Bias Corrections for Historical Sea Surface Temperatures Based on Marine Air Temperatures

THOMAS M. SMITH AND RICHARD W. REYNOLDS

National Climatic Data Center, Asheville, North Carolina

(Manuscript received 8 January 2001, in final form 24 July 2001)

ABSTRACT

Because of changes in SST sampling methods in the 1940s and earlier, there are biases in the earlier period SSTs relative to the most recent 50 years. Published results from the Met Office have shown the need for historic bias correction and have developed several correction techniques. An independent bias-correction method is developed here from an analysis using nighttime marine air temperatures and SST observations from the Comprehensive Ocean–Atmosphere Data Set (COADS). Because this method is independent from methods proposed by the Met Office, the differences indicate uncertainties and similarities indicate where users may have more confidence in the bias correction.

The new method gives results that are broadly consistent with the latest Met Office bias estimates. However, this bias estimate has a stronger annual cycle of bias in the Northern Hemisphere in comparison with the Met Office estimate. Both estimates have midlatitude annual cycles, with the greatest bias in the cold season, and both have a small annual cycle in the Tropics. From the 1850s into the early twentieth century both bias estimates increase with time, although this estimate increases slightly less than the Met Office estimate over that period. Near-global average temperatures are not greatly affected by the choice of bias correction. However, the need for a bias correction in some periods may introduce greater uncertainty in the global averages. Differences in the bias corrections suggest that this bias-induced uncertainty in the near-global average may be 0.1°C in the nineteenth century, with less uncertainty in the early twentieth century.

1. Introduction

Recent estimates of global climate variations (e.g., Houghton et al. 1996) rely on historic estimates of surface temperature as part of their climate assessments. Because of the importance of the ocean in any global surface temperature estimate, sea surface temperatures (SSTs) provide an important contribution to these estimates. However, before merging SSTs and land temperatures, it is necessary to consider some of the biases in the SST data. Our goal here is to examine these biases with emphasis on long-term bias corrections required for in situ SST observations, and to evaluate bias estimates currently used by comparison to an independent estimate.

The longest dataset of SST observations is based on observations made from ships. These observations include measurements of SST alone as well as ocean temperature profiles over depth. However, the observations of SST alone dominate the datasets and account for more than 90% of the observations. Although the earliest observations were taken in the first half of the nineteenth century, sufficient observations to produce a global SST analysis were not available until the late nineteenth century. In addition to the changes in the number of observations, the method of measuring surface marine temperatures changed over the period from the use of uninsulated buckets to the use of insulated buckets, engine intakes, and hull-mounted sensors. Additional in situ SST observations from drifting and moored buoys began to be plentiful in the late 1970s. These observations are typically made by thermistor and usually relayed in real time by satellites. Although the accuracy of the buoy SST observations varies, the accuracies are usually better than 0.5°C, which is better than the accuracy of individual ship reports (Trenberth et al. 1992). In addition, typical depths of buoy observations are roughly 0.5 m rather than the 1-m and deeper depths from modern ships.

It is important to note that accurate SST retrievals from satellites became available in late 1981. Although these retrievals improved the data coverage over that from in situ observations alone, the retrievals have their own instrumental biases (e.g., Reynolds 1993). Because we are interested in long-term climate impacts of SST, we will consider only in situ data here.

In the results that follow, we first examine the historic bias corrections that have been computed at the Met Office. Using these corrections as our starting point, we

Corresponding author address: Dr. Thomas M. Smith, NOAA/NESDIS/NCDC, 151 Patton Avenue, Asheville, NC 28801-5001.
E-mail: tom.smith@noaa.gov

then discuss our own correction methods and the uncertainties of these corrections. We conclude with the impact that the uncertainties may have on global surface temperature.

2. U.K. SST bias corrections

The most important studies of the problem of bias in historic SSTs were performed at the Met Office. Among the published results are Folland et al. (1984), Bottomley et al. (1990), and Folland and Parker (1995). These studies show that before 1942, the global-average SST has a cold bias of between 0.1° and 0.4°C, with respect to the average SST after 1942. The Folland et al. (1984, hereinafter FPK84) bias correction is the simplest of the three. With each new paper, the adjustments were refined. In the last two papers, the bias corrections include models of the evaporative cooling of canvas and wooden buckets. The modeled bias was affected by variables such as the marine air temperature and both ship and wind speed. To properly use the models, it was necessary to estimate how the relative number of canvas and wooden buckets changed with time, as well as how typical ship speeds and deck heights changed with time. These assumptions lead to a comprehensive model for estimating SST bias.

In addition, Bottomley et al. (1990) and Parker et al. (1995) made adjustments to nighttime marine air temperatures (NMAT). They suggest several similar bias-correction schemes for NMAT prior to 1930 and during World War II. Beginning in the nineteenth century, ships gradually increased in size with time, with a corresponding increase in the height of the NMAT. Bottomley et al. (1990) used boundary layer theory to estimate bias induced by these height changes. During World War II nonstandard NMAT measurement practices were used, such as reading the thermometer inside to avoid showing a light on deck (Folland et al. 1984; S. Levitus 2000, personal communication). These practices caused positive bias in NMAT in the early 1940s. In Bottomley et al. (1990), four NMAT bias-correction schemes are described. All four are all similar and give the same general corrections, but details and some local corrections differ. The Bottomley et al. (1990) scheme D corrects for changes in deck heights, World War II practices, and also include some local corrections in the nineteenth century. We applied these adjustments to the Comprehensive Ocean–Atmosphere Data Set (COADS) NMAT to reduce the influence of NMAT bias on our SST bias estimates. In scheme D, some local adjustments for the nineteenth century use the corrected SST anomaly in regions where NMAT is less reliable. For those local adjustments we used the Folland and Parker (1995) SST corrections to adjust NMAT. Thus, for part of the nineteenth century our bias corrections will be influenced by the Folland and Parker (1995) corrections. As we show later, our results are consistent between the nineteenth and twentieth century, suggesting that the influence is small.

The final Met Office adjustments to SST included geographic and seasonal variations, which tended to be larger in extratropical latitudes. The Folland and Parker (1995, hereinafter FP95) bias model was well researched and was shown to give reasonable results based on comparisons of average SST and adjusted NMAT. However, it incorporates many assumptions to compute the bias correction.

The purpose of this study is to independently develop bias corrections using only the available marine air temperature and SST observations. These corrections, referred to as SR, are different from the most recent FP95 estimates because we do not employ models of heat loss from buckets and do not explicitly require assumptions about ship speeds or the types of buckets in use. As noted above, we do incorporate NMAT adjustments that are largest in the nineteenth century and the early 1940s.

Examination of the FP95 bias corrections shows that for each month 99.9% of the variance can be explained by one spatial and temporal empirical orthogonal function. We thus began our own study by defining one spatial and temporal function for each month. However, the SR estimates are developed using relationships between adjusted NMAT and all hours SST. We restricted ourselves to NMAT, as suggested by FP95, to eliminate daytime biases in marine air temperature due to heating of the ship deck. It is also possible for day SST and night SST to have a significant difference due to diurnal heating, which in an extreme case in the Tropics can be as much as several degrees centigrade in some light-wind regions (e.g., Weller and Anderson 1996). Therefore we also tested our method using night SST only, as described in section 4. We also limited ourselves to observations for which both NMAT and SST were present to ensure uniform sampling of both. We have chosen an empirical approach to be as independent as possible from FP95, to help to establish confidence in the historic bias corrections where the independent method yields consistent results.

3. Data

Surface marine observations are obtained from COADS (Slutz et al. 1985) and are available from 1854 to 1997. To compute biases only SST and adjusted NMAT observations are used, with night defined as 1900 to 0700 local time. This time is a compromise day and night definition for all seasons and latitudes. It is a rough estimate of day and night that is best in the Tropics, where day is about 12 h long all year. Near the poleward limits of our bias correction, about 60° latitude, the summer day is between 15 and 19 h long, so there may be some mixing of actual day values into what we call night values at high latitudes in summer. However, this contamination can only occur early or late in the day and does not cause high-latitude noise

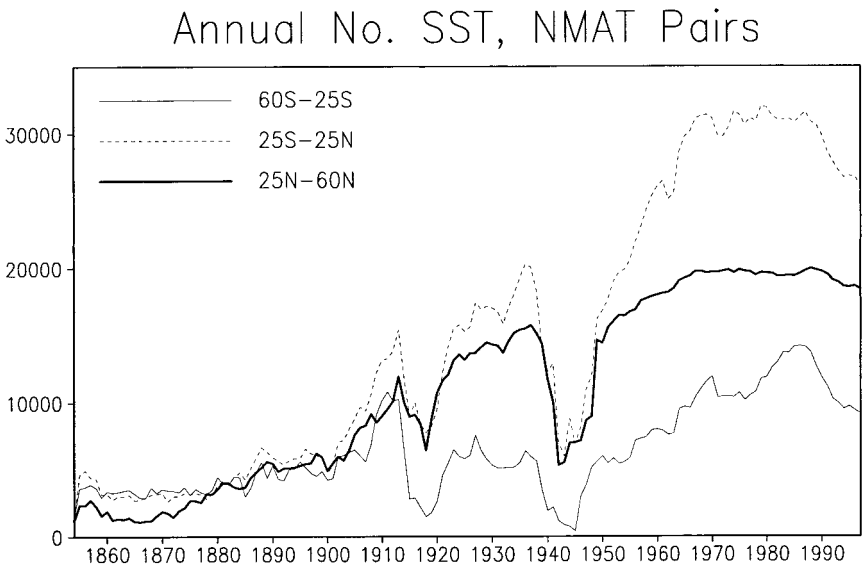


FIG. 1. The annual number of monthly 2° areas with both SST and NMAT observations averaged over 60°–25°S, 25°S–25°N, and 25°–60°N.

in our results. Our purpose was to eliminate the strongest NMAT diurnal heating signal. Thus, for simplicity we use a constant 12-h definition of day. The pairs of SST and NMAT temperatures are compared with the COADS climatological screening limits, which use sextiles and medians (Slutz et al. 1985) to remove bad observations. If either type of observation failed these tests, both were discarded. The remaining observations were averaged onto monthly SST and NMAT 2° gridded arrays (centered on 88°S–88°N and 0°–358°E) and saved with the number of observations averaged in each 2° area.

The total number of 2° pairs of SST and NMAT for each year (Fig. 1) shows that the in situ coverage generally increases with time except for the periods of the two world wars. There are very few data pairs before 1880, and the number of pairs decreases sharply in association with the two world wars, especially in the Southern Hemisphere. Sampling is most dense after 1950. As mentioned earlier, SSTs from drifting and moored buoys become more common after the late 1970s. However, drifting buoys do not include air temperature observations. Thus, only moored buoys are included in the results shown in the figure. The SSTs from temperature profile data (Levitus et al. 1998) are included in COADS. However, they were rarely used in our results because marine air temperature was usually not available with these reports.

To help to justify and verify the bias estimates, we use additional sources of data, which are discussed in the appropriate sections that follow. In particular, in the discussion section we utilize surface land-air temperatures to illustrate the impact of the SST bias corrections on global surface temperatures, combining land and ocean.

4. Methods

a. Analysis of SST–NMAT differences

In order to compute bias corrections we made several assumptions. We assumed that air–sea interactions on large time- and space scales were the same over historical periods as over our most recent period. Specifically, we assume that for each calendar month the relative shape of observed SST–NMAT differences is constant. Any changes in the magnitudes of the patterns were attributed to measurement or instrument changes, and these changes are assumed to affect only SST.

These assumptions are unlikely to be completely true. For example, Bottomley et al. (1990) note problems with NMAT, as discussed in section 2, and we apply their adjustments. Christy et al. (2001) found recent differences between the trends of marine air and SSTs exceeding 0.05°C decade^{-1}. Climate variations such as the North Atlantic oscillation may also cause changes in the SST–NMAT difference. We test these assumptions below to show that the computed spatial patterns of SST–NMAT are stable. Thus, most of the observed difference over the historical record should be related to SST bias. We will proceed with these assumptions with the realization that they introduce some uncertainty in our results.

If the difference, d, between SST and adjusted NMAT for an individual monthly 2° area situated at point x in year y and month m is defined as

$$d_{x,m,y} = \text{SST} - \text{NMAT}$$

then the large-scale climatic difference C for calendar month m is defined as

$$C_{x,m} = I\left(\int_Y \delta_{x,m,y} d_{x,m,y}\, dy \Big/ \int_Y \delta_{x,m,y}\, dy\right). \quad (1)$$

The variable $\delta_{x,m,y} = 1$ if the difference is defined at that spatial–temporal point; otherwise it is 0. The integration over a number of years Y indicates temporal averaging of the available observations. The smoothing operator I, defined below, fills in and smooths the value at point x using data within a local region. Following the results of FP95, we compute an annual cycle of C, which is held constant for all years.

The C patterns are computed by averaging over the most recently available 30-yr climate base period (1968–97), which has dense sampling (presently COADS ends with 1997). This is the temporal averaging in Eq. (1). By definition, this period has zero bias. We require for each month and in each 2° region, that at least 5 yr are defined in order to compute a time average. Data north of 70°N and south of 60°S are excluded from the computation of C because they are extremely sparse at those latitudes and may produce high-latitude noise, which could spread into other regions through the large-scale smoothing.

After the temporal averaging, the climatological d values are spatially smoothed and missing locations are filled using optimal interpolation (OI). This defines the spatial smoothing in Eq. (1). We use OI with Gaussian spatial correlation functions to damp the interpolation estimate at great distances from data, as in Reynolds and Smith (1994). Here the spatial scales are large because we wish to produce smoothed and complete fields. The zonal (meridional) scales used range from about 2000 (1400) km in low latitudes to 500 km at high latitudes. These scales are chosen subjectively, based on scales evident in the FP95 bias fields. Scales that are too small could produce small-scale features that cannot reliably be resolved by the pre-1950 observations, and that would make the difference field noisy. The asymmetry of the zonal and meridional scales at low latitudes reflects the tendency for zonal elongation of climate features at low latitudes.

For historic periods the SST–NMAT difference is estimated from the spatial pattern C, which is constant for each calendar month, scaled by a coefficient A. Estimation of the difference using a coefficient allows a smooth estimate to be computed for the pre-1950 period, when data are sparse. The best-fit coefficient for each time minimizes the global error of the estimate compared to the observations,

$$E^2 = \sum_x \delta_{x,m,y}(d_{x,m,y} - A_{m,y}C_{x,m})^2 a_x,$$

where the summation is performed over the 2° ocean grid boxes, each with an area a_x. The coefficient is computed by

$$A_{m,y} = \frac{\sum_x \delta_{x,m,y} d_{x,m,y} C_{x,m} a_x}{\sum_x \delta_{x,m,y} C_{x,m}^2 a_x}. \quad (2)$$

Thus we assume that the bias in SST has the same relative geographic pattern as the field of SST-adjusted NMAT. These coefficients can become unstable if too few data are available. To prevent that instability, a coefficient is not defined if less than 5% of the variance of C is sampled. The fraction of sampled variance is defined as in Smith et al. (1998),

$$f_v = \frac{\sum_x \delta_{x,m,y} C_{x,m}^2 a_x}{\sum_x C_{x,m}^2 a_x}.$$

b. Consideration of outliers

Because we are interested in the large-scale climatology of C, we exclude extreme values of d from our analysis. These extremes indicate either data errors or overrepresentation of intense synoptic episodes that may not be typical of the month as a whole. Even in the most densely sampled regions and periods, there are usually few observations per month for any given 2° region, so extreme events may have an unduly large influence on the monthly average. There may also be questionable values due to errors not eliminated by the COADS data screening that we employed.

The largest d extremes occur in winter, off the mid-latitude east coast of continents where very cold air may move over much warmer ocean waters (e.g., Trenberth et al. 1992). Weather events responsible for these large d values may have timescales of several days, and sampling during such an event would not be representative of the climatic monthly value of d. Negative d extremes are smaller, and tend to occur in the summer when warm continental air may move over cooler oceans.

Outliers are evaluated by examination of individual COADS SST and marine air temperature pairs for two periods, 1980–89 and 1930–39. For neither period are adjustments to NMAT required. Since we will fit data from the historic period to modern monthly climatological maps, our definition of outliers should be suitable for the historic periods as well as the modern period. Therefore, besides the modern SST–NMAT differences, the differences in the 1930s, a historic period when sampling is relatively high, are used to refine the definition of outliers.

To define outliers, we use all simultaneous nighttime differences for each calendar month and compute frequency distributions of the differences for each month in different regions. Examination of these frequency distributions shows that for all months practically all differences are between −3° and 8°C, with a few odd values having absolute values of 9°C or greater. These frequency distributions, excluding values with a magnitude

TABLE 1. Seasonal and regional 80% confidence intervals (°C) for individual nighttime SST–NMAT differences.

Season	55°–25°S	25°S–25°N	25°–55°N
DJF	[−1.1, 1.9]	[−0.3, 2.3]	[−1.3, 4.9]
MAM	[−0.8, 2.8]	[−0.5, 2.1]	[−0.9, 2.1]
JJA	[−0.9, 3.7]	[−0.2, 1.8]	[−1.5, 1.5]
SON	[−0.7, 2.3]	[−0.3, 2.3]	[−1.3, 3.3]

TABLE 2. Percentage of monthly 2° values discarded for $d < -2$ (Lo) and $d > 4.5$ (Hi) over three regions for the 1968–97 period.

	Lo	Hi
25°–60°N	1.4	5.3
25°S–25°N	0.3	0.3
60°–25°S	2.0	1.4

of 9°C or greater, are used to estimate the mean and standard deviation of the differences in the 1980s and the 1930s. The mean is estimated from the median. The standard deviation is estimated by assuming that the distribution is approximately normal near the median and finding the difference between the 69th and the 31st percentile, which is one standard deviation in a normal distribution. Using the median from the 1980s, we compute seasonal confidence intervals using the standard deviation estimates from the 1930s and the 1980s.

Our goal is to use Eq. (1) to compute a climatology, or monthly mean of differences, and then to fit historic data to that climatology. Therefore, it is important to include the center of the distribution, while the tails are less important. However, the differences may not be normally distributed so we should not be too restrictive. Using the 1980s mean and the 1930s standard deviations, the estimated 80% confidence intervals in most regions are in the range [−1.5, 4.5]°C (Table 1). The 1980s standard deviation estimate is about 10% larger that the 1930s standard deviation estimate, which would slightly expand these limits. We prefer the more restrictive limits suggested by the 1930s data, to avoid allowing outliers through in historic periods.

Because we require limits for screening historic differences, which include a cold SST bias, the lower limit needs to be adjusted to account for the shift in the mean caused by changes in sampling in the early 1940s. The difference between the means for the 1930s and the 1980s suggests that excluding data outside the range [−2, 4.5]°C should eliminate outliers while keeping most good data. For the 1968–97 period, this cutoff criterion leaves more than 90% of the 2° monthly data at all latitudes, for all months (Table 2). Extremes are most rare in the Tropics where >99% of the data pass the test. Most outliers are found in the Northern Hemisphere, where during winter arctic and continental subpolar air masses sometimes move over the warmer oceans. Differences associated with the most intense outbreaks may be damped by this screening, but as Table 2 shows, typical differences will be represented. In the Southern Hemisphere similar situations can occur when Antarctic air masses move over the extratropical oceans in winter, but this situation is less common and the cutoffs have a more modest effect. These cutoffs are used in the remainder of our study.

Annual percentages of monthly 2° differences excluded (Fig. 2) show that Table 2 is representative of most years. However, a larger percentage is discarded in some years. The percentage is larger before 1890, when data are sparse. However, the increase is not con-

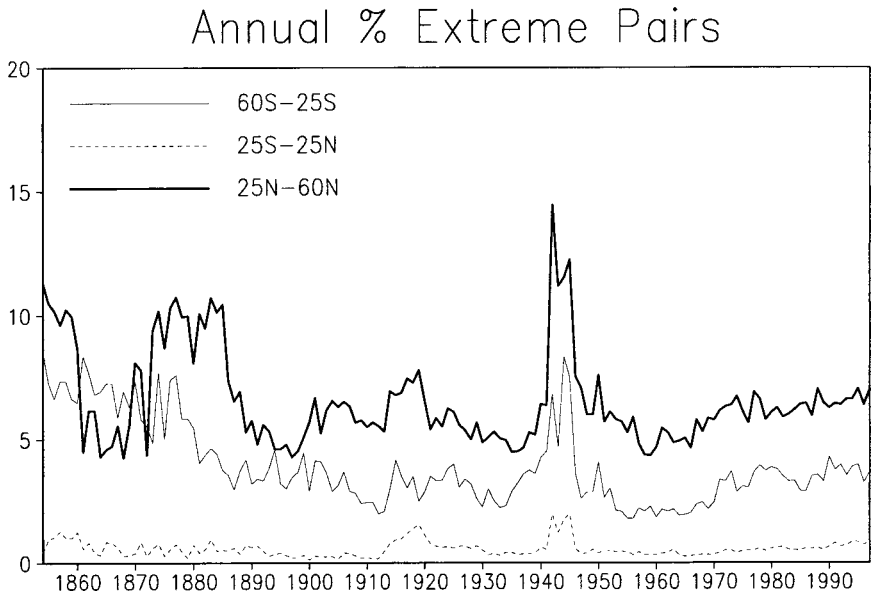


FIG. 2. The annual percent of monthly 2° area SST-adjusted NMAT differences excluded for being outside the range [−2, 4.5]°C, for the areas 60°–25°S, 25°S–25°N, and 25°–60°N.

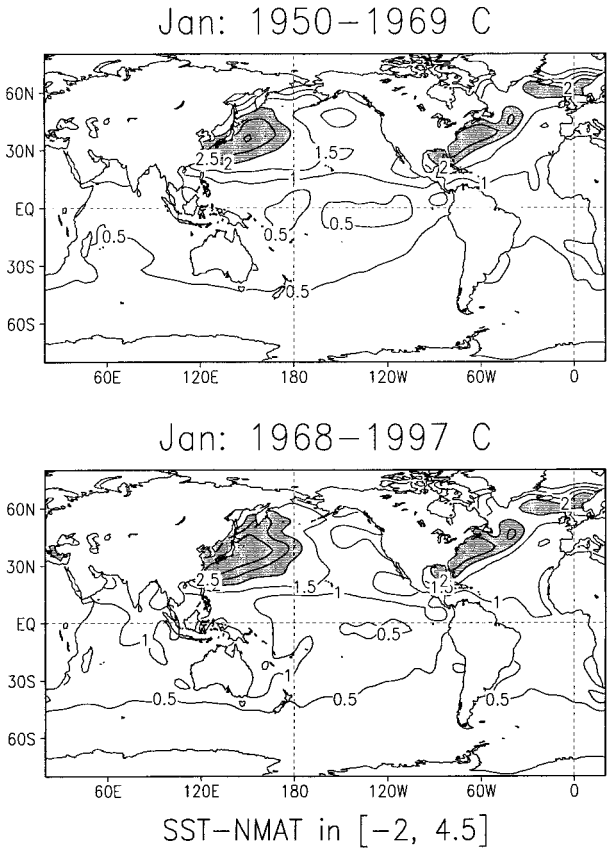


FIG. 3. Jan *C*, excluding differences outside the range [−2, 4.5]°C, computed using data from the periods (top) 1950–69 and (bottom) 1968–97.

sistent across all regions. In the early 1940s the percentages of excluded differences are greater in many regions, suggesting the influence of widespread bad data.

In Fig. 3 the January *C* patterns are shown for two periods, computed using data with the extremes removed. The two January patterns are similar, as are patterns from other months. The spatial correlation between the two maps is 0.96, indicating that they are essentially the same except for a scaling factor (see section 4d). Our *C* estimates for January have largest values in the extratropical Northern Hemisphere, especially in western boundary regions. The approach to the maxima is gradual, with no apparent truncation near the maxima. The maxima are about 1°C less than the data cutoff limit. Without excluding extreme values, the Northern Hemisphere winter maxima in western boundary regions would be slightly larger, which would cause slightly inflated bias estimates in those regions.

c. Bias correction estimates

Using (2) the monthly coefficients $A_{m,y}$ were computed. As shown in Fig. 4a, the coefficients for each month are similar. However, there is some spread among the 12 estimates, especially prior to 1900. Also, some months have insufficient sampling to define a coefficient for the first few years of the record. A smoother and more complete estimate of the coefficient is obtained by simultaneously fitting all months within a given year to obtain an annual coefficient,

$$A_y = \frac{\sum_{m=1}^{12} \sum_x \delta_{x,y,m} d_{x,m,y} C_{x,m} a_x}{\sum_{m=1}^{12} \sum_x \delta_{x,y,m} C_{x,m}^2 a_x}. \quad (3)$$

This annual coefficient is roughly the average of the 12 monthly coefficients (Fig. 4b), and if sampling were equal for every month in the year, it would be exactly the average of the 12 monthly coefficients. The 12-month coefficient estimate is more stable than individual monthly coefficients, as indicated by the figure. This is an important advantage over monthly coefficients when sampling is sparse. Thus, we use the annual coefficient estimate in the rest of this paper. Note that in Fig. 4 the coefficients are scaled by the annual-average *C* pattern, averaged between 60°S and 60°N, to give them units of °C. Thus, the relative difference between the average 1968–97 coefficient and its value at another time gives the size of the annual- and spatial-average SST bias for that time.

Because bias is primarily caused by systematic changes in sampling methods, changes in the coefficient that we relate to SST bias should have little variation on timescales shorter than decades except, as noted by Parker et al. (1995), during the 1940s. In addition, the results of Christy et al. (2001) suggest that differences of less than 0.1°C may not always be related to measurement bias. Thus, we decided to smooth the annual coefficient estimate using straight lines. One line is fit to values between 1854 and 1941 to minimize the mean-squared error of the fit. Because we assume that there is zero bias for the 1968–97 base period, we define the upper line as the average over that period with zero slope. Between 1941 and 1942, the lower smoothed line is forced to join the upper.

In the late 1930s a sharp change in the unsmoothed curve is evident. The FP95 bias is assumed to end after 1941, with zero bias used afterward. As shown by Woodruff et al. (1998), there are dramatic changes in data sources between 1941 and 1942, suggesting that this instantaneous end of the bias may be correct. Most SST data in the late 1930s are from Dutch, German, and Japanese sources. For 1940–41 most of the data is from Japanese sources, while most data from 1942–45 is from U.S. sources. The unsmoothed coefficient changes abruptly around 1940 and it stays below the smoothed curve through the 1950s. Thus, there may be a mix of biased and unbiased data in that period. In the future more work may be needed to better estimate bias between about 1940 and the late 1950s, especially if additional data sources become available.

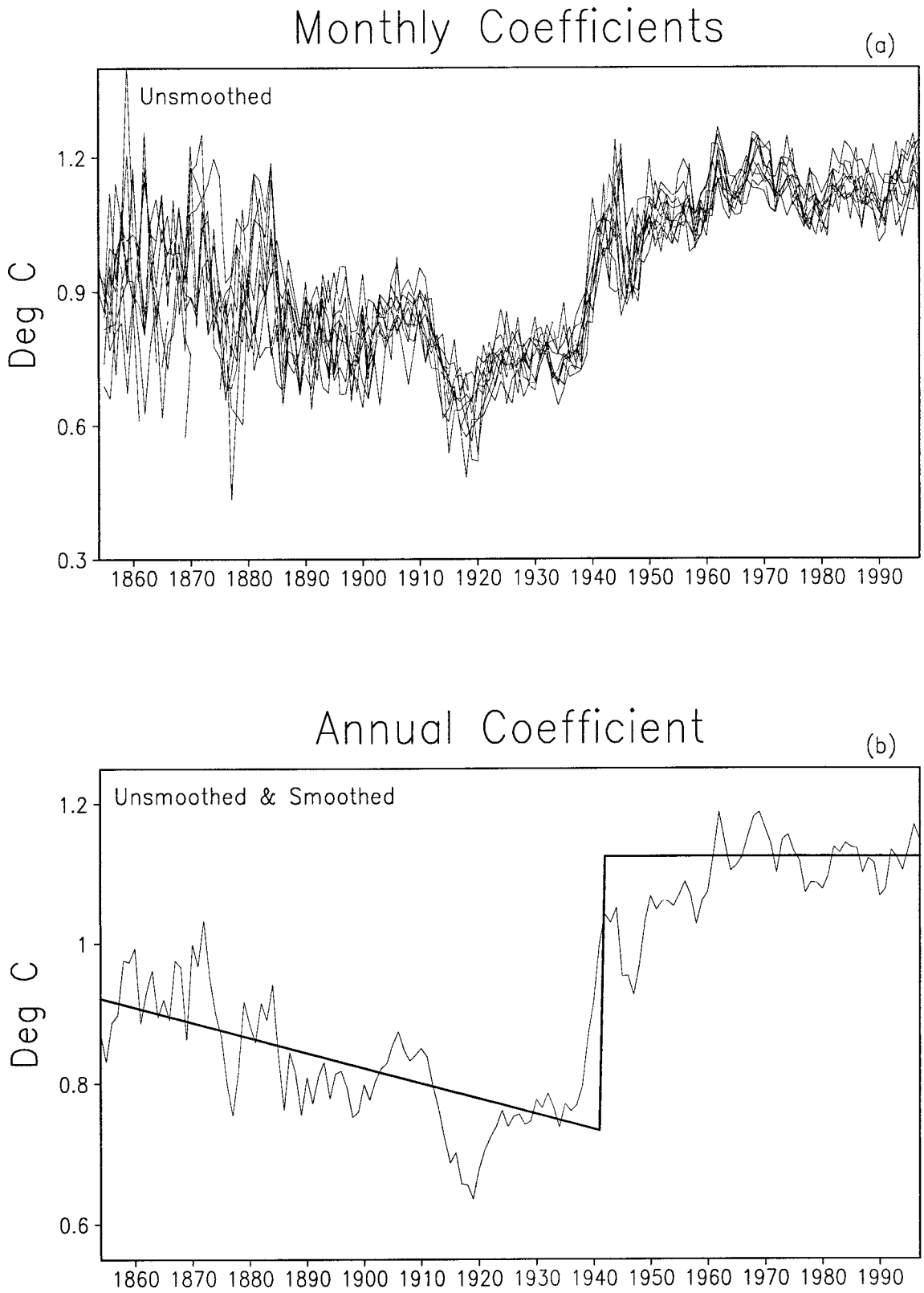


FIG. 4. (a) The 12 monthly coefficients, unsmoothed, and (b) the annual unsmoothed and smoothed coefficients. The coefficients are scaled by the annual and 60°S–60°N average of *C*, to give values with units of degrees Celsius.

We can now compute the historic SST bias correction. To do this we use the smoothed annual coefficient estimate, and define the SR bias correction as

$$B_{x,m,y} = (\overline{A} - A_y)C_{x,m},$$

where $\overline{A}$ is the average coefficient for our base period, 1968–97, which is used to define *C*. The correction *B* is added to the SST data to correct for measurement bias.

d. Evaluation of assumptions

Our major assumption is that the pattern of SST-adjusted NMAT for a given calender month does not change over time except in magnitude, and that long-term changes in the magnitude of this pattern are caused by biases in SST. Since heat loss from buckets is assumed to cause bias in the FP95 model, their patterns of bias are strongly correlated with SST-adjusted NMAT, but they are more similar to scaled maps of latent heat transfer from the ocean (e.g., Higgins et al. 1996). By contrast the *C* patterns that we used to evaluate bias are more similar to sensible heat transfer, which gives greater weighting to areas north of 40°N in the boreal winter than does latent heat transfer. For a given wind speed the sensible heat transfer is proportional to the air–sea temperature difference, so the similarity between patterns of sensible heat and *C* are natural. In reality, bias is caused by latent and sensible heat loss from buckets in the period when that type of sampling dominated. In addition, some bias is caused by the warming of engine-intake water samples in the more recent period, when that type of sampling dominates the record. We discuss the effect of this assumption and the differences from the FP95 bias in section 6.

We test the stability of *C* patterns by comparing them from two independent well-sampled periods, the 20-yr period 1950–69 and the 30-yr period 1968–97. As discussed above, the patterns computed from different periods are similar (Fig. 3). Similarity in shape is demonstrated by pattern correlation (Murphy and Epstein 1989), which for these two January patterns is 0.96. The pattern correlations for all other months are 0.89 or higher. For two completely independent 20-yr periods, 1950–69 and 1978–97, the pattern correlations are 0.87 or higher for all months. Correlations are highest in November–February, where they are 0.95 or greater, and lowest in August–September. Thus, for the independent periods the patterns have essentially the same shape for all months. Slight changes in the magnitudes are not important, since for a given pattern the computed coefficient *A* properly scales the patterns. Thus, the *C* patterns computed from either period would yield similar results.

The assumption that the shape of these patterns does not change in time is only approximately correct, since western boundary currents may meander slightly or form warm or cold rings. Over land, changes in the seasonal cycle may be large (e.g., Kumar et al. 1994), but variations of the seasonal cycle over the oceans are more damped. If we assume that the fundamental oceanic and atmospheric circulations are unchanged over the historic period, then the basic features of these patterns should be unchanged over that period. Thus, changes in the coefficient of the *C* pattern at a given time, relative to the base period, will reflect the bias at that time.

We assume that the daily cycle of SST is on average small enough that we may use both day and night SST together with adjusted NMAT to estimate the historic bias. This is tested by comparison of the 1968–97 *C* patterns computed using day and night SST, compared to patterns computed using only night SST for the same period. The inclusion of day SST makes practically no difference. Spatial correlations for each month are 1.00. Therefore, we may safely use both day and night SST in our analysis and we can ignore the diurnal cycle of SST in the 2° monthly superobservations.

We assume that our 1968–97 base period has a constant level of bias that we can adjust the historic SST against. In the base period, the SST measurements are dominated by intake and insulated bucket temperatures. The stability of the coefficient estimates in the 1968–97 period (Fig. 4) suggests that this is a stable reference period. However, this bias correction does not address individual differences among the different instruments presently used to measure SST.

We assume that the seasonal cycle of bias does not change in phase, allowing the use of annual coefficients. Seasonality in the annual cycle of bias was checked by computing the average annual cycle in the 1930–40 bias using both the annual- and the monthly coefficient estimates. The annual cycle for both had the same phase, so the more stable annual-coefficient estimates were used. This conclusion is consistent with the FP95 bias, which also has a constant phase.

Because we use annual coefficients we have enough data to define the coefficient over the historic period. The final smoothing further reduces variations that may be caused by the use of sparse data. Differences between the smoothed curve and the raw annual fit (Fig. 4b) are usually less than 20% of the size of the main bias change in the early 1940s. The final smoothing does not greatly change the first-order bias estimate, and the difference between the smoothed curve and the unsmoothed annual coefficient is an indication of the uncertainty of the bias. The root-mean-square difference between the smoothed and unsmoothed coefficients in Fig. 4b is about 0.06°C before 1920, going down to 0.04°C in 1930, but increasing to about 0.16°C in the 1940s. In the base period it is about 0.04°C. For the pre-1940 bias, this is similar to or smaller than the standard error of the global bias estimated independently by Folland et al. (2001).

Last, we assume that the Bottomley et al. (1990) version D corrections for NMAT are accurate globally. Recent work by C. K. Folland and D. E. Parker (2001,

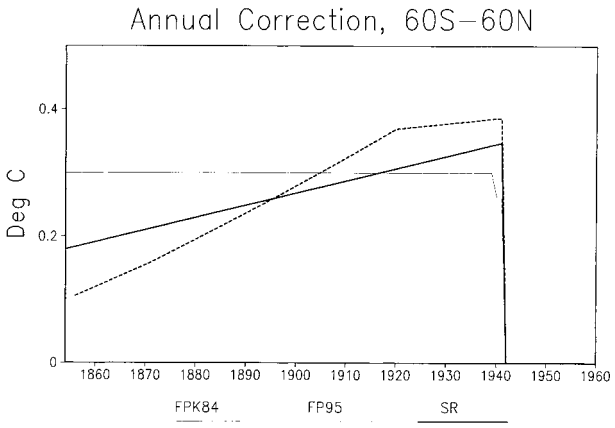


FIG. 5. Annual-average bias corrections over the area 60°S–60°N for the SR corrections (heavy solid), FPK84 (light solid), and the FP95 corrections (dashed).

personal communication) has confirmed those corrections in the late 1930s. However, they also found evidence that there may be some NMAT bias late in the base (1968–97) period, caused by screens now being appreciably higher than in the period for which zero correction was assumed. That can potentially add an additional NMAT bias of 0.05°C to the end of the base period. Averaged over the base period, the uncertainty should be even less, and it would not greatly affect our results.

5. Comparison of SR and FP95 bias corrections

The SR bias corrections are compared with the FP95 bias corrections in order to show consistencies and differences between the two. Average annual bias corrections in the region 60°S–60°N, for both (Fig. 5) shows that the SR corrections are similar to those of FP95. From 1860 to 1920 FP95 increases with a slope of about 0.1°C (25 yr)$^{-1}$. The corrections of Bottomley et al. (1990, not shown) are very similar to FP95 from 1900 to 1941. However, Bottomley et al. (1990) have a smaller slope than FP95 by about one-half between 1860 and 1900. Except for a small decrease near the end of the period, FPK84 is constant from 1860 to 1941. The SR bias correction is slightly stronger than the FP95 estimate in the nineteenth century. Its increase before 1942 is similar to FP95, but with a smaller slope of about 0.05°C (25 yr)$^{-1}$. The SR and FP95 average bias corrections are always within about 0.05°C of each other after 1870.

When we do not adjust NMAT as discussed in section 2, our computed 1854–1941 bias correction is about constant, and the annual and 60°S and 60°N average is similar to the FPK84 average. The slope of the SR SST bias correction shown in Fig. 5 is a direct consequence of the adjustments to NMAT, which are strongest in the nineteenth century.

The FP95 bias is modeled from the estimated proportion of wooden versus canvas buckets used for obtaining surface water samples and from the estimated speed of ships. Both of these change with time, changing the evaporative cooling from buckets, which is responsible for the bias estimate. For the FP95 bias, it is estimated that the percentage of canvas versus wooden buckets increases linearly from 1856 to 1920, and that the speed of ships increases linearly from 1870 to 1940, which produces the two segments of near-linear increase. Parker et al. (1995) and Folland et al. (2001) estimate that the FP95 bias correction has an uncertainty (2σ) of about $\pm$0.15°C in the late nineteenth century. For the near-global averages, the SR bias correction is well within this level of uncertainty.

The average FP95 and SR bias corrections are strong and in good agreement in the 1910–40 period. Fields of bias corrections, average for that period for both analyses, are shown in Fig. 6 for summer and winter. Both analyses show concentrated maxima off the east coast of Asia and North America during the boreal winter. In those regions and times, the prevailing winds bring cold and dry continental air over warm western boundary currents, enhancing the heat loss from buckets due to latent and sensible heat loss. In the austral winter, both estimates show increased bias corrections adjacent to Australia and southern Africa, which produces a Southern Hemisphere maximum in that season. However, the Southern Hemisphere midlatitude winter maximum is less than in the Northern Hemisphere. Changes in the climatological winds also affect the mean monthly heat loss, and may account for the variations in the Tropics.

The SR and FP95 bias estimates for winter and summer are generally consistent, but there are important differences. Differences are largest in the Northern Hemisphere winter, where over most of the western North Pacific and northwest Atlantic the SR estimate is 0.6°C or larger. With the FP95 estimate, the area above 0.6°C is smaller and confined to south of about 40°N. That difference has relatively little practical impact on historic SST, since there are relatively few winter observations from those latitudes. Over most of the oceans the two sets of estimates are more similar. The following section describes comparisons done to help to evaluate differences between the FP95 and SR bias correction estimates.

6. Bias correction evaluations

To help to better evaluate the annual cycle of bias, SSTs from NODC ocean temperature profiles were compared with COADS SSTs for common areas. The NODC data (Levitus et al. 1998) are from hydrographic casts that collected temperature samples at the surface and a number of depths. Here we use the surface NODC temperature observations. These measurements have much lower bias than typical observations from merchant ships because they are taken using scientific instruments from research ships. The bottle-data temperatures are

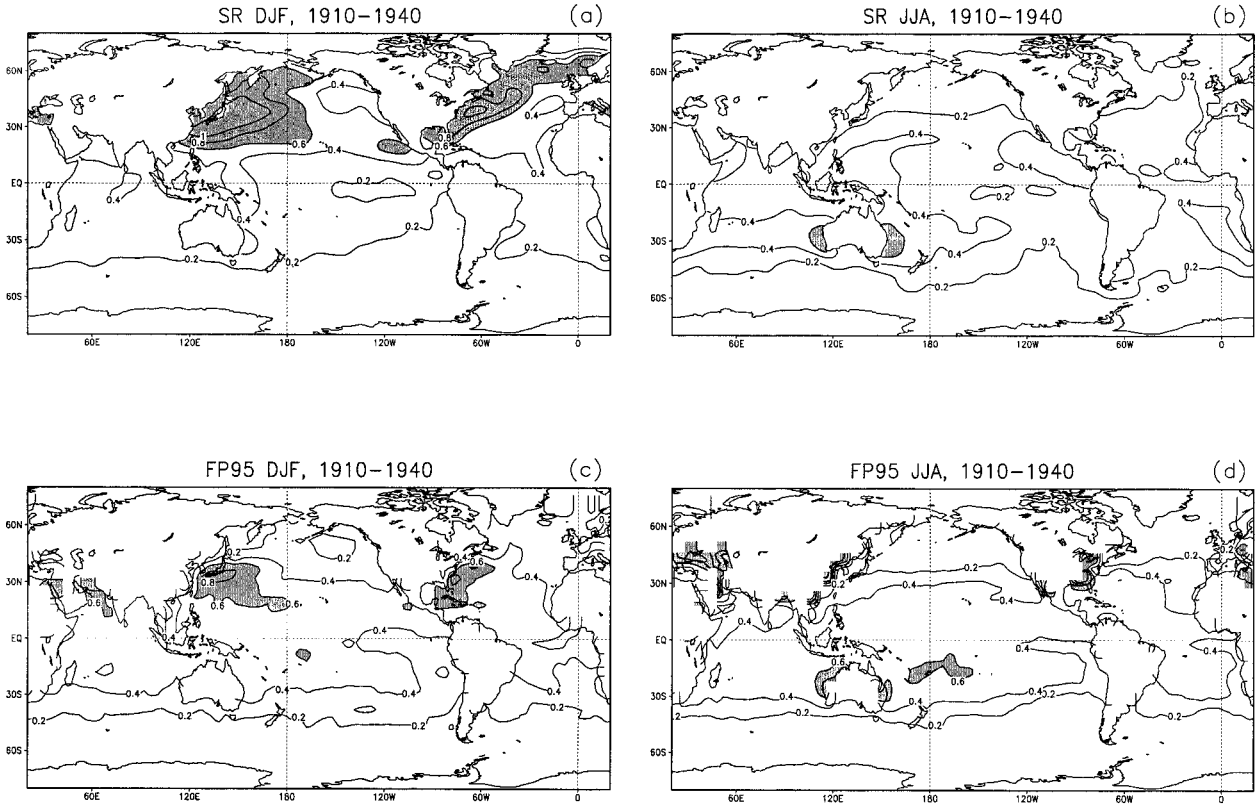


FIG. 6. Average bias corrections over 1910–40 for the seasons Dec–Feb and Jun–Aug, for (a), (b) SR and (c), (d) FP95.

from Nansen-bottle samples using reversing thermometers. These samples do not suffer from heat loss and the thermometers are highly accurate.

Outside of the Northern Hemispheric midlatitudes there are very few NODC data, so we limit our comparisons to the 25°–45°N region. To get the most out of the sparse NODC SSTs, we convert them and the COADS SSTs to anomalies (using the 1961–90 base) and then average the anomalies to 10° latitude–longitude squares using optimal averaging (OA; Kagan 1979; Smith et al. 1994). In an OA, weights for data are computed that minimize the root-mean-square error of the average, provided that covariance statistics and the data error may be estimated. The statistics are similar to those used in the smoothing optimal interpolation analysis discussed in section 4 [see Reynolds and Smith (1994) and Smith et al. (1994) for more discussion of the statistics]. The main advantage of OA over arithmetic averaging (which assumes that all weights are equal) is that an OA takes into consideration the distribution of data within an area. An arithmetic average assumes that all data are equally representative. Anomalies are averaged because they have larger correlation scales than the full SST and, thus, can be more accurately averaged to these relatively large regions.

In all regions where both NODC and COADS SST 10° anomalies are defined the difference is taken. This difference provides an independent bias correction estimate. The mean 1968–97 difference is removed, since we define that period to have zero bias. The 1968–97 difference between COADS and NODC SST is 0.1°C, with COADS SST systematically warmer over this period. The COADS data in this period are affected by engine-intake temperatures, which have a well-documented warm bias (e.g., Folland et al. 1993).

For the annual cycle, averages of each month are computed across the 1930–40 period, when the bias is strong and NODC bottle data are most dense. A three-point binomial filter is applied to the annual cycle. For this comparison, averages of FP95 and SR bias corrections are computed using data only from regions and times when the NODC–COADS bias is defined.

The annual cycle of bias corrections (Fig. 7) is similar for all three. The stronger SR amplitude is closer to the NODC–COADS amplitude than FP95, suggesting that the SR amplitude is more realistic. However, the FP95 phase agrees better with the NODC–COADS phase. The larger Northern Hemisphere SR corrections, as compared with FP95, are concentrated in the western half of the ocean basins in winter (Fig. 6). In the Tropics the FP95 corrections are slightly larger than SR over a broad area. When averaged spatially and over the annual cycle, these differences tend to cancel each other, as shown in Fig. 5.

Similar comparisons are used to evaluate the annual-average bias correction estimates beginning in the

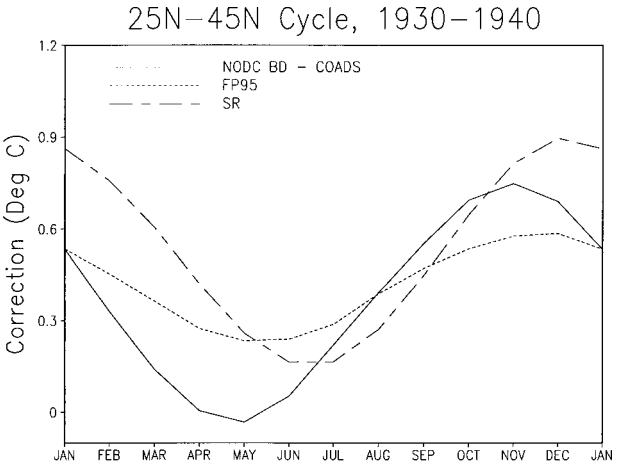


FIG. 7. Annual cycle of bias corrections over the 1930–40 period, averaged over 25°–45°N, from estimates based on NODC bottle data and averages from collocated FP95 and SR corrections.

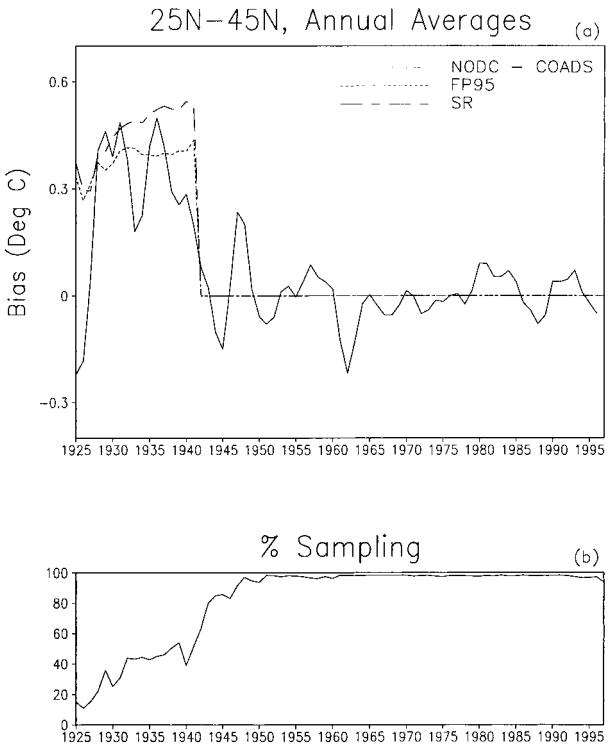


FIG. 8. Annual-average bias correction estimates averaged over 25°–45°N based on NODC data and average corrections from (a) collocated FP95 and SR and (b) the percent of 10° areas sampled by both NODC and COADS data.

1920s. Because of the sparsity of the data, we use both the bottle data and the NODC mechanical bathythermograph (MBT) data, which became available in the 1940s. The NODC–COADS bias correction is defined and averaged over the same region, and annually averaged. A three-point binomial smoother is applied to the time series, and the 1968–97 average difference is removed. Again, averaging is done only using biases collocated with NODC–COADS differences (Fig. 8a). Although the NODC–COADS estimate is noisy, with several large-amplitude swings, it is consistent with the other two estimates. For the early 1940s the NODC–COADS correction decreases abruptly, although it appears to decrease over several years rather than instantly. The high-amplitude variations suggest that NODC–COADS data may not be capable of better resolution.

Before about 1930 the NODC data become very sparse, as indicated by the percent of the ocean area in that latitude band for which a 10° NODC average could be defined (Fig. 8b). We had hoped to use NODC data to investigate bias prior to 1900. However, because the data are too sparse we had to look elsewhere.

In Parker et al. (1995), global and hemispheric averages of corrected SST and NMAT were compared, showing the consistency of their averages in the nineteenth and twentieth centuries. Here we test Northern Hemisphere bias adjustment estimates prior to 1900 using the independent land surface temperature data from the Global Historical Climate Network (GHCN) of station temperature data, adjusted to remove inhomogeneities (Peterson and Vose 1997). Many of the GHCN temperature records extend into the nineteenth century, and they contain many more observations per month at each station as compared with the data available at any one location in COADS. However, the GHCN stations are spatially much more sparse than COADS. The GHCN stations are also subject to diurnal heating and cooling that does not affect SSTs as strongly. Because of these differences from marine data, we do not expect a highly correlated estimate from the two types of data. However, the GHCN data may be suitable for rough validation of the bias corrections.

We define the GHCN bias correction as follows. First the area-average anomalies (with respect to the 1961–90 annual cycle) of GHCN and SSTs are computed over a given latitude range. The difference between them is time averaged to reduce noise. Last, the average difference for the 1968–97 period is removed since we assume that there is no bias in that period. This estimate assumes that all bias is in the SST, that incomplete sampling does not introduce a large average bias, and that urbanization effects in the GHCN have been adequately accounted for. Again we average over the 25°–45°N region, where GHCN sampling is relatively dense, and time average over 30-yr periods. Averages of the FP95 and SR estimates are similarly averaged for comparisons (Table 3).

The GHCN bias correction is weaker than the SR adjustment, and most similar to the FP95 correction from the late nineteenth through the early twentieth century over the extratropical Northern Hemisphere. However, the GHCN correction is weaker than both FP95 and SR averaged over the 1911–40 period. This comparison with GHCN data further demonstrates that, at least in the Northern Hemisphere, the average bias cor-

TABLE 3. Bias estimates averaged over the 25°–45°N area and over 30-yr periods, based on GHCN data in comparison with the FP95 and SR bias estimates (°C).

Period	GHCN	FP95	SR
1861–90	0.19	0.20	0.31
1871–1900	0.13	0.24	0.34
1881–1910	0.19	0.29	0.37
1891–1920	0.27	0.33	0.39
1901–30	0.30	0.38	0.42
1911–40	0.24	0.40	0.45

rections are reasonable in both SR and FP95. If we consider the Folland et al. (2001) uncertainties for the FP95 adjustments and land-air temperatures, then the differences between the three estimates are not significant for the nineteenth and early twentieth century. Sensitivity to sampling of the GHCN correction has been tested by repeatedly computing the GHCN estimate using the 1968–97 data but with sampling from historical 30-yr periods. This test shows that sampling variations may account for variations of 10% or less in the GHCN estimate.

To evaluate the FP95 corrections, Folland et al. (2001) compare observed land-air temperatures to those from a numerical model forced by SST with and without the corrections. For the post-1942 period the area-average model temperatures closely follow the observations. In the pre-1942 period the area-average model air temperatures over land from the run with uncorrected SST are biased with respect to the observations. The area-average model temperatures from the run with corrected SST are similar to the observations, further validating the large-scale corrections.

Evaluations of the FP95 bias corrections around Japan were done by Hanawa et al. (2000) by comparing coastal station SST data with corrected and uncorrected ship reports. The ship SST for comparison with a station were taken from the ocean region near the coastal station, and comparisons were done at five stations they judged to be suitable. In the pre-1942 period they found that the FP95 corrections completely removed bias with respect to coastal SST at three stations in the area 24°–42°N. The annual-average SR correction is 30%–70% larger than the annual-average FP95 correction at those three stations, and thus would overcorrect with respect to coastal SST. At two other stations near 42°–43°N the FP95 corrections removed about one-half the bias. The annual-average SR correction is almost 2 times as large as the annual average from FP95 at those stations, and thus would more completely remove bias with respect to the coastal SST at those two stations. Comparisons with the Hanawa et al. (2000) data lends support for the stronger SR corrections north of 40°N, as compared with FP95. However, these comparisons also suggest that the local maximum SR corrections near 30°N in the western Pacific may be too large.

7. Conclusions

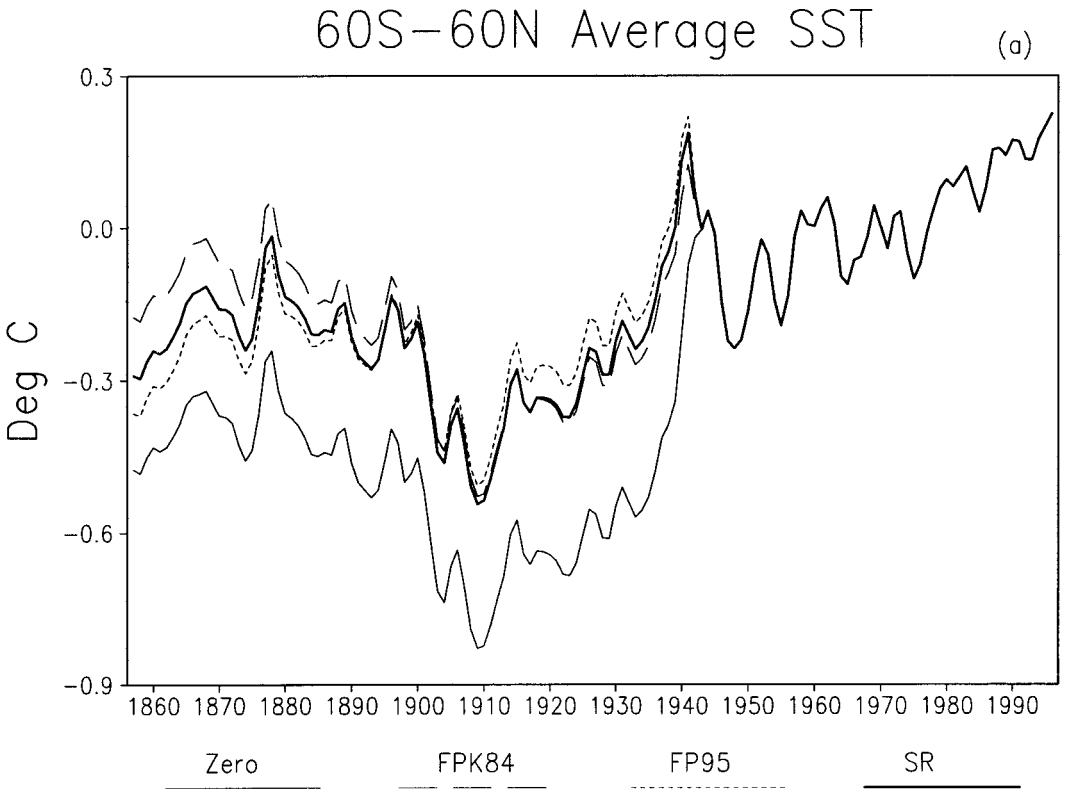
In this study we compared the historic bias corrections of Folland and Parker (1995, FP95) with our own (SR) empirical corrections. The need for bias correction of pre-1942 SST observations, and the approximate size and spatial distribution of corrections, is confirmed by our study. Based on the difference between the SR bias correction and the FP95 bias correction, and also from comparisons with other observations, there is an uncertainty in the magnitude of the strongest bias. In boreal winter that difference can locally be as large as several tenths of a degree (Fig. 6), although the average difference is much less. The correction uncertainty is largest in the nineteenth century and in the early 1940s. In the nineteenth century uncertainty is large because of sparse sampling and the need for strong NMAT adjustments. Increased uncertainty in 1942–45 is caused by decreases in sampling and nonstandard NMAT measurement practices in that period.

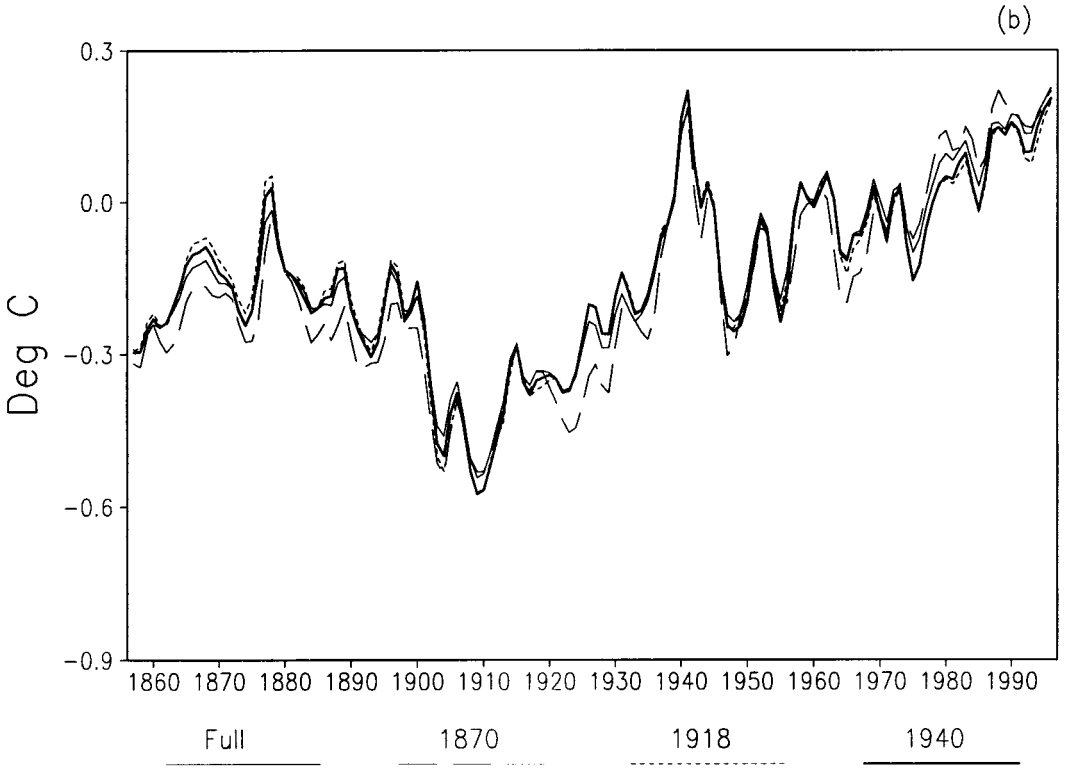
Abrupt changes in the bias are possible when one data source ends and is replaced with another, which happened in 1942. For example, a major source of data in 1941 was Japanese, while in 1942 the major source of data was U.S. (Woodruff et al. 1988). The hydrographic data analyzed here suggest a sudden decrease in the bias may have taken place in the early 1940s. Unfortunately, the hydrographic data are not dense enough to clearly define how rapidly the bias decreases.

To examine how these uncertainties affect evaluation of climate change, we show time series of SST anomalies averaged between 60°S and 60°N (Fig. 9a). The need for bias correction is clear from the jump in uncorrected temperatures around 1940. Corrected temperatures are all similar. In particular, the increasing trend in temperature beginning about 1910 is similar with all bias corrections. The largest differences between the corrected averages occur before about 1890. In that period the SST corrected with the simpler FPK84 corrections are more than 0.1°C larger than the other two corrections. Differences are much less between the analyses employing the FP95 and SR corrections. Their difference has a maximum of about 0.07°C in 1860, decreasing to less than 0.05°C by 1880. The differences are much less than the climate signal of about 0.6°C over the twentieth century, and their consistency in reassuring.

→

FIG. 9. (a) Annual and 60°S–60°N average of SST anomalies (1961–90 base) with no bias correction (''Zero''), and with corrections of FPK84, FP95, and SR. A three-point binomial filter is applied to all time series. (b) Analysis using the SR bias correction and using sampling for the given year.





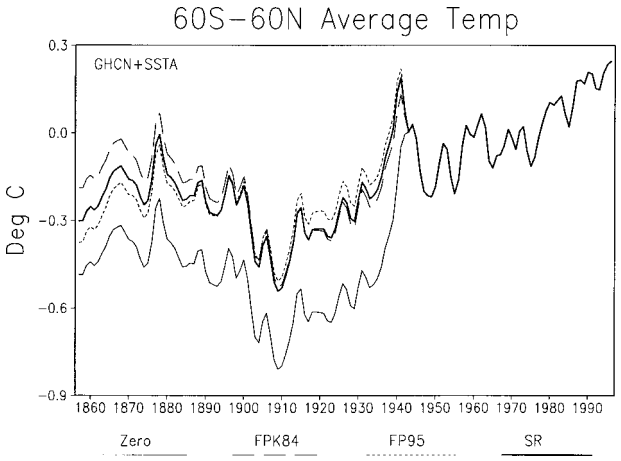


FIG. 10. Annual and 60°S–60°N average of land + SST anomalies (1961–90 base) with no bias correction (''Zero''), and with corrections of FPK84, FP95, and SR. The GHCN land temperatures are used. A three-point binomial filter is applied to all time series.

Sensitivity of the analysis to sampling is approximated using sampling from historical years and repeating the analysis using the SR bias correction (Fig. 9b). For a given time, the reconstruction of SST anomalies is performed using only data from locations where they also exist in the sampling year. This is sometimes referred to as a frozen grid experiment. Variability is greatest using sampling from 1870 and 1918, when the range of different estimates can exceed 0.1°C. This range is comparable to the estimated sampling error for global-average surface temperatures from Folland et al. (2001). This result shows uncertainties in near-global average SST due to sampling. It also shows that much of the near-global signal in SST can be captured using the historic sampling.

If the GHCN land surface temperature anomalies and SST anomalies are combined, the difference in the averaged surface temperature anomalies are slightly less influenced by the different bias estimates (Fig. 10). The series of either near-global SST alone or combined temperatures through the twentieth century are not greatly affected by the choice of the SST bias correction used. Local climate studies may be more strongly affected. For example, studies of the extratropical northern SST in winter will have greater uncertainties because of locally large differences between FP95 and SR corrections. The agreement of the average corrections between 1900 and 1941, and their general agreement with independent bias estimates demonstrates the credibility of the corrections.

We made several assumptions in the development of our corrections. Our most questionable assumption is that historic changes in SST–NMAT reflect only SST changes. However, even with these assumptions our overall bias corrections are substantially the same as the FP95 corrections. Our validation studies suggest that more work is needed to fully understand the local differences. Additional work may also be needed to better evaluate the bias in the 1940s and early 1950s, when the unsmoothed coefficients indicate that there may be some residual bias in the SST observations (Fig. 4).

Because of the many similarities between the SR and FP95 corrections, we could find no clear proof that one is better than the other. For these reasons we recommend that the FP95 bias corrections be used. However, users should be aware of the uncertainties in these bias corrections, especially in the nineteenth century and in the 1940s, and they should also be aware of the effect of those uncertainties on climate-change estimates. In addition, whenever substantial additions to the historical data are made, changes in the bias should be evaluated. The future additions of new historical SST may demand new bias estimates to account for differences in the bias of the new data.

Acknowledgments. We are grateful for the support for this research by NOAA's Office of Global Programs, Grant GC00-140. We also thank Chris Folland and David Parker for their reviews and advice, and for supplying the U.K. bias estimates. We appreciate the encouragement of Tom Karl and Ants Leetmaa to carry out this research. Scott Woodruff, Steve Worley, and Diane Stokes supplied and helped with the COADS data; Syd Levitus supplied the NODC data; Tom Peterson and Jay Lawrimore supplied and helped us with the GHCN data. In addition, we thank Kevin Trenberth, Nick Rayner, Alan Basist, and Russ Vose, and an anonymous reviewer for much constructive advice. This study began while TMS was supported by NCEP/CPC.

REFERENCES

Bottomley, M., C. K. Folland, J. Hsiung, R. E. Newell, and D. E. Parker, 1990: *Global Ocean Surface Temperature Atlas ''GOSTA.''* Met Office and the Massachusetts Institute of Technology, 20 pp. and 313 plates.

Christy, J. R., D. E. Parker, S. J. Brown, I. Macadam, M. Stendel, and W. B. Norris, 2001: Differential trends in tropical sea surface and atmospheric temperatures. *Geophys. Res. Lett.*, **28,** 183–186.

Folland, C. K., and D. E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quart. J. Roy. Meteor. Soc.*, **121,** 319–367.

——, ——, and F. E. Kates, 1984: Worldwide marine surface temperature fluctations 1856–1981. *Nature*, **310,** 670–673.

——, R. W. Reynolds, M. Gordon, and D. E. Parker, 1993: A study of six operational sea surface temperature analyses. *J. Climate*, **6,** 96–113.

Folland, C. K., and Coauthors, 2001: Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.*, **28,** 2621–2624.

Hanawa, K., S. Yasunaka, T. Manabe, and N. Iwasaka, 2000: Examination of correction to historical SST data using long-term coastal SST data taken around Japan. *J. Meteor. Soc. Japan*, **78,** 187–195.

Higgins, R. W., Y.-P. Yao, M. Chelliah, W. Ebisuzaki, J. E. Janowiak, C. F. Ropelewski, and R. E. Kistler, 1996: *Intercomparison of the NCEP/NCAR and the NASA/DAO Reanalyses (1985–1993).* NCEP/Climate Prediction Center Atlas No. 2, 169 pp.

Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell, Eds., 1996: *Climate Change 1995:*

The Science of Climate Change. Cambridge University Press, 572 pp.

Kagan, R. L., 1979: *Averaging Meteorological Fields* (in Russian). Gidrometeoizdat, 212 pp. (English translation 1997, Kluwer Academic.)

Kumar, A., A. Leetmaa, and M. Ji, 1994: Simulations of atmospheric variability induced by sea surface temperatures and implications for global warming. *Science,* **266,** 632–634.

Levitus, S., T. P. Boyer, M. E. Conkright, T. O'Brien, J. Antonov, C. Stevens, L. Stathopolos, D. Johnson, and R. Gelfeld, 1998: *Introduction. Vol. 1, World Ocean Database 1998,* NOAA NESDIS Tech. Report, 346 pp.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Parker, D. E., C. K. Folland, and M. Jackson, 1995: Marine surface temperature: Observed variations and data requirements. *Climatic Change,* **31,** 559–600.

Peterson, T. C., and R. S. Vose, 1997: An overview of the global historical climatology network temperature database. *Bull. Amer. Meteor. Soc.,* **78,** 2837–2849.

Reynolds, R. W., 1993: Impact of Mount Pinatubo aerosols on satellite-derived sea surface temperatures. *J. Climate,* **6,** 768–774.

——, and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate,* **7,** 929–948.

Slutz, R. J., S. J. Lubker, J. D. Hiscox, S. D. Woodruff, R. L. Jenne, D. H. Joseph, P. M. Steurer, and J. D. Elms, 1985: COADS: Comprehensive Ocean-Atmosphere Data Set. Release 1, 262 pp. [Available from Climate Research Program, Environmental Research Laboratories, 325 Broadway, Boulder, CO 80303.]

Smith, T. M., R. W. Reynolds, and C. F. Ropelewski, 1994: Optimal averaging of seasonal sea surface temperatures and associated confidence intervals (1860–1989). *J. Climate,* **7,** 949–964.

——, R. E. Livezey, and S. S. Shen, 1998: An improved method for analyzing sparse and irregularly distributed SST data on a regular grid: The Tropical Pacific Ocean. *J. Climate,* **11,** 1717–1729.

Trenberth, K. E., J. R. Christy, and J. W. Hurrell, 1992: Monitoring global monthly mean surface temperatures. *J. Climate,* **5,** 1405–1423.

Weller, R. A., and S. P. Anderson, 1996: Surface meteorology and air–sea fluxes in the western equatorial Pacific warm pool during the TOGA Coupled Ocean–Atmosphere Response Experiment. *J. Climate,* **9,** 1959–1990.

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.