Correcting Biases in Historical Bathythermograph Data Using Artificial Neural Networks

AARON BAGNELL[a]

Interdepartmental Graduate Program in Marine Science, University of California, Santa Barbara, Santa Barbara, California

TIMOTHY DEVRIES[b]

Department of Geography, University of California, Santa Barbara, Santa Barbara, California

(Manuscript received 19 June 2019, in final form 16 July 2020)

ABSTRACT: Historical estimates of ocean heat content (OHC) are important for understanding the climate sensitivity of the Earth system and for tracking changes in Earth's energy balance over time. Prior to 2004, these estimates 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 with high-quality temperature datasets. Here, we use a new approach to reduce biases in historical BT data after binning them to a regular grid such as would be used for estimating OHC. 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 m. A global correction and corrections optimized to specific BT probe types are presented for the top 1800 m. Our approach differs from most prior studies by accounting for multiple sources of error in a single correction instead of separating the bias into several independent components. These new global and probe-specific corrections perform on par with widely used calibration methods on a series of metrics that examine the residual temperature biases with respect to a high-quality reference dataset. However, distinct patterns emerge across these various calibration methods when they are extrapolated to BT data that are not included in our cross-instrument comparison, contributing to uncertainty that will ultimately impact estimates of OHC.

KEYWORDS: Ocean; Data quality control; In situ oceanic observations; Neural networks

1. Introduction

The oceans play an enormous role in Earth's energy budget, having gained roughly 10 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 2005; Trenberth et al. 2014; Meehl et al. 2005). Reconstructions of ocean heat content 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 ocean heat content (OHC) reconstructions (Lyman et al. 2010; Boyer et al. 2016; Cheng et al. 2016, 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 roughly 60%–70% of the raw temperature casts in a given calendar year (Fig. 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° 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 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, 150, 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

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JTECH-D-19-0103.s1.

[a] ORCID: 0000-0002-4010-2824.
[b] ORCID: 0000-0002-7771-9430.

Corresponding authors: Aaron Bagnell, a-bagnell@ucsb.edu; Timothy DeVries, tdevries@geog.ucsb.edu

DOI: 10.1175/JTECH-D-19-0103.1

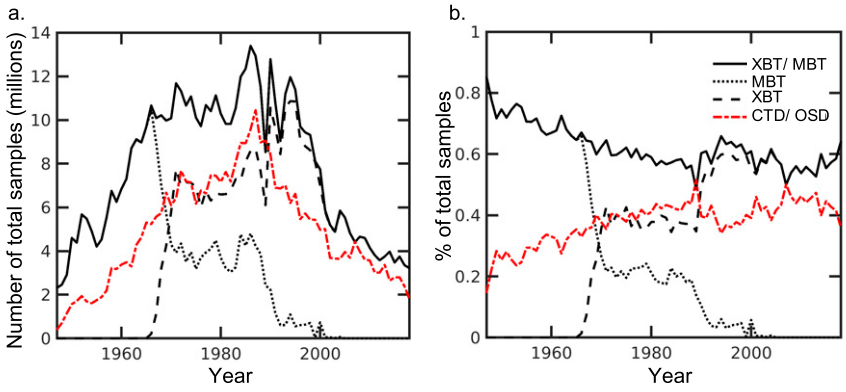

FIG. 1. The relative annual sampling coverage of individual instrument casts for (a) total discrete samples after linearly interpolating to 350 standard depths ($z_{standard}$) and (b) percentage of total samples from a particular temperature instrument.

long and continued period of use of the XBT, with most designs able to reach depths of roughly 400–700 m 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°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 subdecadal time scales. 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 intradecadal variations (Lyman et al. 2010; Cheng et al. 2014, hereinafter CH14), 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 time scales.

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, hereinafter H95) proposed a modified FRE with new coefficients to correct for this depth bias. Against the warning of H95 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 H95 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 themselves, 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, hereinafter GR10; Good 2011; Gouretski 2012; Hamon et al. 2012; Cowley et al. 2013; CH14).

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. GR10 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 naval and merchant ships with minimal adjustments to the vessels' courses or speeds. Nevertheless, these findings were corroborated by CH14 with a much larger global XBT dataset composed of probes deployed under more typical conditions. In addition, CH14 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 H95 depth correction, or subsequent depth corrections that accounted for the effects of near-surface temperature on the probe fall rate (Cheng et al. 2011) and the effect of probe design (CH14), 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 United States 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; CH14) instead of parameters that are constant with time (as with H95). 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, hereinafter L09) 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 2). 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; L09; GR10; CH14). 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 3). 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 3). We close by summarizing our main findings and highlighting unresolved issues in correcting BT biases that should be addressed by the community (section 4).

2. Data and methods

a. 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 30 September 2019). 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 data (OSD) and CTD which we use as reference data (Fig. 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 are modified to have the original manufacturer FRE (MFR FRE hereinafter) 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 H95 FRE in the late 1990s (see section 1 for further details). To modify individual casts from the 1990s onward that have been flagged by the WOD as having the H95 FRE applied, we multiply their sample depths by the factor 0.9675, following GR10, which approximates the depths that would be given by using the original manufacturer FRE. Other studies such as L09 and CH14 have instead opted to use the H95 FRE as a starting point, but some studies (Thadathil et al. 2002; GR10) have indicated that the H95 FRE may increase the time-dependent thermal bias relative to collocated CTD/OSD data as compared with 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 temperature range from −3° to 36°C, because 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 four temperature measurements as well as a measurement in the top 100 m to be included. After applying this quality control (Fig. 2, step 2), there are approximately 2.1 million XBT casts and 2.2 million MBT casts.

Following CH14, 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 (their approximate terminal depths are 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.

CH14 noted that operators have generally considered XBT probes to be able to reach depths that are roughly 20% greater than their listed terminal depths with minimal loss in accuracy. However, because of 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. 2, step 3). These depth levels are spaced 1 m apart in the top 100 m, 5 m apart from 105 to 700 m, and 10 m apart from 710 to 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° of latitude and longitude and within the same 30-day window as a cast from either the XBT or MBT datasets (Fig. 2, step 4). Since multiple

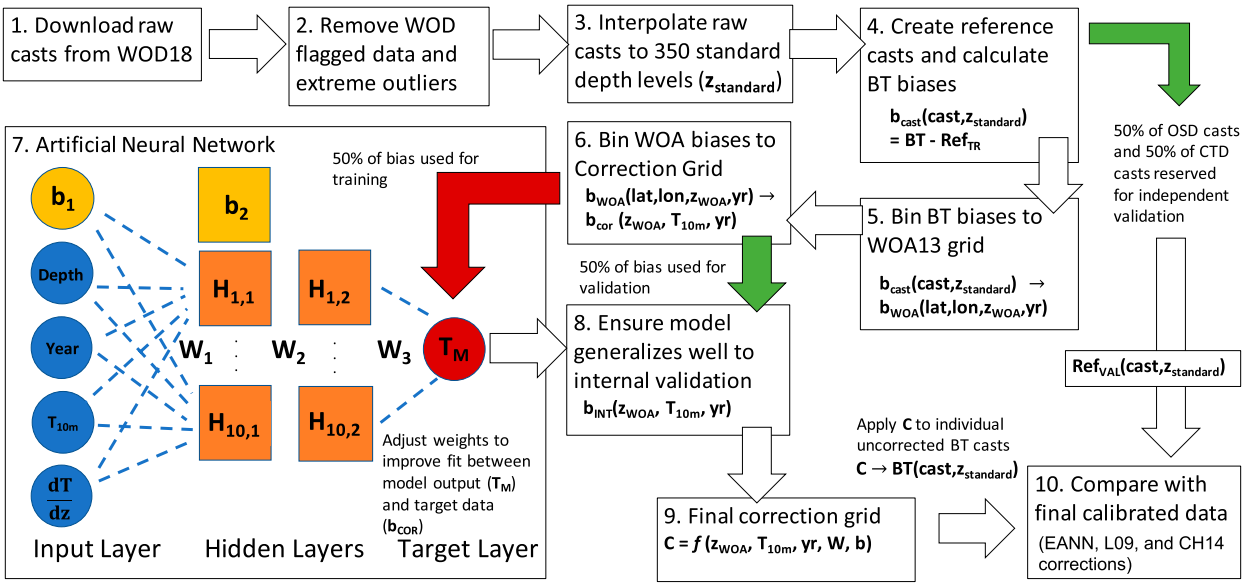

FIG. 2. Schematic of this study's data flow and quality control, as well as the application of an EANN to calibrate bathythermograph data. Steps 6–9 are repeated 30 times to create an ensemble of bias corrections. See the text for additional details on quality control, binning, and network architecture.

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 spatiotemporal sampling histories and vertical resolutions (Cheng and Zhu 2014).

After interpolation of the BT data to standard depth levels, we set aside 50% of the OSD and CTD data in order to use them for validating our calibration method. Because of 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 one-half of all reference CTD casts and one-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 are then concatenated and averaged. We refer to this dataset as 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 3). One caveat to note is that some of the XBT/MBT data were collected by ships of opportunity from areas of the ocean that are far from any contemporaneous CTD/OSD data, so the validation dataset Ref_{VAL} has a different spatiotemporal 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

$$b_{CAST}(cast, z_{standard}) = BT(cast, z_{standard}) - Ref_{TR}(cast, z_{standard}). \quad (1)$$

Because the cast level data are spatially biased toward 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. 2, step 5). Our chosen grid, the *World Ocean Atlas 2013* (*WOA13*) grid, has 1° × 1° resolution and 67 depth layers for 0–2000 m (Locarnini et al. 2013). 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 and 2.5 m fall in the 0-m bin; between 2.5 and 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 subannual time scales the time-varying biases appear to be dominated by changes in water temperature due to the seasonal cycle (GR10) and not by other factors such as changes to probe design that occurred over multiple years (CH14). 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}:

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

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. 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-yr increments) by 67 depths (with depth increments coinciding with the *WOA13*) by 79 temperatures (from −3° to 36°C at 0.5°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. 2005; Reverdin et al. 2009). CH14 used the 0–100-m average temperature taken from the cast itself, but some of these near-surface data contain 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°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. 2) thus yields the value of the bathythermograph instrumental bias on the correction grid, which we refer to as b_{COR},

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

We produced separate correction grids for the XBT and MBT datasets. For the XBT data, we consider the years 1967–2018 in our correction, and for MBT we consider the years 1940–2004. Although certain data exist 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. Creating an ensemble of artificial neural networks

The correction grids that result from the steps described above (Fig. 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. 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. 2, step 8).

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

$$\mathrm{cost} = \sum_{z} \sum_{T_{10\mathrm{m}}} \sum_{\mathrm{yr}} [b_{\mathrm{COR}}(z_{\mathrm{WOA}}, \mathrm{T}_{10\mathrm{m}}, \mathrm{yr}) - C(z_{\mathrm{WOA}}, T_{10\mathrm{m}}, \mathrm{yr})]^2. \tag{4}$$

As such, the ANN is trained to produce a final correction C (Fig. 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. 2005; Reverdin et al. 2009; Cowley et al. 2013; CH14), 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. 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. 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) below 2%, 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 overfitting the model, which occurs when the neural network is overtrained 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 2a, 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. 2, steps 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 are 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. 2, step 8):

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

Steps 6–9 (Fig. 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. As a result of 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 10 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 2a) 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 and 30 different corrected datasets of individual XBT and MBT casts, respectively, 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 with the independent validation set $\mathrm{Ref}_{\mathrm{VAL}}$ (Fig. 2, step 10) as a final test of our method's ability to correct "never before seen" BT biases.

3. Results

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 regard to depth, probe type, year, and latitude, variables on which the bias has been demonstrated to depend (Gouretski and Koltermann 2007; L09; GR10; CH14). We have used four of these metrics to assess our own method's performance when compared to the two top performing calibrations, L09 and CH14, 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-yr 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 2a) 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 ($\mathrm{Ref}_{\mathrm{VAL}}$) after the residual biases of the individual casts are binned to the *WOA13* grid (section 3a). 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 3b). We also consider differences in how these methods extrapolate to locations where there are no collocated reference data (section 3c). 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 (section 3d). Our main analysis and figures use the global (EANN-G) and probe-specific (EANN-P) corrections to XBT data that use 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 1 and 2, which will be described in more detail below).

a. *Assessing global XBT bias correction*

The World Ocean Database 2018 provides XBT data that have been precalibrated using some of the most popular XBT corrections including the CH14 and L09 corrections we consider here. We opted to use these precalibrated XBT data in our comparison, as we could download, quality control, and process them exactly in the same way as the uncorrected data (section 2a). 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 ($\mathrm{Ref}_{\mathrm{VAL}}$) from the interpolated XBT casts.

Next, these biases were binned to the *WOA13* grid using the same procedure discussed in section 2a. 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 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 5-yr 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 5-yr moving filter:

$$\text{Metric 1} = \sigma[\overline{b_{\mathrm{WOA}}}(\mathrm{yr})]. \tag{6}$$

Figure 3a shows the global median extrapolated temporal bias (uncorrected XBT–corrected) of the various calibration schemes for 0–700 m. In our analysis, negative or positive bias respectively implies that the uncorrected XBT data are too cold or too 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 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. 3a). All methods agree that a positive bias reappears in the 2000s, although the magnitude and temporal pattern of the bias differs between the various corrections by as much as 0.05°C or more. Both EANN-G and EANN-P exhibit a smooth leveling off in their extrapolated bias, whereas CH14 and L09 exhibit an oscillatory pattern (Fig. 3a). The choice of FRE also impacts the extrapolated bias of individual XBT probes (Fig. S1a in the online supplemental material). 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 (Fig. S1a).

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. 3b. However, CH14 and the EANN methods appear to underestimate the positive bias in the XBT data for the period after 2010, when data become 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. 3b). The residual biases for all methods, including the uncorrected XBT bias, converge at the 1987 minimum and diverge again afterward. 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

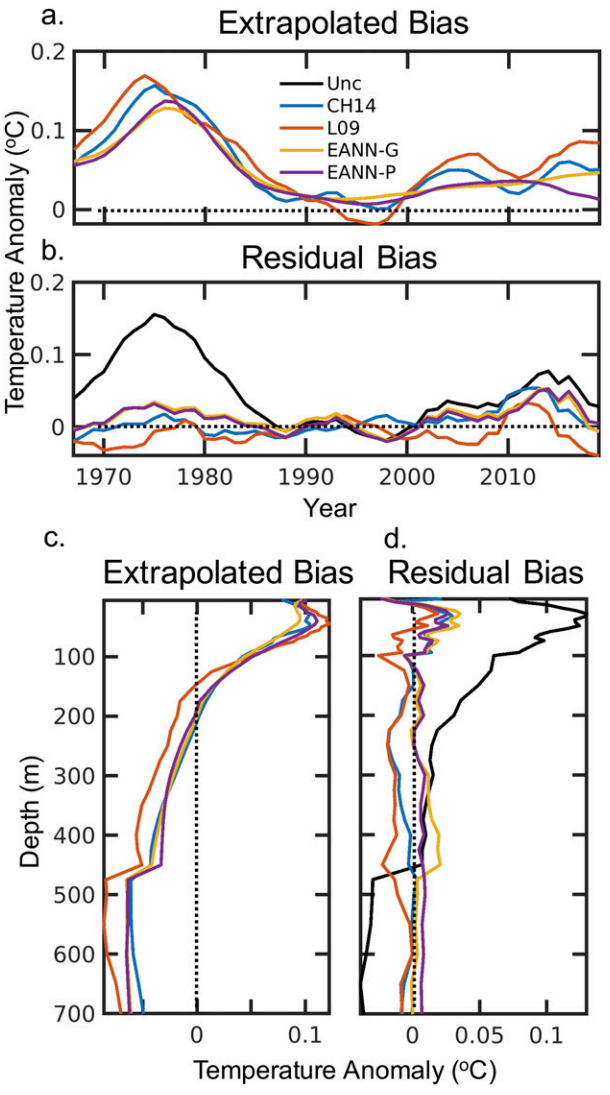

FIG. 3. Global median extrapolated XBT bias (uncorrected XBT − corrected) for 0–700 m by (a) year and (c) depth for various XBT calibration methods (CH14; L09; and our EANN-G and EANN-P) after binning to the *WOA13* grid. Also shown are residual biases with respect to our validation dataset (corrected XBT–$\mathrm{Ref_{VAL}}$) 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 (b) year and (d) depth.

TABLE 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 are not seen by the EANN calibrations). The choice of FRE that each method uses, either the original manufacturer (MFR) or H95, is also listed. Boldface values indicate the best performance for that metric.

	Calibration							
Metric	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

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 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 (Fig. S1b).

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:

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

Figure 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. 3c). All methods exhibit a transition in their extrapolated biases around 450 m, because 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. 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 1). Both the extrapolated and residual depth-dependent biases reveal the impact of the choice of FRE, although the effect is most apparent for lower latitudes (Figs. S1c,d).

Although there are significantly fewer XBT data below 700 m (roughly 3% 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. 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 onward, the CH14 extrapolated bias oscillates with both a similar period and phase to the extrapolated temporal bias in the top 700 m (Fig. 3a), although 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. 4b) for the uncorrected XBT data prior to 1990 likely is not 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. 3b). After 1990 the temporal bias is 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.

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. 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. 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 relative to the uncorrected XBT data above 1000 m (Fig. 4d). Based on the metrics (Table S1 in the online supplemental material), 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.

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 CH14, whose biases with respect to Ref_{VAL} have been separately binned to the *WOA13* grid:

$$\text{Metric } 3 = \sum_{i=1}^{n} |\overline{b_{\text{WOA}}}(z_i, \text{probe})| / n. \tag{8}$$

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. 5 and Table 2). While probe-specific calibrations are better overall (Table 2), the global XBT bias is dominated by only a few probe types that the

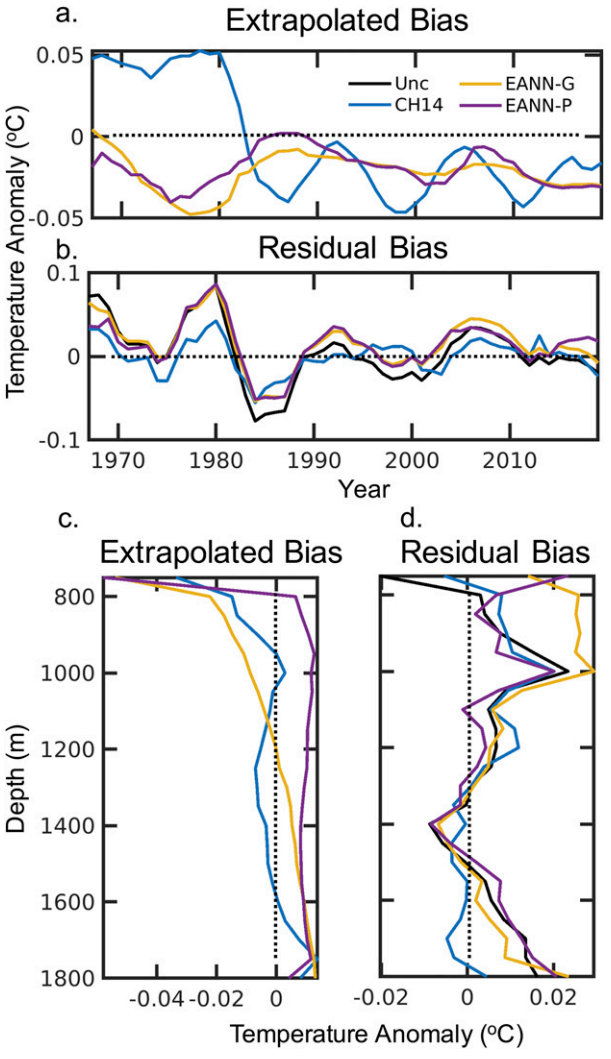

FIG. 4. As in Fig. 3, but for the 700–1800-m depth interval.

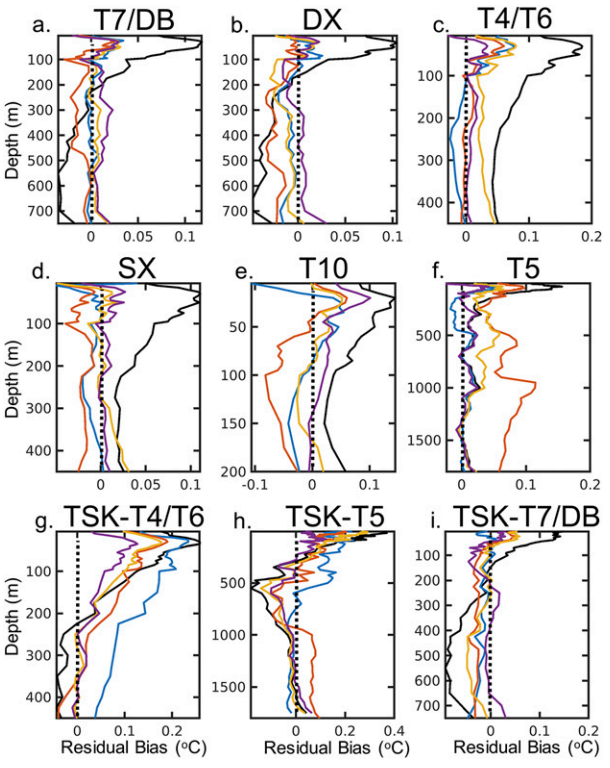

FIG. 5. Residual XBT biases with respect to our validation dataset (corrected XBT − $\mathrm{Ref_{VAL}}$) for various XBT calibration methods (CH14; L09; and our EANNG and EANN-P) after separating the data into nine probe types, binning to the *WOA13* grid, and taking the median by depth.

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 (Figs. 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.

Because of insufficient metadata, nearly one-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 of 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 of less than 930 m, including T7/DB probes and some T5 probes.

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 (Figs. 5c,g), 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. 5d) falls somewhere in between. Global calibrations are also not suitable for T5 (Fig. 5f) or TSK-T5 (Fig. 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. 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 et al. 2009; Cowley et al. 2013; CH14) as well as the vertical temperature gradient (GR10). 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

TABLE 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 FRE that each method uses, either the MFR or H95, is also listed. Boldface values indicate best performance for that probe type.

		Calibration						
Probe group	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

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 (GR10).

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

$$\text{Metric 4} = \sum_{i=1}^{n}\sum_{j=1}^{m} |\overline{b_{\text{WOA}}}(z_i, \text{lat}_j)|/(n \cdot m). \tag{9}$$

The uncorrected zonal XBT bias exhibits a fairly symmetric pattern across the equator (Fig. 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 (GR10, their Fig. 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. 6b). At both high and low latitudes, the L09 method overcorrects, turning positive biases negative and vice versa (Fig. 6c). Both the EANN-G (Fig. 6d) and EANN-P (Fig. 6e) calibrations reduce much of the original spatial bias in the XBT data. However, below 100 m the EANN methods undercorrect 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 (Figs. S2a–c in the online supplemental material). Additionally, these low-latitude negative biases can be eliminated by first applying the H95 FRE and subsequently the EANN correction (Fig. S2d), which based on metric 4 does best at reducing the spatial bias on our combined reference dataset of both CTD and OSD data (Table 1). This indicates that the H95 FRE is effective at reducing spatial biases in the XBT data, even though without additional corrections it worsens the overall XBT quality (Table 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 regard to both year (using a 5-yr moving filter for the temporal binning) and depth, we consider the standard deviation of the

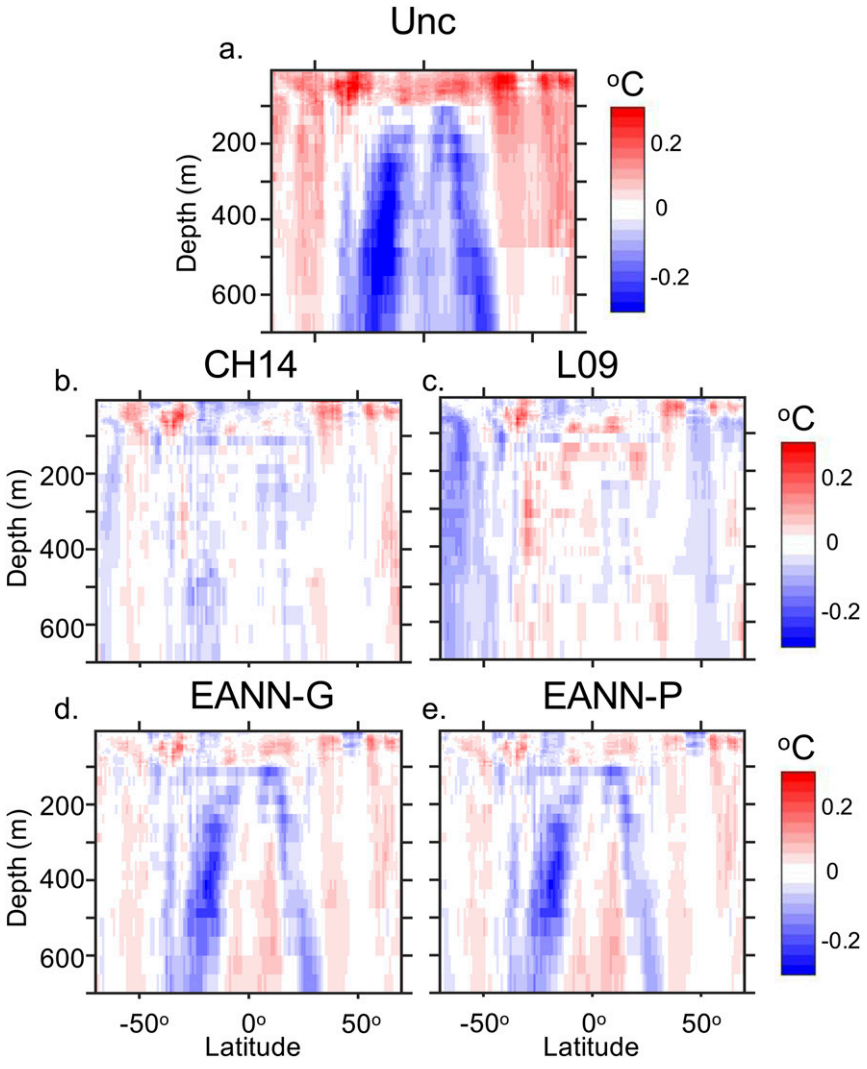

FIG. 6. Residual XBT biases with respect to our validation dataset (corrected XBT − Ref_{VAL}) as a function of both depth and latitude for (a) uncorrected XBT data (Unc), (b) data corrected with the CH14 method, (c) data corrected with the L09 method, (d) data corrected with the EANN-G method, and (e)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

temporal component at each depth bin and then average these values over all depth bins to obtain metric 5,

$$\text{Metric } 5 = \sum_{i=1}^{n} \sigma[\overline{b_{\text{WOA}}}(\text{yr}, z_i)]/n. \tag{10}$$

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, whereas below 100 m the bias shifts from positive in the 1970s to negative in the 1980s to more positive/neutral again after 2000 (Fig. 7a). The L09 method (Fig. 7c) slightly overcorrects the original XBT bias in much of the top 600 m during the 1970–90 period. CH14 (Fig. 7b) and the EANN methods (Figs. 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. 7d) undercorrects 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 very well according to metric 5 (Table 1).

b. *XBT calibration performance in individual basins*

The XBT calibrations differ in their performance on basin scales. To ensure that the performance of the XBT calibrations on metrics 1, 2, 4, and 5 discussed in section 3a 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 Fig. S3 in the online supplemental material. All metrics are calculated the same as in section 3a aside from metric 4, which differs in the number of latitudinal bins for the Pacific (m = 135) and Indian (m = 95) Oceans.

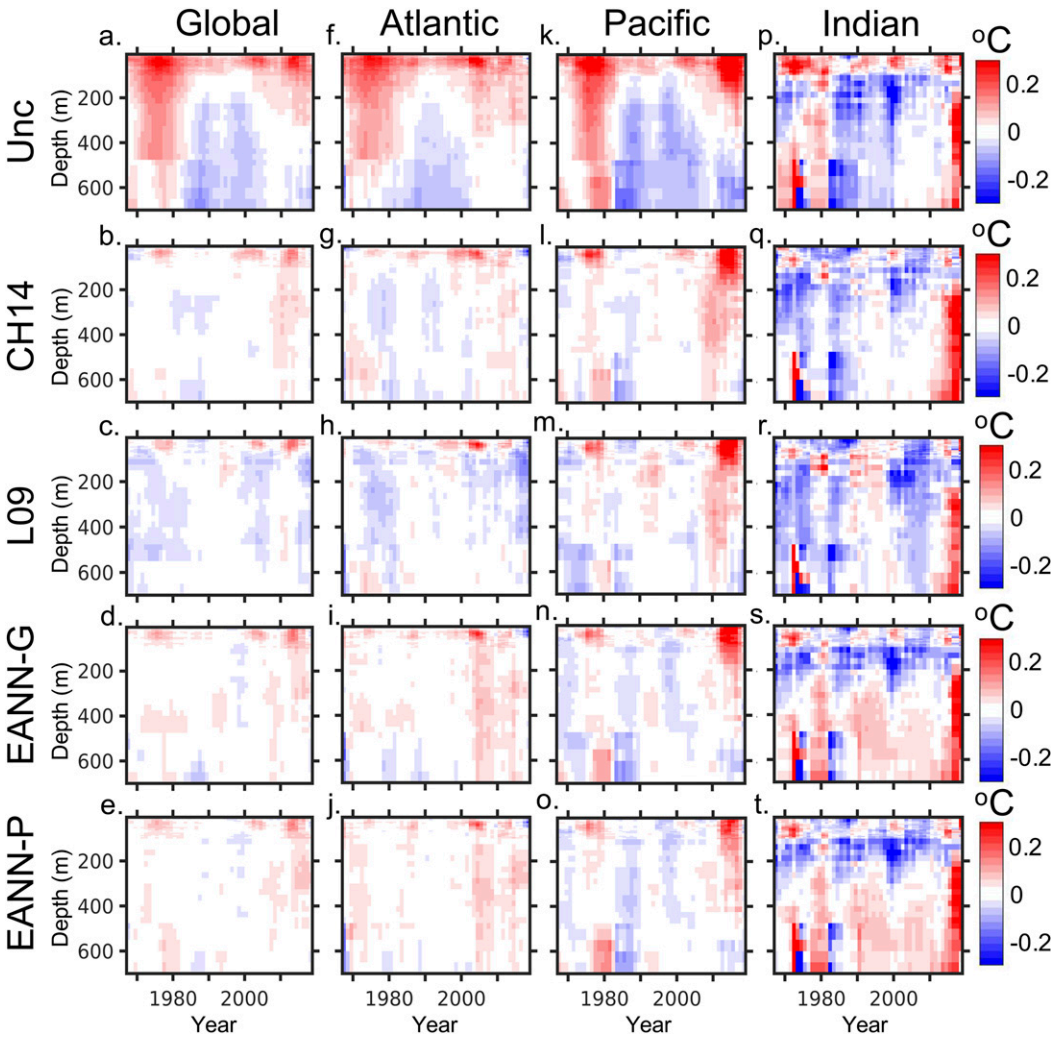

FIG. 7. Residual XBT biases with respect to our validation dataset (corrected XBT − Ref_{VAL}) as a function of both depth and year for uncorrected XBT data (Unc), data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the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

The Atlantic basin exhibits a similar pattern in the uncorrected XBT bias with regard to depth and year as was seen globally, with alternating positive and negative biases over time below 100 m (Fig. 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 (Figs. 7g,h), with the exception of L09 in the top 200 m after year 2000 where it overcorrects. The EANN methods (Figs. 7i,j) reduce the bias more evenly across time and depth but undercorrect significantly after 2010. CH14 also undercorrects 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 1).

The shape of the uncorrected XBT bias in the Pacific (Fig. 7k) dominates the global pattern (Fig. 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 (Figs. 7l–o) undercorrect the positive biases in the 1970s and 2010s, but the EANN-G (Fig. 7n) and EANN-P (Fig. 7o) calibrations do a slightly better job at correcting the biases after 2010 below 200m. Again, based on the metrics, all methods perform similarly (Table 1).

In the Indian Ocean (Figs. 7p–t), the uncorrected XBT biases are larger and noisier than in the other basins because there are less data in this basin. The existing data do 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 make it

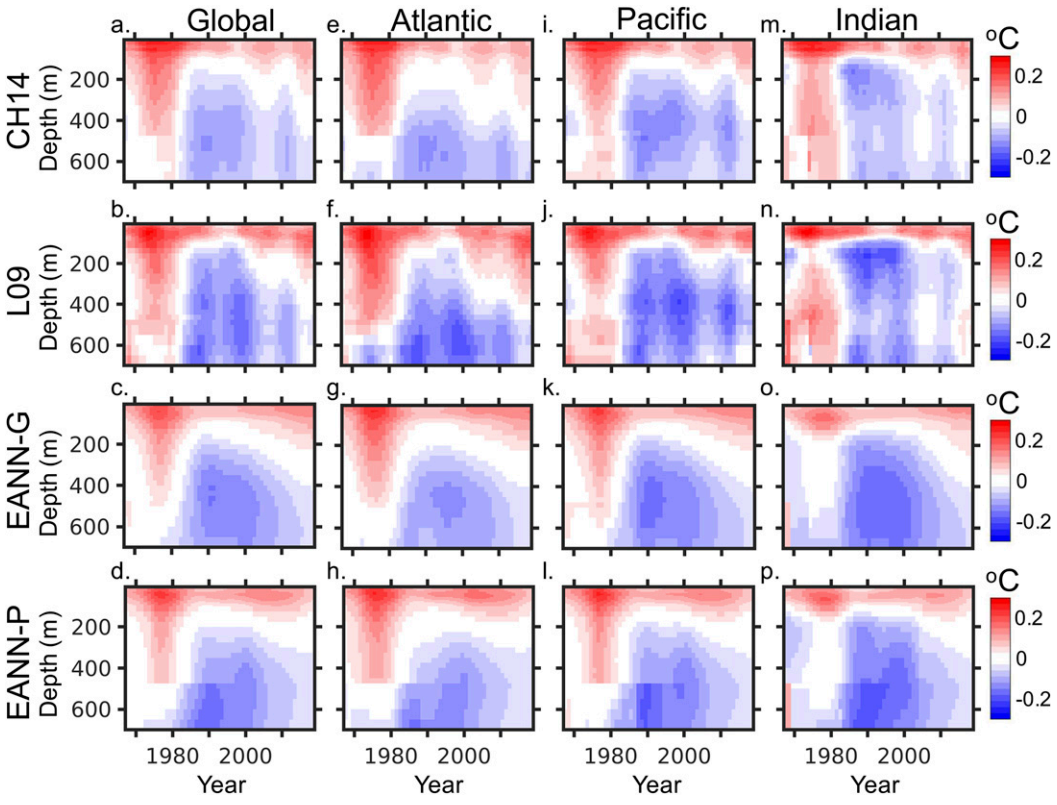

FIG. 8. Median of the extrapolated XBT bias (uncorrected XBT − corrected) as a function of depth and year for data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the EANN-P method, globally and for individual basins (Atlantic,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

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. 7p). In the top 450 m, both CH14 and L09 (Figs. 7q,r) appear to overcorrect the bias prior to 1980 and undercorrect from the 1990s onward. The EANN (Figs. 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. *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 intradecadal time scales 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 (Figs. 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. 8c) and EANN-P (Fig. 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 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 (Figs. 8e,f) extrapolated biases in the Atlantic share a similar form that is mainly distinguished by a

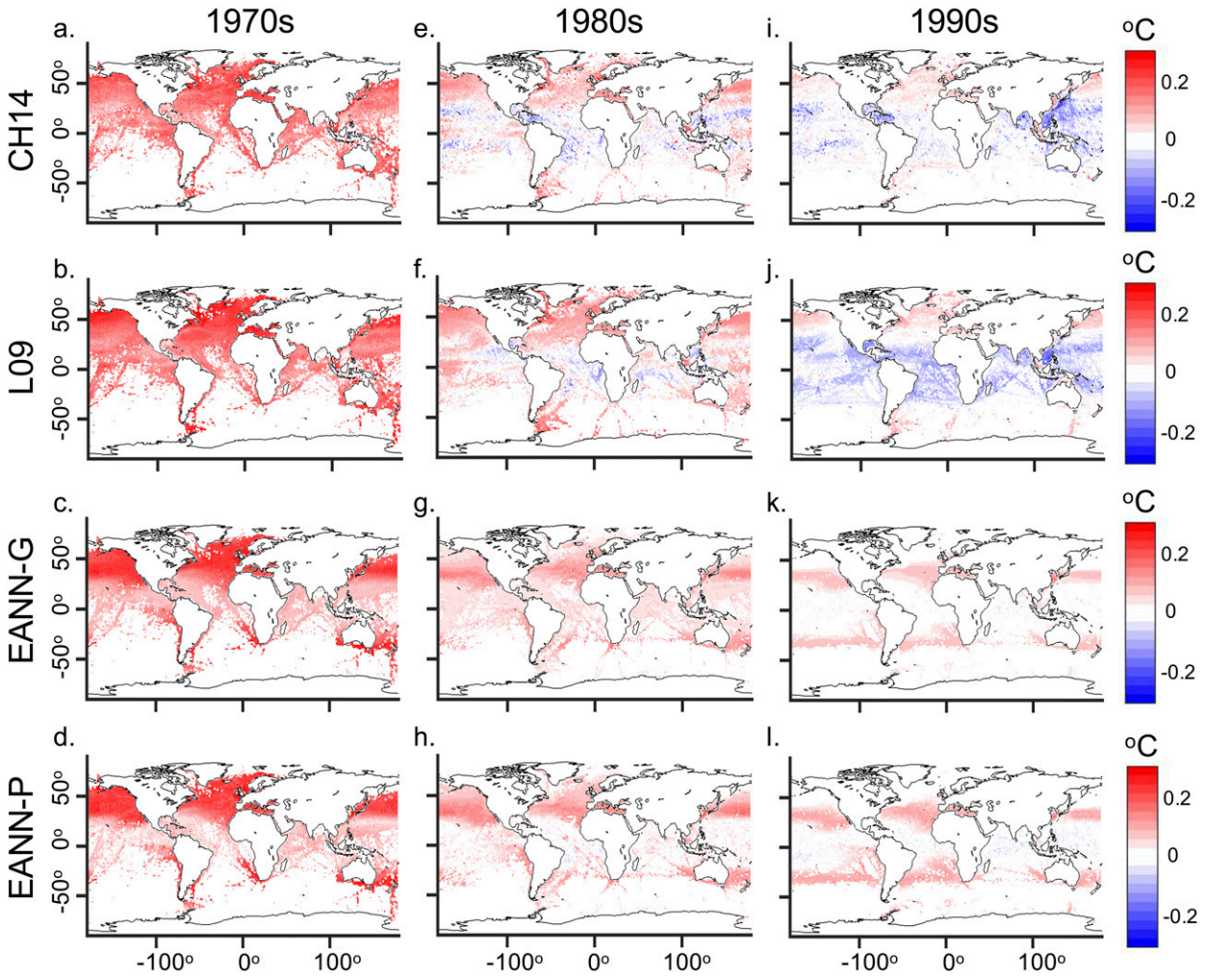

FIG. 9. Decadal averages of the extrapolated XBT bias (uncorrected XBT − corrected) averaged over the top 700 m for data corrected with the CH14 method, data corrected with the L09 method, data corrected with the EANN-G method, and data corrected with the EANN-P method during the 1970s, 1980s, and 1990s.

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

Because of the amount of XBT data in the Pacific, this basin has an outsized contribution 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 (Figs. 8i,j) consider XBT data around 400 m to have a negative bias. The EANN methods (Figs. 8k,l) differ from L09 and CH14 in their characterization of the bias below 450 m in the 1970s, 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. 8m) and L09 (Fig. 8n) methods are highly distinct from those of the EANN methods (Figs. 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. 8n). The CH14 method shows a shallower peak in this negative bias (Fig. 8m), while the EANN-G method spreads the negative bias more evenly across the water column (Fig. 8o), and the EANN-P method concentrates this bias below 450 m (Fig. 8p).

Maps of the extrapolated bias with the median taken over the top 700 m for different decades reveal distinct spatial patterns in the biases over time (Fig. 9). In the 1970s all methods overwhelmingly treat the uncorrected XBT data as warmer

than the corrected, but while the CH14 method (Fig. 9a) and L09 (Fig. 9b) are more homogenous the EANN methods (Figs. 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 (Figs. 9e–h). Both the CH14 (Fig. 9e) and L09 (Fig. 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. 9g) extrapolates positive biases at all latitudes, whereas the EANN-P method (Fig. 9h) indicates a neutral to slightly negative bias in the tropics. For both EANN methods the positive bias peaks near 30°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. 9i) and L09 (Fig. 9j) 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 (Figs. 9k,l) consider the XBT bias to be neutral at most latitudes, aside from bands of positive biases around 30°N and S that mirror those from the 1980s.

d. 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. 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. 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. 10a). The uncorrected time-dependent bias with respect to our validation dataset is also smooth (Fig. 10b), peaking in 1954 around 0.1°C 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 are very little MBT data during this period. GR10 slightly undercorrects the original positive bias in the MBT data with respect to the validation dataset, whereas the EANN-G slightly overcorrects this positive bias (Fig. 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. 10b).

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

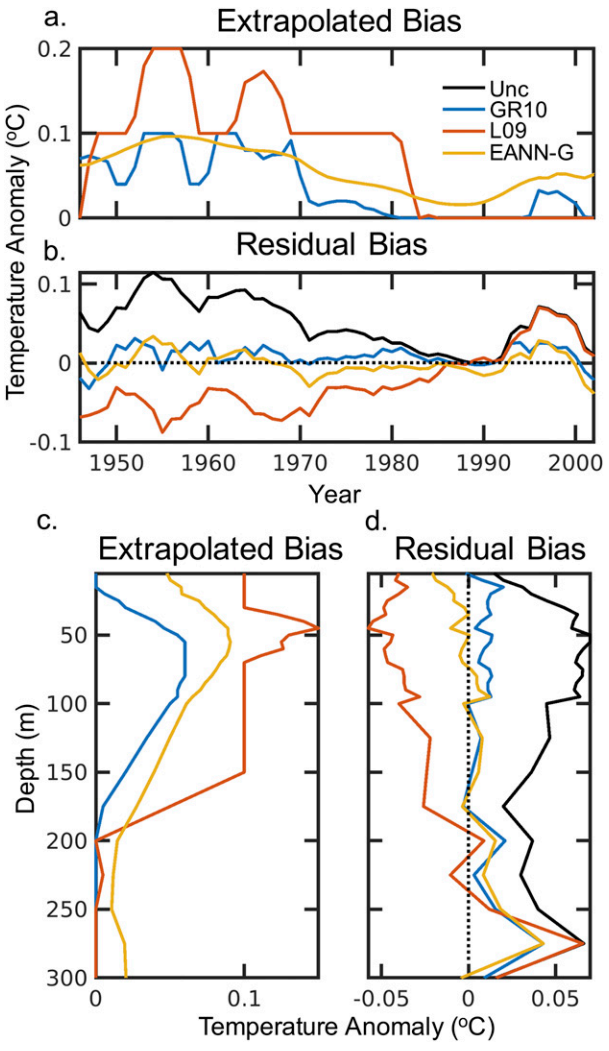

FIG. 10. Global median extrapolated MBT bias (uncorrected MBT − corrected) for 0–300 m by (a) year and (c) depth for various MBT calibration methods (GR10; L09; and our EANN-G) after binning to the *WOA13* grid. Also shown are residual biases with respect to our validation dataset (corrected MBT–$\mathrm{Ref_{VAL}}$) 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 (b) year and (d) depth.

extrapolated bias that also peaks at a shallower depth, leading to negative residual biases in the top 50 m (Figs. 10c,d). L09 creates a larger extrapolated bias than the other methods (Fig. 10c), leading to a residual negative bias with respect to the validation data that is similar in magnitude to the original positive bias (Fig. 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. 10d).

In 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 mostly

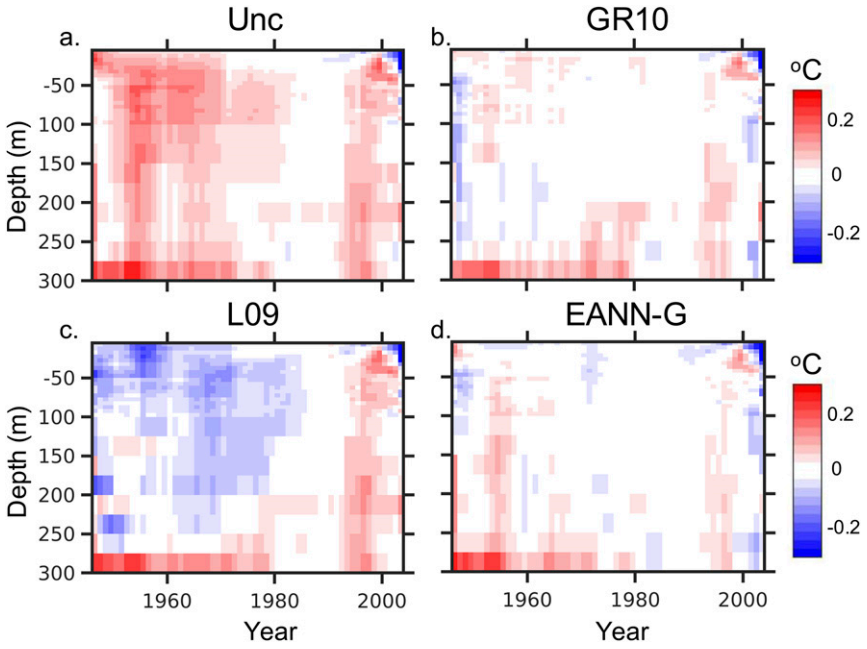

FIG. 11. Global residual MBT biases with respect to our validation dataset (corrected MBT − Ref_{VAL}) as a function of both depth and year for (a) uncorrected MBT data (Unc), (b) data corrected with the GR10 method, (c) data corrected with the L09 method, and (d) data corrected with the EANN-G method.

positive throughout its period of use (Fig. 11a). The GR10 calibration (Fig. 11b) largely reduces these positive temperature biases but slightly overcorrects biases in the 1940s and slightly undercorrects biases in the 1990s. By comparison, the L09 method (Fig. 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. 11d) removes the majority of the systematic bias but undercorrects 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. 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. 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. 11a). A second strong positive bias in the L09 extrapolation occurs in the mid-1960s (Fig. 12b). The EANN-G method (Fig. 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. 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 that 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. On the basis of the comparison presented here, the EANN-G method may present a useful alternative statistical approach to correct the MBT data.

4. Discussion and conclusions

In this study, we developed and implemented a new approach to correct global systematic temperature biases in mechanical and expendable bathythermograph 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, with that of several popular methods (L09 and CH14 for XBT and L09 and GR10 for MBT) using some of the metrics proposed by Cheng et al. (2018). Last, we examined differences in how these calibration methods extrapolate the bias both spatially and temporally.

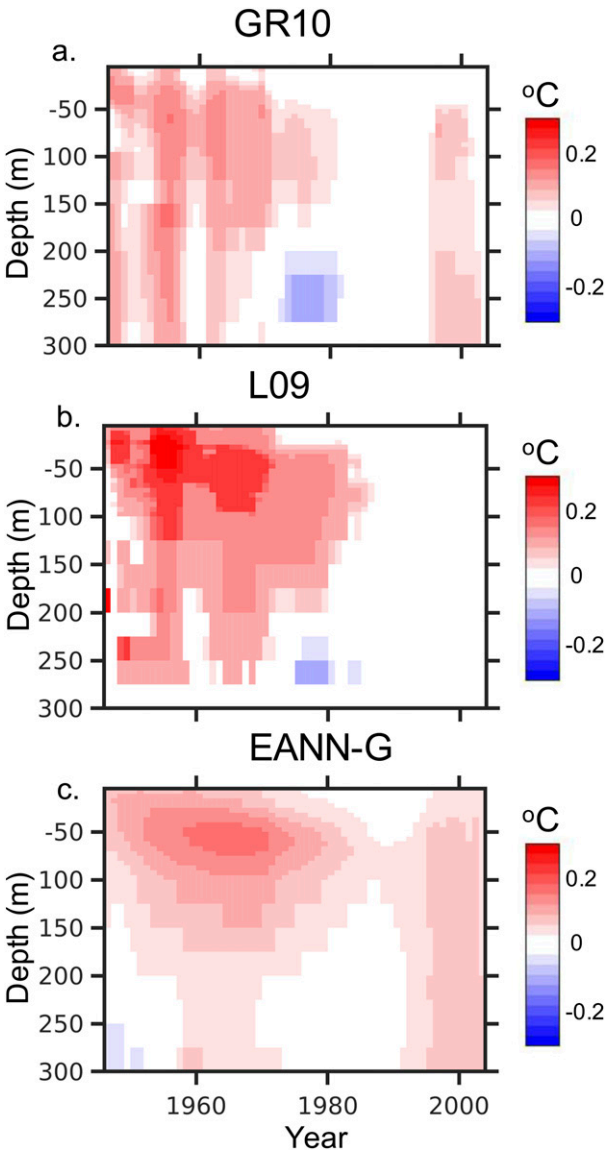

FIG. 12. Median of the global extrapolated MBT bias (uncorrected MBT − corrected) as a function of depth and year for (a) data corrected with the GR10 method, (b) data corrected with the L09 method, and (c) data corrected with the EANN-G method.

Our results demonstrate the following:

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 with 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 toward only a few probe types and the metadata are 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 complementary 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 intradecadal and basin scales.
4) The choice of XBT 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.

Acknowledgments. Timothy DeVries acknowledges support from the National Science Foundation (OCE-1948985). We also thank NOAA NCEI as well as those who constructed and maintain the World Ocean Database. Our appreciation also goes to two anonymous reviewers whose comments and suggestions greatly strengthened this paper. The in situ temperature datasets used in this study, including the reference CTD and OSD casts as well as XBT and MBT casts already corrected by several widely used methods, can be freely obtained from the NCEI website. Data files containing individual corrected XBT and MBT

casts using the EANN methods presented here are freely available online (https://doi.org/10.6084/m9.figshare.12170403).

REFERENCES

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, https://doi.org/10.1002/rog.20022.

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, https://doi.org/10.1175/JCLI-D-15-0801.1.

——, and Coauthors, 2018: *World Ocean Database 2018*. NOAA Atlas NESDIS 87, 207 pp., https://data.nodc.noaa.gov/woa/WOD/DOC/wod_intro.pdf.

Breiman, L., 1996: Bagging predictors. *Mach. Learn.*, **24**, 123–140, https://doi.org/10.1023/A:1018054314350.

Cheng, L., and J. Zhu, 2014: Uncertainties of the ocean heat content estimation induced by insufficient vertical resolution of historical ocean subsurface observations. *J. Atmos. Oceanic Technol.*, **31**, 1383–1396, https://doi.org/10.1175/JTECH-D-13-00220.1.

——, ——, F. Reseghetti, and Q. Liu, 2011: A new method to estimate the systematical biases of expendable bathythermograph. *J. Atmos. Oceanic Technol.*, **28**, 244–265, https://doi.org/10.1175/2010JTECHO759.1.

——, ——, 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, https://doi.org/10.1175/JTECH-D-13-00197.1.

——, and Coauthors, 2016: XBT science: Assessment of instrumental biases and errors. *Bull. Amer. Meteor. Soc.*, **97**, 924–933, https://doi.org/10.1175/BAMS-D-15-00031.1.

——, 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**, e1601545, https://doi.org/10.1126/sciadv.1601545.

——, 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, https://doi.org/10.1175/JTECH-D-17-0122.1.

Church, J. N., and Coauthors, 2011: Revisiting the Earth's sea-level and energy budgets from 1961 to 2008. *Geophys. Res. Lett.*, **38**, L18601, https://doi.org/10.1029/2011GL048794.

Couper, B. K., and E. C. LaFond, 1970: The mechanical bathythermograph: An historical review. *Adv. Instrum*, **25**, 735–770.

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, https://doi.org/10.1175/JTECH-D-12-00127.1.

Domingues, C., J. Church, N. White, P. Gleckler, S. Wijffels, P. Barker. And, and J. Dunn, 2008: Improved estimates of upper-ocean warming and multi-decadal sea-level rise. *Nature*, **453**, 1090–1093, https://doi.org/10.1038/nature07080.

Foresee, F. D., and M. T. Hagan, 1997: Gauss-Newton approximation to Bayesian learning. *Proc. Int. Conf. on Neural Networks*, Houston, TX, IEEE, 1930–1935, https://doi.org/10.1109/ICNN.1997.614194.

Glorot, X., A. Bordes, and Y. Bengio, 2011: Deep sparse rectifier neural networks. *14th Int. Conf. on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, Proceedings of Machine Learning Research, 315–323, https://www.jmlr.org/proceedings/papers/v15/glorot11a/glorot11a.pdf.

Good, S., 2011: Depth biases in XBT data diagnosed using bathymetry data. *J. Atmos. Oceanic Technol.*, **28**, 287–300, https://doi.org/10.1175/2010JTECHO773.1.

Gouretski, V., 2012: Using GEBCO digital bathymetry to infer depth biases in the XBT data. *Deep-Sea Res. I*, **62**, 40–52, https://doi.org/10.1016/j.dsr.2011.12.012.

——, and K. Koltermann, 2007: How much is the ocean really warming? *Geophys. Res. Lett.*, **34**, L01610, https://doi.org/10.1029/2006GL027834.

——, 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. I*, **57**, 812–833, https://doi.org/10.1016/j.dsr.2010.03.011.

——, 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, https://doi.org/10.1175/JTECH-D-19-0205.1.

Hagan, M. T., and M. B. Menhaj, 1994: Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Network*, **5**, 989–993, https://doi.org/10.1109/72.329697.

Hamon, M., G. Reverdin, and P. Le Traon, 2012: Empirical correction of XBT data. *J. Atmos. Oceanic Technol.*, **29**, 960–973, https://doi.org/10.1175/JTECH-D-11-00129.1.

Hanawa, K., P. Rual, R. Bailey, A. Sy, and M. Szabados, 1995: A new depth-time equation for Sippican or TSK T-7, T-6 and T-4 expendable bathythermographs (XBT). *Deep-Sea Res. I*, **42**, 1423–1451, https://doi.org/10.1016/0967-0637(95)97154-Z.

Hansen, J., 2005: Earth's energy imbalance: Confirmation and implications. *Science*, **308**, 1431–1435, https://doi.org/10.1126/science.1110252.

Hansen, L., and P. Salamon, 1990: Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, **12**, 993–1001, https://doi.org/10.1109/34.58871.

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, https://doi.org/10.1007/s10872-009-0027-7.

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, https://doi.org/10.1007/s10872-005-0024-4.

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**, L07608, https://doi.org/10.1029/2008GL037155.

——, and Coauthors, 2012: World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophys. Res. Lett.*, L10603, **39**, https://doi.org/10.1029/2012GL051106.

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

Locarnini, R. A., and Coauthors, 2013: *Temperature*. Vol. 1, *World Ocean Atlas 2013*, NOAA Atlas NESDIS 73, 40 pp., https://data.nodc.noaa.gov/woa/WOA13/DOC/woa13_vol1.pdf.

Lyman, J. M., S. A. Good, V. Gouretski, M. Ishii, G. C. Johnson, M. D. Palmer, D. M. Smith, and J. Willis, 2010: Robust warming of the global upper ocean. *Nature*, **465**, 334–337, https://doi.org/10.1038/nature09043.

MacKay, D. J., 1992: Bayesian interpolation. *Neural Comput.*, **4**, 415–447, https://doi.org/10.1162/neco.1992.4.3.415.

Marquardt, D., 1963: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, **11**, 431–441, https://doi.org/10.1137/0111030.

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, https://doi.org/10.1126/science.1106663.

Prechelt, L., 1998: Automatic early stopping using cross validation: Quantifying the criteria. *Neural Network*, **11**, 761–767, https://doi.org/10.1016/S0893-6080(98)00010-0.

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, https://doi.org/10.1175/2009JTECHO655.1.

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, https://doi.org/10.1175/1520-0426-19.3.391.

Trenberth, K., J. Fasullo, and M. A. Balmaseda, 2014: Earth's energy imbalance. *J. Climate*, **27**, 3129–3144, https://doi.org/10.1175/JCLI-D-13-00294.1.

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, https://doi.org/10.1007/s00382-017-3751-5.

Weigend, A. S., B. A. Huberman, and D. E. Rumelhart, 1990: Predicting the future: A connectionist approach. *Int. J. Neural Syst.*, **01**, 193–209, https://doi.org/10.1142/S0129065790000102.

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, https://doi.org/10.1175/2008JCLI2290.1.