaks at a shallower depth, leading to negative residual biases in the top 50 m (Fig. I.10c-d). L09 creates a larger extrapolated bias than the other methods (Fig. I.10c), leading to a residual negative bias with respect to the validation data that is similar in magnitude to the original positive bias (Fig. I.10d). L09 also does not offer a correction for the sparse data below 250 m. While the residual bias with depth is smallest using the EANN method, the GR10 method performs almost as well (Fig. I.10d).

Examining the residual MBT biases with respect to the validation dataset as a function of both depth and year, we find that the structure of the uncorrected MBT bias remains

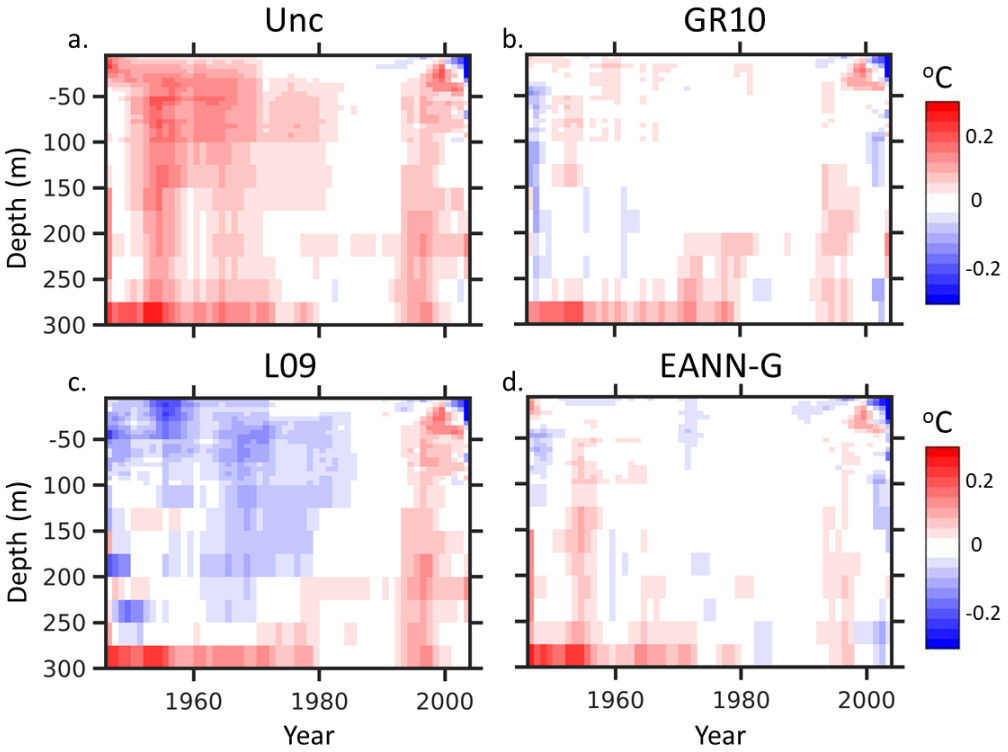


Fig. I.11 Global residual MBT biases with respect to our validation dataset (corrected MBT - Ref_{VAL}) as a function of both depth and year for uncorrected MBT data (Unc) (a), data corrected with the GR10 method (b), data corrected with the L09 method (c), and data corrected with the EANN-G method (d).

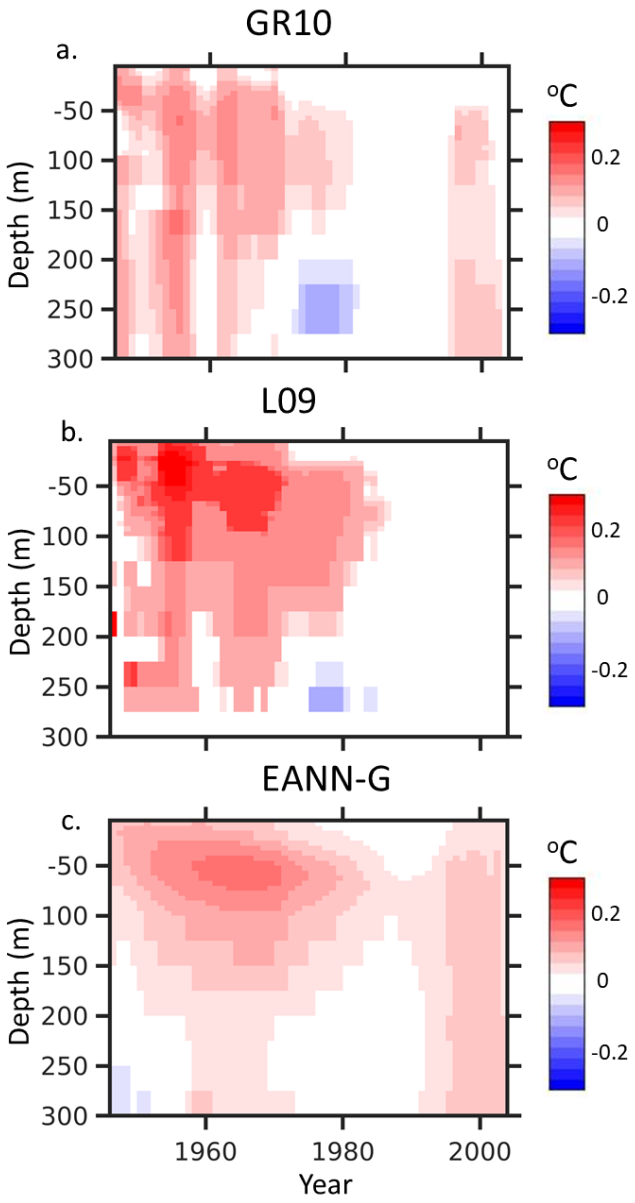


Fig. I.12 Median of the global extrapolated MBT bias (uncorrected MBT - corrected) as a function of depth and year for data corrected with the GR10 method (a), data corrected with the L09 method (b), and data corrected with the EANN-G method (c).

mostly positive throughout its period of use (Fig. I.11a). The GR10 calibration (Fig. I.11b) largely reduces these positive temperature biases but slightly overcorrects biases in the 1940s and slightly under-corrects biases in the 1990s. By comparison, the L09 method (Fig. I.11c) overcorrects MBT biases throughout much of the water column but leaves the biases after 1990 essentially untouched. Similar to the GR10 method, the EANN-G calibration (Fig. I.11d) removes the majority of the systematic bias but under-corrects biases to some extent in the 1950s and 1990s.

Similar patterns appear when examining the extrapolated biases for these methods as a function of both depth and year. The GR10 extrapolation (Fig. I.12a) has a somewhat striated pattern throughout the water column prior to 1975 and suggests no bias in the MBT data during 1980s. The L09 method (Fig. I.12b) produces a maximum extrapolated bias in the mid-

1950s at roughly 50 m, which also roughly coincides with the maximum in the original MBT bias with respect to the validation dataset (Fig. I.11a). A second strong positive bias in the L09 extrapolation occurs in the mid-1960s (Fig. I.12b). The EANN-G method (Fig. I.12c) also produces a maximum extrapolated bias around 50-100 m, but the pattern is much smoother across the 1950s and 1960s than the other methods. EANN-G also does not infer a large positive bias below 200 m in the 1950s, nor does it indicate negative MBT biases around 250 m in the late 1970s, unlike the other two methods (Fig. I.12c).

A recently released study by Gouretski and Cheng (2020) further examined the MBT bias and found similar results regarding the performance of the GR10 and L09 correction schemes as we have here. They also indicate in their study that country of origin for the MBT probes has an impact on the bias history, a factor which we have not considered, and the new corrections they present in their study take an empirical approach when considering the other known sources of bias. After applying some of the metrics from Cheng *et al.* (2018) to the MBT dataset, they found that their corrections outperformed other available methods; our method was not yet available for comparison. Additionally, they concluded that the L09 correction does not reduce the total bias compared to the original uncorrected MBT data, but that the GR10 method is acceptable. Given the performance of L09, the EANN-G method may present an alternative statistical approach to correct the MBT data.

D. Discussion

In this study, we developed and implemented a new approach to correct global systematic temperature biases in mechanical and expendable bathythermograph (BT) datasets using an ensemble of artificial neural networks trained on global data (EANN-G),

and another using probe-specific data (EANN-P). Our method offers a simple correction for the total time-variable BT bias that is explicitly dependent on a combination of year, depth, and water temperature. Additionally, we compared the performance of this method, on both global and basin scales, to that of several popular methods (Levitus *et al.*, 2009 and Cheng *et al.*, 2014 for XBT and Levitus *et al.*, 2009 and Gouretski and Reseghetti, 2010 for MBT) using some of the metrics proposed by Cheng *et al.* (2018). Finally, we examined differences in how these calibration methods extrapolate the bias both spatially and temporally.

Our results demonstrate:

1. The use of EANN-G and EANN-P methods greatly reduces time, depth, and latitudinal components of the XBT bias, performing on par with the best available methods when compared to an independent validation dataset. Based solely on performance it is difficult to distinguish EANN-G and EANN-P, except for their performances correcting the biases of certain probe types. There is some benefit to using a probe-level correction, however the XBT dataset is skewed towards only a few probe types and the metadata is not complete. Both global and probe-level methods would likely need to be incorporated into studies examining OHC in order to fully characterize the uncertainty due to the choice of XBT bias correction.

2. Both the EANN-G and EANN-P calibrations are simple to implement and, like CH14 for XBT, L09 for XBT in the top 700 m, and GR10 for MBT, can extrapolate well to new BT data in areas where we have complimentary CTD/ OSD data,

indicating that both empirical and statistical approaches to calibrating the global BT data are reasonable avenues.

3. All of the calibration methods considered here offer valid corrections for BT biases on global and basin scales, potentially with the exception of L09 for the MBT, but examinations of the extrapolated biases reveal key distinctions among methods, which will contribute to uncertainty in OHC estimates on intra-decadal and basin scales.

4. The choice of XBT fall rate equation (FRE), either opting to use the original manufacturer equation (MFR) or the H95 equation, impacts the extrapolated XBT bias correction, even for what would otherwise be the same calibration method. Our method provides calibrations for both FRE, and we recommend that both corrections be incorporated into global OHC studies in order to fully characterize the uncertainty arising from correcting for historical XBT biases. Considering the effects that different XBT calibrations have on deep OHC may be especially important given the negative impact that the use of the H95 FRE has on the XBT bias for the 700-1800 m depth interval.

There remains room for improvement in both empirical and statistical approaches to reducing biases in historical BT data, and further refinements to existing methods (including the one presented here) could be developed by more deeply examining underlying contributors to the bias, which perhaps can be gleaned from further laboratory studies, numerical simulations, or a comprehensive examination of the available probe metadata. For example, most calibration methods to date (including the one presented here) have assumed

that BT biases depend on the year of deployment, which does not directly represent the underlying technological, manufacturing, and design changes that ultimately drive the time-varying bias. Without additional refinement, existing BT calibration methods are likely only suitable for global or perhaps probe-level calibrations, and corrections to individual casts should not necessarily be considered reliable at this time. For certain geographic regions and the deep ocean especially, where there may not be enough direct data to fully characterize the problem, the community may need to be satisfied with an ensemble of calibration methods that at least impose maximum bounds on our uncertainty.

II. Full-Depth Ocean Heat Content and Earth's Energy Imbalance

Material From: 'Bagnell and DeVries, 20th century cooling of the deep ocean contributed to delayed acceleration of Earth's energy imbalance, Nature Communications, published [2021], [Springer Nature]'

A. Introduction

Global climate change is driven by imbalances in Earth's energy budget due to both anthropogenic and natural influences (Myhre *et al.*, 2013; Meyssignac *et al.*, 2019). Estimating historical changes in Earth's energy imbalance (EEI) is essential for accurately quantifying climate sensitivity to greenhouse gas emissions, benchmarking climate models used in making future climate projections, and for understanding the contribution of natural events and climate patterns to modulating the global climate response to anthropogenic forcing (Meyssignac *et al.*, 2019; Trenberth, 2014). The ocean is currently the largest energy reservoir in the Earth's climate system and is responsible for absorbing and storing more than 90% of the excess heat in the Earth system that results from anthropogenic climate change[2,3,4]. Thus, measurements of the global ocean heat content (OHC) over time provide one of the best ways of estimating historical trends in the EEI (Meyssignac *et al.*, 2019; Trenberth, 2014; von Schuckmann *et al.*, 2016).

Historical changes in global OHC can best be reconstructed from in-situ temperature observations. Over the past 15 years, the Argo program (Roemmich *et al.*, 2019) has deployed thousands of autonomous floats which provide continuous observations of the temperature in the upper half of the ocean, to a depth of 2000 m. This has allowed for a convergence in estimates of OHC over the last fifteen years (Meyssignac *et al.*, 2019;

Roemmich *et al.*, 2019) and increased confidence in calculations of the ongoing EEI in light of independent confirmation from modern satellite observations (Meyssignac *et al.*, 2019; Allan *et al.*, 2014; Loeb *et al.*, 2012). However, several challenges exist for reducing uncertainty in estimates of total ocean warming and extending it over longer time periods. First, the deep ocean below 2000 m remains poorly observed, even during the Argo era, which leads to additional uncertainty on current estimates of total warming. While absolute temperature changes in the deep ocean are small (Purkey and Johnson, 2010), the large volume of the ocean below 2000 m makes it a potentially meaningful contributor to the global heat inventory. Repeat hydrographic sampling indicates that the deep ocean may be warming significantly in some regions (Desbruyeres *et al.*, 2016), particularly the Southern Ocean (Purkey and Johnson, 2010), whereas other regions may still be cooling as a response to cold periods in the past millennium (Gebbie and Huybers, 2019), making it critical to include the heat content of the deep ocean in global estimates of ocean warming. The second issue is that, prior to 2005, data collection was conducted primarily by scientific research vessels and ships of opportunity, leaving areas outside of major trade routes or research transects with few direct observations (Meyssignac *et al.*, 2019; Cheng *et al.*, 2017). This leaves large gaps in the observational record that must be filled in order to estimate OHC.

Several methods have been devised to overcome these gaps in ocean temperature observations and to produce estimates of historical changes in OHC. One common approach applies objective mapping to interpolate the sparse temperature records in space and time (Cheng *et al.*, 2017; Levitus *et al.*, 2012; Ishii *et al.*, 2017). However, while these objective mapping products can reconstruct ocean temperatures back to ~1950, they do not extend below 2000 m due to the sparse sampling at these depths. Dynamical data-assimilation

models offer an alternative approach to objective mapping and provide full-depth estimates of OHC (Boisséson *et al.*, 2018; Palmer *et al.*, 2017), but data sparsity means these models are poorly constrained at depth, leading to large cross-model variance (Palmer *et al.*, 2017). Another approach based on the passive transport of surface temperature anomalies into the interior ocean (Gebbie and Huybers, 2019; Zanna *et al.*, 2019) can also reconstruct full-depth temperature anomalies and OHC changes, but relies on the potentially incorrect assumption of steady-state circulation (Zanna *et al.*, 2019) and is sensitive to the initial condition used in the simulation (Gebbie and Huybers, 2019; Zanna *et al.*, 2019) and to poorly-known surface ocean temperatures dating back several millennia (Gebbie and Huybers, 2019). Finally, statistical methods have been used to detect large-scale trends in the deep ocean temperature from repeat hydrographic sampling (Desbruyeres, *et al.*, 2016), but these have coarse spatial resolution and do not cover the period prior to the mid-1980s. An interpolation product based on in-situ temperature data that covers the deep ocean below 2000 m, allowing for a full-depth OHC estimate, remains crucial to reliably estimating historical changes in EEI (Loeb *et al.*, 2018; Palmer *et al.*, 2011).

Here, we interpolate historical ocean temperature data using an autoregressive artificial neural network (ARANN) to produce a single consistent estimate of the top-to-bottom OHC change for 1946-2019 using in-situ temperature data from the World Ocean Database (Boyer *et al.*, 2018). This approach (Supplementary Figs. II.1-2) adapts an established machine learning method to perform an iterative autoregression that adjusts spatio-temporal correlation scales over time from the in-situ temperature data itself, and effectively propagates information from well-sampled times and regions to more sparsely-sampled areas to produce global maps of temperature anomalies at roughly annual resolution

(Supplementary Fig. II.3). This approach is robust to sparse data, allowing our estimates of OHC change to be extended below 2000 m to the seafloor (Fig. II.1). We have tested the method on datasets from two ocean models used in the Climate Model Intercomparison Project Phase 6 (CMIP6) (Tatebe and Watanabe, 2018; Voldoire *et al.*, 2019), demonstrating the ability to accurately reconstruct OHC changes on both global and basin scales (Supplementary Figs. II.4-7) at all depths of the ocean, and to recreate modeled temperature anomalies at spatial scales of ~1000 km or larger (Supplementary Figs. II.8-11), even in the presence of realistic geophysical noise that is present in the observations but not the models (Supplementary Figs. II.4-12). We apply the ARANN in an ensemble approach designed to take into account sources of uncertainty arising from the sparse distribution of temperature observations (Meyssignac *et al.*, 2019; Cheng *et al.*, 2017), documented instrument biases (Levitus *et al.*, 2009; Cheng *et al.*, 2014; Gouretski and Reseghetti, 2010; Bagnell and DeVries, 2020) (Supplementary Fig. II.13), and choice of reference climatology used to define the temperature anomalies (Cheng and Zhu, 2015; Boyer, 2016) (Supplementary Fig. II.14).

Four instrumental bias corrections and six decadal climatologies are combined with random selections of temperature data to produce the 240 ensemble members used in this study. This ensemble is used to assess the uncertainty of our OHC reconstruction and provide bounds on our estimates of ocean warming. All estimated warming rates come from fitting a linear trend to the mean ARANN OHC estimate and uncertainties in these rates are calculated by taking 2 standard deviations across all ensemble members. Where ranges are given, these compare the mean ARANN estimate to other products. For simplicity, OHC estimates from other studies are not plotted with their respective confidence intervals, as the

methodology for calculating these varies by study, but they generally possess uncertainty levels similar to those provided by the ARANN.

B. Methods

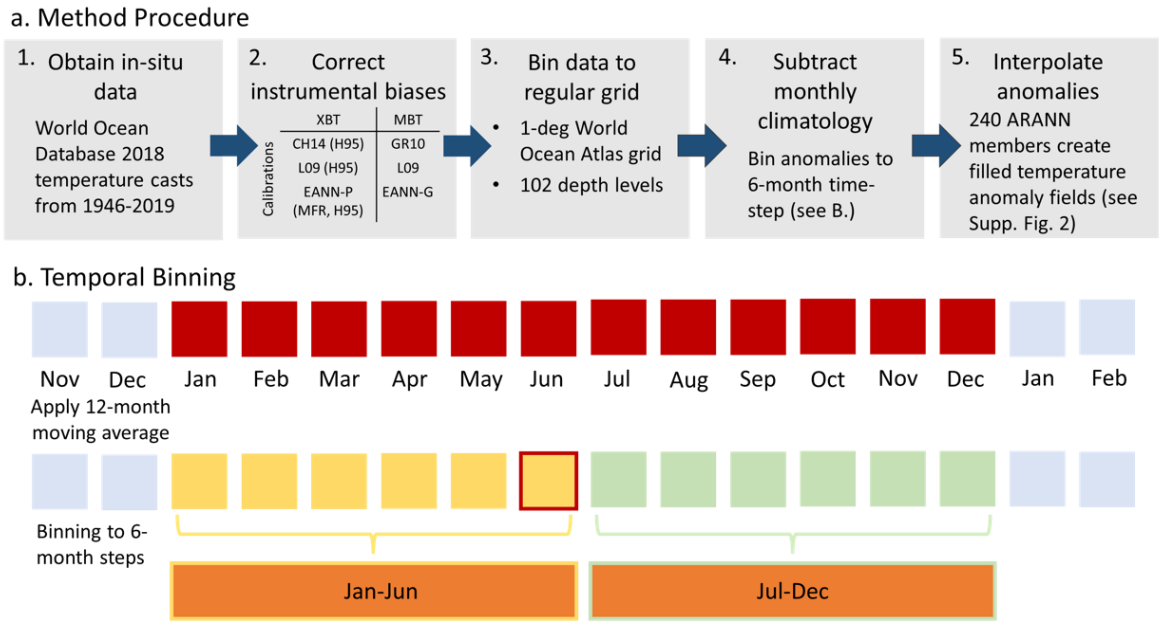
B.1 Observational Datasets and Data Processing

We used a dataset of historical ocean temperature observations from 1946-2019 consisting of individual temperature casts from the World Ocean Database (WOD) 2018 (Boyer *et al.*, 2018) (Supplementary Fig. II.1a, Step 1). We used temperature casts from multiple instruments, including mechanical bathythermographs and expendable bathythermographs (MBT and XBT), ocean station data (OSD), conductivity-temperature-depth (CTD) profiles, autonomous profiling floats (PFL), and autonomous pinniped bathythermographs (APB). We quality controlled these data using the strictest quality control procedures of the World Ocean Database (WOD), which exclude any data with a flag other than 0. Additionally, we excluded casts that have less than 5 discrete temperature samples, while also requiring that one of these samples occurs in the top 100 m. These extra steps reduce the possibility of including data from casts that have had much of their data removed by flags, casts that had insufficient samples to be properly quality controlled to begin with, or casts where much of the data was removed in the near-surface due to an excessive vertical temperature gradient (0.7 ^{o}C m^{-1} for the WOD).

Next, individual casts were linearly interpolated to the 102 standard depths of the World Ocean Atlas[56] (WOA) grid. Before the data from the various instrument types was combined, systematic errors in the bathythermographs were corrected (Supplementary Fig. II.1a, Step 2) using some of the best available calibration methods (see Supplementary Fig. II.13). Then, these data were binned to the WOA grid at monthly resolution based on the

year and calendar month of their collection (Supplementary Fig. II.1a, Step 3). We binned using the median of observations within each grid cell, to reduce the influence of outliers.

After binning the temperature data to the WOA grid, we subtracted a monthly temperature climatology to create a field of monthly temperature anomalies (Supplementary Fig. II.1a, Step 4). For this step, we used one of six WOA decadal climatologies covering years [1955-64, 1965-74, 1985-94, 1995-2004, 2005-2017]. These climatologies are monthly in the top 1500 m and seasonal below that. The choice of climatology produces a difference in the mean total OHC change from 1946-2019 of up to 49 ZJ. The impact of climatological choice is small for the final three decades of the OHC record but has a more significant impact on the global OHC estimate in the deep ocean and further back in time (see Supplementary Fig. II.14).



Supplementary Fig. II.1 (a) Summary of the data processing steps for creating temperature anomaly fields from raw temperature observations, and (b) expansion of Step 4 in (a) demonstrating the procedure for smoothing and binning these temperature anomalies to a 6-month time-step.

Finally, we smoothed the resulting monthly anomaly maps using a 12-month moving average, and then binned the smoothed monthly anomaly maps to 6-month time intervals that span either Jan-Jun and Jul-Dec (Supplementary Fig. II.1b), or Apr-Sep and Oct-Mar. The choice of either Winter/Summer-centered or Spring/Fall-centered anomalies also enters into our ensemble, but has little impact on the final results. The end result is 148 three-dimensional temperature anomaly fields spanning 1946-2019 at 6-month time intervals. Each anomaly field (except for the first and last) contains temperature data from within an

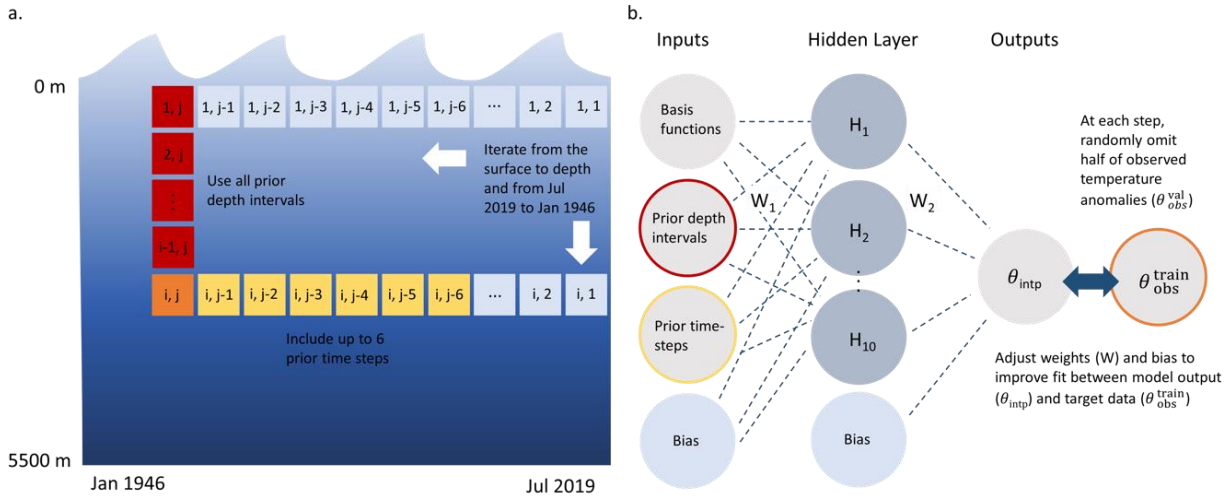
18-month window, with more weight given to observations within each 6-month window. Thus, our final OHC estimates resolve OHC variability at roughly annual resolution.

B.2 Description of the interpolation process

The inputs to the ARANN include a set of basis functions that are used to approximate the spatial autocorrelations of the temperature anomalies. These basis functions consist of a set of sinusoids, which are used to approximate the correlation length scales found in the spatial maps of the temperature anomalies. To obtain these basis functions, we averaged the gridded temperature anomalies over the top 700 m for the years 2005-present and took the first 6 principal components of the resulting anomaly map, which explain > 90% of the variance, in both the meridional and zonal directions. We also found that we obtained very similar principal components if we used the 700-2000 m layer instead of the 0-700 m layer. Using these principal components, we then estimated the periods of their autocorrelations, which then became the periods of our sinusoids. In the meridional direction, the 6 periods of the sinusoids are [360 180 90 60 45 30] degrees and in the zonal direction they are [180 120

90 60 45 22.5] degrees. Both the cosine and sine functions are used for each period, leading to a total of 24 basis functions. Due to how these sinusoids are constructed, they each represent a basis function that is merely a 2-dimensional array of numbers between -1 and 1 with a characteristic length scale ranging from roughly 1,000 km to 18,000 km.

With these basis functions, the ARANN can reconstruct temperature anomalies on horizontal scales of roughly 500-1,000 km. The ARANN is not designed to capture small-scale features and the OHC changes reconstructed by this method should be interpreted on scales of ~1,000 km or larger. Indeed, when the ARANN is applied to temperature anomaly fields from global ocean models, the residuals contain features on the order of several hundred km even where the spatial sampling is relatively dense and evenly distributed (see Supplementary Figs. II.8-11). The presence of geophysical noise and sparse sampling further reduces the scales that we are able to resolve.



Supplementary Fig. II.2 (Schematic of (a) the iterative process of the ARANN method, which propagates information from near the sea surface and in modern times to less sampled time periods and depths, and (b) the architecture of the ARANN, which uses sinusoidal basis functions, prior time-steps, and prior depth levels to estimate temperature anomalies at the current time-step.

While the ARANN uses sinusoidal basis functions to capture spatial variability in the temperature anomalies, there are no explicit variables for time or depth. Instead, the ARANN uses a sweep from higher data coverage to lower data coverage in order to capture the depth and temporal dependencies (Supplementary Fig. II.2a). This is done by slicing the 3-dimensional temperature anomaly fields into two-dimensional "chunks" and applying a separate ANN to each consecutive 2-dimensional temperature anomaly field, using temperature anomalies from previous depth- and time-slices as additional input features. Vertical mixing is important in certain regions and contributes to deep water formation, so surface warming of the ocean would be expected to display some imprint on the layers below. Because the ARANN is iterative, it optimizes for each depth interval during its sweep from the surface to seafloor. Relationships identified by the ARANN at depth will therefore evolve from those found at the surface. In this way, the ARANN mimics in a parameterized way the circulation and mixing processes that connect the surface and deeper layers of the ocean, and encodes some "memory" to capture the temporal evolution and persistence of temperature anomalies. The ARANN has limited memory, or what is sometimes called finite impulse, in that it does not retain knowledge of all prior time-steps but only those that have occurred most recently. This simplification mimics the fact that the correlation between temperature anomalies at a given location in the ocean diminishes over time as heat is circulated and mixed away. Since the autocorrelations in the anomalies evolve over time, the ARANN refines and evolves the relationship between temperature anomalies and the input fields at each step.

The iterative process is described in detail below. At each step of this process, a new ARANN is developed, trained, and validated, following the procedure described in the next

section.

1. Starting with the most recent 6-month period in the time-series (i.e. Jul-Dec 2019), we begin by randomly choosing the size of our initial depth window, which consists of between two and six depth layers. This corresponds to a depth interval of 10-150 m above 300 m. Starting with the depth window closest to the surface, we randomly select 50% of the data within that depth window for training the ARANN, setting aside the remaining 50% for validation (a similar 50/50 training/validation split is maintained through all subsequent steps). We then interpolate the temperature anomalies in this initial near-surface depth window using only the 24 sinusoidal basis functions as inputs to the ARANN. This is iterate (1,1) in Supplementary Fig. II.2a.

2. Moving down to the next depth interval, we again choose a random depth window consisting of between two to six depth layers. The data sampling and interpolation within this depth window are repeated as in Step 1, using the interpolated temperature anomalies from the prior depth window as an additional input to another ARANN (again, a new ARANN is trained at each iteration). This is iterate (2,1) in Supplementary Fig. II.2a.

3. Step 2 is repeated iteratively, moving down an additional depth level at each iterate, and using the interpolated temperature anomalies from all prior depth windows as additional inputs to the ARANN interpolation at each depth level. This means that at each new depth level, there are 24+i inputs to the ARANN, where i is an index corresponding to the number of vertical depth intervals used in the interpolation.

Below 300 m, the depth window is expanded to between six and twelve depth layers (corresponding to a depth interval of 150-1200 m) to ensure that adequate amounts of data are available to the network. Randomizing the size of the depth window at each iteration ensures the model is not fixed to specific depth intervals and will eventually produce a smooth transition in the vertical gradient. Steps 1-3 complete the initial sweep over depth levels and fills in iterates (1,1) to (i,1) in Supplementary Fig. II.2a. The total number of depth intervals used from the surface to the seafloor ranges from 11 to 27 intervals, depending on the (random) choice of depth layers used at each interval.

4. Steps 1-3 are then repeated iteratively, marching backwards in time at 6-month intervals through the time series. At each time step, interpolated temperature anomaly fields from six prior time intervals (or the maximum number available if less than six) are used as additional inputs to the ARANN. Using temperature anomalies from prior time intervals as additional inputs to the ARANN mimics the temporal autocorrelation of temperature anomalies and allows the propagation of information from well-sampled time periods to more sparsely-sampled period. We do not fix the autocorrelation timescale (other than limiting the inputs to 6 previous time steps, or 3 years), but rather let the network decide at each iteration how much weight to put on previous iterates when interpolating anomalies at each time-step. At the conclusion of Step 4, we have 148 gap-filled three-dimensional datasets of temperature anomalies at 6-month resolution from 1946 to 2019.

5. After running the model backwards from 2019 to 1946, steps 1-4 are repeated, this time running time forwards from 1946 to 2019. In this forward sweep, we use

interpolated temperature anomalies from three prior time intervals and three subsequent time intervals as additional inputs to the ARANN at each time step. This forward sweep helps to further propagate information through the network and is particularly helpful for smoothing results from the transient stage at the beginning of the backward run when the model had less information on temporal autocorrelations. This is most important in the abyssal ocean where temporal autocorrelations are longer and data is sparser.

6. Steps 1-5 are repeated 10 times for each possible climatology, generating an ensemble of 60 ARANNs for interpolation, from which we can derive uncertainties related to data sampling, interpolation, and climatology.

7. Steps 1-6 are repeated 4 times, each time applying a different calibration method to the XBT and MBT data (see Supplementary Fig. II.13).

An example of the resulting temperature anomaly field from this mapping method (Supplementary Fig. II.3) reveals how our method takes the original binned temperature anomalies and interpolates the data to produce filled anomaly fields for a single realization of the ARANN. In the upper 50 m for the year 1960, the observational sampling neglects much of the southern hemisphere, but by leveraging information from prior iterations, the ARANN method produces a smooth product that fills gaps that would not be captured by traditional objective mapping. In the year 2010, the upper 50 m has much more regular sampling, and the ARANN method captures the large-scale patterns while smoothing over small-scale features. For 900-1100 m in the year 1960, temperatures are very sparsely sampled, but the ARANN method still produces large-scale regional structures that would

not be captured by traditional objective mapping. The ARANN also smooths over most of the small-scale noise in the observations at these depths. This noise is quite apparent when considering the anomalies fields for 900-1100 m in the year 2010. The ARANN interpolation of this data captures the large, near basin-scale spatial patterns, while ignoring most mesoscale to sub-mesoscale patterns, which can be seen in the residuals (ARANN – Obs.) of the temperature anomalies.

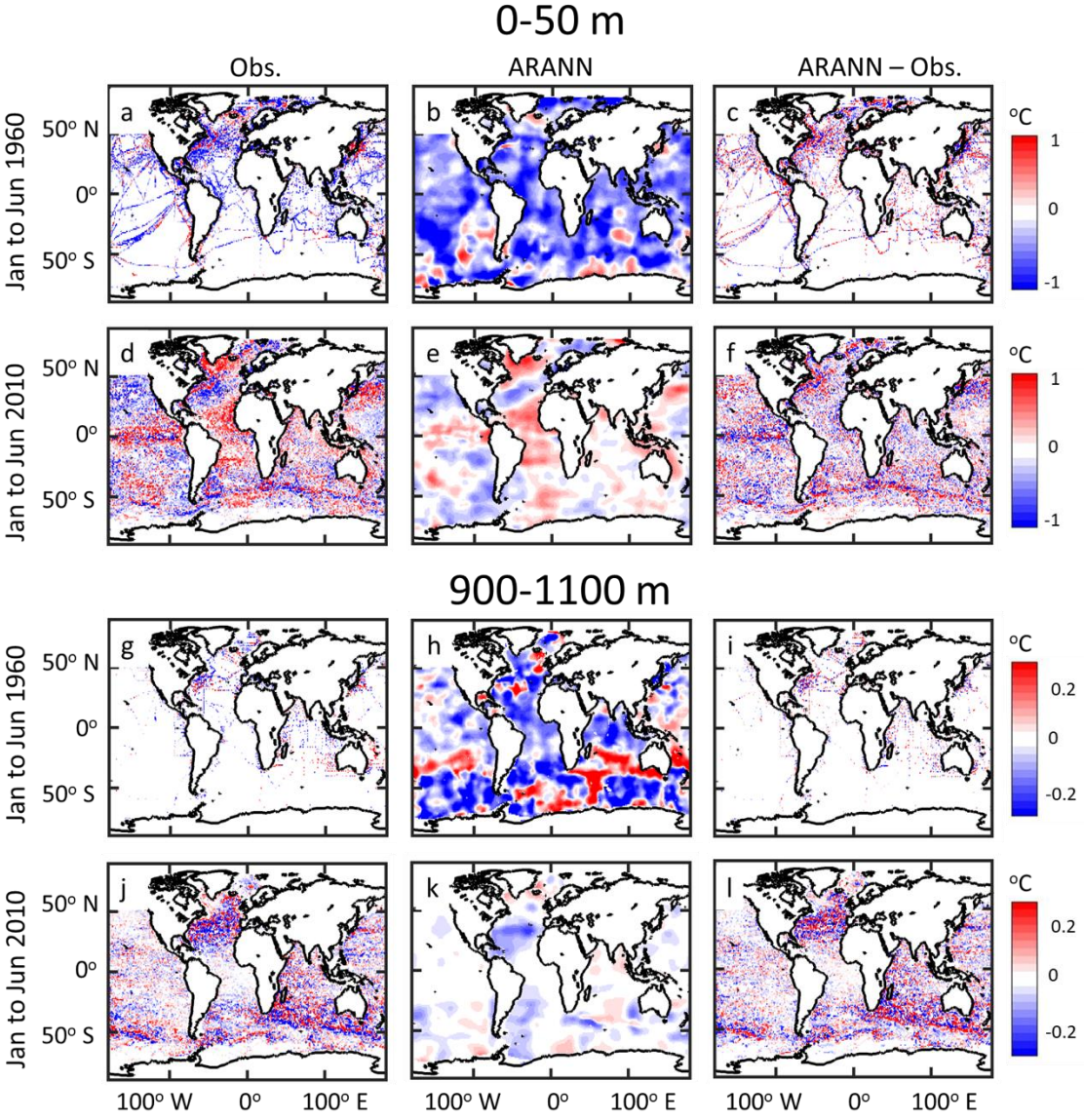
B.3 Architecture of the ARANN and method of solution

Each step of the individual autoregressive artificial neural network (ARANN) (Supplementary Fig. II.2b) consists of an input layer that contains 24 naïve basis functions (*B*), as well as filled temperature anomalies for six adjacent time steps time steps (if available) (θ_t), and anomalies for all prior depth intervals for the current time step (θ_z). Each input is organized as a vector with a length *n*, equal to the number of spatial grid points in the gridded temperature anomaly fields. In total, there are m input fields, where m is between 24 and 46 (24 basis functions, 6 filled temperature anomaly fields from adjacent time-steps, and up to 16 filled temperature anomaly fields from prior depth intervals). These inputs are organized as an array *I* with size (*n* x *m*),

$$I = [B \quad \theta_t \quad \theta_z] \, . \qquad (1)$$

In the ARANN, the input "layer" connects to a single hidden "layer" with 10 nodes, producing a network with (*m* x 10) input "weights" organized as an array (W_1). We selected this number of nodes by experimenting with adding more free parameters, the weights, until

the performance of the ARANN on internal validation sets comprised of data sampled during periods of high data sparsity (pre-2005) no longer improved. The values for the hidden layer (*H*) produced by the ARANN are



Supplementary Fig. II.3 (Top panels) Temperature anomalies for the 0-50 m depth interval for (a-c) January to June 1960 and (d-e) January to June 2010 from (a,d) observations, (b,e) a single realization of the ARANN interpolation of the observed field, and (c,f) the residuals between the ARANN product and the original observations. (Bottom panels) Temperature anomalies for the 900-1100 m depth interval for (g-i) January to June 1960 and (j-l) January to June 2010 from (g,j) observations, (h,k) a single realization of the ARANN interpolation of the observed field, and (i,l) the residuals between the ARANN product and the original observations.

$$H = F(I \cdot W_1 + b_1), \qquad (2)$$

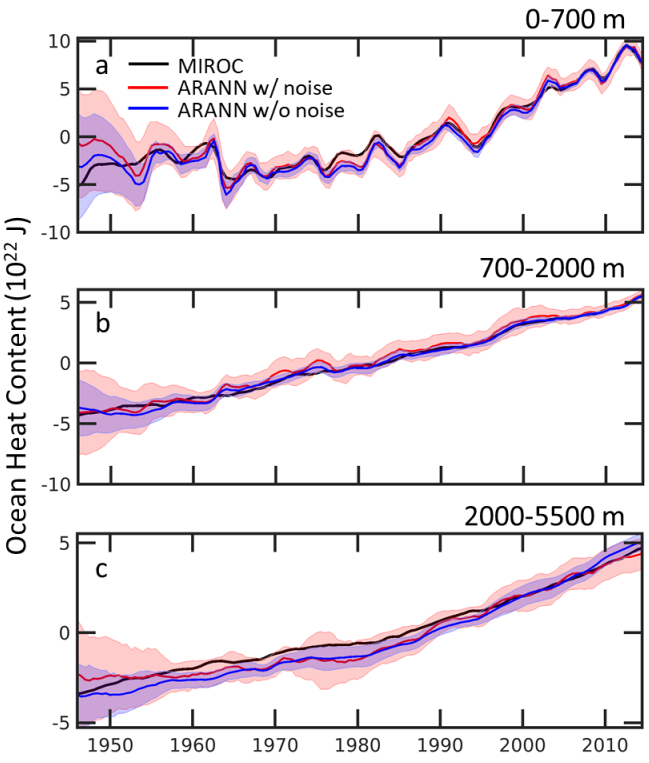
where *F* is the "transfer function" used to propagate information in a non-linear fashion through the network, and b_1 is a [1 x 10] array of "biases", with all values in a given column being identical. We use the hyperbolic tangent as the transfer function, which is commonly employed for interpolation (Krasnopolski *et al.*, 1995; Maier and Dandy, 2001; Hasni *et al.*, 2012) by fully-connected feedforward networks like the ARANN. The output "layer" connects to the hidden layer using another [10 x 1] array of weights W_2 and a single bias [1 x 1] b_2 to produce the predicted temperature anomalies for the interpolation (θ_{intp}),

$$\theta_{\text{intp}} = H \cdot W_2 + b_2 \, . \qquad (3)$$

In all, each network has (10 x m) + 21 free parameters, representing the (10 x m) weights W_1 of the input layer, the 10 bias terms of the input layer, and the one bias term and the 10 weights W_2 of the output layer. These free parameters are iteratively adjusted to achieve a minimum of a cost function that measures the mean sum of squares difference between the interpolated temperature anomalies (θ_{intp}) and the observed temperature anomalies in our training dataset ($\theta_{\text{obs}}^{\text{train}}$),

$$\text{cost} = \frac{\sum_{k=1}^{N}\left(\theta_{\text{intp}}^{k} - \theta_{\text{obs}}^{\text{train},k}\right)^2}{n}, \qquad (4)$$

where *N* is the number of observations within the training dataset at each iteration. As discussed in the prior section, the training dataset consists of a random 50% selection of all available data at each iteration. For the back-propagation algorithm in our network, which iteratively updates the values of the weights to minimize the cost function, we chose the Levenberg-Marquardt algorithm (Marquardt, 1963) due to its improved performance at



Supplementary Fig. II.4 (Global ocean heat content estimates based on an original MIROC CMIP6 climate model simulation (black), the ARANN reconstruction based on modeled temperature anomalies reduced to observational sparsity (ARANN w/o noise, blue), and the ARANN reconstruction based on modeled temperature anomalies with additional noise added to mimic geophysical noise occurring in the observations (ARANN w/ noise, red) for the depth intervals (a) 0-700 m, (b) 700-2000 m, and (c) 2000-5500 m. Error bars for the ARANN reconstructions are the 2 standard deviation range across 30 ensemble members.

reducing the error between predictions and observations versus other common algorithms such as gradient descent (Hagan and Menhaj, 1994).

For each network, we withhold the 50% of data not selected for the training set as validation of the network. This validation dataset ($\theta_{\text{obs}}^{\text{val}}$), is used to prevent overfitting of the network, which occurs if the network is over-trained on a dataset so that it cannot extrapolate well when presented with new data. We use an early stopping technique (Prechelt, 1998) to avoid overfitting, whereby the interpolated temperature anomalies are periodically checked against the validation dataset, and training is terminated when the root-mean squared error of the interpolated temperature anomalies against the validation dataset begins to decrease.

B.4 Required smoothness constraints

After each time-step, we perform a further check on the interpolated temperature anomalies to ensure that they satisfy certain vertical and temporal smoothness conditions in

their basin-averaged temperature anomalies. These checks are performed individually for the Atlantic, Pacific, Indian, and Southern Ocean basins (the Arctic is not considered due to data sparsity) using information about the natural rates of change derived from observations during the Argo Era (defined as 2005 onwards). These checks are only performed for time-steps prior to 2005. The boundaries for these ocean basins are the same as defined in the main text and utilize the WOA 1-degree mask (Garcia *et al.*, 2019).

First, we calculate the change in basin-averaged temperature anomaly at each depth level from the current temporal iteration to the previous temporal iteration, $\Delta\theta_t(\text{basin}, z)$, from the interpolated temperature anomalies during the well-sampled Argo period (2005-2019). We also calculate the difference in basin-averaged temperature anomaly at each iteration from one depth level to the previous depth level, $\Delta\theta_z(\text{basin}, z)$, for the Argo period. This yields 28 values of $\Delta\theta_t$ and $\Delta\theta_z$ for each basin and depth level. We then require that $\Delta\theta_t(\text{basin}, z)$ and $\Delta\theta_z(\text{basin}, z)$ for the current iteration not exceed the 3rd standard deviation (σ) of $\Delta\theta_t(\text{basin}, z)$ and $\Delta\theta_z(\text{basin}, z)$ during the Argo period. That is,

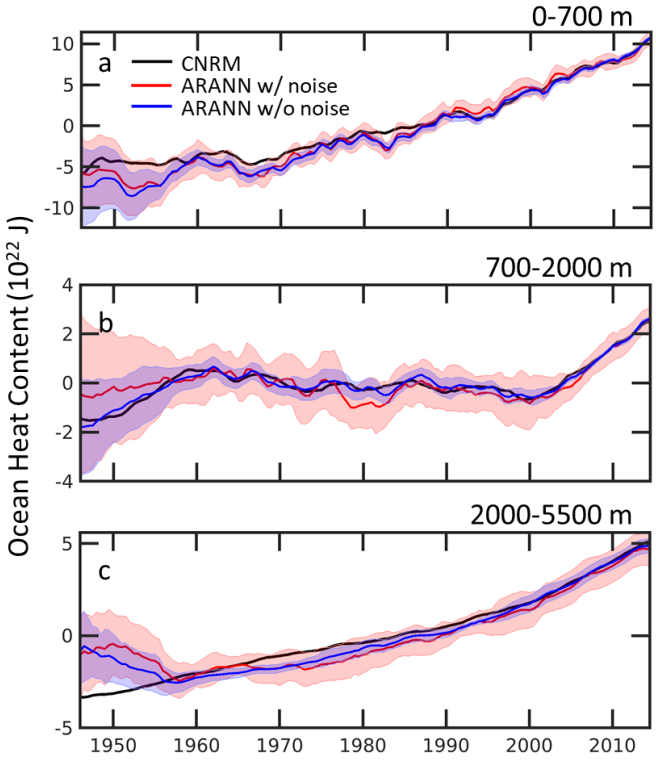
$$\Delta\theta_t(\text{basin}, z)^{\text{current}} < 3\sigma\left[\left(\Delta\theta_t(\text{basin}, z)\right)^{\text{Argo}}\right], \qquad (5)$$

$$\Delta\theta_z(\text{basin}, z)^{\text{current}} < 3\sigma\left[\left(\Delta\theta_z(\text{basin}, z)\right)^{\text{Argo}}\right]. \qquad (6)$$

If a network does not pass both time and depth constraints for all ocean basins, then that network is rejected and the network must start over at the same time-step and depth interval. To allow for some additional flexibility, in the event that a network is rejected more than 5 times in a row, the run that produced the lowest average exceedance of these smoothing constraints is accepted and the procedure continues as usual. This is an uncommon occurrence, appearing only a few times in a single run from 2019 to 1946, but it most often

occurs due to sharp changes in the temperature anomalies of the deep Pacific during the early 1970s.

B.5 Validating the mapping method

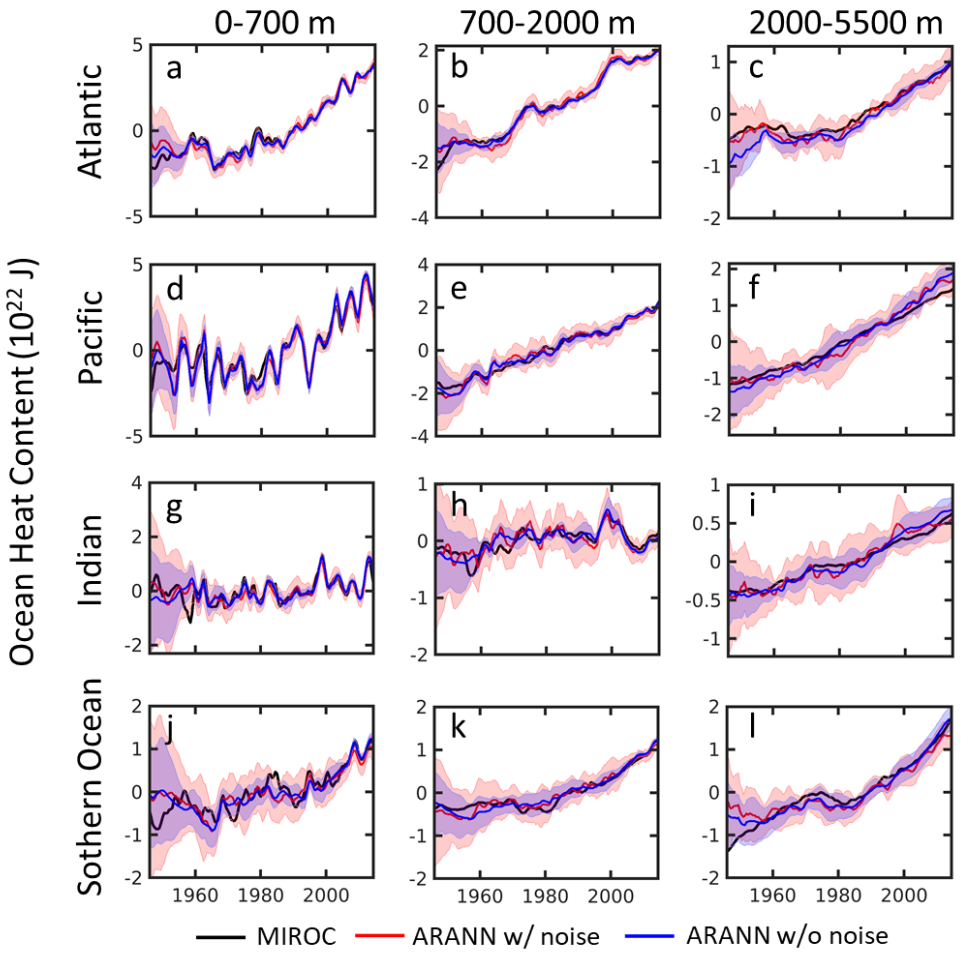


Supplementary Fig. II.5 Global ocean heat content estimates based on an original CNRM CMIP6 climate model simulation (black), the ARANN reconstruction based on modeled temperature anomalies reduced to observational sparsity (ARANN w/o noise, blue), and the ARANN reconstruction based on modeled temperature anomalies with additional noise added to mimic geophysical noise occurring in the observations (ARANN w/ noise, red) for the depth intervals (a) 0-700 m, (b) 700-2000 m, and (c) 2000-5500 m. Error bars for the ARANN reconstructions are the 2 standard deviation range across 30 ensemble members.

One of the major challenges faced by any mapping method for estimating OHC changes are the presence of sampling biases and significant geophysical noise in the observed dataset. Sampling biases arise from the sparse and irregular sampling of ocean temperature observations. Due to the relative difficulty of retrieving observations from remote regions of the ocean, sampling has historically favored coasts, the Northern Hemisphere, and depths shallower than 700 m, leaving much of the ocean unobserved. Given that temperature sampling in the historical record is also heavily biased towards the Argo era (the period after 2005), our method seeks to leverage the autocorrelation of temperature anomalies by using the

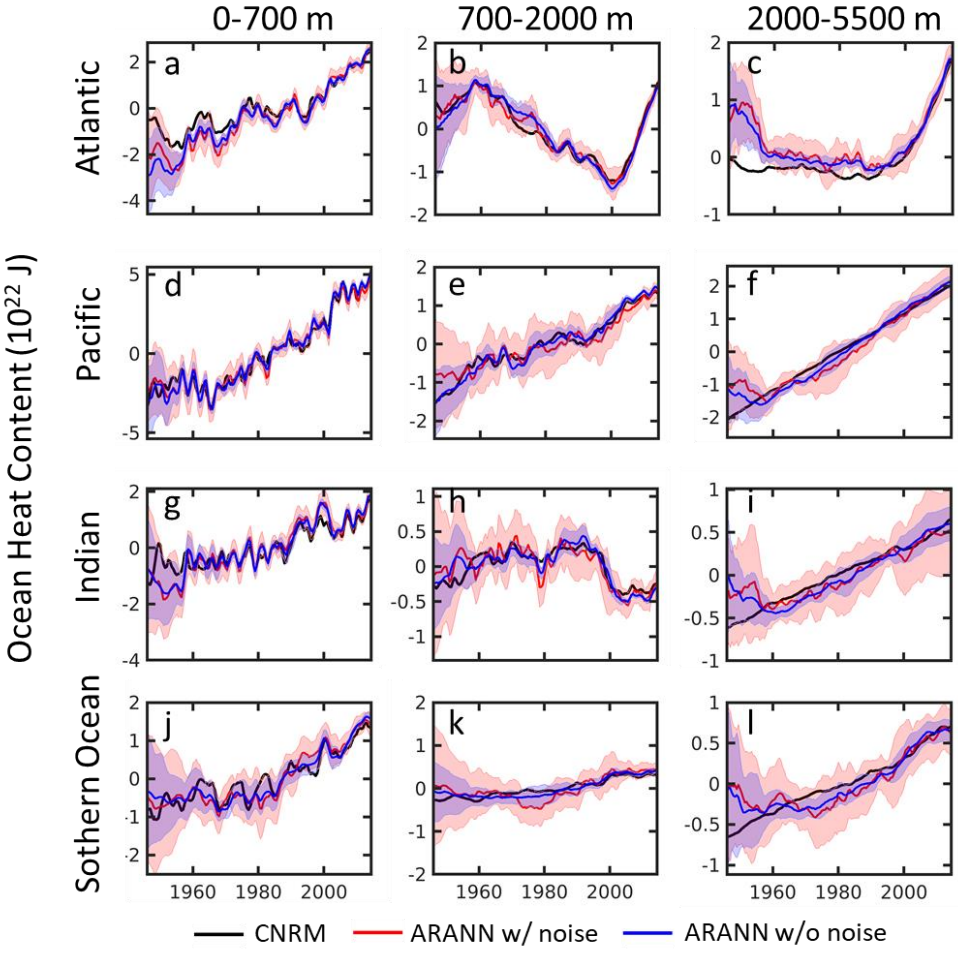
well sampled near-surface and present periods to predict the anomaly fields for the ocean interior at depths greater than 2000 m and for years prior to 2005 when ocean temperature data is sparser.

To demonstrate that our mapping method is robust to irregular sampling coverage, we



Supplementary Fig. II.6 Basin-scale ocean heat content (OHC) anomalies based on an original MIROC CMIP6 climate model run (black), the ARANN reconstruction based on modeled temperature anomalies reduced to observational sparsity (ARANN w/o noise, blue), and the ARANN reconstruction based on modeled temperature anomalies with additional noise added to mimic geophysical noise occurring in the observations (ARANN w/ noise, red). OHC is presented for the Atlantic Ocean at depth intervals (a) 0-700 m, (b) 700-2000 m, (c) 2000-5500 m; for the Pacific Ocean at depth intervals (d) 0-700 m, (e) 700-2000 m, (f) 2000-5500 m; for the Indian Ocean at depth intervals (g) 0-700 m, (h) 700-2000 m, (i) 2000-5500 m; and for the Southern Ocean at depth intervals (j) 0-700 m, (k) 700-2000 m, (l) 2000-5500m. Error bars for the ARANN reconstructions are the 2 standard deviation range across 30 ensemble members. Ocean basins are defined using the World Ocean Atlas mask, with the Southern Ocean considered everything south of 50^{o} S.

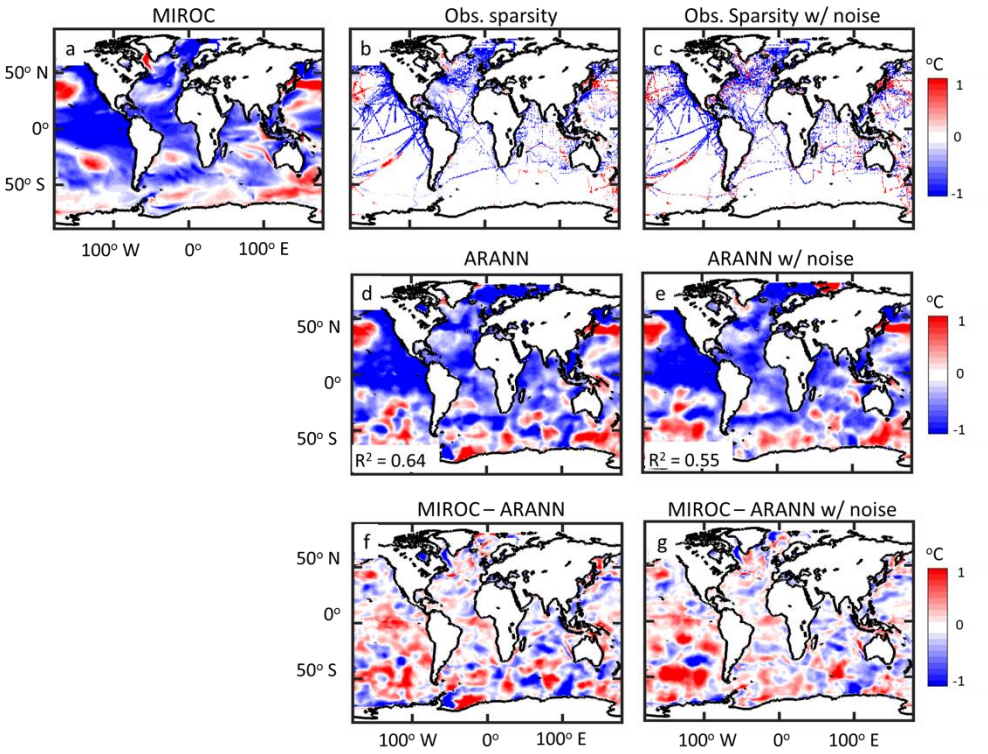
used the sparsity of the combined instrumental datasets to decimate simulated temperature fields from individual historical runs of the MIROC (r1i1p1f1) (Tatabe and Watanabe, 2018) and CNRM (r1i1p1f2) (Voldoire *et al.*, 2019) CMIP6 models, interpolated to the WOA grid. These were obtained from the CMIP6 data archive at https://esgf-node.llnl.gov/projects/cmip6/, last accessed on Jul. 16, 2020. The decimated model temperature fields then have the same data sparsity as the observations on the WOA grid.



Supplementary Fig. II.7 Basin-scale ocean heat content (OHC) anomalies based on an original CNRM CMIP6 climate model run (black), the ARANN reconstruction based on modeled temperature anomalies decimated to observational sparsity (ARANN w/o noise, blue), and the ARANN reconstruction based on modeled temperature anomalies with additional noise added to mimic geophysical noise occurring in the observations (ARANN w/ noise, red). Panels and definitions are the same as those defined for Supplementary Fig. II.6.

We generate a monthly climatology by averaging the model data in the period 2005-2014, which we then subtract from the temperature fields to create the temperature anomalies. Then we apply the same 12-month moving average filter and bin the data to 6-month time-steps, as we do with the observations. We next use the ARANN procedure (Supplementary Fig. II.1) to interpolate the decimated model data in an attempt to recreate the original model temperature anomalies. For our test, we ran the ARANN gap-filling backwards from July 2014 to January 1946, then forwards again across the same time interval, producing 30 ensemble members for each CMIP6 model. We found that the OHC calculated from the ARANN-reconstructed temperature anomalies matched the original modeled CMIP6 OHC very well on global and basin scales (Supplementary Figs. II.4-7) with minimal bias that generally did not exceed the uncertainty (2 standard deviations) of our mapping method.

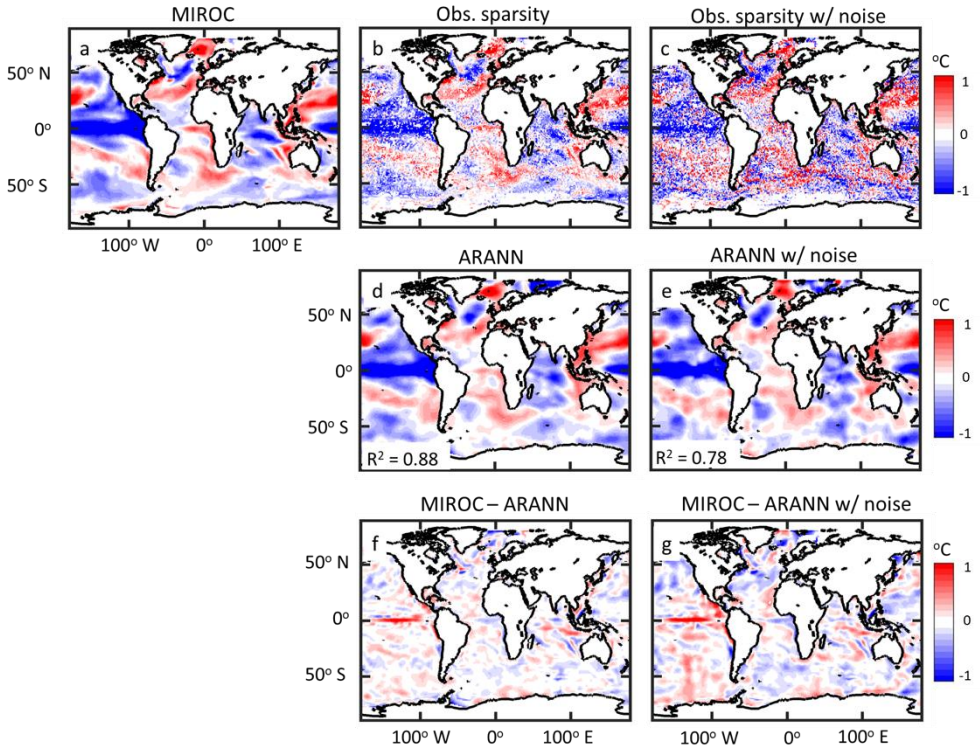
The ability of the ARANN procedure to accurately reconstruct the modeled OHC changes from observed sparsity is reassuring and demonstrates that the ARANN is robust to the presence of large observational gaps. However, the task of accurately reconstructing the modeled temperature anomalies is easier than reconstructing the observed temperature anomalies, since sub-mesoscale processes in the ocean produce additional variability in the temperature observations that is not present in the coarse-resolution models. Comparing the residual error between the ARANN reconstruction and the modeled temperature anomalies (Supplementary Figs. II.8-11, d-e), and the corresponding residual error between the ARANN reconstruction and the observed temperature anomalies (Supplementary Fig. II.3, c, f, i, l), it is clear that the residuals are far larger and more randomly distributed for the



Supplementary Fig. II.8 Temperature anomalies for Jan-Jun 1960 over the 0-50 m depth interval for (a) a MIROC CMIP6 model run, (b) the MIROC model after being reduced to observed sparsity, and (c) the MIROC model at observed sparsity with added geophysical noise. The middle row shows (d) a single realization of the ARANN interpolation of the MIROC temperature anomalies and (e) the ARANN interpolation of the MIROC temperature anomalies with added geophysical noise. The bottom row shows (f) residuals between the original MIROC model and the ARANN interpolation and (g) residuals using an ARANN interpolation with additional geophysical noise added to the modeled temperature anomalies. The middle panels also show the R^2 between ARANN-reconstructed temperature anomalies (shown in each respective panel) and the original MIROC temperature anomalies in panel (a).

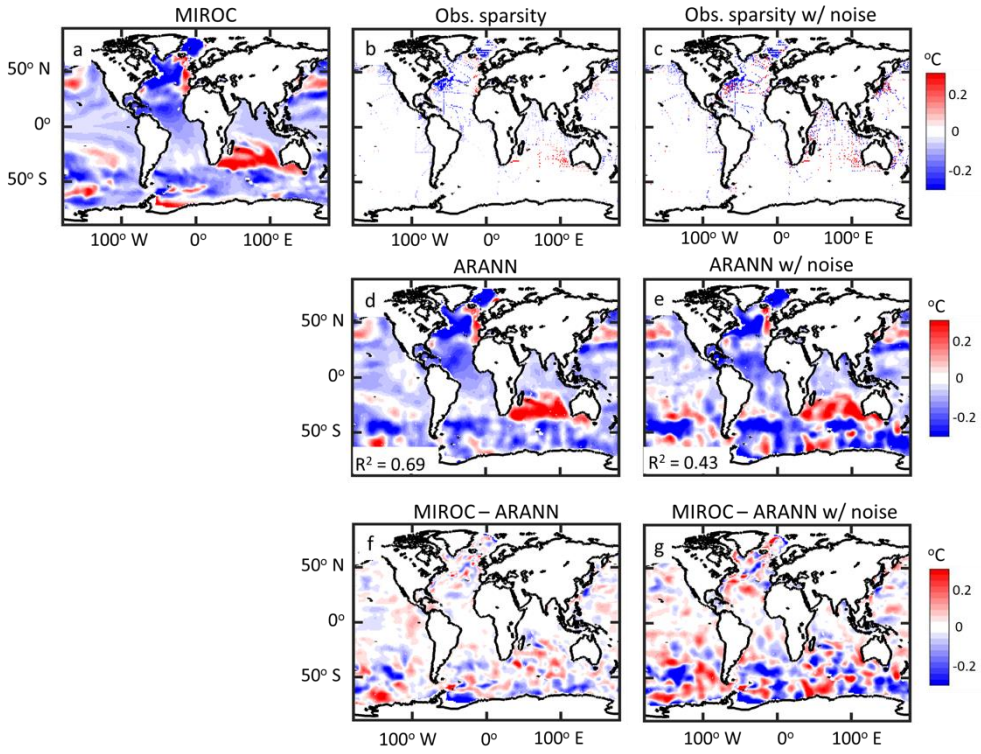
difference between the ARANN and observations, than for the difference between the ARANN and CMIP6 models. This is also demonstrated in Supplementary Fig. II.12 for the residuals between the ARANN reconstruction and the CMIP6 models and observations over time in both the shallow (0-700 m) and deep (700-5500 m) ocean.

To assess the ability of the ARANN to reconstruct the "true" temperature variability in the presence of small-scale variability that affects real-world observations, we added simulated geophysical noise fields to the CMIP6 modeled temperature anomalies. For these



Supplementary Fig. II.9 Temperature anomalies for Jan-Jun 2010 over the 0-50 m depth interval for (a) a MIROC CMIP6 model run, (b) the MIROC model after being reduced to observed sparsity, and (c) the MIROC model at observed sparsity with added geophysical noise. The middle row shows (d) a single realization of the ARANN interpolation of the MIROC temperature anomalies and (e) the ARANN interpolation of the MIROC temperature anomalies with added geophysical noise. The bottom row shows (f) residuals between the original MIROC model and the ARANN interpolation and (g) residuals using an ARANN interpolation with additional geophysical noise added to the modeled temperature anomalies. The middle panels also show the R^2 between ARANN-reconstructed temperature anomalies (shown in each respective panel) and the original MIROC temperature anomalies in panel (a).

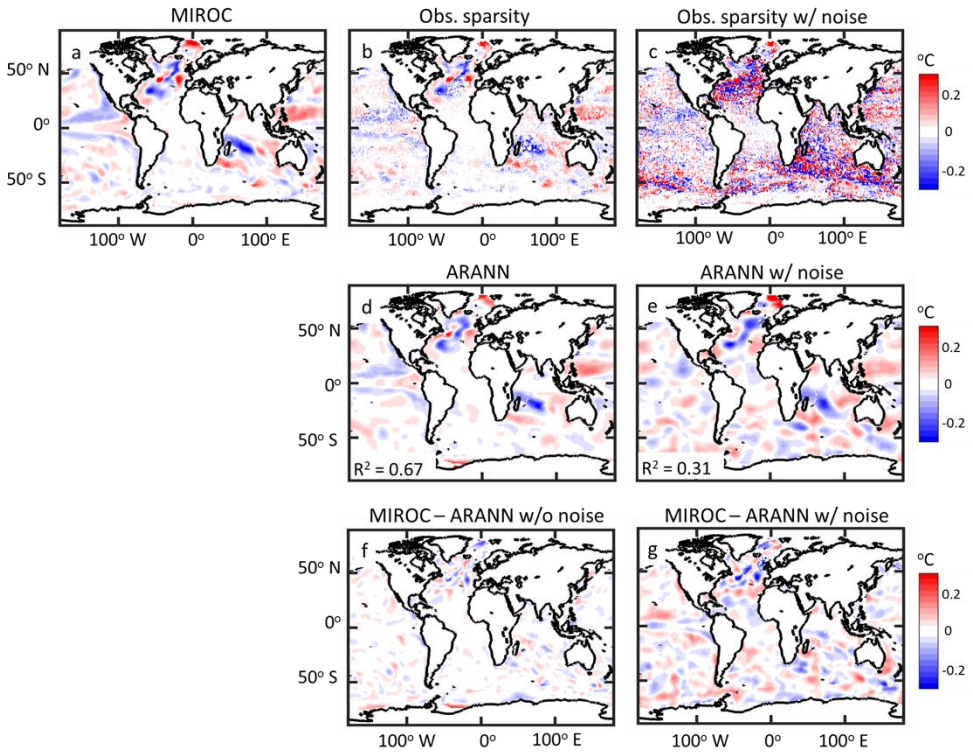
noise fields, we used the residuals between a single member of the ensemble of ARANN-reconstructed temperature anomalies and observed temperature anomalies at each time-step and depth level, for example as shown in Supplementary Fig. II.3 (c, f, i, l) for a couple times and depth levels. We interpret these residuals as due primarily to the geophysical noise present in the observations, although it also includes any additional systematic biases in our interpolation versus the observations. After adding these residuals to the CMIP6 modeled temperature anomaly fields, we then repeated our data processing and interpolation



Supplementary Fig. II.10 Temperature anomalies for Jan-Jun 1960 over the 900-1100 m depth interval for (a) a MIROC CMIP6 model run, (b) the MIROC model after being reduced to observed sparsity, and (c) the MIROC model at observed sparsity with added geophysical noise. The middle row shows (d) a single realization of the ARANN interpolation of the MIROC temperature anomalies and (e) the ARANN interpolation of the MIROC temperature anomalies with added geophysical noise. The bottom row shows (f) residuals between the original MIROC model and the ARANN interpolation and (g) residuals using an ARANN interpolation with additional geophysical noise added to the modeled temperature anomalies. The middle panels also show the R^2 between ARANN-reconstructed temperature anomalies (shown in each respective panel) and the original MIROC temperature anomalies in panel (a).

procedure to see if the ARANN could still faithfully reproduce the original OHC trends in the presence of this more realistic temperature variability.

When interpolating the CMIP6 temperature anomalies with this additional noise that mimics the geophysical noise in the observations, the residuals between the ARANN-reconstructed and the modeled temperature anomalies are larger and more randomly distributed (Supplementary Figs. II.8-11, g) than those reconstructed from the original modeled temperature anomalies (Supplementary Figs. II.8-11,f). The effect of geophysical



Supplementary Fig. II.11 Temperature anomalies for Jan-Jun 2010 over the 900-1100 m depth interval for (a) a MIROC CMIP6 model run, (b) the MIROC model after being reduced to observed sparsity, and (c) the MIROC model at observed sparsity with added geophysical noise. The middle row shows (d) a single realization of the ARANN interpolation of the MIROC temperature anomalies and (e) the ARANN interpolation of the MIROC temperature anomalies with added geophysical noise. The bottom row shows (f) residuals between the original MIROC model and the ARANN interpolation and (g) residuals using an ARANN interpolation with additional geophysical noise added to the modeled temperature anomalies. The middle panels also show the R^2 between ARANN-reconstructed temperature anomalies (shown in each respective panel) and the original MIROC temperature anomalies in panel (a).

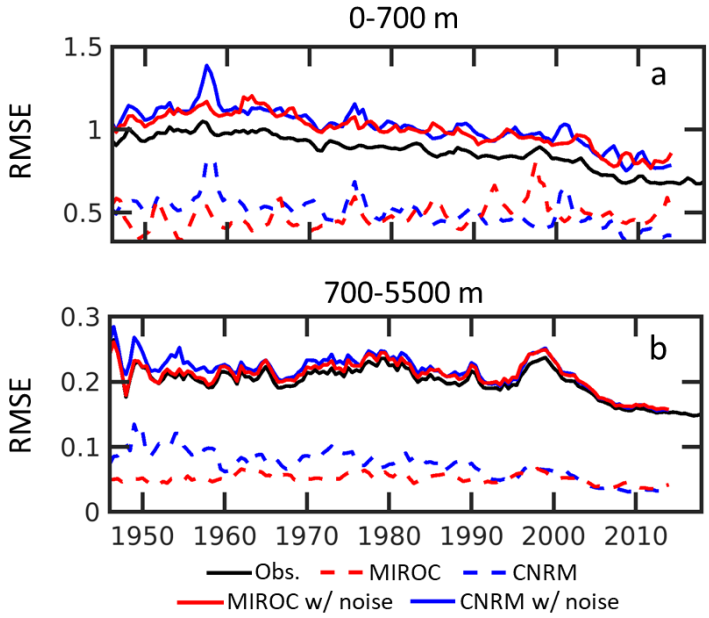
noise on the ARANN reconstruction is also shown in Supplementary Fig. II.12 for the residuals over time: by adding realistic geophysical noise to the CMIP6 models, we obtain residual errors that are similar in magnitude to those obtained from the observations in both the shallow (0-700 m) and deep (700-5500 m) oceans, indicating that small-scale variability has been adequately accounted for in our validation. Even with the larger residuals, the underlying large-scale temperature anomaly patterns are reconstructed well (Supplementary Figs. II.8-11, compare panels d and e to a). The residuals show features on the order of ~10^3

km (Supplementary Figs. II.8-11, f-g), which is smaller than what is needed to assess global and basin-scale OHC. The reconstructions are also robust to regions with large amounts of missing data. For example, at 1000 m in the year 1960 (Supplementary Fig. II.10), sampling biases have left the South Pacific and Southern Ocean with large gaps in coverage. Nevertheless, patterns in these regions are reconstructed with reasonable accuracy.

Importantly for the results discussed in the main text, the bias on the global and basin scale remains relatively small and almost always of lesser magnitude than the uncertainty (Supplementary Figs. II.4-7), although the uncertainty roughly doubles when geophysical noise is included. The performance does somewhat depend on the distribution of temperature anomalies, which is governed by processes in the underlying model. For instance, the ARANN reconstruction (both with and without added geophysical noise) for the CNRM OHC contains a systematic bias that slightly exceeds the estimated uncertainty for the period prior to 1955 and depths greater than 2000 m (Supplementary Fig. II.5). This is not the case with the MIROC reconstructions (Supplementary Fig. II.4), indicating that the impact of sampling bias will depend on the "true" anomaly field.

Further comparing the temperature anomaly fields of the original MIROC CMIP6 model run versus a single realization of the ARANN reconstruction (Supplementary Figs. II.8-11), we find that large scale patterns can be well reproduced both with and without added geophysical noise. For the upper 50 m in the year 1960 (Supplementary Fig. II.8), the ARANN recreates most large-scale patterns, though the presence of geophysical noise yields somewhat larger residuals (Supplementary Fig. II.8f vs. II.8g). The impact of geophysical noise becomes even more apparent in the year 2010, where in the upper 50 m there is very even sampling (Supplementary Fig. II.9). Here the residual temperature anomalies are

smaller than in 1960, but the addition of geophysical noise still leads to anomalies that are both larger in magnitude and spatial extent. In 1960 at 900-1100 m (Supplementary Fig. II.10), sampling coverage is even sparser than in the upper 50 m, but the ARANN reconstruction still performs well at reconstructing the original temperature anomaly fields (R^2 = 0.69), although the addition of geophysical noise reduces the performance substantially (R^2 = 0.43) in the presence of such sparse data. In 2010, there is much better sampling coverage in the deep ocean (Supplementary Fig. II.11), but the ARANN performance is similar to that in 1960, with an R^2 of 0.67 without geophysical noise, and 0.31 with geophysical noise. At all times and depths, the residuals between the ARANN reconstruction and the original modeled temperature anomalies remain relatively small-scale (<~1000 km), even in the presence



Supplementary Fig. II.12 Root-mean squared error (RMSE) between the observed temperature anomalies and the ARANN reconstruction over time (black curve) for (a) the upper 700 m and (b) for the depth range 700-5500 m. Also shown for comparison is the RMSE between the CMIP6 modeled temperature anomalies and the ARANN reconstructions of the CMIP6 models for the MIROC (red dashed) and CNRM (blue dashed) models. Note the much larger RMSE for the ARANN reconstruction of the observed temperature anomalies than for the ARANN reconstruction of the modeled temperature anomalies. When additional noise is added to the modeled temperature anomalies to mimic geophysical noise in the observations, the RMSE between the MIROC (solid red) and CNRM (solid blue) temperature anomalies and the ARANN reconstruction is very similar to the RMSE between the ARANN and the observations (black).

of realistic geophysical noise (Supplementary Figs. II.8-11, f-g). We therefore conclude that the ARANN can overcome both biases in sampling coverage and geophysical noise to reproduce most regional and basin-scale features necessary to assess changes in OHC over time.

B.6 Statistical information

240 ARANN OHC members are used for all depth intervals in this study, representing 10 members for each combination of the four bias corrections (Levitus *et al.*, 2009; Cheng *et al.*, 2014, Gouretski and Reseghetti, 2010; Bagnell and DeVries, 2020) and six decadal climatologies (Garcia *et al.*, 2019). All estimated warming rates for the periods specified in the text are calculated from the mean linear trend of the ARANN OHC members, and the uncertainty in these warming rates are calculated as 2 standard deviations across the individual linear fits for the ensemble members. All estimates in W m^{-2} are for the entire surface area of the Earth. Total warming estimates in ZJ come from multiplying the warming rate derived from a linear fit by the length of the time period under consideration.

Error bars in Fig. II.1, 2, and 4 represent 2 standard deviations computed from the ARANN ensemble members. In Fig. II.3, the trends for each grid cell are considered significantly different from zero only if the trends exceed twice the sum of two components of uncertainty, 1. the ensemble-mean standard error of the linear fit, which is a measure of the amount of uncertainty due to short term variability in the warming rate at a given location, and 2. the standard error of the cross-ensemble linear trends, which is a measure of the uncertainty due to the ARANN mapping method.

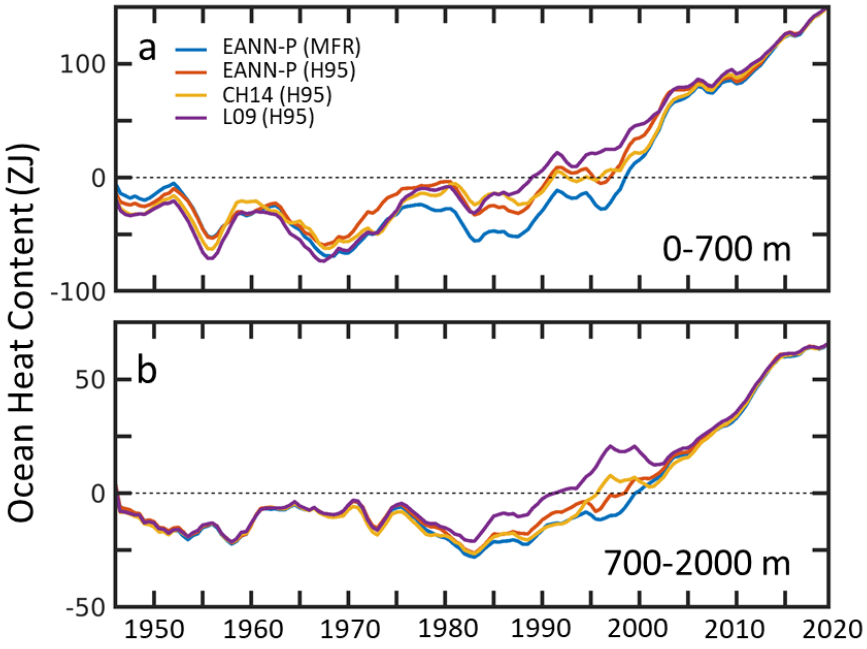
For Fig. II.5, warming rates for the specified periods are calculated from the mean linear trend of the ensemble and the error bars for the ARANN estimates represent the maximum

and minimum warming rates after excluding the 6 members with the highest rates and the 6 with the lowest, representing the 95% confidence interval. Error bars in Fig. II.5 for previously published warming rates come from the uncertainties calculated by these other studies, which may differ somewhat from each other due to methodological choices but roughly represent the 95% confidence interval.

C. Results

C.1 Global and basin-scale OHC changes

Estimates of the global full-depth OHC from the ARANN method show that there was no net ocean warming during the four decades from 1950-1990, but instead the OHC fluctuated by ~50 ZJ on roughly decadal timescales (Fig. II.1a). However, since 1990 there



Supplementary Fig. II.13 Global ocean heat content estimates for the depth intervals (a) 0-700 m and (b) 700-2000 m using various corrections for the systematic biases in the bathythermograph data, with their respective XBT fall rate equation in parentheses. Each curve represents the ensemble mean for a given correction method after running an ensemble of ARANNs comprised of 60 members. The XBT corrections used in this study are from L09, CH14, and EANN-P (see chapter I). Two EANN-P calibrations are applied to the XBT data using either the original manufacturer fall rate equation (MFR) or a modified one (H95).

has been a rapid acceleration in ocean warming, with the ocean gaining 303±56 ZJ of thermal energy in the past three decades (Fig. II.1a). This temporal pattern is roughly consistent throughout the water column, with minor warming prior to 1990 in the upper 700 m, no warming in the 700-2000 m depth range, and cooling in the deep and abyssal layers below 2000 m (Fig. II.1b-d). Warming rates accelerated substantially after 1990 throughout the entire water column, with the deep ocean switching from cooling to warming after 1990 (Fig. II.1b-d).

Passive transport methods Gebbie and Huybers, 2019; Zanna *et al.*, 2019) that propagate surface temperature anomalies to the deep ocean using steady-state ocean circulation patterns provide internally consistent estimates of full-depth OHC that can be directly compared to ARANN, after adjusting their baselines to coincide with the ARANN estimate during the Argo era (2005 onwards) (Fig. II.1a). These passive transport estimates differ from the ARANN and from each other. Both show an earlier onset of ocean warming than the ARANN, with full-depth ocean warming starting in the mid-1970s for the optimized mixing model (OPT-0015) (Gebbie and Huybers, 2019) and going back to the 1950s in the Green's function (GF) (Zanna *et al.*, 2019) product (Fig. II.1a). The OPT-0015 method shows very little ocean warming prior to the mid-1970s, in agreement with the ARANN reconstruction, while the GF method suggests a nearly constant ocean warming trend throughout the past ~70 years.

Estimates of OHC from objective mapping products can be compared to the ARANN estimates in the upper 2000 m (Fig. II.1b-c). There is broad agreement about the total change in OHC among the objective mapping products and with the ARANN ensemble for much of 1980-2019, with the JMA (Ishii *et al.*, 2017) estimate on the very edge of the

ARANN uncertainty range for the 700-2000 m depth interval (Fig. II.1c). Prior to 1980, the mapping methods diverge somewhat in their predictions over the upper 2000 m, and the

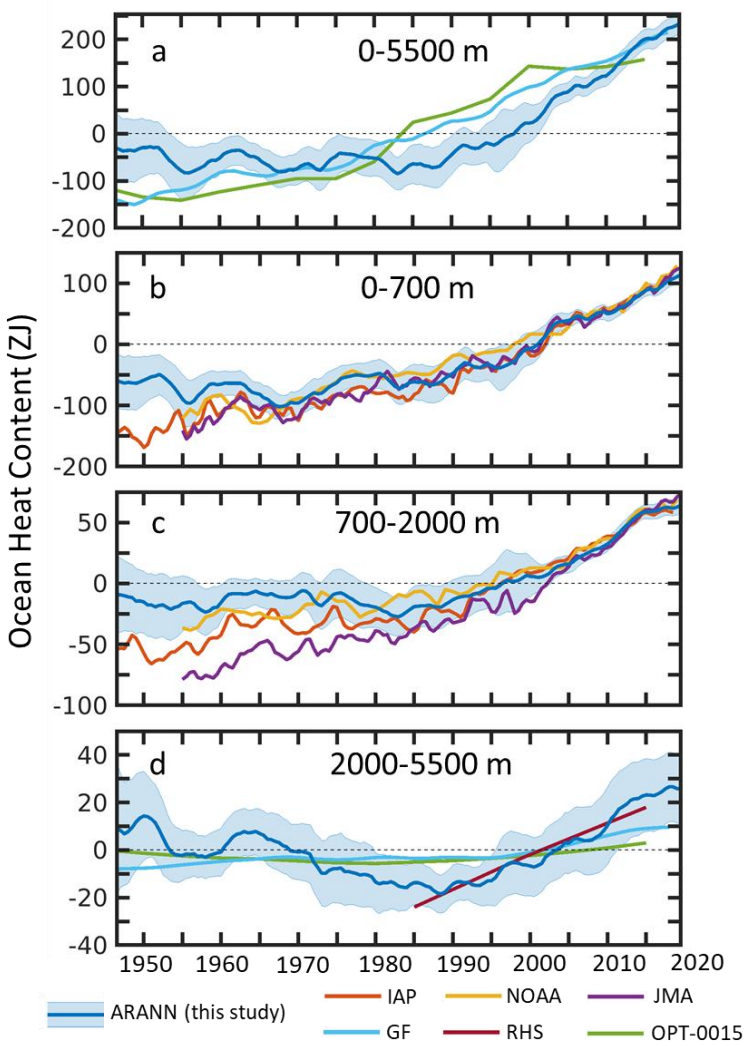
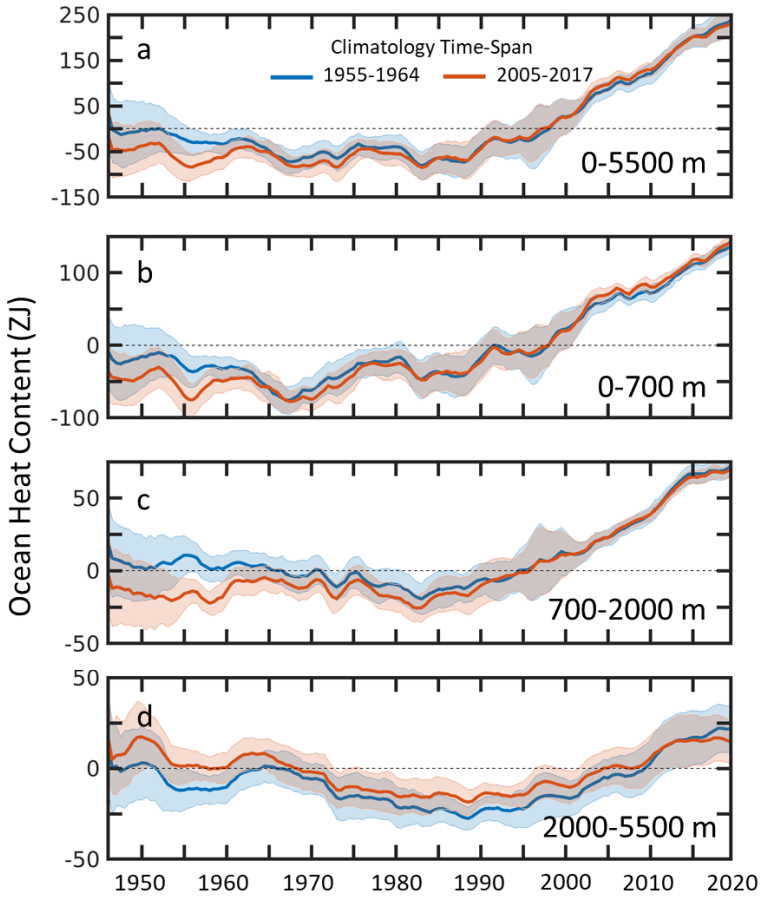


Fig. II.1 Estimates of ocean heat content (OHC) changes for (a) the global ocean from surface to seafloor, (b) the upper 700 m of the ocean, (c) the depth range 700-2000 m, and (d) the depth range 2000-5500 m. The mean estimates derived from this study (ARANN, blue) are shown with shading covering two standard deviations from the mean over the 240 ARANN ensemble members. The zero anomaly is defined such that the mean OHC of the ARANN estimate for the period 1946-2019 is zero. Also shown are the mean OHC anomaly from the IAP (red), NOAA (yellow), and JMA (purple) objective mapping products, which cover the 0-2000 m depth interval as shown in panels (b)-(c). These products have been adjusted to the mean ARANN OHC anomaly for 2005-2019. Shown for (a) the full ocean depth and (d) the deep ocean are OHC anomalies from passive ocean heat uptake models using Green's functions (GF) (light blue) and an optimized mixing model (OPT-0015) (green). The passive ocean heat uptake products are adjusted to the mean ARANN anomaly for 1955-1985. Repeat hydrographic sampling (RHS) gives temperature trends since 1985 in the deep ocean (maroon; panel d). The RHS method gives a linear trend from 1985-2000 and from 2000-2015, which has been adjusted to the mean ARANN anomaly for 1985-2015.

disagreement is most pronounced for the earliest time periods. After adjusting the OHC anomalies of the objective mapping estimates to the mean ARANN 0-2000 m OHC value over 2005-2019, the ARANN OHC in 1955 is 71±58 ZJ greater than that estimated by the IAP (Cheng *et al.*, 2017) product, 51±58 ZJ greater than the NOAA (Levitus *et al.*, 2012) product, and 114±58 ZJ greater than the JMA (Ishii *et al.*, 2017) product (Fig. II.1b-c). The large spread among OHC products prior to 1980 is primarily due to increased data sparsity



Supplementary Fig. II.14 Global ocean heat content (OHC) reconstructions using two different climatologies to form the temperature anomalies before interpolation with the ARANN method. The first climatology uses only data from 2005-2017 (red curve), and the second climatology only uses data from 1955-1964 (blue curve). Thick lines are the ensemble mean, and shading represents two standard deviations across 40 ARANN members using the four bathythermograph corrections shown in Supplementary Fig. II.13. Four more decadal climatologies are combined with the ones shown here to produce the OHC estimates and uncertainties in the main text.

in this period, but the choice of reference climatology also plays a role in enhancing uncertainties in the ARANN during this time period. For years prior to 1970, mean ARANN OHC over the 0-2000 m depth interval can vary by as much as 67 ZJ when using different reference climatologies (Supplementary Fig. II.14), which is a source of uncertainty that has generally been neglected in the other mapping products. Despite these uncertainties, all mapping methods show an acceleration in OHC uptake over time in the 0-2000 m interval. For the mapping products, the underlying subsurface temperature data creates strong constraints that reduce the variance across methods over the last several decades of OHC estimates (Fig. II.1b-c), in contrast to the passive transport products where differences in methodology have large impacts on the inferred OHC trends (Fig. II.1a).

The ARANN yields a global interpolation of deep subsurface ocean temperature data and shows a cooling trend from 1950 to 1990, representing a reduction of OHC by 26±16 ZJ (Fig. II.1d), mainly canceling out the small heat gain in the upper 700 m and contributing to the negligible warming of the global ocean estimated by the ARANN during this time period (Fig. II.1a). The passive transport methods both predict almost negligible changes in deep ocean heat content during this period, with the GF method showing slight warming and OPT-0015 showing slight cooling (Fig. II.1d). The ARANN predicts that the deep ocean has warmed significantly since 1990, gaining 48±19 ZJ, at a rate that closely matches the estimates from repeat hydrographic surveys (Desbruyeres *et al.*, 2016) (RHS). Over the past 30 years the deep ocean has gained back all of the heat lost since 1950, arriving at possibly its warmest level over the entire 75-year record (Fig. II.1d). This rapid warming of the deep ocean is contrary to the slow rise in OHC implied by the passive transport models that use steady ocean circulation (Gebbie and Huybers, 2019; Zanna *et al.*, 2019), suggesting that

transient features of the deep ocean circulation are important for contributing to the warming over the recent three decades (Purkey and Johnson, 2010; Masuda *et al.*, 2010).

Clear regional differences in ocean warming rates emerge on ocean basin scales (Fig. II.2). In the upper 700 m, the Atlantic Ocean has warmed the most of all the major ocean

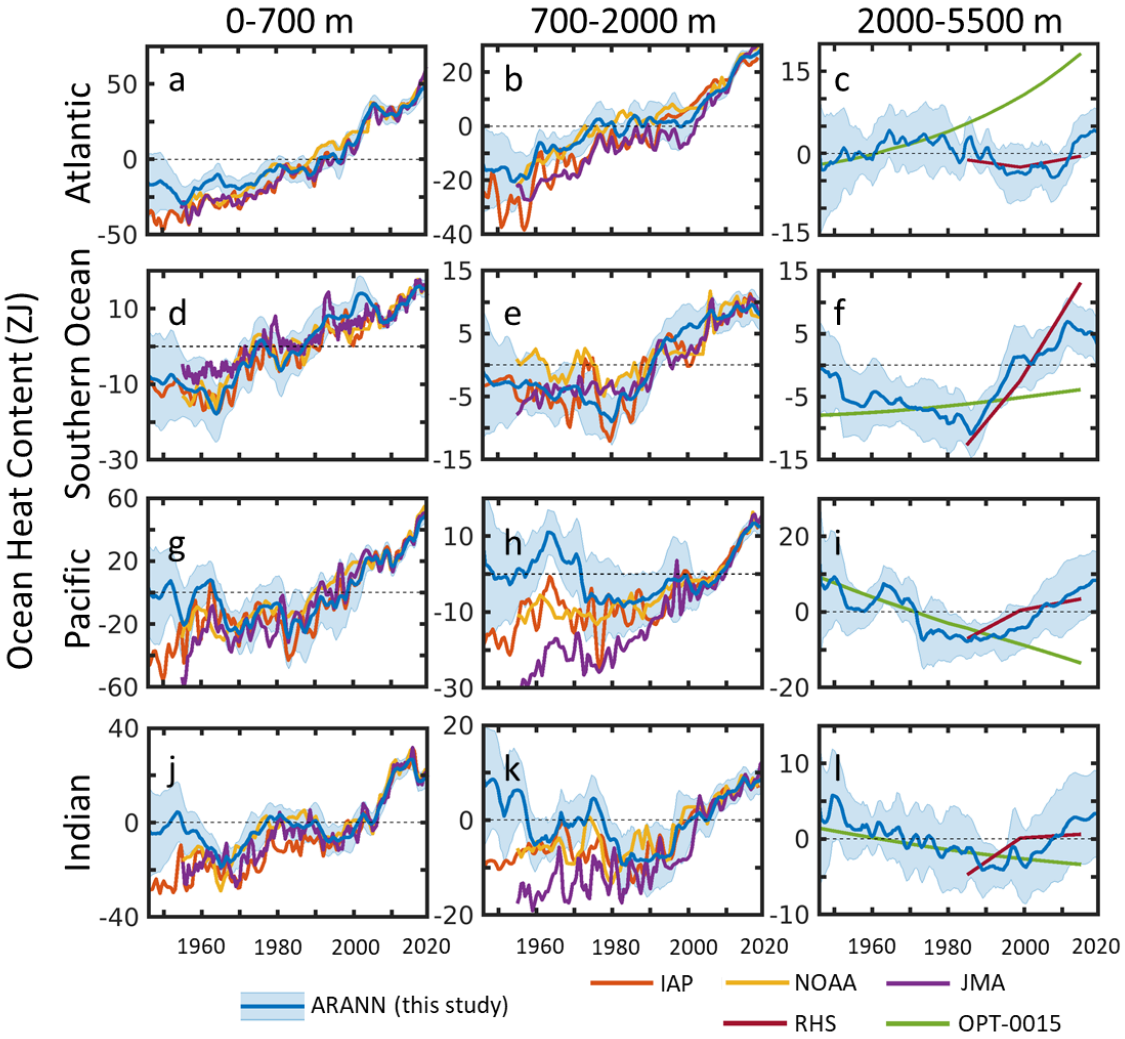


Fig. II.2 Basin-scale ocean heat content (OHC) anomalies for the Atlantic Ocean at depth intervals (a) 0-700 m, (b) 700-2000 m, (c) 2000-5500 m; for the Southern Ocean at depth intervals (d) 0-700 m, (e) 700-2000 m, (f) 2000-5500 m; for the Pacific Ocean at depth intervals (g) 0-700 m, (h) 700-2000 m, (i) 2000-5500 m; and for the Indian Ocean at depth intervals (j) 0-700 m, (k) 700-2000 m, (l) 2000-5500 m. Ocean basins are defined using the World Ocean Atlas mask, with the Southern Ocean considered everything south of 50^{o} S. OHC anomalies and uncertainties are computed as in Fig. II.1 and compared with previous reconstructions as in Fig. II.1. The Green's Function (GF) method does not provide basin-scale estimates of OHC and is omitted in the comparison here.

basins, showing sustained warming since the 1950s and accounting for more than one third of the total warming in the upper 700 m. Objective mapping methods and the ARANN agree for the period after 1980 in the shallow Atlantic, however the ARANN produces 7-10 ZJ less warming than the objective mappings prior to 1985 (Fig. II.2a). The sustained warming of the Atlantic Ocean has penetrated into the intermediate layers (700-2000 m), but methods disagree on the amount of warming prior to 2005, with the JMA and IAP estimates putting as much as 10 ZJ more warming into the intermediate Atlantic than the ARANN and NOAA methods for much of the record (Fig. II.2b). The ARANN reconstruction of deep Atlantic OHC (below 2000 m) shows a slight cooling trend until ~2005 then subsequent warming. This reversal from cooling to warming trends is also captured by the repeat hydrography data (Fig. II.2c). However, the passive transport method OPT-0015 shows an accelerated warming of the deep Atlantic over the entire time period (Fig. II.2c), which may result from biases in its estimated ventilation rates. Passive transport of recent surface warming into the internal Atlantic via a steady-state ocean circulation overlooks natural variability in rates of overturning and deep water formation (Zhang *et al.*, 2019; Polyakov *et al.*, 2005; Kim *et al.*, 2018), possibly explaining an overestimate of the warming trend in the OPT-0015 versus those of subsurface observations such as ARANN and RHS.

Fingerprints of ocean circulation changes are also apparent in the spatial distribution of warming rates (Fig. II.3). Prior to 1990, the upper 700 m of the subpolar and polar North Atlantic was cooling, while the Gulf Stream extension region was warming (Fig. II.3a), consistent with trends that have previously been identified as fingerprints of reduced deep convection in high-latitude deep water formation regions (Robson *et al.*, 2016; Caesar *et al.*, 2018). Since 1990 there has been coherent strong warming throughout most of the Atlantic

basin, except for a small patch of cooling in the center of the North Atlantic subpolar gyre (Fig. II.3d). Warming of the North Atlantic from 1990-2005 has previously been linked to a surge in the Atlantic Meridional Overturning Circulation (AMOC) after 1990 (Robson *et al.*, 2012), although this has been followed by a decline in the AMOC after 2005 and cooling of the subpolar gyre (Robson *et al.*, 2016) potentially contributing to the cooling trend identified over this longer period in the central subpolar gyre (Fig. II.3d). After 1990 there is also pronounced warming at mid depths (700-2000 m) throughout most of the Atlantic, concentrated more strongly in the subpolar north and south Atlantic (Fig. II.3e). This is consistent with the mean overturning circulation transporting surface warming to intermediate waters, since these regions are close to the formation regions for North Atlantic

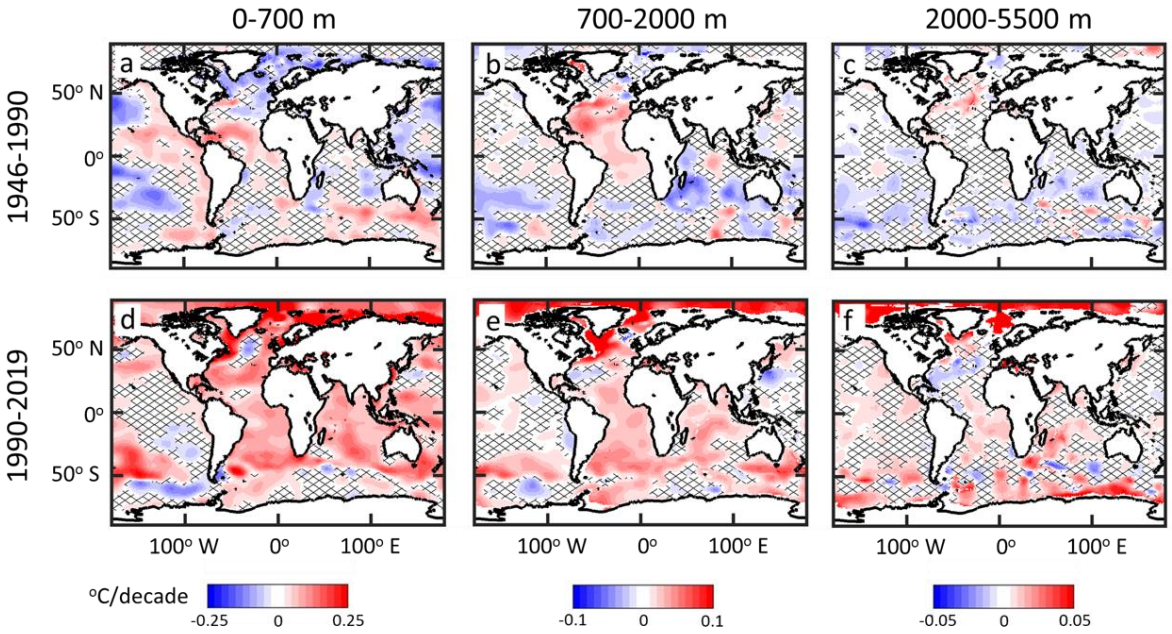


Fig. II.3 Spatial maps of linear warming rates for the period 1946-1990 at depths (a) 0-700 m, (b) 700-2000 m, and (c) 2000-5500m, and for the period 1990-2019 at depths (d) 0-700 m, (e) 700-2000 m, and (f) 2000-5500 m. Warming rates were estimated by averaging temperature anomalies over each depth interval, then applying a linear least-squares fit to the temporal trend of temperature at each 1° grid cell for the specified time periods. Areas without significant trends (95% confidence interval) are cross-hatched. Uncertainties were estimated by calculating the sum of the error of the linear fit and the cross-ensemble uncertainty of the warming rates at each grid cell for the various ARANN ensemble members.

Deep Water in the North Atlantic (Zhang, 2008) and Antarctic Intermediate Waters in the South Atlantic (Sloyan and Rintoul, 2001).

Like the Atlantic, the Southern Ocean (defined here as south of 50^{o} S) has been warming consistently since 1960 in the upper 700 m, a trend seen across multiple reconstruction methods (Fig. II.2d). In the intermediate layers (700-2000 m), large differences in sub-decadal variability across these methods reveal the impact of sparse temperature sampling in the region, but the consensus across methods is that warming of Southern Ocean intermediate waters started in the 1980s, with little warming before that (Fig. II.2e). The ARANN reconstruction shows a cooling trend in the deep Southern Ocean (> 2000 m) until ~1985, followed by rapid warming thereafter (Fig. II.2f). The post-1985 warming trends in the deep Southern Ocean in the ARANN generally agree with the trends derived from repeat hydrography, whereas the OPT-0015 passive transport method shows a very small warming trend over the entire 1946-2019 period (Fig. II.2f). The spatial distribution of warming in the deep Southern Ocean shows that the rapid warming over the past three decades is concentrated along the Antarctic margin, where Antarctic Bottom Waters form in the Weddell Sea, Ross Sea, and other marginal seas along the Antarctic shelf (Orsi *et al.*, 1999; Jacobs, 2004) (Fig. II.3f).

The Southern Ocean is a key global heat sink over the past three decades. Most regions of the Southern Ocean have been warming consistently since 1990 (Fig. II.3d-f), and the 0-700 m, 700-2000 m, and 2000-5500 m depth intervals have all experienced roughly the same amount of OHC change during this period (Fig. II.2d-f). Currently, the entire Southern Ocean from surface to seafloor sits at its warmest levels since at least the 1950s. This rapid warming of the Southern Ocean has been accompanied by a general asymmetrical warming

over the past decade that has favored heating of the Southern Hemisphere. For the period 2005-2019, the ARANN estimates that the Southern Hemisphere accounted for 68% of global OHC change in the top 700 m, 54% for 700-2000 m, and 81% below 2000 m. This asymmetrical warming is consistent with previous analyses that linked the redistribution of heat to internal climate variability (Rathore *et al.*, 2020). This appears to be a recent change in the pattern of ocean heat uptake, however, as Northern Hemisphere waters have consistently warmed over the entire 75-year record of the ARANN while Southern Hemisphere waters cooled on average prior to 1990 (Fig. II.3).

Warming of the Atlantic and Southern Oceans above 2000 m prior to 1990 was counterbalanced by slower rates of warming or by cooling of the Indian and Pacific Oceans over this time period (Fig. II.2g-h,j-k). The Indian and Pacific Oceans show little trend in heat content in the upper 700 m from 1950-1990 for the ARANN, while other mapping products show very slight warming over this period. Instead, this period is mostly marked by decadal-scale oscillations in OHC in the ARANN reconstruction, which also appear to some extent in the objective mapping products (Fig. II.2g,j). Identifying the mechanisms responsible for these oscillations is beyond the scope of this study, but these could be related to changes in upper-ocean overturning circulation associated with the Interdecadal Pacific Oscillation (IPO) (Meehl *et al.*, 2013), which has recently been ascribed to changes in the strength of Pacific trade winds that affect eastern equatorial upwelling (England *et al.*, 2014), as well as modifications to the winds in the tropical North Pacific (Lee *et al.*, 2015). These factors also appear to modulate heat transport into the Indian Ocean via the Indonesian Throughflow (Lee *et al.*, 2015; Liu *et al.*, 2016). Below 700 m, the OHC estimates of the various mapping products greatly diverge prior to 1990. ARANN indicates

that the mid-depth Pacific and Indian Oceans were cooling up until 1990, whereas the IAP and NOAA products show little trend over this period and the JMA method produces unabated warming throughout the record, leading to large disagreements across methods totaling ~45 ZJ of difference in the warming summed over these two basins (Fig. II.2h,k).

The ARANN reconstruction also shows that the deep and abyssal layers of the Pacific and Indian Ocean were cooling up until ~2000 (Fig. II.2i,l). This cooling trend agrees well with the OPT-0015 multi-millennial passive ocean heat uptake reconstruction, which produces a 20th-century cooling trend in the deep Pacific and Indian Oceans in response to cold surface temperatures during the Little Ice Age that occurred from the 14^{th}-19^{th} centuries (Gebbie and Huybers, 2019). This cooling trend has been preserved into the 20th century due to the deep ocean's long overturning timescales (Gebbie and Huybers, 2019).

In the past two decades, the Pacific and Indian Oceans have warmed substantially throughout the water column, contributing ~52% of the global OHC change since the year 2000. This warming has been concentrated in the upper 700 m, in agreement with objective mapping reconstructions (Fig. II.2g,j). While this recent warming is coherent across most of the Indian Ocean, the warming is more concentrated in the central and western Pacific Ocean, with a slight cooling trend in the eastern Pacific (Fig. II.3d). This Pacific pattern is consistent with findings of enhanced meridional temperature gradients in the tropical Pacific Ocean (Karnauskas *et al.*, 2009) which could help explain a shift toward strong basin-wide El Nino events during recent decades (Wang *et al.*, 2019). The ARANN reconstruction also shows substantial warming since 2000 of the mid-depth ocean (700-2000 m) as well as the deep ocean (below 2000 m), in general agreement with objective mapping approaches in the mid-depth layers and with trends derived from repeat hydrographic surveys in the deep

ocean (Fig. II.2 h-i,k-l). However, the recent warming of the deep Pacific and Indian Oceans found by the ARANN is contrary to the continued slow cooling implied by passive heat uptake in the OPT-0015, suggesting that the deep ocean warming since 2000 is related to changes in the transport of deep and bottom waters (Masuda *et al.*, 2010; Kouketsu *et al.*, 2011).

C.2 A shift in Earth's energy imbalance

The time derivative of the full-depth OHC, dOHC/dt, should very closely track the EEI, since all other heat sinks in the climate system are currently an order of magnitude

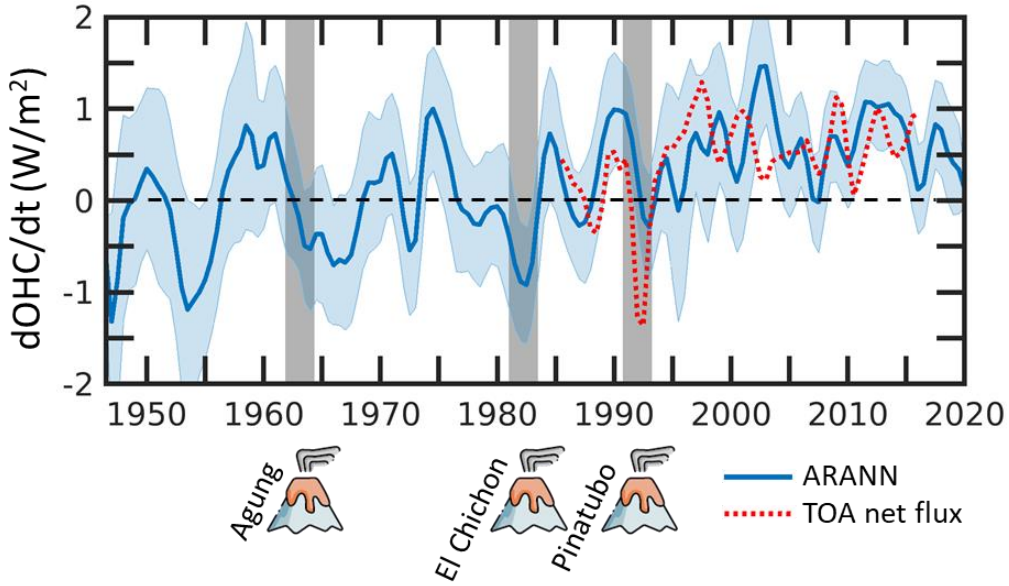


Fig. II.4 Time derivative of the global full-depth ocean heat content (dOHC/dt) from 1946-2019 (blue; mean) with uncertainty (shading; 2 standard deviations) across 240 ARANN ensemble members. The dOHC/dt is computed using a centered difference in time of the global OHC and applying a 1-2-1 filter to smooth the result. This warming rate is divided by the surface area of the Earth so that it can be interpreted as the ocean component of the Earth Energy Imbalance (EEI). This is compared to the top of atmosphere (TOA) net radiative flux (Allan *et al.*, 2014) (red dashed) for years 1985-2016. The timing of 3 major volcanic eruptions is represented by icons with the corresponding volcano's name. Vertical grey bars cover 12 months prior to and 18 months after each eruption to account for the imperfect time resolution of the ARANN dOHC/dt (~12 months) and the e-folding time of volcanic aerosols in the stratosphere (Lambert *et al.*, 1997).

smaller than the ocean (Meyssignac *et al.*, 2019; Trenberth 2014). Periods of positive dOHC/dt represent ocean warming and a positive net energy imbalance at the top of the atmosphere, while negative dOHC/dt represents ocean cooling and a negative EEI (Fig. II.4). The EEI itself reflects a combination of positive anthropogenic forcing due to greenhouse gas emissions and negative forcing due to anthropogenic aerosols (Myhre *et al.*, 2013), as well as natural forcing by volcanic eruptions (Myhre *et al.*, 2013; Trenberth, 2014) and natural variability due to internal dynamics of the climate system (Trenberth, 2014, von Schuckmann *et al.*, 2016). We find that over the entire 1946-2019 period covered by our OHC reconstruction, there are more than a dozen transitions between positive and negative values of the dOHC/dt (Fig. II.4). Some of the most negative dOHC/dt values are coincident with major volcanic eruptions over the past 50 years, including Agun, El Chichon, and Pinatubo (Trenberth, 2014). However, there are more oscillations in the dOHC/dt than can be associated with volcanic aerosol forcing, and in the case of El Chichon the timing of the eruption occurs at a local minimum in the dOHC/dt instead of just prior, indicating that volcanic aerosol forcing may not always dominate over natural variability in the EEI. This supports prior studies that point to internal modes of climate variability such as ENSO (Meyssignac *et al.*, 2019; Loeb *et al.*, 2012) and the IPO (von Schuckmann *et al.*, 2016; Meehl *et al.*, 2013), or natural variability in solar irradiance (Fröhlich, 2006) as factors influencing sub-decadal changes in the EEI for this 75-year record.

Prior to 1990 the dOHC/dt oscillates around zero, averaging -0.04±0.11 W m^{-2} for the period 1946-1990. Without taking into account the deep ocean cooling during this period, the average dOHC/dt would be 0.01±0.09 W m^{-2}. An upward shift in the dOHC/dt is noticeable in the mid-1990s (Fig. II.4), after which the average warming rate in the ARANN

OHC reconstruction is nearly always positive, averaging 0.67±0.13 W m^{-2} for the period 2000-2019. The timing of this shift, the magnitude of the implied EEI, and the temporal variability of the EEI agree well with estimates of the top of atmosphere (TOA) net radiative flux (Allan *et al.*, 2014) for the period 1985-2016. The only time periods where the TOA net radiative flux lies outside 2 standard deviations of the ARANN-estimated EEI are in the early 1990s just after the Pinatubo eruption, when the TOA radiative flux is more negative than the dOHC/dt, and during the early 2000s when the dOHC/dt shows an upward jump that is opposed to a drop in the TOA net radiative flux (Fig. II.4).

The magnitude of the ARANN-estimated dOHC/dt from 2000-2019 agrees within 2 standard deviations with numerous other estimates for this period, including those based on dOHC/dt from objective mapping products (Cheng *et al.*, 2017; Johnson *et al.*, 2016), ocean reanalysis products (Boisséson *et al.*, 2018), and CMIP5 hindcast models (Smith *et al.*, 2015), as well as estimates of EEI from satellite altimeter and gravity data (Meyssignac *et al.*, 2019) and top-of-atmosphere radiative fluxes (Allan *et al.*, 2014) (Fig. II.5). Prior to the year 2000, warming rates estimated from our full-depth OHC reconstruction mostly agree within their respective uncertainties with previous estimates from the IAP mapping product (Cheng *et al.*, 2017), passive transport methods (Gebbie and Huybers, 2019; Zanna *et al.*, 2019), CMIP5 hindcast models (Smith *et al.*, 2015), and an independent estimate of EEI for the period 1990-2016 from measurements of atmospheric composition (Resplandy *et al.*, 2019). For the period 1960-1990, the ARANN approach estimates essentially no warming (-0.03±0.12 W m^{-2}), implying a roughly balanced Earth energy budget over this time period. Objective mapping, passive transport methods, and climate models have generally estimated small warming rates over the 1960-1990 period, but mostly overlap with the ARANN

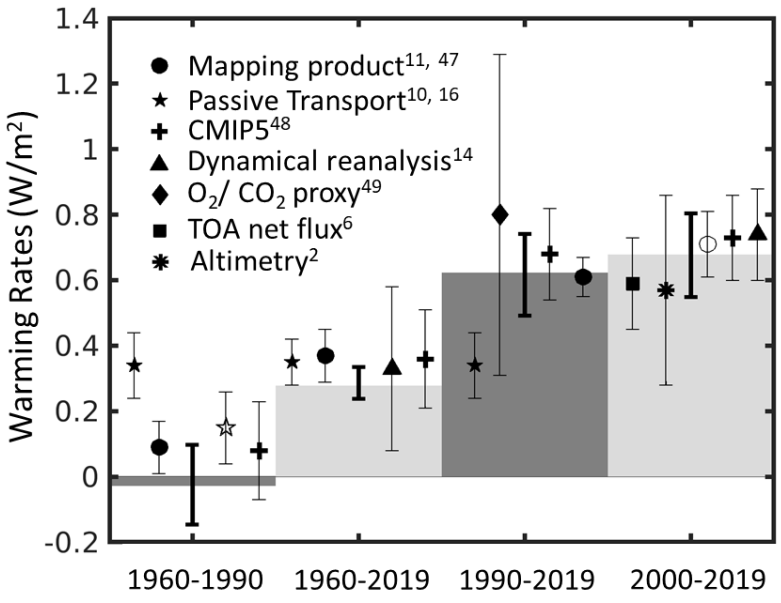


Fig. II.5 Linear ocean warming rates derived from this study (vertical grey bars) with 95% confidence intervals (bold error bars) computed by taking the minimum and maximum warming rates of the middle 95% of 240 ARANN ensemble members. These are compared to other published estimates and uncertainties (symbols and error bars) derived from mapping methods (filled circle Cheng *et al.*, 2017; open circle Johnson *et al.*, 2016), passive transport products (filled star Zanna *et al.*, 2019; open star Gebbie and Huybers, 2019), dynamical reanalyses (Boisséson *et al.*, 2018) (triangle), satellite altimetry (Meyssignac *et al.*, 2019) (asterisk), CMIP5 climate models (Smith *et al.*, 2015) (cross), top of atmosphere (TOA) net radiative flux (Allan *et al.*, 2014) (square), and the chemical composition of the atmosphere (Resplandy *et al.*, 2019) (diamond). Each symbol is associated with one of the grey bars and covers approximately the same time interval.

estimates within their respective uncertainties (Fig. II.5). Almost all methods of estimating EEI agree on an acceleration of the EEI in recent decades, particularly since 1990 (Fig. II.5). A notable exception is the GF passive transport method, which maintains a steady warming rate across the entire 1960-2019 period, implying an important role for ocean circulation changes in controlling the acceleration of the EEI over recent decades. Overall, the ARANN results support a broad consensus across almost all products of accelerated warming over time (Fig. II.5), but they also suggest that previous estimates of ocean warming may have been biased too high prior to 1990, in part due to the neglect of deep ocean cooling. Including the effects of deep ocean cooling, as determined by the mean estimate of the

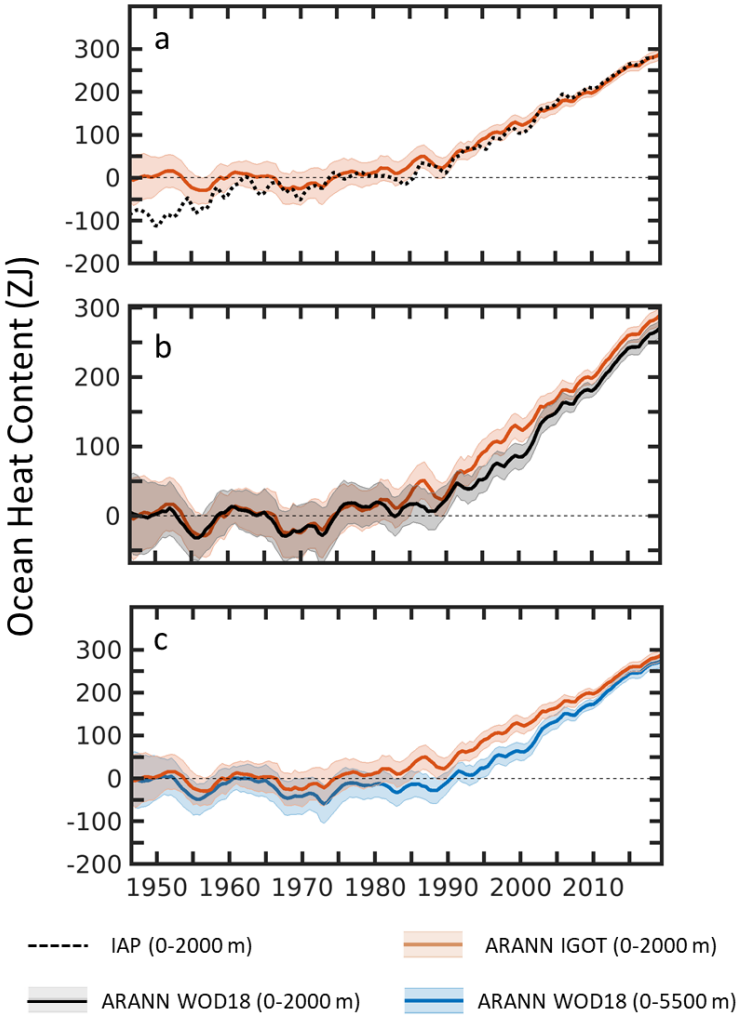
ARANN, would lower the rates of ocean warming prior to 1990 determined by previous objective mapping approaches (Cheng *et al.*, 2017; Levitus *et al.*, 2012, Ishii *et al.*, 2017) by 18 to 32%.

D. Discussion

The ARANN reconstruction of full-depth OHC provides an internally consistent framework for monitoring EEI over time, showing that the Earth energy budget was in quasi-equilibrium, with substantial decadal variability, for the four decades from 1950-1990. The warming rate from the ARANN does not differ from that derived by objective mapping methods with statistical significance, and previous studies already support a slower ocean warming rate for the 1950-1990 period relative to the 21st century (Fig. 5). However, due to the combination of a smaller estimated change in 0-2000 m OHC for 1950-1990 and the contribution of deep ocean cooling, the ARANN implies a stronger and later shift toward accelerated EEI than previously recognized, and raises the question as to what may have caused this climate shift.

Anthropogenic radiative forcing has remained positive and continued to grow in magnitude over the past century (Myhre *et al.*, 2013), so the lack of global ocean warming implied by the ARANN results over the period from 1950-1990 may seem counterintuitive at first. However, Earth's climate system is not currently at equilibrium. Due to the timescales of overturning in the ocean, propagating the entire forced climate signal from the surface to the interior may require decades to centuries to manifest as signals in the deep OHC (Purkey and Johnson, 2010; Gebbie and Huybers, 2019), implying that the EEI is modulated by changes in external forcing on multi-decadal time-scales. In the deep ocean, cooling of the Pacific and Indian over much of the 20th century could result from a past

climate event such as the Little Ice Age (Gebbie and Huybers, 2019). A cooling trend that derives itself from long term modes of climate variability (Jones and Mann, 2004) would not



Supplementary Fig. II.15 Global ocean heat content (OHC) estimates comparing the influence of the 0-2000 m OHC ARANN estimate versus the inclusion of the deep ocean below 2000 m. The official IAP OHC estimate (dashed black) is compared to the 0-2000 m ARANN estimate that utilized the same IGOT dataset employed by the IAP (red), demonstrating differences solely due to methodology employed by the IAP and ARANN estimates (a). The 0-2000 m ARANN estimate using the IGOT dataset (red) is also compared to the estimate from the main text using the WOD18 dataset and the CH14 instrumental bias correction (black) (b). Finally, the 0-2000 m ARANN estimate using the IGOT dataset (red) is compared to the 0-5500 m ARANN estimate using WOD18 dataset and the CH14 instrumental bias correction (blue) (c). The blue curve here is almost identical to that shown in Fig. II.1a, expect that only the CH14 bias correction is used. Anomalies for the ARANN estimates are set zero for year 1946. The IAP estimate is adjusted to the mean anomaly of the ARANN IGOT estimate for 2005-2017. Error bars represent 2 standard deviations across 20 ensemble members.

be reflected in any of the components of the external radiative forcing budget for the 20th century.

Nonetheless, deep ocean cooling does not entirely account for the near zero warming trend in OHC prior to 1990, especially when considering that the 0-2000 m interval shows minimal change in the ARANN OHC estimate as well, averaging just 0.03±0.09 W m^{-2} from 1960-1990. The difference between the ARANN and the IAP reconstruction (Cheng *et al.*, 2017) of OHC in the upper 2000 m, is similar in magnitude to the ARANN estimate of deep ocean cooling (Supplementary Fig. II.15), and in general the spread across OHC estimates in the top 2000 m is larger than the deep ocean cooling trend estimated by the ARANN (Fig. II.1b-c). This spread indicates large uncertainties related to methodological differences in estimating OHC over the latter half of the 20th century. However, if the ARANN estimate of minimal upper ocean warming prior to 1990 is correct, it could indicate that anthropogenic or volcanic aerosol effects are larger than currently estimated for this time period (Storelvmo *et al.*, 2016) or that the transient climate response to anthropogenic forcing is affected by regional feedbacks arising from the pattern of ocean heat uptake (Winton *et al.*, 2010; Armour *et al.*, 2013). Changes in the ocean overturning can also affect the EEI by modifying the rate of ocean heat uptake (Baggenstos *et al.*, 2019), which could also lead to discrepancies between radiative forcing and upper ocean warming.

The recent accelerated warming since 1990 implied by the ARANN is consistent with the dominant effects of anthropogenic greenhouse gas forcing and negligible volcanic aerosol forcing (Myhre *et al.*, 2013; Haywood *et al.*, 2013), as well as estimates of increased radiative forcing (Kramer *et al.*, 2021) during the past three decades. Due to improved ocean temperature sampling over the past several decades, there is high confidence that the top

2000 m of the ocean have been gaining heat at an accelerating rate, as indicated by the convergence of OHC estimates across methodologies during this time period (Fig. II.1b-c). Additionally, the ARANN results suggest that the deep ocean below 2000 m has added 48±19 ZJ since 1990, or about 10 to 28% of the ocean warming above 2000 m during this period, significantly contributing to the accelerating EEI in recent decades. This contribution is larger than that from non-ocean components of the Earth energy budget, including the land surface, cryosphere, and atmosphere, which together account for ~27±8 ZJ of warming since 1990 (von Schuckmann *et al.*, 2020).

In all, the results presented here show that deep ocean cooling during the latter half of the 20th century has given way to deep ocean warming over the past three decades, contributing to a delayed response of the EEI to contemporary radiative forcing effects. If this recent shift toward warming of the deep ocean continues, it will have implications for Earth's climate for decades to centuries to come due to the long overturning timescales of the deep ocean. Continued monitoring of the global ocean heat content, and improved resolution of deep ocean temperature changes, will be key for developing accurate forecasts of Earth's energy budget and future climate change.

III. Global Mean Sea Level Rise Inferred from Ocean Temperature and Salinity

A. Introduction

Global warming due to anthropogenic greenhouse gas emissions is responsible for an increase in global sea level that threatens human life and infrastructure along coastlines. Future sea level rise may expose as many as 160 million more people and an additional $8 trillion in property to episodic flooding by the year 2100 (Kirezci *et al.*, 2020). Global mean sea level rise (GMSLR) measures the rate of increase in the average global sea level

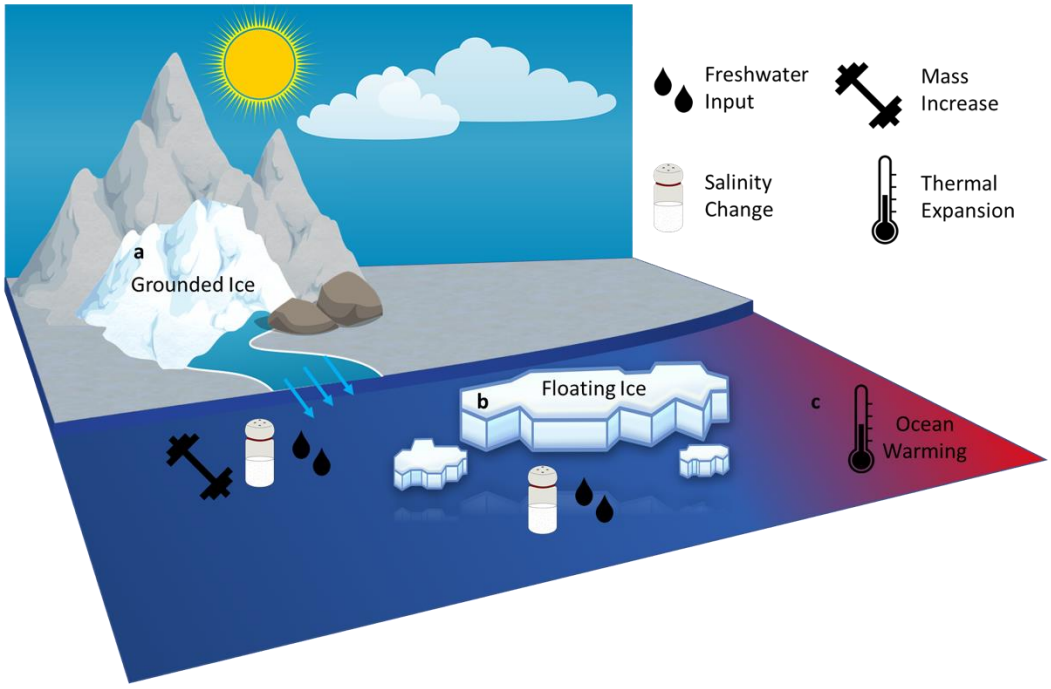
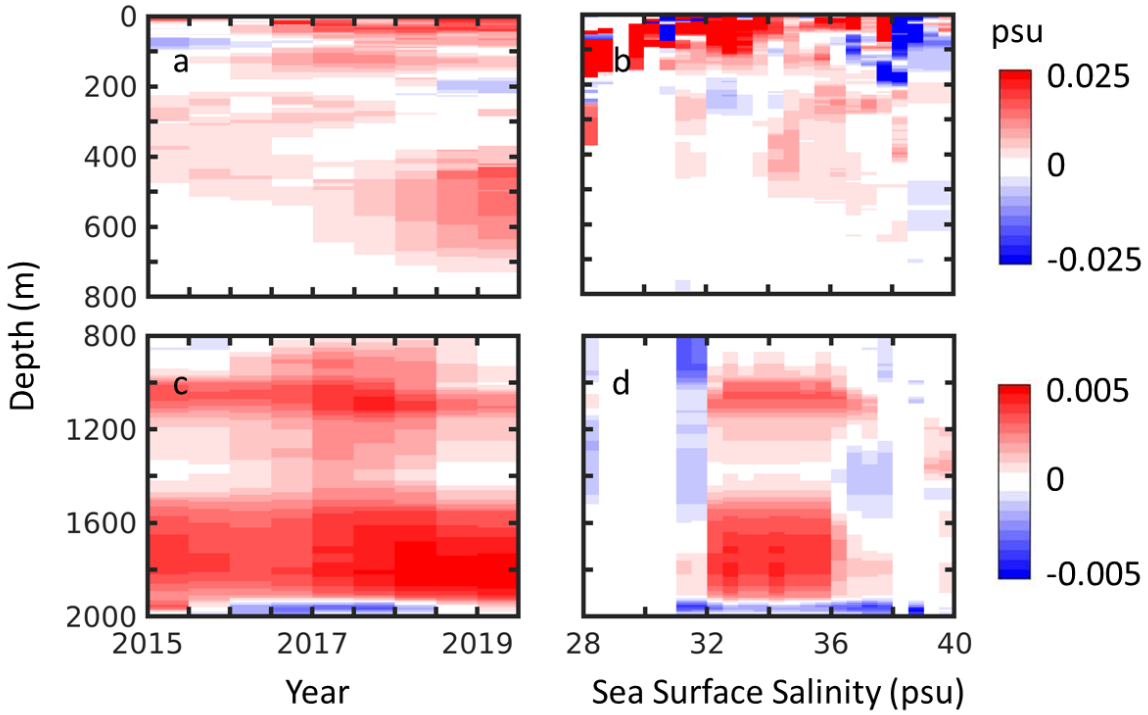


Fig. III.1 Constituents of global mean sea level rise (GMSLR) and freshwater input and their resulting impacts on the ocean. Meltwaters from grounded ice (a) such as glaciers and the continental ice sheets directly contribute to GMSLR by adding mass to the ocean, known as barystatic sea level rise, as freshwater flows from the land to the sea. This meltwater also dilutes the ocean, causing salinity to drop. Floating ice (b) displaces an equivalent mass of seawater *(*Gregory *et al.*, 2019), and so the melting of floating ice does not induce a mass change in the ocean, but it does produce a drop in salinity when freshwater is added. This leads to a very small (generally negligible) GMSLR due to the reduction of seawater density and consequent increase in ocean volume (Jenkins and Holland, 2007). Additionally, ocean warming (c) causes thermosteric sea level rise due to the thermal expansion of seawater as it is heated.

(Gregory *et al.*, 2019), and is a fundamental metric of global climate change (Oppenheimer *et al.*, 2019). GMSLR is currently monitored by orbiting satellite altimeters, which have measured a rate of global sea level rise of 3.4±0.5 mm/yr from 1993-2019 (Beckley *et al.*, 2017).

GMSLR is driven by ocean warming and thermal expansion (thermosteric sea level rise), and by freshwater runoff into the ocean from melting of grounded ice including glaciers and ice sheets (barystatic sea level rise; Fig. III.1) (Gregory *et al.*, 2019). The thermosteric component of GMSLR can be derived from historical measurements of seawater temperature which have indicated about 1.2±0.2 mm/yr of sea level rise (SLR) since 1991 (Cheng *et al.*, 2017; Levitus *et al.*, 2012; Ishii *et al.*, 2017). The barystatic component can in principle be derived from measurements of seawater salinity, since barystatic SLR is driven



Supplementary Fig. III.1 Mean difference in the uncorrected versus corrected ARGO salinity measurements as a function of depth and year for a. 0-700 m depth and c. 800-2000 m depth and as a function of depth and climatological sea surface salinity from the World Ocean Atlas (Garcia, *et al.*, 2019) for b. 0-700 m depth and d. 800-2000 m.

by the input of freshwater into the ocean that reduces ocean salinity (Munk, 2003) (Fig. III.1). However, previous attempts to quantify barystatic SLR using ocean freshening trends have been hampered by the small signal of ocean freshening relative to the large natural variability of regional ocean salinity (Llovel *et al.*, 2019; Wang *et al.*, 2017), by insufficient sampling coverage (Llovel *et al.*, 2019; Wang *et al.*, 2017), and by instrumental biases (Wang *et al.*, 2017). Because of these issues, global ocean mass budgets constructed from ocean salinity measurements show unrealistically large interannual to decadal variability and uncertain long-term trends (Munk, 2003; Llovel *et al.*, 2019; Wang *et al.*, 2017). Therefore, barystatic SLR has been inferred from the difference between the observed SLR and the thermosteric SLR (Llovel *et al.*, 2019), by remote sensing of changes in ocean mass by satellite gravity measurements (Watkins *et al.*, 2015), or by remote sensing of mass losses from glaciers and ice sheets (Bamber *et al.*, 2018).

Here, we develop a new method that produces accurate and precise estimates of barystatic SLR from 2001-2019 using ocean salinity measurements. Our approach first corrects known instrumental biases in ocean salinity measurements from Argo autonomous profiling floats since 2015 (see Methods and Supplementary Fig. III.1). We then combine this corrected dataset with other salinity data from the World Ocean Database (Boyer *et al.*, 2018) and use an ensemble autoregressive artificial neural network (ARANN) to fill gaps in the observational record (Bagnell and DeVries, 2021), providing a salinity reconstruction with global coverage of the ocean from 0-5500 m depth (see Methods). These salinity reconstructions are compared with previous salinity reconstructions that covered the upper 2000 m of the ocean in Supplementary Fig. III.2. The ARANN method is also validated against output from climate model salinity simulations in Supplementary Figs. III.3-6,

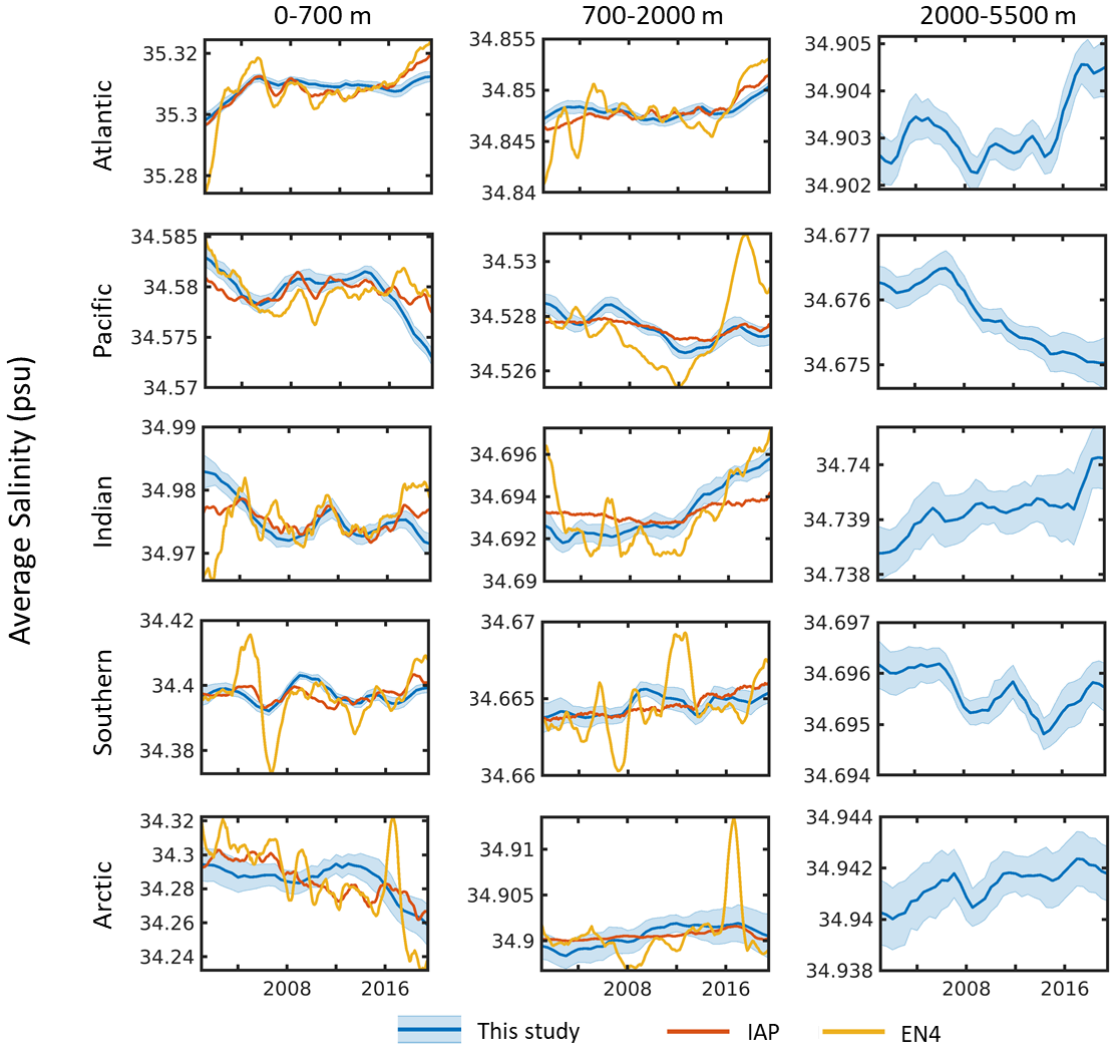
demonstrating the ability to accurately reconstruct the modeled salinity from the sparse distribution of salinity measurements at the global and basin scale.

B. Methods

Our salinity mapping approach utilizes historical observations from 2001-2019 consisting of individual salinity casts from the World Ocean Database (WOD) 2018 (Boyer *et al.*, 2018), including ocean station data (OSD), conductivity-temperature-depth (CTD) profiles, autonomous profiling floats (PFL, primarily Argo floats), and autonomous pinniped bathythermographs (APB). We quality controlled these data using the flags from the World Ocean Database, excluding any data with a flag other than 0, as well as casts that contained less than 5 samples. Individual casts were then linearly interpolated to 5 m standard depths.

Even after this quality control, there remains an existing documented bias in Argo salinity measurements due to instrumental drift for floats deployed after 2014 (Wang *et al.*, 2017). Using high quality in-situ salinity measurements from other instruments as a baseline, we find that the salinity bias in Argo floats after 2014 varies with salinity, depth, and time (Supplementary Fig. III.1). The magnitude of the bias has increased over time, leading to a large rise in global salinity since 2015 in products that use this data without any bias correction (Fig. III.2a). To correct this bias, a high-quality reference dataset of CTD/OSD salinity samples was collocated to the PFL casts to within the same 1 degree of latitude and longitude and 30 calendar days of sampling. The bias was then calculated by taking the difference between the measured PFL salinity and the average of the collocated OSD/CTD salinity. These biases were then binned to a three-dimensional grid with six-month time resolution for sampling dates from Jan 2015-Dec 2019, 5-meter depth resolution

up to a depth of 2000 m, and 0.5 PSU bins for salinity, as calculated from the WOA 2005-2017 climatology (Garcia *et al.*, 2019). No correction was applied to Argo data below 2000 m due to data sparsity. This yields a 3-dimensional lookup table for the PFL salinity bias (Supplementary Fig. III.1), which we subtract from the corresponding Argo salinity measurements for years 2015 and later.



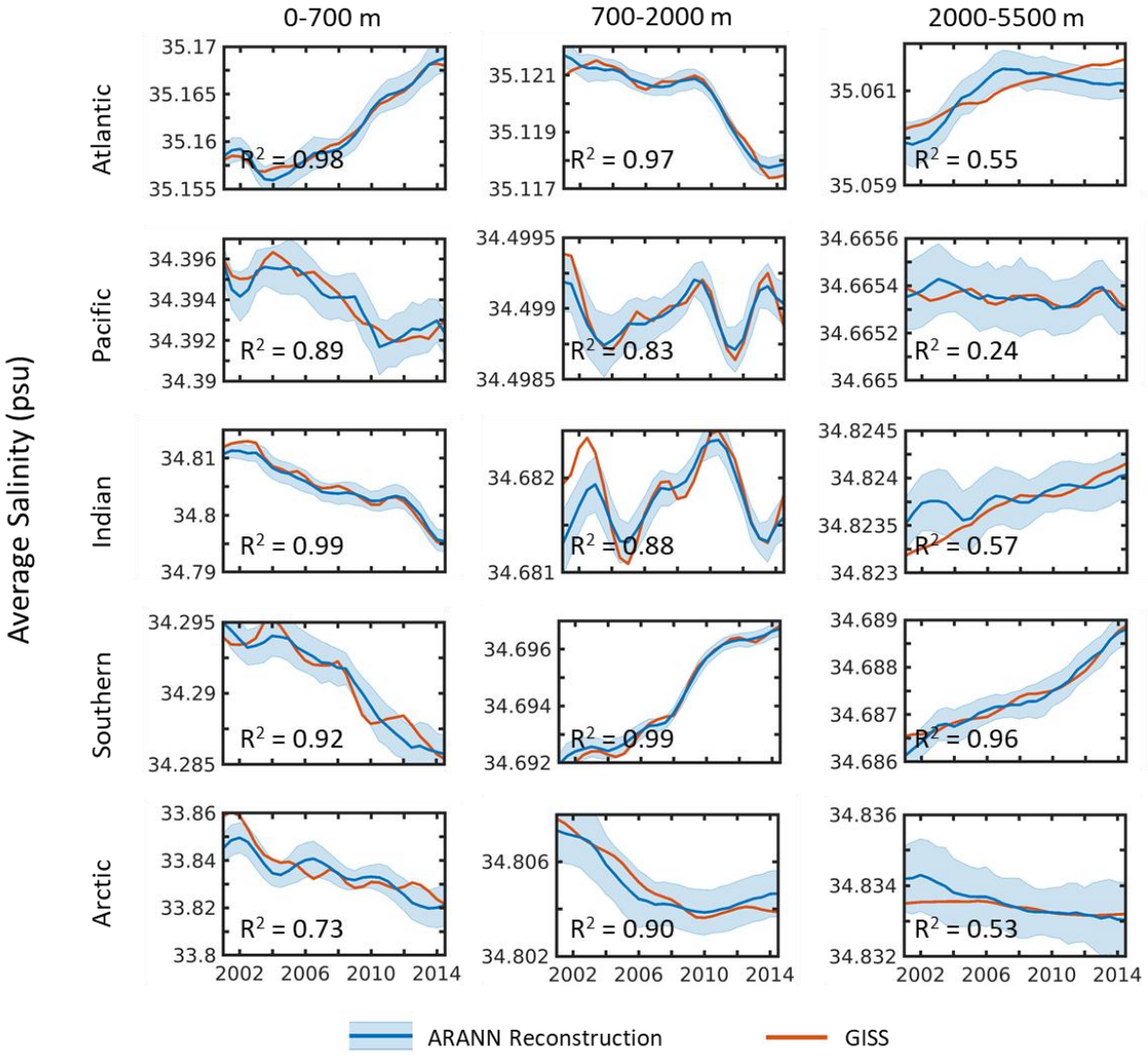
Supplementary Fig. III.2 Volume averaged salinities for the ARANN reconstruction (blue) of salinity observations derived from the World Ocean Database (Boyer *et al.*, 2018) as well as the IAP (Wang *et al.*, 2017) (red) and EN4 (Good *et al.*, 2013) (yellow) salinity reconstructions in the Atlantic, Pacific, Indian, Southern and Arctic Oceans are plotted over the depth intervals 0-700 m, 700-2000m, and 2000-5500 m. Error bars represent 2σ cross-ensemble uncertainty for the ARANN.

This corrected PFL dataset, along with the salinities from the OSD, CTD, and APB datasets were binned, taking the median value, to a regular grid with 100 km horizontal resolution and 132 depth levels (10 m resolution in the top 700 m, 50 m resolution from 700-2000 m, and 100 m resolution from 2000-5500 m) at monthly time intervals. Salinity anomalies were then calculated by subtracting the WOA 2005-2017 monthly climatology from the gridded salinities. Finally, we smoothed the resulting monthly anomaly maps using a 12-month moving average, and then binned the smoothed monthly anomaly maps to 6-month time intervals that spanned Oct-Mar and Apr-Sep.

An ensemble of 60 autoregressive neural networks (ARANN) interpolated the salinity anomalies producing global anomaly fields from 0-5500 m for years 2001-2019. The ARANN method used here follows the approach previously used to interpolate temperature anomalies for 1946-2019 (Bagnell and DeVries, 2021). First, an Artificial Neural Network (ANN) is used to interpolate salinity anomalies for the time-step covering Apr-Sep 2019, using sinusoidal basis functions in an iterative process that sweeps from the surface to 5500 m covering 12-24 depth layers at a time as in Bagnell and DeVries (2021). Then, another ANN is trained to interpolate salinity anomalies for the previous time-step (Oct 2018-Mar 2019), using the interpolated anomalies from the prior time step as an additional input to the ANN (this is the autoregressive step). The process is repeated, continuing backwards to the year 2001 and using up to the last 5 time steps (i.e. 2.5 years) as additional inputs. After this "backwards" run, a forward run is initiated starting from Oct-Mar of 2000-2001 and iterating forward at 1/2-year intervals to the year 2019, this time using up to 2 prior and 2 subsequent time steps, as well as the current time step from the backwards run, as additional inputs to the ANN. This forward run yields the final salinity maps used in this study. This is repeated

60 times, varying the training data used and number of depth layers used at each time step, as in Bagnell and DeVries (2021).

The major modification in this study to the algorithm initially used for assessing ocean heat content (Bagnell and DeVries, 2021) involves the implementation of a constraint that



Supplementary Fig. III.3 Volume averaged salinities for the ARANN reconstruction (blue) of a single run of the GISS CMIP6 climate model and the original modeled salinity (red) in the Atlantic, Pacific, Indian, Southern and Arctic Oceans are plotted over the depth intervals 0-700 m, 700-2000m, and 2000-4800 m. Error bars represent 2σ cross-ensemble uncertainty for the ARANN. R^2 values showing the performance of the ARANN versus the original modeled salinity are given for each basin and depth interval.

prevents excessive variability in the globally averaged salinity from one time-step to another. This constraint helps to avoid unrealistic interannual variability in globally averaged salinity, which is tightly constrained by the mass of ice that undergoes melting/freezing on an annual basis. This constraint is integrated within the cost function of the ARANN model, which is

$$\mathrm{cost}(t,z) = \sum_{i=1}^{M} \left(S(t,z)^{i}_{intp} - S(t,z)^{i}_{obs}\right)^{2} / M + C(t) \qquad (1)$$

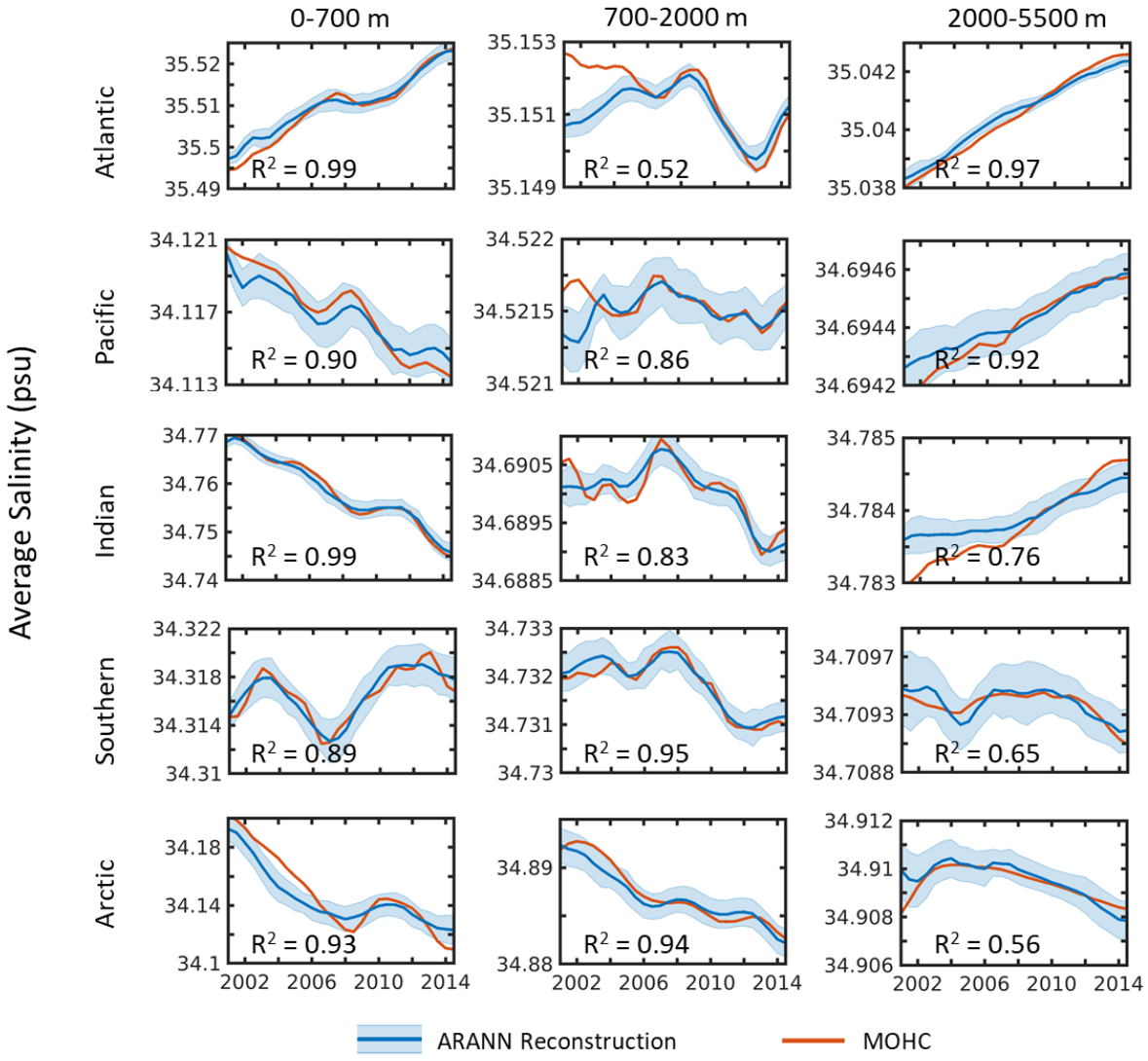
where the cost function calculates the mean squared error of the interpolated anomalies $\left(S_{intp}\right)$ versus the observed anomalies (S_{obs}) over an index (i) with (M) entries corresponding to cells in our regular grid where observations exist for the current time step (t) and depth interval (z). The second term in the cost function is the temporal smoothness constraint, which is

$$C(t) = \left| \sum_{j=1}^{N} S(t)^{j}_{intp} dV^{j} - \sum_{j=1}^{N} S(t-1)^{j}_{intp}\, dV^{j} \right| / \sum_{j=1}^{N} dV^{j} \qquad (2)$$

where the volume weighted sum of interpolated anomalies $\left(S_{intp}\right)$ for the current time step (t) are compared to the interpolated anomalies from the time step prior $(t-1)$ over an index (j) with (N) entries corresponding to all of the ocean grid cells shallower than the deepest layer of the current depth interval for the backwards run or the full depth of the ocean for the forward run. This imposes a global smoothness constraint that can be directly optimized within the cost function. To prevent the smoothing constraint from imposing too strong of a penalty, thus artificially reducing inter-annual variability in the anomalies below observed levels of natural variability, we relax the criteria such that

$$C = \begin{cases} C, & \text{and} C > \sum \frac{\Delta\bar{S}_{obs}}{n} + 2\sigma[\Delta\bar{S}_{obs}] \\ 0, & \text{and} C \leq \sum \frac{\Delta\bar{S}_{obs}}{n} + 2\sigma[\Delta\bar{S}_{obs}] \end{cases} \quad (3)$$

which prevents the smoothing criteria from being less than 2 standard deviations above the average of the time-differenced global mean anomalies ($\Delta\bar{S}_{obs}$), where



Supplementary Fig. III.4 Volume averaged salinities for the ARANN reconstruction (blue) of a single run of the Met Office Hadley Center (MOHC) CMIP6 climate model and the original modeled salinity (red) in the Atlantic, Pacific, Indian, Southern and Arctic Oceans are plotted over the depth intervals 0-700 m, 700-2000m, and 2000-4800 m. Error bars represent 2σ cross-ensemble uncertainty for the ARANN. R^2 values showing the performance of the ARANN versus the original modeled salinity are given for each basin and depth interval.

$$\Delta\bar{S}_{obs} = \bar{S}^{t}_{obs} - \bar{S}^{t-0.5}_{obs} \text{ for } t = 2005, 2005.5 \ldots 2019.5, \quad (4)$$

n is 30, the number of time steps between the years 2005 and 2019.5, and $\bar{S}^{t}_{obs}$ is taken by volume averaging the gridded salinity anomalies where observations exist at each time-step (t).

We produce an ensemble consisting of 60 ARANNs, which randomly vary the salinity data used for training and validation, as well as the size of the depth interval for each time-step. This leads to 60 sets of salinity fields for the years 2001-2019, over which we can quantify uncertainty. Additionally, 240 temperature ensemble members from Bagnell and DeVries (2021) are utilized.

The salinity (temperature) anomalies are converted to densities by adding back the WOA climatology and entering these values into a seawater equation of state calculation (McDougall and Barker, 2011), with climatological values of temperature (salinity) from the WOA used for the halosteric (thermosteric) density calculation.

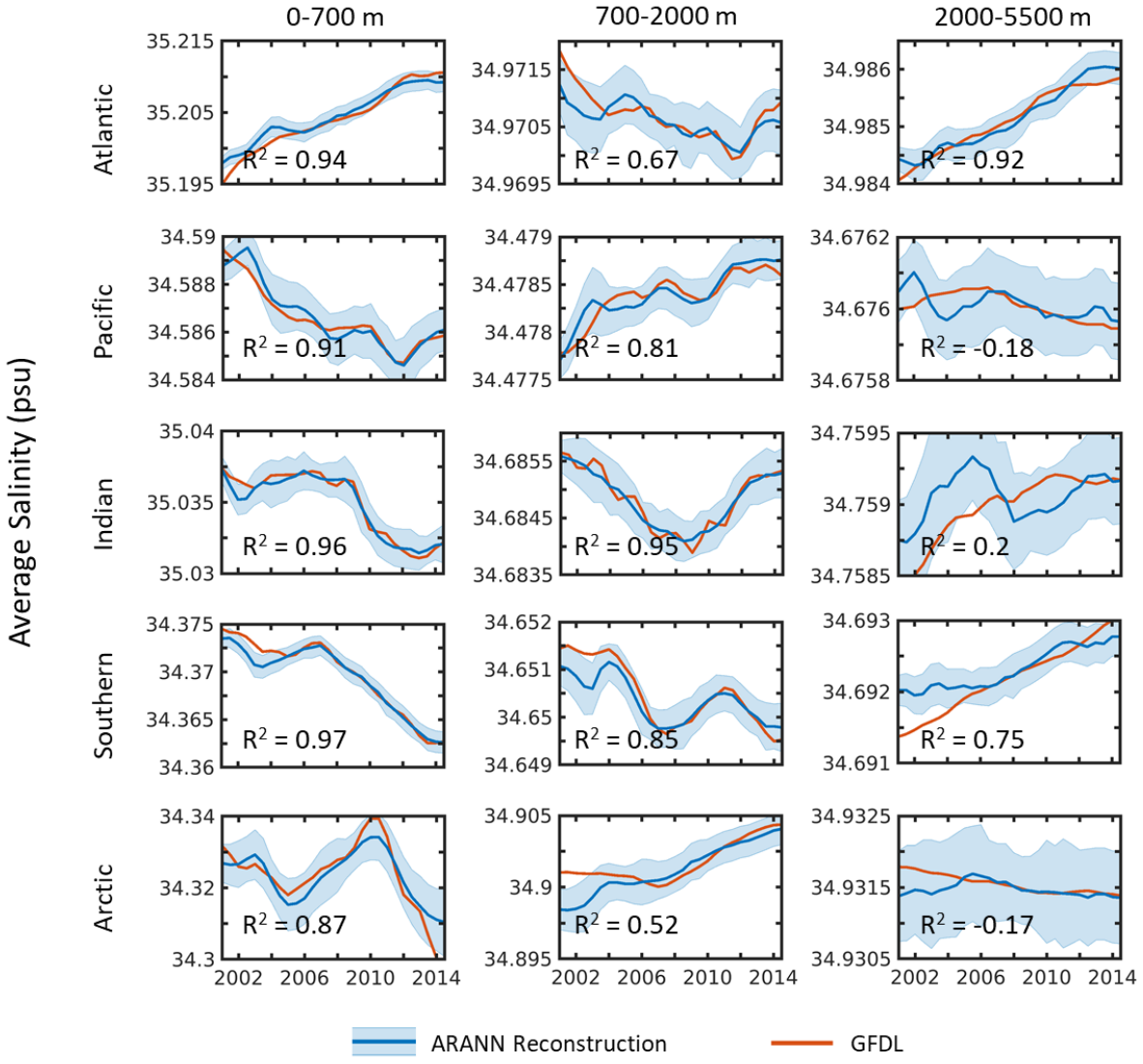
The mass of freshwater input (FW) can be estimated from the halosteric density change as

$$FW = -f_{Munk} \int \Delta\rho_{hal} dV \approx -f_{Munk}\, c\, \bar{S}_{VOL} \quad (5)$$

where the Munk factor (f_{Munk}) arises from conservation of mass (Munk, 2003) in the ocean and is expressed via a density relationship between the density of freshwater (ρ_f) at 1,000 kg/m^3 and the average density of seawater (ρ_*) at 1,028 kg/m^3

$$f_{Munk} = \frac{\rho_*}{\rho_* - \rho_f} = 36.7. \quad (6)$$

This freshwater input can also be approximated by a linearization that includes the Munk factor, the global volume averaged salinity ($\bar{S}_{VOL}$) and a conversion factor (c) of roughly 1100 Tt/psu.



Supplementary Fig. III.5 Volume averaged salinities for the ARANN reconstruction (blue) of a single run of the GFDL CMIP6 climate model and the original modeled salinity (red) in the Atlantic, Pacific, Indian, Southern and Arctic Oceans are plotted over the depth intervals 0-700 m, 700-2000m, and 2000-4800 m. Error bars represent 2σ cross-ensemble uncertainty for the ARANN. R^2 values showing the performance of the ARANN versus the original modeled salinity are given for each basin and depth interval.

In our study, we compare this estimate of freshwater input derived from the salinity budget to one derived from an ice budget approach. This includes freshwater input from floating ice (FW_{float}), whose weight is supported by buoyancy following Archimedes principle and thus does not increase barystatic sea level when melted (Gregory *et al.*, 2019; Llovel *et al.*, 2019; Jenkens and Holland, 2007; Slater *et al.*, 2021). Changes to the mass of the ocean are then

$$\Delta M_{ocean} = FW - FW_{float} \qquad (7)$$

We use the estimate of the freshwater input from floating ice derived from an ice budget approach for 2001-2017 (Slater *et al.*, 2021) and combined estimates of Arctic sea ice (Schweiger *et al.*, 2011) and Antarctic shelf ice for 2018-2019 (Adusumilli *et al.*, 2020), as these separate estimates contain more up to date values. We assume no Antarctic ice shelf melting for year 2019 due to lack of data. This change in ocean mass can be compared to that estimated from the GRACE missions (Watkins *et al.*, 2015).

Barystatic sea level rise is the change in height associated with this mass change if the surface area of the ocean (A) is considered fixed at 510 million km^2

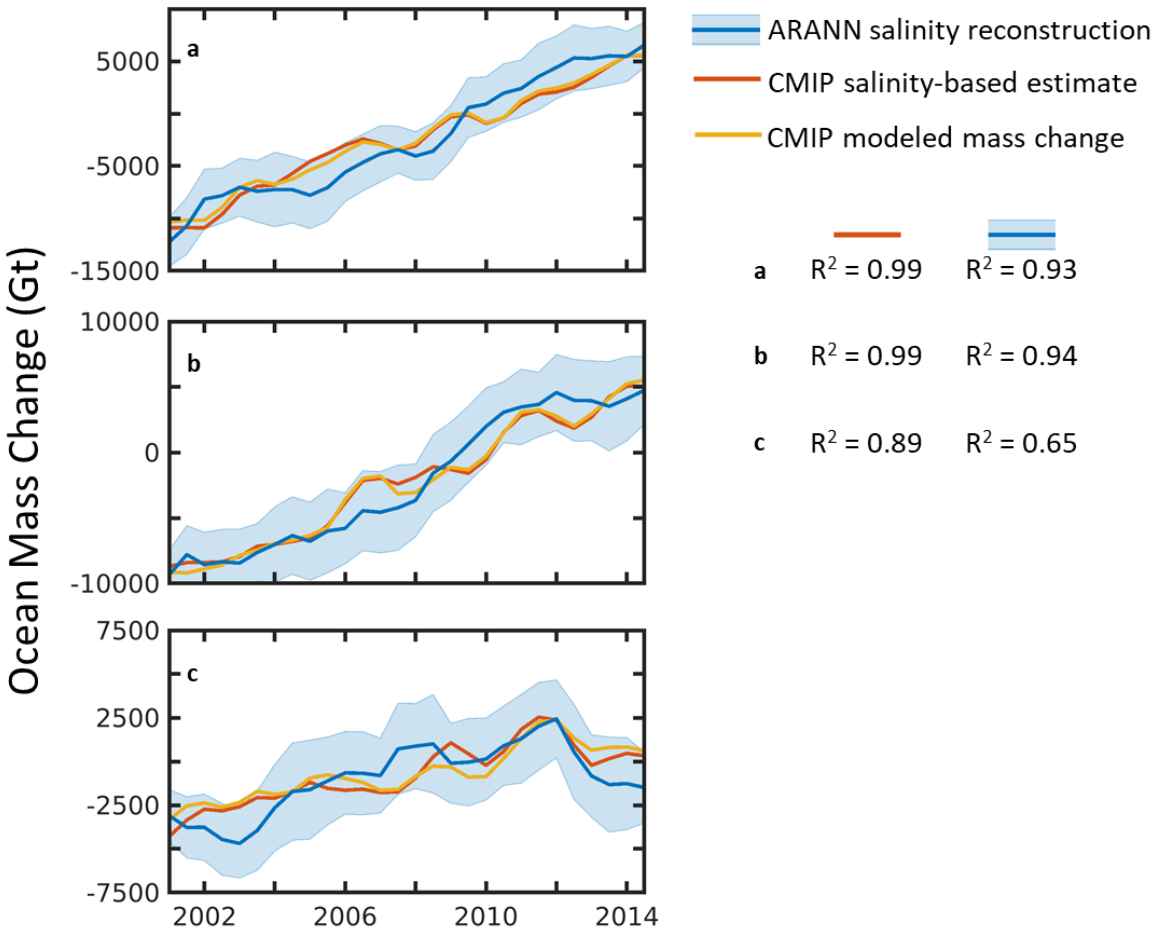
$$\Delta h_{bary} = \frac{\Delta M_{ocean}}{A\rho_f} \qquad (8)$$

Thermosteric sea level rise can be similarly estimated from the thermosteric density changes as

$$\Delta h_{therm} = -\frac{\int \Delta\rho_{therm} dV}{A\rho_*} \qquad (9)$$

Global mean sea level rise is closely approximated by summing the barystatic and thermosteric components (Gregory *et al.*, 2019), which can be compared to the estimate given by satellite altimetry (Beckley *et al.*, 2017).

The halosteric sea level rise is generally considered negligible for the global ocean (Gregory *et al.*, 2019; Munk, 2003). Freshwater input dilutes the concentration of salt in



Supplementary Fig. III.6 Full depth estimates of the ocean mass change for a. a single run of the GISS CMIP6 climate model (Miller *et al.*, 2021), b. a single run of the MOHC model (Martin *et al.*, 2020), and c. a single run of the GFDL model (Dunne *et al.*, 2020). Estimates derived from the ARANN reconstruction of salinity driven density change (blue) or from the models' original salinity fields (red) where the Munk factor (Munk, 2003) is applied are compared to the mass change estimate produced as an output of the model itself (yellow). All estimates are adjusted to have a mean anomaly of zero for the 2005-2017 time period. Error bars represent 2σ cross-ensemble uncertainty. R^2 values for the salinity-based methods versus the model output of ocean mass change are also given.

seawater and leads to a small amount of expansion. However, much of the observed salinity driven density changes correspond to the barystatic component of sea level rise, as it reflects a mass addition of freshwater from grounded ice. This should not be double counted as contributing to halosteric sea level rise (Munk, 2003). The remaining signal from floating ice approximates the global halosteric sea level rise as

$$\Delta h_{hal} = \frac{FW - \Delta M_{ocean}}{f_{Munk} A \rho_*} \qquad (10)$$

This is equivalent to stating that the melting floating ice contributes ~2.7% of its mass to sea level rise (Noerdlinger and Brower, 2007). We find that over 2001-2019, the halosteric component produced 0.8±0.2 mm of sea level rise.

To validate the salt budget approach taken here, we tested our procedure for estimating ocean mass change using three CMIP6 historical climate model runs, the NASA GISS-E2 (r1i1p1f1) (Miller *et al.*, 2021), Met Office Hadley Center (MOHC) HadGEM3-GC3 (r1i1p1f3) (Martin *et al.*, 2020), and NOAA GFDL-CM4 (r1i1p1f1) (Dunne *et al.*, 2020). Salinity anomalies from these runs were gridded to our regular 100 km grid and decimated to the observational sparsity of the in-situ salinity data for the period 2001-2014. We applied the ARANN procedure as outlined above to these modeled anomalies, producing 60 ensemble members for each CMIP6 run. These reconstructed salinities were converted to densities and compared to the density fields derived from the original modeled salinities. We also converted these estimates to ocean mass changes using the Munk factor (Munk, 2003) so that they could be directly compared to the ocean mass changes reported by the CMIP6 models themselves (Supplementary Fig. III.6).

C. Results

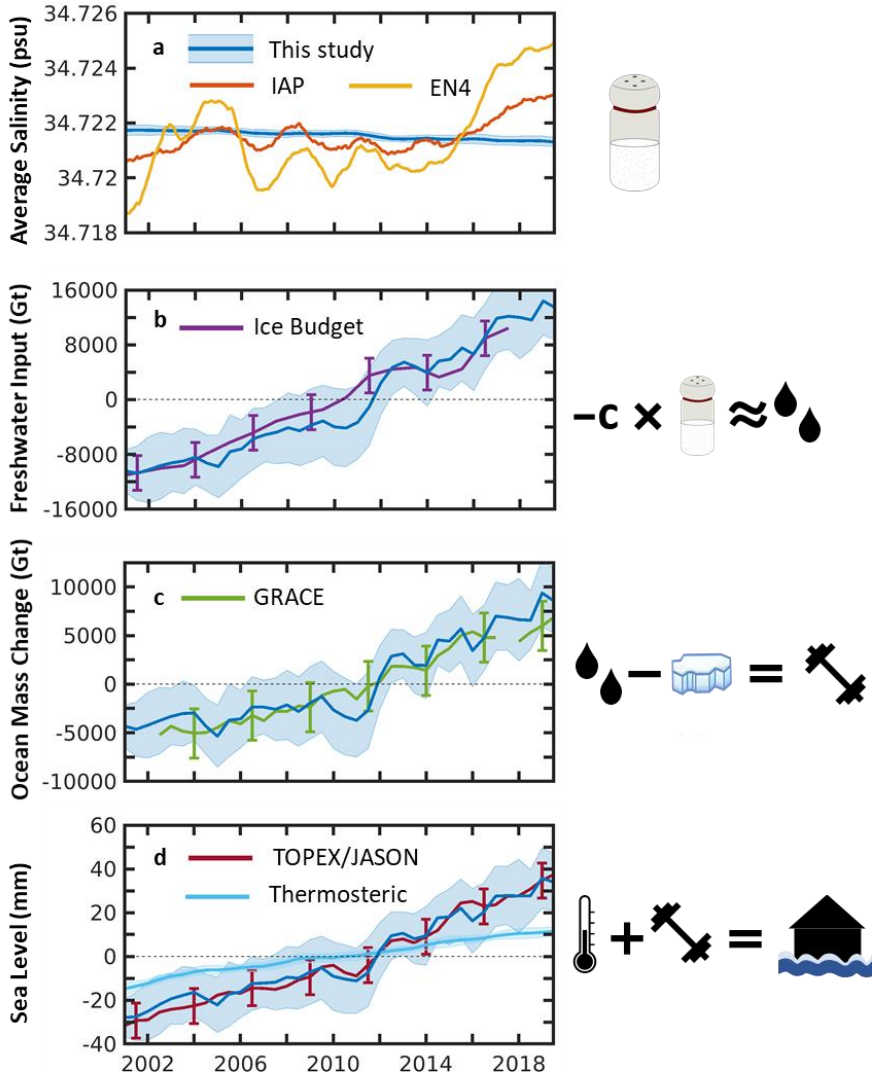


Fig. III.2 Ocean salinity, freshwater input, mass change, and sea level rise from this study and previous estimates. Global volume averaged salinity (a) derived from our study (blue) compared with two previous salinity reconstructions, the IAP red) (Wang *et al.*, 2017) and EN4 (yellow) (Good *et al.*, 2013). Freshwater input (b) as implied by our salinity approach, which is approximately equal to a constant multiplied by the global salinity change, comparted to an ice budget approach (Slater *et al.*, 2021) (purple) that converts the mass loss of grounded and floating ice to freshwater volume. Ocean mass change (c) as implied by our salinity approach resulting from panel (b) with the floating ice mass removed, compared to the GRACE estimate (Watkins *et al.*, 2015) (green), which remotely senses changes in mass above the Earth's reference geoid. Changes in global mean sea level (GMSL) (d) as calculated using the salinity-based ocean mass change from panel (c) and adding the volume from thermally driven density changes (cyan) and dividing by the total area of the ocean, compared to remotely sensed GMSL from TOPEX/JASON satellite altimeters (Beckley *et al.*, 2017) (maroon). Error envelope for the salinity approach represents the 2σ cross-ensemble variance. The 2σ uncertainty for other estimates is represented by error bars at 2.5-year intervals.

The globally-averaged salinity since 2001 from the ARANN reconstruction shows a clear, monotonic decreasing trend that is consistent with the dilution of salinity due to the continued input of freshwater from melting ice (Fig. III.2a). In contrast, previous salinity reconstructions (Wang *et al.*, 2017; Good *et al.*, 2013) do not exhibit a decreasing trend in globally averaged salinity, particularly after 2015 when the uncorrected bias in Argo salinity measurements leads to a rapid increase in salinity that is contrary to the expected freshening due to ongoing global ice melt. Even prior to 2015, there is no clear trend in salinity in these reconstructions due to their large subdecadal variability. Such a large magnitude of variability in global salinity is unrealistic with regard to freshwater fluxes into and out of the ocean due to ice melt and changes in land water storage (Slater *et al.*, 2019; Llovel *et al.*, 2011; Cazenave *et al.*, 2018). In contrast, the ARANN reconstruction displays much smaller variability, with globally averaged salinity dropping by roughly 0.0006 practical salinity units (psu), or 35+10 micro psu/yr, over the 2001-2019 period.

The total mass of freshwater responsible for this salinity change can be approximated by integrating the salinity-driven density changes over the global ocean volume and multiplying by a factor of 36.7, a relationship first expressed by Walter Munk (Munk, 2003) (see Methods, equations 5-6). This calculation yields an independent proxy for ocean freshwater input that we tested against simulated data from three CMIP6 historical climate runs ((Miller *et al.*, 2021; Martin *et al.*, 2020; Dunne *et al.*, 2020).) (Supplementary Fig. III.6), finding very good agreement between the freshwater input calculated from salinity-driven density anomalies and the freshwater input directly reported by the model (R^2 of 0.89 to 0.99, with the higher R^2 values applicable to models with more realistic trends in freshwater input). Applying this calculation to the ARANN reconstructed global salinity, we find that 24±5 Tt

of freshwater entered the ocean from 2001-2019, equivalent to a freshwater input of 1.4±0.4 Tt/yr (Fig. III.2b). Because the majority of the freshwater input to the ocean occurs from melting ice (Slater *et al.*, 2019; Llovel *et al.*, 2011; Cazenave *et al.*, 2018), the freshwater input inferred from salinity provides a strong independent constraint on the global ice budget. The most recent global ice budget (Slater *et al.*, 2019) combines satellite and modeling products to quantify individual contributions to ocean freshwater input from ice sheets, mountain glaciers, ice shelves, and sea ice, and estimates 1.2 ± 0.3 Tt/yr of freshwater input over 2000-2017, in good agreement with the ocean salinity approach (Fig. 2B). Changes in ocean salinity also record changes in the land water storage, which could contribute to the stronger interannual variability of freshwater input inferred from ocean salinity changes than appear in the ice budget (Llovel *et al.*, 2011) (Fig. III.2b). However, the impact of land water storage changes on global ocean freshwater input over multiple decades remains poorly constrained (Cazenave *et al.*, 2018).

The total freshwater input inferred from ocean salinity change includes a contribution from the melting of floating ice, which does not change the mass of the ocean and therefore does not contribute to barystatic SLR (Fig. III.1). To remove this component, we combine estimates of sea ice and ice shelf mass change from refs. (Slater *et al.*, 2019; Schweiger *et al.*, 2011; Adusumilli *et al.*, 2020), which together estimate 11±2 Tt of floating ice melt over 2001-2019. After this correction, the salinity budget approach estimates that the ocean gained a total mass of 13±3 Tt from 2001-2019 (equation 7; Fig. 2C). This estimate provides an independent constraint on the ocean mass change derived from the Gravity Recovery and Climate Experiment (GRACE) satellites which have remotely monitored ocean mass change since mid-2002 (Watkins *et al.*, 2015), with a brief interruption in 2017 during the transition

to the next mission (Fig. 2C). GRACE estimates that the ocean gained 0.8±0.1 Tt/yr from 2003-2016, compared to 0.9 ± 0.2 Tt/yr from the salinity budget approach (after correcting for floating ice melt) over the same period.

The ocean mass change inferred from global ocean freshening amounts to 35±14 mm of barystatic SLR (equation 8) from 2001-2019, with a linear trend of 2.0±0.5 mm/yr (Fig. III.2d). For the same period, 26±2 mm of additional sea level rise occurred due to the thermal expansion of seawater according to a recent full-depth reconstruction of ocean temperature anomalies (Bagnell and DeVries, 2021), for a linear trend of 1.3±0.1 mm/yr (Fig. III.2d). Overall, the combined thermosteric and barystatic SLR components estimated from ocean temperature and salinity estimate a total GMSLR of 61±15 mm over 2001-2019, indicating that global mean sea level rose at a rate of 3.4±0.6 mm/yr over that period. The sea level rise implied by ocean warming and freshening can be compared to altimetry data that has been collected for much of the past three decades by a suite of satellite missions such as TOPEX/Poseidon and Jason (Beckley *et al.*, 2017). These satellite altimetry measurements estimate a rate of GMSLR of 3.6±0.5 mm/yr from 2001-2019 (Fig. III.2d).

These results show that both the barystatic and thermosteric components contributed substantially to GMSLR from 2001-2019. By contrast, the energy absorbed by the ocean over this same period was far greater than that absorbed by the Earth's glaciers and ice caps. To induce 26 mm of sea level rise via thermal expansion, the ocean accumulated 190±30 ZJ of heat from 2001-2019 (Bagnell and DeVries, 2021), whereas the latent heat required to melt 13 Tt of ice and induce 35 mm of barystatic sea level rise would require only an additional 4.3±1.0 ZJ of heat. This translates to roughly 2% of the excess heat from Earth's energy imbalance being responsible for 57% of the GMSLR, which mainly arises because

the ocean contains much more mass than land ice and thus requires a large input of heat to prompt a meaningful change in volume.

In addition to providing accurate estimates of the magnitude and trend of GMSLR from 2001-2019, our salinity- and temperature-based approach can also capture subdecadal variability in GMSL given sufficient sampling coverage and data quality. Subdecadal variability is driven almost exlusively by the barystatic component of GMSL (Fig. III.2d) and so accurate reconstruction of short-term variability in GMSL depends on the availability of high-quality salinity data. These conditions were met during the period 2010-2014, when improvements to sampling coverage of the ocean coincided with the increased density of the Argo array (Roemmich *et al.*, 2019). For example, the salinity approach captures a reduction in sea level due to the 2011 La Nina (Boening *et al.*, 2012) and again captures a slight reduction in GMSL in that appears in the altimeter data in 2013 (Fig. III.2d). After 2015 the trend is reconstructed, but interannual variability becomes more difficult to capture due to biases in Argo salinity measurements that may not be completely corrected by our approach. Nevertheless, these results suggest that with sufficient data coverage and no instrumental biases, ocean salinity measurements will be useful for monitoring not only the magnitude and trend of GMSLR, but the interannual to decadal variability as well, providing independent constraints on the global ice budget and land water storage that drive short-term variations in GMSLR (Slater *et al.*, 2019; Llovel *et al.*, 2011; Cazenave *et al.*, 2018).

D. Discussion

These results demonstrate that GMSLR can be accurately measured using ocean-based observing systems to monitor seawater salinity and temperature. The uncertainty on the

resulting GMSLR trend over the period 2001-2019 is ~0.6 mm/yr (2σ uncertainty level) and arises mainly from the barystatic component which is derived from ocean salinity and has an uncertainty of ±0.5 mm/yr. This uncertainty is significantly less than the uncertainty on the trend of barystatic SLR from 2005-2015 based on a suite of Argo salinity products, which reported a 2σ uncertainty level of ±2.4 mm/yr, approximately the same as the total trend (Llovel *et al.*, 2019). The lower uncertainty in our study arises from an improved autoregressive interpolation method that covers the entire ocean (including the Arctic Ocean and depths below 2000 m) and that reduces unrealistic interannual variability in ocean mass changes (see Methods). Furthermore, our approach is able to extend the analysis period to 2019 by correcting for instrumental biases that are present after 2015.

The great value of this approach is that it provides a completely independent estimate of ocean freshwater input, ocean mass changes, and barystatic sea level rise that can be compared to current estimates that rely mainly on remote sensing instruments (Beckley *et al.*, 2017; Watkins *et al.*, 2015; Bamber *et al.*, 2018; Slater *et al.*, 2019). Having this independent estimate is extremely important for validation of our current understanding of the global budgets of energy, water, and ice, and for validating estimates of the magnitude and mechanisms of global sea level rise. One drawback of the salinity approach for estimating barystatic sea level rise is that the ocean salinity records not only freshwater input from grounded ice, but from floating ice as well. Therefore, an independent estimate of floating ice melt is required in order to derive the barystatic SLR from ocean salinity measurements. Nonetheless, the seawater salinity approach also has significant advantages compared to satellite altimetry or gravity approaches. The latter approaches require a correction for the glacial isostatic adjustment (GIA) of the Earth's surface as it recovers from

the last ice age (Tamisea, 2011), as well as an adjustment for the deformation of the ocean bottom due to increased ocean mass associated with barystatic SLR (Frederikse *et al.*, 2017). These corrections are similar in magnitude to the uncertainty on the salinity-inferred barystatic SLR, and the corrections themselves have substantial uncertainty. The sum of the thermosteric and barystatic SLR from ocean temperature and salinity therefore provides an important validation of the magnitude of the GIA-corrected GMSLR from satellite altimeter and gravity data.

Our results underscore the importance of ocean salinity as a critically important metric of global climate change. As the ocean continues to warm, it will likely become less efficient at absorbing additional heat (Romanou *et al.*, 2017; Newsom *et al.*, 2020), leaving other components of the Earth system such as the cryosphere to take up proportionally more heat. Based on our estimates of barystatic and thermosteric SLR, over 2001-2019 the cryosphere has absorbed significantly more energy per mass unit than the oceans, while air temperature increases in the polar regions have also far outpaced the global average (Cohen *et al.*, 2014). It is therefore likely that future sea level rise will be increasingly dominated by the barystatic component due to the accelerating loss of grounded ice, highlighting the importance of monitoring ocean salinity and freshwater input to the oceans. While the ARANN interpolation method significantly reduces uncertainties in global freshwater input and barystatic SLR derived from salinity, these uncertainties can be further reduced with greater sampling coverage, and by enhancing the accuracy and precision of salinity measurements. It is therefore critical to expand the capabilities of the current Argo observing system (Roemmich *et al.*, 2019) by increasing sampling coverage below 2000 m

and in the polar ice-covered regions, as well as by eliminating systematic instrumental biases.

References

1. Abraham, J. P., and Coauthors, 2013: A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. *Rev. Geophys.*, **51**, 450-483.
2. Adusumilli, S., Fricker, H. A., Medley, B., Laurie, P., and Siegfried, M. R. 2020: Interannual variations in meltwater input to the Southern Ocean from Antarctic ice shelves. *Nat. Geosci.* **13**, 616-620.
3. Allan, R. P., Liu, C., Loeb, N. G., Palmer, M. D., Roberts, M., Smith, D. and Vidale, P. -L. 2014: Changes in global net radiative imbalance 1985-2012. *Geophys. Res. Lett.* 5588-5597.
4. Armour, K. C., Bitz, C. M., Roe, G. H. 2013: Time-varying climate sensitivity from regional feedbacks. *J. Clim.* **26**, 4518-4534.
5. Baggenstos, D., and Coauthors 2019: Earth's radiative imbalance from the Last Glacial Maximum to the present. *Proc. Nat. Academy Sci.* **116**, 14881-14886.
6. Bagnell, A. and DeVries, T. J. 2020: Correcting biases in historical bathythermograph data using artificial neural networks. *J. Atmos. Oceanic Technol.* **37**, 1781-1800.
7. Bagnell, A. and DeVries T. 2021: 20th Century Cooling of the Deep Ocean Contributed to Delayed Acceleration of Earth's Energy Imbalance. *Nat. Comm.* **12** 4604.
8. Beckley, B. D., Callahan, P. S., Hancock III, D.W., Mitchum, G. T., Ray, R. D. 2017: On the "Cal-Mode" Correction to TOPEX Satellite Altimetry and Its Effect on the Global Mean Sea Level Time Series. *JGR Oceans, **122***, 8371-8384.
9. Boening, C., Willis, J., Landerer, F., Nerem, R., Fasullo, J. (2012). The 2011 La Nina: so strong, the oceans fell. *Geophys. Res. Lett.* **39**, GL053055.
10. Boisséson, E., Balmaseda, M., and Mayer, M. 2018: Ocean heat content variability in an ensemble of twentieth century ocean reanalyses. *Clim. Dynam.* **50**, 3783-3798.
11. Boyer, T.P., and Coauthors, 2018: World Ocean Database 2018. A. V. Mishonov, Technical Ed., NOAA Atlas NESDIS 87.
12. Boyer, T.P., and Coauthors, 2016: Sensitivity of global upper-ocean heat content estimates to mapping methods, XBT bias corrections, and baseline climatologies. *J. Climate* **29**, 4817–4842.
13. Breiman, L. 1996: Bagging predictors. *Machine Learning*, **24**, 123-140.
14. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V. 2018: Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature* **556**, 191-196.
15. Cazenave, A., Palanisamy H., and Ablain, M. 2018: Contemporary sea level changes from satellite altimetry: What have we learned? What are the new challenges? *Adv. Space Res.* **62**, 1639-1653.

16. Cheng, L., and Coauthors, 2016: XBT Science: Assessment of Instrumental Biases and Errors. *Bull. Amer. Meteor. Soc.*, **97**, 924-933.

17. Cheng, L., H. Luo, T. Boyer, R. Cowley, J. Abraham, V. Gouretski, F. Reseghetti, and J. Zhu, 2018: How Well Can We Correct Systematic Errors in Historical XBT Data? *J. Atmos. Oceanic Technol.*, **35**, 1103-1125.

18. Cheng, L., J. Zhu, F. Reseghetti, and Q. Liu, 2011: A New Method to Estimate the Systematical Biases of Expendable Bathythermograph. *J. Atmos. Oceanic Technol.*, **28**, 244-265.

19. Cheng, L., J. Zhu, 2014a: Uncertainties of the Ocean Heat Content Estimation Induced by Insufficient Vertical Resolution of Historical Ocean Subsurface Observations. *J. Atmos. Oceanic Technol.*, **32**, 2253-2263.

20. Cheng, L., J. Zhu, R. Cowley, T. Boyer, and S. Wijffels, 2014: Time, Probe Type, and Temperature Variable Bias Corrections to Historical Expendable Bathythermograph Observations. *J. Atmos. Oceanic Technol.*, **31**, 1793-1825.

21. Cheng, L. and Zhu, J. 2015: Influences of the choice of climatology on ocean heat content estimation. *J. Atmos. Oceanic. Technol.* **32**, 388-394.

22. Cheng, L., K. Trenberth, J. Fasullo, T. Boyer, J. Abraham, and J. Zhu, 2017: Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.*, **3**.

23. Church, J., N. and Coauthors, 2011: Revisiting the Earth's sea-level and energy budgets from 1961 to 2008. *Geophys. Res. Lett*., **38**.

24. Cohen, J., Screen, J., Furtado, J., Barlow, M., Whittleston, D., Coumou, D., Francis, J., Dethloff, K., Entekhabi, D., Overland, J., Jones, J. 2014: Recent Arctic amplification and extreme mid-latitude weather. *Nat. Geosci*., **7**, 627-637.

25. Couper, B.K., and E.C. LaFond, 1970: The Mechanical Bathythermograph: An Historical Review. *Adv. Instrum*, **25**, 735-770.

26. Cowley, R., S. Wijffels, L. Cheng, T. Boyer, and S. Kizu, 2013: Biases in Expendable Bathythermograph Data: A New View Based on Historical Side-by-Side Comparisons. *J. Atmos. Oceanic Technol.*, **30**, 1195-1225.

27. Desbruyeres, D. G., Purkey, S. G., McDonagh, E. L., Johnson, G. C. and King, B. A. 2016: Deep and abyssal ocean warming from 35 years of repeat hydrography. *Geophys. Res. Lett.* **43** 10,356-10,365.

28. Domingues, C., J. Church, N. White, P. Gleckler, S. Wijffels, P. Barker. And J. Dunn, 2008: Improved estimates of upper-ocean warming and multi-decadal sea-level rise. *Nature*, **453**, 1090-1093.

29. Dunne, J.P., and Coauthors 2020: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics. *J. Adv. Model. Earth Sys.*, **12**, e2019MS002015.

30. England, M. H., McGregor, S., Spence, P., Meehl, G. A., Timmermann, A., Cai, W., Gupta, A. S., McPhaden, M. J., Purich, A. and Santoso, A. 2014: Recent

intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Clim. Change* **4**, 222-227.

31. Foresee, F. D., and Hagan, M. T., 1997: Gauss-Newton approximation to Bayesian learning. In Proceedings of International Conference on Neural Networks (ICNN'97), **3**, 1930-1935.
32. Frederikse, T., Riccardo, R., and M. A. King. 2017: Ocean mass deformation due to present-day mass redistribution and its impact on sea level observations. *Geophys. Res. Lett.* **44**, 12,306-12,314.
33. Fröhlich, C. 2006: Solar irradiance variability since 1978. *Space Sci. Rev.* **125**, 53-65.
34. Garcia, H. E., Boyer, T. P., Baranova, O. K., Locarnini, R. A., Mishonov, A. V., Grodsky, A., Paver, C. R., Weathers, K. W., Smolyar, I. V., Reagan, J. R., Seidov, D. and Zweng, M. M. 2019: World Ocean Atlas 2018: Product Documentation, A. Mishonov, Tech Ed.
35. Gebbie, G. and Huybers, P. 2019: The Little Ice Age and 20th-century deep Pacific cooling. *Science* **363**, 70-74.
36. Glorot, X., A. Bordes, and Y. Bengio, 2011: Deep sparse rectifier neural networks. *Proc. 14th Int. Conf. Artif. Intell. Stat., PMLR*, **15**, 315–323.
37. Good, S., 2011: Depth Biases in XBT Data Diagnosed Using Bathymetry Data. *J. Atmos. Oceanic Technol.*, **28**, 287-300.
38. Good, S. A., Martin, M. J., and Rayner, N. A. 2013: En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res. Oceans* **118**, 6704-6716.
39. Gouretski, V., 2012: Using GEBCO digital bathymetry to infer depth biases in the XBT data. *Deep Sea Res., Part I*, **62**, 40-52.
40. Gouretski, V., and L. Cheng, 2020: Correction for Systematic Errors in the Global Dataset of Temperature Profiles from Mechanical Bathythermographs, *J. Atmos. Oceanic Technol.*, **37**, 841-855.
41. Gouretski, V., and F. Reseghetti, 2010: On depth and temperature biases in bathythermograph data: Development of a new correction scheme based on analysis of a global ocean database. *Deep Sea Res., Part I*, **57**, 812-833.
42. Gouretski, V., and K. Koltermann, 2007: How much is the ocean really warming? *Geophys. Res. Lett.*, **34**.
43. Gregory, J., Griffies, S., Hughes, C. Lowe, J., Church, J., Fukimori, I., Gomez, N., Kopp, R., Landerer, F., Cozannet, G., Ponte, R., Stammer, D., Tamisea, M., van de Wal, R. 2019: Concepts and Terminology for Sea Level: Mean, Variability and Change, Both Local and Global. *Surv. Geophys*., **40**, 1251-1289.
44. Hagan, M. T. and Menhaj, M. B. 1994: Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Networks*, **5**, 989-993.

45. Hamon, M., G. Reverdin, and P. Le Traon, 2012: Empirical Correction of XBT Data. *J. Atmos. Oceanic Technol.*, **29**, 960-973.

46. Hanawa, K., Rual,P., Bailey, R., Sy, A. and Szabados, M. 1995: A new depth-time equation for Sippican or TSK T-7, T-6 and T-4 expendable bathythermographs (XBT). *Deep Sea Res. Part I,* **42**, 1423-1451.

47. Hansen, J., and Coauthors, 2005: Earth's Energy Imbalance: Confirmation and Implications. *Science*, **308**, 1431-1435.

48. Hansen, L., and P. Salamon, 1990: Neural network ensembles. *IEEE Trans. Patt. Anal. Mach. Int.*, **12**, 993-1001.

49. Hasni, A., Sehli, A., Draoui, B., Bassou, A. and Amieur, B. 2012: Estimating global solar radiation using artificial neural network and climate data in the southwestern region of Algeria. *Energy Procedia* **18**, 531-537.

50. Haywood, J. M., Jones, A. and Jones, G. S. 2013: The impact of volcanic eruptions in the period 2000-2013 on global mean temperature trends evaluated in the HadGEM2-ES climate model. *Atmos. Sci. Lett.* **15**, 92-96.

51. Ishii, M., and M. Kimoto, 2009: Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanogr*., **65**, 287-299.

52. Ishii, M., Fukuda, Y., Hirahara, S., Yasui, S., Susuki, T., and Sato, K. 2017: Accuracy of Global Upper Ocean Heat Content Estimation Expected from Present Observational Data Sets. *SOLA* **13** 163-167.

53. Jenkins, A. and Holland, D. 2007: Melting of floating ice and sea level rise. *Geophys. Res. Lett.* **34**, L16609.

54. Jacobs, S.S. 2004: Bottom water production and its links with the thermohaline circulation. *Antarc. Sci.*, **16**, 427.

55. Johnson, G. C., Lyman, J. M. and Loeb, N. G. 2016: Improving estimates of Earth's energy imbalance. *Nat. Clim. Change* **6**, 639-640.

56. Jones, P. D. and Mann, M. E. 2004: Climate over past millennia. *Rev. Geophys*. **42**.

57. Karnauskas, K. B., Seager, R., Kaplan, A., Kushnir, Y. and Kane, M. A. 2009: Observed strengthening of the zonal sea surface temperature gradient across the equatorial Pacific Ocean. *J. Clim.* **22**, 4316-4321.

58. Kim, W. M., Yeager, S. G. and Danabasoglu, G. 2018: Key role of internal ocean dynamics in Atlantic Multidecadal Variability during the last half century. *Geophys. Res. Lett.* **45**, 13449-13457.

59. Kirezci, E., Young, I. R., Ranasinghe, R., Muis, S., Nicholls, R. J., Lincke, D and Hinkel, J. 2020: Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st Century. *Sci. Rep.* **10,** 11629.

60. Kizu, S., H. Yoritaka, and K. Hanawa, 2005: A New Fall rate Equation for T-5 Expendable Bathythermograph (XBT) by TSK. *J. Oceanogr*. **61**, 115-121.

61. Kouketsu, S. and Coauthors 2011: Deep ocean heat content changes estimated from observation and reanalysis product and their influence on sea level change. *J. Geophys. Res.: Oceans* **116**.

62. Kramer, R. J., He, H., Soden, B. J., Oreopoulos, L., Myhre, G., Forster, P. M., and Smith, C. J. 2021: Observational Evidence of Increasing Global Radiative Forcing. *Geophys. Res. Lett.* **48**, e2020GL091585.

63. Krasnopolsky, V. M., Breaker, L. C. and Gemmel, W. H. 1995: A neural network as a nonlinear transfer function model for retrieving surface wind speeds from the special sensor microwave imager. *J. Geophys. Res*. **100**, 11033-11045.

64. Lambert, A., Grainger, R. G., Rodgers, C. D., Taylor, F. W., Mergenthaler, J. L., Kumer, J. B. and Massie, S. T. 1997: Global evolution of the Mt. Pinatubo volcanic aerosols observed by the infrared limb-sounding instruments CLAES and ISAMs on the Upper Atmosphere Research Satellite. *Atmos.* **102**, 1495-1512.

65. Lee, S. -K., Park, W., Baringer, M. O., Gordon. A. L., Huber, B. and Liu, Y. 2015: Pacific origin of the abrupt increase in Indian Ocean heat content during the warming hiatus. *Nat. Geos.* **8**, 445-449.

66. Levitus, S., and Coauthors, 2012: World ocean heat content and thermosteric sea level change (0-2000 m), 1955-2010. *Geophys. Res. Lett.*, **39**.

67. Levitus, S., J. Antonov, T. Boyer, R. Locarnini, H. Garcia, and A. Mishonov, 2009: Global ocean heat content 1955-2008 in light of recently revealed instrumentation problems. *Geophys. Res. Lett.*, **36**.

68. Lincoln, W. P., and J. Skrzypek. 1990: Systematics of clustering multiple back propagation networks. Advances in Neural Information Processing Systems 2, D.S. Touretzky, Ed., Morgan Kaufmann Publishers, 650-657.

69. Liu, W., Xie, S. -P. and Lu, J. 2016: Tracking ocean heat uptake during the surface warming hiatus. *Nat. Commun.* **7**.

70. Llovel, W., Becker, M., Cazenave, A., Jevrejeva, S., Alkama, R., Decharme, B., Douville, H., Ablain, M., Beckley, B. 2011: Terrestrial waters and sea level variations on interannual time scale, *Glob. Planet. Change* **75**, 76-82.

71. Llovel, W., Purkey, S., Meryssignac, B., Blazquez, A., Kolodziejczyk, N., and Bamber, J. 2019: Global ocean freshening, ocean mass increase and global mean sea level rise over 2005–2015. *Sci. Rep.* **9**, 17717.

72. Locarnini, R. A., and Coauthors, 2013: World Ocean Atlas 2013, Volume 1: Temperature. S. Levitus, Ed., A. Mishonov Technical Ed., NOAA Atlas NESDIS 73, **40**.

73. Loeb, N., Lyman, J., Johnson, G., Allan, R., Doelling, D., Wong, T., Soden, B. and Stephens, G. 2012: Observed changes in top-of-the-atmosphere radiation and upper-ocean heating consistent within uncertainty. *Nat. Geosci.* **5**, 110-113.

74. Lyman J.M., S.A. Good, V. Gouretski, M. Ishii, G.C. Johnson, M.D. Palmer, D.M. Smith, J. Willis, 2010: Robust warming of the global upper ocean. *Nature*, **46**, 334–337.

75. MacKay, D. J., 1992: Bayesian interpolation. *Neural Computation*, **4**, 415-447.

76. Maier, H. C. and Dandy, G. C. 2001: Neural network based modelling of environmental variables: a systematic approach. *Math. Comp. Model.* **33**, 669-682.

77. Marquardt, D. 1963: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. App. Math.* **11**, 431-441.

78. Martin B. Andrews, Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E., Booth, B., Burke, E., Dittus, A., Florek, P., Gray, L., Haddad, S., Hardiman, S., Hermanson, L., Hodson, D., Hogan, E., Jones, G., Knight, J., Kuhlbrodt, T., Misios, S., Mizielinski, M., Ringer, M., Robson, J., Sutton, R. 2020: Historical Simulations With HadGEM3-GC3.1 for CMIP6. *J. Adv. Model. Earth Sys.*, **12**, e2019MS001995.

79. Masuda, S., and Coauthors 2010: Simulated rapid warming of abyssal North Pacific waters. *Science* **329**, 319-322.

80. McDougall, T.J. and P.M. Barker 2011: Getting started with TEOS-10 and the Gibbs Seawater (GSW) Oceanographic Toolbox, **28**., SCOR/IAPSO WG127, ISBN 978-0-646-55621-5.

81. Meehl G., W. Washington, W. Collins, J. Arblaster, A. Hu, L. Buja, W. Strand, and H. Teng, 2005: How much more global warming and sea level rise? *Science* **307**, 1769-1772.

82. Meehl, G. A., Hu, A., Arblaster, J. M., Fasullo, J, and Trenberth, K. E. 2013: Externally forced and internally generated decadal climate variability associated with the Interdecadal Pacific Oscillation. *J Clim.* **26**, 7298-7310.

83. Meyssignac, and Coauthors 2019: Measuring global ocean heat content to estimate the Earth energy imbalance. *Front. Mar. Sci*.

84. Miller, R.L., and Coauthors 2021: CMIP6 Historical Simulations (1850–2014) With GISS-E2.1. *J. Adv. Model. Earth Syst.*, **13**, e2019MS002034.

85. Munk, W. (2003). Ocean Freshening, Sea Level Rising. *Science*, **300**, 2041-2043.

86. Myhre, G., Shindell, D., Bréon, F. -M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J. F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura T. and Zhang, H. "Anthropogenic and Natural Radiative Forcing" in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Stocker, T.F., Qin, D., Plattner, G. -K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P. M. Eds. (Cambridge, 2013), chap. 8.

87. Newsom, E., Zanna, L., Khatiwala, S., Gregory, J. 2020: The Influence of Warming Patterns on Passive Ocean Heat Uptake. *Geophys. Res. Lett.*, **47**, e2020GL088429.

88. Noerdlinger, P. and Brower, K. 2007: The melting of floating ice raises the ocean level. Geophys. J. Internat. **170**, 145-150.

89. Oppenheimer, M., B.C. Glavovic , J. Hinkel, R. van de Wal, A.K. Magnan, A. Abd-Elgawad, R. Cai, M. Cifuentes-Jara, R.M. DeConto, T. Ghosh, J. Hay, F. Isla, B. Marzeion, B. Meyssignac, and Z. Sebesvari, 2019: Sea Level Rise and Implications

for Low-Lying Islands, Coasts and Communities. in *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, H.-O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N.M. Weyer Eds.

90. Orsi, A. H., Jognson, G. C. and Bullister, J. L. 1999: Circulating, mixing and production of Antarctic Bottom Water. *Prog. Oceanography* **43**, 55-109.

91. Palmer, M., McNeall, D. and Dunstone, N. 2011: Importance of the deep ocean for estimating decadal changes in Earth's radiation balance. *Geophys. Res. Lett.* **38**.

92. Palmer, M., D., Roberts, C. D., Balmaseda, M., Chang, Y. -S., Chepurin, G., Ferry, N., Fuji, Y., Good, S. A., Guinehut, S., Haines, K., Hernandez, F., Köhl, A., Lee, T., Martin, M. J., Masina, S., Masuda, S., Peterson, K. A., Storto, A., Toyoda, T., Valdivieso, M., Vernieres, G., Wang, O. and Xue, Y. 2017: Ocean heat content variability and change in an ensemble of ocean reanalyses. *Clim. Dynam.* **49**, 909-930.

93. Polyakov I. V., Bhatt, U. S., Simmons, H. L., Walsh D., Walsh, J. E., Zhang, X. 2005: Multidecadal variability of North Atlantic temperature and salinity during the twentieth century. *J. Clim.* **18**, 4562-4581.

94. Prechelt, L. 1998: Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* **11**, 761-767.

95. Purkey, S. G. and Johnson, G. 2010: Warming of Global Abyssal and Deep Southern Ocean Waters between the 1990s and 2000s: Contributions to Global Heat and Sea Level Rise Budgets. *J. Clim.* **23**, 6336-6351.

96. Rathore, S., Bindhoff, N. L., Phillips, H. E. and Feng, M. 2020: Recent hemispheric asymmetry in global ocean warming induced by climate change and internal variability. *Nat. Commun.* **11**, 2008.

97. Resplandy, L., Keeling, R. F., Eddebbar, Y., Brooks, M., Wang, R., Bopp, L., Long, M. C., Dunne, J. P., Koeve, W. and Oschilies, A. 2019: Quantification of ocean heat uptake from changes in atmospheric O_2 and CO_2 composition. *Scientific Reports* **9**.

98. Reverdin, G., F. Marin, B. Bourlès, and P. Lherminier, 2009: XBT Temperature Errors during French Research Cruises (1999–2007). J. *Atmos. Oceanic Technol.*, **26**, 2462-2473.

99. Robson, J., Ortega, P. and Sutton, R. 2016: A reversal in climate trends in the North Atlantic since 2005. *Nat.Geosci.* **9**, 513-517.

100. Robson, J., Sutton, R., Lohmann, K., Smith, D. and Palmer, M. 2012: Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. *J. Clim.* **25** 4116-4134.

101. Roemmich, D., and Coauthors 2019: On the Future of Argo: A Global, Full-depth, Multi-disciplinary Array. *Front. Mar. Sci.* **6**.

102. Romanou, A., Marshall, J., Kelley, M., Scott, J. 2017: Role of the ocean's AMOC in setting the uptake efficiency of transient tracers. *Geophys. Res. Lett.*, **44**, 5590-5598.

103. Schweiger, A., R. Lindsay, J. Zhang, M. Steele, H. Stern, and R. Kwok 2011: Uncertainty in modeled Arctic sea ice volume, *J. Geophys. Res.* **116**, C00D06.

104. Slater,T., Lawrence, I., Otosaka, I., Shepherd, A., Courmelen, N., Jakob, L., Tepes, P., Gilbert, L., Nienow, P. 2021: Earth's ice imbalance. *Cryosphere*, **15**, 233-246.

105. Sloyan, B. M. and Rintoul, S. R. 2001: Circulation, renewal, and modification of Antarctic Intermediate Water. *J. Phys. Oceanography* **31**, 1005-1030.

106. Smith, D. M., Allan, R. P., Coward, A. C., Eade, R., Hyder, P., Liu, C., Loeb, N. G., Palmer, M. D., Roberts, C. D. and Scaife, A. A. 2015: Earth's energy imbalance since 1960 in observations and CMIP5 models. *Geophys. Res. Lett.* **42**, 1205-1213.

107. Storelvmo, T., Leirvik, T., Lohmann, U., Phillps, P. C. B. and Wild, M. 2010: Disentangling greenhouse warming and aerosol cooling to reveal Earth's climate sensitivity. *Nat. Geosci.* 9, 286-289 (2016). Winton, M., Takahashi, K., and Held, I. M. Importance of ocean heat uptake efficiency to transient climate change. *J. Clim.* **23**, 2333-2344.

108. Tamisiea, A. 2011: Ongoing glacial isostatic contributions to observations of sea level change, *Geophys. J. Internat.***186**, 1036–1044.

109. Tatebe, H. and Watanabe, M. 2018: "MIROC MIROC6 model output prepared for CMIP6 CMIP historical" (Earth System Grid Federation).

110. Thadathil, P., A. K. Saran, V. V. Gopalakrishna, P. Vethamony, N. Araligidad, and R. Bailey, 2002: XBT Fall Rate in Waters of Extreme Temperature: A Case Study in the Antarctic Ocean. *J. Atmos. Oceanic Technol.*, **19**, 391-396.

111. Trenberth, K., J. Fasullo, M. A. Balmaseda, 2014: Earth's energy imbalance. *J. Climate*, **27**, 3129–3144.

112. Voldoire, A. and Coauthors 2019: Evaluation of CMIP6 DECK experiments with CNRM-CM6-1, *J. Adv. Model. Earth Syst.* **11**, 2177-2213.

113. von Schuckmann, K., Palmer, M. D., Trenberth, K. E., Cazenave, A., Chambers, D., Champollion, N., Hansen, J., Josey, S. A., Loeb, N., Mathieu, P. -P., Meyssignac, B. and Wild, M. 2016: An imperative to monitor Earth's energy imbalance. *Nat. Clim. Change* **6**, 138-144.

114. von Schuckmann, and Coauthors 2020: Heat Stored in the Earth System: where does the energy go? *Earth Syst. Sci. Data* **12**, 2013-2041.

115. Wang, G., Cheng, L., Boyer, T., Li, C. 2017: Halosteric Sea Level Changes during the Argo Era. *Water*, **9**, 484.

116. Wang, G., L. Cheng, J. Abraham, and C. Li, 2018: Consensuses and discrepancies of basin-scale ocean heat content changes in different ocean analyses. *Climate Dyn.*, **50**, 2471-2487.

117. Wang, B., Luo, X., Yang, Y.-M., Sun, W., Cane, M. A., Cai, W., Yeh, S.-W. and Liu, J. 2019: Historical change of El Niño properties sheds light on future changes of extreme El Niño. *Proc. Nat. Academy Sci.* **116**, 22512-22517.

118. Watkins, M., Wiese, D., Yuan, D. -N., Boening., C., Lander, F. 2015: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *JGR Solid Earth* **120**, 2648-2671.

119. Weigend, A. S., Huberman, B. A. and Rumelhart, D. E., 1990: Predicting the Future: A Connectionist Approach, *Int. J. Neural Syst.*, **1**, 193–209.

120. Wijffels, S., J. Willis, C. Domingues, P. Barker, N. White, A. Gronell, K. Ridgway, and J. Church, 2008: Changing Expendable Bathythermograph Fall Rates and Their Impact on Estimates of Thermosteric Sea Level Rise. *J. Climate*, **21**, 5657-5672.

121. Zanna, L., Khatiwala, S., Gregory, J. M., Ison, J. and Heimbach, P. 2019: Global reconstruction of historical ocean heat storage and transport. *Proc. Nat. Academy Sci.* **116**, 1126-1131.

122. Zhang, R. 2008: Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation. *Geophys. Res Lett.* **35**.

123. Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y. -O., Marsh, R., Yeager, S. G., Amrhein, D. E. and Little, C. M. 2019: A Review of the Role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and Associated Climate Impacts. *Rev. Geophys*. **57**, 316-375.