Reassessing biases and other uncertainties in sea-surface temperature observations measured *in situ* since 1850, part 2: biases and homogenisation

J. J. Kennedy,[1]N. A. Rayner,[1]R. O. Smith,[2]D. E. Parker,[1]and M. Saunby[1]

Abstract. Changes in instrumentation and data availability have caused time-varying biases in estimates of global- and regional-average sea-surface temperature. The size of the biases arising from these changes are estimated and their uncertainties evaluated. The estimated biases and their associated uncertainties are largest during the period immediately following the Second World War, reflecting the rapid and incompletely documented changes in shipping and data availability at the time. Adjustments have been applied to reduce these effects in gridded data sets of sea-surface temperature and the results are presented as a set of interchangeable realisations. Uncertainties of estimated trends in global- and regional-average sea-surface temperature due to bias adjustments since the Second World War are found to be larger than uncertainties arising from the choice of analysis technique, indicating that this is an important source of uncertainty in analyses of historical sea-surface temperatures. Despite this, trends over the twentieth century remain qualitatively consistent.

1. Introduction

Historical records of sea-surface temperature (SST) are essential to our understanding of the earth's climate. Data sets of SST observations are used to detect climate change and attribute the observed changes to their several causes. They are used to monitor the state of the earth's climate and predict its future course. They are also used as a boundary condition for atmospheric reanalyses and atmosphere only general circulation models (IPCC 2007).

SSTs have been observed by diverse means in the past 160 years. As a result, measurements of SST recorded in historical archives are prone to systematic errors - often referred to as biases - that are of a similar magnitude to the expected climate change signal. That biases exist is well documented. *Folland and Parker* [1995] (FP95) applied temporally- and geographically-varying adjustments to SST data prior to 1942 of several tenths of a degree to correct for the widespread use of canvas and wooden buckets in the collection of water samples. Further calculations by *Rayner et al.* [2006] (R06) showed that the uncertainties in these adjustments were much smaller than the adjustments themselves, an assertion backed up by *Smith and Reynolds* [2002] (SR02) who evaluated the adjustments using an independent method.

The FP95 adjustments stopped in 1942 because they assumed that the SST measurements made after this date were taken using a mixture of methods that thereafter remained more or less unchanged. SR02 identified a possible bias in the post-war period but did not attempt to correct for this. *Kent and Taylor* [2006] conducted a review of literature on SST biases from the 1920s to the present, which indicated significant biases in measurements made using buckets and in the engine rooms of ships - the most common means by which SST measurements have been made *in situ*. *Smith and Reynolds* [2005] allowed for an uncertainty owing to biases of around 0.1K from 1942 "based on typical differences between all observations and ship intake temperatures in ICOADS". More recently, *Thompson et al.* [2008] identified a discontinuity in the record of global-average sea-surface temperature of around 0.3K that coincided with an abrupt change in data sources - from US to UK ships - in the International Comprehensive Ocean Atmosphere Data Set (ICOADS *Worley et al.* [2005]) archive of marine observations in the mid 1940s.

Since the 1970s, a growing number of SST measurements have been made by drifting and moored buoys. *Emery et al.* [2001] identified a warm bias of 0.15K in SST measurements made by ships relative to those made by drifting buoys. *Smith et al.* [2008] predicted that the difference would lead to a growing cool bias in the observed SST record as the number of drifting buoys in the observing array increased. *Kennedy et al.* [2011a] compared global-average SST as measured *in situ* with that retrieved from satellite measurements and estimated that the bias amounts to a shortfall in warming of almost 0.1K between 1991 and 2007.

So, there are significant biases in the SST record after 1941, but despite attempts to quantify the approximate size of these biases on global scales (*Smith and Reynolds* [2005]), no attempt has been made to adjust SST records to account for them. This paper describes a method for estimating the biases in gridded SST products due to known discontinuities in the data base of observations, as well as more general changes in observing practice over time. The method relies on metadata from a variety of sources to build as complete a picture as possible of the way that measurements were made on board ships. Despite the wealth of metadata that is now available, it is not possible to estimate the biases in an exact manner so an attempt has been made to assess the potential uncertainties in the biases that arise from assumptions made in the process of aggregating the information. The bias estimates are used to adjust the SST data to create a new, more homogeneous data set of anomalies relative to the 1961-1990 average. The adjusted data are presented as an ensemble of 100 interchangeable realisations and together with the new uncertainty estimates described in part 1 of the paper (*Kennedy et al.* [2011b]) they constitute the third version of the Met Office Hadley Centre gridded SST data

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK.

[2]Ocean Physics Group, Department of Marine Science, University of Otago, Dunedin, New Zealand

0148-0227/11/$9.00

set, HadSST3. The HadSST3 data set is publically available from http://www.metoffice.gov.uk/hadobs.

It should be noted that the adjustments presented here and their uncertainties represent a first attempt to produce an SST data set that has been homogenized from 1850 to 2006. Therefore, the uncertainties ought to be considered incomplete until other independent attempts have been made to assess the biases and their uncertainties using different approaches to those described here.

The data and metadata used are described in more detail in Section 2. In Section 3 the literature and other metadata are reviewed. The information was used to generate a range of possible estimates for variables - such as the relative fractions of insulated and uninsulated buckets - that might have a significant bearing on the biases. The bias estimation method, described in Section 4, used the ranges to produce multiple realisations of the estimated biases so that the uncertainties in the method could be explored. A more general discussion of the results is found in Section 5 and a number of the remaining issues are discussed in Section 6.

2. Data and metadata

The SST data for 1850-2006 come from version 2.5 of the International Comprehensive Ocean Atmosphere Data Set (ICOADS *Woodruff et al.* [2010]). ICOADS comprises meteorological measurements, principally from ships and buoys. ICOADS also contains a large number of metadata. The metadata relevant to the analysis are: sea-surface temperature method indicator (SI); recruiting country code (C1); platform type (ship, drifting buoy, moored buoy etc., PT); and deck (DCK). The deck identifies the source of the data and refers to the decks of punched cards on which earlier versions of the data set were based. There are three principal types of platform measuring SST *in situ*: ships, drifting buoys and moored buoys.

Most of the ship-based data were taken by ships engaged on other business. The Voluntary Observing Ships (VOS) were recruited into national fleets and issued with standardised equipment and instructions. Although countries standardised equipment within their fleets, there have always been differences between countries concerning best practice. The size of the VOS fleet peaked around 1985, when there were more than 7500 ships in the World Meteorological Organisation's VOS fleet. Numbers have declined since, with fewer than 4000 ships remaining on the list today (http://www.vos.noaa.gov/vos_scheme.shtml).

Some VOS use a bucket to haul a sample of near-surface water to the deck for measurement. During hauling, the water sample in a bucket can lose heat via evaporation and can be cooled or, less commonly, warmed by exchange of sensible heat with the air. The rate of heat loss depends on the type of bucket. Canvas buckets can lose large amounts of heat, whereas buckets made from rubber or wood are effectively insulated and therefore less prone to cooling biases (*Folland and Parker* [1995]). In the nineteenth and early twentieth century wooden and, later, canvas buckets were used. The use of canvas buckets continued until the 1950s when the problem of cooling biases became apparent and they were gradually phased out of use. Rubber, or otherwise insulated buckets, are now the norm.

Other VOS measure the temperature of the water taken in to cool the engines, or for other purposes such as refrigerating cargo. The measurements, often made in the engine room, are known as Engine Room Intake, or Engine Room Inlet (ERI) measurements. As the inlet must be below the water line whatever the loading, it is often several metres below the surface. ERI measurements might therefore be expected to exhibit a cold bias relative to the temperature at the surface because of the greater depth from which the water is drawn. However, the biases associated with ERI measurements have been found to depend on the circumstances peculiar to each vessel (*James and Fox* [1972]) and ERI measurements are most often biased warm due to heating of the water sample by the superstructure of the ship as it passes through pipes and even pumps on its way to the engine room.

The third means by which ships routinely measure SST is via dedicated hull contact sensors, attached either to the outside of the ship or in good thermal contact with the inside of the hull. However, such sensors can be expensive to install and maintain and were not commonly used until recently. Although hull contact sensors are expected to be more consistent than ERI measurements, there are very few studies looking at their accuracy in practice. An analysis of ship observations in the VOSClim project suggested that SST from hull sensors showed some differences to those from engine intakes and were likely to be less noisy (as was found in the VSOP-NA). The analysis did not take in to account any possible regional or environmental effects and did not test the statistical significance of the differences (Elizabeth Kent, personal communication). *Kent et al.* [2010] note the need for further evaluation of hull mounted sensors.

Table 1. List of 5 degree latitude by longitude regions from which observations from deck 732 were excluded at various times. The times at which these regions were excluded are shown in Table 2. W and E refer to the longitudes of the western and eastern edges of the region. S and N refer to the latitudes of the southern and northern edges of the region.

Region	W	S	E	N
1	-175	40	-170	55
2	-165	40	-160	60
3	-145	40	-140	50
4	-140	30	-135	40
5	-140	50	-130	55
6	-70	35	-60	40
7	-50	45	-40	50
8	5	70	10	80
9	0	-10	10	0
10	-30	-25	-25	-20
11	-60	-50	-55	-45
12	75	-20	80	-15
13	50	-30	60	-20
14	30	-40	40	-30
15	20	60	25	65
16	0	-40	10	-30
17	-135	30	-130	40

Table 2. Regions (as defined in Table 1) from which observations from deck 732 were excluded in each year

Year	Regions excluded
1958	1, 2,3,4,5,6,14,15
1959	1,2,3,4,5,6,14,15
1960	1, 2, 3, 5, 6, 9, 14, 15
1961	1, 2, 3, 5, 6, 14, 15
1962	1, 2, 3, 5, 12, 13, 14, 15, 16
1963	1, 2, 3, 5, 6, 12, 13, 14, 15, 16
1964	1, 2, 3, 5, 6, 12, 13, 14, 16
1965	1, 2, 6, 10, 12, 13, 14, 15, 16
1966	1, 2, 6, 9, 14, 15
1967	1, 2, 5, 6, 9, 14, 15
1968	1, 2, 3, 5, 6, 9, 14, 15
1969	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15
1970	1, 2, 3, 4, 5, 6, 8, 9, 14, 15
1971	1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 14
1972	4, 7, 8, 9, 10, 11, 13, 16, 17
1973	4, 7, 8, 10, 11, 13, 16, 17
1974	4, 7, 8, 10, 11, 16, 17

Drifting buoys consist of a plastic ball, approximately 30 cm in diameter, attached to a drogue. The drogue ensures that the buoy remains correctly oriented and that it drifts with the currents in the mixed layer. The SST sensor is embedded in the underside of the buoy and measures at a depth of approximately 25cm in calm seas. Movement of the buoy and the action of waves mean that the measurement is representative of the upper 1m of the water column. The design of drifting buoys was standardised in the early 1990s; consequently, measurements from drifting buoys should be consistent at all times and places thereafter. In contrast, moored buoys come in a wide variety of shapes and sizes, from the 10m discus buoys to the 1.5m fixed buoys deployed in the North Sea. Although individual drifting buoys do exhibit drifts in calibration, the measurements are generally more reliable than ship observations and are thought to give a generally unbiased estimate of SST.

Additional metadata were found in WMO Publication 47, "List of selected, supplementary and auxiliary ships." WMO Pub 47 is now published quarterly, but was published annually before 1998. An almost complete set of annual issues is available in digital format from 1956 to the present. Each entry lists, amongst other things: the ship's name and call sign, the country that recruited the ship, and the method used to measure SST. For a more comprehensive review of WMO Pub 47 see *Kent et al.* [2007]. Metadata in WMO Pub 47 are referenced by the ship's name and callsign. The callsign information is recorded in ICOADS metadata until December 2007. After this date, the callsign information was removed from some real time GTS feeds, and encrypted on others, owing to concerns about ship security. Lack of call sign information meant it was only possible to complete this analysis to 2006.

The ICOADS 2.5 data were quality controlled according to the procedure described in R06. In R06, ships with call sign 0120 were not submitted to track check as this call sign was used by a number of different ships at the same time. In processing the ICOADS 2.5 data additional duplicate call signs (0120, SHIP, PLAT, RIGG, MASKST, "1 ", "58" and, from 1948-1954, "7 ") were identitified and these observations were not submitted to the track check.

A manual scan of the data was performed after gridding the observations at monthly 1 degree latitude by 1 degree longitude resolution. Some observations from deck 732 between 1958 and 1974 were identified as being incorrectly located. A number of these areas were identified in R06. Seventeen 5-degree areas or blocks of 5-degree areas were obviously artificially warm or cold relative to neighbouring areas and relative to other observations within the areas. Data from Deck 732 were not used from the areas specified in Table 1 at the times specified in Table 2.

3. Review of literature and metadata

Bearing in mind the difference in bias between observations taken using insulated and uninsulated buckets, the difficulty of estimating ERI biases and the need to assign metadata to observations, the relevant literature is summarised below with three principal aims:

1. To ascertain how many insulated and uninsulated buckets were in use in the global VOS fleet at any time (Section 3.1).

2. To summarise estimates of biases in ERI measurements (Section 3.2).

3. To maximise the number of observations that can be attributed to a given measurement method (Section 3.3).

In addition, new estimates were made of the biases between drifting buoy and ship observations (Section 3.4).

3.1. Buckets

Dating the switchover from uninsulated canvas buckets to insulated rubber buckets is problematic as it is not clear how quickly the practice of using insulated buckets was adopted. According to the Marine Observers Handbook, buckets issued by the UK Meteorological Office were constructed from canvas until 1957 (*HMSO* [1963]). The Dutch contribution to WMO Technical Report 2 (*WMO* [1954]) stated that "Water samples are taken with an ordinary canvas bucket". Guidance issued to Japanese ships (*Okada* [1922], *Okada* [1927], *Tsukada* [1927], *Okada* [1929], *Okada* [1932], *Horiguchi* [1940]) states that a canvas bucket should be used to collect a water sample. The "Law for Marine Meteorological Elements issued by Kobe Marine Meteorological Observatory", states that canvas buckets were the recommended means of making SST measurements in 1951 and 1956. Thus, it would seem that canvas buckets were in use after the Second World War and until at least 1956.

The British contribution to WMO Technical Report no 2 (*WMO* [1954]) says, "it will be recommended to replace the simple canvas bucket by an insulated one as soon as a satisfactory model becomes available". In Japanese instructions from 1956 onwards, it is stated that a "rubber bucket with doubled layers may be used". In November 1957, the UK Meteorological Office issued its first rubber buckets and large numbers were purchased from at least 1963 onwards. However, *HMSO* [1969] suggests that some canvas buckets were still in use on British ships until the late 1960s. The 1977 edition of the Marine Observer's Handbook still mentions canvas buckets, but the preference is for rubber buckets. Danish mariners were still using canvas buckets as late as 1980 (DMI memorandum, Bennert Machenhauer 1989), but DMI started to provide plastic buckets to its observing ships starting in 1975 and estimate that it took at least 5 years before all canvas buckets were replaced. A Met Office note in our archive dated 1961 shows that Canadian ships had begun to use the 'G.H. Zeal rubber bucket' by at least April 1961. *James and Fox* [1972] analysed observations from 1968 to 1970 and found that around 5% of observations in their sample still came from uninsulated canvas buckets. Canvas buckets might have remained in use on Japanese ships until at least 2004. Instructions after the 1950s stipulate that either a canvas bucket or insulated rubber bucket might be used (Professor Kimio Hanawa personal communication, quoting from Guideline for Observation of Marine Meteorology). However, after the 1970s, the number of buckets in use on Japanese ships was small so it is assumed here for simplicity that the schedule of the changeover on board Japanese ships, and indeed all ships, followed the same pattern.

If a linear switchover is assumed which started in 1954 and was 95% complete in 1969, the middle of the James and Fox study period, then the switchover would have been completed by 1970. Based on the literature reviewed here, the start of the general transition is likely to have occurred between 1954 and 1957 and the end between 1970 and 1980.

3.2. Estimates of ERI bias

The literature contains many contemporary estimates of ERI biases, which are summarised in Table 3. The typical bias is around 0.2°C and most estimates fall between 0.1°C and 0.3°C. The larger biases are typically from small samples of ships or refer to comparisons between ERI and canvas bucket measurements.

The majority of US SST measurements after the Second World War were probably made using the ERI method. Metadata in ICOADS from US Merchant Marine ships (Decks 705-707) identify large numbers of observations made using ERI prior to 1941. The number increases from zero in the 1920s to the majority of ships in 1941. This is consistent

Table 3. Estimates of ERI bias from literature.

Reference	Bias	Number of obs	comparison	Year
Brooks [1926]	0.13F	1 ship	tin bucket	1926
Brooks [1926]	0.5F	1 ship	tin bucket	1926
Brooks [1928]	0.3C	1 ship	tin bucket	1928
	0.8C	1 ship	tin bucket	1928
	0.2C	1ship	tin bucket	1928
Lumby [1927]	0.1F	1 ship	bucket	1927
	-0.2F	1 ship	bucket	1927
Collins et al. [1975]	0.3C	?	Japanese bucket	1927-1933
Wahl [1948]	0.25C	1 ship	German bucket	1948
Roll [1951]	0.07C	1 ship	German bucket	1951
Kirk and Gordon [1952]	0.25F	3 ships	insulated bucket	1952
	1F	many ships	canvas bucket	1952
Amot [1954]	0.1C	2 ships	reversing thermometer	1954
Saur [1963]	1.2F	12 ships 6828 obs	insulated bucket	1963
	1F	?	bucket	1939-1945
Walden [1966]	0.3C	13847 obs	German bucket	1966
Knudsen [1966] in *Tabata* [1978a]	0.1C	?	bucket	1966
Tauber [1969]	0.5-2.3C	several ships	?	1969
James and Fox [1972]	0.3C	13876 obs	bucket	1968-1970
Tabata [1978a]	0.30C	several ship	CTD	1975
Tabata [1978b]	0.2C	several ships	moored buoy	1975-1976
Folland et al. [1993]	0.11C	?	non-bucket vs bucket	1975-1981
Kent et al. [1993]	0.3C	45 ships	forecast model SST	1988-1990
Kent and Kaplan [2006]	0.09C	8410 obs	true bias	1975-1979
	0.15C	11245 obs		1980-1984
	0.18C	11073 obs		1985-1989
	-0.13C	5122 obs		1990-1994

with US Weather Bureau instructions summarised in *Elms et al.* [1993]. The advice given to US observers between 1906 and 1929 (*Bureau* [1925, 1929]; *Heiskell* [1908, 1910]; *Page* [1906]) was to use a bucket. In the 1938 *Bureau* [1938] and 1941 *Bureau* [1941] editions, the advice allows ERI measurements to be made, leaving it to the "judgement of the individual observer". Metadata from decks 705-707 attest to the fact that take up of the more convenient method was rapid after 1938. From 1950 to 1964, only the ERI method is recommended. Around 90% of US observations in the 1955 edition of WMO Pub 47 were made by ERI and only a few percent by bucket. The remaining entries were left blank.

Before and after the Second World War, UK merchant ships were generally instructed to use buckets to make SST measurements. However, newly digitised data from the UK Royal Navy deck logs (deck 245) is of a different character. An inventory of the archives conducted in 1994 identified observations in the Navy deck logs as being made using the ERI method. As it is the deck logs that are digitised in deck 245, observations from this deck ought to be treated as ERI measurements.

It is probable that some observations recorded as being from buckets were made by the ERI method. The Norwegian contribution to WMO Tech note 2 (*Amot* [1954]) states that the ERI method was preferred owing to the dangers involved in deploying a bucket. This is consistent with the first issue of WMO Pub 47 (1955), in which 80% of Norwegian ships were using ERI measurements. US Weather Bureau instructions (*Bureau* [1938]) state that the "condenser-intake method is the simpler and shorter means of obtaining the water temperature" and that some observers took ERI measurements "if the severity of the weather [was] such as to exclude the possibility of making a bucket observation". The only quantitative reference to the practice is in the 1956 UK Handbook of Meteorological Instruments *HMSO* [1956] which states that ships that travel faster than 15 knots should use the ERI method in preference to the bucket method for safety reasons. Approximately 30% of ships travelled at this speed between 1940 and 1970.

3.3. Assigning metadata to observations

In order to estimate the biases in the data, observations must be associated with the methods used to make them. Metadata from ICOADS and WMO Pub 47 were used for the identification. Unfortunately, the metadata from any single source were incomplete. Therefore, it was necessary to combine metadata from different sources to produce a more complete record.

Prior to 1941, all observations were assumed to be bucket measurements unless positively identified as ERI measurements by the ICOADS SI metadata. After 1941 the following procedure was used:

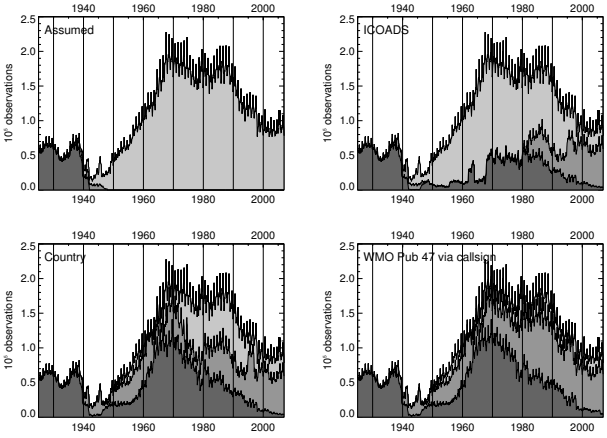

Figure 1. Numbers of SST observations from ships for different measurement methods (1925-2006). Buckets (dark grey), ERI and Hull Contact sensors (mid grey), unknown (light grey). (top left) Observations identified by default: i.e. all observations prior to 1941 are bucket observations unless otherwise stated, Royal Navy observations from World War 2 are ERI measurements. (top right) observations identified using information in ICOADS and by default. (bottom left) observations identified using country information, ICOADS and by default. (bottom right) observations identified using all methods.

Table 4. ICOADS decks associated with ICOADS country indicator and assumed country.

Country	ICOADS C1	ICOADS deck
Netherlands	0	150, 189, 193
USA	2	001-006, 110, 116, 117, 195, 281, 666, 667, 700, 701, 705-707
UK	3	152, 184, 194, 902
Japan	17	118, 119, 187, 762, 898
USSR	25	185, 186, 732, 733
Germany	40	151, 192, 196, 772

1. If the observation was from a drifting buoy or moored buoy, the observation was recorded as a buoy measurement.

2. If the observation was from a ship and a measurement method was present in ICOADS, that was used.

3. From 1956, if no measurement method was given in ICOADS, but the ship's call sign matched an entry in WMO Pub 47 and a measurement method was present for the call sign, then that was used.

4. Otherwise the recruiting country of the ship was found from either the ICOADS country ID (preferred) or the deck ID (see Table 4). The country was used to extract a typical observation method from WMO Pub 47. For example, in 1970, around 50% of Japanese ships were registered in WMO Pub 47 as using buckets and 50% as using ERI. In this case an observation from a Japanese ship, to which no specific measurement method could be assigned, was counted as 0.5 bucket observations and 0.5 ERI observations.

5. If no identification could be made, the measurement method was recorded as unknown.

6. If the observation came from a US ship after 1945 it was counted as an ERI measurement. If the observation came from deck 245 (recently digitised UK Royal Navy data) it was counted as an ERI measurement.

The properties of measurements made using hull contact sensors are not well established, so measurements made using this method were grouped together with the ERI observations. The different types of buckets are not distinguishable using the ICOADS and WMO Pub 47 metadata so they were all considered to be generic bucket observations at this step. In 2007, the call signs of GTS reports were anonymised or encrypted making it impossible to connect metadata with observations. Therefore this analysis ends in 2006. *Rayner et al.* [2009] recommend that the practice of anonymising meteorological reports should be discontinued or circumvented, as this is a barrier to the continued production of climate records from *in situ* data.

Total numbers of SST observations and their breakdown into different observation types are shown in Figure 1. Prior to 1965, the measurement methods for the vast majority of observations were ascertained using the recruiting country of the ship (Step 4 above). After 1946 and particularly after 1965, a large number of observations in ICOADS included measurement method metadata indicating that the observation was a bucket observation (Step 2 above). After 1965, significant numbers of observations could also be associated with entries in WMO Pub 47 via the call sign (Step 3 above). Many of the observations identified via WMO Pub 47 were made by ships using the ERI method. Using all this information meant that a measurement method could be assigned to more than 70% of observations at all times.

Figure 1 shows that when a new source of metadata was added to the analysis, the newly identified observations were not distributed in like ratio to those that had already been identified. It is dangerous, therefore, to assume that observations that cannot be associated with a measurement method should be assigned as bucket and ERI observations in the same ratios as those where the measurement method is known.

The numbers of observations made by each measurement method (or fractions of observations if the method was identified via the recruiting country) were summed onto separate monthly 5° latitude by 5° longitude grids. In each grid box the totals were expressed as a fraction of all observations in the grid box and the contributions of each measurement method to the global-average SST are shown in Figure 2. Canvas, wooden and rubber buckets are not shown separately. From 1900 to 1933 100% of observations in the ICOADS 2.5 data base are considered to be bucket observations. Starting in 1933, there is a rise in the fractional contribution of ERI observations, which starts off slowly, but rises rapidly after 1939 to a peak between 1941 and 1945 of greater than 85%. The ERI observations in this period come principally from ICOADS decks 705-707, containing US Merchant Marine data and deck 245 containing

Figure 2. Fractional contribution to the monthly global-average SST from different measurement methods and platforms (1920-2006): buckets (dark grey), ERI and hull contact (mid grey), Unknown (pale grey) and buoys (cross-hatched). The global-average SST anomaly is effectively a weighted sum of all available observations. The fractional contribution for a given measurement method was calculated as the sum of weights for that observation type.

Table 5. Estimates of the ship minus drifting buoy relative SST bias (°C) 1998-2007, based on observation pairs made within 50km and up to 6 hours before local dawn. SD stands for standard deviation.

Ocean region	Mean bias	SD	SE	Match ups
Globe	0.12	0.85	0.01	21870
N Atlantic	0.13	0.86	0.01	10144
N Pacific	0.11	0.94	0.01	5364
S Atlantic	0.05	0.84	0.04	434
SE Pacific	0.21	0.93	0.10	88
SW Pacific	0.21	0.93	0.06	248
Trop. Atlantic	0.16	0.70	0.02	1200
Trop. W Pacific	0.12	0.79	0.03	866
Trop. E Pacific	0.16	0.89	0.05	325
Trop. Indian	0.23	0.86	0.04	565
Indian	0.13	0.87	0.03	1161
Southern Ocean	0.13	0.43	0.06	55

UK Royal Navy data. In late 1945, there is a sharp drop in the fraction of ERI observations which is compensated by an increase in the fraction of bucket observations. Between 1945 and 1950 the fractional contribution of bucket observations is around 70-80%. Around 1955, the bucket contribution drops to 40%. It increases again to a peak around 1965 after which there is a general decline. By 2005, bucket observations, which were by that time insulated, contribute no more than 3% to the global-average. Since the 1970s, more ships have switched to using ERI and Hull Contact sensors, but overall the contribution of ERI measurements to the global average has declined slightly. The period 1979-2005 also saw the proliferation of drifting and moored buoys. After an early peak in 1979 associated with the First GARP Global Experiment, the combined contribution of moored and drifting buoys rose quickly, reaching almost 70% in 2006.

3.4. Drifting and moored buoys

Drifting buoys and moored buoys are nowadays deployed in large numbers and in 2006 had a 70% weight in the global average (Figure 2). It has been noted a number of times (e.g. *Emery et al.* [2001]) that ships are biased warm relative to drifting buoys and that this relative bias has not changed significantly over the period 1989-2006 (*Reynolds et al.* [2010]). As the numbers of drifting buoys has increased over time and the number of ships has decreased, there is likely to be an artificial reduction of the trend in global average temperatures. Therefore, the difference between ships and drifters needs to be factored into the bias calculation.

A database of nearly coincident ship and buoy observations for the period 1998-2007 was created in which ship-buoy pairs were selected that lay within 50km of one another and on the same day. To avoid complications from diurnal heating, only observations taken close to local dawn were used. The average differences were calculated for each ocean basin, and for the globe. The average difference between ship and drifting buoy observations in the period 1998-2007 was 0.12°C, with ships being warmer than drifting buoys. This estimate is close to that of *Reynolds et al.* [2010] and lower than that made in *Kennedy et al.* [2011a] and likely reflects the different geographical areas sampled by the two methods. However, there were significant regional variations and the results are summarised in Table 5.

4. Bias Estimation

The method presented here uses the relative numbers of measurements made using the different observation methods together with estimates of the biases associated with each measurement type to calculate bias estimates for the gridded analysis that vary in space and time. Section 4.1 describes a general method for estimating the biases in time series of area-averaged data and Section 4.2 explains the specific implementation for HadSST3.

4.1. Basic method

Any ocean area - such as a grid box or ocean basin - can contain SST observations made using a variety of measurement methods. Time series of the fractional contribution, f, to the area average from each of these observation types were created. The fractional contribution is the number of observations taken using a particular method divided by the total number of observations in that area. Each observation type has a characteristic bias associated with it. Here bias means the difference from a notional 'true' SST. There are four principal sources of SST observations:

1. Ships using insulated, rubber or wooden buckets. Fractional contribution f_r, bias B_{tr}.

2. Ships using uninsulated, canvas buckets. Fractional contribution f_c, bias B_{tc}.

3. Ships using Engine Room Intake (ERI) thermometers and Hull Contact sensors. Fractional contribution f_e, bias E.

4. Drifting and moored buoys. Fractional contribution f_d, bias D.

Although the biases are defined as being relative to the 'true' SST - in this case the average SST within the upper few metres of the water column - it is rarely possible to state what that 'true' SST is. For many analyses this is not a problem because the variable of interest is the difference of the SST from the climatological average (typically 1961-1990). The challenge then is one of estimating the relative bias between the climatological average and the measurement. The "bucket corrections" produced by FP95 and R06 are of this kind. They removed the relative bias between bucket observations made in the period prior to 1941 and the average of all observations made in the climatology period (1961 to 1990). In the present analysis, the bucket corrections relative to 1961-1990 are referred to as B_r and B_c for rubber or wooden (insulated) buckets and canvas (uninsulated) buckets respectively. It is important to note that the bucket corrections, B_r and B_c, are not the same as the bucket measurement bias relative to the 'true' SST, B_{tr} and B_{tc}.

In order to calculate the bucket bias from the bucket corrections it is necessary to know E and D. If we assume these are known, then

$$B_r = B_{tr} - avg_{61-90}\left(f_r B_{tr} + f_c B_{tc} + f_e E + f_d D\right) \quad (1)$$

$$B_c = B_{tc} - avg_{61-90}\left(f_r B_{tr} + f_c B_{tc} + f_e E + f_d D\right) \quad (2)$$

$$B_{tr} - B_{tc} = B_r - B_c \quad (3)$$

The first two equations state that the bucket corrections are equal to the bucket biases minus the average bias in the climatology period. The third says that the difference between the insulated bucket bias and the uninsulated bucket bias is the same as the difference between the insulated bucket correction and the uninsulated bucket correction. It is assumed that the bias E is time varying because the general characteristics of the global fleet are liable to change over time and that the bias D is fixed for any given grid-box because the design of buoys changes little over time. Re-arranging the equations for B_r and B_c gives

$$B_{tr} = \frac{B_r\left(1-\bar{f}_c\right) + \bar{f}_c B_c + \overline{f_e E} + \bar{f}_d D}{\left(1-\bar{f}_r-\bar{f}_c\right)} \quad (4)$$

$$B_{tc} = \frac{B_c\left(1-\bar{f}_r\right) + \bar{f}_r B_r + \overline{f_e E} + \bar{f}_d D}{\left(1-\bar{f}_r-\bar{f}_c\right)}, \quad (5)$$

where a bar over a variable means 'take the 1961-1990 average' for that variable. The bias at any time can be written as:

$$B(t) = f_r(t) B_{tr} + f_c(t) B_{tc} + f_e(t) E(t) + f_d(t) D \quad (6)$$

The drifter bias, D, is still unknown, but can be found from the colocated average difference between drifting buoy

and ship observations, Δ_{ds}, during the period when these observations are plentiful: 1998-2006.

$$D - avg_{98-06}\left(g_r B_{tr} + g_e E\right) = \Delta_{ds} \quad (7)$$

Here, g_r and g_e are the fraction of ship observations (as opposed to the fraction of all observations f_r and f_e) made using rubber buckets and ERI thermometers respectively. It is assumed that no canvas bucket measurements were made after 1990. Substituting for D in Equation 4 gives,

$$B_{tr} = \frac{B_r\left(1-\bar{f}_c\right) + \bar{f}_c B_c + \overline{f_e E} + \bar{f}_d\left(\Delta_{ds} + \overline{g_e E}\right)}{\left(1-\bar{f}_r - \bar{f}_c - \bar{f}_d \bar{g}_r\right)} \quad (8)$$

$$B_{tc} = B_{tr} - B_r + B_c \quad (9)$$

where $\bar{g}_r$ and $\overline{g_e E}$ are the averages of the variables over the period 1998-2006.

In order to estimate the biases in the data using this method, it is necessary to know: E, Δ_{ds}, B_r, B_c and the various fractional contributions. With these it is possible to estimate B_{tr}, B_{tc} and hence $B(t)$. Where these estimates come from is described in the next section.

4.2. Implementation

The data were expressed as deviations from the 1961-1990 climatology and averaged onto a monthly five degree latitude and longitude grid. The fractional contributions of each observation type to each grid-box average were calculated. The analysis described in the previous section was applied to each grid box and each calendar month separately.

A variety of variables and parameters had to be estimated to apply the bias calculation method. Many of the variables were uncertain and therefore 100 realisations of the biases were generated by varying the variables within plausible ranges. For each realisation, new values were taken for each of the variables. The parameters that were varied are summarised in Table 6 and described in more detail below.

Realisations of the fields of the bucket corrections B_r and B_c were calculated as in R06 for wooden and canvas buckets respectively. In R06, fields were calculated for fast ships and slow ships separately to reflect the increase in ship speeds in the early record. The fast-ship fields were used to estimate the bucket biases after 1941. Before 1941, the fractions of fast and slow ships were taken from R06. *Kent and Kaplan* [2006] found that the fields of biases for modern buckets were similar to those for the FP95 wooden bucket corrections and so modern rubber buckets were assumed to have the same biases as wooden buckets. The R06 method includes uncertainties due to the changing speeds of ships, the proportions of wooden and canvas buckets and the ambient conditions.

In the R06 analysis, Night Marine Air Temperature-SST anomaly differences were minimised in the tropics in order to estimate the fraction of wooden and canvas buckets in use in 1850. The NMAT data set, MOHMAT, contained fewer data than HadSST2 because it was based on a smaller data base of observations. To test the effect of increasing the number of NMAT observations, an NMAT data set was made using data from ICOADS 2.0. The part of the tropics over which the SST-NMAT comparison was made had to be reduced because the new NMAT data in those regions did not fit the quality criteria described in *Bottomley et al.* [1990]. Data were excluded from the South China Sea and south of the equator in the Indian Ocean between $5°S$ and $0°S$, $55°E$ and $100°E$ and between $10°S$ and $5°S$, $65°E$ and $100°E$. 50 realisations were generated using the R06 bucket corrections and 50 were generated using the bucket corrections calculated using the new NMAT data set.

The ERI biases, E, were derived from the literature (Section 3.2) and used to create plausible realisations of $E(t)$. The best estimate from the literature gives an average value for E of 0.2°C with a likely range of ±0.1°C. Multiple realisations of an AR(1) time series, with a lag 1 correlation of 0.99, were created which had a mean of 0.2°C and a range chosen at random between 0 and 0.2°C. The realisations were used to generate estimates of the uncertainties that arise from our ignorance of the true evolution of $E(t)$. An AR(1) series gives variability on many time scales including decadal time scales. The same series of E(t) was used in all grid boxes except in the North Atlantic between 1970 and 1994 where the estimates of *Kent and Kaplan* [2006] were used in preference to the AR(1) series. *Kent and Kaplan* [2006] estimated ERI biases using a regression technique, based on a database of observations that is likely to overlap considerably with those in ICOADS2.5. The use of Kent and Kaplan (2006) for the North Atlantic caused small discontinuities at the edges of the region so the resulting fields of adjustments were smoothed in space using a box-filter with a width of 5 grid boxes.

Some observations could not be associated with a measurement method. These were randomly assigned to be either bucket or ERI measurements. The relative fractions were derived from a randomly-generated AR(1) time series as above but with range 0 to 1 and applied globally.

It is likely that many ships that are listed as using buckets actually used the ERI method (see end Section 3.2). To reflect the uncertainty arising from this, $30 \pm 10\%$ of bucket observations were reassigned as ERI observations. For example a grid box with 100% bucket observations was reassigned to have, say, 70% bucket and 30% ERI. A single number was chosen for each realisation of the adjustments that applied in all places after 1941. The exact timing of the switch from canvas to rubber buckets is also unknown. The available literature places the start of the transition between

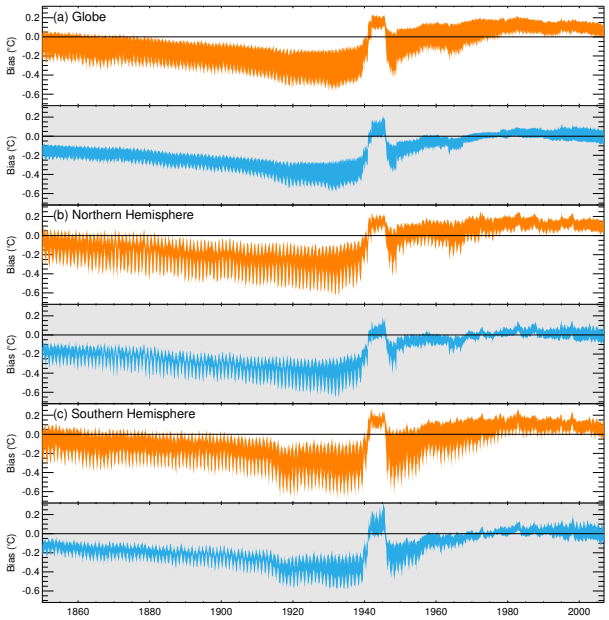

Figure 3. 100 realisations of the monthly biases in the (a) global average, (b) northern-hemisphere average and (c) southern-hemisphere average. The shaded areas represent the 99% uncertainty range of the estimates. The orange areas show the estimated bias relative to the true SST and the blue areas show the bias relative to the average bias for the period 1961-1990 ($b - \bar{b}$ from Equation 10).

Table 6. List of parameters varied for each realisation and permitted values

	Parameter	Brief description of variations
1	NMAT data set	50 realisations used MOHMAT, 50 used new NMAT dataset
2	Bucket corrections, $B_r$and B_c	Realisations generated according to R06. Fast ship corrections used after 1941.
3	$E(t)$	AR(1) series with lag correlation 0.99. Mean 0.2K and range drawn from a uniform distribution between 0 and 0.2. Same value used for all regions except the North Atlantic (see next row).
4	$E(t)$	In North Atlantic 1970-1994, E(t) generated using mean and standard errors from Kent and Kaplan (2006).
5	Unknown measurements	AR(1) series with lag correlation 0.99 and range between 0 and 1 used to assign time varying fraction of unknown to bucket. Remainder were set to ERI. Same value used at all places.
6	ERI recorded as bucket	$30 \pm 10\%$ of bucket observations reassigned as ERI. One value per realisation applied at all times and places after 1940
7	canvas to rubber	Linear switchover. Start point (all canvas) chosen randomly between 1954 and 1957. End point (all rubber) chosen randomly between 1970 and 1980.
8	Δ_{ds}	Values generated for each region described in Table 5. Randomly generated using mean and stated standard errors.

1954 and 1957 and the end between 1970 and 1980 (Section 3.1). The transition was assumed to be the same for all ships and countries and was assumed to be linear with the start and end dates chosen at random between these limits. Δ_{ds} was estimated from the data as described in Section 3.4. Because the value of Δ_{ds} is uncertain, a different value for each region was chosen at random from a normal distribution with mean and standard errors as given in Table 5.

Sometimes the denominator in Equation 8 was close to or equal to zero, for example, grid boxes where all the observations made during the climatology period were made by buckets ($f_r + f_c = 1$). In this case, B_{tr} would be infinite. This did not happen in practice, but there were cases where most observations were made by buckets. Grid boxes in which the calculated bucket biases exceeded 1°C were set to missing. This amounted to fewer than 1% of grid boxes out of a possible 1700 or so in any realisation. In some cases it was not possible to calculate a bias adjustment because there were no observations in that grid box during the climatology period or because there were no observations during the period 1998-2006. These were set to missing.

4.3. Time series of biases

The estimated biases for the global- and hemispheric-average SSTs are shown in Figure 3 (orange areas). They increase from between 0.0 and −0.2°C in the 1850s to between −0.1 and −0.6°C in 1935 as the proportion of both canvas buckets and fast ships increases. From 1935 to 1942, the proportion of ERI measurements increases (see also Figure 2) and the bias approaches zero. Between 1941 and 1945, the biases are between 0.05 and 0.2°C. The positive bias is a result of the large numbers of UK Navy and US ERI measurements in the ICOADS data base during the Second World War. In late 1945, the bias drops sharply and becomes negative again, reflecting an influx of data gathered by UK ships using canvas buckets. The bias then increases from 1946 to the early 1980s - becoming predominantly positive after 1975 - as insulated buckets were introduced and ERI measurements become more common. After 1980, the slow decrease in the bias is caused by the increase in the number of buoy observations.

If it is assumed that there is a true anomaly (A_{true}) and a true climatology (C) then

$$A_{true} = SST - C$$

and the observed anomalies from the observed climatology can be written as

$$A_{biased} = (SST + b) - \left(C + \bar{b}\right)$$

where b is the bias at the time of the anomaly and $\bar{b}$ is the average bias over the climatology period. They are related thus

$$A_{true} = A_{biased} - (b - \bar{b}) \quad (10)$$

The blue areas in Figure 3 show the global average of the quantity $b - \bar{b}$. The quantity $b - \bar{b}$ was subtracted from the anomalies in the gridded ICOADS 2.5 data to produce a data set of less-biased anomalies. This step has the ef-

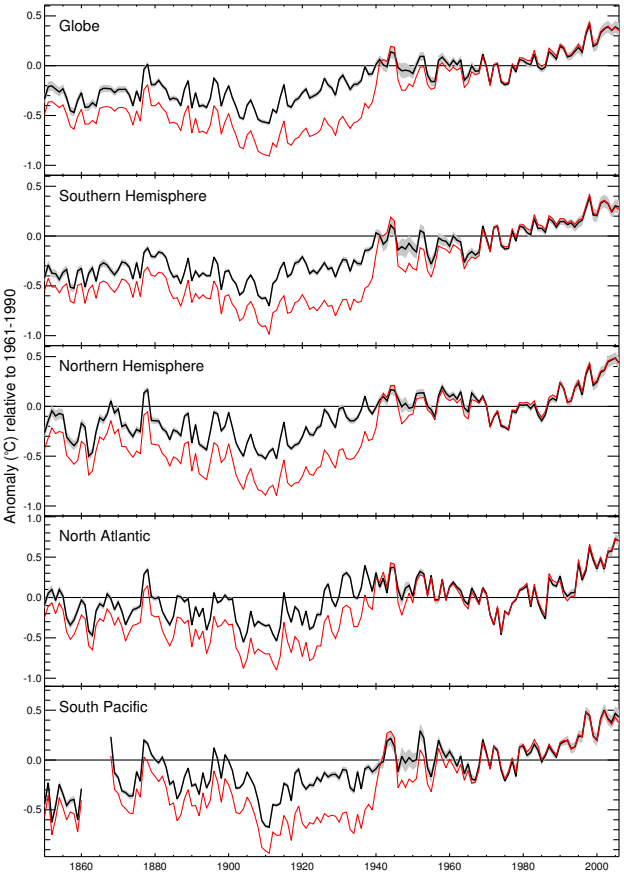

Figure 4. Regional annual-average SST anomalies 1850-2006 (relative to 1961-1990) for 100 realisations of HadSST3 (black line plus 2 standard deviations grey area) and unadjusted gridded data (red line).

fect of narrowing the uncertainty range in the climatology period (compare the orange and blue ranges in Figure 3) because the bias adjusted data must average to zero over this period. Missing bias estimates were then set equal to the adjustments calculated in R06, or to zero after 1941.

A number of tests of the bias adjustment method are provided in the next section. The method adjusts the SST in a very coarse way. It is intended only to estimate large scale differences - at global, hemispheric and ocean-basin scales - between different measurement methods. It does not solve the problem of individually biased ships and buoys, where those biases differ from the population average. The local biases associated with individual platforms are assumed to be an additional source of unresolvable random error and are included in the uncertainty calculations described in part 1 of the paper (*Kennedy et al.* [2011b]). These random errors will be more important for smaller, or more sparsely sampled areas. The resulting data set, combined with the uncertainties in part 1 of the paper (*Kennedy et al.* [2011b]), is HadSST3.

Figure 4 shows 100 realisations of regional-average SST anomalies from HadSST3. The adjusted time series are compared to SST anomalies from the unadjusted gridded ICOADS 2.5 data. The bias adjustments change the character of global SST variability in the second half of the 20th century. The most obvious difference between HadSST3 and the unadjusted data is in the period 1940 to 1975. HadSST3 is characterised by a slight cooling throughout this period, in contrast to the unadjusted data that shows warming after a sharp drop of around 0.3°C in late 1945. The second major difference can be seen in the differing trends over the period 1979-2006. The adjusted data around 1980 are cooler than the unadjusted data. Therefore, in going from the unadjusted data to HadSST3, the trend in observed global-average SST since 1979 has increased somewhat, although the increase falls within the uncertainty range. The uncertainty range in the 2000s is wider than in the climatology period, depsite the fact that there is a greater number of more reliable drifting buoy observations in the modern period. This is due, in part, to the step of setting the average bias adjustment to zero over the 1961-1990 period. The anomaly associated with a drifting buoy observation is therefore equal to the accurately measured buoy SST minus a more uncertain climatological value of the SST at that point.

It is also worth noting that the bias adjustments applied in the northern and southern hemispheres are quite different. Larger adjustments are applied in the southern hemisphere where the estimated proportion of bucket observations both immediately before and after the Second World War is higher.

4.4. Exploring the sensitivity of the bias adjustments

In order to test the reliability of the bias adjustments (MAIN_SET), a number of further tests were carried out. The first was to generate two alternative bias adjustment methods. In the first (SHIP_ONLY), only ship data were used with bias adjustments as described above. Here, the fraction of drifting buoy observations was always zero. In the second (BUOY_FIXED), the method was turned around so that the drifting buoys were the fixed point against which all the other estimates were compared (in the original method this role was played by the ERI biases). This was done by setting D to zero in Equations 4, 5 and 7 and deriving a formula for E. The equivalent formula for B_{tr} was then

$$B_{tr} = \frac{B_r(1-\bar{f}_c) + \bar{f}_c B_c - \frac{\bar{f}_e}{\bar{g}_e}\Delta_{ds}}{(1-\bar{f}_r-\bar{f}_c+\frac{\bar{f}_e\bar{g}_r}{\bar{g}_e})}$$

and E was

$$E = \frac{-(\bar{g}_r B_{tr} + \Delta_{ds})}{\bar{g}_e}.$$

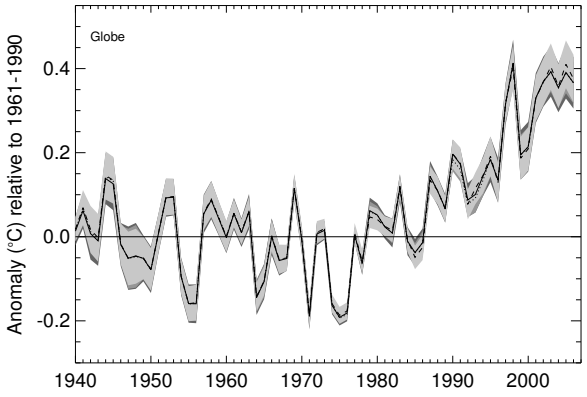

Figure 5. Global-average SST anomalies 1940-2006 for 100 realisations of each of the different bias adjustment methods. The solid line is the original method (MAIN_SET). The dashed line is from the method based only on ship data (SHIP_ONLY). The dotted line uses the alternative method based on drifting buoy data (BUOY_FIXED). The grey areas show the 2-sigma uncertainty ranges for the three data sets.

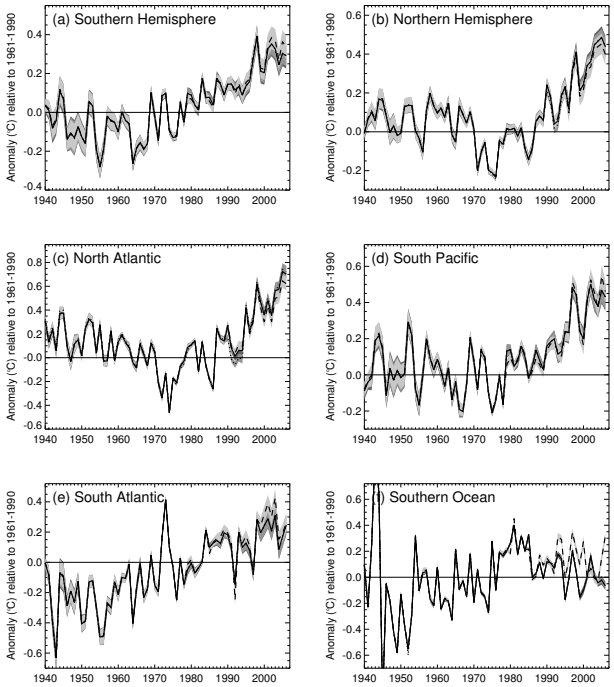

Figure 6. Area-average SST anomalies 1940-2006 for 100 realisations of the three different bias adjustment methods. The solid line is the original method (MAIN_SET). The dashed line is from the method based only on ship data (SHIP_ONLY). The dotted line uses the alternative method based on drifting buoy data (BUOY_FIXED). The grey areas show the 2-sigma uncertainty ranges for the three data sets.

Bias adjusted series using the three different methods are shown in Figures 5 and 6. The bias adjusted global and hemispheric averages agree well with one another and the differences are smaller than the uncertainty range. Agreement is also good in the South Pacific where observations are typically sparse. In the North Atlantic the series in which buoys were assumed to measure the true SST is cooler around 1990 than either the original method or the ship only series. This is because the ERI bias assumed in this method (BUOY_FIXED) takes a single value from 1850 to 2006 and therefore does not accurately reflect the required warming of the adjustments in the original method (MAIN_SET) caused by the cooling of ERI measurements seen in the *Kent and Kaplan* [2006] analysis. In other regions, discrepancies are larger. In the South Atlantic the ship only series is warmer than the the combined ship and buoy series with differences of around 0.1°C after 2000. The differences are not significant, but are noteworthy. In the Southern Ocean the ship-only series is significantly warmer than the two combined ship and buoy series, but this region is sparsely and infrequently observed and the random errors are therefore large (*Kennedy et al.* [2011b]).

The second test was performed by extracting individual observations that were identified as being more than 0.95 bucket or ERI measurements. The series of ERI and bucket observations were gridded separately and then adjusted using the estimated adjustments. The unadjusted (upper panel) and adjusted (lower panel) global and hemispheric series, based on 5 degree areas with both bucket and ERI data, are shown in Figures 7, 8 and 9 along with estimates from buoys. The adjustments bring the three estimates into closer agreement throughout the period 1945 to 2006. In the early 1990s the adjusted buoys are warmer than the ship data in the Southern Hemisphere. At this time, there are fewer buoy observations than in the later period, with patchier coverage, and drifter designs were still being refined. Also in the Southern Hemisphere, there is a divergence between the estimates in the late 1940s. This is due to the small number of coincident ERI and bucket observations in this region at this time. In fact during the two years 1948 and 1949 there were insufficient coincident observations to form an annual average. As before, discrepancies between the estimates are larger where fewer observations are used suggesting that the uncertainties of the bias adjustments are inadequate on their own to describe the full uncertainty in these regional estimates. In order to fully map the uncertainties at small scales, attention must be given to measurement and sampling errors described in part 1 of the paper (*Kennedy et al.* [2011b]).

Smith and Reynolds [2002], hereafter SR02, devised an alternative method for bias adjusting SST data to that of *Folland and Parker* [1995]. In order to assess the effectiveness of the bias adjustments in the present analysis, the SR02 analysis was repeated. The unadjusted and adjusted SST data were used to recreate Figure 4b of SR02, which showed a metric based on air-sea temperature differences. The results are shown in Figure 10. The global average difference is more consistent throughout the record for the adjusted data set, except in two periods: the Second World War, when there are known problems with the NMAT (*Bottomley et al.* [1990]) and in the periods 1900-1905 and 1975-1995 when measured SST anomalies were cooler than NMAT anomalies by around a tenth of a degree. The reason for this difference is not known. The time series of bucket and ERI measurements track each other reasonably closely during this period suggesting that the difference does not lie in one particular SST data source and there are no significant events in the metadata that coincide with this period.

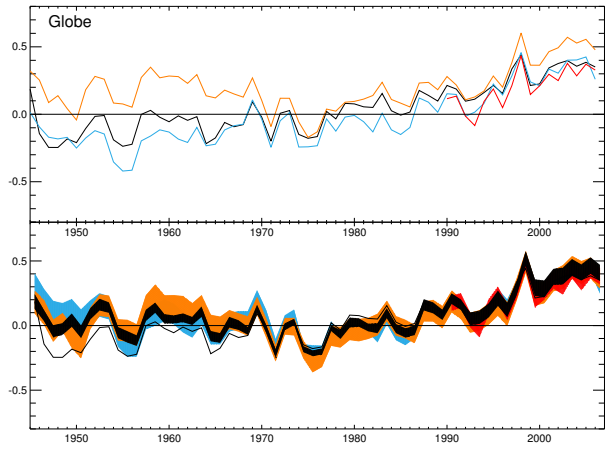

Figure 7. Global annual average sea surface temperature anomalies 1945-2006 from ERI measurements (orange), bucket measurements (blue), buoy measurements (red) and from all observations (black). Unadjusted series are shown in the upper panel and 100 realisations of the adjusted series are shown in the lower panel. The black line in the upper panel is duplicated in the lower panel for comparison. The ERI and bucket observations have been colocated and the buoy observations have been reduced to the common coverage of the bucket and ERI observations. At some times the buoy data have a lower coverage than the colocated ERI and bucket observations. The black lines showing averages based on all observations are also colocated with the ERI-only and bucket-only data.

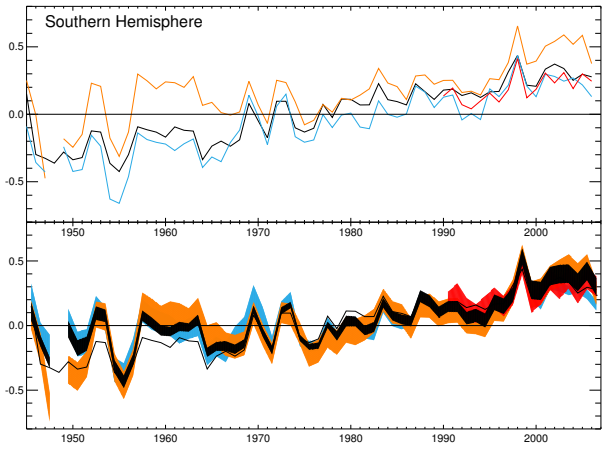

Figure 8. Southern hemisphere annual average sea-surface temperature anomalies 1945-2006 from ERI measurements (orange), bucket measurements (blue), buoy measurements (red) and from all observations (black). Unadjusted series are shown in the upper panel and 100 realisations of the adjusted series are shown in the lower panel. The black line in the upper panel is duplicated in the lower panel for comparison. The ERI and bucket observations have been colocated and the buoy observations have been reduced to the common coverage of the bucket and ERI observations. At some times the buoy data have a lower coverage than the colocated ERI and bucket observations. The black lines showing averages based on all observations are also colocated with the ERI-only and bucket-only data.

The difference between the NMAT and SST is largest in the Northern Hemisphere and larger in the Pacific than in the Atlantic (Figures 13 and 14). SR02 cautions against the interpretation of differences smaller than 0.1K in terms of biases in the data sets on the grounds that such differences might arise from natural variation in the air sea temperature difference. In the immediate post war period the difference is approximately constant for the adjusted SST data, but shows an upward trend in the unadjusted data, which SR02 had interpreted as perhaps being a residual, but insignificant, bias in their original analysis.

5. Key results

There are significant biases in the historical record of SST. Adjusting the data to account for them changes our understanding of the character of observed 20th century SST variability. Because of the incomplete metadata, and the difficulty of estimating biases in historical SST records, the uncertainties of the recent adjustments are relatively large, amounting to almost 0.1°C in the late 1940s and in the 2000s. The estimated uncertainty is consistent with the estimates of *Smith and Reynolds* [2005]. Measurement and sampling errors (derived in part 1, *Kennedy et al.* [2011b] are larger than in previous analyses of SST because they include the effects of correlated errors in the observations. Correlation between the measurement errors leads to an approximate two-fold increase in global- and hemispheric-average uncertainty. A time series of global-average, bias-adjusted SSTs with all uncertainty estimates combined is shown in Figure 11.

The uncertainty of global-average SST is largest in the early record and immediately following the Second World War. The reasons for the large uncertainties are in each case different. In the mid 19th century the largest components of the uncertainty at annual time scales are the measurement and sampling uncertainty and the coverage uncertainty because there were few observations made by a small global fleet. The bias uncertainties are relatively small because it was assumed that there was little variation in how measurements were made. By contrast, in the late 1940s and early 1950s, there is a good deal of uncertainty concerning how measurements were made. As a result the bias uncertainties are larger than the measurement and sampling uncertainties. After the 1960s bias uncertainties dominate the total and are by far the largest component of the uncertainty in the most recent data.

The relative sizes of bias and measurement uncertainties depend on the time scales and regions that are being considered. Measurement errors are correlated from one month to the next because the re-equipping of ships and the failure rate of drifting buoys are characterised by longer time scales. However, measurement errors are almost certainly uncorrelated on decadal timescales. Bias uncertainties, on the other hand, have long-term correlations. As longer and longer periods are considered, the measurement and sampling uncertainties become less important relative to the bias uncertainties, even at times when the measurement uncertainties of individual annual values are large. Therefore, uncertainties in long-term trends are likely to be dominated by uncertainties in the slowly varying biases introduced by changes in instrumentation and analysis methods.

Figures 4 and 11 show the uncertainty range, but it is not clear from the figures what the spread in trends exhibited by the realisations is because they combine both short-term and long-term variations. Figure 12 shows the ordinary least squares (OLS) trends in adjusted and unadjusted global-average SST for different periods all ending in 2006 and compares the trends to those in the previous Met Office Hadley Centre SST data set HadSST2 and those in the drifting buoy data. The adjustments have the effect of

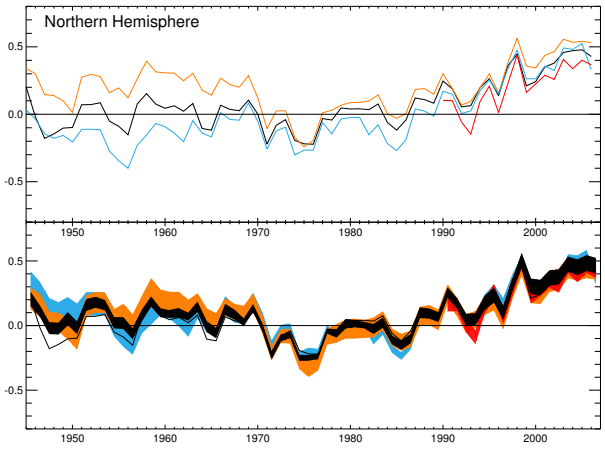

Figure 9. Northern hemisphere annual average sea surface temperature anomalies 1945-2006 from ERI measurements (orange), bucket measurements (blue), buoy measurements (red) and from all observations (black). Unadjusted series are shown in the upper panel and 100 realisations of the adjusted series are shown in the lower panel. The black line in the upper panel is duplicated in the lower panel for comparison. The ERI and bucket observations have been colocated and the buoy observations have been reduced to the common coverage of the bucket and ERI observations. At some times the buoy data have a lower coverage than the colocated ERI and bucket observations. The black lines showing averages based on all observations are also colocated with the ERI-only and bucket-only data.

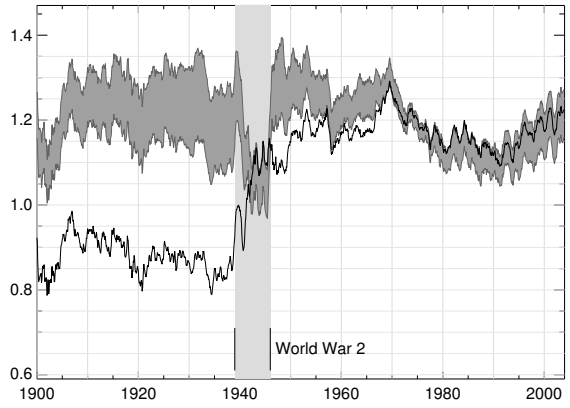

Figure 10. The time series show the annual average area-weighted regression coefficient of the monthly average air sea temperature difference on the climatological average difference. Values are scaled by the 60°S − 60°N average air-sea temperature difference to give values with units of degrees Celsius (cf Smith and Reynolds (2002), Fig. 4b). 100 realisations of the bias adjusted data are shown as the dark grey area and the unadjusted data are shown in black. The pale grey area shows the period during the Second World War when the NMAT observations are adjusted using DMAT observations.

reducing the trend from 1940 to 2006, but generally increase the trend from 1980 to 2006. In the latter case the effects are not significant given the wide uncertainty range. The trends in the adjusted and unadjusted series are, on average, lower than those from HadSST2 but are statistically indistinguishable except for start dates 1935-1955 and from 2003 onwards. Between 1995 and 2001, the trends in the unadjusted data lie at the lower end of the distribution of the trends in the adjusted series reflecting the rapid increase in the number of relatively-cold-biased buoy observations in the record at that time. The buoy data are shown from 1991, but 1996 was the first year in which more than one third of grid boxes were filled in every month. The coverage of buoy observations continued to increase throughout the period 1996-2006. The trends in the buoy data and the HadSST3 data are very similar after 1997 suggesting that the buoy data could be used alone to monitor global-average SST in the future.

At the longest time scales the uncertainties of the trends are much lower. This is partly due to the length of the period. However, one should bear in mind that the diagram does not include two sources of uncertainty. The first is the uncertainty arising from the construction of a global average at times when there are many empty grid boxes. The second is the structural uncertainty in both the bias adjustments and the analysis methods. Figures 13 and 14 suggest that the second component is important at longer time scales.

Figure 13 shows global and regional temperature series and OLS trends for HadSST3 and other marine data sets. Uncertainty of the trend estimation due to serial correlation of the residuals about the trend was not considered because the true variability was common to all data sets. Globally, HadSST3 shows a slight cooling from the late 1940s to the early 1970s followed by a rise in temperature. As the effect of the adjustments has been to remove a warm bias from the SST data around 1970, and a relative cool bias after 1990, the estimated rate of warming over the past 20-30 years has increased on average, although the effect is within the uncertainty range. The rate of warming seen in HadSST2 and other marine data sets over the period 1980-1999 lies below the median for the new HadSST3 data set. In general, however, the relative importance of bias uncertainties and structural uncertainties depends on the length of the period considered and the region.

Over the 20th century, the uncertainty in the global-average SST trend represented by 100 realisations of HadSST3 is smaller than the differences between other data sets that have not been adjusted for changes in instrumentation after 1941: ERSST v3 (*Smith et al.* [2008]), HadSST2 (*Rayner et al.* [2006]), COBE (*Ishii et al.* [2005]) and Kaplan SST (*Kaplan et al.* [1998]). However, it should be noted that these are not all simple gridded aggregates of SST observations. Some are reconstructions based on statistical infilling and the different analysis methods affect the trends represented by the data sets (see also *Rayner et al.* [2009]). In the last two decades of the 20th century, which are the most densely observed, the spread in bias uncertainty in global-average trends is larger than the differences between the data sets, suggesting that changes in measurement method are more important than analysis methods over this period. In many areas the rate of warming in HadSST3 is higher than in other data sets. Over middle length periods, such as 1940-1999 the rate of warming in HadSST3 is generally lower, due to the adjustments applied in the 1950s, which act to increase the estimated SST. Over this period the spread between data sets is approximately equal to the spread arising from bias uncertainties. However, it is noteworthy that

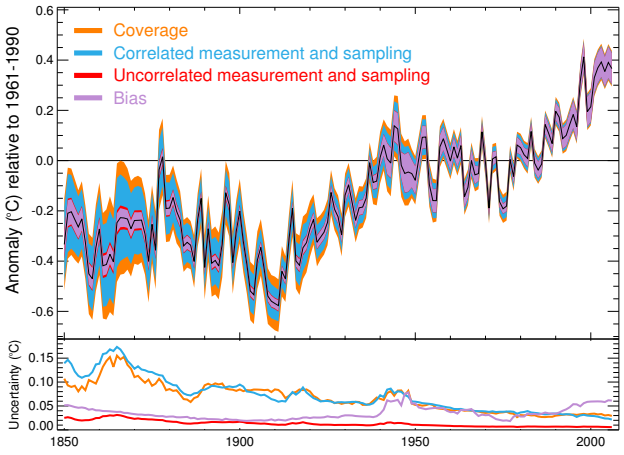

Figure 11. Global-average SST anomaly (relative to 1961-1990) showing the cumulative effect of adding different error components (coloured areas): median HadSST3 value (black) 2-sigma uncertainty arising from assumptions in bias adjustments (purple); measurement and sampling error, assuming these are uncorrelated between grid boxes (red); and the uncertainty including the inter-grid box correlations (blue) and the total uncertainty including all the above terms and the coverage uncertainty (orange). The lower panel shows the sizes of the individual components as 2-sigma uncertainties.

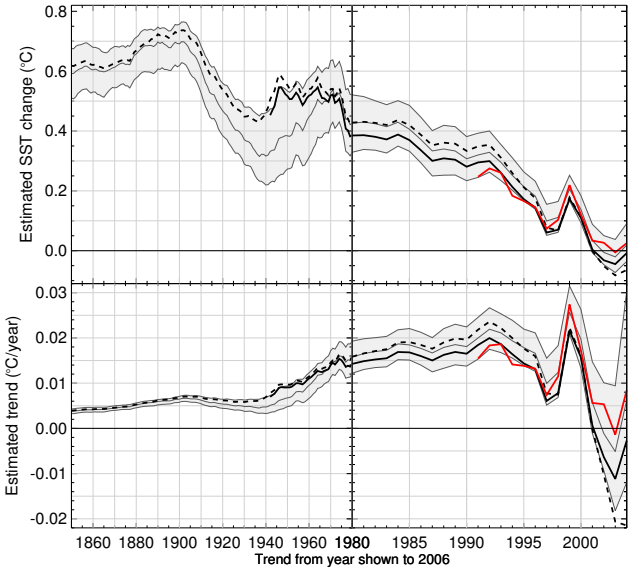

Figure 12. (upper panel) Estimated changes in annual global-average SST calculated by multiplying the Ordinary Least Squares trend (estimated over the period starting at the year shown on the x-axis and ending in 2006) by the length of the period. (lower panel) Estimated OLS trend. The left hand panels shows start dates between 1850 and 1980 and the right hand panels shows start dates between 1980 and 2006 on an expanded scale. The solid black line shows the gridded unadjusted ICOADS 2.5 data (from 1942). The dashed black line shows HadSST2. The grey area and lines show the median and range of the 100 realisations of HadSST3. The red line shows unadjusted data from buoys only. Note that the data are not colocated so there are large differences in coverage between buoy data and all data prior to around 1997.

that the spread of unadjusted data in the late 1940s covers only the very lowest estimates from the adjusted data.

The bias adjustments resulted in an average cooling of SST data during the Second World War. The sudden drop seen in HadSST2 and other marine data sets in late 1945 (*Thompson et al.* [2008]) is largely explained as an artifact of changing measurement method. However, the exact depth of the residual dip remains uncertain. It clearly depends on the actual SST variability, as well as the size of the biases in the measurements immediately before and after. Little additional literature relating to the Second World War period was uncovered during this study and it is possible that the methods used may have differed substantially from what has been assumed here. Night-time marine air temperature measurements made during the war are believed to have been made indoors owing to the dangers of carrying a light on deck to read the thermometer after dark (*Folland and Parker* [1995]). The result was a marked warm bias in the air temperature measurements. The same safety considerations might have meant that ERI measurements were made in preference to bucket measurements. If all measurements between 1941 and 1945 had been made by ERI, the true SST anomaly would be between 0.1 and 0.2°C lower than is allowed for here. If all measurements during the period were made using buckets, which seems unlikely, the true SST might be as much as half a degree warmer. Observation times also changed during the war years from 6-hourly measurements to a thrice-daily regime with many more measurements made at 8am and 8pm local time with another measurement at noon. The net effect of these changes is difficult to assess, but such changes might have led to a coincident step change in many marine variables at the end of the war.

6. Remaining Issues

The understanding of uncertainties associated with *in situ* SST measurements can be improved by increasing the number of observations stored in digital repositories such as ICOADS. The exact amount of undigitised data is unknown but some estimates suggest that the amount of undigitised data from before the second World War is larger than the amount that has already been digitised (Rob Allan personal comm.). As well as digitising observations from log books, metadata are also being systematically scanned and stored online. Additional metadata can inform the assumptions made in estimating data biases and allow a more accurate assessment of the uncertainties. For a much more thorough background to these efforts see *Brohan et al.* [2009]. Of particular interest is the period of the Second World War. It is not clear exactly how measurements recorded in the Met Logs of UK ships (as opposed to the deck logs) were made during this period. An examination of a small number of these logs suggests that they contain a combination of bucket and engine room measurements, sometimes recorded at the same time by the same ship, suggesting that metadata are available that have not been digitised. Such information, gathered on a larger scale could help reduce uncertainties in this key period.

The differences between marine air temperatures and sea-surface temperatures are also of interest, particularly the large difference observed from 1975 to 1995 (Figure 13). An updated data set of marine air temperatures based on ICOADS 2.5 is currently under development and should help to understand those differences. It is also worth noting

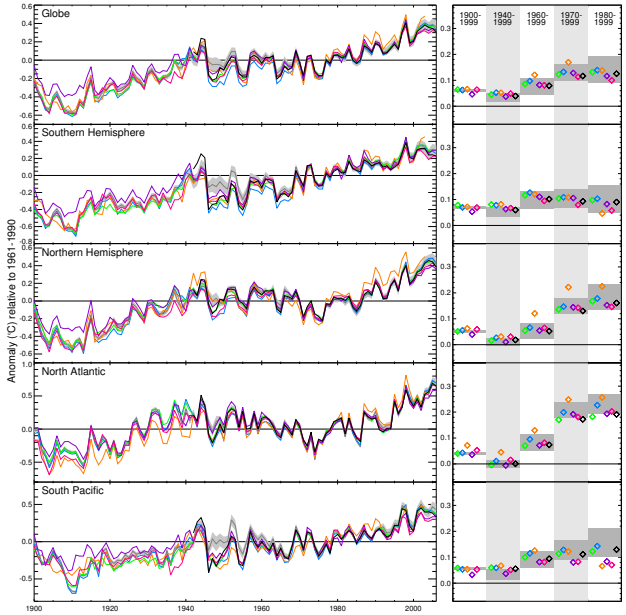

Figure 13. Global- and regional-average marine temperature anomalies 1900-2006 (relative to 1961-1990) from the ERSSTv3 analysis (green *Smith et al.* [2008]), HadSST2 (blue *Rayner et al.* [2006]), MOHMAT (orange *Rayner et al.* [2003]), Kaplan SST (purple *Kaplan et al.* [1998]), COBE (pink *Ishii et al.* [2005]), ICOADS summaries (1941-2006 only, black *Worley et al.* [2005]) and HadSST3 (grey area). All data sets have been reduced to have the same coverage as HadSST3. The panels on the right show trends (°C/decade) for the periods: 1900-1999, 1940-1999, 1960-1999, 1970-1999 and 1980-1999 from the same data sets.

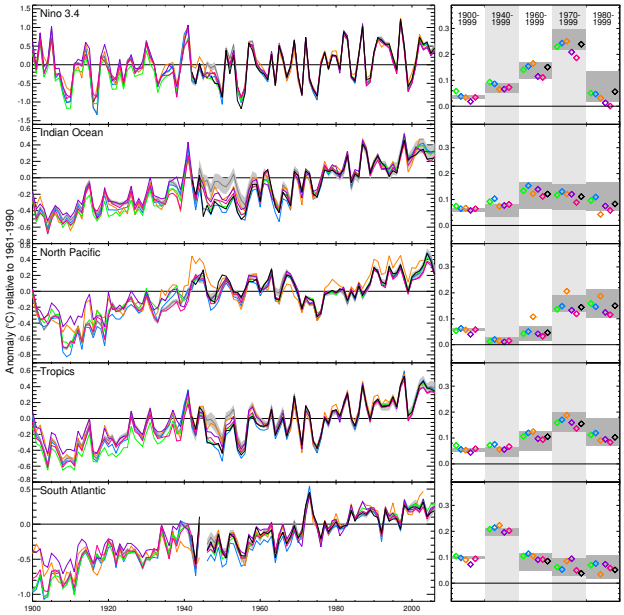

Figure 14. Regional-average marine temperature anomalies 1900-2006 (relative to 1961-1990) from the ERSSTv3 analysis (green *Smith et al.* [2008]), HadSST2 (blue *Rayner et al.* [2006]), MOHMAT (orange *Rayner et al.* [2003]), Kaplan SST (purple *Kaplan et al.* [1998]), COBE (pink *Ishii et al.* [2005]), ICOADS summaries (1941-2006 only, black *Worley et al.* [2005]) and HadSST3 (grey area). All data sets have been reduced to have the same coverage as HadSST3. The panels on the right show trends (°C/decade) for the periods: 1900-1999, 1940-1999, 1960-1999, 1970-1999 and 1980-1999 from the same data sets.

here that we have not allowed for the relatively high speed of modern ships in estimating the insulated bucket adjustments, nor for differences between the many sub-varieties of buckets used historically.

In the nineteenth and early twentieth century, there is an interdependence between the adjustments made to the NMAT and SST. The tropical NMAT are used to constrain the early bucket adjustments (*Rayner et al.* [2006]) and the SSTs are used to adjust for biases in much of the Atlantic NMAT data caused by poor instrument exposure (*Bottomley et al.* [1990]). The SST biases are constrained by the NMAT in the early record leading to a relatively narrow uncertainty range. However, there is value in assessing the uncertainties in the SST record without recourse to the NMAT data. In R06, the fractions of wooden and canvas buckets prior to 1920 were set using the comparison with NMAT. The fractions could instead be varied through some plausible range. The two extreme cases would be that all buckets were considered to be canvas buckets from 1850 to 1940, or that before 1920 all buckets were wooden. The first case would increase SSTs in 1850 by around 0.2C with the increase dropping linearly to zero by 1920. The latter case, which is less likely, would lead to a decrease in estimated global average SST prior to 1920 of between 0.1 in 1850 and 0.3C in 1919. Further work would be needed to refine these limits.

Although independence between SST and NMAT data sets is desirable for some applications it is not desirable for all. The large increase in uncertainty incurred by maintaining strict independence indicates that by considering multiple variables together rather than singly it is possible to reduce significantly uncertainty in estimates of past climate.

Other issues have been highlighted in the Ocean Obs 09 white paper by *Rayner et al.* [2009] including the problem of missing or non-unique call signs, exacerbated in recent years by the decision of several countries to deliberately anonymise their meteorological reports. Because call sign information was unavailable in December 2007, HadSST3 was only produced from 1850 to 2006. In order to bring this analysis up to date it will be necessary to either tackle this problem directly, or find some way around it. The possibility of gaining privileged access to some of the call signs is being pursued, but would have the unfortunate effect of basing part of the analysis on data that are not publicly available. Alternatives might be to form the global average from buoy observations only, or to use a fixed bias field for ship observations after 2006.

Finally, the estimates of biases and other uncertaintes presented here should not be interpreted as providing a comprehensive estimate of uncertainty in historical sea-surface temperature measurements. They are simply a first estimate. Where multiple analyses of the biases in other climatological variables have been produced, for example tropospheric temperatures (*Thorne et al.* [2011]) and ocean heat content (*Palmer et al.* [2009]), the resulting spread in the estimates of key parameters such as the long-term trend has typically been significantly larger than initial estimates of the uncertainty suggested. Until multiple, independent estimates of SST biases exist, a significant contribution to the total uncertainty will remain unexplored. This remains a key weakness of historical SST analysis.

Acknowledgments.

The authors were supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). The ICOADS data for this study are from the Research Data Archive (RDA) which is maintained by the Computational and Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation (NSF). The original data are available from the RDA (http://dss.ucar.edu) in dataset number ds540.0. The WMO publication 47 metadata from 1955 to 1994 were downloaded from the ICOADS web page at http://icoads.noaa.gov/metadata/wmo47/. More recent updates to WMO publication 47 were provided by the WMO. SST data products were downloaded from the SST and sea ice intercomparison site hosted at the GHRSST Long Term Stewardship and Reanalysis Facility (http://ghrsst.nodc.noaa.gov/intercomp.html). Scott Woodruff and Eric Freeman provided observer instructions for US ships and other relevant literature. Volker Weidner provided documentation for German observation practices. Professor Kimio Hanawa provided information on Japanese observing methods. Frits Koek provided information on the Dutch data. Helpful comments were provided by Chris Folland. The authors would also like to thank the three anonymous reviewers for their insightful and constructive comments, which improved the manuscript.

References

Amot, A. (1954), Measurements of sea surface temperature for meteorological purposes. results of observations from ocean weather station m, *Meteorologiske Annaler*, *4*(1), 1–11.

Bottomley, M., C. Folland, J. Hsiung, R. Newell, and D. Parker (1990), *Global Ocean Surface Temperature Atlas*, HMSO.

Brohan, P., R. Allan, J. Freeman, A. Waple, D. Wheeler, C. Wilkinson, and S. Woodruff (2009), Marine observations of old weather, *Bulletin of the American Meteorological Society*, *90*(2), 219–230, doi:10.1175/2008BAMS2522.1.

Brooks, C. (1926), Observing water-surface temperatures at sea, *Monthly Weather Review*, *54*(6), 241–253, doi:10.1175/1520-0493(1926)54¡241:OWTAS¿2.0.CO;2.

Brooks, C. (1928), Reliability of different methods of taking sea-surface temperature measurements, *J. Washington Acad. Sci.*, *18*, 525–545.

Bureau, U. W. (1925), Instructions to marine meteorological observers. w.b. no. 866, Circular-M marine division 4th ed.), *Government printing office*, p. 99.

Bureau, U. W. (1929), Instructions to marine meteorological observers. w.b. no. 991, Circular-M marine division 5th ed.), *Government printing office*, p. 80.

Bureau, U. W. (1938), Instructions to marine meteorological observers. w.b. no. 1221, Circular-M marine division 6th ed.), *Government printing office*, p. 120.

Bureau, U. W. (1941), Instructions to marine meteorological observers. w.b. no. 1221, Circular-M marine division 7th ed.), *Government printing office*, p. 114.

Collins, C., L. Giovando, and K. Abbott-Smith (1975), Comparison of Canadian and Japanese merchant-ship observations of sea-surface temperature in the vicinity of present ocean station P 1927-33, *Can. J. Fish. Aquat. Sci.*, *32*(2), 253–258, doi:10.1139/f75-023.

Elms, J., S. Woodruff, S. Worley, and C. Hanson (1993), Digitizing historical records for the comprehensive ocean-atmosphere data set (COADS), *Earth System Monitor*, *4*(2), 4–10.

Emery, W., D. Baldwin, P. Schlüssel, and R. Reynolds (2001), Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements, *Journal of Geophysical Research*, *106*(C2), 2387–2405, doi:10.1029/2000JC000246.

Folland, C., and D. Parker (1995), Correction of instrumental biases in historical sea surface temperature data, *Quarterly Journal of the Royal Meteorological Society*, *121*(522), 319–367, doi:10.1002/qj.49712152206.

Folland, C., R. Reynolds, M. Gordon, and D. Parker (1993), A study of six operational sea surface temperature analyses, *Journal of Climate*, *6*(1), 96–113, doi:10.1175/1520-0442(1993)006¡0096:ASOSOS¿2.0.CO;2.

Heiskell, H. (1908), Instructions to the marine meteorological observers of the US weather bureau. w.b. no. 397 (Circular-M, 2nd ed.), *Government printing office*, p. 48.

Heiskell, H. (1910), Instructions to the marine meteorological observers of the US weather bureau. w.b. no. 444 (Circular-M, 3rd ed.), *Government printing office*, p. 68.

HMSO (1956), *Handbook of meteorological instruments part 1*, 404–405 pp., HMSO.

HMSO (1963), *Marine Observer's Handbook 8th edition MO 522*, HMSO.

HMSO (1969), *Marine Observer's Handbook 9th edition MO 522*, HMSO.

Horiguchi, Y. (1940), The mean air temperature, cloudiness and sea surface temperature of the north pacific ocean and the neighbouring seas for the year 1939, *Kobe imperial marine observatory memoirs*.

Ishii, M., A. Shouji, S. Sugimoto, and T. Matsumoto (2005), Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection, *Int. J. Climatol.*, *25*(7), 865–879, doi:10.1002/joc.1169.

James, R., and P. Fox (1972), Comparative sea surface temperature measurements in WMO reports on marine science affairs, rep 5, *Tech. Rep. 336*, WMO.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan (1998), Analyses of global sea surface temperature 1856-1991, *Journal of Geophysical Research*, *103*(C9), 18,567–18,589, doi:10.1029/97JC01736.

Kennedy, J., R. Smith, and N. Rayner (2011a), Using AATSR data to assess the quality of in situ sea surface temperature observations for climate studies, *Remote Sensing of Environment*.

Kennedy, J., N. Rayner, R. Smith, M. Saunby, and D. Parker (2011b), Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, part 1: measurement and sampling uncertainties, *JGR Atmospheres*.

Kent, E., and A. Kaplan (2006), Toward estimating climatic trends in SST. Part III: Systematic biases, *Journal of Atmospheric and Oceanic Technology*, *23*(3), 487–500, doi:10.1175/JTECH1845.1.

Kent, E., and P. Taylor (2006), Toward estimating climatic trends in SST. Part I: methods of measurement, *Journal of Atmospheric and Oceanic Technology*, *23*(3), 464–475, doi:10.1175/JTECH1843.1.

Kent, E., P. Taylor, B. Truscott, and J. Hopkins (1993), The accuracy of Voluntary Observing Ships' meteorological observations - results of the VSOP-NA, *Journal of Atmospheric and Oceanic Technology*, *10*(4), 591–608, doi:10.1175/1520-0426(1993)010¡0591:TAOVOS¿2.0.CO;2.

Kent, E., S. Woodruff, and D. Berry (2007), Metadata from WMO publication no. 47 and an assessment of voluntary observing ship observation heights in ICOADS, *Journal of Atmospheric and Oceanic Technology*, *24*(2), 214–234, doi:10.1175/JTECH1949.1.

Kent, E., J. Kennedy, D. Berry, and R. Smith (2010), Effects of instrumentation changes on sea surface temperature measured in situ, *Wiley Interdisciplinary Reviews: Climate Change*, *1*(5), doi:10.1002/wcc.55.

Kirk, T., and A. Gordon (1952), Comparison of intake and bucket methods for measuring sea temperature, *Marine Observer*, *22*, 33–39.

Knudsen, J. (1966), An experiment in measuring the sea surfaec temperature for synoptic purposes, *Tech. Rep. 12*, Det. Norske Meteor. Inst.

Lumby, J. (1927), The surface sampler, an apparatus for the collection of samples from the sea surface from ships in motion. with a note on surface temperature observations, *J. Cons. Perm. Int. Explor. Mer.*, *2*, 332–342.

Okada, T. (1922), On the surface temperature of the japan sea, *Kobe imperial marine observatory memoirs*, *1*, 66–83.

Okada, T. (1927), The mean atmospheric pressure, cloudiness and sea surface temperature of the north pacific ocean and the neighbouring seas for the year 1926, *Kobe imperial marine observatory memoirs*.

Okada, T. (1929), The mean atmospheric pressure, cloudiness and sea surface temperature of the north pacific ocean and the neighbouring seas for the lustrum 1921 to 1925, *Kobe imperial marine observatory memoirs*.

Okada, T. (1932), The mean atmospheric pressure, cloudiness and sea surface temperature of the north pacific ocean and the neighbouring seas for the lustrum 1926 to 1930, *Kobe imperial marine observatory memoirs*.

Page, J. (1906), Instructions to the marine meteorological observers of the US weather bureau (Circular-M), *Government printing office*, p. 46.

Palmer, M., J. Antonov, P. Barker, N. Bindoff, T. Boyer, M. Carson, C. Domingues, S. Gille, P. Gleckler, S. Good, V. Gouretski, S. Guinehut, K. Haines, D. Harrison, M. Ishii, G. Johnson, S. Levitus, S. Lozier, J. Lyman, A. Meijers, K. von Schuckmann, D. Smith, S. Wijffels, and J. Willis (2009), Future observations for monitoring global ocean heat content, in *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 2), Venice, Italy, 21-25 September 2009*, edited by J. Hall, D. Harrison, and D. Stammer, ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.68.

Rayner, N., D. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Journal of Geophysical Research*, *108*(D14), 4407, doi:10.1029/2002JD002670.

Rayner, N., P. Brohan, D. Parker, C. Folland, J. Kennedy, M. Vanicek, T. Ansell, and S. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set, *Journal of Climate*, *19*(3), 446–469, doi:10.1175/JCLI3637.1.

Rayner, N., A. Kaplan, E. Kent, R. Reynolds, P. Brohan, K. Casey, J. Kennedy, S. Woodruff, T. Smith, C. Donlon, L. Breivik, S. Eastwood, M. Ishii, and T. Brandon (2009), Evaluating climate variability and change from modern and historical SST observations, in *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 2), Venice, Italy, 21-25 September 2009*, edited by J. Hall, D. Harrison, and D. Stammer, ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.71.

Reynolds, R., C. Gentemann, and G. Corlett (2010), Evaluation of AATSR and TMI satellite SST data, *Journal of Climate*, *23*(1), 152–165, doi:10.1175/2009JCLI3252.1.

Roll, H. (1951), Water temperature measurements on deck and in the engine room, *Ann. Meteor.*, *4*, 439–443.

Saur, J. (1963), A study of the quality of sea water temperatures reported in the logs of ships' weather observations, *Journal of Applied Meteorology*, *2*(3), 417–425, doi:10.1175/1520-0450(1963)002¡0417:ASOTQO¿2.0.CO;2.

Smith, T., and R. Reynolds (2002), Bias corrections for historical sea surface temperatures based on marine air temperatures, *Journal of Climate*, *15*(1), 73–87, doi:10.1175/1520-0442(2002)015¡0073:BCFHSS¿2.0.CO;2.

Smith, T., and R. Reynolds (2005), A global merged land air and sea surface temperature reconstruction based on historical observations (1880-1997), *Journal of Climate*, *18*(12), 2021–2036, doi:10.1002/wcc.55.

Smith, T., R. Reynolds, T. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), *Journal of Climate*, *21*(10), 2283–2296, doi:10.1175/2007JCLI2100.1.

Tabata, S. (1978a), On the accuracy of sea-surface temperatures and salinities observed in the Northeast Pacific Ocean, *Atmosphere Ocean*, *16*(3), 237–247.

Tabata, S. (1978b), Comparison of observations of sea-surface temperatures at ocean staion P and NOAA buoy stations and those made by merchant ships travelling in their vicinities in the northeast Pacific ocean, *Journal of Applied Meteorology*, *17*(3), 374–385, doi:10.1175/1520-0450(1978)017¡0374:COOOSS¿2.0.CO;2.

Tauber, G. (1969), The comparative measurements of sea surface temperature in the USSR, *Tech. Rep. 103*, WMO.

Thompson, D., J. Kennedy, J. Wallace, and P. Jones (2008), A large discontinuity in mid 20th century global-mean surface temperatures, *Nature*, *453*, 646–649, doi:10.1038/nature06982.

Thorne, P., J. Lanzante, T. Peterson, D. Seidel, and K. Shine (2011), Tropospheric temperature trends: history of an ongoing controversy, *Wiley Interdisciplinary Reviews: Climate Change*, *2*(1), 66–88, doi:10.1002/wcc.80.

Tsukada, K. (1927), On the mean atmospheric pressure, cloudiness and sea surface temperature of the north pacific ocean, *Kobe imperial marine observatory memoirs*, *2*, 163–201.

Wahl, E. (1948), Water temperature measurements on deck and in the engine room, *Ann. Meteor.*, *1*(7).

Walden, H. (1966), On water temperature measurements aboard merchant vessels (in German), *Ocean Dynamics*, *19*, 21–28, doi:10.1007/BF02321345.

WMO (1954), Technical note no 2 methods of observation at sea: part 1 - sea surface temperature No 26 TP 8.

Woodruff, S., S. Worley, S. Lubker, Z. Ji, J. Freeman, D. Berry, P. Brohan, E. Kent, R. Reynolds, S. Smith, and C. Wilkinson (2010), ICOADS release 2.5: extensions and enhancements to the surface marine meteorological archive, *International Journal of Climatology*, doi:doi:10.1002/joc.2103.

Worley, S., S. Woodruff, R. Reynolds, S. Lubker, and N. Lott (2005), ICOADS Release 2.1 data and products, *International Journal of Climatology*, *25*, 823–842, doi:10.1002/joc.1166.

J. J. Kennedy, Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK. (john.kennedy@metoffice.gov.uk)

D. E. Parker, Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK.

N. A. Rayner, Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK.

M. Saunby, Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK.

R. O. Smith, Ocean Physics Group, Department of Marine Science, University of Otago, Dunedin, New Zealand