Measurement Requirements for Climate Monitoring of Upper-Air Temperature Derived from Reanalysis Data

DIAN J. SEIDEL AND MELISSA FREE

NOAA/Air Resources Laboratory, Silver Spring, Maryland

(Manuscript received 15 April 2005, in final form 29 August 2005)

ABSTRACT

Using a reanalysis of the climate of the past half century as a model of temperature variations over the next half century, tests of various data collection protocols are made to develop recommendations for observing system requirements for monitoring upper-air temperature. The analysis focuses on accurately estimating monthly climatic data (specifically, monthly average temperature and its standard deviation) and multidecadal trends in monthly temperatures at specified locations, from the surface to 30 hPa. It does not address upper-air network size or station location issues.

The effects of reducing the precision of temperature data, incomplete sampling of the diurnal cycle, incomplete sampling of the days of the month, imperfect long-term stability of the observations, and changes in observation schedule are assessed. To ensure accurate monthly climate statistics, observations with at least 0.5-K precision, made at least twice daily, at least once every two or three days are sufficient. Using these same criteria, and maintaining long-term measurement stability to within 0.25 (0.1) K, for periods of 20 to 50 yr, errors in trend estimates can be avoided in at least 90% (95%) of cases. In practical terms, this requires no more than one intervention (e.g., instrument change) over the period of record, and its effect must be to change the measurement bias by no more than 0.25 (0.1) K. The effect of the first intervention dominates the effects of subsequent, uncorrelated interventions. Changes in observation schedule also affect trend estimates. Reducing the number of observations per day, or changing the timing of a single observation per day, has a greater potential to produce errors in trends than reducing the number of days per month on which observations are made.

These findings depend on the validity of using reanalysis data to approximate the statistical nature of future climate variations, and on the statistical tests employed. However, the results are based on conservative assumptions, so that adopting observing system requirements based on this analysis should result in a data archive that will meet climate monitoring needs over the next 50 yr.

1. Introduction

In specifying requirements for upper-air temperature observations for climate monitoring, several issues must be addressed. These include the spatial and temporal resolution of the observations, and their accuracy, precision, and long-term stability. To address them requires an understanding of the expected future variations in temperature, the types of climate statistics that will be required from the observations, and the way in which individual observations will be assembled to develop those statistics.

In this study, we develop recommendations for measurement requirements for monitoring upper-air temperature. We use the reanalysis of the climate of the past half century as a model of the spatial and temporal variations in temperature that we might expect over the next half century, from the surface to 30 hPa. We focus on identifying data needs to accurately estimate monthly climatic data (specifically, monthly average temperature and its standard deviation) and multidecadal trends in monthly temperatures, at specified locations. We do not address spatial sampling questions such as the optimal number of stations or their placement, which are topics of other, complementary investigations (Free and Seidel 2005; M. McCarthy, Met Office, 2005, personal communication).

Previous work by Kidson and Trenberth (1988) employed meteorological analyses to quantify the effect of missing data on monthly climate statistics. Unlike any existing observational dataset, analyses offer complete

Corresponding author address: Dr. Dian J. Seidel, NOAA/Air Resources Laboratory (R/ARL), 1315 East–West Highway, Silver Spring, MD 20910.
E-mail: Dian.Seidel@noaa.gov

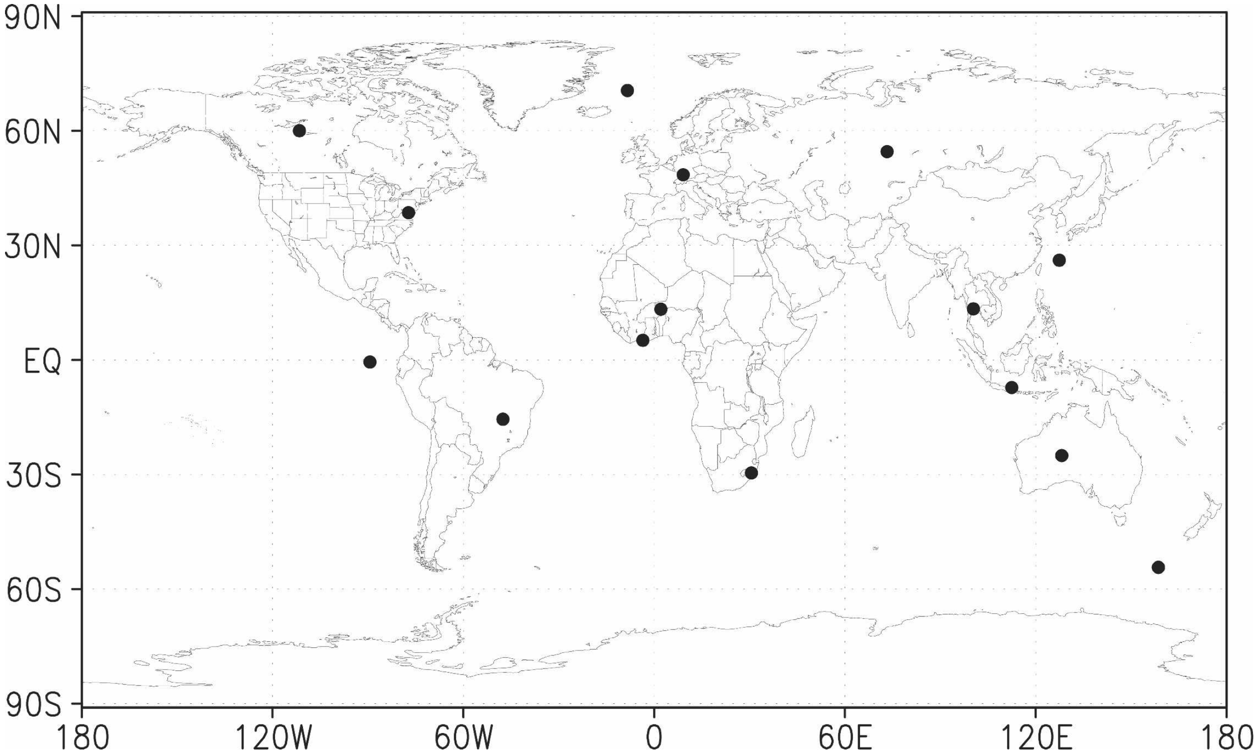

FIG. 1. Map of the 15 sampling locations used in this study and listed in Table 1.

global coverage and sufficient temporal resolution to depict the diurnal cycle. The availability of reanalyses, performed with a consistent assimilation model with data spanning several decades, allows extension of this approach to address long-term climate monitoring issues.

Section 2 describes the reanalysis data and our methodology, including the simulated measurement protocols tested and statistical tests employed. Section 3 presents results of experiments addressing measurement precision, the sampling of the diurnal cycle, the sampling of the month, the long-term stability of the observations, and the observing schedule. Issues that might influence the interpretation of the results are discussed in section 4, and section 5 summarizes our findings.

2. Data and methods

The basis for this study is the assumption that the four-dimensional temperature fields generated by reanalysis of the observed climate of the second half of the twentieth century provide a good approximation of the statistical nature of temperature variations locally and globally, on diurnal to multidecadal time scales. Thus we treat reanalysis as statistical "truth" and perform data sampling experiments to simulate the measurement of the true temperature by the observing system. By varying the simulated observing system's measurement protocols, we can mimic the effects of different choices regarding sampling frequency and measurement error. Comparison of climate statistics based on the simulated measurements with those based on the true reanalysis provides quantitative measures of the ability of a given set of measurement protocols to faithfully reproduce climate statistics.

a. *Reanalysis data*

We employ the temperature data at 6-h intervals (4 times per day) from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis for the period 1948–2003 (Kistler et al. 2001). Data from 15 locations, shown in Fig. 1 and listed in Table 1, were extracted from the global reanalysis archive, as a representative sample of the Global Climate Observing System (GCOS) Upper Air Network (Daan 2002). The locations include continental and maritime sites and range in latitude from the Tropics to the high latitudes. We avoided the polar regions because of concerns about reanalysis data quality there. We selected data from the following 6 of the 18 available vertical levels for analysis: surface and the 850-, 500-, 250-, 100-, and 30-hPa levels.

TABLE 1. List of locations, corresponding to GCOS upper-air network stations, for which reanalysis data were analyzed.

Latitude	Longitude	Station
5.2°N	3.6°W	Abidjan, Ivory Coast
38.6°N	77.3°W	Sterling, VA
70.6°N	8.4°W	Jan Mayen, Norway
7.2°S	112.5°E	Surabaya, Indonesia
29.6°S	30.6°E	Durban, South Africa
26.1°N	127.4°E	Naha, Japan
54.6°N	73.2°E	Omsk, Russia
13.4°N	100.4°E	Bangkok, Thailand
15.5°S	47.6°W	Brasilia, Brazil
0.5°S	89.4°W	San Cristobal (Galapagos), Ecuador
54.3°S	158.6°E	Macquarie Island, Australia
48.5°N	9.1°E	Stuttgart, Germany
13.3°N	2.1°E	Niamey, Niger
25.0°S	128.2°E	Giles, Australia
60.0°N	111.6°W	Fort Smith, NWT, Canada

As an aside, we note that, because of a programming error, in addition to the 15 locations listed in Table 1, we also analyzed data at the same set of latitudes but with longitudes opposite those shown (east and west exchanged). Our conclusions (as summarized in section 5) were completely unchanged after correcting this error, demonstrating the sufficiency of the spatial sampling for the purposes of this analysis.

b. Simulated measurement protocols

Taking the reanalysis temperature data as a true representation of the atmosphere, we simulate temperature measurements, carried out under different protocols, to evaluate their ability to faithfully reproduce true climatological temperatures and temperature variations. While contemporary reanalyses are not perfect representations of past climate variations, particularly in data-sparse regions or during periods of changes in available observations (Stendel et al. 2000), they provide realistic approximations of the statistics of temperature variations and trends (Basist and Chelliah 1997).

Five separate observing system choices are examined in a series of experiments. Each experiment is based on monthly average temperature and its standard deviation.

The first group of experiments tests the sensitivity of climatological statistics to temperature measurement precision, by which we mean the random error of a measurement (as distinct from systematic bias error). Taking the reanalysis temperatures as truth, tests of reduced measurement precision involve rounding the true values to the nearest 0.01, 0.1, 0.5, or 1.0 K, to simulate thermometers with those precisions. For this and subsequent experiments, we then compare monthly averages, standard deviations, and trends based on the unmodified (in this case, maximum precision) data to those based on the experimental data (in this case, with reduced precision).

The second set of experiments examines the sampling of the diurnal cycle. The reanalysis temperatures are available at 0000, 0600, 1200, and 1800 UTC. Using the maximum precision reanalysis data, we perform three subsampling experiments, one with observations twice daily, at 0000 and 1200 UTC, and two with observations once daily, at either 0000 or 1200 UTC.

The third set of experiments considers the number of days per month on which observations are taken. Guided by results of the diurnal sampling experiments, these experiments of the submonthly sampling all involve twice-daily (0000 and 1200 UTC) observations. Four cases test the effects of sampling every day, and every two, three, and seven days. These cases are tested both with the full precision data and with the reduced precisions of 0.1 and 0.5 K.

The final two sets of experiments deal with the long-term stability of the measurements, and so involve time series of monthly means and standard deviations. In the fourth set we imagine an observing system protocol that requires observations to remain stable to within a specified accuracy over a specified period of time. We simulate the effects of artificial inhomogeneities (time-varying systematic biases due to instrument changes, e.g.) on monthly temperature data, for a specific calendar month. Using data segments of 20, 25, 30, and 50 yr, we introduce a step change, or intervention, in the monthly temperature data at one particular time. The time is randomly selected to occur at any point in the time series. The magnitude of the step change is also randomly determined to vary between 0 and a fixed value and can be either positive or negative. Seven different fixed values were used: 0.10, 0.25, 0.50, 0.75, 1.00, 1.50, and 2.00 K. These experiments are performed with monthly values based on twice-daily data, taken every day, every two days, and every three days, and with both maximum measurement precision and reduced precision of 0.1 or 0.5 K. Thus these experiments allow us to examine the combined effects of different measurement protocols. We also examine the effects of multiple interventions on trends.

The fifth set of experiments also focuses on trends, but in this case, rather than imposing a constant step change at random times, we simulate the effects of a change in the observation schedule at the midpoint of each data segment. These experiments are meant to simulate two particular circumstances that could motivate a change in observing schedules. The first is a reduction in the frequency of observations to conserve resources, such as manpower or expendables. We simulate reducing from two to one observations per day and reducing from daily observations to observations every second, third, or seventh day (in all cases retaining two observations per day.)

We also test the effects of changing from one observation per day at 0000 or 1200 UTC to the other observation time at the midpoint of a data segment. This experiment is meant as an (admittedly imperfect) test of the impact of moving from launching radiosondes at a synoptic observing time to timing launches to coincide with overpasses of a polar-orbiting satellite, as has been proposed to maximize the opportunities for calibration and validation of satellite data. In practice, satellite overpasses occur twice daily (except in polar regions where they are more frequent) at fixed local times. Because reanalysis data are only available at four synoptic times, we cannot fully simulate a change from a single synoptic time to a location-specific new time. A switch from 0000 to 1200 UTC observations (or vice versa) is meant to approximate a worst-case scenario of the effects of changing observation time.

c. Statistical tests

The purpose of the experiments outlined above is to determine whether different sampling protocols result in monthly climatological data and multidecadal trends that are representative of the true climatology and trends. We use standard statistical tests to make these determinations, using algorithms given by Press et al. (1989).

To test the null hypotheses that the reanalysis truth and the experimental results have consistent monthly means and variances, we use the Student's t test and the F test, respectively. If the test indicates rejection of the null hypothesis at the $p = 0.05$ level, we consider the means (or variances) to be statistically significantly different. Note that these tests take into account the potentially different sample sizes used to compute the monthly statistics.

We also test whether a trend determined from a time series of monthly values created using one of the experimental sampling protocols is consistent with the true trend based on the fully sampled reanalysis data, with maximum data precision and no artificial step changes. For trend calculations, we employ linear regression to estimate both the trend and its uncertainty (or confidence interval, given as ±2 standard deviations of the trend estimate), with a chi-square merit function that incorporates the errors in the monthly mean temperatures (given by the monthly standard deviations). This is important because different experimental measurement protocols result in different monthly standard deviations, which in turn affects the uncertainty of the trend estimates.

All trends are based on time series of monthly data, and trends are computed separately for each calendar month, rather than for all months or for annual means. This choice offers two major advantages. First, it allows for 12 times as many trends to be computed at a given location and level, giving a larger sample of trend calculations on which to base conclusions. Second, it alleviates the problem of underestimating the uncertainty of the trend estimate due to nontrend-related temporal autocorrelation in the time series.

We test the null hypothesis that a given trend is consistent with the true trend using the t test, which incorporates the trend estimates and their statistical uncertainties, and again use the $p = 0.05$ level to determine statistically significant differences in trend. One potential complication arises when a t test result might indicate that trends are not significantly different because one or both trend estimates have very large uncertainties, but when the interpretation of the two trends would lead to different conclusions about atmospheric temperature changes. For example, data based on one sampling protocol indicate a warming trend with a confidence interval that does not include zero, while data based on another sampling protocol indicate a warming trend with a confidence interval that does include zero. In such a case, a t test could conceivably indicate that the two trends were consistent, but a data analyst might interpret each trend (in the absence of the other) differently, the first but not the second indicating significant warming.

We address such concerns using a contingency table (Table 2) that makes use of both the t-test results and the trend confidence intervals. The situation described in the preceding paragraph corresponds to the second set of possibilities presented in the table (only one trend confidence interval does not include zero), in which case, regardless of the t-test results, we declare the trends to be significantly different. In the first set of possibilities in Table 2 (neither trend confidence interval includes zero), we rely completely on the t-test results. In the third set (both trend confidence intervals include zero), we declare the trends not significantly different regardless of the t test, since a data analyst looking at observations based on either sampling protocol would conclude that no significant temperature change had occurred.

Our use of standard, parametric statistics (averages, standard deviations, and linear regression trends) for this analysis, rather than their nonparametric equivalents (e.g., medians, interquartile ranges, and median-of-pairwise-slopes trends) leads to conservative deci-

TABLE 2. Contingency table used to determine whether two trends (one with absolute magnitude T_1 and 2-sigma confidence interval C_1, and the second with absolute magnitude T_2 and 2-sigma confidence interval C_2) are, or are not, significantly different, depending on the *t* test of the null hypothesis that the two trends are consistent.

	t test of null hypothesis that trends are consistent	
Trends and confidence intervals	Accepted	Rejected
$T_1 > C_1$ and $T_2 > C_2$ (neither trend confidence interval includes zero)	Not significantly different	Significantly different
$T_1 > C_1$ or $T_2 > C_2$ (only one trend confidence interval does not include zero)	Significantly different	Significantly different
$T_1 < C_1$ and $T_2 < C_2$ (both trend confidence intervals include zero)	Not significantly different	Not significantly different

sions regarding the fidelity of results from a given experiment compared with the true reanalysis. Because nonparametric statistics (and nonparametric tests) are robust to the underlying probability model (e.g., non-Gaussian distributions) and resistant to outliers, they are likely to yield more consistent results for some of the experiments in which the within-month sampling frequency is reduced or artificial step changes are introduced into a time series. Therefore, we would more likely accept the hypothesis that the experimental results do not differ significantly from the true values. Because we do not wish to underestimate any potential problems with a given measurement protocol, we base our analysis on parametric statistics.

This same concern influenced our decision not to reduce the *p* value of our statistical tests, as is often recommended when a large number of tests are performed. To avoid the "fishing expedition" problem, in which some tests will give apparently statistically significant results purely by chance if a large enough number of tests is conducted, one can use the Bonferroni inequality (Bonferroni 1936) to make the test more stringent by reducing *p* to p/N, where *N* is the number of tests performed. In our case, we are more concerned with the possibility of missing significant differences in climate statistics than with wrongly identifying differences as statistically significant. Therefore, we report results based on unmodified *p* values.

However, we did perform tests with reduced *p* values, and found the impact to be small. For both a simple reduction of *p* from 0.05 to 0.01, as well as for further reduction using the Bonferroni inequality, the fraction of tests determined to yield statistically significant differences (in means, standard deviations, or trends) was reduced by at most a few percent, and in many cases the change was less than 1%.

3. Results

This section presents the results of the experiments outlined in section 2, with different simulated measurement protocols, including tests of the effects of variable precision of temperature measurements, variable sampling of the diurnal cycle, variable sampling of the month, variable constraints on the long-term stability of measurement accuracy, and variations in observation schedule. We examine climatological monthly statistics first, then trends.

a. *Monthly means and standard deviations*

1) Temperature measurement precision

Figure 2 shows the distribution of the effects on monthly mean temperature and its standard deviation resulting from reducing the precision of the reanalysis temperature. Each box plot represents 60 480 monthly samples (the product of 15 locations, 672 months, and 6 vertical levels). For every single sample, the Student's *t* test and *F* test indicate that there is no significant difference, at the $p = 0.05$ level (or for $p = 0.01$), between the monthly means and standard deviations computed from the full precision and reduced precision data. With measurement precision of 0.01, 0.10, 0.50, or 1.00 K, the means never differ by more than 0.01, 0.01, 0.06, or 0.11 K, respectively. The standard deviation from the reduced precision data is generally within 10% of the true value (Fig. 2, bottom), although, for the 1.00-K precision case, the reduced precision standard deviation can approach 30% larger than the true value, probably because rounding to the nearest whole degree effectively increases the range of the observations.

From this set of experiments, we conclude that, with full temporal sampling of the month (four observations per day every day), reducing the data precision has minor effects on monthly means and standard deviations. To ensure that means are accurate to within ~0.05 K, and standard deviations are accurate to within 10%, measurement precision must be held within 0.50 K.

2) Sampling of the diurnal cycle

The impact of subsampling the diurnal cycle on monthly averages and standard deviations is larger than the impact of reduced data precision. Of the 60 480

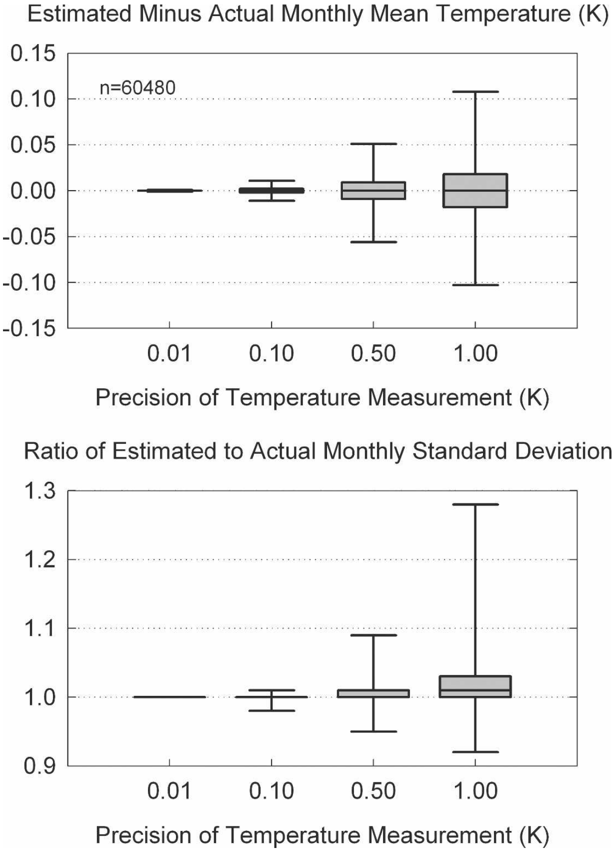

FIG. 2. The effects of reduced measurement precision on monthly means and standard deviations of temperature. (top) The distributions of the differences in monthly mean temperatures, taken as the estimated mean when the precision is reduced (to 0.01, 0.10, 0.50, or 1.0 K) minus the actual monthly mean from reanalysis data. Each box-and-whisker plot is based on 60 480 samples (from data at 15 locations, six vertical levels, and 672 months) and shows the min and max differences, and the 25th, 50th, and 75th percentile values. (bottom) Same as in top, but for the distributions of the ratios of the estimated to actual monthly standard deviations. All monthly means and standard deviations are based on sampling every day of the month, 4 times per day.

cases, sampling twice daily (at 0000 and 1200 UTC) caused monthly means and standard deviations to differ significantly from their values based on four samples per day in 5.5% and 3.9% of the cases, respectively. When only 0000 UTC (1200 UTC) data were used, means were significantly different in 20.2% (25.2%) of the cases, and standard deviations were significantly different in 8.7% (7.3%) of the cases.

The effects of this subsampling depend on the magnitude and shape of the diurnal temperature cycle, which varies seasonally, vertically, and from location to location (Seidel et al. 2005), as well as the timing of the selected observations with respect to the time of maximum and minimum temperature. Examples of the effects at two locations, Abidjan, Ivory Coast, and Sterling, Virginia, are shown in Fig. 3. At Abidjan, near the equator and the Greenwich meridian, sampling at 0000 and 1200 UTC yields monthly means that are generally within 0.2 K, and almost always within 0.5 K, of the values based on four observations per day, but with a small bias toward warmer monthly means at most levels. Sampling only at 0000 UTC (local midnight) yields differences that exceed 0.2 K in more than half the cases for the surface and for the 30-hPa level, with cooler monthly means than those obtained using four observations per day at those levels, and warmer values at midtropospheric levels. Sampling only at 1200 UTC (near local noon) has opposite, and somewhat larger, effects. Standard deviations are generally smaller in the subsampling cases than in the full sampling case, with reductions of more than 5% in half the cases for twice-daily sampling, and reductions of more than 10% in half the cases for once-daily sampling.

At Sterling, the impact of subsampling on monthly mean surface temperature is substantially larger than at Abidjan, with median differences of about 0.6 K for 0000 and 1200 UTC sampling, 1.6 K for 0000 UTC sampling, and 2.8 K for 1200 UTC sampling. (Note the different *x*-axis scales for Abidjan and Sterling in Fig. 3.) In the free atmosphere (850 to 30 hPa), the effects are much smaller, with differences generally <1.0 K. The tendency for reduction in monthly standard deviation is lower at Sterling than at Abidjan. This is probably because synoptic weather variability (from day to day) contributes a greater fraction of the overall variability at Sterling than at Abidjan, where variability associated with the diurnal cycle is a more important factor.

From these experiments, we conclude that sampling once daily introduces systematic biases in monthly mean temperatures and can either inflate or deflate estimates of monthly standard deviations. These effects can be mitigated by sampling twice daily, at 0000 and 1200 UTC, in which case only ~5% of monthly statistics will be significantly different from those based on four observations per day. Based on this conclusion, the remaining experiments are all performed using twice-daily sampling.

3) SAMPLING OF THE MONTH

Figure 4 shows the effects on monthly means and standard deviations of taking (twice daily) observations once every two, three, and seven days, compared with daily. Results are shown for the ensemble of all stations both at all six vertical levels and at three individual levels: 850, 500, and 100 hPa. In each case, the median difference in means is near zero, and the median ratio

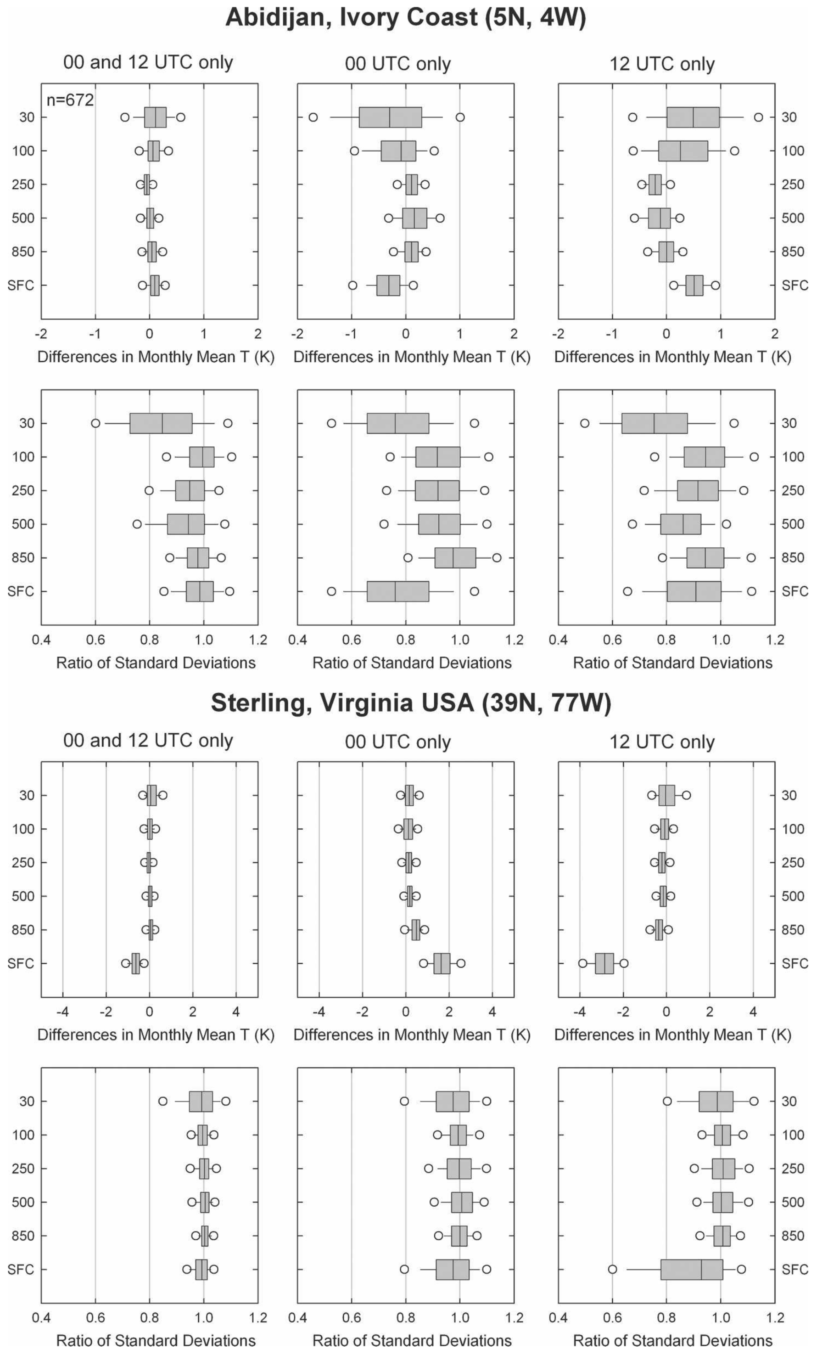

FIG. 3. The effects of reduced sampling of the diurnal cycle on monthly means and standard deviations of temperature. (top row) The distributions of the differences in monthly mean temperatures, taken as the estimated mean based on (left) 0000 and 1200 UTC data, (middle) 0000 UTC data, and (right) 1200 UTC data minus the monthly mean based on 0000, 0600, 1200, and 1800 UTC data, at the location of Abidjan, Ivory Coast, at six vertical levels, with the surface and the five pressure levels (hPa) indicated. Each box-and-whisker plot is based on 672 samples (for 1948–2003) and shows the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile difference values. (second row) Same as top row, but for the distributions of the ratios of the estimated to actual monthly standard deviations. (third row), (bottom row) Same as first and second rows, respectively, but for the location of Sterling, VA. Note the different *x*-axis scales in the first and third rows. All monthly means and standard deviations are based on full-precision data and sampling every day of the month.

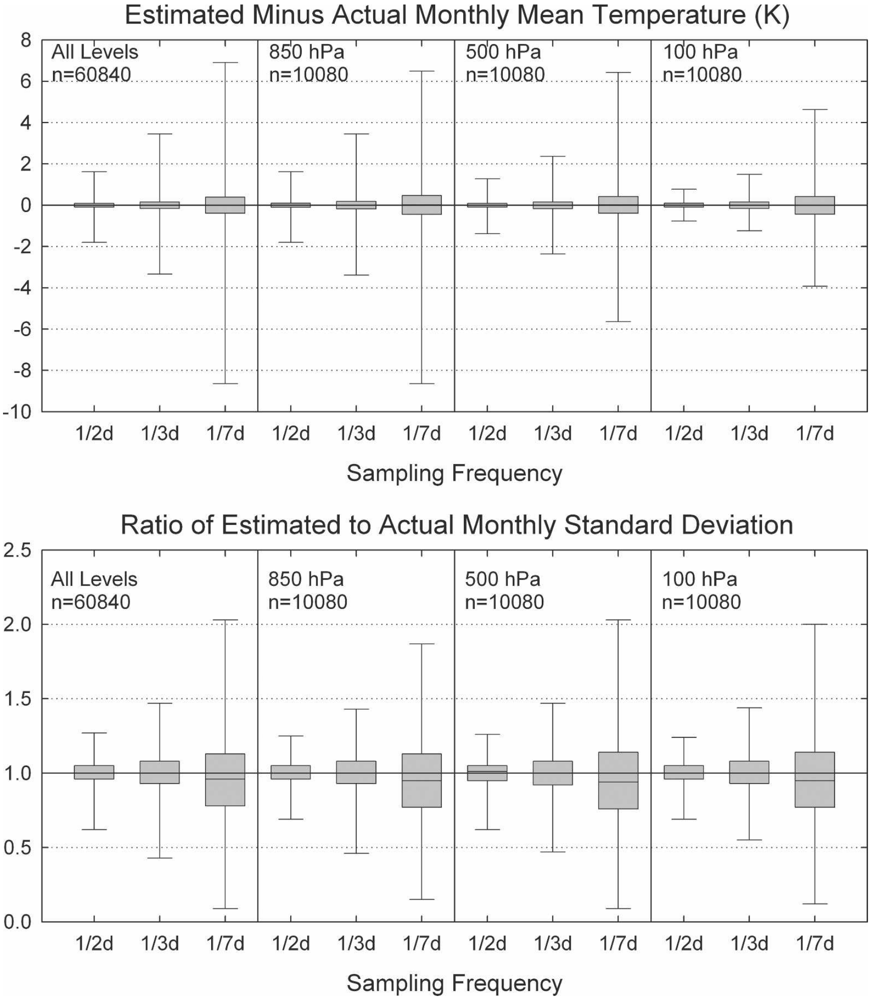

Fig. 4. The effects of reduced sampling of the month on monthly means and standard deviations of temperature. Same as in Fig. 2, but comparing means and standard deviations based on sampling every two, three, and seven days with sampling every day. In all cases, twice-daily sampling (0000 and 1200 UTC data) and full-precision data were used. Leftmost panels show combined results from six vertical levels, and the other three panels show results for (from left to right) 850, 500, and 100 hPa.

of standard deviations is near unity. However, there is considerable spread about the median. For sampling every other day, monthly means are always within 1.8 K of the values based on daily sampling, and they are within 0.09 K more than 50% of the time. For sampling once every three (seven) days, means are always within 3.4 K (8.6 K), and within 0.16 K (0.39 K) more than 50% of the time.

Comparing the results at different pressure levels in Fig. 4 indicates that monthly means at 850 hPa (and at the surface; not shown) are more sensitive to reduced sampling than those at higher altitude. In this regard, subsampling the month is similar to subsampling the day.

The effect of subsampling the month on monthly standard deviations is substantial. Sampling every two or every three days yields standard deviations that are within 10% of the true value more than half the time. But they can be as much as 30% larger or smaller for sampling once every two days, and as much as 50% larger or smaller for sampling once every three days. Weekly sampling can result in standard deviations from 90% smaller to 100% larger than true values.

Rarely are the monthly averages and standard deviations significantly different, according to the *t* test and *F* test, in these submonthly sampling experiments. In the weekly sampling experiment, they are significantly different in 4% of the cases. For sampling every other day or every three days, we find significant differences in less than 0.5% of the cases. This is because large

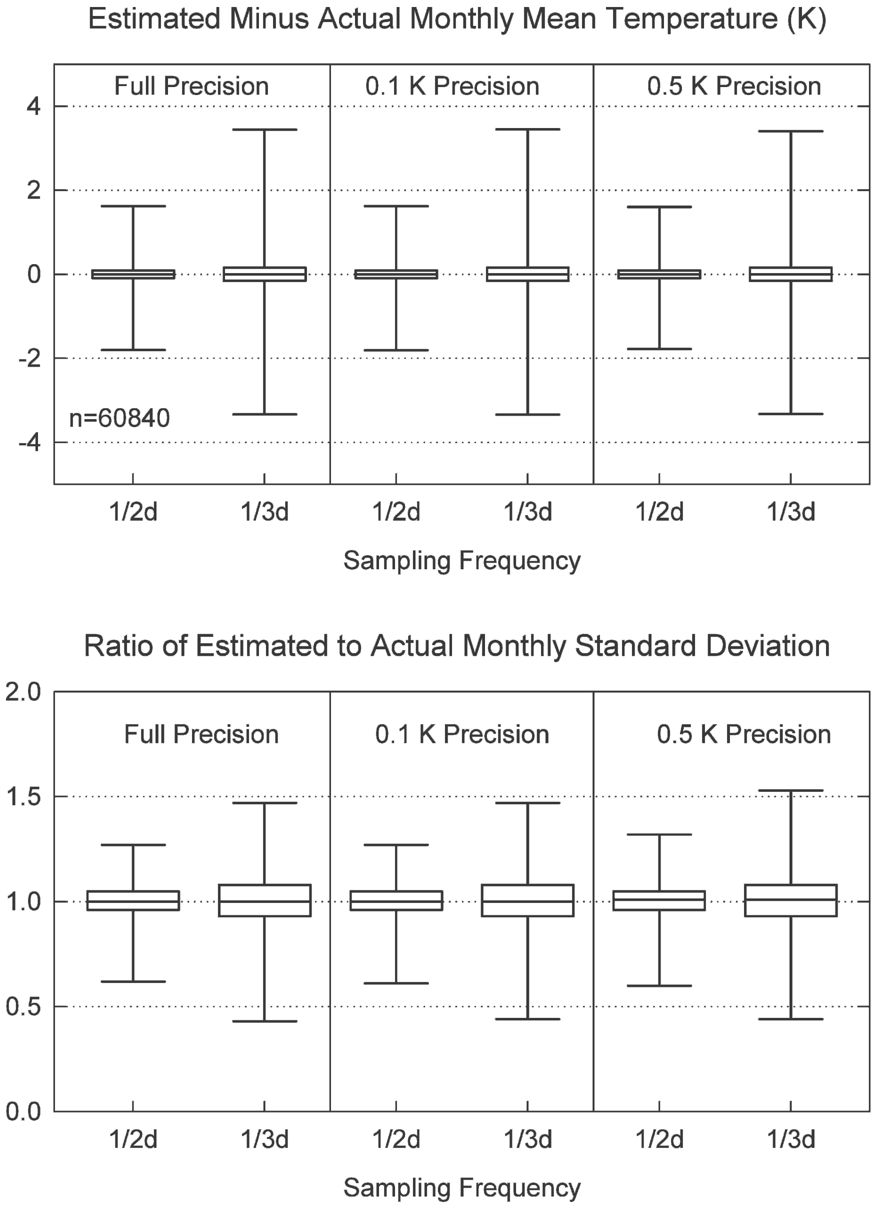

FIG. 5. The combined effects of reduced measurement precision and reduced sampling of the month on monthly means and standard deviations of temperature. (left) Full-precision results from Fig. 4, but showing results only for sampling every two and three days. The effects of reducing measurement precision to (middle) 0.1 and (right) 0.5 K.

changes in means are accompanied by large changes in standard deviations, and because the experiments reduce the number of samples, both of which contribute to lower *t*-test scores.

Results from these tests of subsampling the month, using full precision data, are repeated in Fig. 5, which also shows comparable results using data with reduced precision of 0.1 and 0.5 K. In all cases, two observations per day were used, and Fig. 5 shows the ensemble of results for all the stations and levels. These comparisons indicate that, for a given monthly subsampling protocol, reducing data precision has a very minor impact on the monthly averages and standard deviation, consistent with our results in section 3a(1) above.

From these experiments, we conclude that sampling every other day, or every three days (but not every seven days) yields monthly means and standard deviations that are not significantly different from the true values at least 99.5% of the time, and this is true even if the data precision is reduced to 0.1 or 0.5 K. To

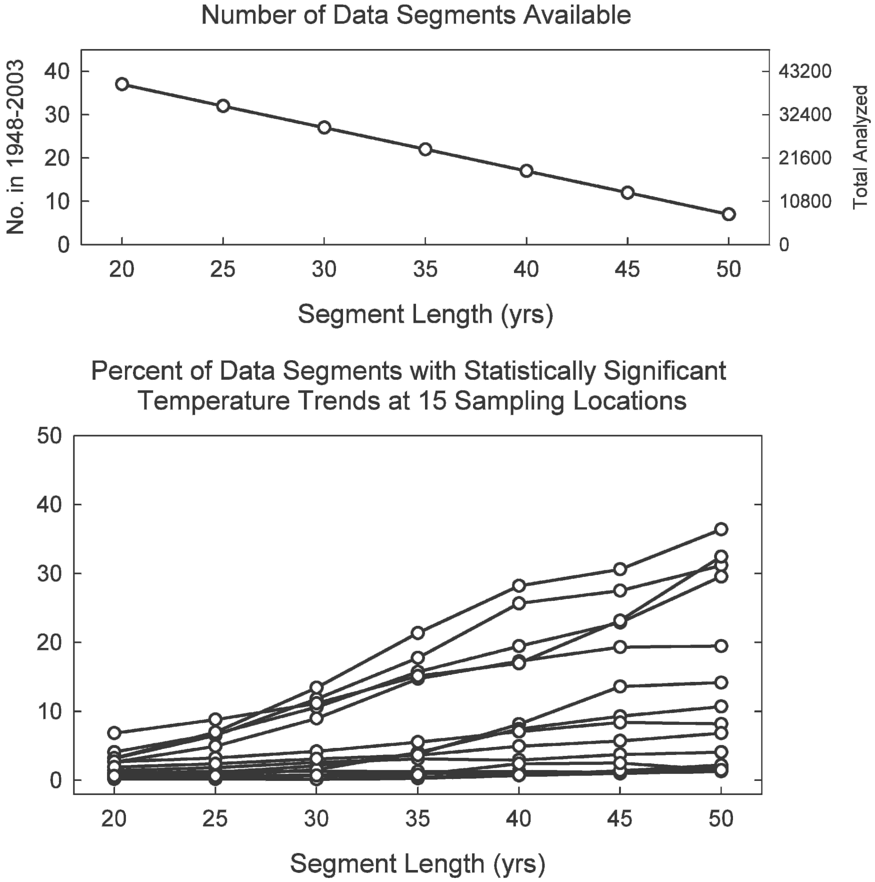

Fig. 6. (top) The number of data segments of various lengths available during 1948–2003. Left axis shows the number of segments, and right axis shows the number of time series analyzed, which is the product of the number of segments and 15 stations, six levels, and 12 calendar months (bottom) The percentage of data segments with temperature trends statistically significantly different from zero. Each trace is for one location, at all six levels and for all 12 calendar months.

ensure that differences in monthly means do not exceed 2 K, sampling must be done at least once every two days. This result—that daily observations are not necessary—is consistent with the findings of Kidson and Trenberth (1988), who stress that subsampling uniformly throughout the month (e.g., taking one observation every 3 days, as we have done here) is substantially less problematic than clumping observations (e.g., taking observations for 10 consecutive days in a month, with no measurements on the other days).

4) Impact of reanalysis data period on climatological results

Several investigators (e.g., Pawson and Fiorino 1999; Kistler et al. 2001; Bengtsson et al. 2004) have pointed out that NCEP–NCAR reanalysis fields exhibit artificial step-like behavior around 1979, the time of the start of assimilation of satellite data, and at other times. We tested whether the results presented in section 3a were sensitive to the selection of reanalysis data period by repeating our analysis for the period 1979–2003, and they were not. Although changes in the input data stream can introduce spurious interannual variations, their effects on data precision, the shape of the diurnal cycle, and submonthly variability appear to be small enough to have no impact on this analysis.

b. Trends

So far we have examined effects of data precision and temporal sampling on monthly climatological statistics. Now we turn to multidecadal trends. In the 56-yr reanalysis record, trends can be computed over various periods, with data segments starting in different years. Figure 6 (top) shows the number of data segments of a given length available in the 56-yr record; these range from 37 twenty-year segments to 7 fifty-yr segments. Considering that we are analyzing data from six vertical levels at 15 locations for 12 calendar months, the right-hand axis of the plot shows the total number of trend estimates possible for a given segment length, ranging from 7560 fifty-year segments to 39 960 twenty-year segments.

Although reanalysis trends are not reliable estimates of true atmospheric trends, for the purposes of this study we are not concerned that every trend for every period and every pressure level be correct. What matters is that the distribution of reanalysis trends provides a reasonable representation of the expected range of atmospheric trends, and that the reanalysis signal-to-noise (trend to shorter-term variability) ratio be realistic.

Before examining the effect of different measurement protocols on trends, we first examine the frequency of statistically significant trends in the reanalysis data. We define a trend as statistically significant if the two standard deviation confidence interval does not include zero. Figure 6 (bottom) shows the percentage of data segments with statistically significant trends, for each of the 15 locations sampled. For 20-yr data segments, less than 10% of the trends are significant. With longer segments, the frequency of significant trends increases, but even for 50-yr segments it is nowhere larger than 40% and is less than 20% in more than half the cases. This result emphasizes the fact that, even with optimal "observations," with perfect precision, full temporal sampling, and no artificial discontinuities, statistically significant temperature trends in the reanalysis are not frequent, especially for short data records.

1) Effects of measurement precision and temporal sampling

Figure 7 shows the effects of various measurement protocols on the detection (or nondetection) of temperature trends. The plots show the trend error rate, defined as the frequency (expressed as a percentage) of erroneous trend estimates, using the criteria described in Table 2 and section 2c above to determine whether trend estimates are consistent. The top panel indicates that reducing the measurement precision has little influence on trend detection, and erroneous estimates are made in less than 1% of the cases, for all trend period lengths, for precisions of up to 1.0 K. If precision is held to 0.5 K, errors are made in less than 0.5% of the cases.

Subsampling the diurnal cycle has a much more significant effect on trend estimates (Fig. 7, second panel). Error frequency increases with increased trend period length, because of differential daytime and nighttime trends at some locations. (The ratio of the number of statistically significant trends at 0000 UTC to the number at 1200 UTC, or the reciprocal, varies between 0.37 and 0.99 among the 15 locations tested, with a median value of 0.80. This dependency of trends on time of observation may be an artifact of the reanalysis but might also reflect actual changes in the amplitude of the diurnal cycle.) For 50-yr periods, sampling twice daily (at 0000 and 1200 UTC) results in erroneous trend estimates in 11% of the cases, and sampling only once daily, at 0000 or 1200 UTC, increases the error rate to 16 or 17%, respectively.

As seen in Fig. 8 (left), these trend error rates (in this case for 50-yr trends) vary with altitude and are smaller at 500 and 250 hPa than at the lower and higher levels. This is probably in part a reflection of the result discussed above—larger errors in monthly means associated with reduced sampling at the lowest levels. The errors in mean values will contribute to errors in trends. The larger error rates at 850 hPa than at the surface may be a reflection of the more realistic representation of the 850-hPa diurnal cycle compared with the surface, and the subsequent larger errors associated with reduced sampling of the diurnal cycle at 850 hPa. Surface temperatures in the NCEP–NCAR reanalysis are not based on surface temperature observations, and, as discussed in section 4 below, the amplitude of the surface diurnal cycle is underestimated.

The high error rates at 30 hPa (up to 44% for sampling only at 0000 UTC) is partially due to this same effect of erroneous means but may also be related to problems with the reanalysis data. Known step-like inhomogeneities in stratospheric temperatures associated with the introduction of satellite data into the assimilation (Pawson and Fiorino 1999; Kistler et al. 2001) will affect trend calculations. If these inhomogeneities have different manifestations at different times of day, trends based on subsampled data will be different from those based on full sampling.

Compared with subsampling the day, even larger error rates are obtained in the experiments in which observations are taken less frequently than daily. In this case (Fig. 7, third panel, and Fig. 8, middle panel), we compare trends based on daily sampling, with two observations per day, with trends based on sampling every two, three, or seven days, also with two observations per day, as before. Again, the error rate increases with increasing period length. For weekly sampling, the overall error rate is 10% for 20-yr periods and 27% for 50-yr periods. For sampling every other day, the error rate remains less than 12% for all periods, and it is about 1% higher for sampling every three days. Note, however, that these error rates for sampling every two or three days are only 1% or 2% higher than those based on daily sampling and are predominantly due to subsampling the diurnal cycle rather than subsampling the month (Fig. 7, second and third panels).

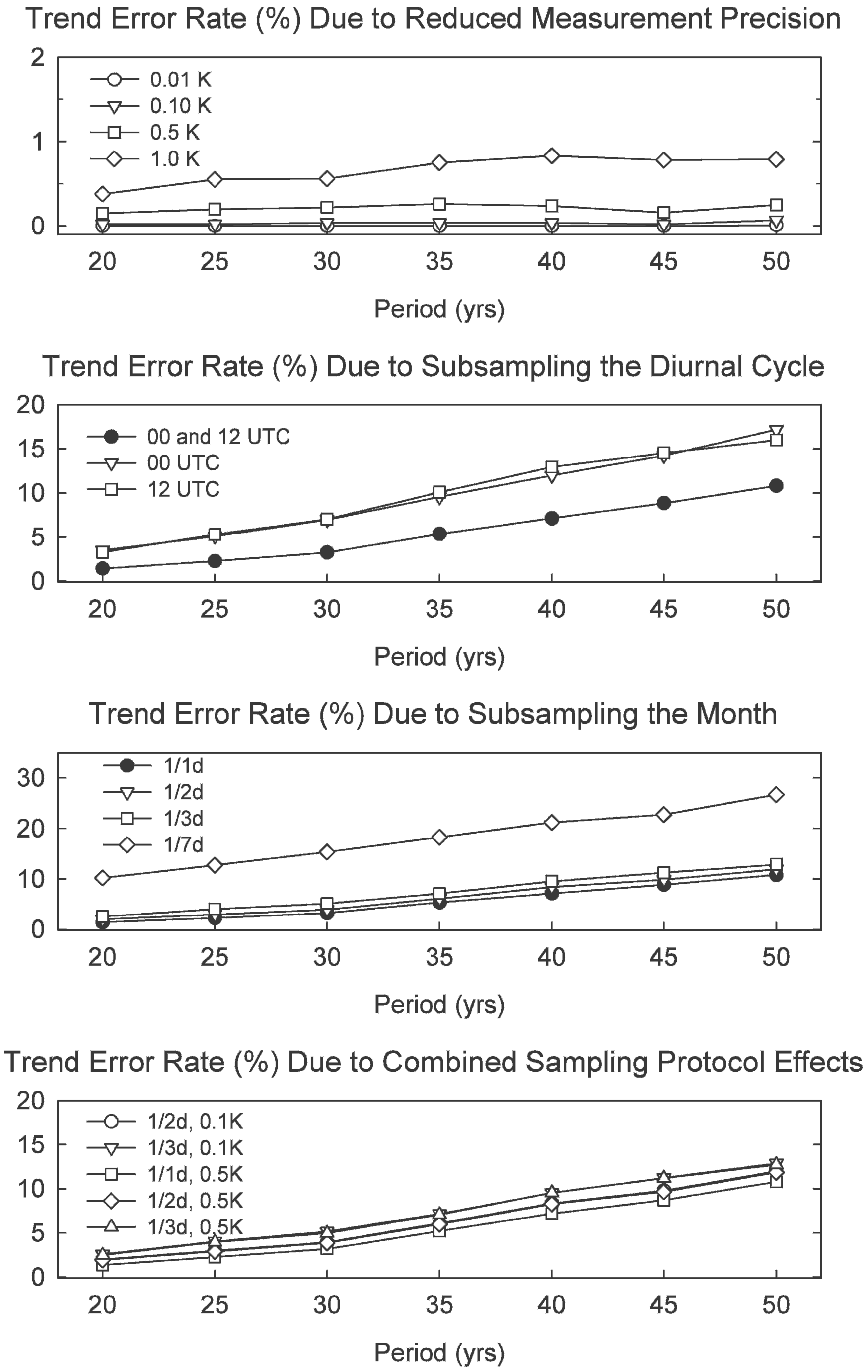

FIG. 7. The effects of different measurement protocols on temperature trend detection, as represented by the percentage of trends that are significantly different from the actual reanalysis trends (error rate), for trend periods ranging from 20 to 50 yr. (See text for discussion of tests for determining significance of trend differences.) (from top to bottom) The effects of reduced measurement precision, the effects of subsampling the diurnal cycle, the effects of subsampling the days of the month, and the combined effects of reduced data precision and subsampling the days of the month. Results in the bottom two panels are based on two samples per day (0000 and 1200 UTC), whereas the top panel is based on four samples per day.

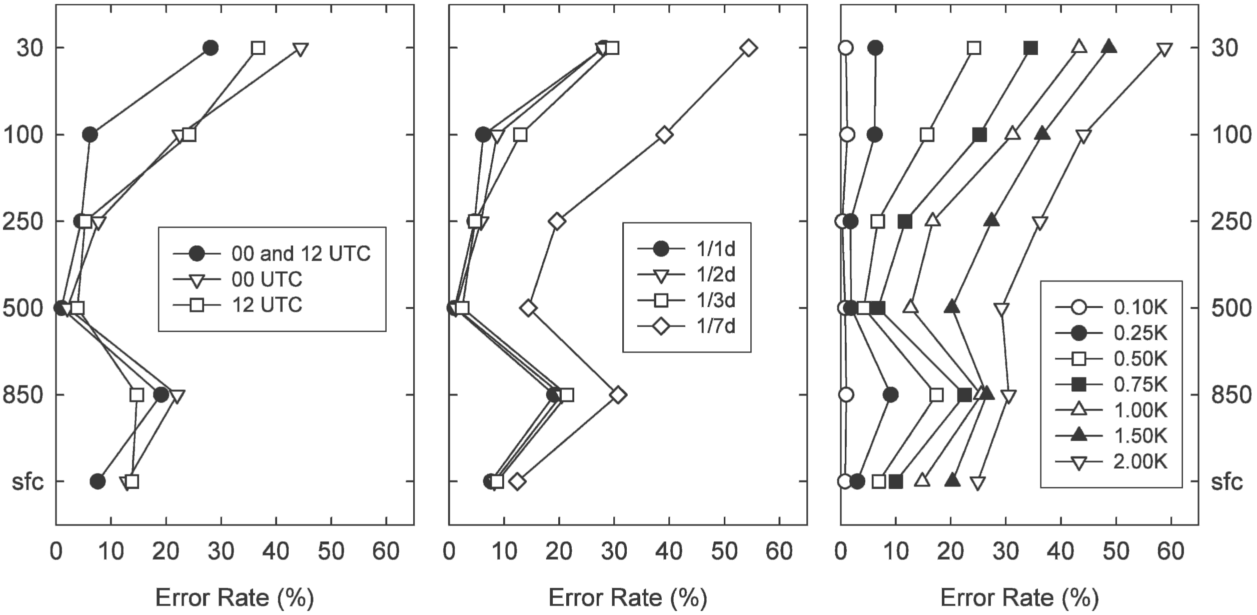

FIG. 8. Vertical profiles of 50-yr trend error rates associated with different measurement protocols. Each plot is an ensemble of results from 15 locations. (left) The percentage of trends that are statistically significantly different from the actual reanalysis trends (error rate) when diurnal sampling is reduced from four observations per day to one or two observations per day. (middle) Error rates due to subsampling the month and using twice-daily observations. (right) Error rates associated with introducing randomly timed interventions of various maximum magnitudes.

The vertical structure of 50-yr trend error rates associated with submonthly sampling is shown in Fig. 8 (middle), which repeats the results for daily sampling at 0000 and 1200 UTC, for comparison. Consistent with the overall results in Fig. 7 (third panel), at all levels sampling only once every two or three days increases the error rate by only a few percent. However, sampling once every seven days increases the error rate by up to 25%.

Combining the effects of measurement precision and submonthly sampling (Fig. 7, bottom) shows once again that measurement precision has a minor effect on error rates, which are dominated by period length (higher error rates for longer periods) and are higher for less frequent sampling.

2) Effects of temporal inhomogeneities

The greatest challenge for detecting temperature trends in observational data is presented not by measurement precision or sampling frequency but by temporal inhomogeneities in time series. These can be introduced by changes in instrumentation, observing practices, or data-processing methods, and result in time-varying biases that can masquerade as, or mask, true trends. We simulate these effects on long-term data stability by randomly introducing step changes, or interventions, in time series, as described in section 2b above. For these experiments, all trends are based on twice-daily observations, and trend errors are assessed by comparing trends for homogeneous and inhomogeneous time series.

Figure 9 shows the effect of interventions on temperature trend detection. The top panel shows the frequency of erroneous temperature trends, as a function of data period length, for seven cases in which the data stability is maintained to within a particular amount (ranging from 0.1 to 2.0 K) over the complete period. In these experiments we have introduced only one randomly timed intervention per time series. Results from all 15 locations, six vertical levels, and 12 calendar months are shown as an ensemble, for each of the seven cases.

The top panel of Fig. 9 indicates that, predictably, the error rate increases as the long-term stability requirement is relaxed. When the stability is maintained to within 0.1 K, error rates are at most 1% for all periods considered. Relaxing the stability requirement to 0.25 K (or 0.5 K) increases the error rate, so that for 50-yr periods, it reaches 5% (12%). When the maximum intervention is 2.0 K, 50-yr trend estimates are erroneous in 35% of the cases.

The bottom panel of Fig. 9 shows comparable results, but instead of using maximum precision data, twice daily, every day (as in the top panel), we sample twice

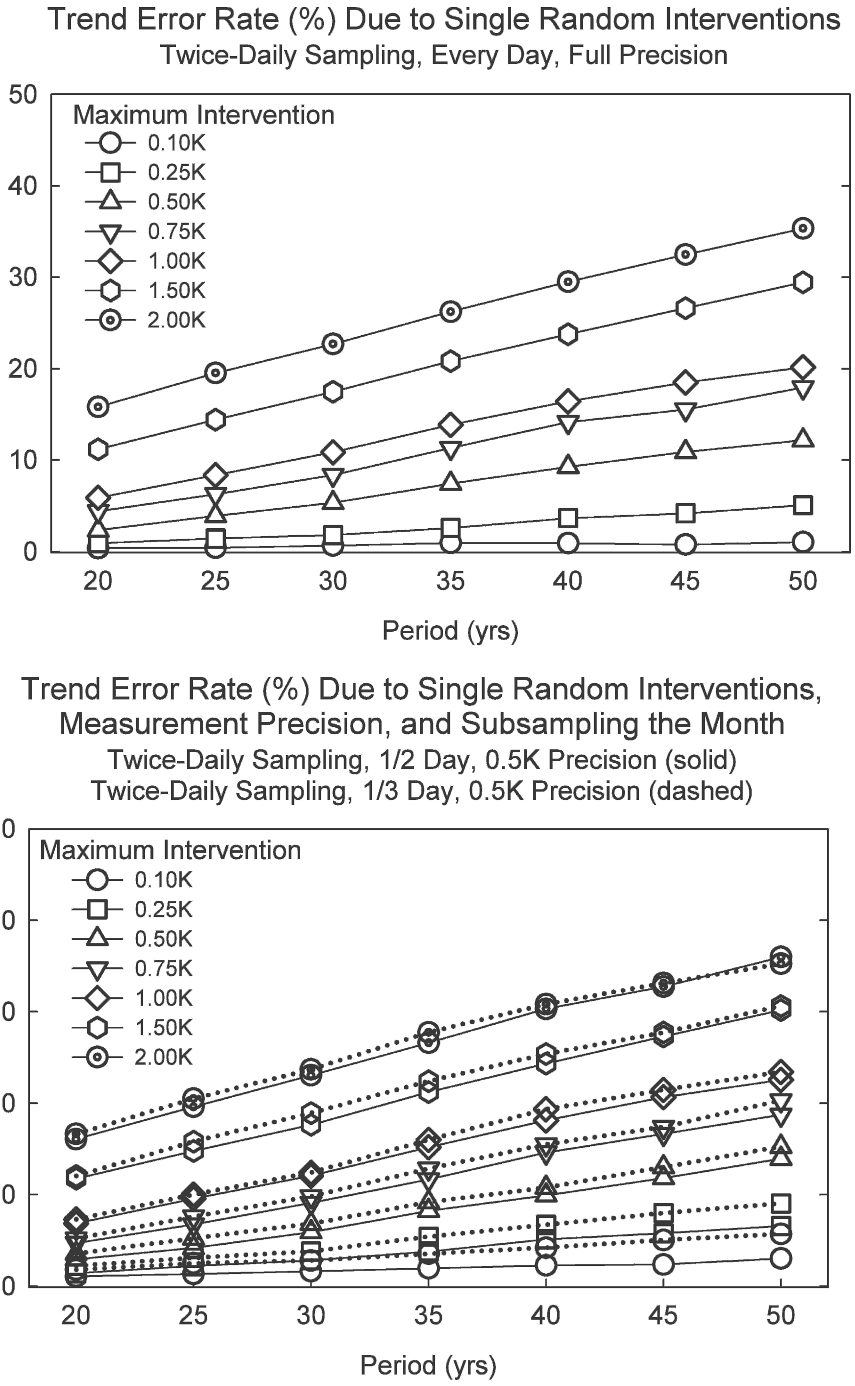

FIG. 9. The effects of long-term data stability on temperature trend detection. (top) The percentage of trends that are statistically significantly different from the actual reanalysis trends, for trend periods ranging from 20 to 50 yr. Each trace shows results from experiments in which a single artificial step change in temperature (of varying maximum magnitudes) was randomly introduced into each time series. All monthly data are based on two samples per day, every day, with maximum precision observations. (bottom) Same as in top, but for 0.5-K measurement precision and reduced sampling of the month. Results for sampling once every other day are shown by solid lines, and those for sampling once every three days are shown by dotted lines. The two sets of results are very similar, with slightly higher percentages of different trends resulting from the once per three-day sampling than for the once per two-day sampling.

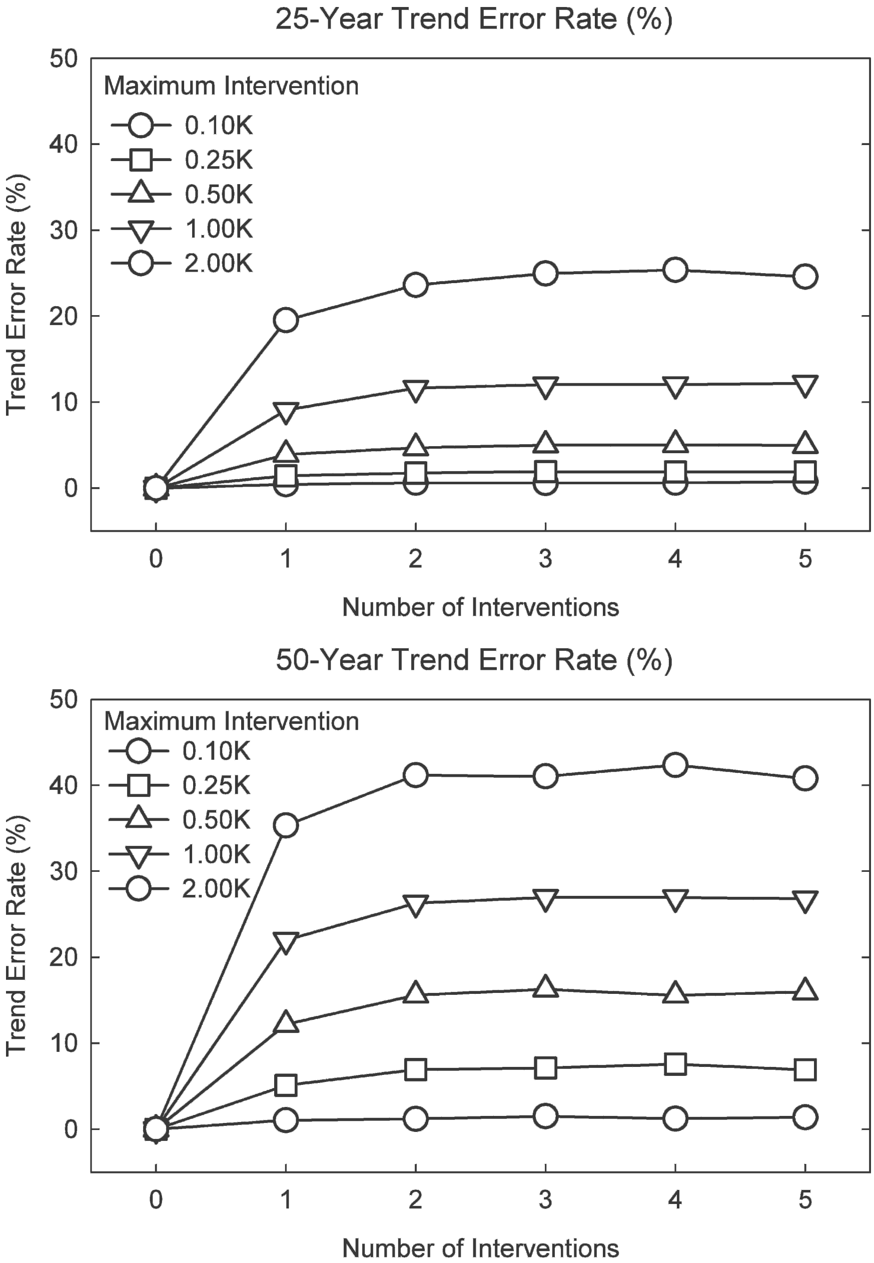

FIG. 10. The effects of multiple interventions on temperature trend detection. The percentage of trends that are statistically significantly different from the actual reanalysis trends, as a function of the number of randomly introduced artificial shifts in time series, for trend periods of (top) 25 and (bottom) 50 yr. Each trace is for a different maximum artificial step change in the observations. All results are based on daily sampling, two observations per day, and maximum precision data.

daily, with 0.5-K precision, either every two days or every three days. The results are almost identical to those in the top panel, reinforcing the point that long-term data stability is the key factor in determining the accuracy of trend estimates.

The vertical structure of the effect of interventions on 50-yr trend error rates is shown in Fig. 8 (right). The largest error rates are at 850 hPa and at 100 and 30 hPa. Error rates are less than 10% at all levels for interventions of 0.25 K or smaller.

Figure 10 examines the impact of multiple interventions on trend error rates for trends computed over 25- and 50-yr periods, again using twice-daily observations, every day, with maximum precision. With zero inter-

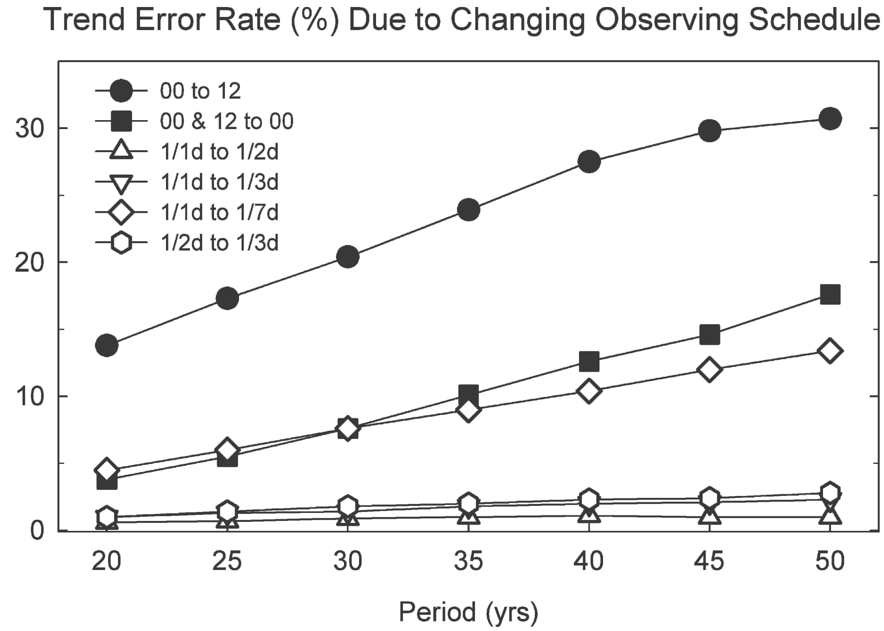

FIG. 11. The effects of changes in observation schedule on temperature trend detection. The percentage of trends in time series with an observing schedule change introduced at the midpoint that are statistically significantly different from trends in unaltered time series, for trend periods ranging from 20 to 50 yr. Effects of changes in the time of observation (from 0000 to 1200 UTC or from 0000 and 1200 UTC to 0000 UTC only) are shown by filled symbols, and changes in the number of days per month sampled are shown by open symbols.

ventions, of course, the error rate is zero. With a single intervention, the 25-yr trend error rate (Fig. 10, top) increases from 0.4% to 20% as the maximum intervention size increases from 0.1 to 2.0 K, as also shown in Fig. 9, and from 1% to 35% for 50-yr trends (Fig. 10, bottom). Introducing a second intervention increases the error rate by up to several percent but has a much smaller impact than the first intervention. Third, fourth, and fifth interventions have negligible effects. This is because the interventions are uncorrelated, to simulate the effects of changes in instruments whose accuracies (bias errors) are unrelated from one type to the next. In this scenario, the first intervention introduces an artificial trend, but subsequent interventions can either aggravate or meliorate the trend error depending on whether the second intervention is of the same or opposite sign as the first.

From this set of experiments, we conclude that the frequency of errors in trend estimates (based on two-standard-deviation confidence levels and *t*-test results significant at the 0.05 level) can be held below 1%, for up to 50-yr data periods, if measurement stability is held to within 0.1 K. The error rate will be at most 5% if measurement stability is held to within 0.25 K. By this we mean that at most one change in bias, of magnitude up to 0.1 K (or 0.25 K), is allowed within the data period. These results can be achieved with twice-daily sampling, every day, with data precision of 0.5 K or better. Reducing the sampling to once every two or three days slightly increases the error rate, so that for 50-yr trends and 0.25-K measurement stability the error rates are 7% or 9%, respectively.

3) EFFECTS OF CHANGES IN OBSERVING SCHEDULE

To simulate the effects of changes in observing schedule, we performed experiments similar to those discussed in the last section but, instead of introducing a constant intervention at a random time in a data segment, we join two time series, each based on a different observing schedule, at the midpoint of each time period examined. The resulting trend error rates are shown in Fig. 11. Changing from twice-daily observations every day to every second or third day results in error rates of less than 5% for trend periods of 20 to 50 yr. Switching from daily to weekly observations (twice daily) results in 20-yr trend errors in 5% of the cases and 50-yr trend errors in 13% of the cases.

Changing the time of observations results in more frequent trend errors. Reducing the observation frequency from two per day (0000 and 1200 UTC) to 0000 UTC only (filled square symbols in Fig. 11) or 1200

UTC only (not shown) results in 20- and 50-yr trend error rates of 4% and 17%, respectively. If a one-observation-per-day schedule is changed from 0000 to 1200 UTC (or vice versa) the error rates are 14% for 20-yr trends and 31% for 50-yr trends. These errors are largest at the lowest and highest pressure levels but are still important in the midtroposphere. At 500 hPa we obtain a 50-yr trend error rate of 11%. Thus it appears that maintaining a constant time of observation is more important than maintaining daily observations for avoiding errors in temperature trend estimates.

4. Discussion

Our conclusions regarding measurement requirements for climate monitoring must be considered in light of a few caveats. First, our results regarding sampling of the diurnal cycle depend on the degree to which the reanalysis data accurately represent the diurnal variation of temperature. A full analysis of this issue is beyond the scope of this paper, and we note that observational data to assess the reanalysis have limitations (Seidel et al. 2005). Nevertheless, it appears that the amplitude of the diurnal cycle in surface temperature may be underestimated by the NCEP–NCAR reanalysis. For example, the average difference between 0000 and 1200 UTC temperatures in the reanalysis at Abidjan (Fig. 3) is about 1 K, whereas climatological surface data suggest it is much larger (closer to 10 K). This may be much less of a problem at levels above the surface, and at other locations, but it highlights a potential shortcoming in reanalysis data, which may warrant further investigation.

It is possible that the fidelity of the diurnal cycle in the NCEP–NCAR reanalysis is worse than, or better than, that in the European Centre for Medium-Range Weather Forecasts (ECMWF) 40-yr reanalysis, and our experiments might yield different results if applied to that reanalysis. Similarly, differences in the magnitude of synoptic-scale temperature variations or of multidecadal trends in a different reanalysis could impact our findings regarding submonthly sampling and long-term data stability, respectively.

We reiterate that our conclusions are based on parametric statistical parameters and tests. More liberal measurement criteria might result from the use of nonparametric statistics. However, we feel that measurement requirements should be based on conservative estimates and arguments, to ensure collection of a robust data archive.

Finally, and perhaps most importantly, recall that our findings depend on the basic assumption that climate variability over the past 50 yr, as depicted in the reanalysis, is a good predictor of its variability over the coming half century. If we have underestimated the variability, particularly the magnitude and detectability of temperature trends, then our conclusion with regard to measurement requirements will be more restrictive than necessary, which might mean unnecessary expenditure of resources, but which would guarantee a useful data archive. If, on the other hand, climate trends are smaller in the future than they are in the reanalysis, our conclusions may result in recommendations that do not ensure that observations will be useful for detecting change. To guard against this possibility, it might be wise to adopt more stringent requirements than this analysis suggests.

We have limited our study to temperature observations. Extension of these methods to other variables (e.g., humidity aloft) should be undertaken only after considering their representation in the reanalysis. Kistler et al. (2001) identify temperature as a “type-A” variable, meaning that the reanalysis fields are more closely tied to observations, and therefore more reliable, than other variables.

5. Summary

We have used the NCEP–NCAR reanalysis upper-air temperature data for the period 1948–2003 to explore the effects of different data collection protocols on climate statistics. Our main findings are as follows:

(a) Reducing the precision (increasing the random error) of temperature data has minor effects on monthly means and standard deviations and is not an important factor in determining multidecadal trends. If individual measurement precision is at least 0.50 K, monthly means are accurate to within ~0.05K, and standard deviations are accurate to within 10%, and the means and standard deviations are never statistically significantly different.

(b) Sampling twice daily, at 0000 and 1200 UTC, ensures that monthly statistics will be statistically significantly different from those based on four observations per day in only ~5% of the cases. However, sampling once daily introduces biases in monthly mean temperatures and can either inflate or deflate estimates of monthly standard deviations, as a result of the diurnal cycle in temperature.

(c) Twice-daily sampling must be done at least once every two days to ensure that monthly means are accurate to within 2 K. Sampling every two days, or every three days (but not every seven days), yields monthly means and standard deviations that are not significantly different from the true values at least 99.5% of the time.

(d) Long-term data stability is the primary determiner of trend estimate accuracy. By maintaining temperature measurement stability to within 0.1 K, for periods of 20 to 50 yr, errors in trend estimates can be avoided in at least 99% of cases. In practical terms, this requires no more than one intervention (e.g., instrument change) over the period of record, the effect of which is to change the measurement bias by no more than 0.1 K. If the stability requirement is relaxed to 0.25 K, 50-yr trend error rates increase to about 5% for twice-daily sampling every day, and about 7% for twice-daily sampling every two days. The impact of second and subsequent interventions on trend error rates is much smaller than the impact of the first intervention, if the sign and magnitude of the interventions are uncorrelated.

(e) Changing observing schedules can introduce data inhomogeneities that lead to errors in trend estimates. Large errors result from changing from 0000 to 1200 UTC observations (or vice versa), from two to one observation per day, or from daily to weekly sampling, which results in 50-yr trend error rates of 31%, 18%, and 13%, respectively. Changing from twice daily observations every day to every second or third day results in errors in trends in less than 5% of the case.

These findings depend on the validity of using reanalysis data to approximate the statistical nature of future climate variations, and on the statistical tests employed. However, we have tried to make conservative assumptions, so that adopting observing system requirements based on this analysis will result in a data archive that will meet climate monitoring needs over the next 50 yr.

Acknowledgments. We are grateful to Jim Angell and Ellen Cooter (NOAA), Peter Thorne (Met Office), and two anonymous reviewers for their constructive suggestions that significantly improved this work.

REFERENCES

Basist, A. N., and M. Chelliah, 1997: Comparison of tropospheric temperatures derived from the NCEP/NCAR reanalysis, NCEP operational analysis, and the Microwave Sounding Unit. *Bull. Amer. Meteor. Soc.,* **78,** 1431–1447.

Bengtsson, L., S. Hagemann, and K. I. Hodges, 2004: Can climate trends be calculated from reanalysis data? *J. Geophys. Res.,* **109,** D11111, doi:10.1029/2004JD004536.

Bonferroni, C. E., 1936: Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. R Ist. Sup. Sci. Econ. Commerc. Firenze,* **8,** 3–62.

Daan, H., 2002: GCOS-73: Guide to the GCOS surface and upper air networks: GSN and GUAN. World Meteorological Organization Tech. Doc. WMO TD-1106, Geneva, Switzerland, 37 pp. [Available online at www.guanweb.com.]

Free, M., and D. J. Seidel, 2005: Causes of differing temperature trends in radiosonde upper-air datasets. *J. Geophys. Res.,* **110,** D07101, doi:10.1029/2004JD00548.

Kidson, J. W., and K. E. Trenberth, 1988: Effects of missing data on estimates of monthly mean general circulation statistics. *J. Climate,* **1,** 1261–1275.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Pawson, S., and M. Fiorino, 1999: A comparison of reanalyses in the tropical stratosphere. Part 3: Inclusion of the pre-satellite data era. *Climate Dyn.,* **15,** 241–250.

Press, W. H., B. F. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1989: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 702 pp.

Seidel, D. J., M. Free, and J. Wang, 2005: The diurnal cycle of upper-air temperature estimated from radiosondes. *J. Geophys. Res.,* **110,** D07101, doi:10.1029/2004JD00548.

Stendel, M., J. R. Christy, and L. Bengtsson, 2000: Assessing levels of uncertainty in recent temperature time series. *Climate Dyn.,* **16,** doi:10.1007/s003820000064.