An Estimate of the Sampling Error Variance of the Gridded GHCN Monthly Surface Air Temperature Data

S. S. P. Shen and H. Yin

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada

T. M. Smith

National Climatic Data Center, Asheville, North Carolina

(Manuscript received 14 September 2005, in final form 25 September 2006)

ABSTRACT

The sampling error variances of the 5° × 5° Global Historical Climatological Network (GHCN) monthly surface air temperature data are estimated from January 1851 to December 2001. For each GHCN grid box and for each month in the above time interval, an error variance is computed. The authors' error estimation is determined by two parameters: the spatial variance and a correlation factor determined by using a regression. The error estimation procedures have the following steps. First, for a given month for each grid box with at least four station anomalies, the spatial variance of the grid box's temperature anomaly, $\hat{\sigma}_s^2$, is calculated by using a 5-yr moving time window (MTW). Second, for each grid box with at least four stations, a regression is applied to find a correlation factor, $\hat{\alpha}_s$, in the same 5-yr MTW. Third, spatial interpolation is used to fill the spatial variance and the correlation factor in grid boxes with less than four stations. Fourth, the sampling error variance is calculated by using the formula $E^2 = \hat{\alpha}_s\hat{\sigma}_s^2/N$, where N is the total number of observations for the grid box in the given month. The two parameters of the authors' error estimation are compared with those of the University of East Anglia's Climatic Research Unit for the decadal data. The comparison shows a close agreement of the parameters' values for decadal data. An advantage of this new method is the generation of monthly error estimates. The authors' error product will be available at the U.S. National Climatic Data Center.

1. Introduction

Gridded monthly temperature data are often used in studies of climate change. The HadCRUT3v dataset archived at the Climatic Research Unit of the University of East Anglia, United Kingdom (Jones et al. 2001), and the Global Historical Climatology Network (GHCN) dataset developed by the U.S. National Climatic Data Center (NCDC) are the two commonly used gridded datasets for studying climate changes. The errors of the datasets are needed to quantitatively assess the uncertainties of the changes. The error variances $\langle E_i^2 \rangle$ were estimated for the decadal time scale for the U.K. dataset (Jones et al. 1997, hereafter referred to as J97). The purpose of this paper is to estimate the sampling error of the GHCN 5° × 5° monthly surface air temperature anomaly data on monthly scales.

J97 was the first publication on the systematic calculation of the error variance of gridded data on the decadal scale by estimating two parameters: the average variance (s_0^2) of all the stations in a grid box, and the average intercorrelation dimensionless percentage (r) of these stations based upon the output of general circulation models (GCMs). Recently, Rayner et al. (2006) used this average variance and average intercorrelation approach and computed the error variances of the sea surface temperature anomalies. However, the sampling error variances at the monthly scale for the U.S. GHCN gridded monthly surface air temperature anomalies over the land have not been systematically estimated. These errors are explicitly needed for many applications of the GHCN data, including optimal spatial average and interpolation. The main result of the

Corresponding author address: Samuel S. P. Shen, Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182.
E-mail: shen@math.sdsu.edu

DOI: 10.1175/JCLI4121.1

present research provides the sampling error variance for each GHCN gridded monthly datum from 1851 to 2001.

The differences between the current dataset and the J97 dataset of error variances are listed below:

(i) Different datasets: Ours is for the GHCN gridded temperature data, and J97's is for the HadCRUT3v data.
(ii) Different time scales: Ours is for the monthly data, and J97's is for decadal data.
(iii) Different methods: Ours is based on spatial variances and a correlation factor estimated by regression, while J97's was based on the spatially averaged temporal variances and averaged intercorrelations.
(iv) Different results: Our error estimate attempts to take nonstationarity into account, and hence, a large spatial variance of a given month of a GHCN grid box leads to a large error variance, while J97 does not have this feature.

The rest of the paper is arranged as follows. Section 2 describes the GHCN monthly surface air temperature data. Section 3 introduces the methodology of error estimation. Section 4 explains the results of error variances. Section 5 contains our conclusions and discussion.

2. Data

The GHCN is a comprehensive, global, and station-based climate dataset composed by the NCDC scientists. This dataset includes temperature, precipitation, and pressure. The GHCN version 1 was released in 1992 and version 2 in 1997 (Peterson and Vose 1997; Peterson et al. 1998). [The adjusted monthly mean station temperature dataset from version 2 is used in this research, and the data were downloaded from the NCDC GHCN Web site http://www.ncdc.noaa.gov/oa/pub/data/ghcn/v2/ghcnftp.html. The data site includes both the "Temperature Station Inventory File" (740 kB) and "Adjusted Monthly Mean Temperature Data" (31 MB).] This version of GHCN records started from January 1702 and ended in February 2004. Before 1835, less than 10 stations existed over the entire globe during a given month. January 1835 had 27 stations (all of them were in the United States and Europe), and the number of stations increased to 158 stations in January 1851, to 446 stations in January 1881, until the maximum of 4230 stations was reached in July 1969, before the number of stations began to decrease. The number of stations dropped sharply during the period from 1990 to 1993, and then from 2003 to 2004. These two sharp decreases were due mainly to the time lag between the NCDC's data archiving and the station observations. The total number of stations contributing to the GHCN temperature data is around 6000. Since few stations were in the GHCN system before January 1851 and after December 2001, this paper assesses the data error between January 1851 and December 2001. The history of the number of stations from January 1835 to February 2004 is shown in Fig. 1. The spatial distributions of the stations of February 1853, August 1891, January 1940, and July 1973, representing both station-sparse and station-dense months, are displayed in Fig. 2.

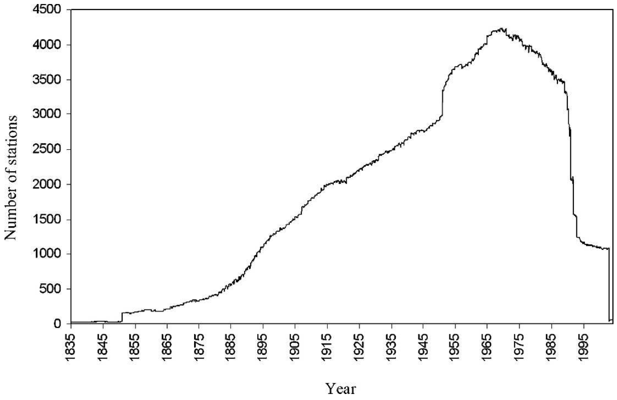


FIG. 1. History of the number of stations in the GHCN network from January 1835 to February 2004.

Two 5° × 5° station-dense boxes in the United States are selected to validate our error-estimation theory. They are (45°–50°N, 120°–125°W) in the western United States and (40°–45°N, 70°–75°W) in the eastern United States. The locations of the two boxes (highlighted) are shown in Fig. 3. The number in each box in Fig. 3 is the total number of stations, each of which appeared once in the box, but might not have continued throughout the entire time period from 1851 to 2001. For a given month, the number of stations with data was usually less than this number, since most stations had incomplete records. The validation box (45°–50°N, 120°–125°W) contained 55 stations. The maximal number of stations appearing in a single month was 52, which was reached in only 42 months including January 1970. The validation box (40°–45°N, 70°–75°W) contained 74 stations. The maximal number of stations appearing in a single month was 68, which was reached in only 78 months including March 1947.

3. Method

a. Basic formulas

It is well known that the formula of the standard error for the spatial average

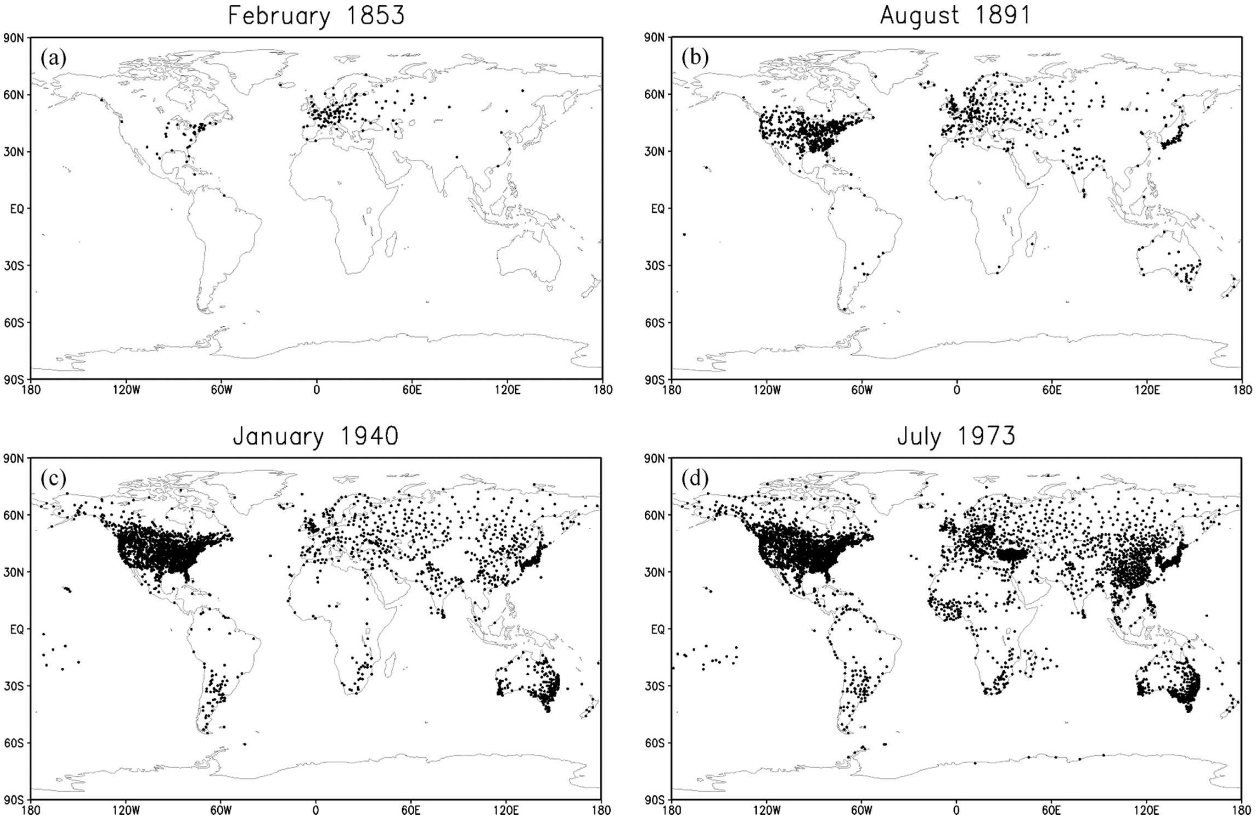


FIG. 2. Spatial distribution of stations.

$$E^2 = \left\langle(\overline{T} - \hat{\overline{T}})^2\right\rangle = \frac{s^2}{N} \tag{1}$$

is applicable to a spatial white-noise field $T(\mathbf{r})$. Here, $\overline{T}$ is the true average of the white-noise field over the spatial domain Ω whose area is $\|\Omega\|$

$$\overline{T} = \frac{1}{\|\Omega\|}\int_{\Omega} T(\mathbf{r})\, d\Omega. \tag{2}$$

Its estimator is

$$\hat{\overline{T}} = \frac{1}{N}\sum_{i=1}^{N} T_i. \tag{3}$$

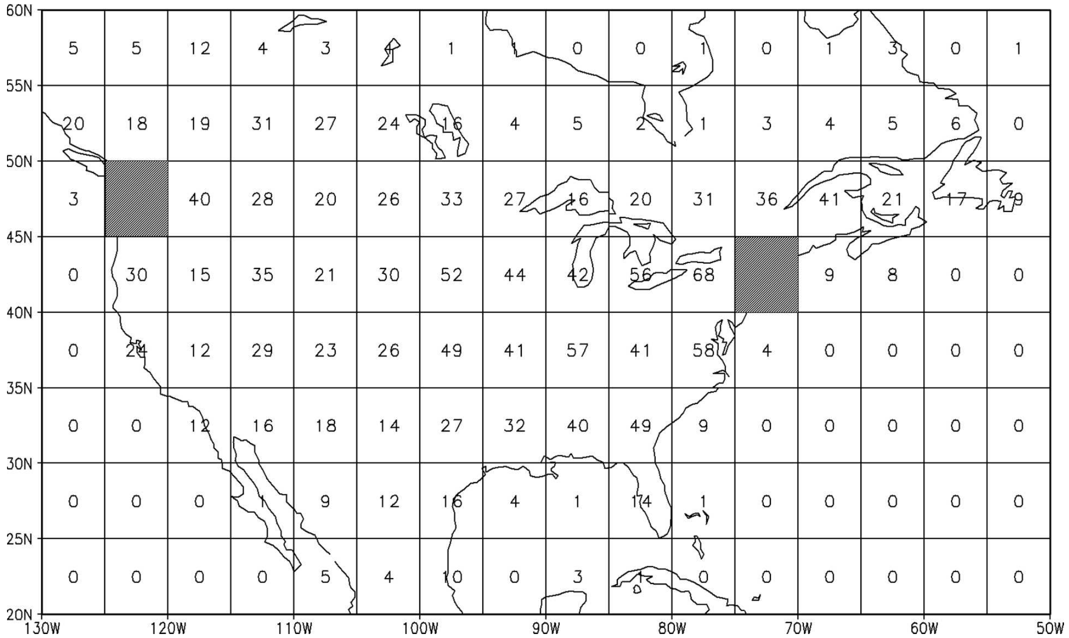
In the above, $T_i = T(\mathbf{r}_i)$ is a sampling datum, N is the number of samples, $s^2 = \langle T^2(\mathbf{r})\rangle$ is the uniform variance of the white-noise field, and $\langle\cdot\rangle$ denote ensemble mean. In statistical climatology, ergodicity is usually implicitly assumed; that is, the ensemble mean is estimated by the temporal mean. Thus, the above variance s^2 refers to the temporal variance. However, the monthly surface temperature anomaly in a 5° grid box is not a spatial white-noise field. One problem is how to quantify the error reduction due to the interstation correlations. A second problem is that the temperature field, even within a 5° × 5° box, may be inhomogeneous, and hence may have nonuniform variances at different station locations within the box, particularly for a grid box in a mountainous or coastal region. When the station data are intercorrelated and the temporal variance at different stations are nonuniform, then the alternative variance quantities of a grid box may be considered. The possible alternative variances include (i) the spatial average of the point temporal variances, (ii) the temporal mean of the spatial variances, and (iii) the relevant covariances. J97 used (i) and (iii), and we choose to use (ii) and (iii) in the current paper. J97 employed the concepts of spatially averaged correlation ($\overline{r}$) and the average of the point variance (s_0^2), and derived the error estimate formula

$$SE^2 = \frac{s_0^2(1 - \overline{r})}{N}, \tag{4}$$

where

$$\overline{r} = \frac{2}{N(N-1)}\sum_{i>j=1}^{N} r_{ij} \tag{5}$$

is the spatially averaged correlation, r_{ij} is the correlation between station i and station j, and



5	5	12	4	3	4	1	1	0	0	1	0	1	3	0	1
20	18	19	31	27	24	16	4	5	2	1	3	4	5	6	0
3	▨	40	28	20	26	33	27	16	20	31	36	41	21	17	9
0	30	15	35	21	30	52	44	42	56	68	▨	9	8	0	0
0	24	12	29	23	26	49	41	57	41	58	4	0	0	0	0
0	0	12	16	18	14	27	32	40	49	9	0	0	0	0	0
0	0	0	1	9	12	16	4	1	14	1	0	0	0	0	0
0	0	0	0	5	4	10	0	3	1	0	0	0	0	0	0

FIG. 3. Grid boxes in the contiguous United States and southern Canada, total number of stations in a box in the GHCN history, and two error-validation boxes in the United States: (45°–50°N, 120°–125°W) with 55 stations and (40°–45°N, 70°–75°W) with 74 stations.

$$s_0^2 = \frac{1}{N}\sum_{i=1}^{N} s_i^2 \tag{6}$$

is the spatial average temporal variance, since s_i^2 is the temporal variance of station i. J97 used the GCM's outputs to help estimate $\bar{r}$ and s_0^2. For a given season in a decade and a given grid box, J97's standard error varies only according to the mean number of stations in the grid box.

This paper uses spatial variances and a correlation factor to estimate the standard error of the grid box data. The appendix gives the following error formula:

$$E^2 = \left\langle (\hat{\overline{T}} - \overline{T})^2 \right\rangle = \alpha_s \times \frac{\sigma_s^2}{N}, \tag{7}$$

where

$$\sigma_s^2 = \left\langle \frac{1}{N}\sum_{j=1}^{N} [T_j(t) - \overline{T}(t)]^2 \right\rangle \tag{8}$$

is the spatial variance, and

$$\alpha_s = 1 + \frac{1}{N}\sum_{\substack{i,j=1\\ i\neq j}}^{N} \left\langle \frac{(T_i - \overline{T})}{\sigma_s}\frac{(T_j - \overline{T})}{\sigma_s} \right\rangle \tag{9}$$

is the correlation factor, which is less than or equal to 1. One can mathematically prove this conclusion.

If the temperature field is homogeneous in a grid box, then Eqs. (7) and (4) are the same. The advantage of error model (7) is its explicit accommodation of the spatial inhomogeneity.

b. Calculation of anomalies and estimation of σ_s^2

The first step is to calculate anomalies from the real readings of the temperature gauges. Our anomaly calculation method is the same as that of J97. For a given month, January to December, if a station has 21 or more years of data (i.e., 21 or more data entries) from 1961 to 1990, the station is then retained. The climatology is then computed as the mean of the at-least 21 data in the 30-yr period. The anomalies for a station are the departures from this climatology.

The error variance is calculated for the anomalies. For a given month in the GHCN history from January 1851 to December 2003, a grid box has K stations, of which N stations have anomalies, where $N \leq K$. The spatial variance σ_s^2 is estimated by

$$\hat{\sigma}_s^2 = \left\langle \frac{1}{N}\sum_{j=1}^{N} \left[T_j(t) - \hat{\overline{T}}(t)\right]^2 \right\rangle. \tag{10}$$

Then, what time window should be used for the temporal mean that replaces the ensemble average? How large should N be when the spatial variance is computed? Following the idea of piecewise stationarity and the moving time window (MTW) of Folland et al.

(2001), a 5-yr MTW is chosen for the temporal mean. For spatial variance, $N = 4$ is chosen as the minimum number of stations within a box, because the regression estimate of α_s needs at least four stations.

The spatial variance for the tth year, for a given month, is computed by

$$\sigma_{s,t}^2 = \frac{1}{N}\sum_{j=1}^{N}[T_j(t) - \overline{T}_N(t)]^2. \qquad (11)$$

This variance $\sigma_{s,t}^2$ varies from year to year. A 5-yr MTW smoothing yields an estimate for σ_s^2:

$$\hat{\sigma}_s^2(t) = \frac{1}{\|\mathrm{MTW}(t)\|}\sum_{\tau \in \mathrm{MTW}(t)} \sigma_{s,\tau}^2. \qquad (12)$$

Here, the 5-yr MTW(t) is centered around year t, and $\|\mathrm{MTW}(t)\|$ denotes the number of years in the MTW(t) with $N \geq 4$. In the 5-yr time window, N may vary from year to year. Thus, $\|\mathrm{MTW}(t)\|$ may be less than 5. If $\|\mathrm{MTW}(t)\| \geq 3$, the above calculation for $\hat{\sigma}_s^2(t)$ is implemented; otherwise, $\hat{\sigma}_s^2(t)$ is not computed.

c. Estimation of α_s

The correlation factor α_s is computed by using a regression. Suppose that a box has N (larger or equal to 4) station anomalies. We treat the data of these N stations as a statistical population. The population mean of the station temperature anomalies in the box is

$$\hat{\overline{T}}_N(t) = \frac{1}{N}\sum_{i=1}^{N} T_i(t). \qquad (13)$$

Simple random sampling of n stations is taken from the population (Cochran 1977). The sample mean of the n stations is

$$\hat{\overline{T}}_n(t) = \frac{1}{n}\sum_{i=1}^{n} T_{n,i}(t), \qquad (14)$$

where $T_{n,i}$ is the ith station's anomaly temperature in the subsample network of size n. The mean square differences between the population mean and the sample mean is estimated by

$$E_n^2 = \frac{1}{1000}\sum_{n \in S_{1000}} (\hat{\overline{T}}_N - \hat{\overline{T}}_n)^2, \qquad (15)$$

where S_{1000} stands for the set of 1000 simple random samples of size n. For a small N, 1000 samples have many repeated samples, while for a large N, 1000 samples do not exhaust all the sampling possibilities. In either case, the sample mean (15) is a good representation of the exhaustive sample mean.

Similar to (12), the 5-yr MTW is applied to E_n^2 to obtain the estimated MSE:

$$\hat{E}_n^2 = \frac{1}{\|\mathrm{MTW}(t)\|}\sum_{\tau \in \mathrm{MTW}(t)} E_n^2(\tau) \approx \langle E_n^2 \rangle. \qquad (16)$$

Again, similar to the estimation (12) for $\hat{\sigma}_s^2(t)$, the estimation (16) is implemented when $\|\mathrm{MTW}(t)\| \geq 3$. Thus, for each month and each grid box of N station anomalies, the $N - 1$ data pairs may be computed:

$$\left(\frac{\hat{E}_n^2}{\hat{\sigma}_s^2}, \frac{1}{n}\right) (n = 1, 2, 3, \ldots, N - 1). \qquad (17)$$

The least square regression between these data pairs estimates the α_s value. This regression is made for every grid box of N greater or equal to 4.

d. Interpolation of $\hat{\alpha}_S$ and $\hat{\sigma}_S^2$

The values of the estimated correlation factor $\hat{\alpha}_S$ have been calculated for the grid box at the month of at least four stations with temperature anomaly data. These values are interpolated onto other grid boxes so that every grid box has an $\hat{\alpha}_S$ value. The interpolation method is a kind of nearest-neighbor-assignment method on a sphere and has two steps. First, the values of $\hat{\alpha}_S$ are interpolated among the boxes of the same latitude. For a given grid box without an $\hat{\alpha}_S$ value, one searches to the west and to the east and assigns the value from the nearest box to it. If two boxes with data are found to be the same distance from the box, then the average of the two values is assigned to the box. This step fills up all the boxes on the 5° latitude band as long as this band contains at least one defined value. Second, for the 5° latitude band that does not contain any grid box with four or more stations, the $\hat{\alpha}_S$ values for the grid boxes of this band are interpolated from the north and south boxes. For any grid box that has not acquired values from step 1, one searches to the south first and assigns the $\hat{\alpha}_S$ value from the nearest grid box to it. If no value is found from the southern boxes, one searches to the north and assigns the value from the nearest grid box to it. The $\hat{\sigma}_s^2$ values are interpolated to the globe in the same way as the α_s values.

Thus, for each month from January 1851 to December 2001, each grid box over the globe has $\hat{\alpha}_S$, $\hat{\sigma}_s^2$, and N values, where N is the actual number of stations in the grid box rather than the number of stations with anomaly data. Of course, when $N = 0$, the error is not defined. The error variance of the GHCN grid box data for a given box and a given month is computed by

$$E^2 = \hat{\alpha}_S \frac{\hat{\sigma}_s^2}{N}. \qquad (18)$$

TABLE 1. Sensitivity test for the values of spatial variance $\hat{\sigma}_s^2$ and the correlation factor $\hat{\alpha}_S$ over the grid boxes (40°–45°N, 70°–75°W) and (45°–50°N, 120°–125°W) when the subsample sizes are n = 30, 20, 10, 9, 8, 7, 6, 5, and 4. The results are for the year 1975.

	Box (40°–45°N, 70°–75°W)				Box (45°–50°N, 120°–125°W)			
No. of stations	Jan		Jul		Jan		Jul	
	$\hat{\sigma}_s^2$	$\hat{\alpha}_S$	$\hat{\sigma}_s^2$	$\hat{\alpha}_S$	$\hat{\sigma}_s^2$	$\hat{\alpha}_S$	$\hat{\sigma}_s^2$	$\hat{\alpha}_S$
30	0.39	0.95	0.30	0.97	0.50	0.98	0.28	0.99
20	0.35	0.92	0.34	0.93	0.47	0.99	0.27	0.97
10	0.40	0.86	0.36	0.93	0.54	1.01	0.32	0.92
9	0.27	0.82	0.26	0.89	0.57	0.96	0.31	1.01
8	0.30	0.83	0.26	0.89	0.55	0.97	0.34	0.87
7	0.34	0.84	0.19	0.84	0.54	0.96	0.20	0.80
6	0.30	0.96	0.22	0.85	0.58	0.97	0.22	0.80
5	0.31	0.99	0.25	0.90	0.64	1.00	0.22	0.82
4	0.34	0.93	0.28	0.95	0.70	1.02	0.23	0.88

e. Sensitivity of $\hat{\sigma}_s^2(t)$ *and* $\hat{\alpha}_S$ *to* N *and validation of the error formula*

Two station-dense validation grid boxes in the United States (40°–45°N, 70°–75°W) and (45°–50°N, 120°–125°W) are used to test the sensitivity of $\hat{\sigma}_s^2(t)$ and $\hat{\alpha}_s$ to N and to validate the error Eq. (18). Each box contained over 30 stations in the MTW centered around 1975.

Table 1 shows the results of $\hat{\alpha}_S$ values for the sizes of the full samples N = 30, 20, 10, 9, 8, 7, 6, 5, and 4 stations for the two validation grid boxes for January and July 1975. Of course, the largest full sample gives the best approximation to the "true" α_s value for the grid box at a given month. The samples of 20, 10, . . . , 4 are subsets of the 30-station sample and are picked visually to maintain an even spatial distribution. The $\hat{\alpha}_S$ value differences among the nine samples are less than 20%. Considering the observational uncertainties of the monthly temperature, this difference may be attributed to the random errors. A few $\hat{\alpha}_S$ values in Table 1 are slightly greater than 1.0 due to the errors of regression estimation. The same test is conducted for $\hat{\sigma}_s^2$. The results in Table 1 indicate that $\hat{\sigma}_s^2$ fluctuations according to the sample size may be attributed to the sampling errors due to sampling locations rather than the sample size. Thus, this table supports the finding that $\hat{\sigma}_s^2$ is insensitive to the number of stations when the number is sufficiently large and the stations are well distributed.

The following will validate the regression error Eq. (18). Only the stations with complete records from 1959 to 1992 are used in this step. The grid box (45°–50°N, 120°–125°W) had 33 stations in January and 28 stations in July. The grid box (40°–45°N, 70°–75°W) had 41 stations in January and 40 stations in July. The simple average of all the stations is considered the "true" average, that is, the true grid box value. The square differences of the true average and the average of the subsamples enable the computation of the true mean square error (MSE). The 5-yr MTW mean of the true MSE $\hat{D}_n^2$ is computed for January and July for every year from 1961 to 1990. The solid curve of Fig. 4a shows the mean of the 30 values of $\hat{D}_n^2$, and the bars show the one standard deviation of the 30 values on each side. The same mean is computed for the estimated MSE $\hat{\alpha}_S\hat{\sigma}_s^2/n$ and is depicted by the dashed line in Fig. 4a. The closeness of the solid and dashed lines and the reasonable range of the one standard deviation imply the good fit of $\hat{\alpha}_S\hat{\sigma}_s^2/n$ to the MSE $\hat{D}_n^2$.

Figures 4b–d are obtained in a similar way. Again, these results all support the $\hat{\alpha}_S\hat{\sigma}_s^2/n$ error variance model.

4. Results

a. Global results of the sampling error variance on each grid box

According to the procedures described in the above section, the sampling error variance can be calculated according to Eq. (18) for each grid box that has stations from January 1851 to December 2001. Each month from January 1851 to December 2001 has an error map showing the error variance on each grid box with station data. Four error maps of selected months (February 1853, August 1891, January 1940, and July 1973) from data-sparse to data-dense cases are shown in Fig. 5. It is obvious that the fewer the number of stations, the larger the error variances. For the Northern Hemisphere, the errors are usually large in the north where the numbers of stations are few and spatial variances are large. Some coastal grid boxes also have large error variances, which are attributed mainly to strong temperature inhomogeneity and, hence, large spatial temperature variances. The large errors of the northern and coastal grid boxes for the Northern Hemisphere imply that smaller weights should be assigned to these grid boxes when their data are used to calculate the global or regional average or in spatial interpolation. Optimal averaging and interpolation methods should take these errors into account.

The maximal sampling error variance in a grid box from January 1851 to December 2001 was 9.506 $(°\mathrm{C})^2$ in January 1935 in some grid boxes in the latitude band 60°–75°N where the grid box had only one station and a large spatial variance. The minimal sampling error variance was 0.001 $(°\mathrm{C})^2$ and occurred in October 1993 due to large number of stations in this grid box.

To examine the temporal variations of the error vari-

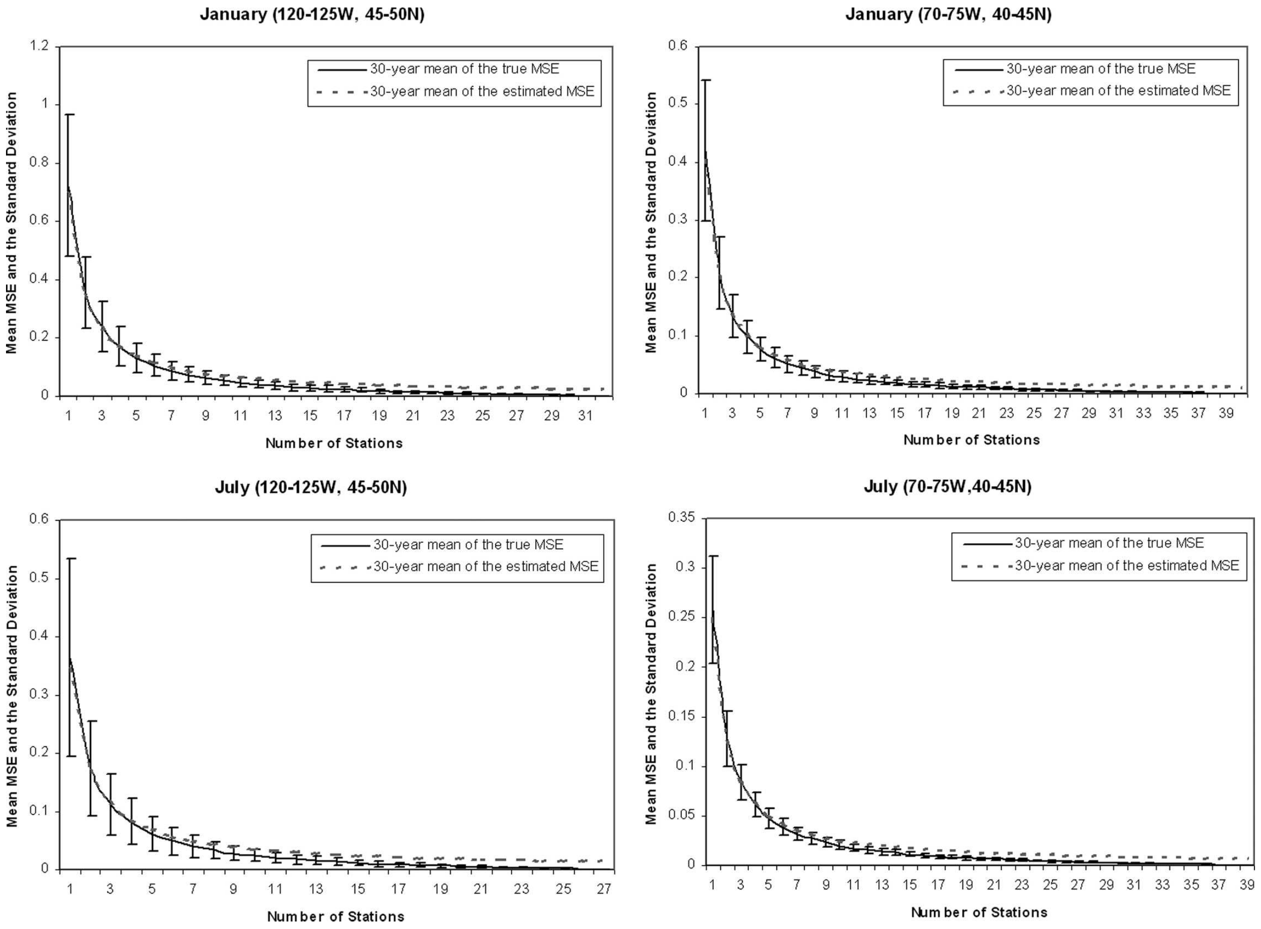


FIG. 4. The 1961–90 30-yr mean of the "true" MSE (°C^2; solid line) and the 1961–1990 30-yr mean of the estimated MSE (°C^2; dashed line) for the grid boxes (45°–50°N, 120°–125°W) and (40°–45°N, 70°–75°W). The bars on each side of the mean are the 1 std devs of the 30 "true" MSE values.

ance of a grid box in detail, the validation grid box (45°–50°N, 120°–125°W) is used. The monthly time series of the error variance for this grid box starts in December 1849 and ends in December 2001 (Fig. 6). This time series demonstrates two properties of the error: seasonality and station density. The error is larger in the winter months than in the summer months due to the larger spatial variances of the winter temperature. After 1890, the station density of this validation box became large, and the seasonal fluctuation of the error variance was effectively suppressed and became very small.

The global area-weighted average of the error variance for all the station-covered grid boxes is displayed in Fig. 7. The ratio of the station-covered areas to the earth's surface area is also displayed in the same figure. The clear seasonality of the globally averaged error variance is attributed mainly to the Northern Hemisphere data. After the early 1980s, despite the sharp decrease of the station-covered areas, the station density changed little for the grid boxes with data; hence the area-weighted average of the error variance did not go up sharply.

b. Comparison with J97's data

The outputs of J97's error estimates were in seasons defined by each consecutive three months starting from December to February (DJF) in every decade from 1851 for two time scales: interannual and interdecadal. The error variances were calculated for every 5° × 5° grid box with (87.5°N, 177.5°W) as the center of the first grid box, (87.5°N, 172.5°W) as that of the second grid box, and (87.5°S, 177.5°E) as that of the 2592th (i.e., the last) one. Error variance was computed for even the grid boxes without observational data, where it was slightly smaller than the spatially averaged temporal variance.

The output of our error values for the monthly 5° × 5° data also begins in January 1851, when the GHCN network had 158 stations. For each month from Janu-

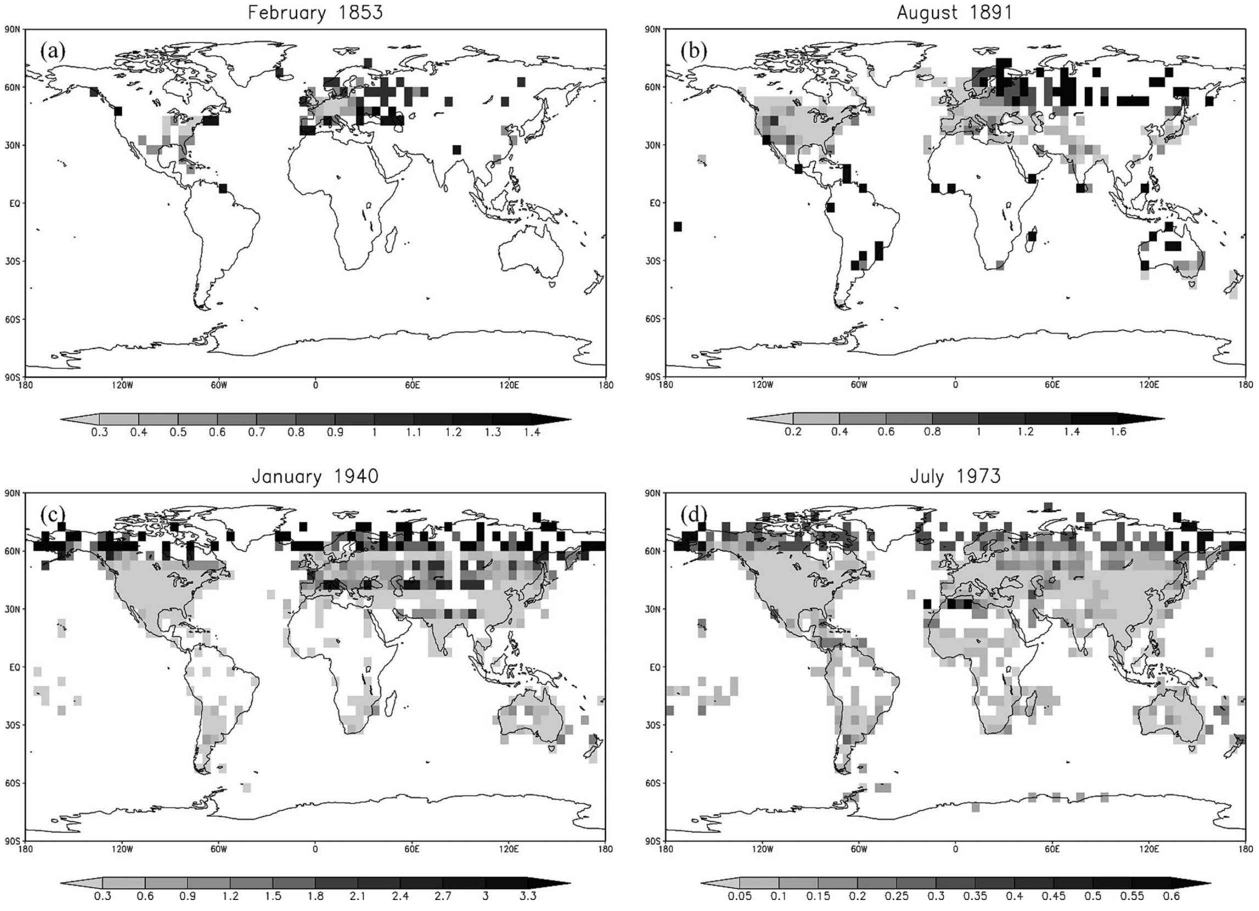


FIG. 5. Maps of the estimated sampling error variances of the grid boxes with data (°C^2) for four selected months.

ary 1851 to December 2001, if a grid box had observed data, this box's datum is assigned an error variance. A grid box without observations in a certain month has no error variance values and is assigned −999.000. The 1st and the 2592th grid boxes are arranged in the same order as those of J97.

Because of different temporal time scales and different output, it is not fair to directly compare the values of the error variances of J97 and ours from grid to grid and from month to month. A method of fair comparison is to recalculate the parameters of the error formulas of J97 and ours under the same time scale and the same number of sampling stations. From J97's error Eq. (4) and our error Eq. (18), the comparison can thus be made by comparing the values of the following two quantities, $s_0^2(1 - \bar{r})$ and $\hat{\alpha}_s\hat{\sigma}_s^2$, which are the error variance of a single station. Again the station-dense grid boxes (45°–50°N, 120°–125°W) and (40°–45°N, 70°–75°W) are used for the comparison since they allow one to estimate the "true" error variances and, hence, to compare J97's and our own error results with the true errors. The year 1975, which is in the middle of climatology period 1961–90, is chosen for the comparison.

Since the interdecadal and interannual means were used in J97, two MTWs of different window lengths are considered here: one of 5 yr and another of 29 yr. The 29-yr MTW covers almost the entire climatology period of 1961–90. The results computed from the 29-yr MTW are included in the round brackets (Table 2). The last column contains the values of the true error variance.

Table 2 implies the following: (i) Our errors and those of J97 are of the same order, but ours are consistently larger in both the 5- and 29-yr MTWs. The differences are large and in the range of 30%–100% in the 5-yr MTW, and they become small in the 29-yr MTW. (ii) In most cases, our error variances are closer to the true error variances than those of J97 (this difference is due mainly to our regression method of error estimate since our errors are the fits to the true errors when more than four stations are in a grid box). (iii) The parameter estimates are smoother for the 29-yr MTW than for the 5-yr MTW, but the realistic error variances

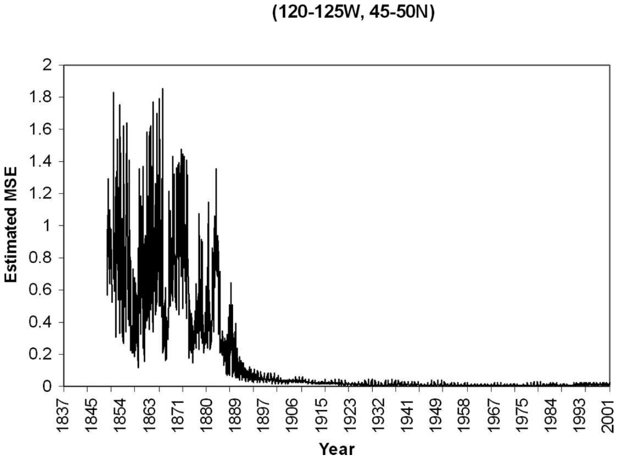


Fig. 6. Monthly time series of the estimated MSE (°C^2) for the grid box (45°–50°N, 120°–125°W) from December 1849 to December 2001.

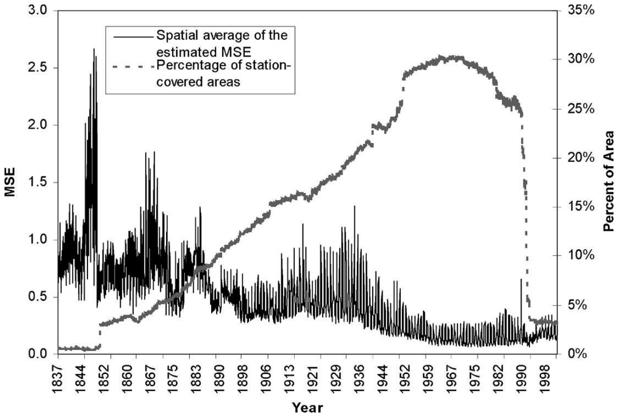


Fig. 7. Time series of the spatial average of the estimated MSE (°C^2; solid line) and the time series of the ratio of the station-covered areas to the earth's surface area (dimensionless percentage; dashed line) from January 1837 to December 2001. A grid box is station covered for a given month if it contains at least one station.

of the GHCN grid box temperature anomaly data should correspond to the less-smooth results because of the anomalies' spatial variations. Thus, the final product of our error variances for the GHCN data is calculated by using the 5-yr MTW.

For the interannual seasonal data, the maximal standard error (i.e., the square root of the error variance) of J97 is 4.184°C in DJF for the grid box (80°–85°N, 50°–55°E) for the periods of 1851–1920 and 1961–94, and

Table 2. Comparison between the results from Eq. (18) and J97's Eq. (4). The comparison is done for the year 1975. The 29-yr MTW results are in parentheses, and the other results are from the 5-yr MTW. The MSE(1) is the "true" MSE computed from the mean of the 1000 random one-station samplings.

	Box (45°–50°N, 120°–125°W)						
	s_0^2	r	$s_0^2\,(1-r)$	$\hat{\alpha}_S$	$\hat{\sigma}_s^2$	$\hat{\alpha}_S\hat{\sigma}_s^2$	MSE(1)
Jan	1.27 (4.18)	0.73 (0.86)	0.34 (0.58)	0.98 (0.97)	0.50 (0.75)	0.48 (0.73)	0.50
Feb	1.94 (2.70)	0.83 (0.84)	0.34 (0.45)	0.98 (0.97)	0.49 (0.52)	0.48 (0.50)	0.51
Mar	0.70 (1.53)	0.83 (0.86)	0.12 (0.22)	0.98 (0.97)	0.24 (0.23)	0.23 (0.22)	0.24
Apr	1.33 (1.38)	0.90 (0.87)	0.13 (0.18)	0.97 (0.97)	0.22 (0.19)	0.22 (0.19)	0.21
May	0.81 (0.98)	0.89 (0.80)	0.09 (0.19)	0.97 (0.97)	0.20 (0.20)	0.19 (0.20)	0.20
Jun	0.92 (1.54)	0.87 (0.81)	0.12 (0.29)	0.98 (0.97)	0.30 (0.31)	0.29 (0.31)	0.31
Jul	0.60 (0.99)	0.71 (0.68)	0.18 (0.32)	0.98 (0.97)	0.29 (0.32)	0.28 (0.31)	0.29
Aug	2.12 (1.54)	0.91 (0.82)	0.20 (0.28)	0.98 (0.97)	0.30 (0.31)	0.29 (0.30)	0.30
Sep	1.18 (1.75)	0.84 (0.85)	0.19 (0.27)	0.97 (0.97)	0.29 (0.31)	0.28 (0.30)	0.31
Oct	0.28 (1.15)	0.40 (0.79)	0.17 (0.24)	0.97 (0.97)	0.22 (0.26)	0.21 (0.25)	0.22
Nov	1.31 (2.69)	0.89 (0.90)	0.14 (0.26)	0.97 (0.97)	0.26 (0.31)	0.26 (0.30)	0.28
Dec	0.56 (3.35)	0.47 (0.86)	0.30 (0.48)	0.96 (0.97)	0.48 (0.61)	0.46 (0.59)	0.47
	Box (40°–45°N, 70°–75°W)						
Jan	8.05 (5.61)	0.96 (0.93)	0.30 (0.39)	0.97 (0.97)	0.43 (0.42)	0.42 (0.40)	0.44
Feb	2.29 (4.91)	0.89 (0.93)	0.25 (0.36)	0.95 (0.96)	0.35 (0.44)	0.34 (0.42)	0.36
Mar	3.54 (2.74)	0.95 (0.89)	0.19 (0.30)	0.97 (0.96)	0.44 (0.35)	0.43 (0.34)	0.45
Apr	2.66 (2.07)	0.94 (0.84)	0.16 (0.33)	0.97 (0.96)	0.34 (0.37)	0.33 (0.35)	0.36
May	2.70 (2.11)	0.89 (0.83)	0.30 (0.37)	0.96 (0.95)	0.44 (0.38)	0.42 (0.36)	0.43
Jun	1.27 (1.04)	0.85 (0.74)	0.19 (0.27)	0.97 (0.95)	0.30 (0.27)	0.30 (0.26)	0.31
Jul	0.71 (0.79)	0.74 (0.67)	0.19 (0.26)	0.96 (0.97)	0.29 (0.26)	0.28 (0.25)	0.29
Aug	0.85 (1.17)	0.80 (0.80)	0.17 (0.23)	0.98 (0.97)	0.31 (0.25)	0.30 (0.24)	0.30
Sep	0.48 (1.65)	0.62 (0.84)	0.18 (0.27)	0.96 (0.96)	0.28 (0.28)	0.27 (0.27)	0.26
Oct	2.56 (2.36)	0.92 (0.90)	0.20 (0.24)	0.97 (0.97)	0.33 (0.26)	0.32 (0.25)	0.32
Nov	3.17 (2.25)	0.95 (0.89)	0.15 (0.25)	0.96 (0.97)	0.26 (0.26)	0.25 (0.25)	0.26
Dec	4.53 (5.83)	0.94 (0.92)	0.29 (0.46)	0.96 (0.97)	0.48 (0.54)	0.46 (0.52)	0.48

the minimal standard error is 0.007°C in June–August (JJA) for the grid box (25°–30°N, 75°–80°W) for the period of 1991–94. For the interdecadal data, the maximal standard error of J97 is 3.267°C in DJF for the grid box (75°–80°N, 10°–15°E) for the decades of 1851–1920, while the minimal error is 0.002°C in several grid boxes and decades. Our maximal standard error is smaller than J97's even in the interannual scale, and our minimal standard error is larger than that of J97's interdecadal results but smaller than their interannual results.

5. Conclusions and discussion

The sampling error variances of the GHCN 5° gridded data have been estimated by using a regression approach. Our method yields results that are similar to those of J97, but our error values are consistently larger.

From our error estimation method, one might conclude that the errors are simply the results of data fitting, but this conclusion would be incorrect. The errors for most grid boxes are from the extrapolation or interpolation of two parameters: the spatial variance and the correlation factor. We examined the history of the number of grid boxes with the two parameters being computed and the total number of grid boxes with at least one station. The latter was usually over three times more than the former. A question arises: Is the interpolation of the two parameters valid? An answer to this question might need high-resolution (a grid box of 1° × 1° or finer) multiple GCM simulation results and the use of the spectral method for MSE over each grid box (Shen et al. 1994, 1998). The MSE expression by EOFs and their corresponding eigenvalues that were designed for global and regional optimal averages can be applied to the spatial average over a grid box. The simulation data from atmospheric GCMs with high resolution are needed to prepare the EOFs and eigenvalues. Coupled GCMs ought not to be used to generate this kind of EOF because the coupled GCMs do not yield a strict correspondence between the spatial and temporal grids between the model results and the observations. The land-surface-dependent EOFs derived from the high-resolution GCM results can explicitly take account of the distribution geometry of the station locations, the redundancy of clustering stations, and the contributions from the stations in the neighboring grid boxes.

Another approach that takes account of the geometry of the grid box and the station distribution is the geostatistical method. It uses empirical semivariograms for different regions (Isaaks and Srivastava 1989). This method is similar to the optimal average method used by Smith et al. (1994) and to objective analysis, which was developed by Gandin (1963). However, the assumed semivariograms cannot take atmospheric dynamics into account as well as the EOFs derived from the GCM output.

The error variances are needed in many applications of the gridded data, including the optimal average, optimal interpolation, data assimilation for climate models, and Bayesian posterior estimate for combining two preliminary estimates (Houghton et al. 2001; Shen et al. 2004; Smith et al. 1994).

Last, we discuss the size of errors for E in (18) introduced in the section 3d's interpolation of $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ for the grid boxes with less than four stations. Our spatial cross-validation implies that the standard error E using the interpolated $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values is likely to be in the range of (50%, 200%) of the actual E. The cross-validation procedure is as follows: (a) withhold the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values of a grid box with at least four stations, (b) interpolate the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values in the remaining grid boxes onto this box, and (c) compute the ratio R_E of E from the withheld $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values to the E from the interpolated $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values. For a given month, this is done for every grid box with at least four stations. Four particular months are analyzed in detail: July and January 1971, and July and January 1901. The results suggest that the errors introduced in the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ interpolation are mostly attributed to those for $\hat{\sigma}_s^2$ and the large errors are related to mountain and coastal grid boxes. For July 1971 and July 1901, all the boxes, except one, with at least four stations have their R_E within (50%, 200%). For January 1901, only two boxes with at least four stations are outside the range. For January 1971, the R_E ratio is in the range (40%, 310%), 190 boxes out of 209 are in the range (50%, 200%), and 138 boxes are in the smaller range of (75%, 150%). The largest $\hat{\sigma}_s^2$ difference is 7.7 (°C)^2, occurred over the grid box centered at (47.5°N, 112.5°W). The corresponding $\hat{\alpha}_S$'s difference is only 0.03. This box's new $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ values are interpolated from its western neighbor (47.5°N, 117.5°W). The average elevation of the (47.5°N, 112.5°W) box is about 1300 m, while that of the (47.5°N, 117.5°W) box is only 700 m, and the latter has smaller spatial variance. Consequently, $R_E = 220\%$. We may conclude that for over 90% of grid boxes the errors introduced in the $\hat{\alpha}_S$ and $\hat{\sigma}_s^2$ interpolation are very likely to limit the grid boxes' E values in the range (50%, 200%) of the actual ones, and the grid boxes over relatively flat regions like the eastern United States have even smaller ranges. Because the errors are mainly due to the $\hat{\sigma}_s^2$ interpolation, a more accurate assessment of the $\hat{\sigma}_s^2$ values will be desirable, and high-resolution and accurate atmospheric GCM output will be helpful in this assessment.

Acknowledgments. This study was supported in part by the National Oceanographic and Atmospheric Administration's OGP/CCDD program. Shen also thanks the U.S. National Research Council for its Senior Associateship award at NASA, the Chinese Academy of Sciences for an Overseas Assessor's research grant and for the Well-Known Overseas Chinese Scholar award. This work has been developed from Alan Basist, Samuel Shen, Claude Williams, and Guilong Li's initial research in 1999 on the GHCN error assessment. Discussions with Richard Reynolds and Thomas Karl were very helpful. The suggestions from anonymous referees improved the paper's presentation.

APPENDIX

Derivations of the Error Formula

The mathematics in the following derivation are motivated by the calculations of the ground truth problem for satellite observations (North et al. 1994). The mean square error is

$$
\begin{aligned}
E^2 = \langle(\overline{T} - \hat{\overline{T}})^2\rangle &= \left\langle \left[\frac{1}{N}\sum_{i=1}^{N}(\overline{T} - T_i)\right]^2 \right\rangle = \frac{1}{N}\left[\left\langle \frac{1}{N}\sum_{i=1}^{N}(\overline{T} - T_i)^2 \right\rangle + \left\langle \frac{1}{N}\sum_{\substack{i\neq j\\ i,j=1}}^{N}(\overline{T} - T_i)(\overline{T} - T_j) \right\rangle\right] \\
&= \frac{1}{N}\left[\sigma_s^2 + \left\langle \frac{1}{N}\sum_{\substack{i\neq j\\ i,j=1}}^{N}(\overline{T} - T_i)(\overline{T} - T_j) \right\rangle\right] = \frac{\sigma_s^2}{N}\left[1 + \frac{1}{N}\sum_{\substack{i\neq j\\ i,j=1}}^{N}\left\langle \frac{(\overline{T} - T_i)}{\sigma_s}\frac{(\overline{T} - T_j)}{\sigma_s} \right\rangle\right].
\end{aligned}
$$

REFERENCES

Cochran, W. G., 1977: *Sampling Techniques*. 3d ed. Wiley, 428 pp.

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

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields* (in Russian). Gidrometeoizdat, 238 pp. [English translation, 1966, Israel Program for Scientific Translation, 242 pp.]

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., 2001: *Climate Change 2001: The Scientific Basis*. Cambridge University Press, 881 pp.

Isaaks, E. H., and R. M. Srivastava, 1989: *An Introduction to Applied Geostatistics*. Oxford University Press, 561 pp.

Jones, P. D., T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate,* **10,** 2548–2568.

——, ——, ——, C. K. Folland, E. B. Horton, L. V. Alexander, D. E. Parker, and N. A. Rayner, 2001: Adjusting for sampling density in grid box land and ocean surface temperature time series. *J. Geophys. Res.,* **106,** 3371–3380.

North, G. R., J. B. Valdes, E. Ha, and S. S. P. Shen, 1994: The ground-truth problem for satellite estimates of rain rate. *J. Atmos. Oceanic Technol.,* **11,** 1035–1041.

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

——, T. R. Karl, P. F. Jamason, R. Knight, and D. Easterling, 1998: First difference method: Maximizing station density for the calculation of long-term global temperature change. *J. Geophys. Res.,* **103,** 25 967–25 974.

Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. Ansell, and S. F. B. Tett, 2006: Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *J. Climate,* **19,** 446–469.

Shen, S. S. P., G. R. North, and K.-Y. Kim, 1994: Spectral approach to optimal estimation of the global average temperature. *J. Climate,* **7,** 1999–2007.

——, T. M. Smith, C. F. Ropelewski, and R. E. Livezey, 1998: An optimal regional averaging method with error estimates and a test using tropical Pacific SST Data. *J. Climate,* **11,** 2340–2350.

——, A. N. Basist, G. Li, C. Williams, and T. R. Karl, 2004: Prediction of sea surface temperature from the Global Historical Climatology Network data. *Environmetrics,* **15,** 233–249.

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