Reassessing biases and other uncertainties in sea-surface temperature observations measured *in situ* since 1850, part 1: measurement and sampling uncertainties

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

Abstract. New estimates of measurement and sampling uncertainties of gridded *in situ* sea-surface temperature anomalies are calculated for 1850 to 2006. The measurement uncertainties account for correlations between errors in observations made by the same ship or buoy due, for example, to miscalibration of the thermometer. Correlations between the errors increase the estimated uncertainties on grid-box averages. In grid boxes where there are many observations from only a few ships or drifting buoys, this increase can be large. The correlations also increase uncertainties of regional, hemispheric and global averages above and beyond the increase arising solely from the inflation of the grid-box uncertainties. This is due to correlations in the errors between grid boxes visited by the same ship or drifting buoy. At times when reliable estimates can be made, the uncertainties in global-average, southern-hemisphere and tropical sea-surface temperature anomalies are between two and three times as large as when calculated assuming the errors are uncorrelated. Uncertainties of northern hemisphere averages approximately double. A new estimate is also made of sampling uncertainties. They are largest in regions of high sea-surface temperature variability such as the western boundary currents and along the northern boundary of the Southern Ocean. The sampling uncertainties are generally smaller in the tropics and in the ocean gyres.

1. Introduction

In order to understand changes in sea-surface temperature (SST) at global and regional levels, it is important to quantify the expected uncertainties in the observations. A number of studies have attempted to quantify the measurement errors in observations made by ships, drifting buoys and moored buoys. Typically, studies have considered two separate, but related, problems. One is the problem of estimating uncertainties associated with random measurement errors, which are assumed to be uncorrelated from one observation to the next (for example *Emery et al.* [2001], *Kent and Challenor* [2006], *Rayner et al.* [2006]). The second is that of identifying biases in the data due to changes in the way that measurements were taken. For example, many early SST measurements were made by collecting water samples in poorly insulated buckets. The buckets lost heat as they were hauled to the deck thus introducing a persistent cold bias into records of SST (e.g. *Folland and Parker* [1995], *Smith and Reynolds* [2002]). Part 2 of this paper (*Kennedy et al.* [2011a]) deals with estimating biases and their uncertainties on long space and time scales due to changing measurement methods.

However, the two types of errors cannot be separated so easily. Although measurements from ships exhibit characteristic average biases arising from the methods used to make the measurements, no two ships or buoys are identical. Therefore, one would expect biases that vary from ship to ship, or from drifting buoy to drifting buoy. By comparing SST observations to SST fields output from the Met Office Numerical Weather Prediction model, *Kent and Berry* [2008] estimated the standard deviations of these micro-biases, which they referred to as inter-platform errors to distinguish them from actually-random intra-platform errors. They found significant inter- and intra-platform errors in observations from ships, drifting buoys and moored buoys. *Kennedy et al.* [2011b] found similar errors in comparisons of *in situ* observations with measurements made by the Along Track Scanning Radiometer. The inter-platform and intra-platform designations can be confusing so here we refer to micro-bias errors and random measurement errors instead.

In addition to measurement uncertainties, gridded data sets of SST also incur an additional uncertainty from estimating area averages from a finite number of observations. *Rayner et al.* [2006] estimated combined measurement and sampling uncertainties from the SST data, but did not explicitly estimate the sampling uncertainty alone. Sampling uncertainties for land air temperatures have been estimated by *Jones et al.* [1997], *Brohan et al.* [2006] and *Shen et al.* [2007]. The formalism used in these studies depends on the average correlations between points within a grid box. The situation is slightly more complicated for marine data because the observations are randomly placed in time as well as in space. *Morrissey and Greene* [2009] extended the average correlation concept to marine data, which is the starting point for this analysis.

In the present paper, the estimates of micro-bias and random measurement uncertainties are used to estimate the uncertainties of gridded SST fields and on global and regional average SSTs for the period 1850 to 2006. Section 2 briefly describes the data source used in the analysis. Section 3 details the theoretical basis for the error model used in the subsequent sections. Section 4 deals with the practical problem of estimating uncertainties particularly when it is not possible to identify individual ships because the ships call sign, or ship name is not contained in the meteorological reports. Section 5 describes the results. In part 2 of the paper *Kennedy et al.* [2011a] a bias adjusted version of the

[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

Met Office Hadley Centre SST data set is described. The data set and uncertainty estimates together constitute version 3 of that data set. HadSST3 runs from 1850 to 2006 and includes bias adjustments and more comprehensive error estimates.

2. Data

The sea-surface temperature data for 1850 to 2006 come from version 2.5 of the International Comprehensive Ocean Atmosphere Data Set (ICOADS, *Woodruff et al.* [2010]). ICOADS comprises meteorological measurements originating from ships, oceanographic stations, moored buoys, drifting buoys and research vessels. SST data from ICOADS were quality controlled and processed according to the method detailed in *Rayner et al.* [2006] to produce monthly 5° latitude × 5°longitude grids.

There are three principal types of platform measuring SST *in situ*: ships, drifting buoys and moored buoys. Ships and buoys are identified by a unique call sign, or other identifier. For many historical reports, however, this information is absent. Where present, call sign information is recorded in ICOADS metadata and was recorded in Global Telecommunication System (GTS) reports until December 2007. After this date, the callsign information was removed from some GTS feeds and encrypted on others owing to concerns about ship security. Due to the lack of public call sign information it was not possible to complete this uncertainty analysis after 2006.

Most of the ship-based data in ICOADS 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 for their use. Although countries made an effort to standardise equipment and instructions within their fleets at any one time, 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 on the World Meteorological Organisation's VOS fleet list. Numbers have declined since, with fewer than 4000 ships remaining on the list today (http://www.bom.gov.au/jcomm/vos/).

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 (*Lumpkin and Pazos* [2007]). 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.

The sampling characteristics of these three platform types are quite distinctive. Ships travel between ports making regular observations so the observations from a single ship can provide a representative sample for a large area. However, most ship observations are made in the standard shipping lanes so the coverage obtained in this way can be limited. Drifting buoys, as their names suggests, drift. How far they drift depends on the prevailing currents, but during a month they do not often travel far. They typically take hourly observations and consequently provide dense sampling of a limited area in any given month. Drifters are generally deployed to provide a quasi-uniform coverage of the oceans. Moored buoys take regular measurements at a fixed point, typically in coastal areas, but there are also a number of tropical moorings in the open oceans.

3. Error model - theory

Consider an SST measurement, O_{ij}, taken by ship i at point j that has been converted to an anomaly from a reference climatology. This observation is a combination of the true SST anomaly at that point, T_{ij} , a random measurement error, M_{ij}, that is different for every observation and a constant offset - the micro-bias error - for measurements from that ship, B_i. Here "ship" is used to refer to any single entity - be it ship, drifting buoy, or moored buoy - that takes SST measurements. O_{ij} can thus be written as:

$$O_{ij} = T_{ij} + M_{ij} + B_i \tag{1}$$

M_{ij} has mean zero and standard deviation σ_{m_i} and is different at each point j. B_i is drawn for each ship, i, from a sample with mean zero and standard deviation σ_{b_i}. A mean of zero assumes that the bias adjustments discussed in Part 2 of this paper have adjusted successfully for the mean bias for each measurement type. The indices i and j can be used to keep track of observations in a single grid box. In this instance $i = 1, 2, ..., m;\ j = 1, 2, ..., n_i$. i.e. there are m ships and ship i takes n_i observations. The total number of observations is n. The grid-box average, G, is therefore,

$$G = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(T_{ij} + M_{ij} + B_i). \tag{2}$$

It is assumed that the Ms and Bs are independent of the Ts and the Ms are independent of the Bs. Hence

$$\mathrm{var}(G) = \frac{1}{n^2}[\mathrm{var}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(T_{ij}) + \mathrm{var}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(M_{ij}) \tag{3}$$

$$+\mathrm{var}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(B_i)] \tag{4}$$

and all the covariance terms between the different sources of error are equal to zero. The variance of the first term in the square brackets can be expressed (*Kagan* [1966]; *Yevjevich* [1972]) as,

$$\frac{1}{n^2}\mathrm{var}\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n_i}}(\mathrm{T_{ij}}) = \frac{1}{n}\sigma_s^2\left[1 + (n-1)\,\bar{r}\right] \tag{5}$$

where σ_s^2 is the standard deviation of the SST anomalies at a fixed point, which is assumed to be a constant in any given grid box, and $\bar{r}$ is the average correlation of the SST anomalies measured at any pair of points within the grid box. The Ms are independent of each other, as are the Bs, so we have

$$\frac{1}{n^2}\mathrm{var}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(M_{ij}) = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{var}\,(M_{ij}) = \frac{1}{n^2}\sum_{i=1}^{m} n_i\sigma_{m_i}^2 \tag{6}$$

and

$$\frac{1}{n^2}\mathrm{var}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(B_i) = \frac{1}{n^2}\sum_{i=1}^{m}\mathrm{var}[\sum_{j=1}^{n_i}(B_i)] = \tag{7}$$

$$\frac{1}{n^2}\sum_{i=1}^{m}\mathrm{var}[n_iB_i] = \frac{1}{n^2}\sum_{i=1}^{m} n_i^2\sigma_{b_i}^2. \tag{8}$$

The total variance of the grid-box average is therefore

$$\sigma^2_{tot} = \frac{1}{n^2}\sum_{i=1}^{m} n_i \sigma^2_{m_i} + \frac{1}{n^2}\sum_{i=1}^{m} n_i^2 \sigma^2_{b_i} + \frac{1}{n}\sigma^2_s \left[1 + (n-1)\,\bar{r}\right]. \tag{9}$$

The total variance includes a component representing the variability of the true grid-box average from one month to the next. This variability is given by the constant term ($\sigma^2_s \bar{r}$) and represents the climate signal. Subtracting $\sigma^2_s \bar{r}$ from the total variance leaves only those terms that together provide an estimate of the excess variance due to under-sampling errors and measurement errors:

$$\sigma^2_{error} = \frac{1}{n^2}\sum_{i=1}^{m} n_i \sigma^2_{m_i} + \frac{1}{n^2}\sum_{i=1}^{m} n_i^2 \sigma^2_{b_i} + \frac{1}{n}\sigma^2_s \left[1 - \bar{r}\right] \tag{10}$$

This formula gives the uncertainty on the grid-box average for a given set of observations made within that grid box and tends to zero as n and m increase. To better understand the formula it is instructive to consider the case where each of the m ships takes the same number ($n_i = n/m$) of observations and has the same random measurement and micro-bias error characteristics, then the formula above reduces to

$$\sigma^2_{error} = \frac{\sigma^2_m}{n} + \frac{\sigma^2_b}{m} + \frac{\sigma^2_s\,(1-\bar{r})}{n}. \tag{11}$$

The three terms correspond to the random measurement uncertainty, the micro-bias uncertainty and the sampling uncertainty respectively. This is the minimum uncertainty for a given number of observations n and ships m and shows that it is only possible to reduce significantly the uncertainty of the grid-box average by increasing both the number of observations (n) and the number of ships (m) making those observations. It is also interesting to consider the limiting cases for a given number of observations. The uncertainty of the grid-box average will be a minimum when $m = n$ and a maximum when $m = 1$. Between these extremes, the uncertainty will tend to reduce with n more slowly than n^{-1}. The micro-bias error term will also lead to correlations between the errors in two grid boxes where the same ship makes observations in both. This will be explored further in the next section.

3.1. Inter-grid box error correlations

The micro-bias error, $\sigma^2_{b_i}$, for a given ship in equation 10 is correlated both within a grid box and between any two grid boxes that the ship visits. Random measurement and sampling errors are uncorrelated between grid boxes. Therefore the covariance, $C_{p,q}$, between two grid boxes is:

$$C_{p,q\,p\neq q} = \frac{\sum_k n_{p_k} n_{q_k} \sigma^2_{b_k}}{N_p N_q} \tag{12}$$

where n_{p_k} and n_{q_k} are the number of observations from ship k in grid boxes p and q respectively. k runs from 1 to the number of ships that are present in both grid boxes. N_p and N_q are the total number of observations in boxes p and q. If there are no ships that made measurements in both grid boxes then $C_{p,q}$ is zero. $C_{p,p}$ is given by the error variance on the grid-box average in Equation 10.

To do this calculation it is necessary to know which ships took measurements in each grid box, how many observations were made by each ship (n_{p_k} and n_{q_k}) and the estimated bias error associated with that ship ($\sigma^2_{b_k}$). Given this information, the covariance matrix can easily be calculated during gridding. Unfortunately, as was noted above, not all SST observations made by ships can be unambiguously associated with an individual vessel. This problem is discussed in Section 4.2.

3.2. Sampling error

Sampling errors arise when the area-average of a spatially-varying quantity is estimated from a finite number of observations. Even if the individual measurement errors were zero, the mean of the observations in the grid box would not be equal to the true spatial-average. As discussed above, the total variance of a grid-box average of n perfect observations is given (*Kagan* [1966]; *Yevjevich* [1972]) by

$$\sigma^2_{grid} = \frac{\sigma^2_s}{n}\left(1 + \bar{r}\,(n-1)\right). \tag{13}$$

The formula assumes that the variance at every point within the grid box is equal to σ^2_s and that the points are randomly distributed. These are likely to be good approximations for most grid boxes, but this additional uncertainty is discussed in Section 4.5. The sampling uncertainty term is the part that depends on n i.e. the excess variance caused by under sampling:

$$\sigma^2_{se} = \frac{\sigma^2_s}{n}\left(1 - \bar{r}\right). \tag{14}$$

It is possible to estimate σ^2_s for a given grid box by calculating the grid-box average variance for months when n is large.

$$\sigma^2_{n=\infty} = \sigma^2_s \bar{r} \tag{15}$$

The value of $\bar{r}$ can be estimated from the data as in *Jones et al.* [1997] and *Brohan et al.* [2006] who used the correlations between grid boxes to estimate the average correlation within a grid box for land stations. Land stations generally make two or more observations a day every day, thus ensuring perfect temporal sampling. The only concern then is that the measurements made at the station are not representative of the whole area of the grid box. Estimating $\bar{r}$ is slightly more complicated for marine data because observations are sampled randomly in both space and time. *Morrissey and Greene* [2009] examined the case of SST sampling and extended the $\bar{r}$ concept to include a time dimension. The special case where the observations are randomly distributed and there is no correlation between the space and time components is used here. It will tend to lead to a slightly larger estimate of the sampling error than for the case where the two are correlated.

In *Jones et al.* [1997] $\bar{r}_{space}$was estimated by calculating a correlation decay length from inter-grid box correlations. A similar method is followed here. The correlation between two points is assumed to be of the form:

$$r = \exp(\frac{-x}{x_0}) \tag{16}$$

where x is the distance between two points and x_0 is the characteristic length scale. In most grid boxes an exponential form for r gives a better fit to the data than a gaussian form.

SST data were taken from 1961 to 2010, using HadSST2 to extend the series from 2007-2010. The data were detrended using a 5-term polynomial fit to the time series of observations in each grid box. Winsorised correlations (*Wilcox* [2001]) were calculated between the de-trended time series of SST from a given grid box with de-trended time series in all grid boxes within $\pm 100°$of longitude and $\pm 20°$of latitude. A minimum of 30 months of coincident data were

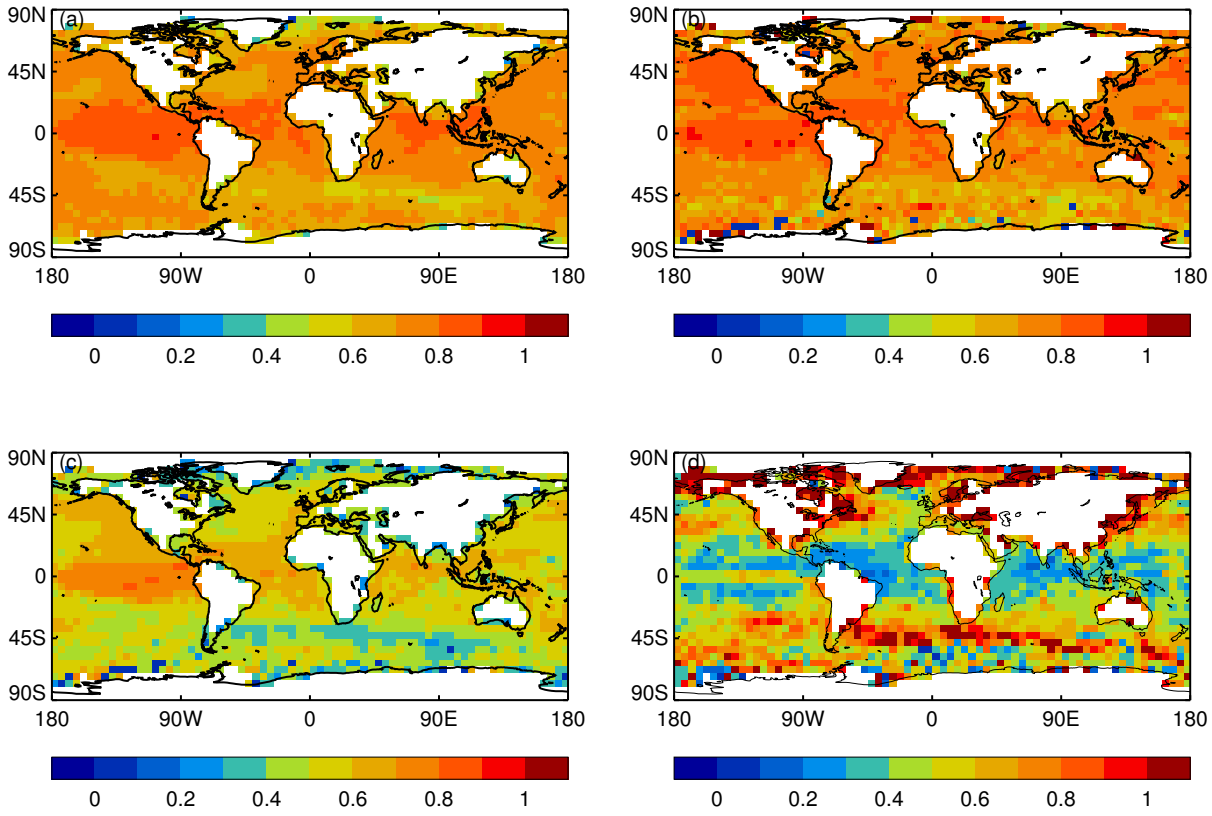

Figure 1. (a) $\bar{r}_{space}$, (b) $\bar{r}_{time}$, (c) $\bar{r}_{all}$ and (d) $\sqrt{\sigma_s^2(1-\bar{r})}$, the sampling error (°C) on a single observation.

required to estimate the correlations. The value of x_0 that minimised the RMS difference between the observed correlations and the model was calculated. The average spatial correlation, $\bar{r}_{space}$, within each grid box was estimated by choosing 100000 pairs of points from within the grid box, calculating the correlations using equation 16 and taking the average. The map of $\bar{r}_{space}$ is shown in Figure 1.

$\bar{r}_{time}$ was calculated in an analogous manner by fitting an exponentially decaying lag correlation function to the de-trended (using a low-pass filter) monthly time series of SST anomalies in each grid box. The resulting field of $\bar{r}_{time}$ was multiplied by $\bar{r}_{space}$to get the final estimate of $\bar{r}_{all}$ for each grid box (Figure 1).

Typically the average correlations are highest in the tropics, particularly in the eastern Pacific. They are lower in the region of the western boundary currents, on the northern boundary of the Southern Ocean and at the edges of the field near areas of seasonal sea ice cover.

The true standard deviation at a point in a given grid box, σ_s^2, was calculated from $\sigma_{n=\infty}^2 = \sigma_s^2\bar{r}$. This was estimated by *Rayner et al.* [2006] from the observed data. A map of $\sqrt{\sigma_s^2(1-\bar{r})}$, the sampling uncertainty of a grid box average calculated from a single observation, is shown in Figure 1. Sampling uncertainties are highest in the western boundary currents and along the northern boundary of the Southern Ocean where SST can change rapidly and spatial temperature gradients are strong. Sampling uncertainties are smaller in the tropics - particularly in the Indian and Atlantic Oceans - where the spatial and temporal correlations are large. In the eastern tropical Pacific there is an area of higher variability in the region of the El Niño cold tongue. In the North Pacific the Kuroshio extension increases sampling uncertainty as far east as Hawaii. The pattern and magnitude of the sampling uncertainty is similar to that calculated from sub-sampling complete satellite data (*Rayner et al.* [2009]).

3.3. The full error covariance matrix

Combining all the terms of the error covariance matrix gives

$$C_{pp} = \frac{1}{n^2}\sum_{i=1}^{m} n_i\sigma_{m_i}^2 + \frac{1}{n^2}\sum_{i=1}^{m} n_i^2\sigma_{b_i}^2 + \frac{1}{n}\sigma_s^2\left(1-\bar{r}\right) \quad (17)$$

along the diagonal, and

$$C_{pq,\,p\neq q} = \frac{\sum_k n_{p_k} n_{q_k}\sigma_{b_k}^2}{N_p N_q} \quad (18)$$

on the off-diagonal, which can be evaluated only when we have call sign information. When calculating area-averages and other useful quantities, it is necessary to keep track of the uncertainties through the calculation. For a generic function $f = f(x_1, x_2 ... x_n)$, where the covariance of x_p and x_q is given by $C_{p,q}$, the standard expression for the error variance of f, σ_f^2 is given by

$$\sigma_f^2 = \sum_{p=1}^{n}\sum_{q=1}^{n}\left(\frac{\partial f}{\partial x_p}\frac{\partial f}{\partial x_q}C_{pq}\right). \quad (19)$$

When f is the weighted average of n grid boxes

$$f = \frac{\sum_{p=1}^{n} a_p x_p}{\sum_{p=1}^{n} a_p} \quad (20)$$

where a_p is the weight of grid box p and

$$\frac{\partial f}{\partial x_p}\frac{\partial f}{\partial x_q} = \frac{a_p a_q}{\left(\sum_{l=1}^{n} a_l\right)^2} \quad (21)$$

Therefore, in matrix notation,

$$\sigma_f^2 = \frac{aCa^T}{(\Sigma_{l=1}^{n} a_l)^2} \quad (22)$$

where a is a vector of weights $a = (a_1, a_2, ..., a_n)$. For calculating area averages of the SST data, the values of a are the grid box areas, or set to zero where data are missing.

3.4. Uncertainties of temporal averages

It is useful to know the uncertainty of annual average values as well as monthly average values. The correlated errors that lead to non-zero covariance in the spatial fields of uncertainty will also lead to non-zero covariances in the full space-time covariance matrix. The non-zero covariances need to be accounted for when propagating uncertainties from monthly to annual averages. There are two extreme cases: first that the monthly errors are independent; second, that they are completely correlated. All other assumptions being correct, the correct error variance will lie somewhere between the two corresponding estimates.

It is impractical to create the covariance matrix for the whole year. For a 5°latitude × 5°longitude monthly analysis such a matrix would contain nearly a billion $((72\times36\times12)^2)$ elements. Instead, the weight that each observation takes in the global average was calculated and the uncertainty was built up in the following way.

$$\sigma^2_{global\,average} = \frac{1}{(\Sigma w_{ij})^2}\Big(\sum_{i=1}^{m}\Big(\sum_{j=1}^{n_i} w_{ij}\Big)^2\sigma^2_{b_i} \quad (23)$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}^2\sigma^2_{m_i} \quad (24)$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}^2\sigma^2_{se_{ij}}\Big) \quad (25)$$

where w_{ij} is the weight that observation j from ship i takes in the global annual average. $\sigma_{se_{ij}}$ is the sampling error in the grid box containing observation ij. w_{ij} depends on how the annual average is calculated. Here, the annual global average is calculated by first taking the area-weighted average of the 12 monthly SST fields and then averaging the 12 numbers to get the annual average. Therefore the weight of each observation is

$$w_{ij} = \frac{area_{i,j}}{n_{gridbox}n_{months}area_{total,month}} \quad (26)$$

where $area_{i,j}$is the area of the grid box containing observation i, j, and the $area_{total,month}$ in the denominator is the total area of occupied grid boxes in the field for a given month. $n_{gridbox}$ is the number of observations in the monthly grid box containing observation j from ship i and n_{months} is the number of months in the average, usually 12.

4. Error model - implementation

The method was applied to data from ICOADS 2.5 from 1850 to 2006. Each year was split into 12 *pseudomonths* comprising six five-day *pentads* with the exception of August which has seven. Leap years have an extra day in the final February pentad. The data were processed following *Rayner et al.* [2006]. In *Rayner et al.* [2006] the grid box averages were calculated by first taking the winsorised average of the SST anomalies within each 1x1xpentad grid box. The winsorised average of the 1x1xpentad grid boxes within each final 5x5xmonthly grid box were then calculated. Despite the averaging method differing from that assumed in the error model, equation 9 provides a good fit to the grid box variances. This was checked by repeatedly resampling a grid box in the North Sea which contains more than 1000 observations each month (not shown). In this way it was possible to ensure that the model fitted the data for a wide range of n and m.

The values of $\sigma^2_{m_i}$ and $\sigma^2_{b_i}$ are taken from *Kennedy et al.* [2011b] for ships and drifting buoys and *Kent and Berry* [2008] for moored buoys. In *Kennedy et al.* [2011b] observations from each unique ship or drifting buoy were compared to a background SST field from the Advanced Along-Track Scanning Radiometer (AATSR) instrument and the difference time series were used to calculate representative values of $\sigma^2_{m_i}$ and $\sigma^2_{b_i}$ for ships and drifting buoys. The differing values, shown in Table 1, reflect the relative accuracies of ship and buoy measurements. *Kent and Berry* [2008] calculated $\sigma^2_{m_i}$ and $\sigma^2_{b_i}$ for moored buoys by comparing their measurements to SST fields taken from the Met Office Numerical Weather Prediction system.

In the following analysis all ships are considered to have the same error characteristics i.e. a fixed value of σ_m and σ_b. Drifting and moored buoys have their own characteristic values. The values used are summarised in Table 1.

The error model as described above assumes that it is possible to unambiguously identify each ship, or buoy. However, this is not always possible. Often the call signs by which individual ships can be identified are missing from the ICOADS reports or a generic call sign is used instead. In other cases the ship identifier is incomplete. For example, between 1870 and the 1910s many observations have call sign "0120". Many of these observations have the same time stamp, but different locations so the call sign cannot be a unique identifier. The non-unique call signs identified were 'SHIP', 'SHIPX', '0120', 'PLAT', '1', '58', '7', and ' ' (i.e. no call sign).

Excluding these observations from the uncertainty calculation would lead to an underestimate of the uncertainties. An alternative approach might be to assume that the observations with non-unique call signs did indeed come from the same vessel, but this would lead to a very large overestimate of the uncertainties at those times when many observations cannot be uniquely identified with a particular ship. The following sections attempt a more reasonable estimate of the uncertainty component associated with the unidentifiable observations.

4.1. Grid-box average

The first difficulty in implementing the error model is in estimating the micro-bias uncertainty term,

$$\sum_{i=1}^{m} n_i^2\sigma^2_{b_i} \quad (27)$$

Table 1. Estimated random measurement (σ_m) and micro-bias (σ_b) uncertainties for ships and drifting buoys from *Kennedy et al.* [2011b] and for moored buoys from *Kent and Berry* [2008].

Platform	σ_m(°C)	σ_b(°C)
Ship	0.74	0.71
Mooring	0.30	0.20
Drifter	0.26	0.29

when the ships cannot be identified and the n_i are therefore unknown. However, what is known in all situations is n, the total number of observations in the grid box. The micro-bias term can be estimated using only this information.

The value of Equation 27 depends on how the n observations are partitioned between the ships that made them. Assuming that the partitioning of observations between ships is roughly constant for a given grid box over time, $\sum_i n_i^2$ can be estimated using a function of the form

$$an^b \approx \sum_i n_i^2 \tag{28}$$

where n is the total number of observations in the grid box. The parameters a and b can be estimated from those observations for which the call signs are known and for which $\sum n_i^2$ can therefore be evaluated. Values of a and b were estimated for each grid box for each year using data from alternate months (January, March, May...) for the nine year period centred on that year. The logarithm of $\sum n_i^2$ was regressed on the logarithm of n.

$$\log \sum_i n_i^2 = \log(a) + b \log(\mathrm{n}) \tag{29}$$

Ships that took hourly observations, such as ocean weather ships, were excluded because they tended to skew $\sum n_i^2$ towards high values that were not representative of typical SST observations. Months where the value of $\sum n_i^2$ was in the upper 15% were also rejected because these were often dominated by a few exceptionally high values. The model parameters were verified using the remaining months (February, April, June...). Where the correlation of the estimate from the model with the actual value was below 0.5, a and b were flagged as missing. Values of a and b are shown in Figure 4.

Where observations were available from many ships, b was close to 1. For example, in the case where each of m ships takes g observations, n is equal to mg and $\sum n_i^2$ is equal to mg^2, so $a = g$ and $b = 1$. Where most of the observations come from a single ship, b approaches 2. Therefore, along the well travelled shipping lanes, the observed value of b tends to be around 1. At high latitudes and in other areas where shipping traffic is more sporadic, b tends to be higher, reflecting the fact that in these areas the grid-box average is likely to be based on observations from a smaller number of vessels. When b is close to 2, the value of a is generally close to 1, as expected.

Where it was not possible to estimate a and b, it was assumed that b was equal to 2 and a was equal to 1. This is the most conservative case and reflects the fact that in areas where observations were not sufficiently numerous to calculate reliable estimates of a and b, the observations were likely to have come from a single ship. In the 1860s there were few observations and most had no call sign, so many grid boxes had to be filled in this way. Once a and b had been calculated, they were used to estimate the micro-bias uncertainty term thus:

$$\sum_{i(ID\,known)} n_i^2 \sigma_{b_i}^2 + a n_{unknown}^b \sigma_b^2 \tag{30}$$

All observations with missing or generic call signs were assumed to come from ships and not from drifting or moored buoys. This information can then be combined with the sampling and random measurement uncertainties, like so

$$C_{pp} = \frac{1}{n^2} \sum_{i=1}^{m} n_i \sigma_{m_i}^2 + \frac{1}{n^2} \sum_{ID\,known} n_i^2 \sigma_{b_i}^2 \tag{31}$$

$$+ \frac{1}{n^2} a n_{unknown}^b \sigma_b^2 + \frac{1}{n} \sigma_s^2 (1 - \bar{r}) \tag{32}$$

along the diagonal.

4.2. Regional average

Because of the problem of missing call signs, it is not possible to explicitly calculate the off-diagonal elements of

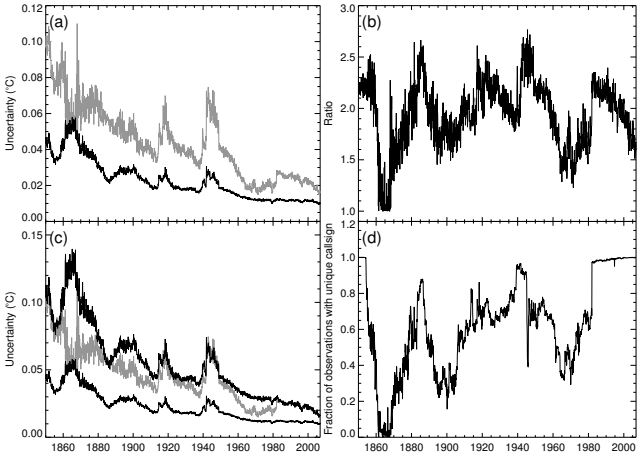

Figure 2. (a) Uncertainty of global monthly average SST anomaly for the error model including inter-grid box correlations (grey) and for the error model with no inter-grid box correlations (black). (b) Ratio of the two lines in panel (a). (c) as for panel (a) but including an estimate of the full uncertainty of the global monthly average SST anomaly (upper black line). This is 2.4 times the lower black line prior to 1982 and equal to the grey line after 1982. (d) Fraction of observations with unique call signs.

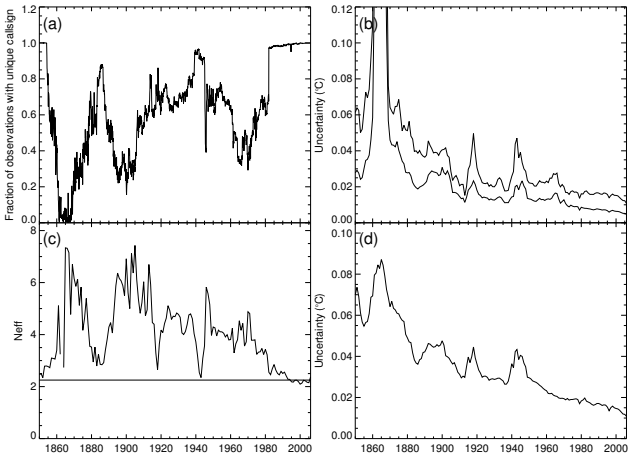

Figure 3. (a) fraction of observations in each month from 1850 to 2006 with unique call signs. (b) Uncertainty of the global annual average sea-surface temperature calculated from only those observations with unique call signs. The upper line shows the full error calculation described in Section 4.3, and the lower line was calculated from the uncertainties on the monthly averages assuming that they are independent. (c) N_{eff} calculated from the uncertainties shown in panel (b). The horizontal line is at $n_{eff} = 2.25$. (d) The uncertainty of the global annual average sea-surface temperature calculated using all observations and assuming $n_{eff} = 2.25$.

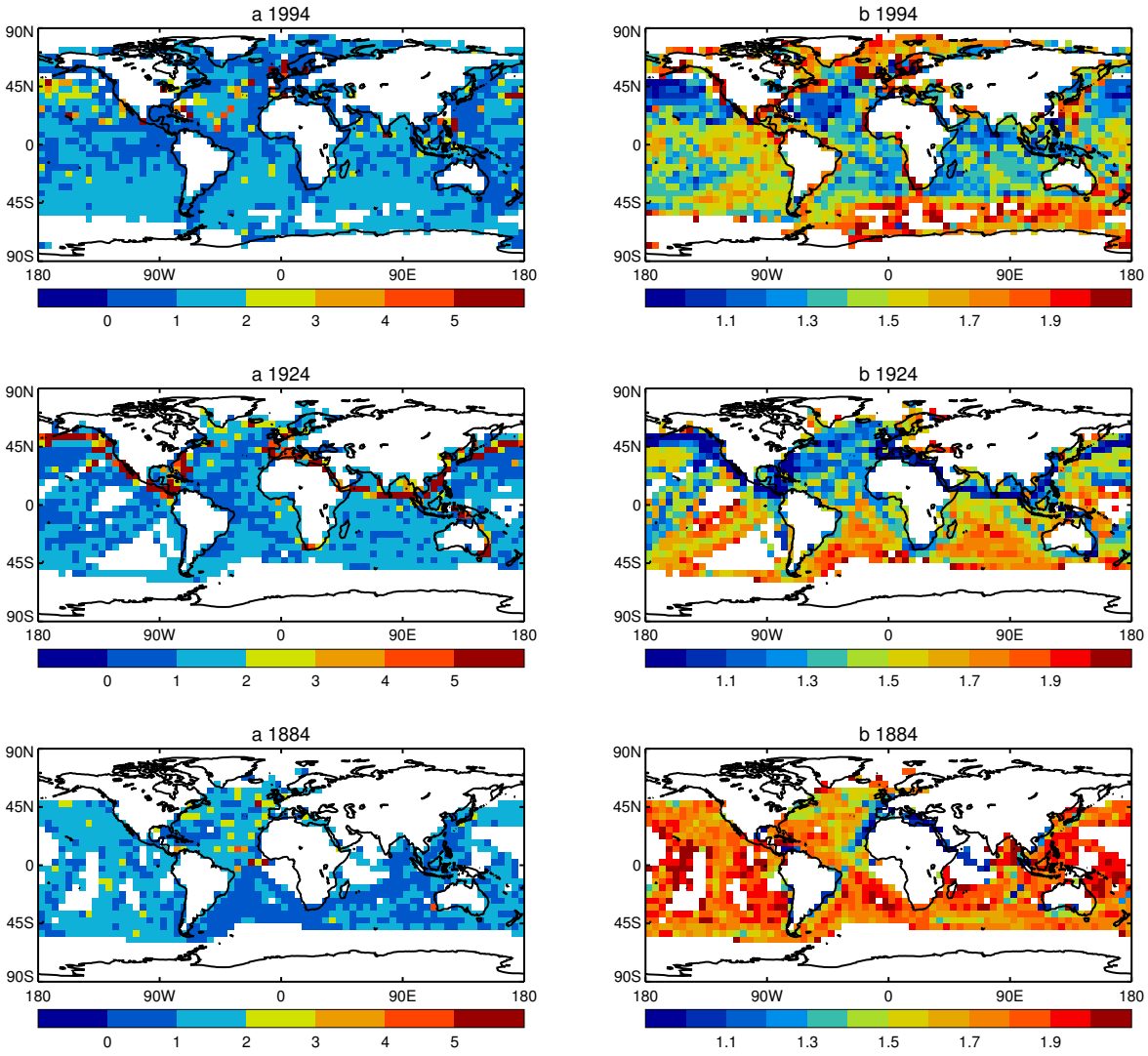

Figure 4. Values of a (left) and b (right) calculated for three different nine-year periods centred on: 1994 (top), 1924 (middle) and 1894 (bottom). a and b are dimensionless quantities.

the covariance matrix (Equation 12) for all months. To estimate the size of this effect, the off-diagonal elements, C_{pq}, were calculated using only data from ships that could be identified unambiguously. Generic call signs (SHIP, SHIPX, 0120, PLAT etc.) were not included, neither were observations with no call sign. A second estimate was made using all the data, but assuming that all the off-diagonal elements were set to zero. Figures 2(a) and (b) show the uncertainty of the global average calculated using these two methods and the ratio between them. Also shown are the fractions of observations with unique call signs (Figure 2(d)).

In the 1860s and 1970s there are few observations with a unique call sign. In these cases C_{pq} could only be estimated in very few cases so the grey line runs close to the black line and is an underestimate of the true uncertainty. In contrast, there are a number of periods during which the number of identified callsigns is large relative to the number of unidentified call signs: 1850-1854, 1880-1890, 1915-1945 and from around 1982. Before 1982, there is an approximately linear relationship between the fraction of observations with unique call signs and the ratio shown in Figure 2(b). Extrapolating a linear fit to the case where all observations have unique call signs gives an estimated ratio of 2.4 between the uncertainties based only on the diagonal elements and the uncertainty from the full error model in the case when all call signs are known.

An estimate of the uncertainty of the global average was made by multiplying the series of uncertainties based only on the diagonal elements by 2.4. After 1982, the estimates from the full error model were used because the number of non-unique call signs is relatively small and the sampling characteristics changed markedly with the introduction of drifting and moored buoys. The resulting combined series is shown in the upper black line of Figure 2(c) which is 2.4 times the lower black line until 1982. The multiplier is different for each region, reflecting the nature of the local shipping and observational coverage. Prior to 1982, a factor of 2.4 was used for the globe and southern hemisphere and a factor of 2.2 was used for the tropics. A multiplier of 2.1 was more appropriate for the better observed northern hemisphere, again, prior to 1982.

4.3. Temporal average

The calculation of the uncertainty of an annual average, σ^2_{annual}, is once again complicated by the presence of missing call signs. This is dealt with in a similar way to the regional

averages, by comparing the annual and monthly uncertainties calculated from observations for which the call signs are all known. There are two limiting cases

$$\frac{\sum \sigma^2_{monthly}}{12^2} \leq \sigma^2_{annual} \leq \frac{\sum \sigma^2_{monthly}}{12} \quad (33)$$

where $\sigma^2_{monthly}$ are the error variances of the monthly averages. In the first case the monthly uncertainties are assumed to be independent. In the latter case, they are assumed to be perfectly correlated. In practice, the truth will lie somewhere between these extremes. An effective number of months, n_{eff}, was defined such that

$$\sigma^2_{annual} = \frac{\sum \sigma^2_{monthly}}{12 n_{eff}} \quad (34)$$

To estimate n_{eff}, Equation 25 was used to estimate the uncertainty on the global annual average, σ^2_{annual}, calculated only from those observations with unique callsigns. The sum of the error variances of the monthly global averages was calculated from only those observations with unique call signs and divided by twelve times the error variance on the annual average to get n_{eff}. The results are shown in Figure 3. Figure 3(b) shows the uncertainty of the annual average calculated only for those observations with call signs and the lower line shows the uncertainty on the annual average assuming that the monthly errors are independent. Panel (c) shows the calculated value of n_{eff} for each year. The lowest values of n_{eff} are around 2.25 in the late 1990s, suggesting that in this period there was significant correlation between errors in individual months. At other times values as high as 7.5 were recorded. The values can be used to estimate uncertainties on averages at annual time scales for the full data set by assuming a constant value for n_{eff} and using this to estimate the uncertainty on the annual average from the uncertainties on the monthly averages (Figure 3d). In this case n_{eff} was chosen to be 2.25, as this is a reasonably conservative estimate that is also consistent with well observed periods in the record.

4.4. Coverage uncertainty

When calculating area averages from a gridded data set there is an additional uncertainty that arises because there are often large areas, and consequently, many grid boxes, which contain no observations. Such uncertainties are referred to here as coverage uncertainties. In *Brohan et al.* [2006] coverage uncertainties were estimated by subsampling reanalysis data. A similar method is used here. SST anomalies from the globally complete HadISST1 data set (*Rayner et al.* [2003]) were used in the place of reanalysis data. For example, to calculate the uncertainty on the March 1973 monthly average for the North Pacific a time series of North Pacific average SST anomalies was calculated using HadISST from 1870 to 2010. The coverage of HadISST at all time steps was then reduced to that of HadSST3 for March 1973. The North Pacific time series was recalculated from the sub-sampled data and the standard deviation of the difference between the series from the complete and subsampled series was used as an estimate of the uncertainty for March 1973. Data from the ERSSTv3b (*Smith et al.* [2008]) and COBE (*Ishii et al.* [2005]) data sets were also used in place of HadISST1 and gave similar results suggesting that the uncertainties do not depend strongly on the statistical assumptions made in creating HadISST1. Coverage uncertainties calculated using HadISST1 are shown in Figure 6.

4.5. Additional uncertainties

A number of additional sources of uncertainty are not included in this analysis. These include, but are not limited to, the following.

1. Uncertainties associated with the adjustments for biases in the data are dealt with in part 2 of this paper (*Kennedy et al.* [2011a]). On larger spatial and longer temporal scales, they are comparable to, or of greater importance than the measurement and sampling errors discussed here.

2. No estimate has yet been made of uncertainties in the estimate of the climatological average SST that has been used to convert actual SST measurements to anomalies. In sparsely observed regions, such as the Southern Ocean, the climatological uncertainty could be large. For future applications it might be wise to estimate climatological averages from a modern well-observed period and make use of satellite data. There is an open problem in how regions that have, until recently, been covered by sea ice should be treated. Is it meaningful to assign a sea-surface temperature anomaly to a region that, during the climatology period, was covered with sea ice? This question is of particular significance if estimates of SST anomalies are combined with land surface air temperature anomalies to create an estimate of global average temperature.

3. Uncertainties inherent in the estimated values for σ_m, σ_b, σ_s and $\bar{r}$ have not been included. For example, due to random measurement and sampling errors, the estimates of $\bar{r}$, the average correlation of two points within a grid box derived in Section 3.2, might be underestimated leading to a sampling uncertainty that is too large. This effect has been reduced by the use of a robust measure of correlation. Another compensating factor might be that micro-bias errors lead to a slight increase of the correlation of separated grid boxes. It has also been pointed out (*Rayner et al.* [2009]) that sampling within a 5 degree grid box might not be random as supposed here. *Morrissey and Greene* [2009] developed a more general means of estimating the sampling error that takes into account the locations of the observations within the grid boxes. Clustering of the observations would typically lead to an underestimate of the sampling uncertainty, which currently assumes that the observations are randomly distributed. Work by *Kent et al.* [1993] and *Kent and Challenor* [2006] shows that the measurement uncertainties of SST observations from ships using buckets to make their SST measurements were different from those made by ships using engine intake water. They also found variations between the ships recruited by different countries

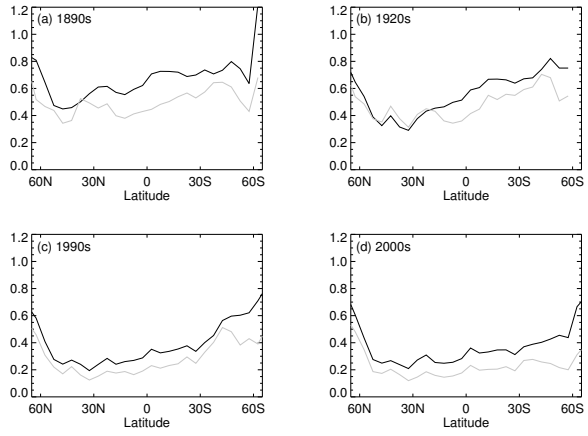

Figure 5. Zonal average of 5° latitude by 5° longitude grid box 1-sigma uncertainties (°C) for: (a) 1890s, (b) 1920s, (c) 1990s and (d) 2000-2006. The black line is the new estimate of the uncertainty and the grey line is the estimate from Rayner et al. (2006).

and that many observations were only recorded to whole or half degrees. More detailed analysis using metadata for different ships could be used to further refine the analysis in future.

4. Uncertainties arising from outliers, or otherwise non-normal deviations, in the data have not been considered although their effects have been quantified elsewhere (*Kennedy et al.* [2011b]).

5. Results

Figure 5 shows the zonal mean of the individual grid-box uncertainties for four different periods. Despite improvements in data coverage associated with using ICOADS 2.5 rather than ICOADS 2.0, the grid-box uncertainties are generally larger than the estimates made in *Rayner et al.* [2006] owing to the correlation of errors within the grid boxes, which was neglected in *Rayner et al.* [2006].

In the 1890s, the uncertainties are higher at most latitudes reflecting the sparseness of the observations and the small number of platforms. In the 1920s, the mid-latitude northern hemisphere was more densely observed, principally as a result of increased shipping rather than an increase in the number of observations per ship, and consequently there is little difference between the estimates. The latitude at which the lines cross is also marked by high variability as it is the latitude of the Gulf Stream and Kuroshio. In these regions, the uncertainty of a grid-box average based on a single observation is often higher in HadSST2 than in HadSST3. However at lower latitudes and in the southern hemisphere the effects of data sparsity and the lack of diversity in the observing network are still apparent. By the 1990s, the density of observations had increased at most latitudes except in the southern hemisphere, southward of about 40°S, where uncertainties remain large. There is a relative increase in the average uncertainty (between HadSST2 and HadSST3) in the mid-latitude northern hemisphere in the 1990s that was larger than in the 1920s. This arises because, although more observations were made in the 1990s, this was due mainly to an increase in the number of observations made per ship without a proportionate increase in the number of ships. Since 2000, the difference is more uniform and reflects a move to fewer platforms (principally drifting buoys) making more observations. *Kent and Berry* [2008] saw a similar increase in grid box uncertainties relative to those in *Rayner et al.* [2006].

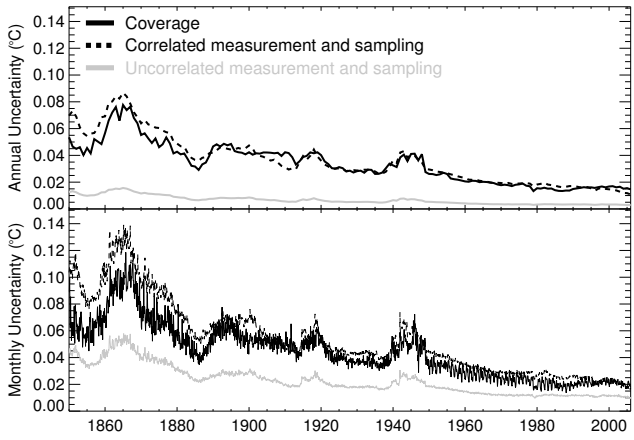

Figure 6. 1-sigma uncertainties of annual-average (upper panel) and monthly-average (lower panel) global SST anomalies. The three lines in each panel correspond to the coverage uncertainty (solid black line), the correlated measurement and sampling uncertainty (dashed black line) and the uncorrelated measurement and sampling uncertainty (solid grey line).

A significant increase in global and regional uncertainties arises from the inter-grid box correlations. The off-diagonal terms of the error covariance lead to approximately a factor of 2.4 increase in the global-average uncertainty relative to the case where they are assumed to be uncorrelated between grid boxes. This effect can be seen in the time series of uncertainties in global, hemispheric and tropical average sea-surface temperature anomalies shown in Figure 7. The red line, which is not corrected for missing call signs, is likely to be an underestimate at times when there are many missing call signs. This is most apparent between 1860 and 1870 when the off-diagonal component is almost zero. Very few call signs are associated with the observations in this data-sparse period. The sudden apparent increase in uncertainty in 1982 arises because there are far more observations with call signs available after 1982 than before. The estimated full error range which accounts for the absent and generic call signs is given by the upper blue line in each diagram.

Overall, there is a general decrease in the uncertainty of the global average SST from the mid-nineteenth century to the present, interrupted by periods of increased global uncertainty during the two world wars. The uncertainty is highest in the 1860s, around 0.13°C, when there are few observations and fewer extant metadata. The dip in southern hemisphere uncertainty in 1979 is due to the temporary deployment of many drifting buoys during the First GARP Global experiment. The fall in total uncertainty at the very end of the series occurred after the number of drifting buoys in the oceans was almost doubled. This had a particularly striking effect in the southern hemisphere. Measurement and sampling uncertainties are comparable to the coverage uncertainties (Figure 6) throughout much of the record with each exhibiting similar temporal correlations.

It is interesting to compare the SST uncertainties with comparable estimates of uncertainty on global and hemispheric averages of land surface air temperature. In *Brohan et al.* [2006], in which land surface air temperatures (LSAT) and their uncertainties were calculated, uncertainties on SST (from *Rayner et al.* [2006]) were generally much smaller than uncertainties on LSAT. This was due to a large extent to the higher variability of temperature anomalies over land. Although the new uncertainties on SST measurements are significantly larger than those in *Rayner et al.* [2006] they are still typically lower than hemispheric estimates of land surface air temperature uncertainties.

6. Summary

An error model that accounts for correlated (micro-bias) and independent (random measurement) random errors in sea-surface temperature measurements was described. The correlated errors lead to an increase in uncertainties of grid-box average and regional-average SSTs compared to previous estimates. It was not possible to estimate the full correlated error term at all times because that requires that each contributing ship can be unambiguously identified. At times this was not possible because the call sign information was not included in the meteorological report. From December 2007, the call signs of GTS reports were removed from some feeds and encrypted on others, so this analysis could only be completed for the period 1850 to 2006. New estimates of sampling uncertainty were also calculated.

The new estimates of the grid-box uncertainties for HadSST3 are typically larger than estimates made in

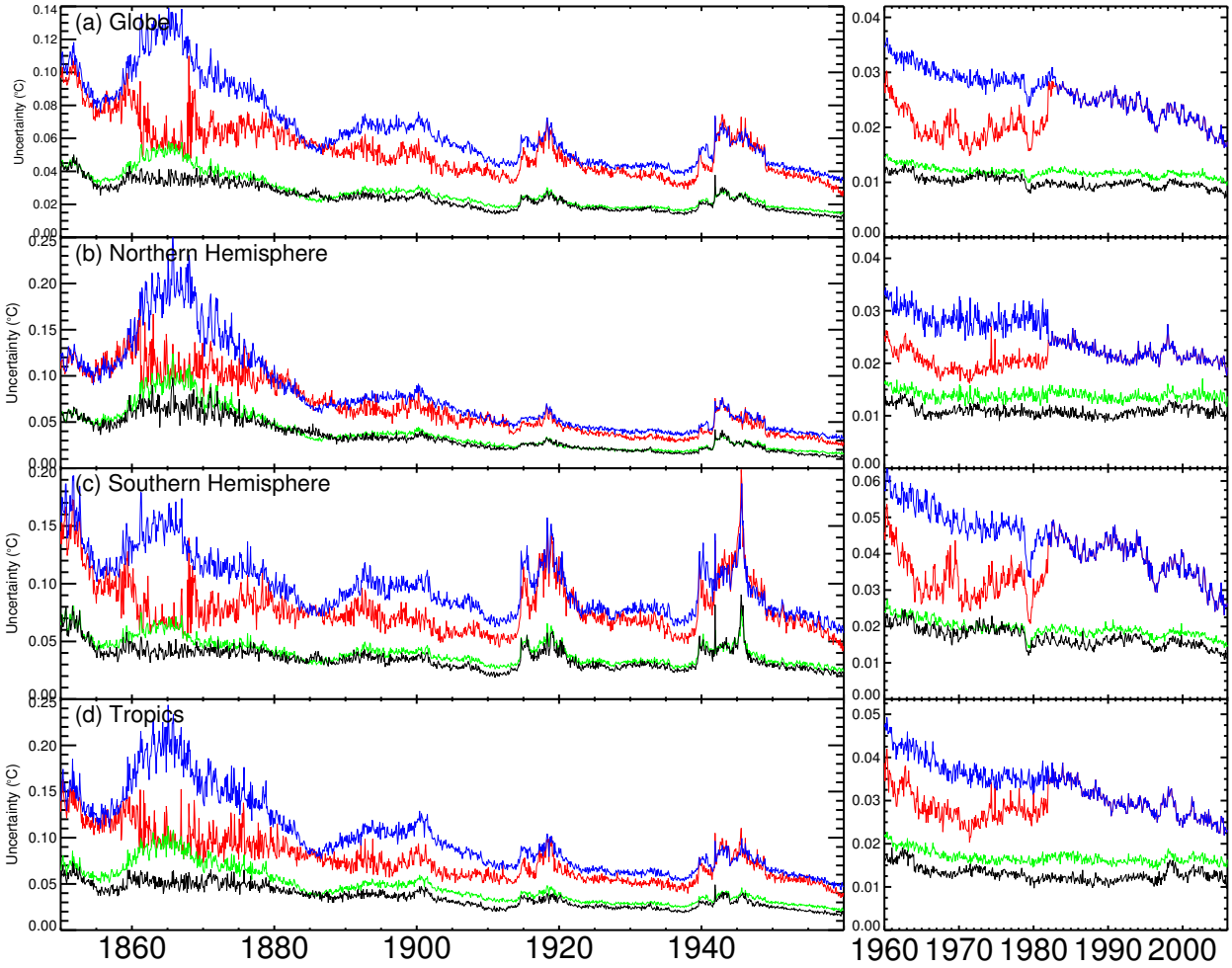

Figure 7. 1-sigma uncertainties on monthly SSTs for (a) Global-, (b) Northern Hemisphere-, (c) Southern Hemisphere- and (d) Tropical-averages. The blue line shows the uncertainty calculated using the full error model including inter grid box correlations corrected for missing callsigns. The red line shows the uncertainty calculated using the full error model including inter grid box correlations, but not corrected for missing call signs. The green line shows the uncertainty if the grid-box errors are assumed to be uncorrelated. The lower black line is the Rayner et al. (2006) estimate. Note the different scales on each diagram. The right hand column is shown on an expanded scale because the uncertainties are generally smaller in the later period.

HadSST2. The differences are largest in areas where the observing network was less diverse and observations were made by only a small number of ships or buoys; for example, at high latitudes, in the earliest part of the record and during the Second World War. The new estimates of the uncertainties of regional averages are also larger owing to correlations between grid boxes. At times when reliable estimates can be made, the uncertainties in global-average, southern-hemipshere and tropical-average sea-surface temperatures are between two and three times as large as the case where errors are considered to be uncorrelated. Uncertainties of northern hemisphere averages approximately double.

There is one additional point to note. Uncertainties on derived quantities calculated using the new error model will not necessarily increase in all situations. When differences between grid box values are taken, the correlated component will cancel to a certain extent. Persistent biases, and biases arising due to generic means of measuring SST are dealt with in part 2 of the paper.

The gridded SST anomalies, time series and uncertainties are available via http://www.metoffice.gov.uk/hadobs

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. Ian Jolliffe made helpful comments on the notation of the error analysis. The authors would also like to thank the three anonymous reviewers for their insightful and constructive comments, which improved the manuscript.

References

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850, *Journal of Geophysical Research*, *111*(D12106), doi:10.1029/2005JD006548.

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.

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.

Jones, P., T. Osborn, and K. Briffa (1997), Estimating sampling errors in large-scale temperature averages, *Journal of Climate*, *10*(10), 2548–2568, doi:10.1175/1520-0442(1997)010¡2548:ESEILS¿2.0.CO;2.

Kagan, R. (1966), An evaluation of the representativeness of precipitation data (in Russian), *Gidrometeoizdat*, p. 191.

Kennedy, J., N. Rayner, R. Smith, M. Saunby, and D. Parker (2011a), Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, part 2: biases and homogenisation, *JGR Atmospheres*.

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

Kent, E., and D. Berry (2008), Assessment of the marine observing system (ASMOS): final report, *Tech. Rep. 32*, National Oceanography Centre Southampton.

Kent, E., and P. Challenor (2006), Toward estimating climatic trends in SST. Part II: Random errors, *Journal of Atmospheric and Oceanic Technology*, *23*(3), 476–486, doi:10.1175/JTECH1844.1.

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.

Lumpkin, R., and M. Pazos (2007), *Measuring surface currents with Surface Velocity Program drifters: the instrument, its data, and some recent results*, Cambridge University Press.

Morrissey, M., and J. Greene (2009), A theoretical framework for the sampling error variance for three-dimensional climate averages of ICOADS monthly ship data, *Theoretical and Applied Climatology*, *96*(3-4), 235–248, doi:10.1007/s00704-008-0027-3.

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.

Shen, S., H. Yin, and T. Smith (2007), An estimate of the sampling error variance of the gridded GHCN monthly surface air temperature data, *Journal of Climate*, *20*(10), 2321–2331, doi:10.1175/JCLI4121.1.

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., 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.

Wilcox, R. (2001), *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, New York: Springer.

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.

Yevjevich, V. (1972), Probability and statistics in hydrology, *Water resources publications*, p. 302.

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