Improved Extended Reconstruction of SST (1854-1997)

Thomas M. Smith and Richard W. Reynolds

National Climatic Data Center

151 Patton Ave., Asheville, NC 28801

December 11, 2003

ABSTRACT

An improved SST reconstruction for the 1854-1997 period is developed. Compared to the version 1 analysis, in the western tropical Pacific, the tropical Atlantic and Indian oceans, more variance is resolved in the new analysis. This improved analysis also uses sea-ice concentrations to improve the high-latitude SST analysis, and a modified historical bias correction for the 1939-1941 period. In addition, the new analysis includes an improved error estimate. Analysis uncertainty is largest in the 19th century and during the two world wars, due to sparse sampling. The near-global average SST in the new analysis is consistent with the version 1 reconstruction. The 95% confidence uncertainty for the near-global average is 0.4°C or more in the 19th century, near 0.2°C for the first half of the 20th century, and 0.1°C or less after 1950.

1. Introduction

Sea surface temperatures (SST) are important to the study of the earth's climate. For this purpose it is important that the SST analyses be global and as consistent as possible for the historic record. This is difficult because of changes in the observing system. As discussed in many sources (*e.g.*, Smith, *et al.* 1994) the historic distribution of in situ SST data from ships has varied with time due to a variety of economic and political changes (the opening of new canals, world wars, improved communication, *etc.*). In addition, biases in the ship in situ data have occurred as observational techniques have changed, and those biases must be corrected, as discussed in Folland and Parker (1995) and Smith and Reynolds (2002). Beginning in the 1970s, SST *in situ* data began to be available from drifting and moored buoys. Initially the number of observations from buoys was small, although the present number is comparable to the number of ship observations. Infrared (IR) satellite SST retrievals became available in November 1981, as discussed by Reynolds *et al.* (2002, henceforth REA), and are now supplemented with satellite SST microwave retrievals. The coverage from satellite data greatly improves SST coverage and now overwhelms the *in situ* data. This is an improvement for climate SSTs but may also be a problem when biases occur as discussed by Reynolds (1993) and by Reynolds *et al.* (2003).

Smith and Reynolds (2003) introduced the Extended Reconstruction SST analysis (ERSST), a global SST analysis that attempts to be consistent over time, here referred to as ERSST version 1 (ERSST.v1). This was done by using the REA analysis from the period with satellite data to develop global spatial correlation scales defined by a set of spatial modes. However, the ERSST.v1 analysis is always computed by a fit to *in situ* data, even in the period with satellite data. The UK Met Office computed their own analysis (Rayner *et al.* 2003) using

a similar technique but using all available data, both *in situ* and satellite. Both of these analyses and others (*e.g.*, Smith *et al.* 1996, REA) were used for intercomparisons. The results showed several shortcomings in the ERSST.v1 analysis. The purpose of this paper is to correct these shortcomings with an improved version of the analysis, ERSST version 2 (ERSST.v2). In the sections which follow we first briefly review the methods used to develop ERSST.v1 so that this paper can be more easily understood. We then discuss the problems in ERSST.v1. In the following sections, we develop the modifications used to create ERSST.v2 and show intercomparisons of ERSST.v2 and some of the other analyses mentioned above.

2. ERSST.v1

The ERSST.v1 analysis is produced from the latest version of the Comprehensive Ocean Atmosphere Data Set (COADS, Slutz *et al.* 1985, Woodruff *et al.* 1998). This version is called COADS release 2.0, and it is used for both ERSST.v1 and ERSST.v2. The analysis uses monthly and 2° spatial super observations, which are defined as individual observations averaged onto our 2° grid. The super observations are produced after data screening, or quality control (QC), which is needed to eliminate outliers. The super observations are also corrected for historical biases before 1942 by the method described in Smith and Reynolds (2002). The combined satellite and *in situ* analysis of REA is used to develop spatially-complete statistics for our reconstruction. We averaged the monthly 1982-2000 REA analysis to the same 2° grid that we use for the COADS data. In addition, we computed a SST climatology for the same period.

The ERSST.v1 analysis is performed separately for the low- and high-frequency

components, which are then added together to form the total SST anomaly. The low- and high-frequency components are separated because the stationary statistics used for the high-frequency analysis are based on only nineteen years of SST anomalies, and thus may not adequately span interdecadal variations. The low-frequency analysis is computed by smoothing and filtering our QC'd super observation anomalies over 10° spatial regions using 15 years of data, to generate one low-frequency analysis per year. This low-frequency anomaly is removed from the observed SST super observations before the high-frequency analysis is computed.

The analysis of high-frequency anomalies uses our 2° version of the 1982-2000 REA analysis to define a set of anomaly-increment modes, or spatial patterns. Increments are defined as the difference between the present analysis or super observation and the previous month's analysis. The modes are computed using the method of empirical orthogonal teleconnections (EOTs, van den Dool *et al.* 2000), which are similar to empirical orthogonal functions. The EOTs used here are localized so that teleconnections of more than 5000 km are damped, and those of more than 8000 km are set to zero. Localization prevents the analysis of anomalies that are not locally supported by the observations. An objective method is used to select the set of EOT modes that are adequately sampled by the super observations. For the selected modes, a weight for each mode is found by fitting the set of modes to the super observations. Variance associated with the unselected modes is damped. The super observations used to compute the set of weights have been adjusted in two ways. First the low-frequency analysis has been subtracted from the observations. Second, the observations have been converted to data-increments. The high-frequency component of ERSST.v1 is then computed from the weighted sum of the EOT modes. The complete analysis is the high-frequency analysis plus the low-frequency analysis.

For the satellite period the standard deviation of ERSST.v1 (σ_{Ev1}) is similar to that of the NOAA OI data (σ_{NOI}) of REA in most regions. However, there are some regions where σ_{Ev1} is weaker than σ_{NOI}. These include the western tropical Pacific and Atlantic and the tropical Indian (Fig. 1). For comparison, the standard deviation from an *in situ* only OI analysis (σ_{IOI}) is also shown. The *in situ* OI analysis is produced using methods similar to the REA NOAA OI analysis, except that only *in situ* SST data are used and monthly anomalies are analyzed directly on the 2° grid. The 1982-1997 period is fairly well sampled, and all three have similar standard deviation in the Northern Hemisphere, where sampling is best. Both of σ_{NOI} and σ_{IOI} are comparable in the tropics, but σ_{Ev1} is weaker than the other two in the western tropical Pacific, Atlantic, and in the Indian. Those regions have low variations in all three, but σ_{Ev1} is weakest, which may affect the monitoring or modeling some climate variations.

As discussed in the following section, the ERSST.v2 analysis method will be devised to correct the low tropical variance. In addition, other improvements will be added. The first is to add sea-ice information to the analysis. This will improve the analysis at high latitudes where other *in situ* data are sparse. The second is an adjustment of the historical SST bias correction for the 1939-1941 period, which affects results most strongly in 1941. The third is to add improved error statistics as well as a spatial gridded field of the monthly error. This will provide an objective way to determine how the analysis accuracy changes with time and space.

3. ERSST.v2

We now discuss the changes that have been made in the ERSST.v2 analysis: improved

high-frequency analysis, use of a sea ice to SST conversion algorithm, adjustment of the bias correction, and improved error analysis.

3a. Improved high-frequency analysis

In many ways ERSST.v2 is nearly identical to ERSST.v1. They both separate the analysis into low- and high-frequency components, which are summed for the total analysis. Both use the same COADS SST anomaly super observations and the same low-frequency anomaly analysis. The difference between the two versions is the high-frequency analysis.

In ERSST.v1 anomaly increments were analyzed using anomaly-increment modes. There are some advantages to analyzing increments with increment modes. Increments are designed to include temporal information, and fewer increment modes may be needed provided that they can be accurately defined (Thiébaux 1997). However, month-to-month anomaly increments generally have a much weaker signal than the anomalies themselves. In addition, the modes used for analysis are computed from anomalies containing some random error, and the difference of two such anomalies can have twice the random error. Thus, the anomaly increments used for computing modes will tend to have a lower signal/noise ratio than the anomalies alone. In regions where the anomaly changes slowly this can make it difficult to accurately compute increment modes. This explains why ERSST.v1 had difficulty analyzing the western-tropical Pacific, where month-to-month variations tend to be small.

To overcome problems associated with increments, ERSST.v2 analyzes anomalies using anomaly modes. As with ERSST.v1, a satellite-based SST analysis is used to compute a set of M spatial modes, except that ERSST.v2 is based on anomalies. The high-frequency anomaly for

ERSST.v2 is then estimated using

$$A_h(x,t) = \sum_{m=1}^{M} E_m(x) w_m(t). \tag{1}$$

Here $E_m(x)$ is the mode, m, at each spatial area x, and the analysis weight for time t and mode m is $w_m(t)$.

For each time step the high-frequency anomaly is analyzed by screening out modes not adequately sampled, with screening done as in ERSST.v1. Screening rejects modes that have less than a critical percent of their variance sampled. For the selected modes we find the set of weights that minimize the squared error of the fit, compared to the available data, as in the ERSST.v1 analysis. The maximum number of anomaly modes used for analysis, 130, is larger than the maximum number of increment modes in ERSST.v1, which is 75 modes. More anomaly modes are used to ensure that we better resolve tropical variations. Note that 130 is the maximum number of modes used. Under sampled modes are screened out, with sampling checked at each time step. The method for determining if a mode is adequately sampled is discussed below and in Smith and Reynolds (2003, hereafter referred to as SR).

When a mode is under sampled its ERSST.v2 weight is not defined by fitting to the data. In those cases the weight is estimated using the autocorrelation for that mode, similar to ERSST.v1. The autocorrelations are estimated from the 1950-1997 period, when weights for all modes can be computed most of the time. For each mode, the undefined weights are estimated using the nearest defined weights in both the forward and backward directions and the

autocorrelation of the mode. If the time lags to the nearest defined weights in the forward and backward directions are *lp* and *lm*, and the lag-1 autocorrelation for the mode is r_m, then the damped estimate for the undefined weight is set to $w_e(t) = 0.5\ (w(t+lp)\ r_m^{lp} + w(t-lm) r_m^{lm})$. For months far from any defined weights, the estimate will damp to near zero. Undefined weights that are close to defined weights will be more continuous.

In addition to the autocorrelation damping described above, a way to include temporal information is to pool anomalies from several months about the analysis month. Here we test the analysis two ways: using only anomalies from the analysis month and using pooled anomalies from three months centered on the analysis month. For the pooled anomalies, where the anomaly is defined for the analysis month we use that value. Where the anomaly is not defined for the analysis month but is defined in one or both of the adjacent months, the anomalies from the adjacent months are used, averaging them if both are defined. As discussed below, both methods give similar results, but the pooled data gives slightly stronger analyses when sampling is most sparse.

Besides the quality control described by SR, the 2° monthly anomalies used for the ERSST.v2 high-frequency analysis are further screened. This was found to be necessary by first performing the analysis without the second screening. In that analysis, there were a few times and locations where the analyzed anomalies were much larger than other anomalies at that location. These anomalies were apparently due to biased data. For example, in the NINO 4 area (5°S-5°N, 160°E-150°W), the range of anomalies is nearly always between -1.5°C and 1.5°C. However, without the second screening, there is a time in the late 1910s when the NINO 4 average is less than -3°C. In that period sampling is sparse and the unusually large anomaly

appears to be caused by a few biased data that passed the initial quality control. There must be several observations with a similar bias or else the first screening would have removed them, while random errors are filtered out by fitting to modes. This was not a problem in ERSST.v1, which used fewer modes which had larger spatial scales than some of the modes used in ERSST.v2.

The additional screening removes super-observations with a high-frequency anomaly magnitude of greater than $k\sigma_{NOI}$, where k a positive number. We also retain all super observations with anomaly magnitudes of 1.0°C or less to avoid excessive screening where the variance is weak. To assign k, screening with k = 4 , 4.5, and 5 was tested and the change to the analysis was evaluated. For k = 4, the resulting ERSST.v2 analysis still contained occasional outliers like the NINO 4 outliers discussed above. For both k = 4.5 and 5, the NINO 4 example was cleaned up without damping the analysis, and there were few other large-magnitude anomalies in the analysis. We used the less restrictive k = 4.5 value.

Most anomalies removed by the additional screening are at high latitudes, where SST gradients are strong and small errors in latitude can produce large anomaly errors. In low latitudes relatively few are removed except for before 1880, around 1915, and around 1940. Analysis sampling is determined by how many modes are selected with the available sampling. The second screening usually only reduces the number of modes selected by one or two (out of the 70 or more modes typically selected). Therefore, the second screening does not greatly affect sampling error in the analysis. Note that the low-frequency analysis is exactly the same as was used in ERSST.v1, and does not use the second screening. Additional screening is not needed for the low-frequency analysis, since that analysis filters out the effects of the outliers.

The ERSST.v2 analysis was tuned to determine: 1) the critical sampling level needed to accept a mode in the analysis; and 2) whether to use 1 month of data or data pooled from 3 months. In ERSST.v1 SR used cross validation tests to determine that at least 15% of the variance of each mode should be sampled for the mode to be used in the analysis. Here we tested ERSST.v2 using 3-months of pooled data and critical values of 15%, 25%, and 35% of the variance sampled for each mode.

The variance from 15-year running periods for each test is computed, and averaged over the 60°S-60°N region. The square root of the ratio of that variance, to the average 1982-1997 NOAA OI variance, gives the relative strength of the analyses, here called the relative standard deviation. For all three critical values, the variance is similar after 1950, when sampling is relatively dense. Compared to the recent period, the 15% critical value gives high variability before 1950, when sampling is sparse (Fig. 2). The relative standard deviation is as much as 25% stronger than in the recent period, indicating that some poorly sampled modes may be fit in that period. For a 35% critical-sampling value, variability is lower before 1950, compared to the recent period. With a 25% critical-sampling value the relative standard deviation is nearly constant except before 1880, when it decreases due to sparse sampling. In addition, the relative standard deviation from ERSST.v1 is nearly the same as for the 25% critical sampling in ERSST.v2. Therefore, we use a 25% critical-sampling value for ERSST.v2.

The 1982-1997 standard deviation map for the ERSST.v2 using 3-month pooled data (Fig. 3) compares well to the NOAA OI standard deviation for the same period. It is an improvement over ERSST.v1, which has more damped tropical standard deviation. For the 1-month analysis the standard deviation in this period is nearly identical to the 3-month pooled

analysis. For both, nearly all modes are chosen in this recent period, but in earlier periods many more modes are lost in the 1-month analysis compared to the 3-month analysis. In the first half of the 20th century, using 3-month pooled data increases the number of selected modes by more than 20%. The increase is even greater in the 19th century. In addition, the spatial standard deviation for the 60°S-60°N region for ERSST.v1 and for the 3-month analysis are comparable (see Murphy and Epstein 1989 for a definition of spatial variance). However, for the 1-month analysis the spatial standard deviation is lower before about 1910. For these reasons we use the 3-month pooled data analysis for ERSST.v2.

3b. Merging of sea-ice information

As noted above, the REA analysis incorporates estimates of sea-ice concentrations into its high-latitude SST analysis. An algorithm that converts sea-ice concentrations to SSTs was developed by Rayner *et al.* (2003) for an extended historical SST analysis (HadISST). The HadISST analysis is based on in situ, and when available satellite based observations. Although different from COADS, for most of the analysis period that data are comparable to COADS. With the Rayner *et al.* (2003) method the SST is forced to the freezing temperature of sea water, T_f, as the ice concentration approaches 0.9 (*i.e.* 90% ice cover). The freezing point of sea water is here set to −1.8°C, except in the Great Lakes where it is set to 0°C. A relationship between co-located SST and sea-ice concentration is derived by a quadratic adjustment. The adjustment is defined as

$$T_Q = aI^2 + bI + c \qquad (3)$$

where I is the ice concentration and a, b, and c are empirically-derived coefficients, derived with

the constraint that $T_Q(I=0.9) = T_f$. Those coefficients are defined using the available data from the marginal ice zone (MIZ). *In situ* data in the MIZ are sparse, so both *in situ* and AVHRR satellite-based SSTs are used to derive the coefficients for a recent period. In the Arctic, coefficients are computed as a function of both month and longitude, with separate coefficients for several Northern Hemisphere regions (the Great Lakes, Baltic Sea, Sea of Okhotsk, Sea of Japan, and Gulf of Alaska). In the Antarctic the coefficients are computed only as a function of month because of limited data.

In Rayner *et al.* (2003), adjusted temperatures with concentrations less than 0.15 were smoothed into the no-ice analysis using Poisson blending. In our ice blending we test the quadratic adjustments, computing them the same way except that we simplify the ice-edge smoothing of T_Q so that the adjustment linearly merges with the no-ice analysis as I goes from 0.2 to 0.

Quadratic adjustments gives reasonable results, but there are potential problems because of reliance on region and month dependent empirical coefficients, which were computed to fit the available data in a fixed base period. If conditions change significantly over the analysis period from the base conditions, the method may not fully represent those changes. In addition, some of the coefficient fits were computed using satellite SST retrievals which have been shown to be biased, because of difficulty in distinguishing ice from open water (Rayner 2003, personal communication). Therefore, we developed a simplified adjustment method, which we here examine along with the quadratic method.

To test the sea ice to SST conversion algorithms, we use the sea-ice concentrations of Rayner *et al.* (2003). Those concentrations include Arctic summer concentrations from

microwave sea ice retrievals, which tend to underestimate summer concentrations due to the formation of melt ponds on the ice. Those satellite retrievals are combined with other observations, values taken from atlases, and climatology when there are insufficient data. The methods tested all set the analysis SST to the freezing point of sea water for ice concentrations of more than 0.9.

We test a piece-wise linear adjustment defined as

$$T_P = \begin{cases} T_o & I \le A \\ T_o + (I - A)(T_f - T_o)/(0.9 - A) & A < I < 0.9 \end{cases} \tag{4}$$

where A is the pivot point where the two pieces join, and is set between 0 and 0.9. The unadjusted analysis temperature is T_o. For A = 0, this gives a linear adjustment with no pivot. For A = 0.9 there is only an adjustment at I = 0.9. This gives a linear adjustment from T_o to T_f, as the concentration I ranges between A and 0.9. In practice the piece-wise linear method is similar to the quadratic method, which gives slowly changing temperatures at low concentrations and more rapid changes at high concentrations. However, the piece-wise linear adjustments require only one (monthly and regionally constant) parameter, compared to two monthly and spatially dependent parameters used for the quadratic adjustment.

To evaluate the different methods, we bin SST values from the unanalyzed COADS 2° super observations, along with the ERSST.v2 analysis adjusted using several sea ice to SST algorithms. For different ice-concentration intervals (i.e., intervals of I) the average SST from the algorithms is computed and compared to the COADS super observations, at locations where COADS observations are available averaged over time and spatially. Values are binned within

ice-concentration intervals of 0.1. Here we discuss comparisons for the 1980-1989 period over regions poleward of 60° latitude. Comparisons for other decades and also some regional comparisons showed similar results.

Table 1 gives the root-mean-squared difference (RMSD) of adjusted SST from COADS super observations, comparing concentrations between 0.1 and 0.8. The RMSD is weighted by the number of comparisons in each ice-concentration interval. For the 60°N-90°N region in summer, winter, and all months, the piece-wise linear RMSD decreases with increasing A, from $A = 0$ (referred to as PL0) $A = 0.9$ (referred to as PL9). As discussed below, we believe that the high-concentration COADS values are questionable, which makes the lower RMSD values for PL9 questionable. The quadratic adjustment gives similar RMSD values to PL6, but PL0 is clearly inferior to the quadratic adjustment. From comparison in the 60°S-90°S region, a similar conclusion can be drawn.

In the warm seasons for both hemispheres, the COADS super observations tend to not decrease very much with increasing concentration (Fig. 4). This may be due to more noise at high concentrations because there are fewer temperature observations with high concentrations (shown on the lower half of the panels), or due to bad match ups between the ice-concentration data and COADS from errors in either data set, or due to SST biases in high-concentration observations. Physically, the SST must approach T_f as I approaches 0.9, so the high-temperatures with high sea-ice concentrations are questionable.

The quadratic-adjusted SST and the piece-wise linear adjusted SST with $A = 0.6$ (PL6) give similar variations of temperature with ice concentration. Based on these comparisons, we use the PL6 adjustment for ERSST.v2. The same adjustment is also applied to the *in situ* OI

comparison analysis. The entire 1854-1997 period is adjusted using the sea-ice concentrations of Rayner *et al.* (2003).

3c. Bias-correction adjustment

Corrections of SST for historical biases are needed before 1941 due to differences in measurement practices before and after that year (Folland and Parker 1995). Before World War II most SST measurements were computed from water samples taken on deck in buckets. After the war, ship-engine intake measurements were more common. Folland and Parker (1995) suggest a set of bias corrections to account for the different measurement practices. Those corrections abruptly end after 1941, and there are no suggested corrections for after that year. Following Folland and Parker (1995), Smith and Reynolds (2002) developed a set of bias corrections using different methods. Due to sparse data during World War II, they were not able to resolve that period well enough to justify a more gradual end to the corrections. Thus, they also end their corrections abruptly after 1941. Those Smith and Reynolds (2002) corrections were used in ERSST.v1.

Around 1940, COADS release 1 was dominated by Japanese data (see Fig. 3 of Woodruff et al. 1998), but new (primarily US) data altered the release 2.0 COADS data used for ERSST. Over much of the oceans, the 1941 ERSST.v1 anomalies are noticeably warmer that the HadISST anomalies for that year. An analysis by Folland (2003, personal communication) indicates that COADS release 2 SST has a positive bias relative to the data originally considered by Folland and Parker (1995). This bias begins in 1939 and gradually increases until 1941, ending after 1941.

Folland (2003) found that most differences between COADS and the data previously considered by Folland and Parker (1995) disappear if the bias correction is reduced to zero linearly over the 1939-1941 period, rather than as a step after the end of 1941. Therefore, for that period we weight of the Smith and Reynolds (2002) bias correction by 1-*m*/36, where *m* = 1 for January 1939 to *m* = 36 for December 1941. That adjusted bias correction is used for ERSST.v2. In the future, more careful evaluation of biases in this period is planned.

3d. Error analysis

Analysis errors can be separated into three types of error: random error, sampling error, and bias error. These three components are independent, as discussed by Kagan (1979). Thus the total analysis error variance may be written as their sum

$$\varepsilon^2 = \varepsilon_R^2 + \varepsilon_S^2 + \varepsilon_B^2 \qquad (5)$$

where the subscripts *R*, *S*, and *B* represent random, sampling, and bias error variances, respectively. Random error variance in an analysis, ε_R^2, is from random errors in the input data which are analyzed. Since those errors are random they can be mostly averaged or filtered out of an analysis that incorporates enough data. Sampling error variance, ε_S^2, reflects the representativeness of the available grid of sampling and how that sampling changes in time. If correlation scales are very large then few data may be used to represent a region. With smaller correlation scales more data distributed over the region are needed. While the random error for a region can be reduced with more data at the same place, reducing sampling error requires that additional data be spread over the region. Bias error variance, ε_B^2, is due to systematic biases in the data or due to the analysis method. The pre-1942 SST data are corrected for known

systematic biases with respect to the later data (Folland and Parker 1995, Smith and Reynolds 2002). However, these bias corrections are not perfect. In addition, there can be systematic uncertainties in the data other than those corrected, since COADS is formed by merging different data sets from the ships of different nations, which used different sampling methods.

The ERSST.v2 analysis removes almost all random error. As discussed above and in SR, the low-frequency component of the analysis averages and filters a large number of data. The low frequency analysis requires a minimum of over 100 observations, and in addition there is spatial and temporal binomial smoothing as part of that analysis. So random error variance in the low-frequency analysis is reduced a factor of more than 100 compared to the random noise of individual observations. The individual observations are themselves quality controlled, so the low-frequency random error variance is very small.

The high-frequency component of the analysis is computed by fitting data to a set of spatial modes representing the covariance from a well-sampled period. Random errors should not project onto these modes, and therefore they should be filtered out of the high-frequency analysis. For ERSST.v1, SR computed the signal/noise variance ratio for several 30-year periods in the 19th and 20th century, and gave the 60°S-60°N spatial average of the ratio for each period. In ERSST.v1 the ratio was found to be about 30 for all periods. The noise error variance is here represented by ε_R^2, and this implies that the signal overwhelms that component of the error in ERSST.v1. In ERSST.v2 the ratios were similarly computed, and found to be between 32 and 34. This indicates that there may be some residual random error variance in ERSST.v2, but it is only about 3% of the analysis variance, σ_{Ta}^2. This is for the analysis with 2° spatial and monthly resolution. For spatially or temporally averages of the analysis, the noise error variance

is further reduced by dividing it by the number of 2° monthly regions averaged.

To better describe the error, let T be the true SST anomaly and T_a be the analysis SST anomaly. Then by definition, the total error variance is

$$\left\langle \left(T - T_a\right)^2 \right\rangle = \sigma_T^2 + \sigma_{Ta}^2 - 2\sigma_T \sigma_{Ta} r_{T,Ta} + \left(\left\langle T \right\rangle - \left\langle T_a \right\rangle\right)^2 , \qquad (6)$$

where the brackets denote averaging. The first three terms on the right-hand side of (6) involve the variance of the true and analysis SST anomalies and the correlation between them, $r_{T,Ta}$. Those three terms represent sampling and random error variance. The last term, $(\langle T\rangle - \langle T_a\rangle)^2$, is the bias error variance.

For analysis of the error it is useful to write the analysis SST anomaly, T_a, in terms of the true anomaly, T, as

$$T_a = \alpha T + \beta + R \qquad (7)$$

where α and β are constants, and R is the random noise. Random noise includes all of the uncorrelated differences, which can not be accounted for by $\alpha T + \beta$. Note that α determines the relative strength of the analysis, and the sampling error, ε_S^2, is reflected by α. The systematic bias is represented by β. Since R is random, its mean is zero and it is uncorrelated with anything else, and its variance is ε_R^2. Using this equation (7) we can define

$$\sigma_{Ta}^2 = \alpha^2 \sigma_T^2 + \varepsilon_R^2 ,$$

where σ_T^2 is the variance of the actual anomaly and $\alpha^2\sigma_T^2$ is the analyzed signal variance. Using these definitions, the correlation between the true SST anomaly and the analysis can be expressed as

$$r_{T,Ta} = \frac{\alpha\sigma_T}{\sqrt{\alpha^2\sigma_T^2 + \varepsilon_R^2}}.$$

As discussed above, the ERSST.v2 signal/noise variance ratio is very large, so $\alpha^2\sigma_T^2 \gg \varepsilon_R^2$, and $r_{T,Ta} \approx 1$. Thus, for ERSST.v2 $\varepsilon_S^2 + \varepsilon_R^2 \approx \varepsilon_S^2$ and we can simplify equation (6) to

$$\left\langle \left(T - T_a\right)^2 \right\rangle \approx \left(\sigma_T - \sigma_{Ta}\right)^2 + \left(\langle T \rangle - \langle T_a \rangle\right)^2.$$

The ERSST.v2 sampling error is caused by incomplete analysis of the true variations, due to analysis filtering or incomplete data. Even when there are enough data for all modes to be sampled, there may be variations not spanned by the set of modes. Thus, there may always be some residual sampling error. In addition, the true SST anomaly variance is not stationary over the analysis period, largely because of interdecadal variations. Because changes in the true SST anomaly variance are difficult to detect, we first estimate σ_T - σ_{Ta} for the high frequency. This gives the expression

$$(\varepsilon_S^2)_{hf} = \left(\sigma_T - \sigma_{Ta}\right)_{hf}^2$$

to compute only the high-frequency (*hf*) sampling error. As with the analysis, we separately compute the low-frequency sampling error, and add the independent components for the total sampling error.

The NOAA OI from 1982-1997 is used to estimate the true high-frequency SST anomaly variance, $(\sigma_{NOI}^2)_{hf} \approx (\sigma_T^2)_{hf}$, and we assume that it is stationary. Although the NOAA OI analysis

contains some noise due to its use of different data types and bias corrections for satellite data, it is dominated by satellite data and gives a good estimate of the truth. Since we use this for estimating the high-frequency error, we remove the linear trend from the NOAA OI data before computing the variance to minimize its interdecadal variations. Similarly, for the analysis we remove the linear trend for running 15-year periods centered on the time of interest, and compute the analysis high-frequency variance in that period to estimate the high-frequency sampling error. Because of the possibility of seasonality in the sampling error, we use only data from the appropriate season for seasonal or monthly estimates.

The low-frequency sampling error is computed by estimating how well the available sampling can estimate a linear trend. A trend of 0.5°C/100 years is assigned everywhere on the globe. That is approximately the magnitude of the trend of global-average temperature over the 20th century. For each year the low-frequency sampling error is estimated by sampling that constant trend using the observed sampling for the year, with sampling held constant over 100 simulated trend years. The mean-squared difference between the actual trend and that sub-sampled low-frequency analysis of the trend defines the low-frequency sampling error variance for the given year. Repeating the process using the sampling of other years gives the low-frequency sampling error for those years. Note that this estimate of the low-frequency sampling error is similar to the estimate in SR and Smith *et al.* (2002), who estimated the low-frequency sampling error using coupled GCM SST output that was sub-sampled to match historical sampling.

The bias error variance, ε_B^2, is estimated similar to in SR, using the mean-squared-difference between the Folland and Parker (1995) bias correction and the adjusted Smith and

Reynolds (2002) bias correction. This difference is assumed to represent uncertainty in the bias correction. Both bias corrections are zero after 1941. Since there may be some residual bias after 1941, we reduce the difference linearly from its 1941 value to zero at 1945. In addition, to account for the 1939-1941 adjustment, we increase the difference in that period by a factor proportional to the magnitude of the adjustment. With maximum adjustment the factor is 2 in December 1941. The mean-squared difference is computed over the time period used to compute σ_{Ta}^2. A minimum ε_B value of 0.015°C is assigned whenever the measured value falls below the minimum, as discussed by SR.

The three error components, for annual and 60°S-60°N averages (Fig. 5), show that most of the error of the average is from the low-frequency analysis. Because of its damping when data are sparse, the low-frequency analysis may underestimate interdecadal variations. Problems are greatest before 1900 and in periods associated with the two world wars. The bias error can also be large prior to 1940, but it is generally less than the low-frequency error. The high-frequency error is usually the smallest, although it is occasionally larger than the bias error. Our bias error is similar in size to the estimate of Folland *et al.* (2001). However, because of the large low-frequency error estimate the overall ERSST.v2 error estimate for the near-global average is larger than their estimate. Sampling of the low-frequency (trend) part of SST needs to be greater than for the high-frequency part since the trend is not described by a stationary set of basin-scale modes. In Folland et al. (2001) a 52 year filled analysis was developed to describe covariance associated with both the low and high frequency, allowing potentially better low-frequency analysis. However, that analysis assumes that the 52-year period spans all important inter-decadal variations, while our analysis is slightly more conservative with the low-frequency

analysis.

4. Results

Regions where ERSST.v1 variability is too weak include the western tropical Pacific and Atlantic (Fig. 1). Time series of the NINO 4 area SST anomalies (Fig. 6) from both ERSST.v1 and ERSST.v2 show differences in the new analysis and differences from the HadISST analysis of Rayner *et al.* (2003). All three show similar interannual variations. Both ERSST.v1 and ERSST.v2 are more damped than HadISST before 1900. The HadISST analysis more closely follows the available observations when data are sparse, to maximize analysis signal, while the ERSST analyses are more conservative in sparse-sampling situations in order to minimize noise. The ERSST.v1 anomalies are more damped than the ERSST.v2 anomalies in this west-central tropical Pacific region, especially before 1960.

In section 3a (Fig. 2) we discussed the relative standard deviation for the 60°S-60°N region. Here we show the relative standard deviation for three smaller regions: 23°N-60°N, 23°S-23°N, and 60°S-23°S (Fig. 7). In both of the extra-tropical regions the ERSST.v2 analysis is slightly weaker than ERSST.v1. However, in the northern extra tropics ERSST.v2 is stronger than ERSST.v1 in the most recent period, when sampling is best. The Northern Hemisphere stronger variance in ERSST.v1 before 1960, when sampling is less, may be due to slightly less filtering of noise in that analysis. The Southern Hemisphere ERSST.v2 variance is smaller and changes less over time than the ERSST.v1 variance. Variance differences are larger in the Southern Hemisphere than in the Northern Hemisphere, and differences are largest in the most recent period, suggesting that ERSST.v1 is more sensitive to changes in sampling in that region.

For the tropics and Northern Hemisphere, HadISST variations are stronger after 1950. The HadISST analysis attempts to make maximum use of the available data, so the strength of its variations are more sensitive to changes in sampling.

In the tropics the ERSST.v2 variance is always larger than for ERSST.v1 due to better representation of the western tropical Pacific and Atlantic, as well as the tropical Indian. Both analyses have fairly constant tropical variance. The better representation of the tropics is the main advantage of ERSST.v2 compared to version 1. The HadISST analysis tends to be weaker between before 1950 and its strength increases as sampling increases later.

The RMSD between ERSST.v2 and HadISST (over the same 15-year running periods) shows an uncertainty similar to our total standard error (Fig. 8), which includes bias and sampling error. Differences are largest in periods of poor sampling, especially before 1900 and around the 1940s. Besides poor sampling around the 1940s, that period also has a large difference between the two SST bias corrections, which further increases the analysis differences. Since it is difficult to know which analysis is better, the RMSD between them could also be used as a measure of analysis uncertainty. The consistency of the standard error estimate with the RMSD suggest that our error estimate is reasonable.

Maps of monthly standard sampling error indicate regions where sampling is relatively good or bad. The ERSST.v2 normalized standard sampling error for January of 1900 and 1940 (Fig. 9, left panels) indicate normalized error less than 0.3 over most regions for those years. This includes both the low- and high-frequency sampling error. Standard errors discussed here are normalized with the NOAA OI standard deviation for the month. For 1900 the sampling problems are largest in the western tropical Pacific. For 1940 the largest sampling problems are

in the Southern Hemisphere. Both of these years have relatively large sampling error (Fig. 8), and sampling errors after 1950 are much smaller. The comparable sampling error maps for the *in situ* OI (Fig. 9, right panels) clearly shows the sampling along ship tracks. In that analysis, only densely-sampled regions have a normalized sampling error of less than 0.3, and there are many more regions with error above 0.9. Contrasted to ERSST.v2, the *in situ* OI analysis does not use large-scale spatial modes and therefore does not incorporate as much spatial-covariance information. The *in situ* OI also does not include as much temporal-covariance information. Because of its lesser use of these statistics, the *in situ* OI requires more dense sampling. The ERSST.v2 analysis requires less sampling, with the trade off that it filters the input data more than the *in situ* OI analysis.

5. Conclusions

The new analysis (ERSST.v2) is an improvement over version 1 because of its stronger variance in the western-tropical Pacific, its inclusion of ice-concentration information, and its improved error estimates. Averaged over large regions there is relatively little difference between ERSST.v2 and version 1. However, there are at times large differences between ERSST.v2 and the HadISST analysis of Rayner *et al.* (2003). Those differences generally lie within the 95% confidence interval for the analysis. For example, for the 60°S-60°N annual averages (Fig. 10, upper panel) there are large differences before 1900. In that period sampling is sparse, as reflected by the wide confidence intervals. The confidence interval is most narrow after 1950, when sampling is better than in earlier periods.

Temporal correlation of monthly HadISST and ERSST.v2 anomalies is highest in the

most recent decades and away from the Southern Oceans (Table 2). Except for that region and the Arctic Ocean, typical correlations are 0.7 or higher since 1940, but less than 0.6 before 1940. Correlations are lowest near Antarctica, as would be expected from the sparse data there.

Although HadISST is almost always within the 95% confidence intervals for ERSST.v2, an exception occurs in the 60°S-23°S annual averages for the period after 1980 (Fig. 10, lower panel). HadISST incorporates different types of data, including satellite estimates of SST when they are available, to help the analysis where data are sparse. Biases in the satellite data may cause the southern ocean cool bias relative to the COADS-based ERSST.v2. For the 60°S-60°N area the warm spike around 1940 is much more pronounced in ERSST.v2 compared to HadISST. That spike is due to an increase in tropical SSTs. Historical and paleo data suggest that there was a very warm El Niño episode at that time (*e.g.*, Rasmusson 1984, Evans *et al.* 2000). The COADS data indicate a stronger episode than HadISST, but the errors are large in that period due to sparse sampling.

For the 1982-1997 period, when the NOAA OI and the historical analyses overlap, comparisons can be made to the NOAA OI analysis. Since the NOAA OI analysis incorporates satellite and *in situ* data, we consider it to be a close approximation of the truth. The global spatial RMSD between the NOAA OI analysis and ERSST.v1, ERSST.v2, and HadISST (Fig. 11) show that both ERSST.v1 and HadISST have comparable overall skill. They both also show a large annual cycle in the RMSD, with larger differences around the middle of the year (when southern scean ship sampling is reduced). The ERSST.v2 analysis has consistently higher skill (lower RMSD), and also a smaller cycle in the RMSD. The improvement in ERSST.v2 RMSD is nearly the same at all latitudes except in the Arctic Ocean, where there is almost no difference

between all three.

All of these historical analyses show broadly similar variations which give us confidence in the results. Because data are sparse over much of the historical period, methods incorporating large-scale spatial modes of variation are needed to make the most efficient use of those sparse observations. In our comparisons, all methods except the *in situ* OI use large-scale spatial modes. However, even with the most efficient use of the available data, there are periods such as the 1940s and before 1880 when data are very sparse, and analyses from those times should be used with caution.

The monthly average ERSST.v2 analysis will be available on line along with the monthly error estimate at the NCDC web site. Analysis begins January 1853 and will be updated monthly. (The site address has not yet been created, but should be available to list before publication.)

Acknowledgments. We thank Vern Kousky, Yan Xue, Sam Shen, Sharon LeDuc, Scott Woodruff, and an anonymous reviewer for discussions and comments. In addition, the NOAA office of Global Programs provided support for some of this work.

REFERENCES

Evans, M.N., A. Kaplan, and M.A. Cane, 2000: Intercomparisons of coral oxygen isotope data and historical sea surface temperature (SST): Potential for coral-based SST field reconstructions. *Paleoceanography*, **15**, 551-562.

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

Folland, C.K., N.A. Rayner, S.J. Brown, T.M. Smith, S.S. Shen, D.E. Parker, I. Macadam, P.D. Jones, R.N. Jones, N. Nicholls, and D.M.H. Sexton, 2001: Global temperature change and its uncertainties since 1861. *Geophys. Res. Letts*, **28**, 2621-2624.

Kagan, R.L., 1979: *Averaging of Meteorological Fields*. Gidrometeoizdat, 212 pp. In Russian. [English translation, 1997. Kluwer, 212 pp.].

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

Rasmusson, E.M., 1984: El Niño: The ocean/atmosphere connection. *Oceanus*, **27**, 4-12.

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

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

Reynolds, R.W., N.A. Rayner, T.M. Smith, D.C. Stokes, and W. Wang, 2002: An improved in

situ and satellite SST analysis. *J. Climate*,**15**, 1609-1625.

Reynolds, R.W., C. L. Gentemann and F. Wentz, 2003: Impact of TRMM SSTs on a climate-scale SST analysis. Submitted *J. Climate*.

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.

Smith, T.M., R.W. Reynolds, R.E. Livezey, and D.C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9**, 1403-1420.

Smith, T.M., and R.W. Reynolds, 2002: Bias corrections for historic sea surface temperatures based on marine air temperatures. *J. Climate*, **15**, 73-87.

Smith, T.M., T.R. Karl, and R.W. Reynolds, 2003: How accurate are climate simulations? *Science*, **296**, 483-484.

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

Thiébaux, H.J., 1997: The power of the duality in spatial-temporal estimation. *J. Climate*, **10**, 567-573.

van den Dool, H.M., S. Saha, and Å. Johansson, 2000: Empirical orthogonal teleconnections. *J. Climate*, **13**, 1421-1435.

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.

FIGURE CAPTIONS

Fig. 1. Standard deviation of monthly SST anomalies over the 1982-1997 period from the NOAA OI, ERSST.v1, and the *in situ* OI. Values less than 0.2°C and more than 0.6°C are shaded.

Fig. 2. Ratios of temporal standard deviations from 15-year running periods, to the 1982-1997 NOAA OI standard deviation, averaged over 60°S-60°N. The ERSST.v2 ratios with different critical values are shown. For comparison the ERSST.v1 ratio is also shown.

Fig. 3. Standard deviation of monthly SST anomalies over the 1982-1997 period from the ERSST.v2 analysis using 3 months of data centered on the analysis month. Values less than 0.2°C and more than 0.6°C are shaded.

Fig. 4. Average 60° latitude to the pole SST from COADS and ERSST.v2 using the quadratic method and a piece-wise linear method. Temperatures are shown for the warm season in each hemisphere, and are only averaged where co-located with a COADS observation (upper). Also shown is the number of seasonal observations averaged for each SST (lower). The horizontal axis is sea-ice concentration in tenth.

Fig. 5. Standard error types for the ERSST.v2 annual and 60°S-60°N average. Error types are the low-frequency analysis error (L.F.), the high-frequency analysis error (H.F.), and the bias-correction error.

Fig. 6. NINO 4 area annual average SST anomalies from ERSST.v2, ERSST.v1, and HadISST.

Fig. 7. Temporal standard deviation ratios, as with Fig. 2, except for the areas 23°N-60°N, 23°S-23°N, and 60°S-23°S, for ERSST.v2, ERSST.v1, and HadISST.

Fig. 8. For the ERSST.v2 annual and 60°S-60°N average, the total standard error and the root-mean-square difference (RMSD) from the average HadISST analysis.

Fig. 9. For ERSST.v2 (left panels) and the *in situ* OI (right panels), monthly normalized sampling standard error for January of 1900 and 1940. Values below 0.3 and above 0.9 are shaded.

Fig. 10. Annual and spatial averaged ERSST.v2 with its 95% confidence interval (heavy lines) and the average HadISST anomaly (light line), for the 60°S-60°N region (upper panel) and the 60°S-23°S region (lower panel).

Fig. 11. Monthly values of global spatial RMSD from the NOAA OI analysis, for the 1982-1997 period. The RMSD for ERSST.v1, ERSST.v2 and HadISST are given. Time averages over the period are also given.

TABLES

Table 1. Average RMSD (°C) from COADS super observations for ERSST.v2 SST adjusted using the quadratic method (Quad), and the piece-wise linear method with *A* = 0.0 (PL0), 0.4 (PL4), ..., 0.9 (PL9). Results are given for the 1980-1989 period for December-February (DJF), June-August (JJA), and all months (All). Results are given for the two polar regions as indicated.

60°N-90°N

	Quad	PL0	PL4	PL5	PL6	PL7	PL8	PL9
DJF	1.55	1.68	1.45	1.42	1.38	1.34	1.32	1.32
JJA	1.77	1.84	1.53	1.50	1.47	1.45	1.44	1.44
All	1.63	1.68	1.46	1.43	1.41	1.39	1.38	1.38

60°S-90°S

	Quad	PL0	PL4	PL5	PL6	PL7	PL8	PL9
DJF	0.68	0.76	0.62	0.61	0.60	0.60	0.60	0.60
JJA	1.10	0.88	0.85	0.85	0.84	0.84	0.85	0.85
All	0.71	0.71	0.62	0.62	0.62	0.62	0.62	0.62

Table 2. Spatial averages of temporal correlation of monthly SST anomalies. The HadISST and ERSST.v2 monthly anomalies are correlated for the indicated periods, averaged over the listed regions.

	1881-1940	1941-1997
60°S-60°N	0.53	0.70
60°N-90°N	0.66	0.61
20°N-60°N	0.54	0.77
20°S-20°N	0.59	0.77
60°S-20°S	0.44	0.57
90°S-60°S	0.22	0.36