INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **25**: 895–911 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.1171

ASSESSING BIAS CORRECTIONS IN HISTORICAL SEA SURFACE TEMPERATURE USING A CLIMATE MODEL

CHRIS FOLLAND*
Hadley Centre, Met Office, Exeter EX1 3PB, UK

Received 5 July 2004
Revised 23 November 2004
Accepted 2 December 2004

ABSTRACT

Analyses of simulations of variations in global and large-regional land surface air temperature (LSAT) for 1872–1998 using the HadAM3 atmospheric general circulation model are reported. The analyses are designed to test the accuracy of bias corrections to sea-surface temperature (SST) used in the Hadley Centre's global sea ice and SST (GISST3.1) data set, the more recent Hadley Centre sea ice and SST (HadISST) data set, and in the underlying Met Office historical SST (MOHSST and HadSST1) data sets. The tests are important because SST corrections considerably affect estimates of the magnitude of global warming since the late 19th century. Two ensembles of simulations were created using GISST3.1 as the lower boundary condition. The first ensemble, of six integrations, was forced using GISST with bias-corrections applied from 1871 until 1941, and was continued with no bias corrections to 1998. A second ensemble of four integrations, for 1871 to 1941, was forced with uncorrected GISST data. Simulations with uncorrected GISST show a substantial and often highly significant cold bias in simulated global and large-regional annual mean LSAT changes before 1942 relative to a 1946–65 reference period. By contrast, corrected SST data led to simulations of LSAT changes that are generally insignificantly different from those of observed LSAT in most regions before 1942. Tests on extratropical hemispheric scales generally validate the seasonal variation of the bias corrections, though less clearly before 1890 in some seasons. Issues about the quality of the LSAT data are raised by the results in a couple of regions. Over Australia, the model may have reconstructed LSAT changes using bias-corrected GISST with greater accuracy than the observations before about 1910. © Crown Copyright 2005. Reproduced with the permission of Her Majesty's Stationery Office. Published by John Wiley & Sons, Ltd.

KEY WORDS: sea surface temperature; air temperature; climate model; biases

1. INTRODUCTION

Sea-surface temperature (SST) is a major component of global surface temperature used in analyses of climate change, including the series used by the first to third assessments of the Intergovernmental Panel on Climate Change (IPCC; e.g. Folland *et al*., 2001a). In addition, SST data are also vital to many studies of climate variability and seasonal to decadal predictability. Before 1942, quite large corrections are applied to these data in Met Office global SST data sets (Folland and Parker, 1995), such as the HadSST1 and HadISST (Rayner *et al*., 2003) data sets and the earlier MOHSST data sets (e.g. Bottomley *et al*., 1990, and Parker *et al*., 1995). The reason for the corrections is that there were substantial changes in SST measurement practice over the period from the beginning of the data set to 1941. In the mid to late 19th century, wooden buckets are thought to have been widely used to collect sea water to make the measurement of SST on a ship's deck. These buckets lose heat due to evaporation from the free water surface and to a limited extent through the wooden bucket walls. However, the heat losses are relatively small, so cooling of the sea water prior to measurement of its temperature was relatively small, but not negligible. As time went on, a greater proportion of measurements are thought to have been made using uninsulated canvas sea-water buckets, which lose

* Correspondence to: Chris Folland, Hadley Centre, Met Office, Exeter, EX1 3PB, UK; e-mail: chris.folland@metoffice.gov.uk

heat more readily due to evaporation and sensible heat losses from the usually thin canvas bucket walls. So, cooling (under most conditions) of the sea water prior to measurement of its temperature was substantially faster than for wooden buckets. Soon after World War II started, buckets ceased to be used, being replaced by engine-intake thermometers, or were used much less, presumably due to the dangers involved. Thus, the artificial cooling of the sea water prior to measurement largely ceased.

The corrections are not fully dependent on the assumption that there were no engine intakes used before 1942. It is likely that some engine intakes were actually used, at least from the 1920s. The method that Folland and Parker (1995) use to make corrections implicitly allows for this possibility, at least to an extent. The correction technique essentially depends on the minimization in the extratropics of the additional annual cycle of SST that the wooden or uninsulated bucket measurements create compared with the annual cycle of SST in a given region in the reference period 1951–80. The exact method is slightly more complicated, for statistical reasons; Folland and Parker (1995) give the details. If some engine intakes were used, then this additional or artificial annual cycle would be reduced compared with the situation where only buckets were used. This would result in the physical model used to correct the annual cycles being integrated for a smaller time, reducing the calculated correction sizes. If engine intakes introduced their own artificial annual cycle that was different from that expected from buckets, then the method would not work so well. However, the evidence is that any artificial annual cycle in engine-intake data is considerably less in magnitude than for the uninsulated buckets. (It is possible for engine-intake data to have an artificial annual cycle, as they tend to sample deeper water than buckets and may not fully measure the amplitude of the seasonal thermocline.) Thus, the method is, to an extent, self-calibrating. Similarly, if, in the 1951–80 reference period, there was a mixture of observational types, with at least some buckets present (there is very good evidence for this; J. Kennedy, personal communication), then the calculated corrections are simply those needed relative to the mix of data in 1951–80. Note that in equatorial areas, where annual cycles are too small for reliable correction of SST directly, a slightly different approach is used by Folland and Parker (1995). In these regions, the time for which the bucket model was integrated in the extratropics to minimize the artificial annual cycle is used to calculate the correction directly. This time is the mean time of exposure of the bucket before the measurement was taken, averaged over the globe. This often leads to large annual mean corrections in the tropics, because of the strong influence of evaporation on the cooling of the bucket in warm climatic conditions.

Another assumption was that the ratio of the number of wooden bucket to canvas bucket measurements changed globally in a uniform way in early decades. In addition, the corrections calculated depend on assumptions, based on historic records, about the changes in the mean speed of ships. This affects the ventilation and heat loss from the bucket, both on deck and during the bucket hauling phase. Again, global mean changes in this speed were assumed in the absence of better information. Finally, the heat losses calculated worldwide from buckets used atmospheric and SST climatological data using a 1951–80 climatology. There were insufficient data to make any other choice; but, because of the partially self-calibrating characteristics of the method, this is not likely to be a serious problem. However, in future it would be desirable, if sufficient data were to become available, to estimate earlier climatologies, though the current SST bias corrections themselves would influence the result slightly. Sufficient data are unlikely to become available to correct each SST observation individually; indeed, the current correction technique does not allow that to be done in principle. For further details, see Folland and Parker (1995).

It is important to realize that the SST bias corrections are completely independent of any land surface air temperature (LSAT) data set, including that used later in this paper, and are also independent of the night marine air temperature (NMAT) data set except in one respect. NMAT data is used in a limited way to estimate the changing mix of wooden and canvas buckets between 1856 and 1920 as discussed in Folland and Parker (1995). Smith and Reynolds (2002) use the NMAT data to create an alternative set of bias corrections to SST for the same period. Quite apart from the assumptions made in such a technique, changes in NMAT due to better data or changes in their own bias corrections, e.g. as developed in Rayner *et al*. (2003), will feed back on the SST corrections. Thus, the corrections assessed here are independent of all other data sets, depending on a combination of a physical model of the cooling of wooden or canvas buckets and the statistical characteristics of the SST data themselves. As mentioned above, it is known that some canvas buckets were used after World War II; some, but not all, of them were insulated. The problem of possible bias corrections

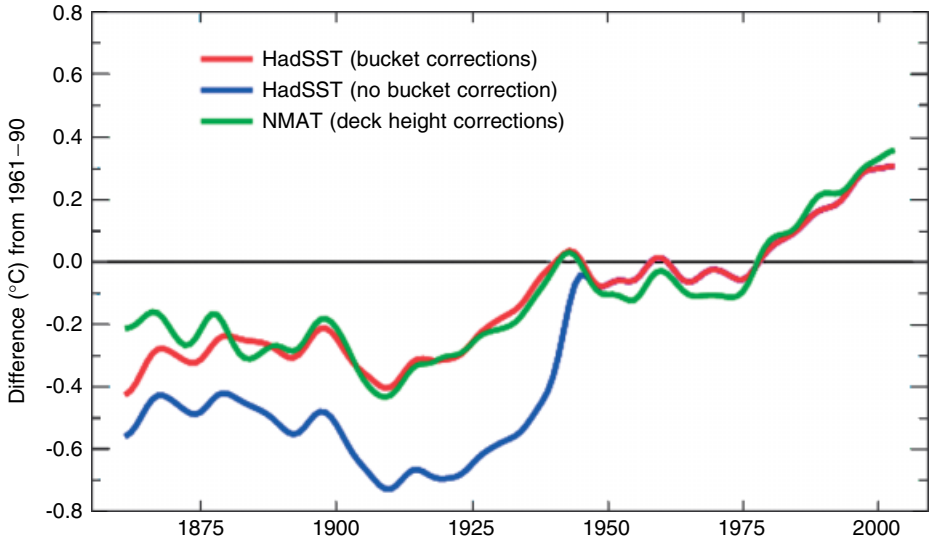

Figure 1. Influence of bias corrections on SST. The globally averaged HadSST1 data set with and without bias corrections (corrections cease at the end of 1941) is compared with bias-corrected global NMAT (HadMAT1). The data are approximately decadally averaged, using a 21-point binomial filter

to SST in recent decades, generally considerably smaller than before World War II, is still being researched, and no corrections are currently made. So, the period after World War II is not important in this paper.

On a global average, time-varying SST bias corrections have a considerable influence on assessed climate change. Figure 1 illustrates this globally by comparing the HadSST data set with and without bias corrections, along with bias-corrected NMAT (Rayner *et al*., 2003). The NMAT data strongly support the bias-corrected SST. However, bias-corrected SST itself has been used to improve NMAT in some areas before 1893 (Parker *et al*., 1995) in Figure 1 (Rayner *et al*., 2003), so it is desirable to validate SST independently of NMAT.

Therefore, in this paper we present a more complete analysis building on the tests of the SST bias corrections using a climate model reported briefly by Folland *et al*. (2001b). Our SST data will soon be superseded by new data sets based on the International Comprehensive Ocean Atmosphere Data Set (ICOADS; Diaz *et al*., 2002). However, it is expected that the SST bias corrections in ICOADS to 1941 will be similar to those discussed here, except for about 1939–41 (Rayner *et al*., 2005). So, the tests we describe will be largely relevant to future SST data sets, though additional tests will be desirable.

2. MODEL EXPERIMENTS TO SIMULATE LSAT

2.1. Design of the model experiments

A method of testing the SST bias corrections independently of NMAT is to force an atmospheric climate model in ensemble mode with both time-varying bias-corrected and time-varying uncorrected SST. A test is then made to determine whether a suitable measure of the simulated climate variability is more realistic when forced with the bias-corrected SST. This measure is chosen to be the LSAT of Jones *et al*. (1999). A limitation is that the observed LSAT data sets used here (nominally at 1.5 m above ground level) are less complete than those more recently developed (Jones and Moberg, 2003). They also lack the variance correction procedure of Jones *et al*. (2001) that was applied to individual $5^\circ \times 5^\circ$ areas. On the relatively large space scales analysed here, and especially for decadal averages, lack of this procedure will only have very small effects.

The HadAM3 atmospheric model (horizontal resolution 2.5° latitude $\times$ 3.75° longitude, 19 levels in the vertical; Pope *et al*., 2000) was forced with two versions of the Hadley Centre's global sea ice and SST GISST3.1 data, one with and one without bias corrections, starting in 1871. An ensemble of six runs covering

1871–1998 was made with the standard GISST data set that includes SST bias corrections until December 1941, but thereafter has no corrections. These runs were started from different initial atmospheric conditions. A further ensemble of four runs, starting from a further set of different atmospheric starting conditions, was created with a version of GISST without bias corrections. The sea-ice extent data in both experiments was the same. The latter set of runs was terminated at 1944; 1942–44, which had no bias corrections, were used to investigate how fast the model recovered from a lack of corrections before 1942. This recovery supports the idea that commencing the analysis in 1872 after allowing the model to adjust for 1 year is sound. No changes in greenhouse gases, aerosols or other forcings are included. This choice was made because, over the period when bias corrections are applied, these effects are quite small in an atmospheric model. Another factor is that the specification of some of the forcings is rather uncertain and we did not want the results to be affected by such uncertainties.

Table I lists globally averaged corrections in the globally interpolated GISST3.1 data set for selected years. For the best comparison with noninterpolated data used in Folland *et al*. (2001b), we have calculated the mean global GISST correction for 60 °N to 50 °S. Table I shows that the corrections more than double between 1872 and 1940, though they are globally nearly constant after 1920.

To test the veracity of the bias corrections, we have used simulations of LSAT anomalies referred to an average LSAT for the model for 1946–65, a reference period early enough that lack of applied anthropogenic forcings is unlikely to cause serious problems. We chose LSAT as our test variable because it is most likely to be influenced by SST in a quasi-deterministic way, though internal atmospheric variability in the model may still be important. Internal variability is reflected as variations in the simulated LSAT between the ensemble members. We have minimized this problem by using the averages of ensembles of runs when calculating the extent to which the bias corrections can be verified, though the ensemble members are used to estimate uncertainties in the ensemble mean results. In addition, internal variability is likely to increase as the spatial scale of the analysis decreases. So, we have used a number of large land regions varying in scale from the globe to much of Europe. To further minimize internal variability, we mainly analyse annual values but analyse 2-month seasons for the two large regions of extratropical Northern Hemisphere land and extratropical Southern Hemisphere land to test the seasonal variation of the corrections. Table II shows the areas where simulated LSAT changes are analysed. Note that Antarctic land has no influence before 1942 because there are no LSAT data.

2.2. Tests of model surface air temperature climatology and the overall sensitivity of the modelling method

First, we test the veracity of the model's LSAT climatology and gain an overview of the sensitivity of the method. A seriously incorrect model LSAT climatology could result in an incorrect sensitivity to changing SST, e.g. through gross errors in snow cover or soil moisture. We test the model LSAT climatology using the difference between the annually averaged climatological LSAT of HadAM3 for 1961–90 (including a full set of greenhouse forcings for maximum realism over this period) and the observed LSAT climatological temperature for the 11 large regions in Table II. This climatology is taken from the interpolated LSAT data of

Table I. Globally averaged SST corrections in GISST3.1

Year	Bias correction (°C)
1872	0.17
1880	0.20
1890	0.25
1900	0.30
1910	0.34
1920	0.39
1930	0.40
1940	0.40

Table II. Large land regions whose LSAT is used to test annual SST bias corrections

Region	Land areas
Globe	
Northern Hemisphere	
Southern Hemisphere	
Tropics	20 °N to 20 °S
Extratropical Northern Hemisphere	20 °N to 90 °N
Extratropical Southern Hemisphere	20 °S to 90 °S
Extratropical North America	30 °N to 90 °N, 170 °W to 60 °W
Europe	40 °N to 90 °N, 20 °W to 30 °E
Extratropical Eurasia	30 °N to 90 °N, 20 °E to 180
Extratropical South America	30 °S to 90 °S, 90 °W to 30 °W
Australia	10 °S to 50 °S, 110 °E to 160 °E

Table III. Differences in absolute 1.5 m climatological temperature, model minus observations, 1961–90

Area (annual)	Difference (°C)	Area (seasonal)	Difference (°C)
Globe	−0.89	Extratropical Northern Hemisphere Jan and Feb	−2.92
Northern Hemisphere	−1.01	Extratropical Northern Hemisphere Mar and Apr	−0.89
Southern Hemisphere	−0.50	Extratropical Northern Hemisphere May and Jun	1.11
Tropics	−0.57	Extratropical Northern Hemisphere Jul and Aug	0.95
Extratropical Northern Hemisphere	−1.16	Extratropical Northern Hemisphere Sep and Oct	−1.47
Extratropical Southern Hemisphere	−0.36	Extratropical Northern Hemisphere Nov and Dec	−3.18
Extratropical North America	−0.78	Extratropical Southern Hemisphere Jan and Feb	0.69
Europe	−1.79	Extratropical Southern Hemisphere Mar and Apr	−0.82
Extratropical Eurasia	−1.57	Extratropical Southern Hemisphere May and Jun	−2.53
Extratropical South America	−0.38	Extratropical Southern Hemisphere Jul and Aug	−1.97
Australia	0.40	Extratropical Southern Hemisphere Sep and Oct	0.70
		Extratropical Southern Hemisphere Nov and Dec	1.34

New *et al*. (2000). We also include differences for the six individual 2-month seasons and two large regions used in the seasonal tests in Section 4. The data in Table III are calculated using the average of the cosine weighted $5^\circ \times 5^\circ$ differences that contribute to a region.

Table III reflects a well-known overall cold bias in LSAT in most climate models, with only Australia being warmer than the observed climatology. However, the global annual mean bias in LSAT of −0.89 °C is quite small. It is unlikely that the biases on the left-hand side of Table III will cause many problems, noting that Europe has the largest annual bias at −1.79 °C. The biases on the right-hand side of Table III show that the seasonal cycle of extratropical LSAT in both hemispheres is too large, especially in the Northern Hemisphere, where the influence of the imposed observed SST is less. Because of biases, the least reliable model data are likely to be in Northern and Southern Hemisphere winter, though even here the biases of about −3 °C are not excessive, though large enough for some caution. Seasonal biases in smaller areas are greater, so we do not use such areas to test the seasonal cycle of simulated LSAT. Note that the transition seasons have least, and rather little, bias. Given these modest caveats, the strong criticism by Soon *et al*. (2004) of the use of HadAM3 to test SST bias corrections in the paper by Folland *et al*. (2001b) seems unjustified. Indeed, these authors could not have known about the relatively good performance of the HadAM3 LSAT climatology and base their criticisms on a blanket rejection of all dynamical climate models as having unrealistic climatologies.

To test the overall sensitivity of the method, Figure 2(a) firstly shows the worldwide distribution of mean differences in simulated land and ocean surface air temperature for the decade 1930–39 between the ensemble

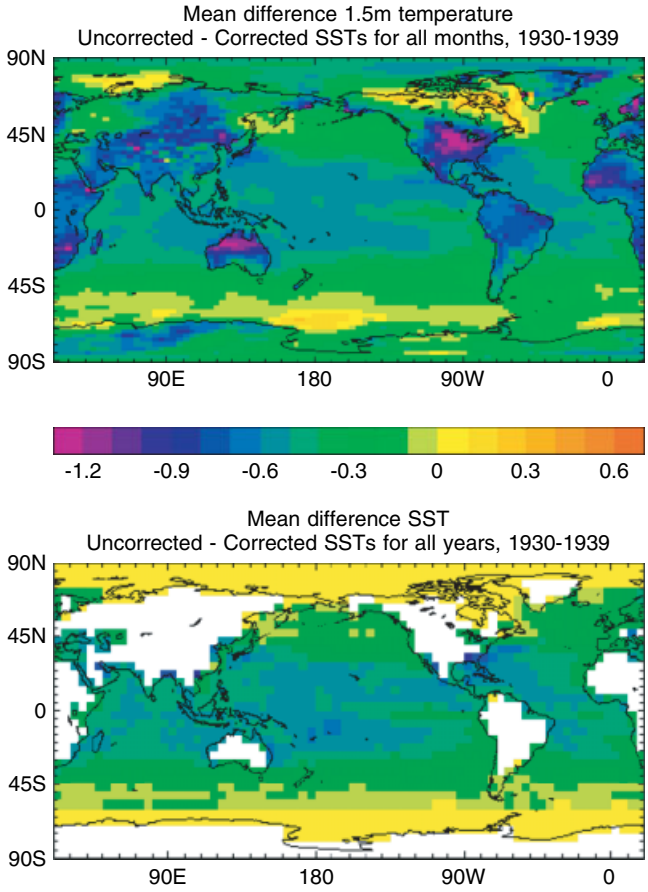

Figure 2. (a) Ensemble mean differences in 1.5 m simulated temperature for the decade 1930–39 between runs of HadAM3 forced with uncorrected GISST and bias-corrected GISST (°C). (b) Corresponding differences in imposed SST (°C)

mean of runs forced with uncorrected GISST and that forced with bias-corrected GISST. The SST difference field is shown in Figure 2(b). This decade has close to the largest corrections (Table I). The general relative coolness of global land is apparent when uncorrected GISST is used, with few areas warmer despite internal atmospheric variability. Substantial land areas are more than 0.6 °C colder with uncorrected SST, with a few isolated regions more than 1 °C colder. This demonstrates the truly global impact of SST bias corrections on modelled surface air temperature. The largest impacts on average are in the tropics and the Northern Hemisphere middle latitudes, where annual mean SST bias corrections are greatest (Folland and Parker, 1995). Figure 2(a) also demonstrates that for the ensemble and region sizes used here, internal atmospheric variability, as demonstrated by substantial local variations in surface air temperature anomalies, will not be a serious problem on the decadal time scale even over a relatively small region like Europe, and over the globe it will be negligible. As indicated in Figure 2, internal variability is likely to be largest in higher latitudes.

The impression gained from Figure 2 is that the cooling of the land surface exceeds that of the SST when there are no SST bias corrections. Over 1930–39, when corrections are largest, the change in global SST is approximately 0.40 °C (Table I), but the change in simulated LSAT is larger at 0.57 °C, a ratio of 1.42. Over the other decades 1880–89, 1890–99,1900–09,1910–19,1920–29 the ratios are 1.29, 1.20, 1.25, 1.32 and 1.30 respectively, giving a mean ratio over 1880–1939 of 1.30. This is not surprising, as a change in SST will also change the water vapour content of the air over land and, therefore, its greenhouse effect over land. Therefore, we may expect about 30% larger simulated LSAT than SST changes on a global scale for the relatively small SST changes seen in this paper. Regionally, this might not be true over individual decades

because of atmospheric circulation changes between the runs with and without bias adjustments (as seen in Figure 2(a)). However, the tendency to larger changes in simulated LSAT than the changes in applied SST is an advantage for the method. It tends, on average, to increase differences by about 30% between LSAT ensemble members simulated by a given change in corrected to uncorrected SST, helping to reduce any confounding effects of internal model variability.

3. TESTS OF ANNUAL MEAN BIAS CORRECTIONS

3.1. Using collocated unfiltered model members

Because observed LSAT data are spatially incomplete, but model data are spatially complete, we concentrate on comparisons of model LSAT collocated (to a resolution of $5^\circ \times 5^\circ$) with those observed, these also being on a $5^\circ \times 5^\circ$ grid. Collocated modelled data are likely to be noisier than complete data, but they are a better sample of the observations. In practice, the ensembles are large enough and spatial correlations between sparsely observed annual or seasonal data are often sufficiently large that it is not surprising that tests (not shown) indicate that little extra noise is introduced by collocation. Accordingly, collocated data have been used throughout the paper.

Figure 3 shows time series, for all 11 regions, of unfiltered simulated annual LSAT anomalies relative to 1946–65 for all individual ensemble members compared with observed anomalies. Ensemble members have anomalies calculated from the 1946–65 average of the relevant ensemble mean, and observations are referred to the observed mean. This removes the effects of biases in model LSAT when we compare modelled and observed *changes* in temperature from the climatology (noting the caveats above). Thus, each ensemble mean anomaly has a value of zero over 1946–65, as does the observed mean.

Observed increases in greenhouse gases might have additional influences on observed LSAT that are not contained in the SST in recent decades. Such additional warming clearly occurs after the 1970s, as shown by Folland *et al*. (1998) and Sexton *et al*. (2003). Thus, after that time, model LSAT time series would be warmer than those shown here if a full range of changing, mainly anthropogenic, atmospheric forcings were added. This extra warming is statistically significant on many space scales, as shown by Sexton *et al*. (2003). Thus, the time series shown here cannot be used to infer relationships between imposed SST and LSAT after about 1980. However, this is not a significant problem for the purposes of this paper.

For the globe (Figure 3(a)), the simulations of LSAT by members of the ensemble using bias-corrected SST (red) are generally separated from those of the ensemble members using uncorrected SST (blue, to 1941 only). An exception is perhaps in the 1870s, when bias corrections are smallest. The same is generally true of Figure 3(b)–(d) for the two hemispheres and the tropics. In the tropics, which has the largest annual mean bias corrections (Figure 2(b)), the members of a given ensemble are remarkably similar and very distinct from the other ensemble. Internal variability is noticeably greater in Figure 3(e) for the extratropical Northern Hemisphere; nevertheless, there is little overlap between bias-corrected and uncorrected members after 1890. The extratropical Southern Hemisphere shows a fairly similar picture, but with somewhat more overlap. In the Southern Hemisphere, especially in the extratropics, pre-1890 observed LSAT values look surprisingly warm, and are distinctly warmer than the bias-corrected simulations of LSAT. The five smaller regions (Figures 3(g)–(k)) show more internal variability, though simulated LSAT forced with the bias-corrected SST data are mostly warmer. Although most of the regions show a significant and generally high correlation between simulated and observed annual LSAT, two regions, extratropical South America and extratropical North America, show no or a negative annual correlation. The pre- 1910 observed LSAT in Australia (Figure 3(k)) contributes disproportionately to Southern Hemisphere LSAT at this time and is notably warmer than the model LSAT forced with bias-corrected SST data, and, surprisingly, even warmer before 1890 than for the recent observed warm decade. There is good evidence that observed Australian data are too warm before 1890 (Nicholls *et al*., 1996, personal communication, and Section 3.3) due to bad exposures of thermometers, and are probably too warm before 1910. Thus, the Jones *et al*. (1999; Jones and Moberg, 2003) annual mean LSAT data are very likely to be biased too warm before 1910 in Australia. Accordingly, pre-1910 Australian LSAT is not used in most climate monitoring by the Bureau of Meteorology

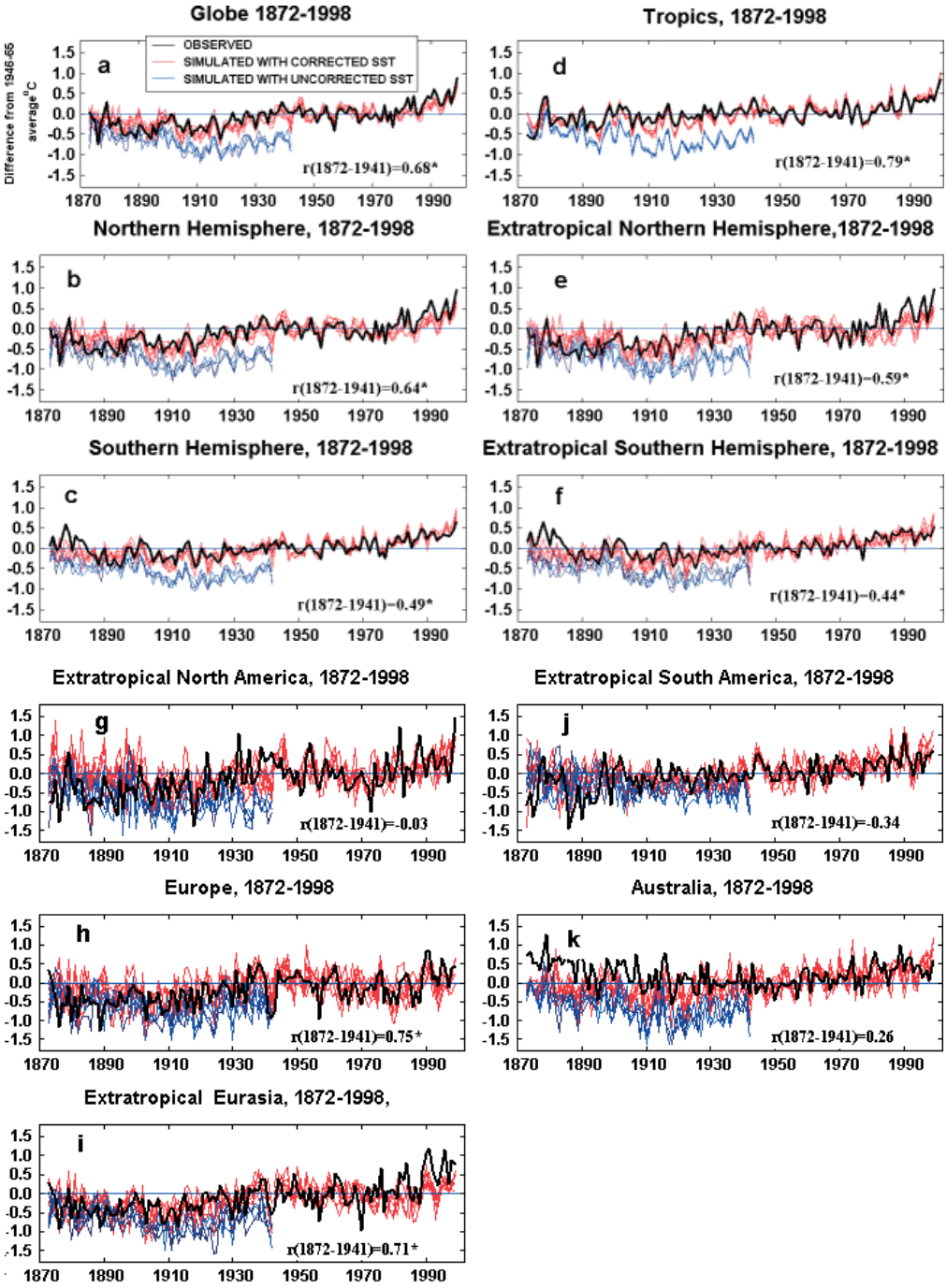

Figure 3. Unfiltered collocated simulations and observations (black curve) of annual mean LSAT anomalies from a 1946–65 average for 11 regions and for each HadAM3 ensemble member, 1872–1998. The six red curves in each plot derive from simulations with SST bias corrections up to 1941 inclusive, with no corrections thereafter. The four blue plots are derived from simulations without bias corrections up to 1941

(e.g. Australian Bureau of Meteorology, 2004), though a limited number of improved station series have been created before 1910 (e.g. Della-Marta *et al.*, 2004).

3.2. Using collocated decadally filtered ensemble means

To clarify these results, we show time series of ensemble *means* of LSAT anomalies in Figure 4, filtered on approximately the decadal time scale using a 21-term binomial filter. Running estimates of twice the standard error of the ensemble member spread have been added. (Values at the beginning and end of each smoothed series are estimated by first extending the series by the average of the first and last five values.) Uncertainty estimates are made more stable by calculating those for both filtered uncorrected and filtered bias-corrected SST before 1942 separately and then pooling the variances at a given time. This is justified, because variations between ensemble members are due to internal variability and can be expected to be virtually the same at a given time whether bias-corrected or uncorrected SST data are used. When the collocated data are particularly sparse in earlier decades, the uncertainties can be seen to increase, though not greatly.

Figure 4(a) shows that decadally averaged bias-corrected and observed simulated global LSAT are generally within 0.1 °C of each other globally, and very clearly separated from the uncorrected simulations of LSAT. A slight apparent cold bias in the bias-corrected simulations in the tropics after 1920 (Figure 4(d)) may be due to warm biases in the observations as reported by Parker *et al.* (1995) due to the use of thatched shed thermometer screens in some tropical regions. Recent differences, particularly over the Northern Hemisphere land, are likely to reflect additional forcing of the observed land surface from increasing greenhouse gases, as mentioned above. The relative warmth in early observed Southern Hemisphere data, especially in the extratropics, is influenced by the biases in observed Australian data indicated in Figure 3(k). Figure 4(g) shows a rather discordant simulation of extratropical North American temperatures, as noted above, on the interannual time scale. By contrast, the bias-corrected decadal mean simulation for Europe is remarkably similar to the observations and clearly warmer than the uncorrected simulation. Similarly consistent results are seen in extratropical Eurasia before 1941. Note the lack of simulated warmth after 1970 in extratropical Eurasia; this is partly due to non-inclusion of the direct effect of greenhouse gases, but is also likely to be due to errors in simulating too weakly the winter Arctic oscillation trends that have resulted in greater warm air advection in winter. Other analyses clearly show this problem (A. Scaife, personal communication), much as seen in the earlier HadAM2a model (Folland *et al.*, 1998). Figure 4(j) shows a reasonable bias-corrected simulation of temperature changes between 1900 and 1940 in extratropical South America. Before that date the observed data are very sparse, giving large observed interannual to decadal fluctuations. Consequently, this period should probably be ignored. Figure 4(k), for Australia, shows the previously highlighted problems in the observed data. Despite this, the simulated LSAT data based on bias-corrected SSTs are clearly warmer and a better representation of the observations than the uncorrected data. Nicholls (personal communication) believes that the decadally averaged bias-corrected LSAT estimates for Australia provided by HadAM3 may be nearer the true values before 1910. This is discussed further in Section 3.3.

Figure 5(a) and (b) summarizes regional differences between the observations and the bias-corrected and uncorrected simulations, and their statistical significance. We have only used the ensemble means at this stage. The mean differences between the simulations and the observations have been calculated for the whole period 1872–1941. A t test has been applied, allowing for serial correlation in the differences. Given the remarks above, results for extratropical North America and Australia should be regarded with great caution.

The bias-corrected simulations of LSAT (Figure 5(a), also shown in Folland *et al.* (2001b)) have mean anomalies over 1872–1941 close to the observations (generally within 0.1 °C) and are either not significantly different or only just significantly different at the 95% confidence level, except for Australia, where the observations are significantly warmer. Even simulations for North America are not significantly different over the whole period. Nicholls *et al.* (1996) ascribe part of the errors in the observations to the widespread use of open thermometer screens before about 1910, which were too warm by day, especially in the warm season, due to the heating influence of diffuse shortwave radiation or longwave radiation from the warmer ground. However, these are not the only factors, as Figure 7(k) for the extratropical Southern Hemisphere suggests that the problem is worst in September–October, i.e. southern spring. Figure 5(b) shows that, over the same period, simulations using uncorrected SST are significantly too cold in all regions except extratropical South America (which is still cold). A t-test of the mean differences between the uncorrected and corrected simulations of surface air temperature over 1872–1941 shows that all regions, except extratropical South America, have significantly warmer simulated LSAT when forced with corrected SSTs.

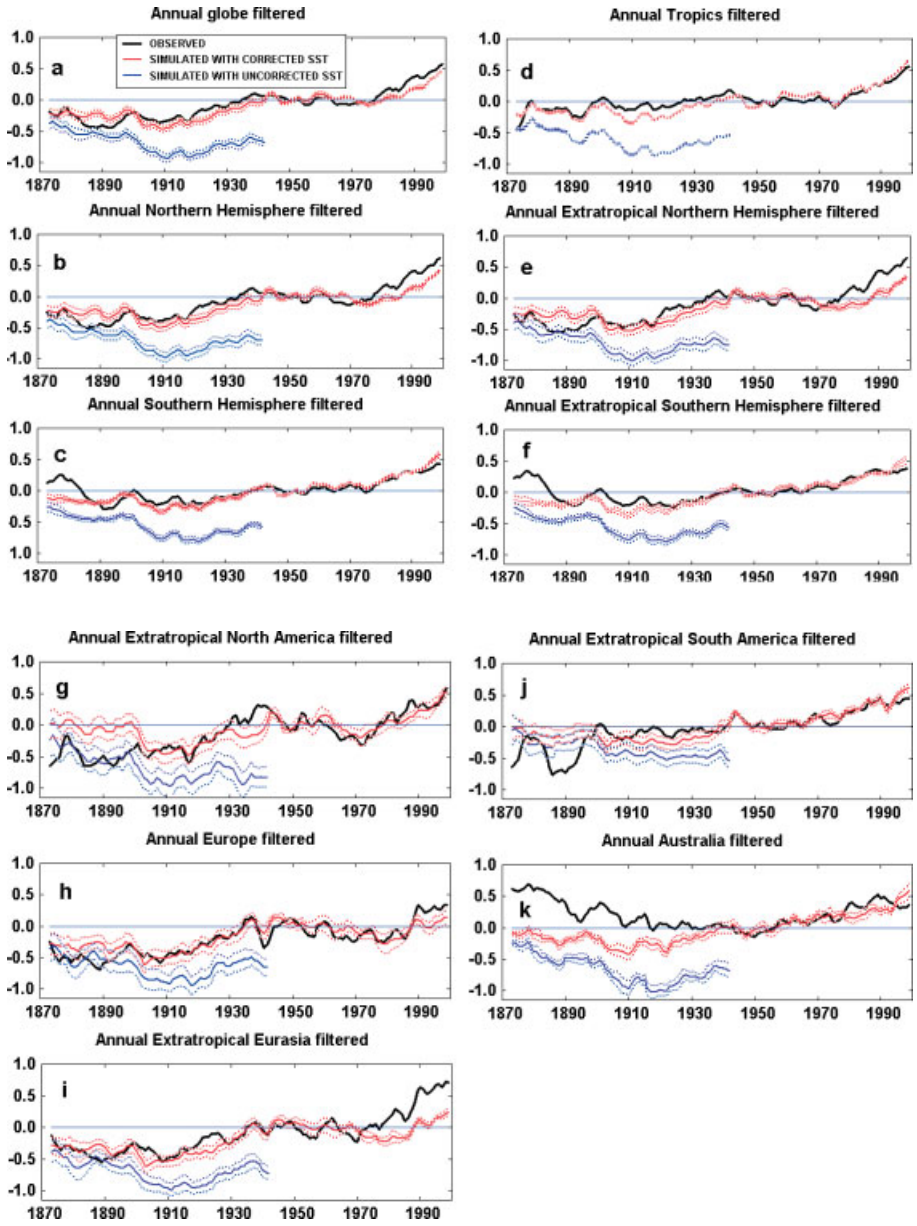

Figure 4. Near-decadally filtered (21-point binomial filter) collocated observations (black curve) and HadAM3 ensemble means of annual mean LSAT anomalies from a 1946–65 average for 11 regions, 1872–1998. Red (blue): with (without) bias corrections to SST. Red and blue dotted curves are ±2 standard errors of the ensemble means, calculated as in the text

3.3. Further investigation of apparent observed Australian temperature warm biases before about 1930

Della-Marta *et al*. (2004), based on earlier work by S. Torok (personal communication), have carried out studies of the problems of early Australian temperature data and have attempted corrections. Here, we take three station series that they have adjusted for various biases in southeast Australia. The model data were collocated as far as possible with the stations, allowing for the periods when each station was available. This was done by only using those model 2.5° × 3.75° boxes that included a station. Only an annual analysis of mean temperature is presented; the noise in the model ensemble means for such a small area is too large for seasonal analyses given the small ensemble sizes available here. Model data for this region are only

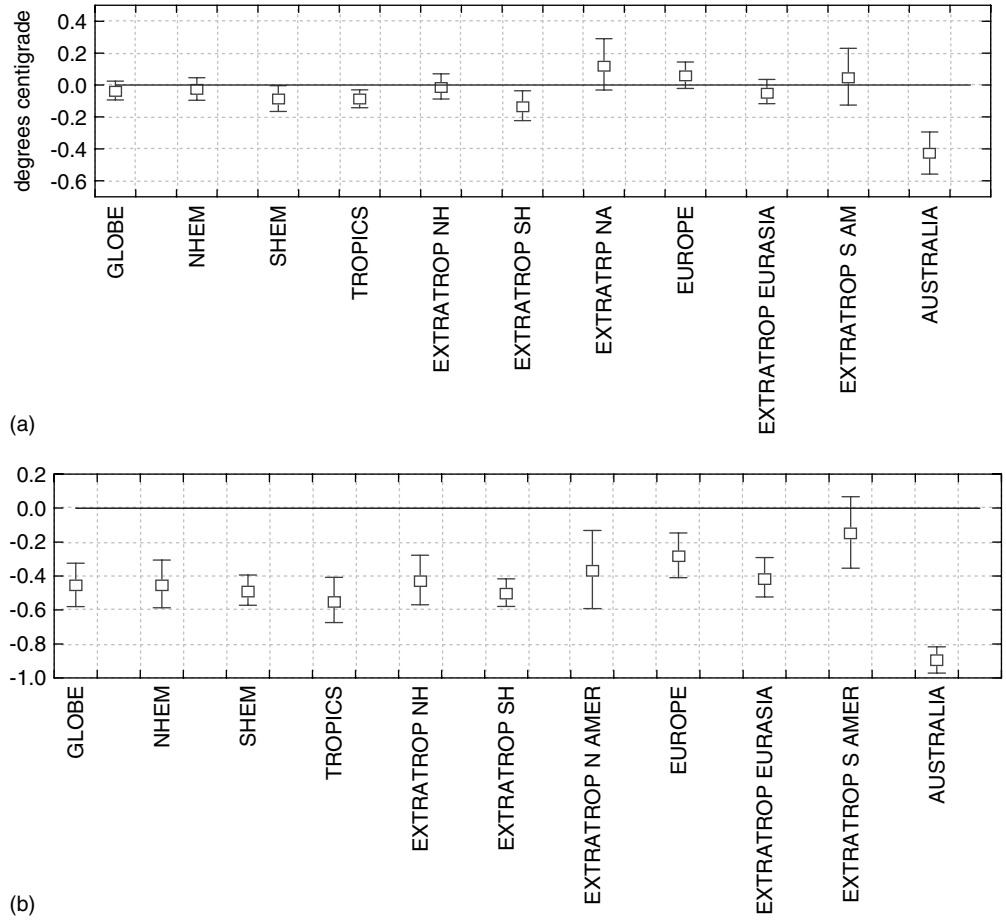

Figure 5. (a) Differences between bias-corrected simulations of annual mean LSAT anomalies from a 1946–65 average and observed (Jones *et al*., 1999) annual mean LSAT anomalies from the same period over the whole period of bias corrections, 1872–1941. (b) As (a), but simulations using uncorrected SST data

available for 1879–1990 because an early form of the southeast Australian data covered this period, and later model data were not kept. Data are available from Cape Otway (38°52′S, 143°31′E) and Wilson's Promontary (39°08′S, 146°25′E) from 1879, but with a gap in 1884 at Cape Otway. Richmond (33°36′S, 150°47′E) began in 1907. Decadally filtered data (Figure 6) were used, calculated as in Figure 4.

While Figure 6 shows that the average of these stations agrees somewhat better than the Jones *et al*. (1999) Australian data (Figure 4(k)) with the simulations using bias-corrected SST, the overall characteristics are the same in both figures. The simulations show cooler conditions than observed before 1930, though the offset is only significant before 1920, and then only marginally. Near 1900 the simulations do not differ significantly from the observations. However, before 1890, the simulations are clearly cooler. Importantly, the simulations show the coldest temperatures around 1910, whereas the observations are coolest much later, near 1950. For Australia as a whole using the Jones *et al*. (1999) data (Figure 4(k)), the bias-corrected simulations were significantly colder than the observations before 1930, with differences near $-0.8\,^{\circ}$C in 1880 and $-0.5\,^{\circ}$C around 1900–10. Figure 6 shows a difference of $-0.6\,^{\circ}$C around 1880 and nearly $-0.5\,^{\circ}$C around 1910. Simulated temperatures using uncorrected data are clearly colder still, although because of the greater noise the 95% confidence limits of simulations using corrected data overlap through most of the period.

We conclude that use of corrected observed data for southeast Australia reduces the differences between the observations and model simulated temperatures compared with Australia as a whole based on the Jones

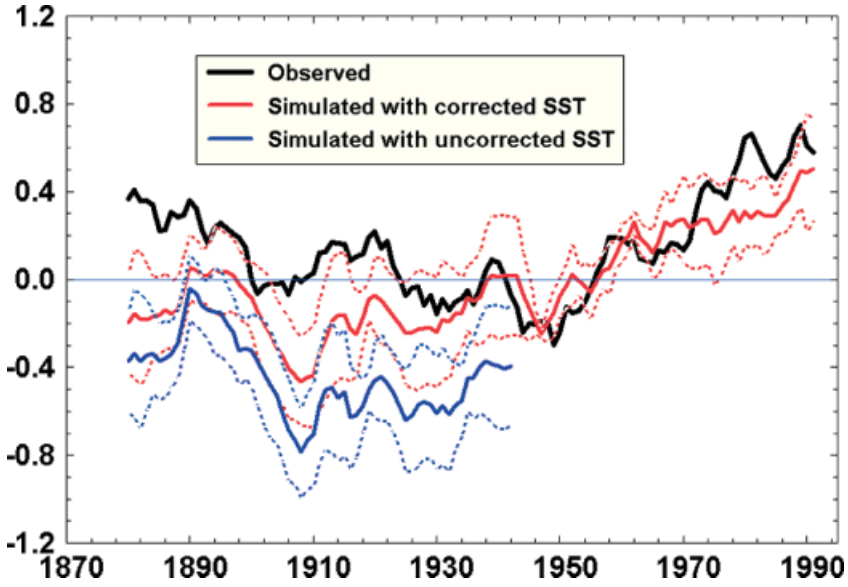

Figure 6. Approximately decadally filtered (21-point binomial filter) quasi-collocated observations (black curve) and HadAM3 ensemble mean anomalies from a 1946–65 average of annual mean LSAT for three corrected stations in southeast Australia, 1879–1990. Red (blue): with (without) bias corrections to SST. Light red and blue curves are ±2 standard errors of the ensemble means

et al. (1999) data set, but a substantial fraction of the disagreement remains before 1920. The disagreement is important, as a rather different picture emerges of the timing of lowest temperatures in southeast Australia and Australia as a whole during the last 120 years. However, the SST reconstruction used to force the model has greater uncertainty in such a region in the late 19th and early 20th centuries than, say, near Europe, because of data sparsity. An idea of these uncertainties in SST can be gained from observational analyses of SST and island surface air temperature anomaly time series published by Folland *et al.* (2003) for the adjacent southwest Pacific region. Here, all the main forms of uncertainty are assessed and uncertainty ranges are calculated for time series representing separate regions in the southwest Pacific, one extending back to 1871.

4. TESTS OF SEASONAL BIAS CORRECTIONS

To test the seasonal variation of the corrections, we assess the simulated LSAT over the extratropical Northern and Southern Hemispheres using decadally averaged LSAT as calculated in Figure 4. We use 2-month seasons January–February, March–April, etc. to give a good resolution of the seasonal cycle, but at the same time minimizing noise (Figure 7). The decadally averaged values are often more variable than those in Figure 4 because of the use of intrinsically more variable seasonal data. We note, as before, that the winter simulations may be the least accurate because of the cold bias in the model climatology of LSAT in both hemispheres. The overall impression from Figure 7 is that, after about 1900, seasonally varying uncorrected LSAT simulations are too cool compared with observations, but the bias-corrected LSAT simulations are generally not significantly different from observations in all seasons and both hemispheres. There is slight evidence of undercorrection in the extratropical Northern Hemisphere in summer in the 1920s and 1930s, and possibly in autumn. Otherwise, the modelled LSAT is not clearly distinguishable from the observed LSAT after about 1900.

Before 1900 the bias corrections are smaller, and both the SST and LSAT data are less numerous, especially for analysis of 2-month seasons as opposed to annual means. In the extratropical Northern Hemisphere, the cold values of observed LSAT in November–February reflect the unexplained relative coldness of the interior land surfaces relative to coastal land over much of this period noted by Parker *et al.* (1995). However, the

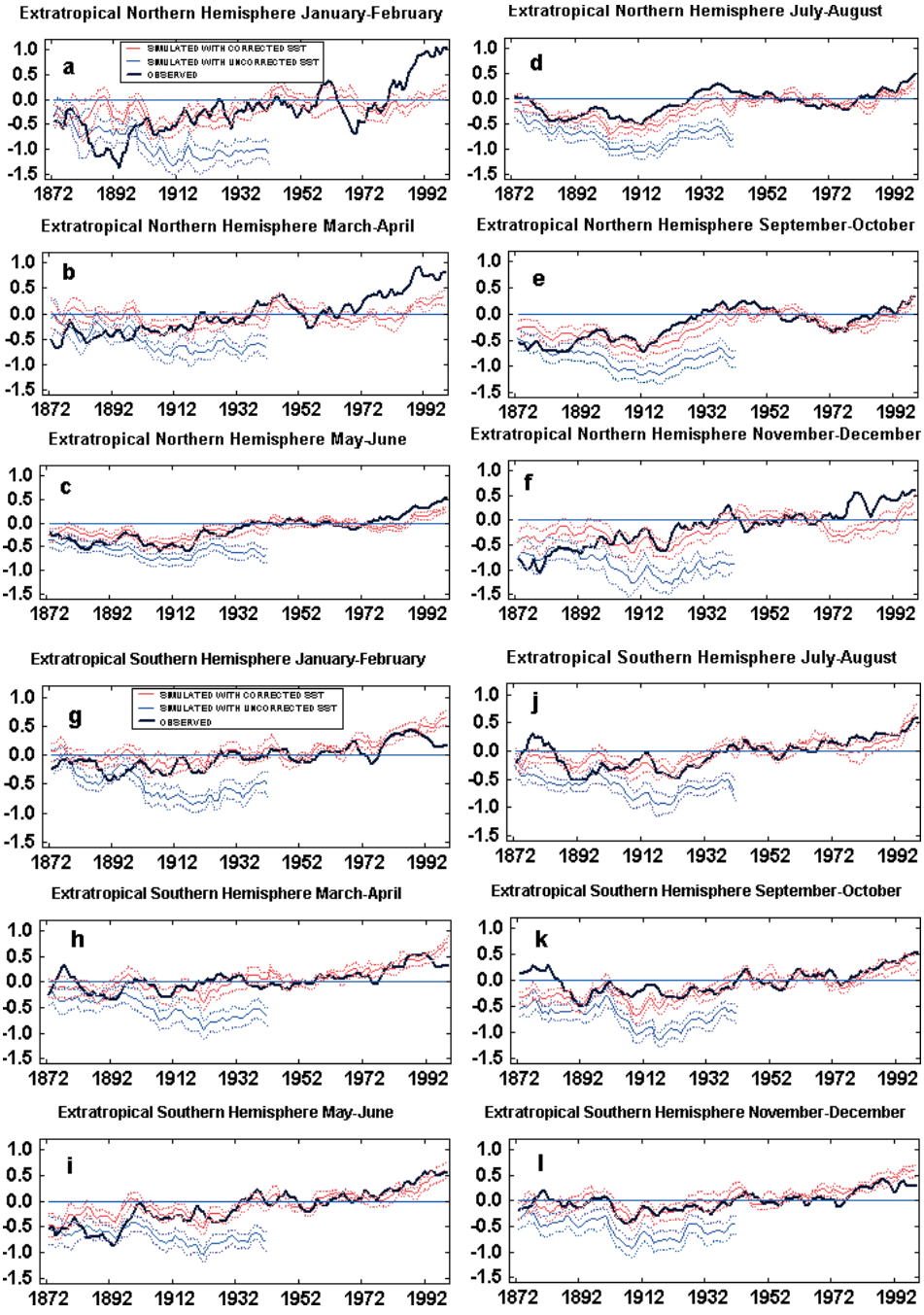

Figure 7. Simulations of seasonally varying approximately decadally averaged LSAT using SST with and without bias corrections for the extratropical Northern and Southern Hemispheres, 1872–1998, using the conventions of Figure 4. Two-month seasons January–February, March–April, etc. The observations are in black

results in spring and autumn also indicate slight overcorrection of bias-adjusted SST, perhaps by 0.1 °C, as reflected in the simulated LSAT. Figure 7 seems to indicate, however, that generally accurate adjustments were made in the summer season at this early period. A limitation on the interpretation of these results is that we do not have estimates of the uncertainty in the Jones *et al.* (1999) LSAT data. In the Southern Hemisphere

before 1900, the influence of the observed Australian warm LSAT is seen in many of the seasons (a seasonal analysis for Australia, not shown, confirms this).

We can summarize Figure 7 by noting that the seasonally varying corrections to SST on the space scale of the extratropical hemispheres seem well supported overall, but an overcorrection of perhaps 0.1–0.15 °C before the late 1890s (allowing for the additional sensitivity of simulated LSAT to SST changes) may exist outside the summer season in the extratropical Northern Hemisphere. However, error estimates are needed for the improved LSAT data now available before this assessment can be confirmed. If our results are confirmed, then the SST component of extratropical Northern Hemisphere surface warming since 1861, but not since 1901, may be underestimated by, say, typically 0.1 °C outside the 4-month summer season. This would give about 0.05 °C lack of warming since 1861 in combined observed LSAT and SST in September to April, and rather less than 0.05 °C in the annual full Northern Hemisphere mean of both quantities. The most likely reason for a slightly cooler late 19th century extratropical ocean than currently assessed is an underestimate of the fraction of late 19th century sea-temperature buckets made of wood by Folland and Parker (1995) and, therefore, an overestimate of the those made of canvas that lose more heat. An error of about −20% in the percentage of all (wooden and canvas) buckets that are assessed to be wooden centred around 1875–80 might account for these results. The relative number of these types of bucket at that time is admitted to be uncertain in that paper, and remains so today. The effect of such an error would be most obvious in Northern Hemisphere winter given the seasonal and geographical distribution of bias corrections reported in that paper.

Smith and Reynolds (2002) have tested the annual mean bias corrections, and their seasonal cycle, in another way. They use NMAT data, as mentioned before, using the version D NMAT bias corrections calculated by Bottomley *et al*. (1990) as their reference. They concluded that, overall, their method gave fairly similar results globally to that of Folland and Parker (1995), though there are differences regionally. Thus, the NMAT method only implicitly allows for evaporative effects in the biases of canvas buckets and tends to estimate lower corrections where evaporation is important, e.g. in the tropics. Smith and Reynolds (2002) noted that their seasonal cycle averaged over the zone 25–45 °N was generally larger than that of Folland and Parker (1995). In the period 1930–40, when corrections were largest but no wooden buckets were used by Folland and Parker (1995), their seasonal range of corrections was about twice as large, varying from +0.9 °C in December to 0.2 °C in July, whereas the Folland and Parker (1995) corrections ranged from just over 0.5 °C in December to about 0.25 °C in May and June. Smith and Reynolds (2002) used hydrographic data from Nansen reversing bottles to check this conclusion, as these data would be expected to be highly accurate. Insufficient data existed prior to 1930–40, so only that decade could be tested. This test also suggests that the Folland and Parker (1995) corrections have too low a seasonal cycle, but that their phasing is better than that of Smith and Reynolds (2002). The hydrographic data suggest that the largest correction should be 0.75 °C in November and zero in May (Smith and Reynolds (2002: Figure 7)). However, Figure 7 of this paper, based on the model simulations, does not support the hydrographic data in the 1930s, suggesting by contrast that the Folland and Parker (1995) corrections have about the correct seasonal cycle, except possibly a slight undercorrection in the period July–October.

The main message is that the twice standard error uncertainties in SST bias corrections may be quite large from about 1870 to 1890, especially in individual seasons, and any requantification of the uncertainties presented in Folland *et al*. (2001b) needs to allow for the uncertain mix in bucket types at this time. The seasonal cycle clearly needs retesting in the future in other decades using better LSAT data, as well as the new ICOADS SST data.

We conclude this discussion by noting that, during the period for which bias corrections are made (as well as afterwards), there is considerable observational consistency between collocated coastal corrected SST and coastal land air temperature anomalies on decadal time scales over most of the globe after about 1870. This was shown initially by Parker *et al*. (1995: Figure 17) when averaged for the Northern and Southern Hemispheres separately, for the tropics and for the globe. The tropics showed less consistency before about 1880, when the coastal tropical land was significantly colder, though the assessed uncertainties were very large due to few data. More recently, a high level of consistency in surface temperature trends across the land–ocean boundary has been shown by Folland *et al*. (2001a: Figure 2.9) in the 20th century on a 5° × 5° resolution. There are no discernable mismatches in annual trends across this boundary anywhere in the world

for 1910–45 or for 1901–2000, periods that include the SST bias corrections, or in periods that start later; however, this situation now seems to be changing, as LSAT appears since about 1990 to be warming decisively faster than SST (S. Tett, personal communication). This is likely due to the recent onset of faster warming of the land than the ocean that is expected from enhanced greenhouse gas forcing (e.g. Cubasch and Meehl, 2001). However, surface temperature trends in the interior of a given continent can also vary from those near its coasts as seen in Folland *et al*. (2001a: Figure 2.9) on time scales of a decade. This may explain some of the less-good regional results obtained above using the model, and particularly the winter results in the extratropical Northern Hemisphere in the late 19th century.

5. CONCLUSIONS

We have reported in detail a novel method of testing bias corrections to pre-1942 observed SST. This data set has a near 70% weight in estimates of changes in observed global surface temperature and, therefore, contributes substantially to the conclusion of the IPCC Third Assessment Report, that global temperatures rose over the 20th century, or from the late 19th century to the end of the 20th century, by $0.6 \pm 0.2\,^{\circ}\mathrm{C}$ (Folland *et al*., 2001a). The method described here uses an atmospheric climate model forced with observed SST and sea-ice extents with and without corrections to the SST, and is run in ensemble mode. The simulated LSAT is compared between these two ensembles and shown to follow the observed LSAT rather well globally and in many large geographical regions when the forcing SST data are corrected. The method is more sensitive than might be expected, as a given error in SST results in a rather larger error (about 30% larger) in simulated LSAT, at least on a global scale. This is likely to be due mainly to the additional influence of the greenhouse effect of changing water vapour.

The assumptions made in the correction procedure are mostly vindicated, and any changes to the corrections in future may be quite small unless substantial amounts of new SST data have different characteristics. The new raw ICOADS SST data generally seem to have similar bias characteristics, except around 1939–41 (Rayner *et al*., 2005). There is some evidence, however, that corrections in the late 19th century may be a little too high in the Northern Hemisphere winter, perhaps by a maximum of $0.15\,^{\circ}\mathrm{C}$. This could be due to an underestimation in the fraction of wooden sea-temperature buckets used at that time, at least in the extratropical Northern Hemisphere, which need less-positive corrections for their cool biases than canvas buckets. Any such bias disappears around 1900. Such an underestimation would have only small effects on a global annual average. However, as we note above, other authors come to oppositely signed conclusions with a comparable magnitude. Thus, SST bias corrections to at least 1890 have significant uncertainties that are probably greater than those assessed at that time by Folland *et al*. (2001b). We also noted that biases in observed LSAT in the late 19th century cannot be discounted, which might account for these results. In particular, Australian LSAT data in the Jones *et al*. (1999) (and later) data sets are very likely to be erroneously warm at that time and need correcting. This is important not only for Australian climate, as noted by Nicholls *et al*. (1996), but also because Australian LSAT anomalies make a disproportionate contribution to Southern Hemisphere LSAT anomalies in the late 19th century. Tests using corrected observed temperature data for southeast Australia only give moderately better agreement. Therefore, it is possible that reconstructed SST in the Australian region may be too cold in the early 20th century. This could be due to a lack of representativeness of the often sparse SST data or to insufficiently small positive bias corrections in the lower mid latitudes of the Southern Hemisphere at that time. The latter could be contributed to by an absence of wooden buckets in that region at that time, rather than the time-varying fraction currently assumed. However, this effect is too small to account for much of the observed difference.

Limitations of the method of testing SST bias corrections presented here are the lack of uncertainty estimates in the observed LSAT data and the fact that a better observed LSAT data set (Jones and Moberg, 2003) has recently become available. A smaller limitation is that absolute estimates of LSAT in the model are not quite the same as those of the real world, so the sensitivity of simulated LSAT to changes in SST may not be quite correct. However, Section 2.2 shows that these biases are mostly fairly small in HadAM3, with the possible exception of winter in both hemispheres, where they still only reach about $-3\,^{\circ}\mathrm{C}$ on the large space scales we use to test the seasonal bias corrections. However, a model with a better seasonal cycle of LSAT would help.

It is recommended that the tests carried out here are repeated on the new ICOADS SST analysis (Rayner *et al*., 2005) using the Jones and Moberg (2003) LSAT data set with uncertainty estimates. The new HadGAM1 atmospheric model is expected to be completed soon and could be used in preference to HadAM3, especially if its LSAT climatology is better. It would also be advantageous if pre-1910 Australian LSAT data could be corrected for their warm biases across the whole continent and made available to the Jones and Moberg (2003) global data set. It would also be very desirable to include estimates of the uncertainty in the SST analyses themselves. Such estimates are currently being made for the ICOADS SST analyses (Rayner *et al*., 2005).

ACKNOWLEDGEMENTS

I acknowledge support from the UK Government Meteorological Research Contract and the UK Department of the Environment, Food and Rural Affairs contract PECD/7/12/37. Thanks are also due to Ian Macadam and Simon Brown for extracting the model data, Jeff Knight for data on the model temperature climatology, Simon Torok and Dean Collins for data on southeast Australian temperatures and David Parker, Tara Ansell and two anonymous reviewers for helpful comments.

REFERENCES

Australian Bureau of Meteorology. 2004. Climate Monitoring Bulletin Australia. (Available monthly from: Climate Analysis Section, National Climate Centre, Bureau of Meteorology, Melbourne, Australia.)

Bottomley M, Folland CK, Hsiung J, Newell RE, Parker DE. 1990. *Global Ocean Surface Temperature Atlas "GOSTA"*. HMSO: London.

Cubasch U, Meehl G. 2001. Projections of future climate change. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden P, Dai X, Maskell K, Johnson CI (eds). Cambridge University Press: Cambridge; 525–582.

Della-Marta P, Collins D, Braganza K. 2004. Updating Australia's high-quality annual temperature dataset. *Australian Meteorological Magazine* **53**: 75–80.

Diaz H, Folland CK, Manabe T, Parker DE, Reynolds R, Woodruff S. 2002. Workshop on Advances in the Use of Historical Marine Climate Data. *CLIVAR Exchanges* **25**: 71–73.

Folland CK, Parker DE. 1995. Correction of instrumental biases in historical sea surface temperature data. *Quarterly Journal of the Royal Meteorological Society* **121**: 319–367.

Folland CK, Sexton DMH, Karoly DJ, Johnson CE, Rowell DP, Parker DE. 1998. Influences of anthropogenic and oceanic forcing on recent climate change. *Geophysical Research Letters* **25**: 353–356.

Folland CK, Karl TR, Christy JR, Clarke RA, Gruza GV, Jouzel J, Mann ME, Oerlemans J, Salinger MJ, Wang S-W. 2001a. Observed climate variability and change. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden P, Dai X, Maskell K, Johnson CI (eds). Cambridge University Press: Cambridge; 99–181.

Folland CK, Rayner NA, Brown SJ, Smith TM, Shen SS, Parker DE, Macadam I, Jones PD, Jones RN, Nicholls N, Sexton DMH. 2001b. Global temperature change and its uncertainties since 1861. *Geophysical Research Letters* **106**: 2621–2624.

Folland CK, Salinger MJ, Jiang N, Rayner N. 2003. Trends and variations in South Pacific island and ocean surface temperature. *Journal of Climate* **16**: 2859–2874.

Jones PD, New M, Parker DE, Martin S, Rigor IG. 1999. Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics* **37**: 173–199.

Jones PD, Osborn TJ, Briffa KR, Folland CK, Horton EB, Alexander LV, Parker DE, Rayner NA. 2001. Adjusting for sampling density in grid box land and ocean surface temperature time series. *Journal of Geophysical Research* **106**: 3371–3380.

Jones PD, Moberg A. 2003. Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. *Journal of Climate* **16**: 206–223.

New M, Hulme M, Jones PD. 2000. Representing twentieth century space–time climate variability, II. Development of monthly grids of terrestrial surface climate. *Journal of Climate* **13**: 2217–2238.

Nicholls N, Tapp R, Burrows K, Richards D. 1996. Historical thermometer exposures in Australia. *International Journal of Climatology* **16**: 705–710.

Parker DE, Folland CK, Jackson M. 1995. Marine surface temperature: observed variations and data requirements. *Climatic Change* **31**: 559–600.

Pope VD, Gallani ML, Rowntree PR, Stratton RA. 2000. The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dynamics* **16**: 123–146.

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research* **108**(D14): 4407. DOI: 10.1209/2002JD002670.

Rayner NA, Brohan P, Parker DE, Folland CK, Kennedy J, Vanicek M, Ansell T, Tett SFB. 2005. Improved analyses of changes and uncertainties in sea surface temperature measured *in situ* since the mid nineteenth century. *Journal of Climate* submitted.

Sexton DMH, Grubb H, Shine KP, Folland CK. 2003. Design and analysis of climate model experiments for the efficient estimation of anthropogenic signals. *Journal of Climate* **16**: 1320–1336.

Smith TM, Reynolds RW. 2002. Bias corrections for historical sea surface temperatures based on marine air temperatures. *Journal of Climate* **15**: 73–87.
Soon WW-H, Legates DR, Baliunas S. 2004. Estimation and representation of long-term (>40 years) trends of Northern-Hemisphere-gridded surface temperature: a note of caution. *Geophysical Research Letters* **31**: L03 209. DOI: 10.1029/2003GL019141.