Toward Estimating Climatic Trends in SST. Part II: Random Errors

ELIZABETH C. KENT AND PETER G. CHALLENOR

National Oceanography Centre, Southampton, United Kingdom

(Manuscript received 24 February 2004, in final form 15 July 2005)

ABSTRACT

Random observational errors for sea surface temperature (SST) are estimated using merchant ship reports from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) for the period of 1970–97. A statistical technique, semivariogram analysis, is used to isolate the variance resulting from the observational error from that resulting from the spatial variability in a dataset of the differences of paired SST reports. The method is largely successful, although there is some evidence that in high-variability regions the separation of random and spatial error is not complete, which may have led to an overestimate of the random observational error in these regions. The error estimates are robust to changes in the details of the regression method used to estimate the spatial variability.

The resulting error estimates are shown to vary with region, time, the quality control applied, the method of measurement, the recruiting country, and the source of the data. SST data measured using buckets typically contain smaller random errors than those measured using an engine-intake thermometer. Errors are larger in the 1970s, probably because of problems with data transmission in the early days of the Global Telecommunications System. The best estimate of the global average random error in ICOADS ship SST for the period of 1970–97 is 1.2°C if the estimates are weighted by ocean area and 1.3°C if the estimates are weighted by the number of observations.

1. Introduction

Estimates of random errors in measurements of marine surface meteorological variables are important for generating and understanding fields of these variables. We define random errors as the unpredictable component of repeated independent measurements of the same parameter. In this definition we include those errors that vary from ship to ship (such as instrument calibration errors) and are systematic errors for an individual ship, but random errors for an ensemble of ships. Information about random errors is necessary for the construction of climatological fields (e.g., Rayner et al. 2003; Kaplan et al. 1998), data assimilation (e.g., Yu and O'Brien 1995), the generation of model forcing fields (e.g., Reynolds et al. 2002; Smith and Reynolds 2003), and for analysis of the bias in the variables themselves (Kent and Kaplan 2006, hereafter Part III).

Estimates of random errors in SST have previously been made by Gleckler and Weare (1997), Wilkerson and Earle (1990), and Kent et al. (1999, hereafter KCT). Previous estimates are reviewed by KCT who used the semivariogram method of Lindau (1995, 2003) to estimate random errors within the Comprehensive Ocean Atmosphere Data Set (COADS; Woodruff et al. 1998) for four sample months. Reynolds et al. (2002) use random error estimates of 1.3°C for ship SST, 0.5°C for both drifting and moored buoy SST, and 0.5° and 0.3°C for day and nighttime Advanced Very High Resolution Radiometer (AVHRR) satellite retrievals, respectively, in their optimal interpolation of SST to give gridded weekly $1° \times 1°$ area fields.

This paper will extend the error estimates of KCT using the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; Diaz et al. 2002) and all data from 1970 to 1997. The choice of this period is because of the availability of metadata giving information about the methods of measurement described in Kent and Taylor (2006, hereafter Part I). This will enable us to identify SST reports that are made by particular methods, in particular using buckets or engine intakes. Part III describes the use of the derived random errors in a statistical method to assess biases re-

Corresponding author address: Dr. Elizabeth C. Kent, National Oceanography Centre, European Way, Southampton SO14 3ZH, United Kingdom.
E-mail: Elizabeth.C.Kent@noc.soton.ac.uk

sulting from heat loss on bucket SST by comparison with engine-intake SSTs.

2. Data and quality assurance

a. ICOADS data and WMO metadata

This study uses the ICOADS dataset (Diaz et al. 2002) in combination with metadata from the World Meteorological Organization (e.g., WMO 1997), described in detail by Part I. ICOADS collates reports of surface meteorological variables from a range of different sources, but this study describes only the error characteristics of SST from voluntary observing ships (VOSs). Methods of SST measurement, which will be shown to have a significant impact on data quality, are derived both from the ICOADS "SST Indicator (SI) flag" and from the WMO metadata.

b. Quality assurance

1) ICOADS DUPLICATE ELIMINATION AND TRIMMING

ICOADS contains a "platform type" identifier that we have used to restrict our analysis to ship data rather than include other data, such as those from buoys or island stations. An important part of the ICOADS quality assurance (QA) is the removal of erroneous and duplicate reports (Woodruff et al. 1998). Because ICOADS contains data from many sources, reports received from more than one source need to be identified and the report that is expected to be of the best quality is selected. ICOADS includes the source of the data using a deck identifier and a source identifier. The "deck" identifier (originally "punched card deck") indicates the source of the data, for example, "U.S. National Centers for Environmental Prediction (NCEP) ship data," "Russian Marine Meteorological Dataset (MORMET)," or "Australian." The source identifier gives further information such as a period of receipt or the original data format. Combinations of decks and sources that are known to contain erroneous data, or data that are available elsewhere with better quality or more completely, are first eliminated. The remaining data are checked for duplicates using matches between seven weather elements for data within the same 1° area box. True duplicates with all identical elements are easy to eliminate; the problem arises when elements differ either because of corruption or the source of the data. Tolerances are allowed for the matching of weather elements to account for differences in transmission method [such as truncation for the Global Telecommunications System (GTS) reports compared with logbook reports] or differences in conversion methodology. Any data identified as duplicate have the likely better quality report selected depending on the deck and source identifiers. Duplicates with exact matches in space, time, and ship call sign are also removed. For the period of 1980–89, nearly two-thirds of the reports were removed. Full documentation of the duplicate elimination procedure is available on the ICOADS Web site (http://icoads.noaa.gov/e-doc/).

ICOADS contains QA flags resulting from a "trimming" procedure (Wolter 1997). The QA is based on differences from climatological 2° monthly means. Flags indicate whether the observation is within 2.8, 3.5, or 4.5 climatological standard deviations from the climatological mean. Although the trimming QA removes much of the poor-quality data within ICOADS, it is known to remove good data (Wolter 1997) or retain bad data (KCT) in climatologically extreme months. An alternative method of ICOADS QA for SST is described by Smith and Reynolds (2003). This method allows the mean to vary rather than imposing a climatological monthly mean that should better cope with climatological extremes. However, flags for this QA scheme were not available in the version of ICOADS used in this study. In addition to the standard ICOADS QA described above, we have also tested some additional approaches to QA, described in the following two sections.

2) SHIP TRACKING

Before the SST values themselves are examined, we have checked the ship positions for consistency. This was done for each ship report with a valid call sign. Any reports without a call sign, or with a generic call sign such as "SHIP" are left unchecked. In 1970 14% of the ship observations have a valid call sign, rising to more than 60% by the mid-1970s. The percentage of valid call signs then decreases after 1978 to a minimum of 40% in 1981. After the mid-1980s between 85% and 90% of ship observations contain a valid call sign. To perform the tracking all of the data with a common call sign were extracted in date order. The data were checked in blocks; a block was defined as a succession of reports with no gap longer than a week. Using the latitude, longitude, and time of each report the ship's speed is calculated between successive reports, assuming a spherical earth. If the required speed between positions is greater than 100 km h^{-1} one of the reports is flagged as being mispositioned. It is necessary to identify the first good report in each block because if the first report is erroneous then subsequent good data could be thrown away. The method chosen uses the first 10 reports (or all data in the block if there are less than 10 reports) and calculates the number of valid speed

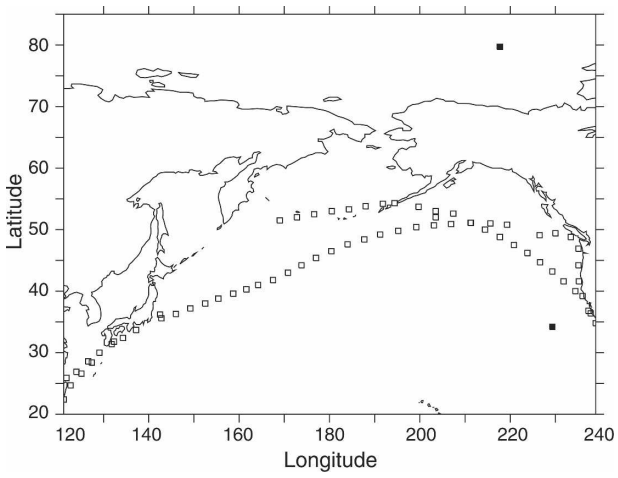

FIG. 1. Example of the tracking procedure. Reports for the *Sealand Trade*, call sign WTEI from January 1980. The open squares are those reports that passed the tracking check; the solid squares are those that failed.

comparisons within the block in turn for each of the 10 reports paired with each of the other reports. The first good report in each block is then chosen as the first report in the block that gives the maximum number of valid speed comparisons with the remaining data in the block. The rest of the block is then checked using this first good report as a starting point. The maximum allowed speed is relatively unimportant because most of the small percentage of reports that fail the track check are actually duplicates (with an identical time and call sign), and therefore have an infinite speed. The ICOADS duplicate elimination procedure only considers reports within the same 1° area as possible duplicates [see section 2b(1)]. While this is effective at removing some duplicates (such as the same GTS report coming from multiple sources or identifying a GTS report that should be replaced by a delayed-mode logbook report), when the position is corrupted the ICOADS check can fail. It was noted that often the report that contained a corrupted or misreported position also had other elements of the report that were corrupt. The ship tracking removes data that are often significantly different from the surrounding values. In addition, many of the reports identified as being mispositioned and therefore excluded from further analysis are close to, or actually on, land because ICOADS uses a fairly coarse 2° mask to identify landlocked points.

Figure 1 shows as an example, data from the ship *Sealand Trade* from January 1980. Marked as open squares are the positions of reports passing the track check; the two dark squares are those that failed. The southern mispositioned report has latitude reported as 34.2°N instead of 43.2°N; the latitude for the northern report has not been corrupted in an easily identifiable manner. In the northern report the dewpoint is corrupted. The ICOADS trimming has identified all of the variables in the northern report as erroneous, however for the southern report only air temperature is flagged as erroneous. For January 1980 overall 3860 reports are flagged as being mispositioned, of those with an SST report 27% pass the trimming check at the 4.5-sigma level. For air temperature the figure is 89%, for dewpoint and pressure 96%, and for wind speed 94%. The tracking procedure is therefore useful in identifying erroneous reports missed by the ICOADS-trimming QA; however, SST reports show the least improvement.

3) QUALITY ASSURANCE USING LOCAL MONTHLY MEAN

In performing QA for error analyses we wish to retain observations containing biases that result from the use of different types of instruments in different environmental conditions, along with the distribution of the random uncertainty associated with these biased data. The data to be excluded are those that result from misreporting, miscoding, transmission or keying errors, large biases resulting from poorly calibrated instruments, or any other gross errors.

An alternative QA scheme to trimming was performed globally on all reports passing the track check in the following way. First, all reports were binned into 30°-wide latitudinal bands and any reports within the band falling outside 4.5 standard deviations of their band mean were flagged as erroneous. This was necessary to remove mispositioned reports in the Arctic regions not yet removed by the track checking (possibly because the report does not contain a call sign). Wide latitude bands were required because erroneous data can be in the majority in data-sparse regions. The process was repeated, recalculating the standard deviation without the extreme points until no data were removed. The next pass iteratively removes for each band all data outside 4.5 standard deviations of the mean of the 10°-wide latitudinal band. The third pass removes all data outside 4.5 standard deviations of the local 10° × 10° area mean. The biases from individual ships were then examined to identify ships that consistently made reports significantly different from those of its neighbors. Any ship that within a particular month had more than half of its reports outside 3 standard deviations of the local mean had all of its reports for that variable flagged as invalid. The removal of data for the three differently sized areas was then repeated. The QA is included as a difference of the report from the local 10° area monthly mean normalized by the local monthly standard deviation. Data can therefore be excluded with varying cri-

teria as required. The sizes of the grids on which the quality control was performed was chosen to allow all reports, even those in very data-sparse regions, to receive some QA. More passes with smaller grids could have been performed in some regions with a high data density, such as the northern midlatitudes, but it was decided to keep the criteria broadly similar for all but the most data-sparse regions.

3. Calculation of random error estimates

a. The semivariogram method

KCT, following Lindau (1995, 2003), use the semivariogram method to estimate errors in surface meteorological variables from VOSs. The semivariogram method attempts to remove spatial variability from error estimates calculated from differences of individual reports at known separation. The VOS reports are first paired to give a dataset containing reports made at the same observation hour and at separations of 300 km or less. The differences between these nearby observations are partly made up of real spatial and temporal differences between the reports, and partly from errors in each observation. The variation of the squared observational differences with distance is used to try to isolate these two contributions, and hence estimate typical differences between the observations at zero separation—the observational error. KCT showed that the increase in the spatial component of the error was approximately linear up to separations of 300 km. In the following sections we therefore describe methods of estimating the error at zero separation using data pairs at separations below 300 km although the effect of using alternative maximum separations was tested.

KCT performed a simple linear regression on individual data pairs within 300-km separation. However, for a least squares linear regression to be statistically valid the distribution of data points around the regression line should be approximately Gaussian. We have a distribution of squared differences with large peaks at squared integer differences resulting from preferential reporting (Part I). The least squares regression is not therefore appropriate for these data. KCT found that their error estimates were, on occasion, significantly affected by small numbers of outliers. In an attempt to reduce the influence of outliers the semivariogram analysis was performed using a generalized linear model with a gamma function error distribution (McCullagh and Nelder 1989). The gamma function error distribution allows a longer tail of large squared differences and is closer to the observed distribution of squared differences than the normal distribution assumed when performing least squares regression. The residuals however still fail tests for goodness of fit following Yan et al. (2002), partly because of the truncation of the observed distribution at zero and partly because of preferential reporting. An alternative to these two methods of performing the semivariogram regression calculates a regression following the averaging of the data into distance bins (Lindau 2003). The central limit theorem states that the distribution of an average tends to be normal, even when the distribution from which the average is computed is nonnormal. This can be used to justify the use of least squares regression on the mean squared differences, even though the squared differences themselves have a distribution that cannot be simply described by a standard statistical distribution. The results for all three types of regression are however very similar. Mean differences for the SST error estimates calculated for 30° areas, using data pairs closer than 300 km, are less than 0.01°C, and the standard deviation of the differences is less than 0.1°C. It was decided therefore to perform the regression analysis using mean squared differences binned in distance ranges because this is the most statistically sound method.

b. Choice of regression

Having decided from statistical considerations to use a binned regression we need to further determine the maximum separation to use and the number of bins. It was thought that, especially in regions of localized high variability, the choice of maximum separation might be important. However, maximum separations ranging from 180 to 300 km were tested with little difference in results, even in high-variability regions. Use of smaller maximum separations resulted in fewer valid error estimates, and so 300 km was used as a maximum separation, as in KCT. Changing the binning interval between 10 and 30 km also had almost no impact on the error estimates: 20 km has been chosen for this study.

c. Choice of quality assurance

How the data are quality assured will obviously have a strong impact on the errors we calculate. We can therefore use the error estimates to guide our choice of QA criteria. The best criteria will produce a small ratio of estimated random error to the quantity of data accepted. The methods tested are the ICOADS trimming with levels of 4.5, 3.5, and 2.8 standard deviations; local quality control with 5, 3, and 1 standard deviation limits [section 2b(3)]; and the use of ship tracking to remove mispositioned reports [section 2b(2)].

Compared with the ICOADS 3.5-sigma-trimming level the 4.5-sigma-trimming level included 3% extra

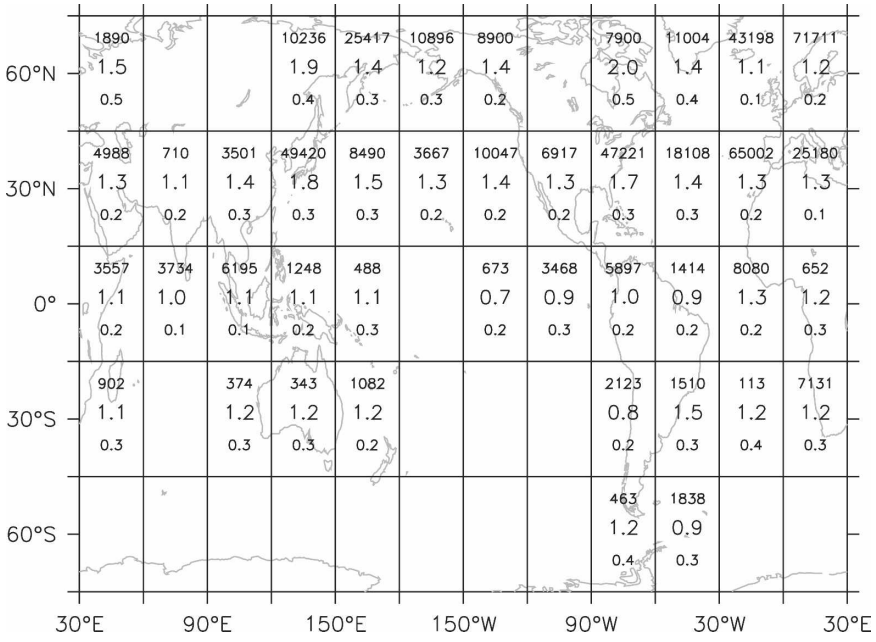

FIG. 2. Random error estimates for SST (°C) calculated using the semivariogram method. Errors have been calculated each month for 30° areas and averaged over the period of 1970–97 where there were 36 or more monthly values (middle, large-sized number). Also plotted is the monthly average number of observation pairs used in the analysis (top number) and the std dev of the monthly values (lower number).

data and the estimated error increased by 8%. The 2.8-sigma-trimming level excluded 5% extra data and the estimated error reduced by 8%. The local quality control performs less well than does the ICOADS trimming. Removing data outside 3 standard deviations of the local mean results in similar amounts of data to that of the ICOADS 3.5-sigma-trimming level, but the resulting error estimates are similar to the ICOADS 4.5-sigma-trimming level. Tracking removes about 3% of the data but makes very little difference to the error estimates. This is because some of the data that are removed by the ship tracking are mispositioned duplicates. Any of these reports that fall within the 300-km limit will act to artificially reduce the error estimate; those falling outside 300 km will increase the error estimate.

We will use the ICOADS 3.5-sigma-trimming flags for our QA and exclude mispositioned data using the tracking flag. Random errors for the 4.5- and 2.8-sigma levels can be estimated by increasing or reducing the error estimates presented by 8%, respectively.

4. Results: SST data quality characteristics

a. Estimates of SST random error

Figure 2 shows monthly error estimates calculated for all ship data in 30° areas averaged over the period of 1970–97. The error estimates are typically slightly smaller than those of KCT, and as expected from the much larger quantity of data analyzed, are more smoothly varying and globally complete. As in KCT the error estimates are larger in the midlatitudes than in the Tropics and the largest error estimates are in the Arctic and the Gulf Stream and Kuroshio regions.

Figure 3 shows the average of northwestern North Atlantic monthly error estimates calculated in the same way as those in Fig. 2, but for 2° regions. This clearly shows that the largest error estimates are associated with high variability. The most likely explanation is that the semivariogram method has not removed all of the spatial variability from the error estimates. The method used to calculate the estimates makes no allowance for different spatial variation in different directions, and this will be significant in boundary current regions. It seems likely therefore that the error estimates presented are an upper limit of the true random error in an individual SST observation and in some regions may have been increased by an element of spatial variability.

b. Time variation of random errors

Figure 4 shows how the error estimates vary with time. Two estimates of the global error are shown: the

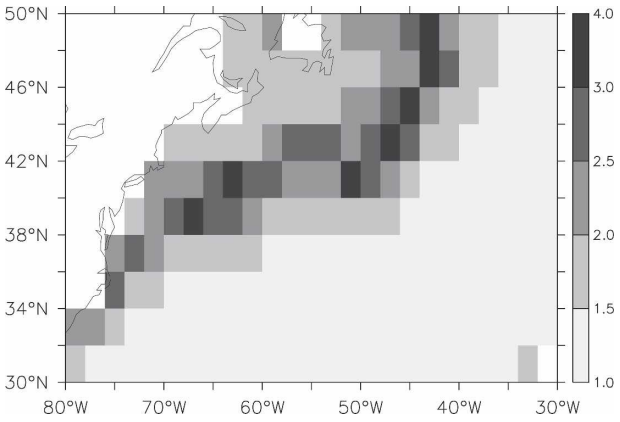

FIG. 3. SST random error estimates (°C) calculated as in Fig. 2, but for 2° areas in the North Atlantic.

error estimates weighted by the number of observations in each 30° region and month, and the error estimates weighted by the ocean area in each 30° region. It has only been possible to use the error estimates between 45°S and 75°N. Weighting is by the actual number of observations in each 30° region in each month rather than the number of data pairs used in the calculation of the error estimates. Globally representative average errors have been calculated in the following way. Small temporal gaps of 1 month were filled by linear interpolation in time. Remaining missing error values in each 30° region were then replaced with a combination of the 12-month running mean and annual cycle (calculated from all available data in the region) of the error estimates for that 30° region. Any remaining missing values were in extremely data-sparse regions where a long-term mean could not be estimated and were filled by zonal linear interpolation. The resulting complete fields of error estimates were weighted to give either a value that represents the observational density or the ocean area. For an error estimate representative of the observational density the interpolated complete fields were weighted by the monthly mean number of ICOADS VOS SST observations within each 30° area (note that this is not the number observational pairs shown in Fig. 2). To obtain a geographically representative error estimate the same interpolated fields were weighted by the fraction of ocean area, calculated from the ETOPO5 dataset (National Geophysical Data Center 1993) within each 30° area.

In northern high and midlatitudes there is a seasonal cycle in the random error estimates, which is more obvious in the global estimate when the observations are weighted by the number of observations. Errors are larger in the Northern Hemisphere summer than in the winter, but the size of the annual signal varies throughout the period, with the amplitude being largest toward the end of the period analyzed. The analysis cannot determine the reasons for the larger random error in summer observations. If a global average error for the period of 1970–97 and the region between 45°S and 75°N is calculated, including weighting for the number of observations in each 30° area, the average random error is 1.3°C with a standard deviation of 0.3°C. When the weighting is by ocean area the global average random error is 1.2 ± 0.4°C, because of the lower errors in the Tropics compared with high latitudes. KCT estimated the global error in VOS SST to be 1.5 ± 0.1°C. Reynolds et al. (2002) use a value of 1.3°C in their global analysis.

The change in the long-term mean with time can be related to the sources of the data making up ICOADS over the period. Figure 5 shows how the data in ICOADS are broken down by deck [see section 2b(1)].

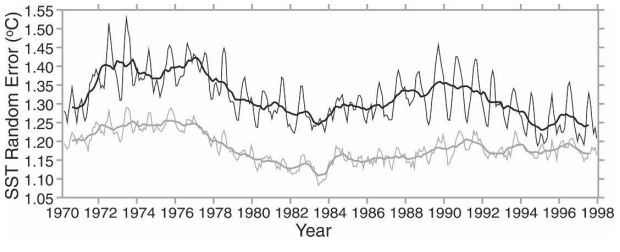

FIG. 4. Time series of average SST monthly random error estimates (°C) for the region from 45°S to 75°N. Average of 30° area error estimates weighted by number of observations (black line) and weighted by the ocean area (gray line). The thick lines have been smoothed with a 12-month running mean and the thinner lines with a 3-month running mean.

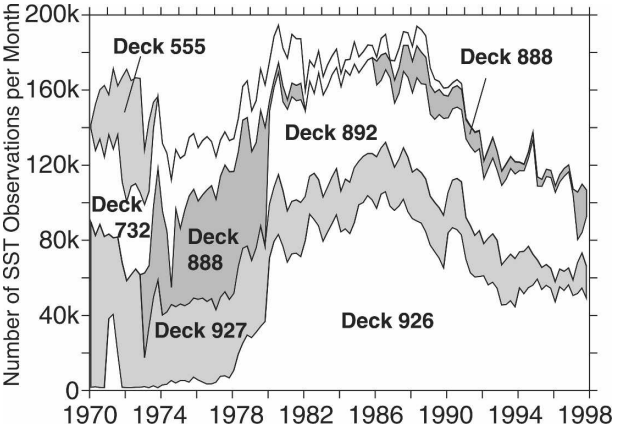

FIG. 5. The main sources of data within ICOADS for the period of 1970–97. The number of SST reports from the six main decks contributing during this period have been plotted cumulatively. The decks are as follows: 555, U.S. Navy FNMOC; 732, Russian MORMET; 888, U.S. Air Force GWC; 926, IMM; 927, keyed logbook data; and 892, NCEP data.

The errors calculated for each deck vary with time. In the early 1970s the SST data are mainly keyed (deck 927), but with significant contributions from Russian data (MORMET, deck 732) and the U.S. Navy Fleet Numerical Meteorology and Oceanography Center (FNMOC; deck 555). All of these data contain similar random errors that are typical of the overall figure for this period. Data from U.S. Air Force Global Weather Central (GWC; deck 888) is present in large quantities between 1974 and 1980 and is of relatively poor quality, with typical random errors being about 15% greater than average. In March, April, and May 1973 the SST data from GWC are of very poor quality, probably resulting from errors in coding or data transmission. However, QA removes most of these gross errors. Both the quantity and quality of the keyed data (deck 927) decrease with time: typical random errors for the 1970s are of average quality, for the 1980s errors are 30% greater than average, rising to 50% greater in the 1990s. International Marine Meteorological data (IMM; deck 926) become the most common source of data in 1980. It is of good quality with a typical random error that is 20% smaller than the average, which reduces the overall error. NCEP data (deck 892) appear in 1980 and form a substantial part of the data from then on; it is generally of a poorer quality than IMM data, with typical random errors 10% greater than average. There is some evidence that data from NCEP improves in quality toward the end of the period analyzed.

c. SST error estimates by observation method

Figure 6a shows error estimates similar to those in Fig. 2, but calculated only for pairs of SST reports where both ships used buckets. It should be noted that there are a substantial number of observations with unknown methods, particularly in the early period (Part I), and these will have been excluded from the analysis. The monthly mean random errors for ships reporting bucket SSTs are in all cases smaller than, or the same as, the estimates calculated using all of the data. Again, increased errors are seen in high-variability regions and the lowest error estimates are in the Tropics. The errors calculated using only reports from engine intakes are shown in Fig. 6b. The engine-intake error estimates are larger than those from buckets and larger than, or the same as, those calculated using all data.

Figure 7 shows how the estimated random errors in bucket and engine-intake SSTs vary with time. The difference in the mean error estimates for the two different methods is clear. For the estimates weighted by the ocean area (black lines) the bucket random errors are 1.1° ± 0.3°C, and those for the engine intake are 1.3° ± 0.4°C. When the estimates are weighted by the number of observations (gray lines) the bucket random errors are 1.1° ± 0.3°C, and those for the engine intake are 1.4° ± 0.3°C. The problem with the GTS reports in the 1970s identified in the previous section is shown to largely affect the bucket SST reports; however, the random errors are still smaller than those for the engine intakes. There is some evidence that the bucket SSTs are worsening in quality with time, although they are still of a better quality than the engine-intake SSTs. As for all of the data (Fig. 4), there is a seasonal cycle in the error estimates both for buckets and engine-intake SST. This seasonal cycle is larger for the bucket SST than for the engine-intake SST, probably because of the shallower measurement depth for the bucket observations.

d. Data quality by country

Error estimates were made for SST reports from different recruiting countries. However, there were insufficient data to make reliable error estimates with pairs containing data from only one country. A different approach was therefore taken to assess the quality of data for each country, using the estimate of data quality described in section 2b(3). A paired file of collocated reports was generated where the reports were from the same reporting hour and separated by 50-km separation or less. For this analysis for which spatial variability is not accounted we need to use a separation smaller than the 300 km used for the semivariogram analysis. Each pair contained one report from a ship reporting bucket SST and one from a ship reporting engine-intake SST. The data chosen covered the period of 1990–94. The accuracy of a report was assessed using the difference of each data point from the monthly mean of the 10° area normalized by the standard deviation of the data in that area in that month. The country of origin of each report was determined either from the ICOADS country code where available, or estimated from the call sign where the country code was absent. Figure 8 shows examples, for a selection of recruiting countries, of histograms of normalized differences separately for the bucket and engine-intake report. There are significant differences between the qualities of SST reports from different countries. In general, SST reports from northern European countries have the narrowest range of normalized differences and are therefore of the best quality, for example, France, the Netherlands, and the United Kingdom are shown. Reports from Japan and the United States are more scattered but are present in significant quantities and are therefore still valuable. Still more scattered are the SST reports for which a recruiting country cannot be identified (not shown). The histogram for Russia

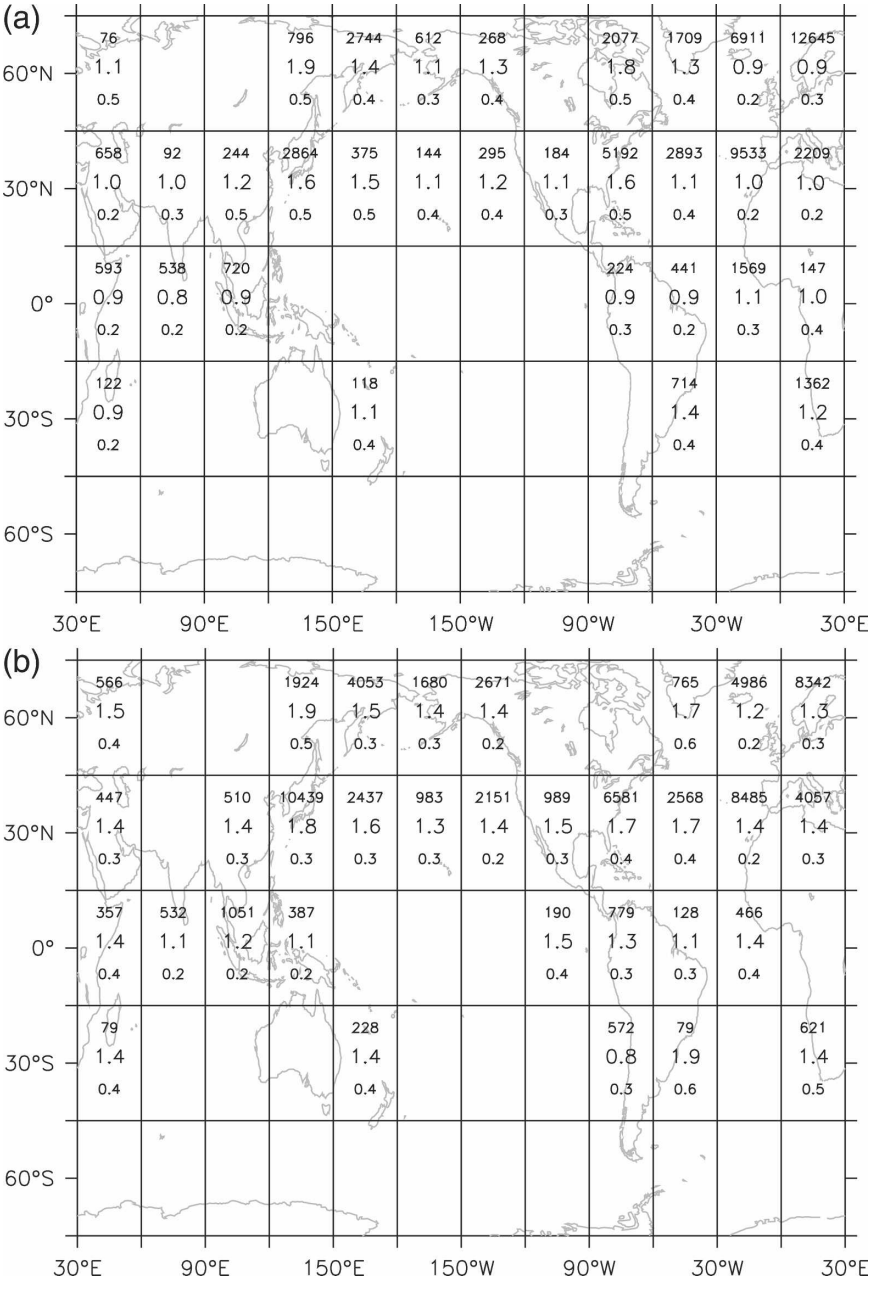

FIG. 6. (a) SST random error estimates (°C) calculated as in Fig. 2, but for bucket SST reports only. (b) SST random error estimates (°C) calculated as in Fig. 2, but for engine-intake SST reports only.

shows an interesting feature: the Russian bucket reports are biased noticeably colder than the engine-intake reports, perhaps showing cooling of the buckets in the high-heat-loss regions in which the Russian ships often operate. This is also shown to a lesser extent by the Netherlands bucket data. Analysis of systematic errors is described in Part III. Figure 8 shows that the difference between the quality of SST observations is more dependent on the recruiting country than on the method of measurement. For example, bucket reports from the United States are more scattered than engine-intake reports from France.

5. Discussion and conclusions

Statistical estimates have been made of the random errors in SST observations from ICOADS for the pe-

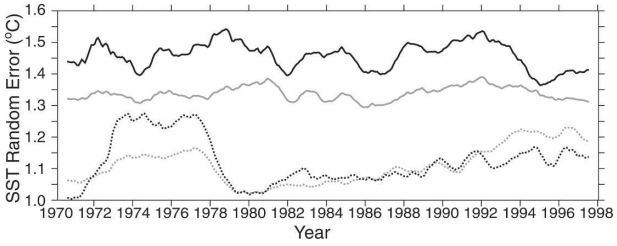

FIG. 7. Time series of bucket and engine-intake random errors as in Fig. 4 with a 12-month running mean filter. Solid lines are random errors from the engine-intake SST, and dotted lines are random errors for the bucket SST. Estimates weighted by the number of observations are shown by the dark lines and those weighted by ocean area are shown as the gray lines.

riod of 1970–97. The error estimates vary with space and time, but the global average random error in SST is estimated to be 1.2° ± 0.4°C if weighted by ocean area and 1.3° ± 0.3°C when weighted by the number of observations. The largest random error estimates are in high-variability regions and the lowest in the Tropics. This suggests that the semivariogram method may not be fully separating the spatial and random components of the SST variability. If this is the case, the error estimates presented may overestimate the random errors in SST. The global average monthly mean SST error estimate shows variation on times scales greater than a year, which can be related to the changes in the sources of data making up ICOADS. SST data quality, particularly bucket SST, in the mid-1970s is degraded by poor-quality GTS data. Data from IMM produce the smallest error estimates. The size of the random errors depends on the QA chosen. The errors quoted are for trimming limits of 3.5 sigma. Errors for 2.8-sigma limits are approximately 8% smaller and those for 4.5-sigma limits approximately 8% larger.

The bucket reports are much more consistent than the engine-intake reports. Bucket SST random errors however show more seasonal variability with larger random errors in the summer in high latitudes. It is not clear whether the decrease in consistency of bucket reports in the summer months is because of warming of buckets or the water sample on deck or to the formation of a diurnal warm layer increasing small-scale variability.

The variation of random errors by country suggests however that the variations in data quality depend on the care with which the measurement was taken rather than the method itself. When SST reports made by ships of a particular country are compared there are much smaller differences between the different types of observation for each country than between the different countries (Fig. 8). For example, both bucket and engine-intake SST reports made by ships recruited by the United States have a similar spread of errors as do those from ships recruited by France. However, the

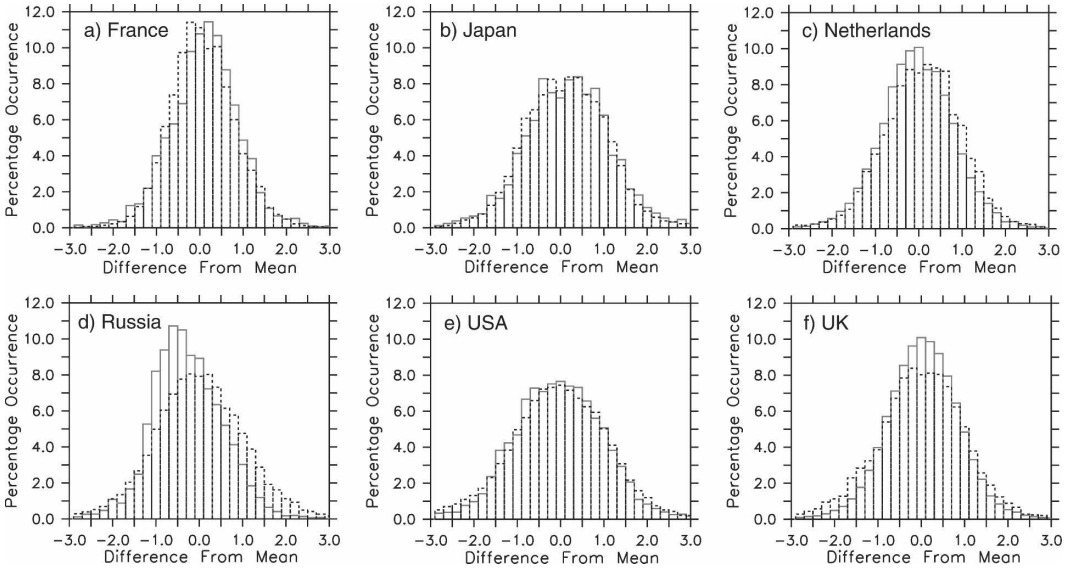

FIG. 8. Histograms of the deviation of SST measurements from the local 10° area monthly mean normalized by the local standard deviation for the period of 1990–94. The differences from the mean are therefore dimensionless. The gray bars represent reports from buckets, and the dashed bars represent reports from engine intakes. Histograms have been plotted separately for data from (a) France, (b) Japan, (c) the Netherlands, (d) Russia, (e) the United States, and (f) the United Kingdom.

French ships, regardless of method, have a narrower distribution of errors than the U.S. ships using either method. The difference in quality between bucket and engine-intake reports may therefore result from the preferred measurement methods of different countries. Countries, such as France, Germany, the Netherlands, and the United Kingdom, which have a national preference for bucket SST reports, also make good-quality engine-intake reports. Ships from Japan and the United States usually report engine-intake SST and their SST reports from both engine-intake and bucket methods are more scattered. However, it should be noted that reports from all of these countries are of better quality than those for which a recruiting country or measurement method cannot be identified.

Previous studies have suggested that engine-intake SSTs are of poor quality (Saur 1963; Walden 1966; Tauber 1969). However, the study of James and Fox (1972) showed that errors in engine-intake SST reports can be small if the observations are taken with precision thermometers positioned 3 m or less from the intake (see discussion in Part I). A high level of preferential reporting [where an observer preferentially rounds observations to integer, half-integer, or favored decimal digits (Part I)] degrades the data quality but does not affect the random error estimates. This study supports the conclusions of James and Fox (1972) that good-quality SST reports can be made using engine-intake thermometers, but that often the measurements are not made to the desired accuracy.

Changes in data quality in the period of 1970–97 seem therefore to be more closely related to changes in the data stream rather than changes in the observations themselves. Certainly, the largest changes in quality are related to problems arising after the observation has been taken. It should be noted that because of the more common use of dedicated SST sensors and automatic meteorological systems by the VOS there is likely to be an improvement in VOS data quality in the period after 1997.

Acknowledgments. We thank all three reviewers for their help in improving this paper. This work was supported by funding from the U.K. Government Meteorological Research Programme. The authors thank Steven Worley of the National Center for Atmospheric Research Data Support Section for providing the ICOADS and Scott Woodruff of the NOAA–CIRES Climate Diagnostics Center for help and advice on ICOADS. Peter K. Taylor and Zhongwei Yan of the Southampton Oceanography Centre provided help with the analysis. WMO metadata ASCII files were provided by Joe Elms at the National Climatic Data Center (Asheville, North Carolina) and the WMO. The Ferret program (a product of NOAA's Pacific Marine Environmental Laboratory; information available online at http://www.ferret.noaa.gov/) was used for some analysis and graphics in this paper.

REFERENCES

Diaz, H. F., C. K. Folland, T. Manabe, D. E. Parker, R. W. Reynolds, and S. D. Woodruff, 2002: Workshop on advances in the use of historical marine climate data (Boulder, Co., USA, 29th Jan–1st Feb 2002). *WMO Bull.,* **51,** 377–380.

Gleckler, P. J., and B. C. Weare, 1997: Uncertainties in global ocean surface heat flux climatologies derived from ship observations. *J. Climate,* **10,** 2764–2781.

James, R. W., and P. T. Fox, 1972: Comparative sea surface temperature measurements. World Meteorological Organization Reports on Marine Science Affairs, Rep. 5, WMO 336, 27 pp.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. Benno Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.,* **103** (C90), 18 567–18 589.

Kent, E. C., and A. Kaplan, 2006: Toward estimating climatic trends in SST data. Part III: Systematic biases. *J. Atmos. Oceanic Technol.,* **23,** 487–500.

——, and P. K. Taylor, 2006: Toward estimating climatic trends in SST data. Part I: Methods of measurement. *J. Atmos. Oceanic Technol.,* **23,** 464–475.

——, P. G. Challenor, and P. K. Taylor, 1999: A statistical determination of the random observational errors present in voluntary observing ships meteorological reports. *J. Atmos. Oceanic Technol.,* **16,** 905–914.

Lindau, R., 1995: A new Beaufort equivalent scale. *Proc. Int. COADS Winds Workshop*, Kiel, Germany, Institut für Meereskunde and NOAA Environmental Research Labs, 232–252.

——, 2003: Errors of Atlantic air-sea fluxes derived from ship observations. *J. Climate,* **16,** 783–788.

McCullagh, P., and J. A. Nelder, 1989: *General Linearised Models*. 2d ed. Chapman and Hall, 511 pp.

National Geophysical Data Center, 1993: Digital relief of the surface of the earth. NOAA Data Announcement 93-MGG-01, 1 pp.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of SST, sea ice and night marine air temperature since the late 19th century. *J. Geophys. Res.*, **108,** 4407, doi:10.1029/2002JD002670.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Q. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate,* **15,** 1609–1625.

Saur, J. F. T., 1963: A study of the quality of sea water temperatures reported in the logs of ships' weather observations. *J. Appl. Meteor.,* **2,** 417–425.

Smith, T. M., and R. W. Reynolds, 2003: Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J. Climate,* **16,** 1495–1510.

Tauber, G. M., 1969: The comparative measurements of sea surface temperature in the U.S.S.R. World Meteorological Organization Tech. Note 103, 151 pp.

Walden, H., 1966: On water temperature measurements aboard merchant vessels (in German). *Dtsch. Hydrogr. Z.,* **19,** 21–28.

Wilkerson, J. C., and M. D. Earle, 1990: A study of differences between environmental reports by ships in the Voluntary Observing Program and measurements from NOAA buoys. *J. Geophys. Res.,* **95** (C3), 3373–3385.

WMO, 1997: International list of selected, supplementary and auxiliary ships. WMO Rep. 47.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate,* **10,** 1980–1997.

Woodruff, S. D., H. F. Diaz, J. D. Elms, and S. J. Worley, 1998: COADS release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth,* **23,** 517–527.

Yan, Z., S. Bate, R. E. Chandler, V. Isham, and H. Wheater, 2002: An analysis of daily maximum wind speed in northwestern Europe using generalized linear models. *J. Climate,* **15,** 2073–2088.

Yu, L., and J. J. O'Brien, 1995: Variational data assimilation for determining the seasonal net surface heat flux using a tropical Pacific Ocean model. *J. Phys. Oceanogr.,* **25,** 2319–2343.