Robust Automated Quality Assurance of Radiosonde Temperatures

By

Imke Durre, Russell S. Vose, and David B. Wuertz

National Climatic Data Center, Asheville, North Carolina

Submitted to the

Journal of Applied Meteorology and Climatology

June 2007

Revised December 2007

Corresponding author's address:

Imke Durre
National Climatic Data Center
151 Patton Avenue
Asheville, NC 28801
Phone: 828.271.4870
Fax: 828.271.4328
E-mail: Imke.Durre@noaa.gov

Abstract

This paper presents a description of the fully-automated quality assurance (QA) procedures that are being applied to temperatures in the Integrated Global Radiosonde Archive (IGRA). Since these data are routinely used for monitoring variations in tropospheric temperature, it is of critical importance that the system be able to detect as many errors as possible without falsely identifying true meteorological events as erroneous. Three steps were taken to achieve such robust performance. First, a total of 14 tests for excessive persistence, climatological outliers, and vertical and temporal inconsistencies were developed and arranged into a deliberate sequence so as to render the system capable of detecting a variety of data errors. Second, manual review of random samples of flagged values was used to set the "thresholds" for each individual check so as to minimize the number of valid values that are mistakenly identified as errors. Finally, the performance of the system as a whole was also assessed through manual inspection of random samples of the quality-assured data. As a result of these efforts, the IGRA temperature QA procedures effectively remove the grossest errors while maintaining a false-positive rate of approximately 10%.

1. Introduction

In this era of strong interest in climate change studies, there exists an ever-growing need for high-quality historical and real-time meteorological observations. One parameter that is increasingly important is air temperature measured by radiosondes (Free et al. 2005; Thorne et al. 2005). Extending back to the early 1940s, radiosonde observations constitute the longest available record of temperature in the free atmosphere. As such, they are the primary source of information on historical variations in the vertical temperature profile and thus are central to the assessment of differences between surface and tropospheric temperatures.

The largest readily available collection of radiosonde observations is the Integrated Global Radiosonde Archive (IGRA, Durre et al. 2006), which consists of over 30 million soundings from 1500 stations worldwide. Temperature data from IGRA have been employed in a number of climate change applications, including the "HADAT3" gridded dataset of adjusted monthly-mean temperatures (Thorne et al. 2005) as well as the Radiosonde Atmospheric Temperature Products for Assessing Climate (Free et al. 2005). Both products are used to monitor upper-air temperatures in real time.

Unfortunately, when first received, radiosonde observations often not only are characterized by inhomogeneities resulting from changes in instrumentation, measurement practice, or station location, but also contain a variety of gross random and systematic errors. The latter can be caused by problems in equipment calibration, sensor performance, data transmission, or data processing. Consequently, in addition to techniques for homogenizing the radiosonde record (Free et al. 2002, 2005; Thorne et al. 2005; Haimberger 2007), procedures that can detect basic data quality problems in

historical and real-time radiosonde data are also of critical importance from the perspective of climate-change detection. Such fundamental Quality Assurance (QA) procedures have been implemented in IGRA, with particular attention to the climatically important variable of temperature. In brief, the IGRA QA system includes tests for impossible and implausible values; for internal, vertical, and temporal inconsistency; and for excessive temporal and vertical persistence.

This paper describes the logic and performance of the QA procedures applied to temperatures in IGRA. The philosophy used in designing and evaluating the checks is reviewed in Section 2, and an overview of the system is provided in Section 3. Specific checks are described in Sections 4-7. The overall performance of the system is discussed in Section 8, and concluding remarks are offered in Section 9.

2. Design and evaluation principles

The challenge in developing an automated QA system for global radiosonde data is that the system must be capable of detecting a wide variety of errors in the context of considerable variability in climatological conditions and data resolution. For instance, radiosonde observations are taken at stations located in climate zones ranging from the tropics to Antarctica; likewise, record length, temporal resolution, and vertical resolution and extent vary considerably among stations and across time (Durre et al. 2006). Furthermore, errors in the data include physically impossible values, implausible repetitions of the same value, and improbable vertical profiles (Gandin 1988; Schwartz and Doswell 1991; Gandin et al. 1993; Loehrer et al. 1996; Collins 2001b). To accommodate this diversity in data completeness and error characteristics, the IGRA QA

system was developed with two fundamental principles in mind. First, the system should consist of a suite of specialized algorithms applied in a deliberate sequence. Second, each individual algorithm, and the system as a whole, should be rigorously evaluated to ensure satisfactory performance.

In accordance with these principles, the design and evaluation strategies outlined in Durre et al. (2008) were employed during system development. Specifically, the development process included the following steps:

- the design of tests to detect known data problems;
- the use of manual review of samples of flagged values for the selection of test thresholds that yield a low false-positive rate for each check;
- the identification of any undetected types of errors through manual review of samples of the quality-assured data;
- the development of additional QA procedures as long as significant numbers and types of gross errors remained undetected; and
- the estimation of the overall type I error and false-positive rates for the final combination of checks through the manual review of a random sample of the values flagged by the entire system.

The goal of this process is to produce a QA system that requires no manual intervention during operational data processing but reflects logic that would be employed by human validators during a typical semiautomatic QA approach (e.g., Loehrer et al. 1996). By combining multiple tests with low false-positive rates into one system, it is possible to

compensate for the limitations of any individual check and exploit the differing error detection capabilities of each test while minimizing the risk of inadvertently flagging unusual meteorological events.

The resulting set of QA checks are applied in sequence, with each procedure ignoring values that have been flagged by preceding tests. Checks based on climatological statistics follow those testing for more basic plausibility, so that statistics required by a particular procedure are computed from as clean a dataset as possible. The primary benefit of this approach is that it allows each component check to detect a specific error regardless of whether other tests can be applied to the same value. An alternative approach would be to apply the technique of Complex Quality Control, a method in which the final QA decision for each value is made by a decision-making algorithm that takes into account the results from multiple tests (Gandin 1988; Eskridge et al. 1995; Collins 2001a; Graybeal et al. 2004a). Unfortunately, however, the decisions made by such a complex system can be compromised when incomplete data prevent the application of certain tests, and, therefore, the sequential approach appears preferable for radiosonde data that frequently are not serially or vertically complete.

3. System overview

Table 1 lists a total of 14 temperature QA procedures that are part of the IGRA system. The tests, which are described in detail in subsequent sections, are designed to check any combination of mandatory-and significant-level reports against certain fundamental, logical, or physical principles. For the sake of discussion, the procedures have been grouped into four general categories: checks for runs, outlier checks, vertical

consistency checks, and temporal consistency checks. The four "runs tests" identify cases in which the same value is repeated across an excessive number of soundings or pressure levels. Physically implausible values, as well as temperatures that deviate excessively from station-specific climatological parameters, are detected by the four outlier checks. Finally, four vertical consistency checks and two temporal consistency checks are employed in order to address outliers that pass the climatological checks but are vertically or temporally inconsistent. The procedures are applied in the order in which they are listed in Table 1. With the exception of the "gross plausibility check" and "Tier-2 climatological check," the descriptions in Sections 4-7 also follow this order.

Each of the procedures relies on one or two key parameters for determining whether a particular value or sounding constitutes a data error (see column 2 of Table 1). For example, the relevant parameter in the Tier-1 climatological check is the z-score obtained by standardizing a given temperature relative to its climatology. The temperature fails this climatological check if its z-score exceeds the test threshold of 6.0.

The tests' thresholds (column 4 of Table 1) are set such that the vast majority of values exceeding the thresholds constitute true errors, i.e., the false-positive rate of each check is kept as low as possible (Durre et al. 2008). In each case, the appropriate threshold was determined by means of visually assessing the validity of samples of flagged values, soundings, or time series for a range of plausible parameter thresholds (column three of Table 1). During this threshold selection process, the validity of data points was judged on the basis of geographical location, season, and consistency within the profile, and temporal consistency with the previous and subsequent soundings reported at the same site.

In the case of the Tier-1 climatological check, for example, the test was first applied to all temperature data in IGRA using z-score thresholds ranging from 2.0 to 7.0. An initial manual inspection then indicated that values with z-scores greater than 7.0 were clearly erroneous while those with z-scores less than three were clearly plausible. Consequently, thresholds of 3.0, 4.0, 5.0, and 6.0 were examined in detail by inspecting time series plots for selected representative levels (surface, 500 hPa, and 50 hPa) for all IGRA stations. The final threshold of 6.0 for this check is based on the observation that the false-positive rate would increase significantly if a lower threshold were chosen. Errors with smaller z-score magnitudes are therefore more effectively detected by other (subsequent) tests.

4. Checks for excessive persistence

One type of error that appears in many types of digital meteorological data is the repetition of the same value in time or space (Reek et al. 1992; Peterson et al. 1998; Graybeal et al. 2004a). In the IGRA data, this problem was found to occur in the form of excessive temporal and vertical persistence. Temporal persistence implies that at a particular pressure level, the same value is repeated across an extended sequence of soundings. Vertical persistence means that identical values are found at a large number of consecutive levels within the same sounding. Excessive persistence of either kind usually reflects a systematic data problem that may not necessarily result in outliers detectable by conventional climatological or consistency tests. Consequently, specialized checks are needed to detect these types of problems.

a. Checks for runs in time

Two procedures were implemented to check for runs of the same value across soundings. The first considers all soundings together, whereas the second considers soundings from each hour separately. In either case, temperatures from the surface and each mandatory level are analyzed. At any one of these levels, a run is terminated by a change to another temperature, but not by the mere absence of an observation.

When runs of different lengths were inspected, it was found that runs shorter than 15 observations sometimes corresponded to events in tropical environments where natural day-to-day variability tends to be low, and therefore persistence is not surprising. This was particularly true when the precision of the data was low, e.g., 1°C rather than the usual 0.1 or 0.2°C. By contrast, runs longer than 15 were clearly erroneous; the values comprising these runs tended to be climatologically unusual and/or vertically inconsistent with the profiles to which they belong. An example of such a run is shown in Figure 1.

Erroneous runs at a specific hour (not shown) were found to occur particularly when a station's record was a mix of radiosonde observations at 0000 and 1200 Coordinated Universal Time (UTC) and pilot balloon observations at 0600 and 1800 UTC. In such cases, the pilot balloon observations sometimes contain one fixed temperature value at the top pressure level, while the radiosonde observations from the same days report temperatures of a more realistic magnitude and variability at the same level. Based on these findings, both tests for runs in time were set to remove runs consisting of 15 or more values (Table 1) since their implausibility was judged to be largely independent of geographical location, the observation times considered, and the quantization of the data.

b. Checks for runs in the vertical

Two procedures were developed to test for excessive vertical persistence within a sounding. One identifies runs across mandatory levels, while the other operates exclusively on significant levels. This approach was chosen in favor of one that considers all levels simultaneously because runs are sometimes confined to only one of these two types of levels and therefore are interrupted by values from the other type of level.

The manual review of vertical runs of different lengths suggested that for both the mandatory-level and significant-level procedures, runs extending over fewer than five levels tended to occur in layers with closely-spaced levels as well as in the near-surface or tropopause regions where isothermal layers can be expected. On the other hand, runs spanning five or more levels were indicative of a data problem in a significant portion of the sounding and frequently were accompanied by other egregious errors in the remainder of the profile. An erroneous vertical run occurs, for example, when temperature is reported as zero throughout an entire sounding. Based on these findings, both tests consider a run across five or more levels erroneous (Table 1) and remove the entire temperature profile when such a run is detected.

5. Outlier checks

QA systems frequently include a test for "climatological outliers," i.e., values that exceed the corresponding long-term mean by a specified amount (Kahl et al. 1992; Parker and Cox 1995; Wolter 1997; Peterson et al. 1998). To take into account geographical and seasonal differences in variability, the departure from the mean is usually expressed as a

multiple of the standard deviation (STD) of the data. In practice, this implies that a value being tested is first converted to a standard z-score and is then identified as an error if the z-score exceeds a specified threshold. Although this approach is statistically and physically intuitive, it is not without complications that can compromise its effectiveness. First, the test can be applied only when there exists a sufficient number of data values for computing the required statistics, an issue that is particularly relevant at locations with short records and at pressure levels where observations are intermittent. Furthermore, the requisite means and STDs can be contaminated when they are computed from data containing large numbers of gross outliers. Finally, the process of normalizing values by the STD can lead to overflagging in environments with extremely low variability and when the distribution of measured values is positively skewed relative to the normal distribution (Wolter 1997).

The IGRA QA process is designed to address these issues. Specifically, the system contains four outlier checks with different data requirements, including one testing for absolutely implausible values, one relying on a station- and level-specific climatology, and two employing a climatology that also varies with time of year and time of day. To minimize the impact of erroneous data values on the effectiveness of the climatological tests, means and STDs are calculated from data that have passed through basic QA procedures, and the computations are performed using biweight statistics (Lanzante 1996) which tend to be less sensitive to outliers than their conventional counterparts.

a. Gross plausibility check

The gross plausibility check uses absolute thresholds to identify temperatures that are clearly erroneous regardless of location, altitude, or time. Specifically, temperatures colder than -120°C or hotter than 70°C are rejected. These thresholds represent values that are well outside the temperatures typically observed in the coldest and hottest regions of the earth, namely, the tropical tropopause and the subtropical deserts, respectively. The test therefore functions as a fundamental "house-cleaning" measure that has the advantage of operating irrespective of the amount of data available. As such, it constitutes the first procedure in the IGRA QA system (Table 1), preceding even the checks for excessive persistence discussed in the previous section.

b. Tier-1 climatological check

The Tier-1 climatological check is based on a climatology that varies with location and altitude, but not with season or time of day. A significant advantage of this test is that it can be applied to records in which observations are too sparse for the development of the seasonally- or diurnally-varying statistics used by the subsequent "Tier-2" climatological check. In addition, the Tier-1 climatological check is of value even when the more stringent test is possible because it provides a means for excluding grossly implausible values from the Tier-2 climatological statistics. The trade-off, however, is that the limits of the test have to be sufficiently wide to avoid erroneously flagging the extremes of the seasonal and diurnal cycles, as is illustrated by the dotted lines in Figure 2.

The procedure is applied to each value that can be converted to a standard z-score, i.e., whenever sufficient data are available for deriving the corresponding mean and STD

(in this case, 120 observations). The method for deriving the required statistics depends on whether the observation is reported at a mandatory or significant level as well as on the location of the level within a sounding. For mandatory levels, the requisite biweight mean and biweight STD are computed from all observations at the relevant location and pressure level regardless of the time of year or time of day. The mean and STD for a significant level are derived as needed by interpolating linearly with respect to the logarithm of pressure between the corresponding statistics at the nearest adjacent mandatory or surface levels, if available. (This approach for estimating a climatology at significant levels was chosen in favor of a direct calculation since observations at a particular significant level tend to be less frequent and more prone to reporting biases than those at mandatory levels.) At levels with pressures higher than the relevant mean surface pressure, the statistics computed for the surface level are used. For levels above the top mandatory level, the test references the statistics for the top mandatory level (provided the pressure difference is less than 30 hPa).

Based on the threshold selection analysis (Table 1), the z-score threshold for the Tier-1 climatological check was set to 6.0; virtually all values exceeding this threshold were erroneous. An example of such a point is the -55.9°C 500-hPa temperature reported at 1200 UTC on 4 August 1972 at Fort Sill, Oklahoma (Fig. 2), which has a Tier-1 z-score of -7.54.

c. Tier-2 climatological check

In contrast to the Tier-1 climatology, the Tier-2 climatology contains both a seasonally-varying and a diurnally-varying component (gray lines in Fig. 2) and

incorporates more stringent data requirements. Specifically, means and STDs for the surface and mandatory levels are computed from observations within a running 45-day window and within the appropriate fixed three-hour period. The 45-day window is centered on the date of the sounding being tested, and the appropriate three-hour window is one of eight periods centered on 0000, 0300, 0600, 0900, 1200, 1500, 1800, and 2100 UTC. The minimum number of values required for the Tier-2 climatology was set to 150. As in the Tier-1 climatology, significant-level statistics are estimated by interpolating or extrapolating from statistics calculated for adjacent mandatory levels. However, rather than extrapolating to any level within 30 hPa of the top-climatology level, the top Tier-2 climatology is extrapolated only to those levels where the ratio of the pressure to the top pressure with climatology is equal to at least 0.9.

The difference between the final Tier-2 threshold of 5.0 (Table 1) and the corresponding Tier-1 threshold is illustrated in Figure 2. The temperature observed on 4 August is a sufficiently large outlier that it exceeds the Tier-1 threshold and thus could be detected even if an insufficient amount of data prevented the computation of the Tier-2 climatology. The value shown for 25 July, on the other hand, cannot be considered an outlier relative to the time series as a whole, but does fall outside of what is considered reasonable for its time of year and time of day and thus exceeds only the Tier-2 threshold.

d. Whole-profile climatological check

An additional check in the IGRA system tests the degree to which an entire profile deviates from the Tier-2 climatology. This check identifies soundings in which a certain measurement or processing problem has resulted in all or most of the

temperatures being far warmer or far colder than expected for the given season in time of day, yet not all affected temperatures exceed the thresholds of the other outlier checks. An example of such a temperature profile is shown in Figure 3a. Throughout much of the troposphere and stratosphere, the temperatures in this sounding are far too warm for the location (Tunis, Tunisia) at this time of year (March). If the Tier-2 climatological check is applied to this profile, the temperatures at the surface, 200 hPa, 179 hPa, and 170 hPa survive because the magnitudes of their z-scores (Fig. 3b) are less than 5.0. What remains is a rather disjointed profile in which three of the four temperatures have z-scores that are not far below the Tier-2 threshold. For soundings such as this, a test that identifies the entire sounding as erroneous is therefore preferable to one that checks each value individually.

The test developed for this purpose uses the median of the absolute values of the temperature z-scores at all available levels in a sounding as a measure for how anomalous a profile is relative to climatology. In other words, the temperature at each level is first standardized using the Tier-2 climatology, and then the median of all of the absolute values of these standardized temperatures is calculated. For the sounding in Figure 3, for example, this "median absolute z-score" is equal to 9.95, a value that far exceeds the chosen test threshold of 4.0 (Table 1). The test removes all temperatures in such soundings.

6. Vertical consistency checks

Most QA systems for radiosonde data include some form of procedure that tests profiles for vertical inconsistencies (Kahl et al. 1992; Eskridge et al. 1995; Loehrer et al. 1996; Collins 2001a). Vertical inconsistencies can arise from single temperatures that are

outliers from the rest of the profile, from a cluster of values at consecutive levels that are vertically consistent with each other but deviate considerably from the rest of the profile, and from excessive "zigzag" fluctuations of temperatures throughout a sounding. To address these types of data problems, the IGRA QA system contains four types of vertical consistency checks that are applied iteratively to a single sounding until no further vertical inconsistency is found. They include a check for excessive level-to-level fluctuations throughout a sounding, a check for spikes in the vertical profile, and two tests for multipoint anomalous portions of a profile. All four procedures operate on temperatures that have been standardized relative to the Tier-2 climatology described in Section 5c above. As in the case of the climatological checks, the standardization increases the reliability of the tests by reducing geographical differences in variability. In order to minimize the risk of overflagging in soundings with poor vertical resolution, each procedure has certain minimum data requirements that are tailored to its potential vulnerabilities.

a. Test for excessive fluctuations

The first of the vertical consistency checks tests for excessive level-to-level temperature fluctuations such as those shown in Figure 4a. Although many of the positive and negative extremes in the sounding would be identified as errors by the Tier-2 climatological check (Fig. 4b), some unlikely fluctuations would remain in the lower troposphere. To avoid retaining such profiles in the data, the test for excessive fluctuations, like the whole-profile climatological check, identifies soundings in which errors are so pervasive that the entire temperature profile is called into question.

The procedure uses the median of the absolute differences between the temperature z-scores at consecutive levels as a climatologically-independent measure of the degree to which temperatures change from level to level. Temperature profiles with a "median absolute z-score difference" greater than 3.0 (Table 1) are removed from the data. For example, in the rather extreme case shown in Figure 4, the absolute level-to-level differences range from 0.02 to 10.28, and the median of all 22 differences is equal to 6.13, a value that clearly exceeds the test threshold.

b. Gap checks

Another form of vertical inconsistency arises when an entire section of a sounding deviates significantly from the rest of the profile. For example, in the sounding shown in Figure 5a, the temperatures at the top two levels are much warmer than the temperatures at the levels immediately below. Perhaps even more striking is that their z-scores are clearly separated from all other z-scores in the sounding (Fig. 5b). It is this principle of comparing a z-score to the frequency distribution of z-scores at other levels that is utilized by the two gap checks.

The first of these checks considers the frequency distribution of z-scores from the entire sounding. The procedure sorts the standardized temperatures of a particular sounding and looks for an unusually large gap in the resulting frequency distribution. If such a gap is located in the upper or lower tail of the distribution, then all values on the far side of the gap are considered to be invalid. In the case of Figure 5b, there exists a gap of 3.73 between the second- and third-largest z-scores.

A more localized variant of this procedure compares the z-score at each level to the distribution of z-scores at surrounding levels. In this variant, a level is included in the calculation of the z-score distribution only if the ratio of its pressure to the pressure of the level being tested lies between 0.5 and 2.0. This ensures that the frequency distribution of z-scores is computed only from levels within a similar portion of the atmosphere. Such a restriction can be useful, for example, when the stratosphere is unusually warm relative to the troposphere, and there also exists a warm outlier at the surface (Fig. 6). When testing the surface point in this example, the pressure ratio limitation implies that only the points up to the 500-hPa level can be considered. As a result, the gap between the surface z-score and the next largest z-score is more than twice as large as it would be if the z-score distribution were computed from the entire profile (6.07 vs. 2.99).

For both gap checks, the key parameter is the magnitude of the gap. The threshold for this parameter was set to 3.5 for the whole-profile version and to 2.0 for the partial-profile version (Table 1). During the threshold selection process, it was found that the two algorithms are prone to overflagging when the z-scores of a sounding are far from normally distributed, as is sometimes the case during certain meteorological situations and when the sounding is incomplete. To guard against this problem, the gap checks are applied only to those profiles in which at least two-thirds of the points are clustered near the center of the distribution (i.e., within 1.5 units of the median z-score), and values are flagged only when the identified gap falls entirely outside the central two-thirds of the distribution.

c. Spike check

Perhaps the simplest form of vertical inconsistency is a spike (or dip) created by a single data point that does not fit into an otherwise reasonable profile (e.g., Fig. 7). In the strictest sense, outliers of this type tend to be characterized by an unusually rapid change in temperature in the layers immediately below and above the level in question. Making use of this characteristic, the spike check, testing one value at a time, checks whether the z-score differences to the levels above and below exceed a certain "absolute z-score difference threshold," and are of opposite sign.

An important distinction between this test and the other procedures in the IGRA system is that the test threshold itself is a function rather than a constant. Since the spacing between levels varies widely throughout historical radiosonde data, the absolute z-score difference threshold is calculated as a function of the ratio of the pressures of the levels being tested. Consequently, it is not the absolute z-score difference itself, but the shape and coefficients of this function that are the fixed characteristics of the test. Rather than requiring the selection of a single threshold, the thresholds selection process therefore consisted of two steps: 1) the identification of an appropriate absolute z-score difference threshold for each of several pressure ratios (Table 1), and 2) fitting a function to the resulting (pressure ratio, absolute difference threshold) points. The function chosen based on this analysis is linear with an intercept of 6.42 and a slope of -3.52 (Table 1). Thus, the absolute z-score difference threshold increases from a minimum of 2.90 to a maximum of 6.42 as the distance between levels increases. This function is applied regardless of location, time, and vertical resolution.

The spike test then works as follows. For each level in a sounding, the following quantities are calculated: the z-score difference and pressure ratio between level i and the

next lower level; the z-score difference and pressure ratio between the next higher level and level i; and the corresponding z-score difference thresholds for the two pairs of levels. The temperature at level i fails the test if three conditions are true: 1) the absolute value of the z-score difference between levels i+1 and i exceeds the corresponding threshold; 2) the absolute value of the z-score difference between levels i and i-1 exceeds its respective threshold; and 3) the two z-score differences are of opposite sign. In Figure 7b, for example, the z-score at 400 hPa exceeds the z-scores at the levels immediately below and above, yielding z-score differences of 3.48 and -3.52, respectively. Since these differences are of opposite sign, and the corresponding absolute difference thresholds based on the relevant pressure ratios are 2.65 and 2.58, the 400-hPa data point fails the test.

While both the spike check and the gap checks described in the previous subsection are effective at detecting isolated outliers in relatively complete soundings, they complement each other in terms of their ability to detect vertical inconsistencies under other circumstances. The spike test detects isolated errors in profiles whose number or distribution of data points does not permit the application of one or both of the gap checks. The gap checks, on the other hand, are able to detect groups of outliers at consecutive levels and identify errors at the top or bottom level of a sounding where the spike check cannot be applied.

7. Temporal consistency checks

The final set of IGRA QA procedures for temperature consists of two variants of a test for temporal inconsistencies. While such tests are sometimes applied to hourly and

daily observations of surface temperature (Reek et al. 1992; Graybeal et al. 2004b), they are not usually employed in QA systems for radiosonde data. Most such systems are designed to check the quality of data from one instance in time, thus making the application of a temporal consistency check impractical. Even when historical radiosonde observations are being tested, the varying temporal and vertical resolution of the data make it difficult to apply a test that compares data from consecutive soundings. However, as illustrated in Figures 8 and 9, the application of runs, outlier, and vertical consistency checks is not sufficient for identifying all values that appear as clear outliers when viewed from a time series perspective.

Based on the above considerations, the IGRA temporal consistency checks are equipped with several safeguards. First, they are applied only to the surface and mandatory levels where time series are far more continuous than at significant levels. Second, a value is compared to all observations within a specified time window rather than solely to the observations immediately preceding and following it. Finally, like the vertical consistency checks, the tests for temporal inconsistencies operate on standardized temperatures rather than on raw observations, thus reducing the influence of geographical differences in variability. For the temporal consistency checks, however, standardization is performed based on the overall mean and STD for each station and level, as for the Tier-1 climatological check. Although this approach implies that the annual cycle is retained in the standardized values, it also has the distinct advantage of allowing for the application of the test to a larger number of time series then would be possible if the more precise Tier-2 climatology were used.

Specifically, the algorithm itself works as follows. The z-score of the value being tested is compared to all other z-scores within a specified time window centered on the relevant day. If the z-score is found to differ by more than a specified number of STDs from both the next largest and next smallest z-score within the window, then the procedure flags not only the tested temperature, but also all other temperatures within the window that are identical to it. This approach allows for the identification of both isolated outliers and erroneous clusters of identical temperatures that are not detected by the other checks. To improve the efficiency of the check and reduce the risk of overflagging in cases of unusual meteorological situations or gaps in the data, a temperature qualifies for this test only if its z-score exceeds a certain value and if there are a sufficient number of observations within the time window.

The algorithm thus depends on four parameters: the length of the time window, the number of observations within the time window, the z-score threshold that must be exceeded in order for a temperature to be tested, and the z-score difference that identifies a temperature as an error. In order to accommodate different levels of temporal completeness, two variants of this procedure are applied, one with a time window of 45 days, and the other with a time window of five years. Based on initial testing with different thresholds for the minimum number of observations, the completeness requirement in both cases stipulates that the test can be applied only when z-scores are available on at least half of the days in the window. For the two remaining parameters, the following thresholds were chosen based on a systematic evaluation of time series with outliers of different magnitudes (Table 1): For the 22.5-day window, the tested temperature must have a z-score of at least 2.5 and must differ from the other z-scores by

at least 2 STDs in order to be considered erroneous. For the five-year test, the z-score threshold is also set to 2.5, but the difference threshold is equal to only 1.0. These thresholds allow for the detection of values that are clearly erroneous from a time series perspective while limiting the risk of labeling as erroneous sharp features associated with frontal passages or other phenomena.

As an example of how the temporal consistency checks operate, consider the outliers shown in Figures 8 and 9. In Figure 8, the time series of 100-hPa temperatures at Lajes, Portugal, has a mean of -63°C and an STD of 3.77°C, implying that temperatures above -54.6°C or below -72.4°C are tested for temporal consistency. The -50°C temperature reported at 1100 UTC on 13 September 2000 is identified as temporally inconsistent by the 45-day check because within 22.5 days before and after this time, there are no other temperatures within two STDs of this value. Similarly, the five-year check labels the 250-hPa temperature reported at Goose Bay, Canada, on 11 March 1982 (Fig. 9) as erroneous since there are no other temperatures within one STD (5.38°C) of it during the five-year period centered on this date.

8. System performance

As part of the overall IGRA processing system, the suite of temperature checks flags approximately 0.27% of the more than 500 million temperature observations in the entire database. The climatological checks set approximately two-thirds of these flags, while the vertical consistency checks account for much of the remainder (Table 1). Even though the runs and temporal consistency tests detect considerably fewer errors, they fulfill a significant need by identifying problems that could affect studies of variability or

extremes. This disparity in flag rates illustrates that the effectiveness of a QA system or any of its components should be measured not so much in terms of the percentage of data flagged, but in terms of the degree to which obvious errors are removed and valid values are left intact. With this in mind, several measures were taken to assess the performance of the IGRA QA system as a whole (Durre et al. 2008). First, the basic integrity of the procedures was tested by means of several "sanity" checks of the final quality-assured data. In addition, a random sample of flagged values was visually examined in order to obtain an estimate of the overall false-positive rate of the system. Finally, a similar random inspection of unflagged values was performed to ensure that the percentage of undetected errors in the quality-assured data was not unreasonably high.

The sanity checks included visual examinations of the spatial and temporal distribution of flagged values, of maps of long-term monthly means at the surface and mandatory pressure levels, and of selected time series and soundings at locations known to experience somewhat unusual meteorological conditions. The purpose of all of these checks was to identify cases in which the automated system either fails to detect a large number of errors or misidentifies a significant number of valid values as erroneous. For example, stations where systematic errors remain in the data may appear as geographically inconsistent on a map of climatological means. On the other hand, if a QA system had a tendency to misidentify a sudden stratospheric warming as a data problem, this tendency might manifest itself as the frequent flagging of unusually warm stratospheric temperatures, such as those found in early 1989 at Jan Mayen (Fig. 10). In the case of the IGRA temperature QA procedures, the sanity checks revealed no systematic tendency of either type.

A more quantitative assessment of the overall system performance was obtained through a final set of manual inspections during which the false-positive and miss rates of several groups of procedures were determined. The runs, gross plausibility, and Tier-1 climatological checks were not included in this final evaluation since their thresholds had been chosen so as to avoid all false positives and leave the detection of any errors missed to the subsequent tests. The remainder of the checks were divided into two groups: 1) procedures that look at the character of a sounding (i.e., the climatological whole-profile check and the four vertical consistency checks) and 2) procedures that look at data purely from a time perspective (i.e., the Tier-2 climatological and temporal consistency checks). For each of the two groups, one flagged and one unflagged value was randomly selected from mandatory-level observations at each GCOS Upper Air Network (GUAN) station. This yielded a sample size of 130 flagged and 130 unflagged values for each group. For each of the chosen observations, a plot of the relevant vertical profile and time series was generated. One of the co-authors (RSV) then subjectively identified each of the data values as either valid or invalid, without knowledge of the assessment made by the automated system. During this evaluation, 6% of the flags generated by the sounding-based checks and 20% of the values flagged by the Tier-2 climatology and temporal consistency checks were subjectively judged to be valid. Taking into account sampling variability, the statistical 95% confidence limits on these two false-positive rates are approximately ±2% and ±4%, respectively. Since these two groups of tests account for approximately one-fourth and one-third of the total number of values flagged, respectively (Table 1), these findings imply that for the system as a whole, approximately

9 out of every 10 values flagged can be expected to be errors. At the same time, 1.1% of the unflagged values inspected were judged to be marginally erroneous.

These results suggest that the combination of different types of QA checks is effective at removing gross data errors without compromising unique meteorological events. For example, features such as the period of strong warming in 50-hPa temperatures found in early 1989 at Jan Mayen (Fig. 10) are left intact, while the isolated outliers found earlier in the time series are removed. The presence of marginal errors in the evaluated sample is indicative of the fact that the QA system has been designed to detect only the grossest errors. Many of the additional humanly identifiable errors reflect unique situations that would require the development of highly specialized checks for the identification of only a few additional errors. The alternative of lowering the thresholds of the existing tests, while significantly increasing the number of errors detected, would result in a much larger increase in the number of valid values flagged. Consequently, we consider the current system to be the most desirable compromise among error detection, false-positive rate, and system complexity.

9. Concluding remarks

The QA procedures described in this paper constitute a fully automated, robust system for quality-assuring radiosonde temperature measurements. Since the system is intended for application to historical and real-time radiosonde data from around the globe, the procedures are designed to complement each other in terms of the types of data errors they detect and to compensate for each other's limitations. The runs, outlier, vertical consistency, and temporal consistency checks are applied in a sequence (Table 1)

in which the removal of the grossest errors by the earlier tests benefits the performance of the later tests. At a minimum, each temperature is subjected to the gross plausibility and runs tests. Each additional test is applied only when the relevant climatological statistics are available and any requirements for vertical or temporal resolution are met.

The suite of temperature QA tests is part of the system that processes the IGRA data (Durre et al. 2006). Careful manual inspection of random samples of values flagged in this dataset indicates that the overall false-positive rate of the temperature checks lies around 10% out of the total number of values flagged. At the same time, the error detection capabilities of the different tests make it possible for the entire system to effectively identify the gross errors in the data.

This robust performance notwithstanding, additional checks may seem desirable for certain applications. For example, a possible approach for checking for systematic errors in station records would be to compare the radiosonde data with reanalysis products. In areas where the station network is sufficiently dense, a carefully-designed test for spatial inconsistencies may be capable of detecting additional errors. Similarly, a check for hydrostatic consistency among pressure, temperature, and geopotential height, as described by Gandin (1988), may further enhance the error detection capabilities of the system at locations and altitudes where temperature tends to vary linearly with height.

Acknowledgments

We thank Jon Burroughs and Dr. Xungang Yin for their assistance in the preparation of figures. We also thank the reviewers for comments on earlier drafts of this manuscript. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (Grant number DE-AI02-96ER62276).

References

Collins, W. G., 2001a: The operational complex quality control of radiosonde heights and temperatures at the National Centers for Environmental Prediction. Part I: Description of the method. *J. Appl. Meteor.*, **40**, 137–151.

_____, 2001b: The operational complex quality control of radiosonde heights and temperatures at the National Centers for Environmental Prediction. Part II: Examples of error diagnosis and correction from operational use. *J. Appl. Meteor.*, **40**, 152–168.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. *J. Climate*, **19**, 53-68.

_____, M. J. Menne, and R. S. Vose, 2008: Strategies for evaluating quality assurance procedures. *J. Appl. Meteor.*, in press.

Eskridge, R. E., O. A. Alduchov, I. V. Chernykh, P. Zhai, A. C. Polansky, and S. R. Doty, 1995: A comprehensive aerological reference data set (CARDS): Rough and systematic errors. *Bull. Amer. Meteor. Soc.*, **76**, 1759-1775.

Free, M., I. Durre, E. Aguilar, D. Seidel, T. C. Peterson, R. E. Eskridge, J. K. Luers, D. Parker, M. Gordon, J. Lanzante, S. Klein, J. Christy, S. Schroeder, B. Soden, L. M. McMillin, and E. Weatherhead, 2002: Creating climate reference datasets: CARDS workshop on adjusting radiosonde temperature data for climate monitoring. *Bull. Amer. Meteor. Soc.*, **83**, 891–899.

_____, D. J. Seidel, J. K. Angell, I. Durre, and T. C. Peterson, 2005: Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new dataset of large-area anomaly series. *J. Geophys. Res.*, **110** (D22), doi:10.1029/2005JD006169.

Gandin, L. S., 1988: Complex quality control of meteorological observations. *Mon. Wea. Rev.*, **116,** 1137–1156.

_____, L. L. Morone, and W. G. Collins, 1993: Two years of operational comprehensive hydrostatic quality control at the National Meteorological Center. *Wea. Forecasting*, **8,** 57-72.

Graybeal, D. Y., A. T. DeGaetano, and K. L. Eggleston, 2004a: Complex quality assurance of historical hourly surface airways meteorological data. *J. Atmos. Oceanic Tech.*, **21**, 1156-1169.

_____, _____, and _____, 2004b: Improved quality assurance for historical hourly temperature and humidity: Development and application to environmental analysis. *J. Appl. Meteor*., **43**, 1722-1735.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate*, **19**, 1377-1403.

Kahl, J. D., M. C. Serreze, S. Shiotani, S. M. Skony, and R. C. Schnell, 1992: In-situ meteorological sounding archives for Arctic studies. *Bull. Amer. Meteor. Soc.*, **73**, 1824-1830.

Lanzante, J. R., 1996: Resistant, robust and nonparametric techniques for analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.*, **16**, 1197-1226.

Loehrer, S. M., T. A. Edmands, and J. A. Moore, 1996: TOGA COARE Upper-Air Sounding Data Archive: Development and quality control procedures. *Bull. Amer. Meteor. Soc.*, **77**, 2651–2672.

Parker, D. E., and D. I. Cox, 1995: Towards a consistent global climatological

rawinsonde data-base. *Int. J. Climatol.*, **15**, 473-496.

Peterson, T. C., R. S. Vose, V. N. Razuvaev, and R. L. Schmoyer, 1998: Global Historical Climatology Network (GHCN) quality control of monthly temperature data. *Int. J. Climatol.*, **18**, 1169-1179.

Reek, T., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the Cooperative Network. *Bull. Amer. Meteor. Soc.*, **73**, 753-765.

Schwartz, B. E., and C. A. Doswell III, 1991: North American rawinsonde observations: Problems, concerns, and a call to action. Bull. *Amer. Meteor. Soc.*, **72**, 1885-1896.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, P. Brohan, and J. R. Knight, 2005: Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J. Geophys. Res.*, **110**, doi:10.1029/2004JD005753.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate*, **10**, 1980-1997.

List of tables

Table 1. QA checks applied to IGRA temperatures. Procedures are listed in the order in which they are applied. Flag rates are based on IGRA data through 2005.

List of figures

Figure 1. Time series of 250-hPa temperature (January-December 2003) at Campo Grande, Brazil, showing a run across 35 soundings from all times of day during 10-29 January.

Figure 2. Time series of 500-hPa temperature for 1200 UTC (January-December 1972) at Fort Sill, Oklahoma, USA, showing outliers detected by the Tier-1 and Tier-2 climatological checks. Also plotted are the upper and lower limits for the Tier-1 (black dotted lines) and Tier-2 (gray solid lines) checks. The -55.9°C temperature on 4 August falls outside both limits, whereas the -15.5°C temperature on 25 July falls inside the Tier-1, but outside the Tier-2 limits. See text for details on how the limits are determined.

Figure 3. Example of a sounding that the whole-profile climatological check identifies as erroneous. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 24 March 1996 at Tunis, Tunisia. The profile fails the test because its median absolute z-score (9.95) exceeds the test threshold of 4.00.

Figure 4. Example of a sounding that the check for excessive level-to-level fluctuations identifies as erroneous. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 25 December 1974 at Atyray, Kazakhstan. The sounding fails the test because its median absolute level-to-level z-score difference of 6.13 exceeds the test threshold of 3.00.

Figure 5. Sample profile in which temperatures at the top two levels are identified as errors by the whole-profile gap check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 1200 UTC on 20 March 1986 at Tura, Russia. Since more than two-thirds (87.1%) of the z-scores lie within 1.5 units of the median z-score (-0.82), the profile qualifies for the test. The temperatures at 30 and 20 hPa fail the test because their z-scores are separated from the other z-scores by a gap of 3.73, which exceeds the test threshold of 3.50.

Figure 6. Profile in which a surface temperature is identified as an error by the partial-profile gap check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 1200 UTC on 31 December 1971 at Petropavlovsk, Russia. See text for details.

Figure 7. Sample profile in which a temperature is identified as an error by the vertical spike check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 15 September 1964 at Jan Mayen, Norway. The temperature at 400 hPa fails the test because its z-score (2.34) exceeds the z-scores at the levels immediately below (-1.14) and above (-1.18) by more than permitted by the test. See text for details on how the thresholds are determined.

Figure 8. Time series of 100-hPa temperature (January-December 2000) at Lajes, Portugal, showing an outlier identified by the 45-day temporal consistency check. The -50°C temperature reported at 1100 UTC on 13 September fails the test because there are

no corroborating points within two STDs of this temperature during the 45-day window centered on the point. The temperature limits and time window are indicated by the box surrounding the outlier.

Figure 9. Time series of 250-hPa temperature (January 1979-December 1984) at Goose Bay, Canada, showing an outlier identified by the five-year temporal consistency check. The -74.9°C temperature reported at 0000 UTC on 11 March 1982 fails the test because there are no corroborating points within one STD of this temperature during the five-year window centered on the point. The temperature limits and time window are indicated by the box surrounding the outlier.

Figure 10. Time series of 50-hPa temperature (January 1980-December 1989) at Jan Mayen, Norway, (a) prior to and (b) after the application of all QA procedures. Note that several outliers before 1985 are removed by the QA process, while the more coherent feature of unusually warm temperatures in early 1989 is retained.

Table 1. QA checks applied to IGRA temperatures. Procedures are listed in the order in which they are applied. Flag rates are based on IGRA data through 2005.

Type of error detected	Key test parameters	Parameter values evaluated	Threshold chosen	% of values flagged
Gross plausibility check	Temperature	-120°C and 70°C	-120°C and 70°C	0.0035%
Run across soundings of any hour	Number of consecutive identical temperatures at a particular mandatory level	5-30, increment 5	15	0.0002%
Run across soundings at a specific hour	Number of consecutive identical temperatures at a particular mandatory level and hour	5-30, increment 5	15	0.0002%
Run across mandatory levels within a sounding	Number of consecutive identical mandatory-level temperatures	3, 5, 10, 15	5	0.0001%

Run across significant levels within a sounding	Number of consecutive identical significant-level temperatures	3, 5, 10, 15	5	0.0002%
Tier-1 climatological outlier	z-score based on all values at a particular station and level	2.0-7.0, increment 1.0	6.0	0.1094%
Whole-profile climatological outlier	Median of absolute Tier-2 z-scores in sounding	3.0-6.0, increment 0.5	4.0	0.0078%
Tier-2 climatological outlier	z-score based on 45-day and 3-hour windows at each station and level	3.0-6.0, increment 0.5	5.0	0.0819%
Excessive fluctuations in profile	Median of absolute Tier-2 z-score differences	1.0-6.0, increment 1.0	3.0	0.0107%
Whole-profile gap	Size of largest gap in the profile's Tier-2 z-score distribution	2.0-5.0, increment 0.5	3.5	0.0002%
Partial-profile gap	Size of largest gap in Tier-2 z-score distribution	1.0-4.0, increments 0.5	2.0	0.0474%

Spike in profile	Absolute tier-2 z-score differences (dz) with next higher and lower levels as a function of pressure ratio (rp)	dz of 2.0-5.0 for rp of 0.5-0.95	dz = 6.42-3.52*rp	0.0111%
Temporal inconsistency within 45-day window	Smallest absolute tier-1 z-score difference (dz) with other values in time window; magnitude of z-score (z)	dz of 1.0-4.0; z of 2.0-4.0	dz=2.0; z=2.5	0.0035%
Temporal inconsistency within 5-year window	Smallest absolute tier-1 z-score difference (dz) with other values in time window; magnitude of z-score (z)	dz of 0.5-3.0; z of 2.0-4.0	dz=1.0; z=2.5	0.0007%

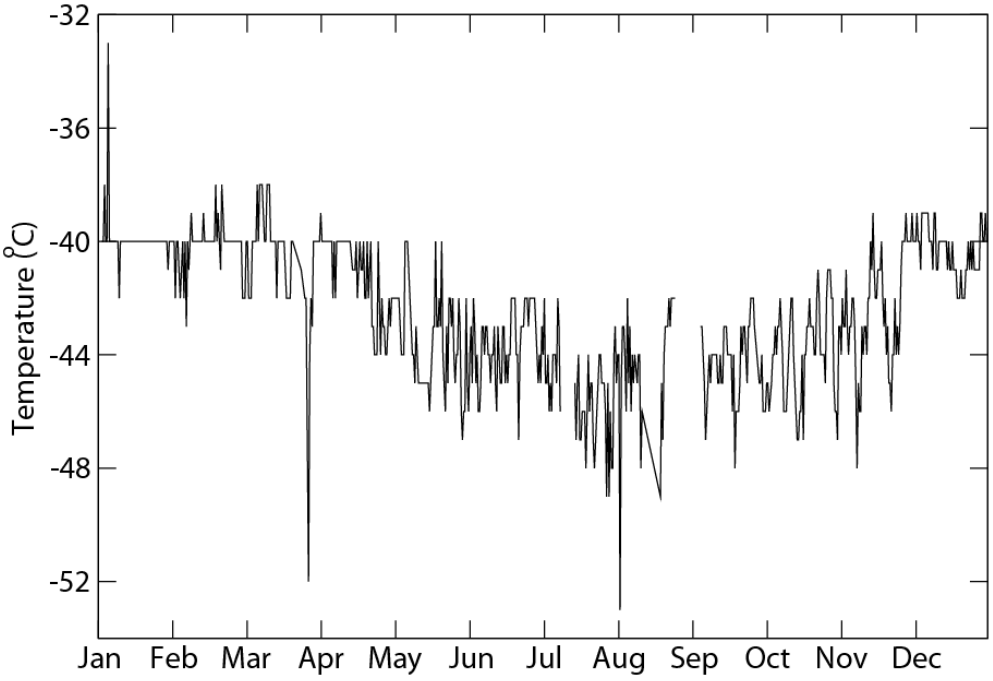

Figure 1. Time series of 250-hPa temperature (January-December 2003) at Campo Grande, Brazil, showing a run across 35 soundings from all times of day during 10-29 January.

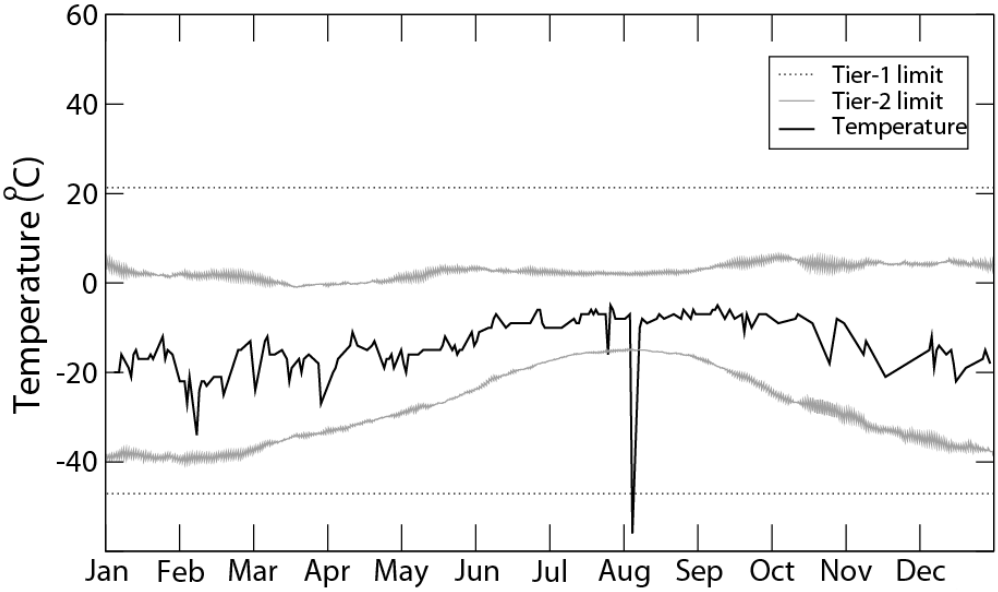

Figure 2. Time series of 500-hPa temperature for 1200 UTC (January-December 1972) at Fort Sill, Oklahoma, USA, showing outliers detected by the Tier-1 and Tier-2 climatological checks. Also plotted are the upper and lower limits for the Tier-1 (black dotted lines) and Tier-2 (gray solid lines) checks. The -55.9°C temperature on 4 August falls outside both limits, whereas the -15.5°C temperature on 25 July falls inside the Tier-1, but outside the Tier-2 limits. See text for details on how the limits are determined.

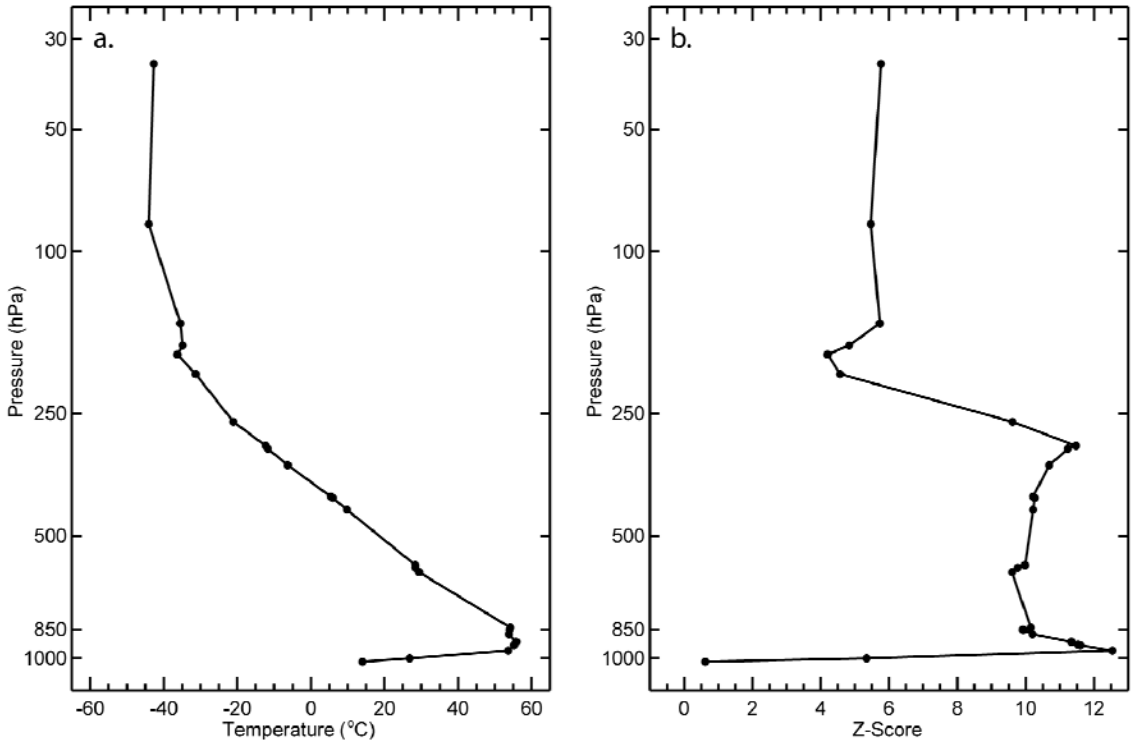

Figure 3. Example of a sounding that the whole-profile climatological check identifies as erroneous. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 24 March 1996 at Tunis, Tunisia. The profile fails the test because its median absolute z-score (9.95) exceeds the test threshold of 4.00.

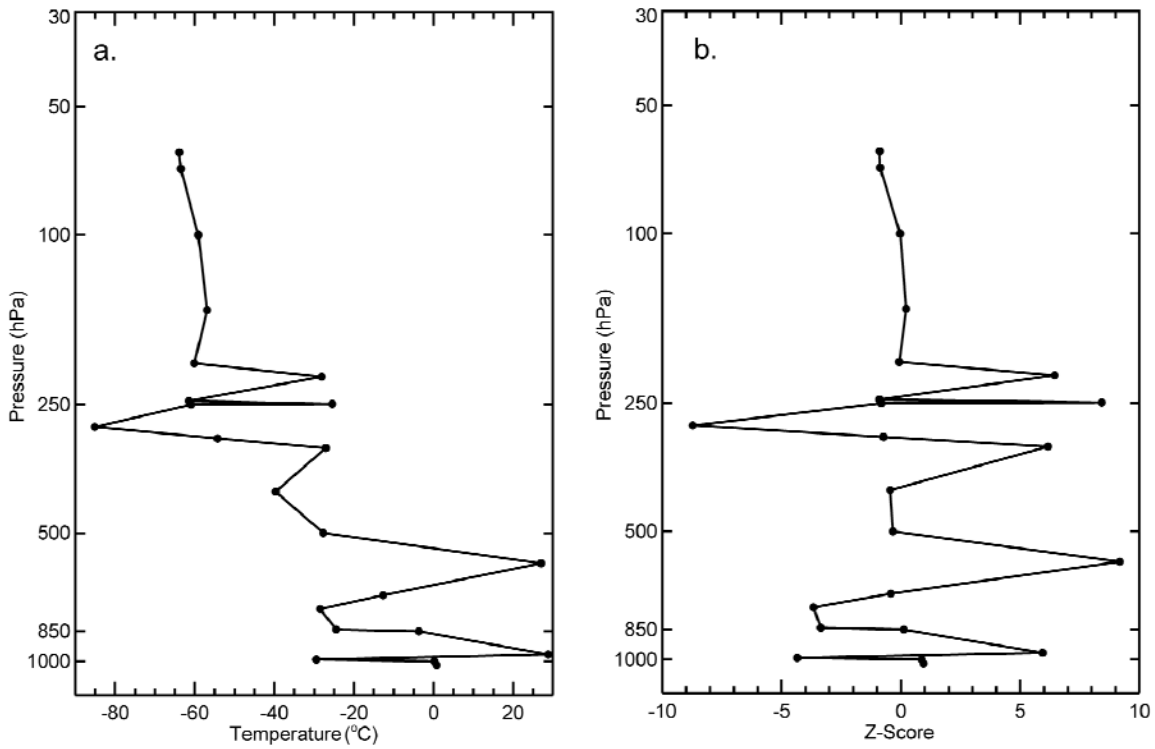

Figure 4. Example of a sounding that the check for excessive level-to-level fluctuations identifies as erroneous. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 25 December 1974 at Atyray, Kazakhstan. The sounding fails the test because its median absolute level-to-level z-score difference of 6.13 exceeds the test threshold of 3.00.

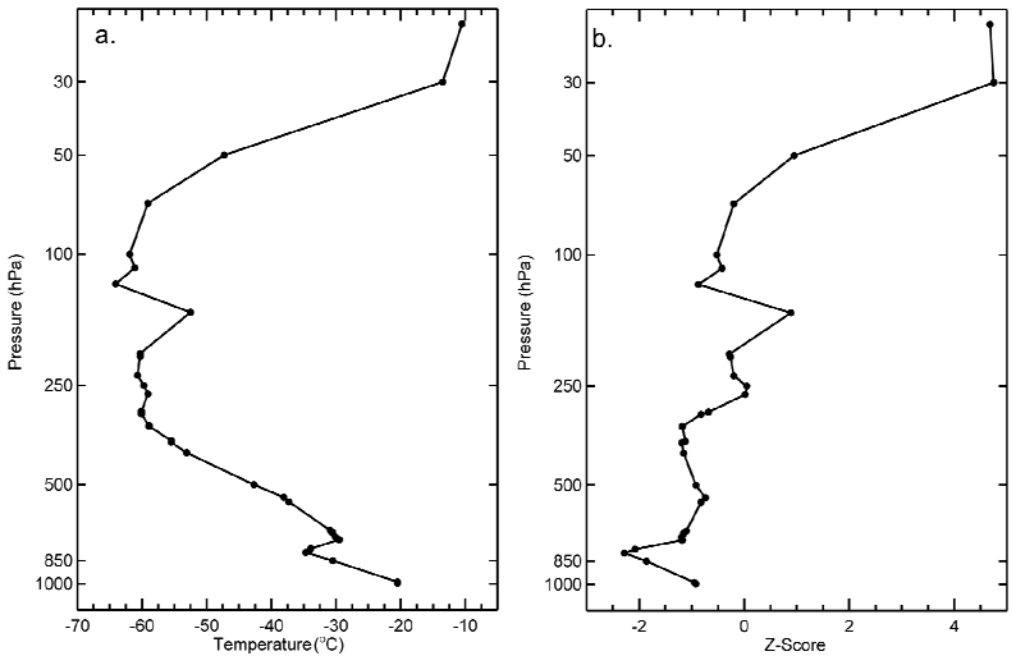

Figure 5. Sample profile in which temperatures at the top two levels are identified as errors by the whole-profile gap check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 1200 UTC on 20 March 1986 at Tura, Russia. Since more than two-thirds (87.1%) of the z-scores lie within 1.5 units of the median z-score (-0.82), the profile qualifies for the test. The temperatures at 30 and 20 hPa fail the test because their z-scores are separated from the other z-scores by a gap of 3.73, which exceeds the test threshold of 3.50.

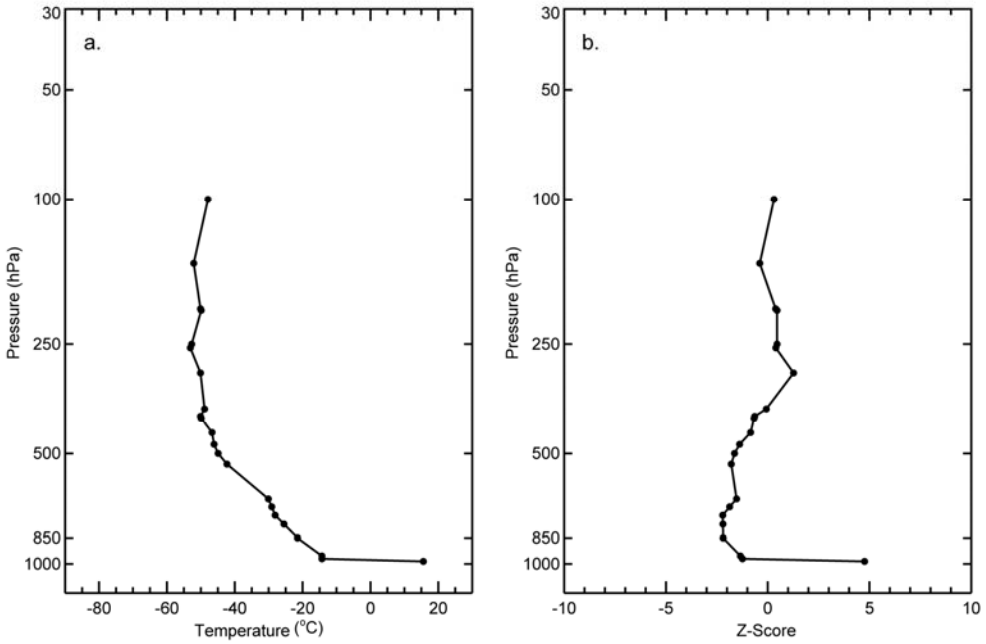

Figure 6. Profile in which a surface temperature is identified as an error by the partial-profile gap check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 1200 UTC on 31 December 1971 at Petropavlovsk, Russia. See text for details.

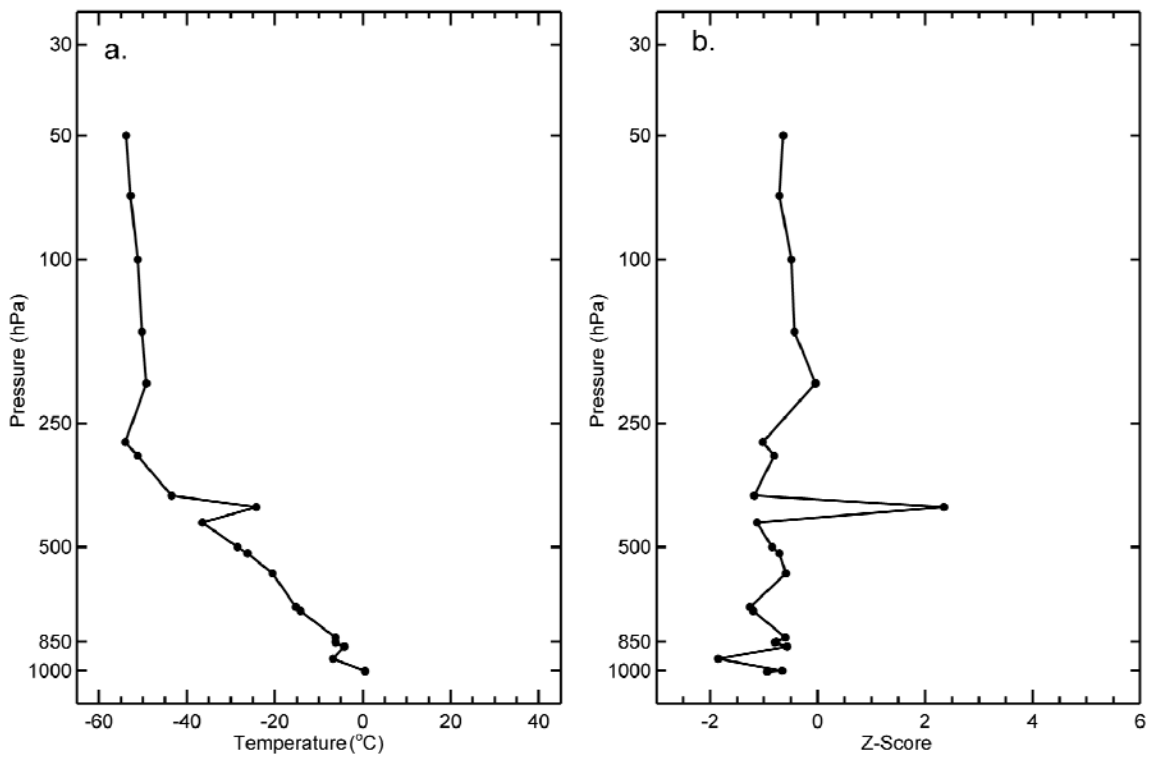

Figure 7. Sample profile in which a temperature is identified as an error by the vertical spike check. (a) Temperatures and (b) corresponding Tier-2 z-scores for 0000 UTC on 15 September 1964 at Jan Mayen, Norway. The temperature at 400 hPa fails the test because its z-score (2.34) exceeds the z-scores at the levels immediately below (-1.14) and above (-1.18) by more than permitted by the test. See text for details on how the thresholds are determined.

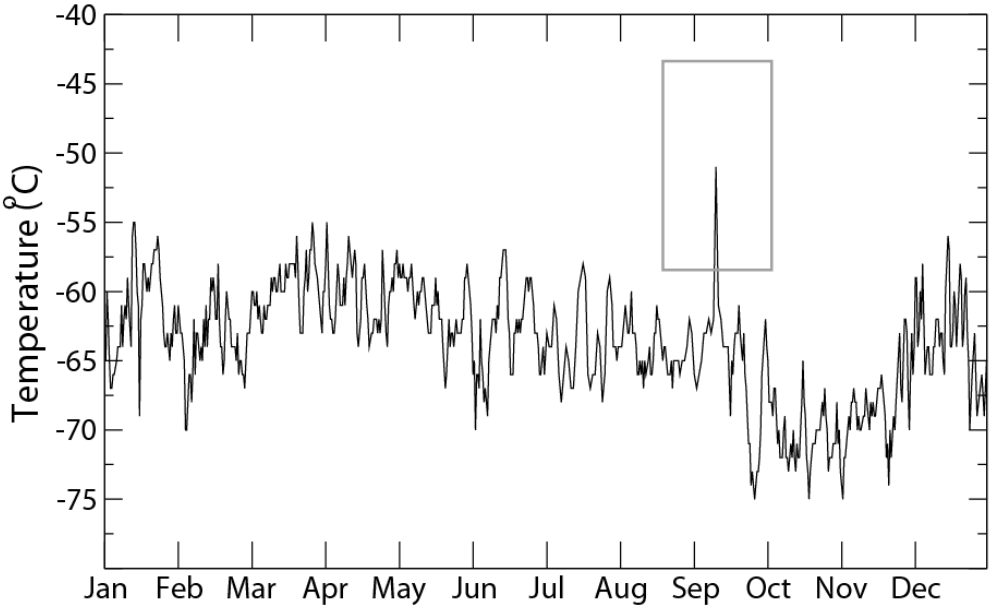

Figure 8. Time series of 100-hPa temperature (January-December 2000) at Lajes, Portugal, showing an outlier identified by the 45-day temporal consistency check. The -50°C temperature reported at 1100 UTC on 13 September fails the test because there are no corroborating points within two STDs of this temperature during the 45-day window centered on the point. The temperature limits and time window are indicated by the box surrounding the outlier.

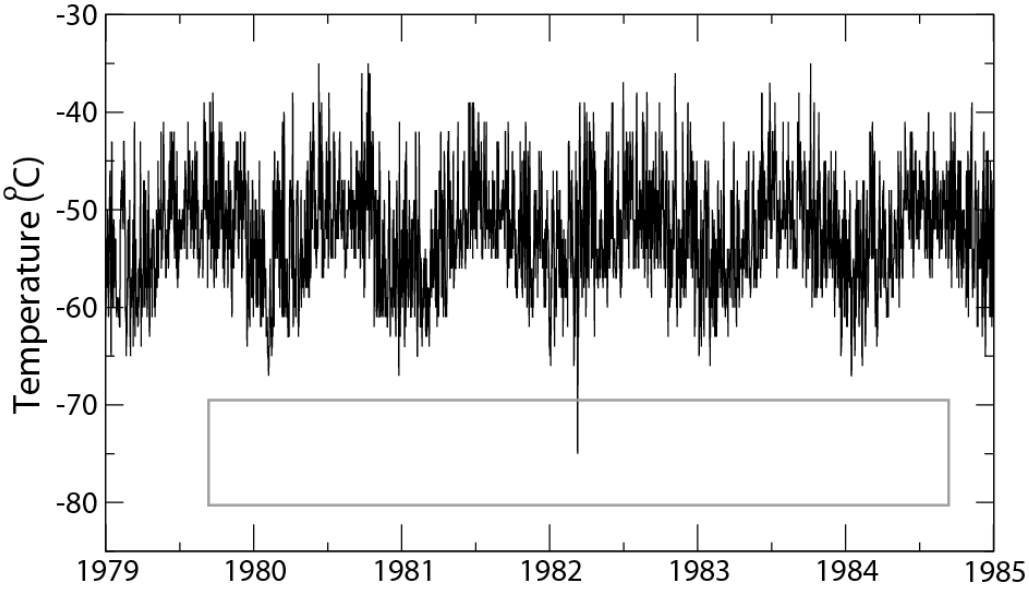

Figure 9. Time series of 250-hPa temperature (January 1979-December 1984) at Goose Bay, Canada, showing an outlier identified by the five-year temporal consistency check. The -74.9°C temperature reported at 0000 UTC on 11 March 1982 fails the test because there are no corroborating points within one STD of this temperature during the five-year window centered on the point. The temperature limits and time window are indicated by the box surrounding the outlier.

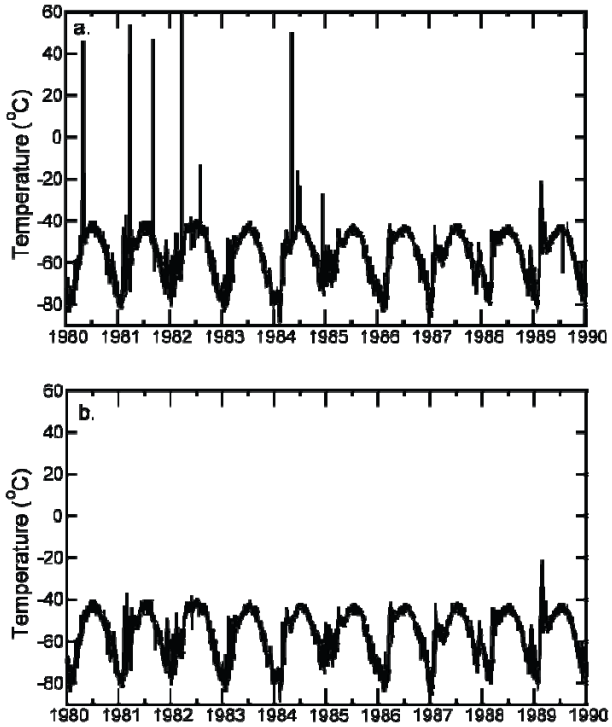

Figure 10. Time series of 50-hPa temperature (January 1980-December 1989) at Jan Mayen, Norway, (a) prior to and (b) after the application of all QA procedures. Note that several outliers before 1985 are removed by the QA process, while the more coherent feature of unusually warm temperatures in early 1989 is retained.