Correction for Systematic Errors in the Global Dataset of Temperature Profiles from Mechanical Bathythermographs

VIKTOR GOURETSKI

*International Center for Climate and Environmental Science, Institute of Atmospheric Physics,
Chinese Academy of Sciences, Beijing, China*

LIJING CHENG

*International Center for Climate and Environmental Science, Institute of Atmospheric Physics, and
Center for Ocean Mega-Science, Chinese Academy of Sciences, Beijing, China*

(Manuscript received 15 December 2019, in final form 14 March 2020)

ABSTRACT

A homogeneous, consistent, high-quality in situ temperature dataset covering some decades in time is crucial for the detection of climate changes in the ocean. For the period from 1940 to the present, this study investigates the data quality of temperature profiles from mechanical bathythermographs (MBT) by comparing these data with reference data obtained from Nansen bottle casts and conductivity–temperature–depth (CTD) profilers. This comparison reveals significant systematic errors in MBT measurements. The MBT bias is as large as 0.2°C before 1980 on the global average and reduces to less than 0.1°C after 1980. A new empirical correction scheme for MBT data is derived, where the MBT correction is country, depth, and time dependent. Comparison of the new MBT correction scheme with three schemes proposed earlier in the literature suggests a better performance of the new schemes. The reduction of the biases increases the homogeneity of the global ocean database being mostly important for climate change–related studies, such as the improved estimation of the ocean heat content changes.

1. Introduction

The ocean plays a crucial role in the climate system as a key component of global energy, water, and carbon cycles (IPCC 2019). Due to its large heat capacity, the ocean stores more than 90% of the excess heat trapped by greenhouse gases (Rhein et al. 2013, 261–265). This highlights the importance of the global hydrographic temperature profile dataset as invaluable source of information in describing climate change, because the ocean heat content time series is a key measure of Earth's energy imbalance (Meyssignac et al. 2019; Cheng et al. 2017; Levitus et al. 2012). Besides, an unbiased ocean temperature profile data is crucial in weather forecast

models, data assimilation systems, gridded climatologies, and numerous other applications.

A rapid accumulation of temperature profile data permitted the calculation of the first ocean heat content (OHC) time series (Levitus et al. 2000, 2005), which showed a progressive warming of the global ocean since the 1950s. These first time series also showed a strong variability on decadal scales, with a local heat content maximum around 1980, which was not reproduced by climate models (Levitus et al. 2000, 2005; IPCC 2007).

Gouretski and Koltermann (2007) estimated heat content time series separately for different instrumentation types and found that both the mechanical bathythermograph (MBT) and the expendable bathythermograph (XBT) data were warm biased with the biases varying with time. It was this warm bias that introduced artificial decadal-scale variability into the OHC time series, especially the OHC maxima around 1980.

This work triggered a series of studies where different kinds of XBT bias correction schemes were developed (i.e., Wijffels et al. 2008; Gouretski and Reseghetti 2010;

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JTECH-D-19-0205.s1>.

Corresponding authors: Viktor Gouretski, lunegur@yahoo.de; Lijing Cheng, chenglij@mail.iap.ac.cn

Good 2011; Gouretski 2012; Cowley et al. 2013; Cheng et al. 2018). It was found that the total temperature bias is due to both depth bias and a thermal bias. Both bias components were found to be time dependent, with several factors affecting the biases. The XBT science community (Cheng et al. 2018) conducted a comparison of the correction schemes, with Cheng et al. (2014) being ranked as the most effective scheme.

The XBT bias issue was important because the XBTs contributed to 38% of the ocean profile data within the period 1970–2001 (Cheng et al. 2016). Possible biases in other instrumentation types received less attention in the literature. Along with hydrographic casts and XBTs, two other instrumentation types contributed significantly to the global temperature profile dataset: MBTs and satellite-related dataloggers attached to marine animals. MBTs provided the majority of the temperature profile data for the upper 200 m between 1941 and the beginning of the 1970s, until the XBTs started to replace them. In the World Ocean Database 2018 (WOD18) (Boyer et al. 2018), MBTs contribute to ~68% of data within 1940–66, dominating the upper-ocean temperature archive within that period.

There is strong consensus in the atmosphere and ocean climate community that improving the data quality is a vital task to enhance our scientific understanding of the climate, the ocean, and the ocean ecosystems (IPCC 2019). The WOD18 contains data from 10 different instrumentation types, which have different measurement accuracy. Whereas the automatic quality control procedures are able to filter out random data outliers rather successfully, the assessment of systematic errors in the data requires a different approach. Although there was an active and successful treatment of the biases in the XBT data, which ensured their reliability (Goni et al. 2019), the quality assessments of the MBT data are less numerous. Three MBT bias correction schemes were suggested by Ishii and Kimoto (2009), Levitus et al. (2009), and Gouretski and Reseghetti (2010).

In this paper, we examine the quality of the MBT temperature profiles by comparing these data with other more accurate observations, with data and methods being introduced in section 2. We identify significant systematic errors in MBT data, and new bias correction schemes are derived for MBT (section 3). The robustness of the new correction scheme is evaluated by a “throwaway” test, where a part of the data is used for the derivation of the depth and temperature corrections, and the other part is used to evaluate the performance of the scheme performance. Further, the performance of the new correction scheme is compared with the

existing correction schemes for MBTs in section 3. The conclusions are provided in section 4.

2. Data and methods

a. The MBT instrument

The MBT was developed by Spilhaus (1938) shortly before the Second World War and used for determining temperature down to a maximum depth of approximately 300 m. During the war years the MBTs were used to locate thermoclines mainly for hydro-acoustic purposes, but after the war they were extensively used in nonmilitary applications. Most MBTs were designed for the temperature range from -2° to 30°C with an accuracy of 0.1°C in temperature and $\sim 1\%$ in depth (UNESCO 1966).

The MBT consists of a thermal element and a depth element. The thermal element is attached to a stylus that scribes a trace across a glass slide, whereas a depth element moves the slide perpendicular to the motion of the stylus as the instrument is lowered into the water. The trace on a slide is read using a magnifying grid viewer, individually calibrated for each instrument. The figures and details of MBT construction can be found in Spilhaus (1938) and State Oceanographic Institute (2016). A description of the MBTs manufactured in the United States and the practices of work at sea is given in the book on historical oceanographic observations (Shor 1978).

The MBT data are subject to several kinds of error linked to the instrument and data acquisition and processing (Bralove and Williams 1961; Hazelworth 1964, 1966): 1) temperature element response delay, 2) pressure element response delay, 3) errors due to field operator, 4) errors during processing of the slide, 5) inaccurate calibration of the grid viewer, and 6) reading, coding, key-punching errors. An additional source of error could arise when the MBT temperatures were corrected by means of a reference temperature. Usually two kinds of independent references were used: vessel’s water intake temperature and the cruise bucket water temperature. Unfortunately, both references are known to have their own bias, so that these adjustments in many cases have introduced additional temperature biases (T. Winterfeld 1963, unpublished manuscript). The WOD18 (similar to other major databases) contains no information on corrections applied to the MBT profiles, and all types of errors named above contributed to the mean diagnosed MBT bias.

According to UNESCO (1966), at least nine manufacturers produced different kinds of MBTs over the

FIG. 1. (a) Percentage and (b) number of MBT profiles for the main contributing countries.

years: three manufacturers in the United States (Belfort, GM, and Kahl), one in the United Kingdom (Brown), two in France (Jules Richard and Mecabolier), one in the former Soviet Union (Mashpribor), and one in Japan (TSK). Most manufacturers used the U.S. patent.

All manufacturers (except for Mashpribor and Mecabolier) built three MBT modifications for three depth ranges: 0 to 50–75, 0 to 135–150, and 0 to 200–300 m. Unfortunately, there is no information on the MBT type modification in the WOD18 archive.

b. MBT data

WOD18 (Boyer et al. 2018) with the updates as of January 2019 provided the data basis for this study. The automated quality control procedure of Gouretski (2018) was applied to MBT temperature profiles at observed levels. The procedure includes several data quality checks, with the observation being rejected if at least one quality check fails. The percentages of rejected MBT observed levels are 3.85%, with 14.10% of profiles having at least one flagged level. Data outlier frequency histograms are presented in Fig. S1a in the online supplemental material. A total of 2419 739 MBT profiles were retained for the analysis with yearly numbers of profiles shown in Fig. 1.

The MBT instrumentation type of the WOD18 contains profiles from different MBT modifications. The majority of the profiles came from the MBT instruments of the A. Spilhaus design, but the WOD18 contains a much smaller amount of micro-MBT and digital MBT profiles classified under the same instrument type. Though these two types currently belong to

the MBT instrumentation type in the WOD18 they are different in their design and construction and are not considered in this study.

Inspection of the yearly profile numbers for different countries showed that MBTs were no longer in use on U.S. ships after 1974, being substituted by XBTs. However, there is a number of U.S. profiles linked to the MBT instrument type obtained after 2000. According to T. Boyer (2019, personal communication), these profiles belong to a different instrument type (attributed to MBTs in WOD18 database) and were not used in our analysis.

Whereas MBTs were used on ships of many countries, the United States, the former USSR (Soviet Union), Japan, Canada and United Kingdom contributed 95% of all WOD18 MBT profiles (Table 1). The United States contributed the largest number of MBT profiles for the WOD18. Performance of the MBTs manufactured in the United States was investigated in several studies (Dinkel and Stawnichy 1973; Casciano 1967; Hazelworth 1966; R. L. Stewart 1963, unpublished manuscript). Each of these studies considered only a limited set of the MBT profiles, so that it is not clear if the respective conclusions are valid for the global MBT dataset.

The MBTs were also extensively used on the USSR ships, where the MBTs of the type GM-9-III was manufactured, with the maximum depth of 200 m (State Oceanographic Institute 2016). Pavlov and Kuksa (1957) estimated the accuracy of the MBT by the comparison with side-by-side Nansen casts. They found significant differences between upward and downward MBT casts, with the upward (downward) casts underestimating (overestimating) temperature measured by the Nansen cast thermometers. They recommended to use the mean temperature of the downward and upward casts. The third-largest MBT data contributor is Japan, where several types of MBTs were developed. The Tsurumi Seiko company was the only manufacturer in Japan. The production was based on the U.S. patent. Gautam and Thadathil (2016) found systematic positive errors of about 0.6°C in the MBT data from the Indian Ocean between 1979 and 1984, with the temperature bias caused by systematic depth errors.

Unfortunately, there is no information for other countries regarding the MBTs and the working procedure. According to J. Gould (2019, personal communication) most of MBTs used in the United Kingdom were manufactured by the British Brown company, with the majority of the data coming from Navy vessels. The MBTs used on the German vessels were produced by the Wallace-Tiernan Company (J. Meincke 2019,

TABLE 1. Statistics for the MBT profiles from the WOD18 archive.

Country	All profiles	Fraction of total (%)	Collocated pairs	Fraction of total (%)	Side-by-side pairs	Fraction of total (%)
US	1 169 350	48.325	283 379	31.392	75 155	29.952
SU	448 908	18.552	227 994	25.257	77 997	31.085
JP	365 508	15.105	194 304	21.525	29 345	11.695
CA	195 946	8.098	104 053	11.527	41 927	16.71
GB	118 638	4.903	40 304	4.465	3787	1.509
Other	121 389	5.017	52 673	5.835	22 704	9.048

personal communication; D. Machochek 2019, personal communication).

The histogram of the sampled depth (Fig. 2) implies country-specific strategies to read data from the glass slide, as implemented on the ships of different nations. For instance, on the Japanese ships the readings from the MBT slides were done predominantly at a widely spaced set of levels. Also on the Russian ships the readings at standard depths dominate. On the ships from the United States, Canada, and United Kingdom the observations usually are reported at a larger number of levels, and the respective histograms imply rather similar observation practices. Percentage of profiles versus maximum observed depth (Fig. 2) also indicates different measurement strategies and different characteristics of the MBTs used by different countries. Significant drops of the observed level percentages occur at 200 and 250 m for the Russian and

Japanese data, and at 120 m (close to 400 feet) for the data from the United States, Canada, and United Kingdom. According to Shor (1978) the processing group there handled bathythermograph data from many other institutions in several countries, and this can partly explain the commonalities between the observed levels. Unfortunately, missing metadata do not permit a certain conclusion.

The UNESCO report (UNESCO 1966) provides useful information on MBT manufacturers and on the maximum observation depth for different MBT modifications (see their Table S1-009). Since this kind of information is not available in the WOD18, we used this table along with the information on MBT manufacturers to arrive at the country-specific MBT maximum observed depths, which is set to be 250, 200, 200, 270, and 274 m for the MBT profiles from the United States, the former Soviet Union, Japan, Germany, and

FIG. 2. Percentage of MBT profiles vs maximum observed depth (curves, left y axes) and percentage of observations for each 1 m level (vertical bars, right y axes).

the United Kingdom, respectively. Temperature profiles with the deepest level exceeding the abovementioned limits were not used for the analysis. Such profiles represent less than 0.1% of all available MBT profiles. Based on the manufacturer-specific differences in MBT design and country-specific observational strategies we assume that the MBT depth and temperature biases are *country specific*, which is further confirmed by our analyses below.

In summary, there is only some basic knowledge available for each archived MBT profile (i.e., time, location, platform, cruise, and country). The absent metadata does not permit the construction of a bias correction scheme based on the knowledge on instrument type, data acquisition and handling practice. This is the same challenge as experienced with the development of the XBT data correction schemes. Therefore, only an empirical bias correction scheme can be proposed.

c. Reference data

For the calculation of the systematic offsets in the MBT data, reference data are required. We assume Nansen bottle cast data and CTD data to be bias free because both the reversing thermometers attached to the bottles and the CTD sensors were normally calibrated in the laboratory. For instance, the initial accuracy of the SBE3 temperature sensors implemented on the CTDs manufactured by Sea-Bird Scientific is about 0.001°C with sensor drift being less than 0.001°C during a 6-month screening period (<https://www.seabird.com>). If CTD temperature sensors were factory calibrated close to cruise time and if the correct calibration coefficients were entered into the data acquisition software, then the CTD temperatures should be affected by only small random error. This is the case for the cruises occupied during such programs like World Ocean Circulation Experiment (WOCE), Climate and Ocean: Variability, Predictability and Change (CLIVAR) Programme, and Global Ocean Ship-Based Hydrographic Investigations Program (GO-SHIP). Unfortunately, no respective metadata are available for many historical CTD data. Figures S1b and S1c show data outlier frequency histograms for bottle and CTD data, respectively. Comparison of the outlier rates for MBT and bottle temperature profiles for the Second World War years reveals the Nansen cast profiles having severe quality problems with often more than 50% of all data failing to pass quality checks. In contrary, the MBT profiles show an increased outlier percentage only for the levels below 200 m, with just a small number of MBT profiles exceeding this depth during the war years. We conclude that the reference bottle data basis for the war years is less reliable

compared to the following years, and the MBT bias corrections are calculated only for the years after 1946.

Between the end of the 1960s and the beginning of the 1990s both CTD and Nansen bottle casts were used for temperature observations. Unlike the CTDs, which measure pressure with the typical error of 0.2% of the full range, the depth determination on the Nansen casts was less accurate, with the depths of observations being determined with paired protected and unprotected reversing thermometers (Wüst 1933; Warren 2008). However, the paired thermometers were usually applied only at sample depths greater than 100–300 m, and also not at each Nansen bottle. Before the introduction of the thermometric method (and also, later, at least for shallow casts) the depth of Nansen bottles was estimated from the wire paid out and the angle of the wire from the vertical at the ship deck. For instance, on Russian ships two depth corrections were applied: one accounting for the wire angle at the deck height and the other taking account for the parabolic shape of the wire below the sea surface (Soskin 1977; Kirejev 1939). Respectively, the Nansen bottle casts are characterized by a less accurate sample depth estimation compared to the CTDs. Generally, the available WOD18 metadata provide no information on the use of paired thermometers, the length of the wire out, and the wire angle.

Here we first intercompare the reference data to check their consistency/homogeneity. A total of 94 791 collocated pairs of the Nansen casts and CTD casts were found for the time period 1967–93 (Fig. 3a) and the mean and median offsets $T_{\text{BOT}} - T_{\text{CTD}}$ were calculated. According to our calculations the statistical distribution of the offsets is skewed, with the typical median (mean) offset in the upper 300 m layer about 0.01°C (0.05°C) (Fig. 3b). Since both the reversing thermometers at Nansen casts and the CTD sensors undergo laboratory calibrations we do not expect a thermal offset between the bottle and the CTD temperatures. Therefore, the small positive temperature offset may imply a slight systematic depth overestimation by the Nansen bottle casts. As the typical MBT offset is on the order of 0.1°C , we neglect the bottle–CTD offset and both the CTD and the bottle data are used together as the reference dataset.

d. Simultaneous and quasi-collocated profile pairs

For the estimation of the (potential) temperature bias and the derivation of the respective corrections we used the so-called collocation method (Ishii and Kimoto 2009; Gouretski and Reseghetti 2010; Cheng et al. 2018). In this method, a biased profile (e.g., MBT) is compared with the reference (Nansen bottle and CTD) profiles within a certain temporal and spatial vicinity of the

FIG. 3. (a) Map of Nansen cast (bottle) and CTD collocated profiles. (b) Mean and median overall temperature offsets between Nansen cast and CTD data vs depth.

biased profile. The choice of the collocation parameters is explained below in this section. For the subsequent analysis, the retained temperature profiles are interpolated on 1 m levels using the parabolic method by Reiniger and Ross (1968). A part of the biased profiles has one or several hydrographic (reference) profiles in their temporal and spatial vicinity, and for a fraction of those, there are reference profiles obtained almost simultaneously at the same geographical location. Two kinds of datasets were prepared for the analysis.

1) SIDE-BY-SIDE DATASET

The profile pairs of the MBT and reference profiles that are not more than 1 day and 4 km apart were selected to create a dataset amounting to 251 025 pairs. In most cases these pairs represent the MBT and hydrographic casts made from single ship. In this study, the side-by-side MBT dataset is mainly used to check the performance of the bias elimination models and provide additional tests for the bias model ranking.

2) QUASI-COLLOCATED DATASET

For this dataset, all pairs (including the side-by-side pairs) within a certain temporal-spatial bubble were selected. Figure 4 shows the dependence of the root-mean-square temperature difference $\Delta T = T_{\text{MBT}} - T_{\text{reference}}$ on the size of the time - distance influence bubble. The figure served as a guidance for the cases when several reference profiles are available for a single MBT profile. To reduce the effect of the temporal and spatial variability, the time-distance weighted mean of individual biases was calculated with weights decreasing with the increasing distance and the time separation. In this study, the profiles within the distance of 55 km from each other and having a time difference of not more

than 15 days are considered to be quasi collocated. This is a trade-off between the noise added through the time-spatial variability and the number of collocation pairs needed for the analysis. Figures S2a and S2b show spatial distributions of the MBT collocated and side-by-side profile pairs for main contributing countries for the whole period of MBT observations. Typically, less than 10 collocated pairs are available per one $2^\circ \times 2^\circ$ geographical box. The number of collocations per a $2^\circ \times 2^\circ$ box amounts to several hundred near the coasts of the North America, Europe, and Japan. There is a total of 902 134 MBT collocated pairs. The global-scale collocated datasets are used to finally derive at the optimal parameters of the bias models.

FIG. 4. Root-mean-square offset for all collocated pairs vs the size of the collocation bubble.

FIG. 5. Overall median MBT-bottle temperature bias in depth-temperature gradient bins. Bin size is $10 \text{ m} \times 0.02^\circ\text{C m}^{-1}$.

For each MBT profile having collocated reference stations the estimation of the *original individual profile temperature bias* is done by calculating the *median of all* MBT minus reference temperature differences. The median is used throughout the paper to reduce the impact of big outliers, which is the accepted best practice in XBT studies (e.g., Levitus et al. 2009; Cowley et al. 2013; Cheng et al. 2014).

3. Derivation and evaluation of the MBT bias correction

Temperature differences averaged within depth-vertical temperature gradient bins over all global-scale pairs are shown in Fig. 5. Especially below the seasonal mixed layer (e.g., below about 50 m) and for the typically negative temperature gradients the plot clearly demonstrates the dependence of the MBT temperature bias on the vertical gradient: for stronger gradients a larger positive temperature bias is observed, suggesting a systematic depth overestimation in the MBT profiles. The dependence is less clear in the regions with temperature inversions, at least partly because there are less observations available in these regions, for example, in the Southern Ocean.

a. MBT bias models

Since MBTs possess both temperature and pressure (depth) sensors, we investigate in the following the presence of both thermal and depth biases in the original temperature profiles similar to the XBT bias studies. We also assume that both biases are time variable and country specific. The goal is to arrive at yearly depth and temperature bias corrections. In all three models, yearly biases are calculated taking the data from a certain time

window. After some experiments, we set a 3-yr time window as a trade-off between the time resolution and the necessity to have enough collocated profiles. We considered three bias models in this study:

Model 1: MBT data are subjected only to the depth-dependent depth bias $\delta_Z(z)$ (D model).

In this model, the depth bias $\delta_Z(z) = A + Bz + Cz^2$, where z is the depth (m). Ishii and Kimoto (2009) also used the parabolic bias approximation but with the offset term $A = 0$.

The optimum sample depth correction $\delta_Z(z)$ at each level z is calculated as the depth shift, which is necessary to apply to each MBT profile within the time window in order to obtain the smallest residual absolute median temperature bias. The yearly depth correction profiles $\delta_Z(z)$ are finally approximated by a second-order polynomial versus depth.

Model 2: MBT data are subjected both to the depth-dependent depth bias $\delta_Z(z)$ and the depth-independent pure thermal bias δ_{TH} (DT model).

For this model, the thermal bias δ_{TH} is estimated as the median of all temperature differences $\Delta T = T_{MBT} - T_{reference}$ within the upper 7 m layer. As a guidance for the choice of the layer, we used the plot of the absolute median temperature difference between the surface and the lower levels (Fig. S3) based on reference data. For the depth of 7 m the median difference is close to 0.2°C . This value is often used as a criterion for the estimation of the upper mixed layer depth. Due to a negligibly small temperature gradient within the upper mixed layer, the calculated bias in this layer is most probably due to the pure thermal bias, solely. The yearly thermal bias is subtracted from the observed temperature at all observed levels of each MBT profile in the time window. The value of $\delta_Z(z)$ is calculated similar to the D model.

Model 3: MBT data are subjected only to the depth-dependent temperature bias $\beta_T(z)$ (T model).

For the T model the total yearly temperature bias $\beta_T(z)$ is calculated for each level as the *median of all* $T_{MBT} - T_{reference}$ differences and no depth correction is introduced. This is similar to the XBT and MBT bias corrections suggested by Levitus et al. (2009).

b. Metrics for the evaluation of the correction scheme performance

Here we follow Cheng et al. (2018) to quantify the performance of the correction schemes described in the literature and developed in this study. We introduce the following four metrics that provide four different integral characteristics of the original and residual biases

FIG. 6. Application of the DT model to the country-specific datasets. For each country, shown are (from left to right) derived depth bias (m), optimal thermal bias ($^{\circ}\text{C}$), original temperature bias ($^{\circ}\text{C}$), residual temperature bias ($^{\circ}\text{C}$) after depth and temperature correction, bias reduction factor, and the number of collocated pairs. Also shown are values of the metric M_3 and the bias reduction factor.

$B(\{\dots\})$ denotes the arithmetic mean, and $|\dots|$ denotes the absolute value):

- (i) Metric 1: $M_1 = \{ \{ |B(i, z)| \}_i \}_z$, with averaging over all individual biases (i) and 1 m levels (z)
- (ii) Metric 2: $M_2 = \{ \{ |B(i, \gamma)| \}_i \}_\gamma$, with averaging over all individual biases (i) and within $0.01^{\circ}\text{C m}^{-1}$ bins of the vertical temperature gradient $\gamma = \text{DT}/\text{DZ}$
- (iii) Metric 3: $M_3 = \{ \{ |B(\text{year}, z)| \}_i \}_z$, with averaging over all year-mean bias values and 1 m levels (z)
- (iv) Metric 4: $M_4 = \{ \{ |B(\text{latitude}, z)| \}_\varphi \}_z$, with averaging over 2° latitude (φ) bins and 1 m levels (z)

All metrics are calculated both for the original profiles and for the corrected profiles, with the bias reduction factor R defined for each metric as $R = \text{Metric}_{\text{original}}/\text{Metric}_{\text{residual}}$, with a higher reduction factor indicating a better scheme performance. $R < 1$ indicates that the bias correction model does not achieve bias

reduction and the residual bias exceeds the original bias. All four metrics are applied to all three models.

c. Country-specific MBT bias corrections

Bias corrections were derived for five countries that jointly contributed about 95% of all MBT profiles. Figure 6 shows yearly values of the depth correction, thermal bias, original total T bias, residual T bias, bias reduction, and the number of collocated profiles versus depth for model 2, the DT model (section 3a). As will be shown later, this model achieves the best bias reduction among other models tested. For each country, the bias reduction R is calculated by taking into account the year–depth bins with the number of collocated pairs $N \geq 500$.

Figure 6 demonstrates that the model successfully reduces the total temperature bias for all countries considered. The reduction ratio R varies between 2.319

for the British data and 6.222 for the U.S. data, with the reduction factor decreasing with the number of available collocated pairs. Both the amplitude and the original bias pattern differ between the countries. A pronounced time variability is also revealed (see Fig. 6, third column). Thus, the U.S. MBT data become increasingly warm biased between 1940s and 1970s. The Soviet MBT data are characterized by a larger positive bias during the 1970s and by a smaller bias during the 1990s. There is a sudden change from the predominantly negative bias between 1960 and mid-1970s and a predominantly positive bias after the mid-1970s for Japanese profiles. The MBT profiles obtained from the Canadian ships are typically characterized by larger positive biases compared to the other countries. The British MBT data in contrast tend to negative bias values. All these differences can be attributed to the differences in the instrumentation, calibration procedure, and to the working practices, which could be different on the ships of different nationalities. Unfortunately, no or very little information is available on changes in the MBT manufacturing process, types of MBTs, or MBT observational practices used on the British and Canadian ships.

d. Comparison of the available MBT bias correction schemes

Three MBT bias correction schemes were described in the literature. Ishii and Kimoto (2009) assumed the MBT data to be depth biased and suggested depth (z) correction in the form $\delta_z = Dz + Cz^2$, with the polynomial coefficients C and D provided for the years between 1950 and 1994. The bias derivation is based on the calculation of the depth differences between the box-averaged MBT and CTD plus bottle temperature profiles. Gouretski and Reseghetti (2010) also used box-averaged differences between the MBT and reference data to derive yearly depth and thermal corrections for the time period 1941–2003. Levitus et al. (2009) defined the total MBT temperature bias as the global MBT – bottle or CTD temperature difference. They provided temperature corrections for MBT data at 11 standard levels and for the years between 1947 and 1994.

Since the beginning of the 1980s, the mechanical bathythermographs were no more in use on American, British, Canadian, Australian, German, and French ships. Among the important MBT data contributors Japan obviously used MBTs until the end of the 1990s, whereas the USSR stopped the MBT observations around 1993. At least since the publication of the last World Ocean Database, the oceanographic MBT archive does not undergo any changes, so that all

FIG. 7. Overall median temperature bias $T_{\text{MBT}} - T_{\text{reference}}$ vs depth for the original profiles and after the application of the different bias correction models. Also shown are the values of the metric M_1 .

published MBT correction schemes described above used essentially the same MBT dataset.

Depth and temperature corrections provided by the six correction methods were also applied to the side-by-side dataset and the respective metrics were calculated. Due to an almost 4-times-fewer amount of profile pairs compared to the quasi-collocated dataset, all schemes show lower reduction factors.

Figure 7 shows overall median biases at depths for the original and corrected profiles. All schemes except that of Levitus et al. (2009) result in a bias reduction. The overall reduction is less successful below 200 m, as there is a sharp decrease of available collocated pairs for the deeper layers (Fig. 2). However, even for the deeper levels, the T, DT and D models result in the reduction of the original bias to the levels better than 0.01°C . The T model gives the lowest overall residual bias, with the DT model ranking second.

For the second metric (Fig. 8) the DT model produces the best result, with Gouretski and Reseghetti (2010) corrections ranking second. However, even the DT model is not successful in bias reduction for vertical temperature gradients less than $-0.15^\circ\text{C m}^{-1}$, especially within the upper 30 m layer, showing a certain over-correction here. Also, for the regions with positive temperature gradients the remaining positive bias is observed. This suggests the necessity for further improvements in the bias correction scheme.

Distribution of biases versus year and depth for original and corrected data (metric 3) is shown in Fig. 9.

FIG. 8. Overall median MBT temperature bias in depth–vertical temperature gradient bins for the original profiles and after the application of the bias correction models. Shown are the values of the metric M_2 and the corresponding bias reduction factors.

The MBT original bias shows significant time variation: larger positive biases before 1980 (0.05°C – 0.2°C), followed by smaller biases until 1993 ($<0.01^{\circ}\text{C}$), and larger positive biases during the midst of 1990s. The T model automatically becomes the best ranking one because the correction at each level and for each year is set equal to the original bias. The DT and D models also achieve the reduction of the original bias according to metric 3 having the second and third best score, respectively. Other correction schemes [again except that of [Levitus et al. \(2009\)](#)] also reduce the original bias. [Levitus et al. \(2009\)](#) overestimate the bias for almost all periods, leading to overall negative residuals $<-0.01^{\circ}\text{C}$ (sometimes as large as -0.15°C).

Since the optimal bias corrections in all schemes are derived by minimizing the residual biases in the year–depth plane, all schemes are less successful in the bias reduction in the latitude–depth plane (metric 4, [Fig. 10](#)). We note that all schemes perform better in the Northern Hemisphere where the majority of the collocated pairs are situated ([Fig. S2a](#)). All schemes do not reduce the negative bias in the 55° – 65°S latitude band and all show a prevailing positive residual bias between 55° and 10°S . The best performance is shown by the DT and D models. The T model is not capable of reducing the bias within the tropical 20°S – 20°N band, where a strong thermocline

exaggerates the translated positive temperature bias. In this region the bias models that take the depth bias into account [DT model, D model, and [Gouretski and Reseghetti \(2010\)](#) corrections] produce better results. [Levitus et al. \(2009\)](#) bias model overcorrects except for the tropical subsurface bands (20°S – 20°N , below 100 m). We suggest the different performance of the [Levitus et al. \(2009\)](#) correction scheme compared to other schemes is in part due to the differences in the bias estimation methods, the quality control procedure, and the reference data. The MBT dataset have not experienced significant changes since the publication of the World Ocean Database (2013).

Finally, a respective rank is ascribed to each correction scheme for each metric, and a total score for each scheme is calculated as the sum of ranks ([Table 2](#)). According to the table, the DT, T, and D models demonstrate superior results (total scores 12, 19, and 25) followed by the correction schemes suggested by [Gouretski and Reseghetti \(2010\)](#) and [Ishii and Kimoto \(2009\)](#) with the total scores of 30 and 34, respectively. The [Levitus et al. \(2009\)](#) corrections do not reduce the original bias. We also note here that the performance assessment based on the metric 3 might not be fair to other correction schemes, because the data used to derive the new corrections are the same as the data used for the performance assessment.

FIG. 9. (top) Original and (bottom) residual median MBT temperature bias vs depth and year for different bias correction schemes. Shown are the values of the metric M_3 .

Therefore, metric 3 is not used to calculate correction scheme ranks. The other metrics provide the evaluation, which is partially dataset independent, since none of the methods explicitly accounts for the geographical variation of the bias. Nevertheless, a full intercomparison of the available schemes should be made in the future in a more comprehensive way based on a community-agreed dataset similar to the

dataset used in [Cheng et al. \(2018\)](#) for the XBT bias correction schemes.

e. Assessment of the correction scheme robustness

To test the robustness of the correction schemes, a “throwaway” test was adapted, similar to [Cheng et al. \(2018\)](#). Here we divide the total set of collocated pairs in two subsets. For each of the six main data contributors

FIG. 10. Median MBT original and residual bias vs depth and latitude for different correction schemes. Shown are the values of the metric M_4 .

TABLE 2. Scores of different correction schemes and corresponding reduction factors (in parentheses). Total score is based on metric 1, metric 2, and metric 4 (see text). The best scores are marked in boldface.

	Collocated dataset				Side-by-side dataset				Total score
	Metric 1	Metric 2	Metric 3	Metric 4	Metric 1	Metric 2	Metric 3	Metric 4	
DT model	2 (22.700)	1 (2.953)	2 (2.895)	1 (1.783)	2 (5.028)	1 (2.167)	2 (1.577)	1 (1.539)	8
T model	1 (133.989)	3 (1.869)	1 (—)	4 (1.487)	1 (9.446)	5 (1.498)	1 (1.739)	3 (1.444)	17
D model	3 (13.129)	5 (1.700)	3 (2.373)	2 (1.616)	3 (4.383)	4 (1.614)	3 (1.518)	2 (1.475)	19
GR2010 model	4 (3.310)	2 (2.478)	5 (1.186)	3 (1.520)	4 (3.048)	3 (1.622)	5 (1.084)	4 (1.407)	20
IK2009 model	5 (2.777)	4 (1.671)	4 (1.236)	5 (1.234)	5 (1.409)	2 (1.826)	4 (1.087)	5 (1.334)	26
Levitus et al. (2009) model	6 (0.959)	6 (1.069)	6 (0.750)	6 (1.007)	6 (0.508)	6 (0.748)	6 (0.748)	6 (1.022)	36

the total yearly number of profiles was calculated and the eastern dataset containing 50% of all data was created by summing up the profiles moving eastward from the Greenwich meridian. The remaining part of the profiles builds the western dataset. The Greenwich meridian is the western boundary for all six country-specific eastern datasets, and the country-specific western datasets have the Greenwich meridian as the eastern boundary. The geographical distribution of the collocated profiles for the eastern and western datasets is shown in Figs. S2c and S2d.

Here, we tested only the DT model for robustness since this model is characterized by the best ranking according to the model comparison test. The DT model was applied independently to the eastern and western datasets. Figure 11 shows original and residual biases after the application of the global and dataset-specific corrections to each of the datasets, along with the metrics and the reduction factors. First, we note that the original biases derived separately for the eastern and western datasets demonstrate a high degree of similarity. Thus, a gradual decrease of the warm bias from about 0.15°C to close to 0°C is observed from the 1950s to the 1990s (Figs. 11a,e). High positive bias values are found in the tropical belt between about 25°N and 25°S and in the latitude belt of 35°–50°S (Figs. 11i,m). Also, for both datasets an increase of the warm bias with the increasing temperature gradient is observed (Figs. 11q,u). This similarity suggests that the MBTs from the two datasets are characterized by similar biases.

The global thermal and depth corrections being applied to the each of the datasets results in the reduction of the total temperature bias (Figs. 11b,f,j,n,r,v). For each metric, the bias reduction is higher for the eastern dataset (Fig. 11, top row of panels for each metric). We explain this both by the differences between the datasets and by the different input from the contributing countries each having specific biases. In the western dataset, for instance, the majority of the MBT profiles comes from the Atlantic Ocean, which is characterized by a

larger spatial variability compared to the other oceans. If the eastern corrections are applied to the eastern dataset (Figs. 11c,k,s) and the western corrections are applied to the western dataset (Figs. 11g,o,w) the residual metrics and the reduction factors are close to the case of the global corrections. Application of the eastern corrections to the western dataset and vice versa leads to the highest residual metrics and the lowest reduction factors (except for the second metric when eastern corrections are applied to the western dataset). Since in all cases the corrections lead to the bias reduction, we conclude that the bias correction model is robust and does not critically depend on the choice of the dataset used to derive the corrections.

4. Discussion and conclusions

In this study we developed new bias correction schemes for the temperature profiles obtained by means of mechanical bathythermographs. The MBT temperature profiles are compared with the collocated CTD and Nansen cast profiles from the World Ocean Database 2018 to derive optimal depth and temperature corrections for systematic errors. The reference profiles were checked for consistency, and the systematic offsets between the reference data types were found to be negligible for the purpose of this study.

For the MBT profiles, we found that the derived optimum depth and temperature corrections are country specific. Therefore, the new corrections have been calculated separately for five countries (United States, USSR, Japan, Canada, and United Kingdom) whose contribution amounts to 95% of the total MBT profiles. The sixth correction group includes all other countries. Because of the absence of the necessary metadata on manufacturer, probe type, observational practice, and calibration procedure we are unable to explain the differences between the country-specific depth and temperature offsets.

The new MBT correction schemes were compared with three corrections schemes available in the literature. We found that except for the scheme by Levitus et al. (2009) all

FIG. 11. Median MBT original and residual bias for the eastern dataset vs depth and year: (a) original bias; (b) residual with global corrections; (c) residual with eastern corrections; (d) residual with western corrections. Median MBT original and residual bias for the western dataset vs depth and year: (e) original bias; (f) residual with global corrections; (g) residual with western corrections; (h) residual with eastern corrections. (i)–(p) As in (a)–(h), respectively, but for the biases vs depth and latitude. (q)–(x) As in (a)–(h), respectively, but for the biases vs depth and vertical temperature gradient. Also shown are the values of the respective metrics and the bias reduction factors.

other schemes reduce the total temperature bias, with the schemes from this study being superior according to the ranking.

To objectively assess the performance of the schemes, four metrics were introduced and for each correction scheme bias reduction factors were calculated. The bias reduction is observed for all metrics both for the quasi-located and for the side-by-side data. Not unexpectedly, the bias reduction factor is lower for the side-by-side dataset, which has 3.6 times fewer pairs compared to the quasi-located dataset. We found, that even after application of the corrections a positive bias (up to 0.1°C) is still observed in the Southern Hemisphere, probably because the bias estimation in the Southern Hemisphere is less reliable due to a much fewer number of available located pairs. Based on the model intercomparison results, we recommend applying MBT corrections using the DT model, which takes into account both the depth bias and the thermal bias. Recommended corrections are given in the Table S1.

Acknowledgments. This study is funded by National Key R&D Program (2016YFC1401800), Key Deployment Project of Centre for Ocean Mega-Research of Science, CAS (COMS2019Q01) and Chinese Academy of Sciences (CAS) President's International Fellowship Initiative (PIFI). We are thankful to the staff of the National Center for Environmental Information for preparing the invaluable collection of the hydrographic data that served as a basis for this study. We are thankful to John Gould, Shoichi Kizu, Jens Meincke, Detlef Machozek, and V. Maslennikov, who provided information about the MBT manufacturers and practices and pointed to literature references not known to us. We also acknowledge Jiang Zhu for his support during the preparation of this work. We appreciate useful comments on the accuracy of CTD temperature measurements provided by Steve Diggs, Tim Boyer, and Jim Swift. Numerous comments from Peter Koltermann helped to improve the original version of the manuscript. The data used in this study can be found in <http://159.226.119.60/cheng/>. Finally, our thanks go to two anonymous reviewers whose detailed comments and suggestions were highly beneficial for the manuscript.

REFERENCES

- Boyer, T. P., and Coauthors, 2018: World Ocean Database 2018. NOAA Atlas NESDIS 87, 207 pp.
- Bralove, A. L., and E. I. Williams Jr., 1961: A study of errors of the bathythermograph. National Scientific Laboratories Rep., 51 pp.
- Casciano, L., 1967: Mechanical bathythermographs. *Geo-Mar. Technol.*, **3**, 19–21.
- 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, <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. E. Trenberth, J. T. Fasullo, T. Boyer, J. Abraham, and J. Zhu, 2017: Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.*, **3**, e1601545, <http://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>.
- Cowley, R. S., 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>.
- Dinkel, C. R., and M. Stawnychy, 1973: Reliability study of mechanical bathythermographs. *Mar. Technol. Soc. J.*, **7**, 41–47.
- Gautam, S., and P. Thadathil, 2016: Temperature and depth error in the mechanical bathythermograph data from the Indian Ocean. *Indian J. Geo-Mar. Sci.*, **45**, 1288–1291.
- Goni, G., and Coauthors, 2019: More than 50 years of successful continuous temperature section measurements by the global expendable bathythermograph network, its integrity, societal benefits, and future. *Front. Mar. Sci.*, **6**, 452, <https://doi.org/10.3389/fmars.2019.00452>.
- Good, S., 2011: Depth biases in XBT data diagnosed using bathymetry data. *J. Atmos. Oceanic Technol.*, **28**, 287–300, <https://doi.org/10.1175/2010JTECH0773.1>.
- 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>.
- , 2018: World Ocean Circulation Experiment–Argo Global Hydrographic Climatology. *Ocean Sci.*, **14**, 1127–1146, <https://doi.org/10.5194/os-14-1127-2018>.
- , and K. P. 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>.
- Hazelworth, J. B., 1964: Statistical analysis of the thermal structure at ocean weather station Echo. U.S. Naval Oceanographic Office Tech. Rep. 146, 76 pp., <https://doi.org/10.5962/bhl.title.47867>.
- , 1966: Quantitative analysis of some bathythermograph errors. U.S. Naval Oceanographic Office Tech. Rep. 180, 27 pp.
- IPCC, 2007: *Climate Change 2007: Synthesis Report*. IPCC, 104 pp.
- , 2019: Summary for policymakers. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, IPCC, 27 pp.
- 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>.
- Kirejev, I. A., 1939: Tables for actual sample depth of the series of bathometer (in Russian). *Sev. Morskoy Put*, **14**, 85–99.

- Levitus, S., J. I. Antonov, T. P. Boyer, and C. Stephens, 2000: Warming of the World Ocean. *Science*, **287**, 2225–2229, <https://doi.org/10.1126/science.287.5461.2225>.
- , —, and —, 2005: Warming of the World Ocean, 1955–2003. *Geophys. Res. Lett.*, **32**, L02604, <https://doi.org/10.1029/2004GL021592>.
- , J. Antonov, T. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov, 2009: Global ocean heat content 1955–2007 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.*, **39**, L10603, <https://doi.org/10.1029/2012GL051106>.
- Meyssignac, B., and Coauthors, 2019: Measuring global ocean heat content to estimate the Earth energy imbalance. *Front. Mar. Sci.*, **6**, 432, <https://doi.org/10.3389/fmars.2019.00432>.
- Pavlov, V. M., and V. I. Kuksa, 1957: Experience of work with shipboard thermobathygraph TB-52 (in Russian). *Trudy Inst. Okeanol.*, **25**, 88–97.
- Reiniger, R. F. and C. K. Ross, 1968: A method of interpolation with application to oceanographic data. *Deep-Sea Res. Oceanogr. Abstr.*, **15**, 185–193, [https://doi.org/10.1016/0011-7471\(68\)90040-5](https://doi.org/10.1016/0011-7471(68)90040-5).
- Rhein, M., and Coauthors, 2013: Observation: Ocean. *Climate Change 2013: The Physical Science Basis*, Cambridge University Press, 255–315.
- Shor, E. N., 1978: *Scripps Institution of Oceanography: Probing the Oceans 1936 to 1976*. Tofua Press, 502 pp.
- Soskin, I. M., 1977: *Handbook of Hydrological Studies in Oceans and Seas* (in Russian). Gidrometeoizdat, 719 pp.
- Spilhaus, A. F., 1938: A bathythermograph. *J. Mar. Res.*, **1**, 95–100, <https://doi.org/10.1357/002224038806440647>.
- State Oceanographic Institute, 2016: *Manual of Hydrographic Work at Sea* (in Russian). State Oceanographic Institute, 537 pp.
- UNESCO, 1966: Oceanographic instruments and equipment. *Int. Mar. Sci.*, **IV**, **3**, 1–43.
- Warren, B., 2008: Nansen-bottle stations at the Woods Hole Oceanographic Institution. *Deep-Sea Res. I*, **55**, 379–395, <https://doi.org/10.1016/j.dsr.2007.10.003>.
- Wijffels, S. E., J. Willis, C. M. Domingues, P. Barker, N. J. White, A. Gronell, K. Ridgway, and J. A. 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>.
- Wüst, G., 1933: Thermometric measurement of depth. *Hydrogr. Rev.*, **10**, 28–49.