A New Approach to Homogenize Daily Radiosonde Humidity Data

AIGUO DAI AND JUNHONG WANG

National Center for Atmospheric Research,* Boulder, Colorado

PETER W. THORNE[+] AND DAVID E. PARKER

Met Office Hadley Centre, Exeter, United Kingdom

LEOPOLD HAIMBERGER

Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

XIAOLAN L. WANG

Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, Ontario, Canada

(Manuscript received 29 April 2010, in final form 1 October 2010)

ABSTRACT

Radiosonde humidity records represent the only in situ observations of tropospheric water vapor content with multi-decadal length and quasi-global coverage. However, their use has been hampered by ubiquitous and large discontinuities resulting from changes to instrumentation and observing practices. Here a new approach is developed to homogenize historical records of tropospheric (up to 100 hPa) dewpoint depression (DPD), the archived radiosonde humidity parameter. Two statistical tests are used to detect changepoints, which are most apparent in histograms and occurrence frequencies of the daily DPD: a variant of the Kolmogorov–Smirnov (K–S) test for changes in distributions and the penalized maximal F test (PMFred) for mean shifts in the occurrence frequency for different bins of DPD. These tests capture most of the apparent discontinuities in the daily DPD data, with an average of 8.6 changepoints (~1 changepoint per 5 yr) in each of the analyzed radiosonde records, which begin as early as the 1950s and ended in March 2009. Before applying breakpoint adjustments, artificial sampling effects are first adjusted by estimating missing DPD reports for cold ($T < -30°C$) and dry (DPD artificially set to 30°C) conditions using empirical relationships at each station between the anomalies of air temperature and vapor pressure derived from recent observations when DPD reports are available under these conditions. Next, the sampling-adjusted DPD is detrended separately for each of the 4–10 quantile categories and then adjusted using a quantile-matching algorithm so that the earlier segments have histograms comparable to that of the latest segment. Neither the changepoint detection nor the adjustment uses a reference series given the stability of the DPD series.

Using this new approach, a homogenized global, twice-daily DPD dataset (available online at www.cgd.ucar.edu/cas/catalog/) is created for climate and other applications based on the Integrated Global Radiosonde Archive (IGRA) and two other data sources. The adjusted-daily DPD has much smaller and spatially more coherent trends during 1973–2008 than the raw data. It implies only small changes in relative humidity in the lower and middle troposphere. When combined with homogenized radiosonde temperature, other atmospheric humidity variables can be calculated, and these exhibit spatially more coherent trends than without the DPD homogenization. The DPD adjustment yields a different pattern of change in humidity parameters compared to the apparent trends from the raw data. The adjusted estimates show an increase in tropospheric water vapor globally.

---

1. Introduction

Water vapor is the single largest contributor to the natural greenhouse effect in the atmosphere. Its content is expected to rise as greenhouse gas (GHG)-induced global warming continues in both models (Held and Soden 2000; Dai et al. 2001; Meehl et al. 2007) and the real world (Ross and Elliott 2001; Zhai and Eskridge 1997; Trenberth et al. 2005; Durre et al. 2009; McCarthy

et al. 2009). How much water vapor increases is a key feedback that determines climate sensitivity (Held and Soden 2000; Dessler and Sherwood 2009), particularly above the middle troposphere, where the effect of water vapor is not radiatively saturated (Solomon et al. 2010). Water vapor also plays a key role in the global hydrologic cycle (Trenberth et al. 2007), because it acts as a medium for water and heat exchange and transport within the climate system and also because it is linked to the formation of clouds and precipitation. Thus it is crucial to monitor long-term changes in atmospheric water vapor content using various types of observations. Since the middle of the twentieth century, many countries have made routine observations of atmospheric humidity using balloon-borne radiosondes. These radiosonde humidity observations provide the only record that has high vertical resolution and is long enough for quantifying multidecadal changes in atmospheric water vapor, although other observations, such as surface humidity data (Dai 2006; Willett et al. 2008) and recent satellite (Trenberth et al. 2005) and global positioning system (GPS) observations (Wang et al. 2007), can also provide useful information.

The raw radiosonde humidity records contain large errors and biases. They are notoriously inhomogeneous because of difficulties in accurately measuring humidity, changes in hygrometers and observing practices, and other factors, such as changing methods of converting relative humidity (RH) to dewpoint depression (DPD) for archives (Gaffen et al. 1991; Gaffen 1993, 1994, 1996; Elliott and Gaffen 1991, 1993; Zhai and Eskridge 1996; Elliott et al. 1998; Wang et al. 2003; Wang and Zhang 2008; McCarthy et al. 2009). These problems, especially the temporal inhomogeneities, have severely hampered the application of radiosonde humidity data in climate studies and atmospheric reanalyses.

Partially motivated by a need to reconcile tropospheric warming trends from radiosondes, satellites, and models with available surface temperature trend estimates (Santer et al. 2008; Thorne et al. 2011), there have been many efforts devoted to homogenizing global radiosonde *temperature* data (U.S. CCSP 2006; Thorne et al. 2011 provide historical overviews). To our knowledge, however, there have been very few attempts to homogenize the global radiosonde *humidity* records, and no published studies have homogenized radiosonde humidity *daily* data. Two recent studies by Durre et al. (2009) and McCarthy et al. (2009) have adopted methods developed for homogenizing temperature data to detect and adjust radiosonde humidity *monthly* data over the Northern Hemisphere. These two studies both employed a neighbor-based difference approach, in which a homogeneous reference series of humidity is assumed to be available from neighboring station data, but with very different approaches for deriving these neighbor sets.

In reality, however, many countries have switched their sounding systems near the same time for most stations. Furthermore, similar discontinuities often occur at nearby stations and in current reanalysis humidity data even across geopolitical boundaries. In addition, the radiosonde network is very sparse, and the spatial correlation scales and therefore the potential neighbor pool are smaller for humidity variables than for temperature (McCarthy et al. 2008). There are often no neighboring stations within hundreds to thousands of kilometers, especially over the Southern Hemisphere. Further confounding such efforts, the available humidity records often contain many gaps that considerably shorten the record length for the neighbor-based difference series. Thus, the neighbor-based difference approach is impractical for homogenizing radiosonde humidity data over many regions. This issue is exacerbated for daily data whose correlation distances are much smaller and for which the impacts of synoptic variations are more pronounced than monthly data.

Unlike radiosonde temperature records, whose discontinuities primarily result from sensor-dependent biases, the discontinuities in monthly humidity time series result not only from hygrometer-dependent biases (measurement bias) (Wang et al. 2003; Sapucci et al. 2005; Wang and Zhang 2008) but also from changes in sampling (sampling bias). For example, as radiosonde hygrometers have been improved over time to have shorter response time and smaller errors under cold conditions, there has been increased sampling–reporting of cold and dry conditions at a given pressure level. Because of this, any attempts to homogenize the humidity data without first improving the homogeneity of sampling are likely to adjust the individual reported daily values farther away from the truth. Therefore, previous homogenization methods based on monthly data (e.g., Lanzante 1996; Durre et al. 2009) are not applicable to the daily data. On the other hand, homogenized daily data are needed for many applications, such as atmospheric reanalysis and analyses of extreme events and their physical causes. Current atmospheric reanalyses ingest the unadjusted radiosonde humidity and a host of other data. Combined with the varying observational constraint from the rest of the observing system over time, the discontinuities in these radiosonde data induce temporal discontinuities in reanalysis water vapor and other fields that often make them unsuitable for trend analyses (Trenberth et al. 2005; Qian et al. 2006).

Previous studies have shown that the saturation level of water vapor in the atmosphere, as measured by the RH near the surface (Dai 2006; Willett et al. 2008) and in the troposphere (Trenberth et al. 2005), has stayed fairly constant since the 1970s despite the rapidly rising air

temperature, although an apparent downward trend since around 2003 was reported in surface RH over land (Simmons et al. 2010). Physically, RH is not controlled by air temperature (Sherwood et al. 2006), unlike specific humidity (q) and precipitable water (PW). Therefore, long-term trends are likely to be much smaller in RH than in q and PW. This implies that one may not need a reference series to remove the nonstationary components in the tested time series if RH or a closely related parameter is analyzed. Furthermore, one can also avoid the need of a reference series by examining statistics other than the data time series itself. For example, by analyzing the histograms and time series of occurrence frequency of the daily data, we show that a reference series is not necessary in changepoint detection and subsequent adjustments for the radiosonde humidity data.

RH measurements by hygrometers were usually converted into DPD at each station (Elliott and Gaffen 1993), and only DPD is archived in most radiosonde databases, such as the Integrated Global Radiosonde Archive (IGRA) (Durre et al. 2006). We therefore focus on the homogenization of the individual DPD reports at 0000 and 1200 UTC (referred to as daily data in this paper). By combining them with homogenized daily air temperature data (e.g., Haimberger et al. 2008), other humidity variables (RH, q, PW) can then be derived from the adjusted DPD values. This is in contrast to the previous efforts by Durre et al. (2009) who homogenized PW, and McCarthy et al. (2009) who performed separate homogenizations for different humidity variables. For q and PW, complicated discontinuity patterns may arise from different inhomogeneities in humidity and temperature measurements, making the detection and adjustment harder (Lanzante 1996).

Homogenization of daily data often involves comparing and adjusting histograms or cumulative distribution functions (CDFs) over different periods (e.g., Della-Marta and Wanner 2006; Wang et al. 2010) under the assumption that sampling is homogenous over time. In many cases, such as measurements of light precipitation (Wang et al. 2010) and atmospheric humidity (McCarthy et al. 2009), sampling changes systematically over time. This presents a big challenge for homogenizing the radiosonde humidity data. The homogenized data may still contain systematic biases if the data series is adjusted to a reference period that contains such biases, although these biases usually have small effects on estimated long-term trends.

In the following, we first illustrate the major discontinuities in radiosonde daily DPD data using examples (section 2), and then we describe two methods to detect changepoints in the DPD data (section 3). We subsequently propose new methods to improve the homogeneity of DPD sampling (section 4) and a procedure to make adjustments (for all standard pressure levels up to 100 hPa) to remove any remaining shifts at the locations of detected changepoints (section 5). In section 6, some results are shown to illustrate the impact of the adjustments on the estimated trends for DPD, q, and RH. A summary is given in section 7. The goal of this paper is to present a new approach for homogenizing radiosonde humidity daily data for use in climate analyses and atmospheric reanalyses. Evaluations with other PW measurements, such as those from ground-based GPS (Wang et al. 2007; Wang and Zhang 2008), and analyses of long-term PW trends will be done elsewhere.

2. Known discontinuities in radiosonde humidity data

McCarthy et al. (2009) have described two major sampling problems that cause discontinuities in radiosonde humidity data. The first sampling issue is related to humidity values under dry conditions (RH < 20%) for U.S. stations, which were set to a DPD = 30°C (or RH = 19%) in radiosonde archives from about 1973 to about October 1993 (Wade 1994). This is clearly shown by strings of DPD = 300 (in units of 0.1°C) in IGRA. We found that many stations in Australia and some stations in South America, the Pacific islands, Europe, and northern Africa (where carbon hygristor was used) also show abnormally high occurrence of DPD = 300 before the early or mid-1990s in IGRA. This appears to have resulted from a practice built into certain types of radiosonde systems used in the United States and several other countries from the 1970s until the 1990s. Figure 1a shows an example of the DPD = 30°C problem at Norman, Oklahoma (taken from IGRA), which is typical for all U.S. stations. Obviously, the histogram for the 1980s is unrealistic, because it does not show any cases with DPD above about 20°C, except for the artificially assigned value of 30°C. This practice ended around early October 1993 at most U.S. stations (Wade 1994), but we found that it generally ended around 1988 in Australia when a switch from Mark II to Vaisala RS80-15 radiosondes occurred (Wang et al. 2001) and as late as 1995 in the other countries.

Clearly, these 30°C DPD values should not be used in any applications, and it is unclear whether they were used in current atmospheric reanalyses. Simply discarding them would induce a bias for the earlier records toward wetter conditions relative to more recent observations (Fig. 1). On the other hand, accepting the early DPD = 30°C reports as true observations would result in spurious downward trends in DPD since the 1970s or 1980s. In some previous analyses, a universal [e.g., RH = 16% in Ross and Elliott (1996)] or a local (e.g., McCarthy et al. 2009) constant was used to replace
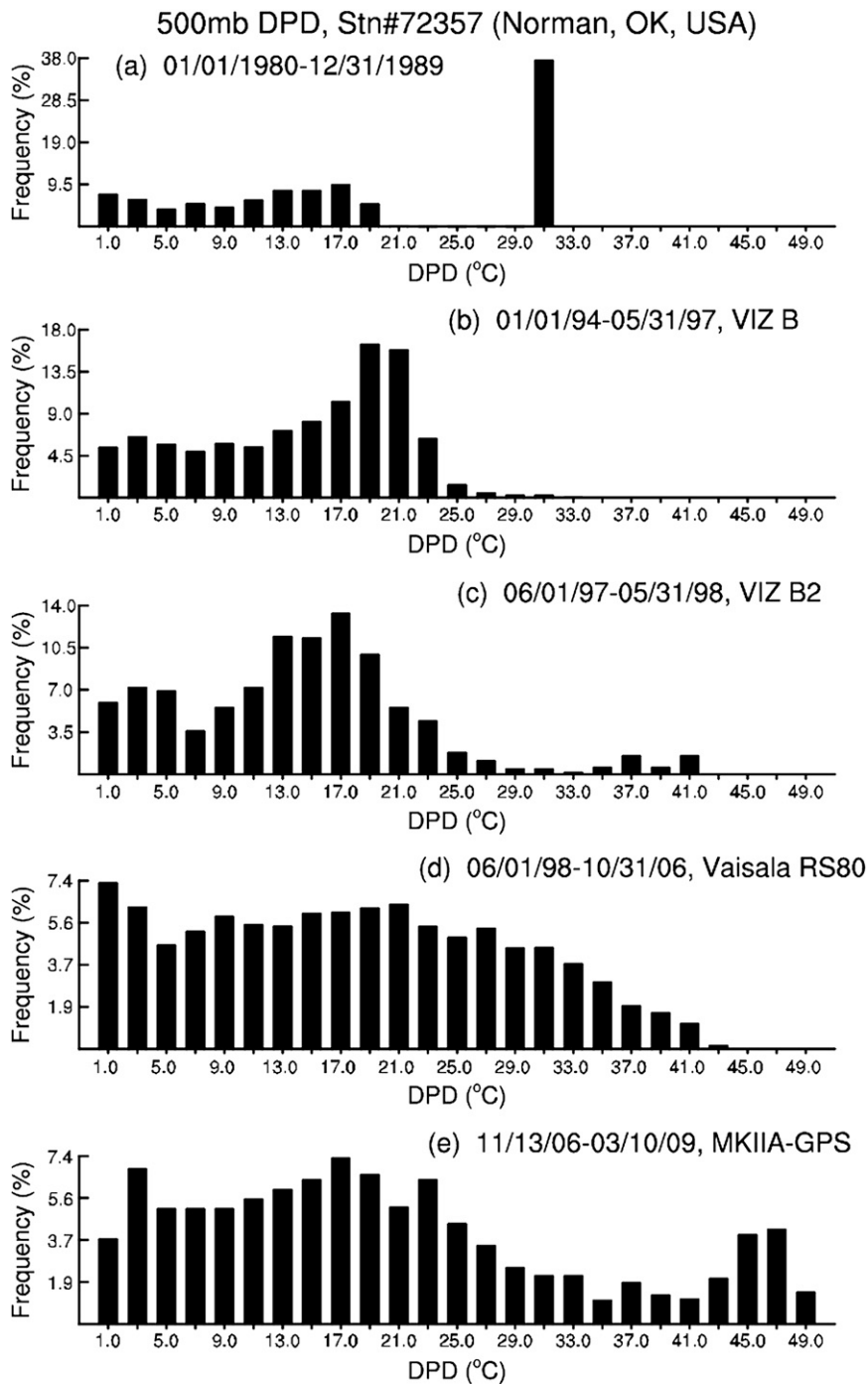
FIG. 1. Histograms of radiosonde DPD in 2°C bins, including both 0000 and 1200 UTC reports, at the 500-hPa level from a station in Norman during five different periods with different humidity sensors: (a) the 1980s, (b) January 1994–May 1997, (c) June 1997–May 1998, (d) June 1998–October 2006, and (e) November 2006–March 2009.

these RH = 19% or DPD = 30°C cases, whereas in other studies this issue was not addressed (and it is unclear whether these reports were discarded or used). An alternative solution is presented in section 4.

The second sampling issue is related to measurements under cold (ambient air temperature $T < -40°C$) conditions. Most early radiosonde hygrometers were considered unreliable under such cold conditions (Remsberg
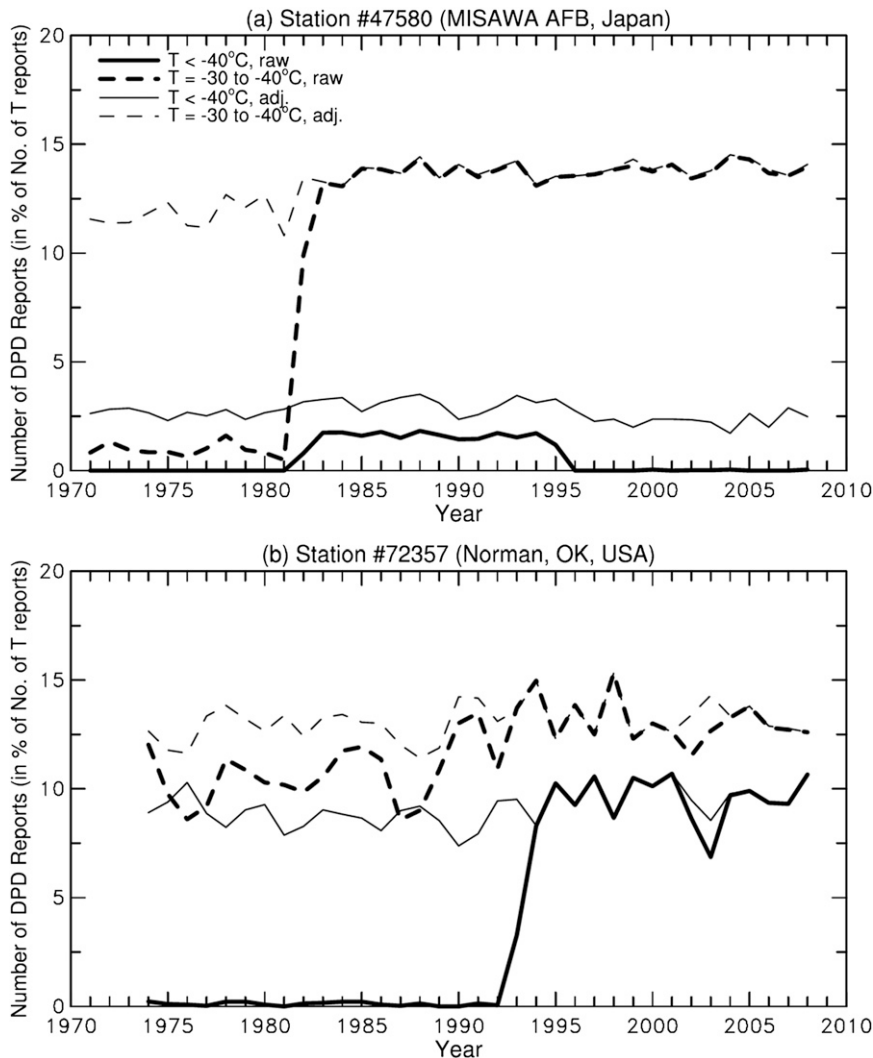
FIG. 2. Time series of annual DPD reports (expressed as a percentage of the total number of T reports) from all levels at 1000, 850, 700, 500, 400, and 300 hPa under cold conditions with $T < -40°C$ (solid lines) and $-40°C \leq T < -30°C$ (dashed lines) from the raw radiosonde data (thick lines) and the cold–dry bias-corrected DPD data (thin lines) at (a) one Japanese station and (b) one U.S. station.

et al. 2000), and it was a standard practice until 1993 for U.S. stations to report humidity measurements as missing when T was below $-40°C$. This practice may have continued until at least 2006 at more than 200 radiosonde stations in South America, Australia, India, China, and other countries, even though some modern hygrometers are able to measure RH when T is as low as $-100°C$ (Clough and Padovani 2008). To minimize the effect of this practice, McCarthy et al. (2009) first rejected all the humidity reports when $T < -40°C$ in >5% of a season's observations, and then they used the linear regression coefficients between monthly-mean T and RH, and between monthly-mean T and the natural logarithm of q in the 1995–2003 observations to estimate the mean RH and

q of the missing (or rejected) observations from the mean of the associated T for all years. Our analysis (Fig. 2) revealed that there were fewer DPD reports during the early years even for conditions with T from $-30°$ to $-40°C$. Here we made use of the anomaly relationship between instantaneous T and vapor pressure (e) in the recent records to estimate missing DPD for cold ($T < -30°C$) conditions (section 4).

Besides these sampling problems for dry and cold conditions that can cause discontinuities in monthly-mean time series, different radiosonde hygrometers have different sensitivity and response time that often lead to different measurement errors under wet (e.g., when exiting from clouds) and dry conditions even during recent years.
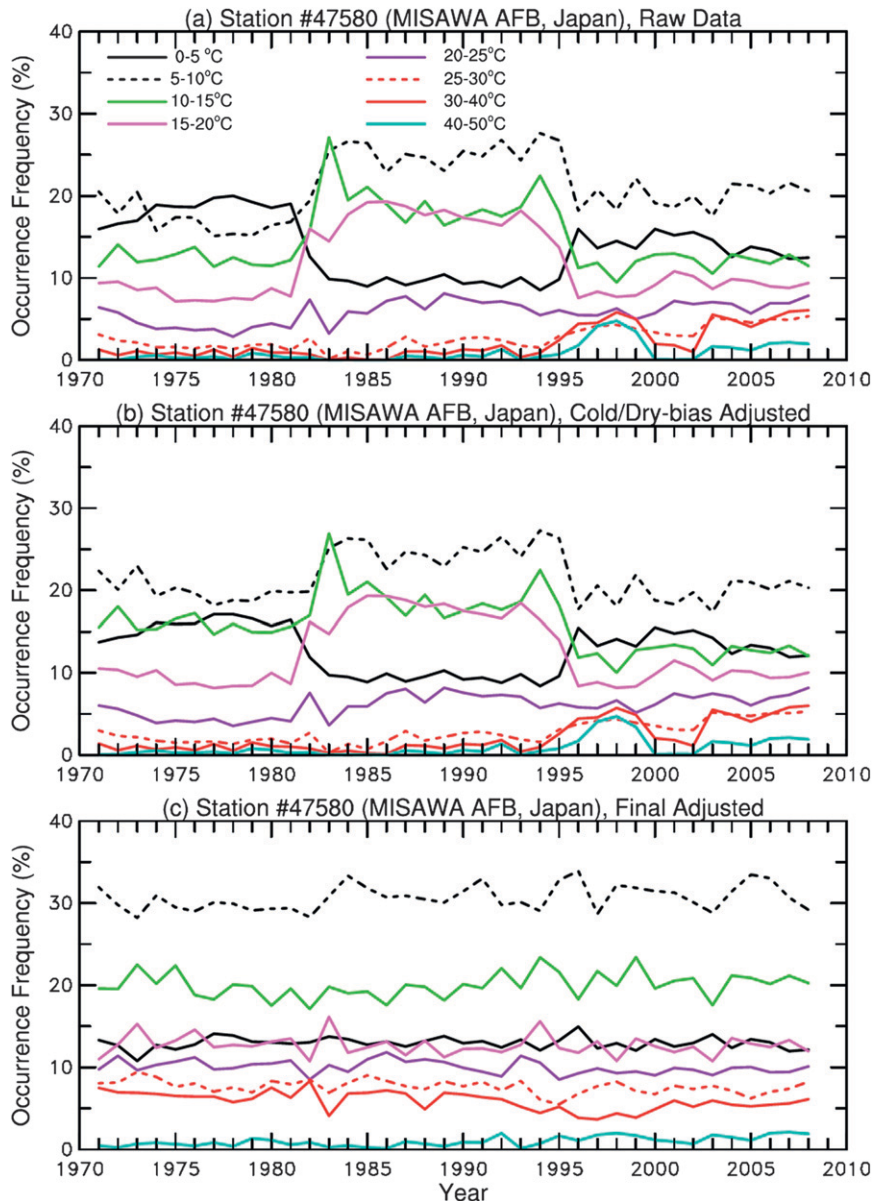
FIG. 3. Annual time series of DPD occurrence frequency as a percentage of the total number of the DPD reports from all levels at 1000, 850, 700, 500, 400, and 300 hPa for the listed bins of DPD at a Japanese station from (a) raw, (b) cold and dry bias-adjusted, and (c) final adjusted DPD data. The scale for the black line needs to be multiplied by 3.

At U.S. stations, for example, several types of radiosondes have been used since the early 1990s and the newer ones tend to have more reports of very dry (DPD ≤ 25°C) cases (Fig. 1). The end effect is that the occurrence frequency for a given bin of DPD may change because of either changes in the radiosonde types and/or practices, as discussed below. Such changes are often much larger and more abrupt than any real-world changes in DPD could conceivably be (Figs. 2 and 3).

3. Detection of changepoints

Figures 2 and 3a show that the frequencies exhibit stepwise changes not only for dry and cold conditions but also for most other categories, such as DPD = 0°–5°, 5°–10°, 10°–15°, and 15°–20°C, resulting primarily from instrumental or observational changes (Fig. 3a). The normalization by the number of temperature reports in Fig. 2 minimizes the effect of the varying total number of

soundings, and thus this frequency (referred to as sampling frequency) may be used for assessing the homogeneity of sampling. On the other hand, the frequency shown in Fig. 3 (referred to as occurrence frequency) reveals the relative homogeneity among the different DPD categories (e.g., dry versus humid) and thus can be used for identifying changes in the histograms of the daily DPD data. To make full use of the shifts seen in the histograms (Fig. 1) and occurrence frequencies (Fig. 3) in the detection of changepoints, we applied an improved variant (namely, Kuiper's statistic) of the Kolmogorov–Smirnov (K–S) test for differences in two distributions (Press et al. 1992, p. 621) to the DPD daily data and the penalized maximum F test (PMFred) of Wang (2008a,b) to the time series of monthly occurrence frequency (Fig. 3) for specified bins of DPD (see appendix A for details of the tests). To help the statistical tests locate the changepoints associated with the sampling biases discussed in section 2, we did not apply the correction for the cold and dry conditions discussed in section 4 before the changepoint detection. Although the adjustment made in section 4 may remove some of the discontinuities (mostly for DPD bins under cold and dry conditions) associated with the changepoints resulting from the sampling biases, the histogram-based adjustment done in section 5 may still be needed at these changepoints for other DPD quantiles.

Before applying the K–S and PMFred tests, we filled gaps in the IGRA twice-daily DPD data from January 1945 (but around 1973 for most stations) to March 2009 with additional data from 125 Chinese stations (T. Zhao et al. 2010, unpublished manuscript) and the Comprehensive Aerological Reference Dataset (CARDS; Eskridge et al. 1995). Fifteen stations with short records and close to other stations were excluded. Stations with fewer than 120 months of DPD records during January 1945–March 2009 were also excluded. Furthermore, records for 16 pairs of closely located stations with similar station names and matched end and start of records were merged. We ended with 881 stations (Fig. 4) to which the K–S and PMFred tests were applied. We found that surface DPD data included in the radiosonde archives often contained changepoints different from the upper-air record at many stations, because the surface data usually came from surface measurements at the station, not from radiosonde measurements, and also because of the higher sensitivity of surface observations to station relocation and local nonclimatic effects. Thus, the detection was done separately for the surface and upper-air records. Of the 881 stations, only 776 had some metadata (mostly before ~1993) as part of the IGRA dataset. Despite very substantial efforts, most notably by Gaffen (1993, 1996), these metadata are known to be grossly incomplete, especially

outside the northern midlatitudes, and they usually contain only the year of change without specific dates. The metadata also do not contain error information.

Because of these reasons, we did not use the metadata directly in the changepoint detection; instead, they were used to validate the detected changepoints. Thus, our tests were for detecting unknown changepoints, which involves searching over most of the data series to locate the data points associated with a significant shift in the mean. Testing for a mean shift at an unknown (or undocumented) location is different from testing for a mean shift at a known location because the critical value for the former is much larger than for the latter (Wang 2008a,b). This mainly affects the number of identified changepoints with small to moderate shifts, because large shifts will be identified using the critical values either for known or unknown changepoints.

In our tests, changepoints with small shifts are limited to a manageable number (less than ~15 at most stations based on our visual examination of station data series) by our choice of significance level and by how we combine the changepoints detected at different levels and for different DPD bins (see appendix A for details). We made these choices based on comparisons with available metadata after a number of experiments with different values for the significance level and other criteria. As in all homogenizations, there are large uncertainties in detecting small shifts, and we had to make a choice with regard to how many such changepoints are retained. We recognize that our choice may not be optimal for some stations. This could lead to either overadjustment if there are too many changepoints or underadjustment if there are too few of them. Improved metadata may help us in detecting and retaining the changepoints with small shifts in the future.

About 50% of the changepoints detected by the K–S test of the DPD histograms were confirmed within 12 months by the PMFred test of monthly occurrence frequency anomaly series. This number increases to about two-thirds for major changepoints detected by the K–S test (defined as those with the test statistic D being 50% above the critical value). In many cases, the K–S test was able to find the exact date at which an instrumental change was indicated by the available metadata for some U.S. stations during recent years. The changepoints from the K–S and PMFred tests were merged to produce a final list of changepoints, as described in appendix A.

About 45% of our final detected changepoints in upper-air DPD data were confirmed by the available metadata within 12 months, compared with about 41%–42% in Durre et al. (2009), and about 72% of the changes indicated by the available metadata were found to be associated with a significant discontinuity by our
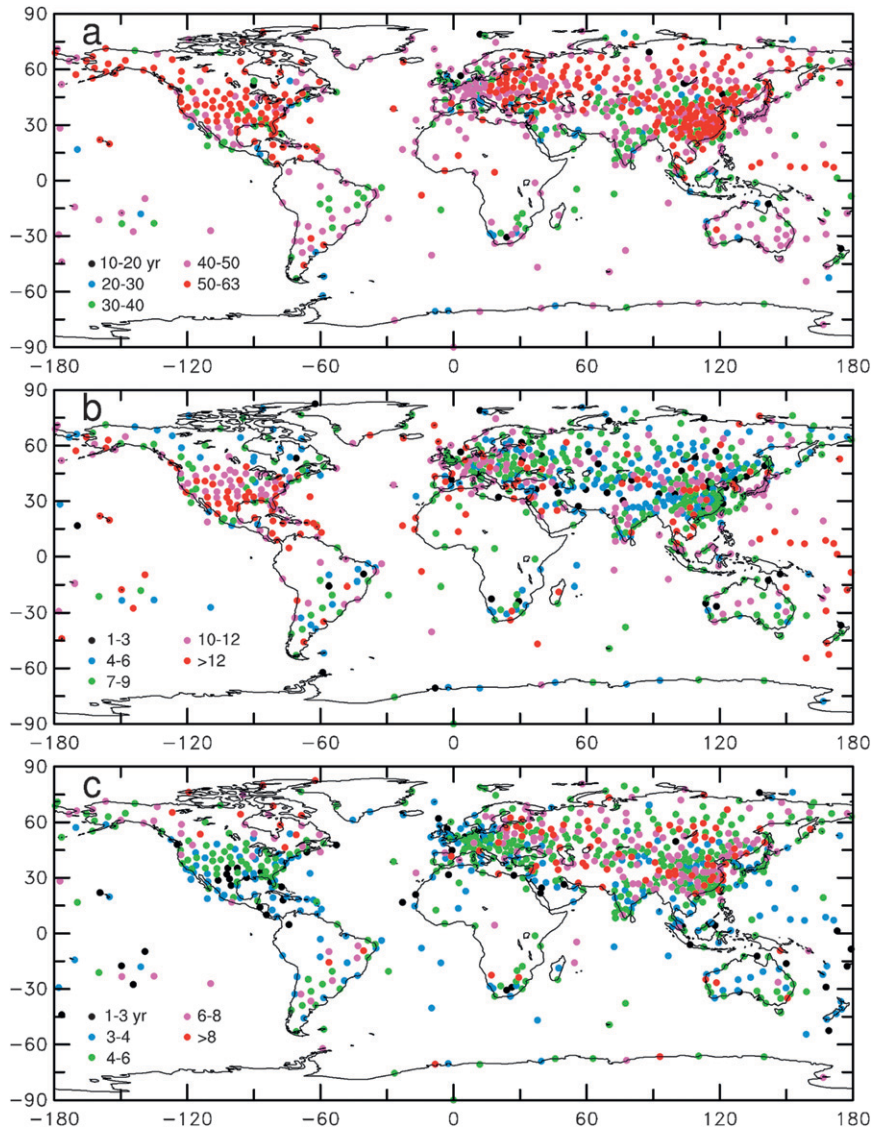
FIG. 4. (a) Radiosonde record length (years, before March 2009) for the 881 stations analyzed in this study; (b) number of changepoints in upper-air daily DPD series based on the K–S and PMFred tests; and (c) mean segment length (years) for the DPD series.

tests. Figure 4 shows the DPD record length, number of detected changepoints, and the mean segment length at each of the 881 stations analyzed. The mean segment length is relatively short (4–6 yr) over most stations in the United States and western Europe, compared to more than 6–10 yr at many stations in Asia and eastern Europe. This may reflect a tendency to upgrade the sounding systems more frequently in the United States and western Europe. On average, there were about 8.6 changepoints in each of the upper-air station records with a mean rate of approximately one change per 5 yr. This propensity for changepoint detection and the metadata congruence of identified changepoints are also broadly

similar to those documented in radiosonde temperature homogenization efforts (Thorne et al. 2005; Haimberger 2007). Visual examinations of the DPD time series from select stations confirmed that a discontinuity is likely at many of the detected changepoints.

Six such examples are shown in Figs. 5–7 for upper-air DPD series. The discontinuities at these stations are often typical for other stations in their respective countries. Available metadata generally confirmed the detected changepoints, especially for the large shifts. For example, at station 47580 (Fig. 6a), the detected changepoints generally agree with known radiosonde type or software changes, such as the change from Meisei RSII-56 to RSII-80 around
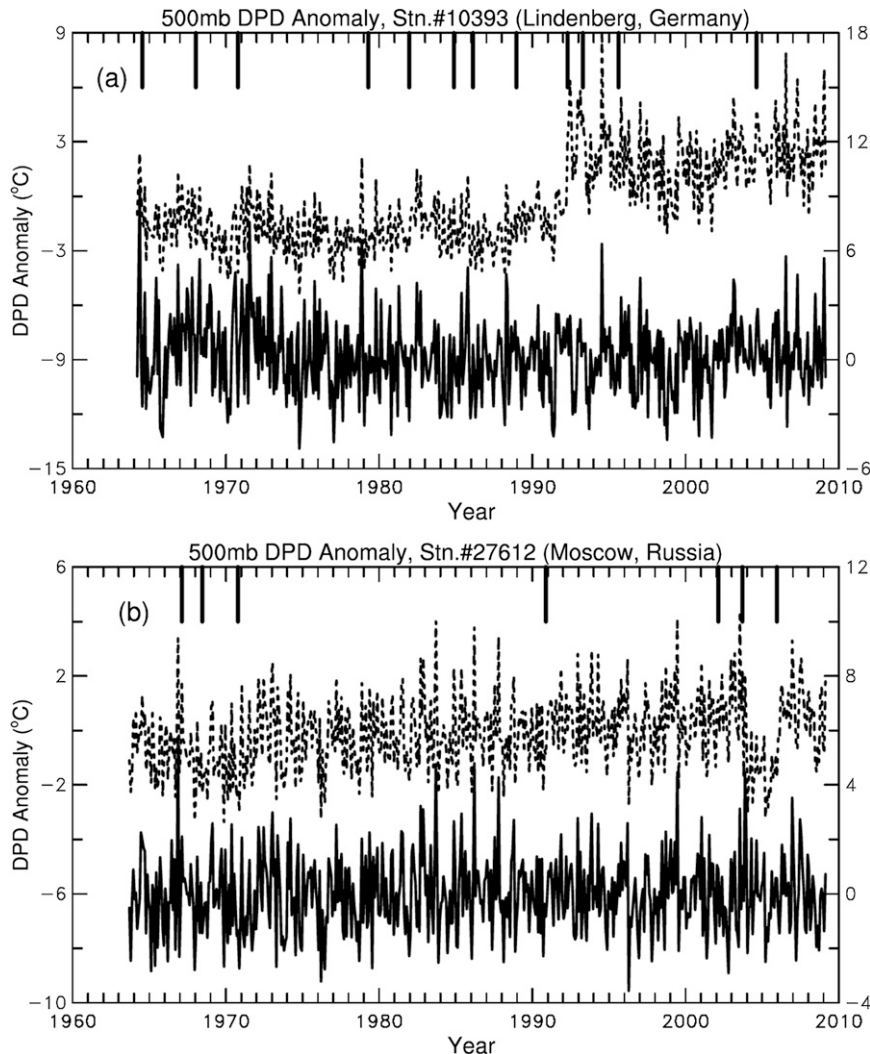
FIG. 5. Monthly anomaly time series of the raw (dashed line, left ordinate) and final adjusted (solid line, right ordinate) at 500 hPa from a station in (a) Lindenberg, Germany and (b) Moscow, Russia. The vertical bars indicate the locations of the detected changepoints. The DPD after the cold and dry bias adjustment overlays the dashed line and thus is not shown.

1982, the introduction of MEIR91 in 1996, the improvement to MEIR91 hygrometer near the end of 1999 to reduce its dry bias (Nakamura et al. 2004), and the implementation of RH wet bias correction at $T < 0°C$ started in February 2003 (Ishihara 2004). At the Norman station (Fig. 7a), VIZ OMEGA, VIZ B, Vaisala RS80, Vaisala RS80 with the sensor boom cover (Wang et al. 2002), and MkIIA radiosondes were introduced in 1986, on 24 March 1989, on 1 June 1998, in late 2000 and on 13 November 2006, respectively; and all these changes were detected by our tests to within a few months. Detailed metadata for station 94910 from R. Atkinson and S. Allen (2010, personal communication) confirmed five of the changepoints shown in Fig. 7b. For example,

the lithium chloride used in radiosondes before 25 July 1982 introduced a systematic moist bias because of its very slow response time; and the K–S test detected a shift at 31 July 1982 (Fig. 7b). Figure 7 shows that the cold and dry bias corrections had a large impact on the monthly DPD at the U.S. station before the early 1990s and at the Australian station in Wagga in the 1980s. The sampling bias corrections had little effect on the monthly series at the other stations shown in Figs. 5 and 6.

Although a change in radiosonde type or observational practice may have varying impacts on different variables (e.g., T, DPD, winds, geopotential height) in the same radiosonde record, in many cases the change
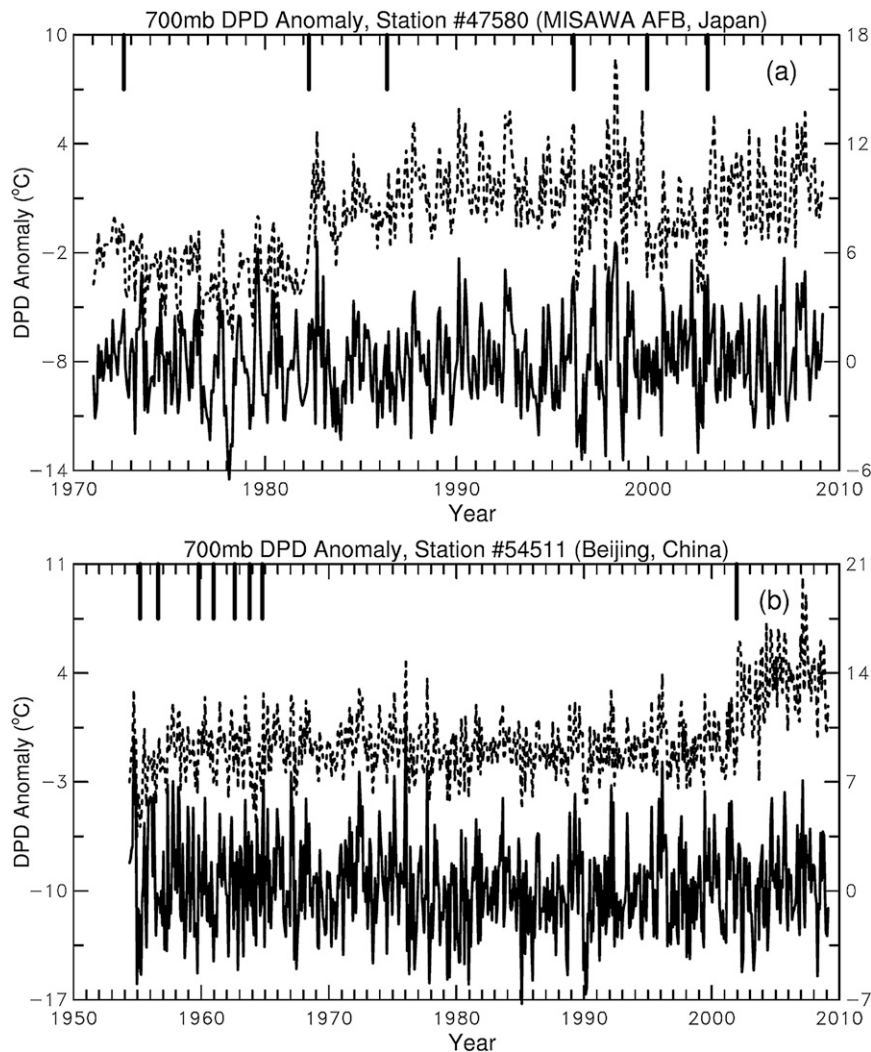
FIG. 6. As in Fig. 5, but for 700-hPa DPD from a station in (a) Japan and (b) Beijing, China. The DPD after the cold and dry bias adjustment overlays the dashed line and thus is not shown.

will likely induce discontinuities not only in the DPD series but also in some of the other variables at the same time. And because the DPD discontinuities are often more pronounced than in T and other variables, the detected changepoints from our DPD analysis may be helpful for identifying changepoints in radiosonde data series for the other variables.

4. Adjusting for artificial DPD sampling biases

After the changepoints had been detected, the homogeneity of the DPD sampling was improved by replacing the missing or bad DPD reports under the cold and dry conditions discussed in section 2 with empirically derived temperature-based estimates. These estimates used an anomaly (relative to a multiyear mean denoted by subscript "$_o$") relationship (Fig. 8) between e and homogenized T from Haimberger et al. (2008). The empirical de versus dT relationship was derived separately for dry (RH $\leq$ 20% or DPD $\geq$ 20°C) and cold ($T < -30$°C) conditions using observations made during recent years, which were 2000–08 except for U.S. stations, which used 1999–2005 because of changes of radiosonde types at many U.S. stations during 2006–08. Some stations (e.g., in Japan, cf. Fig. 2) still had no humidity observations for $T < -40$°C conditions in most recent years, in which cases the cold de versus dT relationship was based on conditions with -40°C $< T < -30$°C.

We tried various other forms of e versus T relationship besides the one shown in Fig. 8, for example, applying the natural logarithm to the y axis plus a constant; however, we failed to improve the correlation despite the apparent
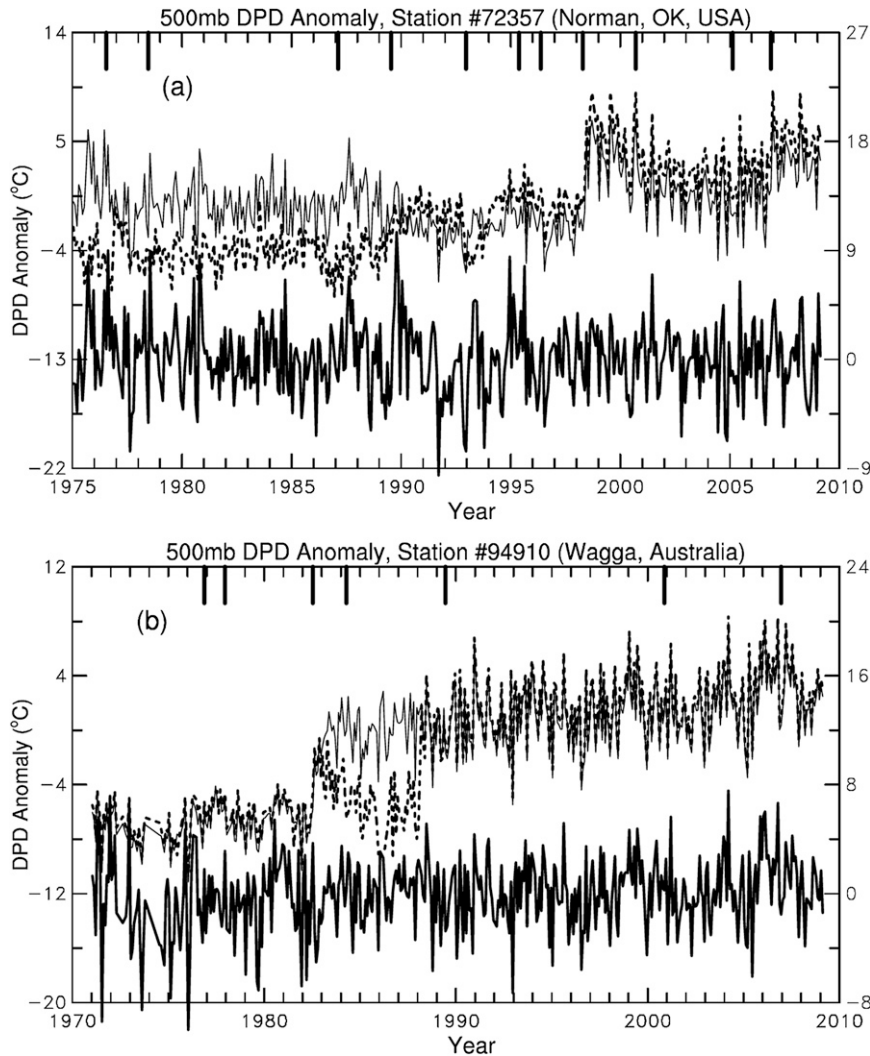
FIG. 7. As in Fig. 5, but from a station in (a) Norman and (b) Wagga. The thin solid line is the DPD after the cold and dry bias adjustment.

nonlinearity in Fig. 8. Because of this, a simple linear regression was done between de and $e_o dT/T_o^2$ at each station for both the dry and cold cases separately. This de versus $e_o dT/T_o^2$ relationship is an approximation to the Clausius–Clapeyron equation ($de_s = \text{const} \times e_s dT/T^2$) for saturation vapor pressure. Daily variations in RH contribute to the scatter shown in Fig. 8. For the majority of the 881 stations analyzed, the correlation between de and $e_o dT/T_o^2$ was between 0.4 and 0.8. For about 20% of the stations, however, the correlation was weak (below 0.2) for unknown reasons, and the regression from the closest station was used in these cases. After many tries, we used upper-air observations at or below the 300-hPa level for the recent years together in the regression analysis, with the mean e_o and T_o for each level being removed in deriving the de and dT.

The regression for dry conditions was used to estimate vapor pressure e ($=e_o + de$, in hPa) from air temperature and then DPD for 175 stations, mostly in the United States and Australia. These stations had abnormally high frequency of DPD = 30°C cases before a specific date around the late 1980s to mid-1990s. We used the following equation for vapor pressure over pure water from WMO (2008) in the regression analysis: $e = 6.112$ $\exp[17.62 \times T_d/(243.12 + T_d)]$, $T_d = T - \text{DPD}$, where T and T_d are air temperature and dewpoint temperature in degrees Celsius, respectively. For estimating our final humidity variables, however, this e estimate was multiplied by a factor of $[1.0007 + (3.46 \times P \times 10^{-6})]$, where P is air pressure in hectopascals (Buck 1981). For cases where T was below $-30°C$ and DPD was missing in all the years, the regression under the cold conditions at
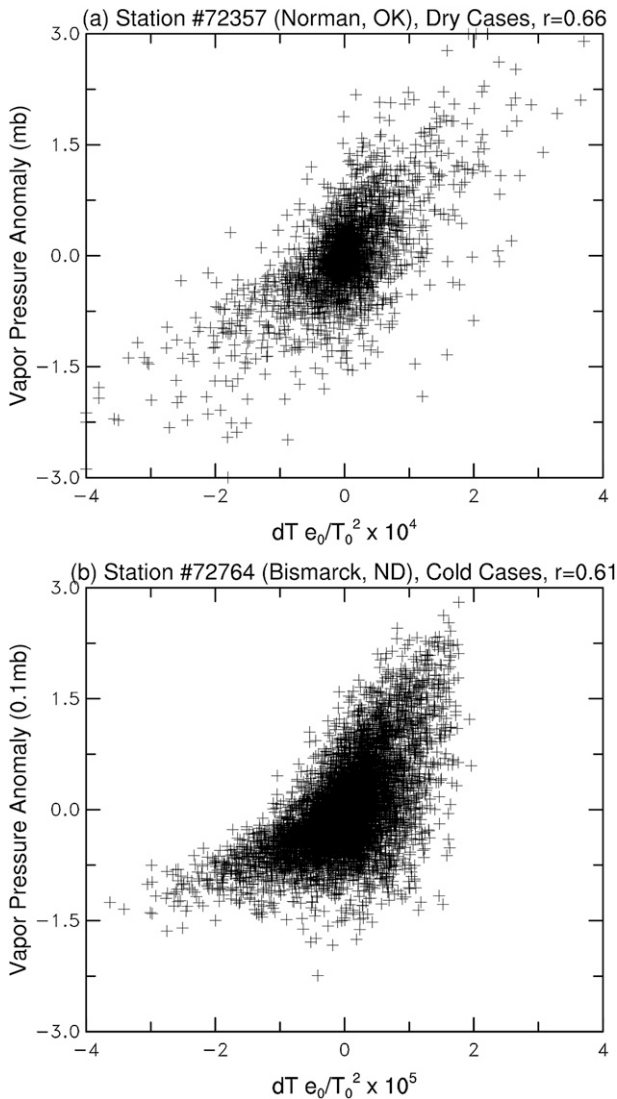
FIG. 8. Scatterplots between vapor pressure anomaly (relative to climatological mean e_o for each day and level) and air temperature anomaly (multiplied by e_o/T_o^2, where T_o is the climatological mean temperature for the same day and level) for (a) dry conditions (RH $\leq$ 20% or DPD $\geq$ 20°C) and (b) cold conditions ($T < -30°C$) at two U.S. stations. Also shown is the correlation coefficient (r) between the anomaly data.

each station was used to estimate the e and then DPD from air temperature.

Figure 2 illustrates the effect of the above adjustments for the dry and cold sampling biases at one Japanese and one U.S. station. It shows that the DPD sampling under both conditions with $T = -30°$ to $-40°C$ and $T < -40°C$ is much more homogeneous after the sampling bias correction (thin lines) than the original data (thick lines). For the Japanese station, the big jump around 1982 resulted from a documented radiosonde change, which also caused occurrence frequency changes for many DPD

categories (Fig. 3a). Even at the U.S. station, there is an apparent shift in the DPD frequency for $T = -30°$ to $-40°C$ around the early 1990s, besides the previously noticed change in DPD reporting for $T < -40°C$. We also examined DPD sampling frequency for $T > -30°C$ conditions, but we did not find large discontinuities.

Because the adjustment discussed in the next section requires that the data are homogeneously sampled—that is, not biased toward certain DPD quantiles—we need to assess the homogeneity of DPD sampling after the cold and dry bias corrections. Here we examined the time series of the annual number of DPD reports as a percentage of the total number of T reports for various DPD categories. The number of T reports below 100 hPa since the early 1970s is fairly stable at most stations. Thus, it is reasonable to assume that radiosonde sampling for T is homogeneous, that is, not biased toward certain synoptic conditions. Because the main humidity sampling problem results from the changing number of reports of dry conditions (DPD $> \sim 20°C$) over time, here we focus on the sampling for DPD below and above 20°C cases.

Figures 9 and 10 show two examples for stations 10393 and 47580. At the Lindenberg station in Germany (Figs. 9a and 9b), the DPD sampling looks homogeneous besides the two major shifts around 1992 and 2004, and the bias corrections are trivial and had little impact. At the Japanese station (Figs. 10a and 10b), the reports of DPD $< 20°C$ cases show a large increase around ~ 1982 when the reports for DPD $\geq 20°C$ do not change much (Fig. 10a). The cold bias correction removes this sampling discontinuity around 1982 (Fig. 10b). Without this bias correction, the quantile-matching (QM) algorithm used in section 5 would mistake the drop in the number of DPD $< 20°C$ cases before 1982 as a shift toward reporting higher DPD values and will try to decrease the DPD values in all DPD $\geq 20°C$ categories so that the contribution from DPD $< 20°C$ cases will be comparable to later years. In other words, it will mistakenly adjust the correct measurements to make up the missed measurements under cold conditions.

We also visually examined plots from other stations and found that the sampling bias corrections significantly improve the homogeneity of DPD sampling, and the bias-adjusted DPD appears to be homogeneously sampled. We emphasize that the DPD data after the bias corrections can be homogeneously sampled even if the frequency series after the bias corrections (Figs. 9b and 10b) still contain changes, as long as these changes offset each other (due to discontinuities in the measured data that are dealt with in the next section) or occur similarly in all DPD categories (e.g., due to random higher missing data rates for older hygrometers at U.S. and other stations).
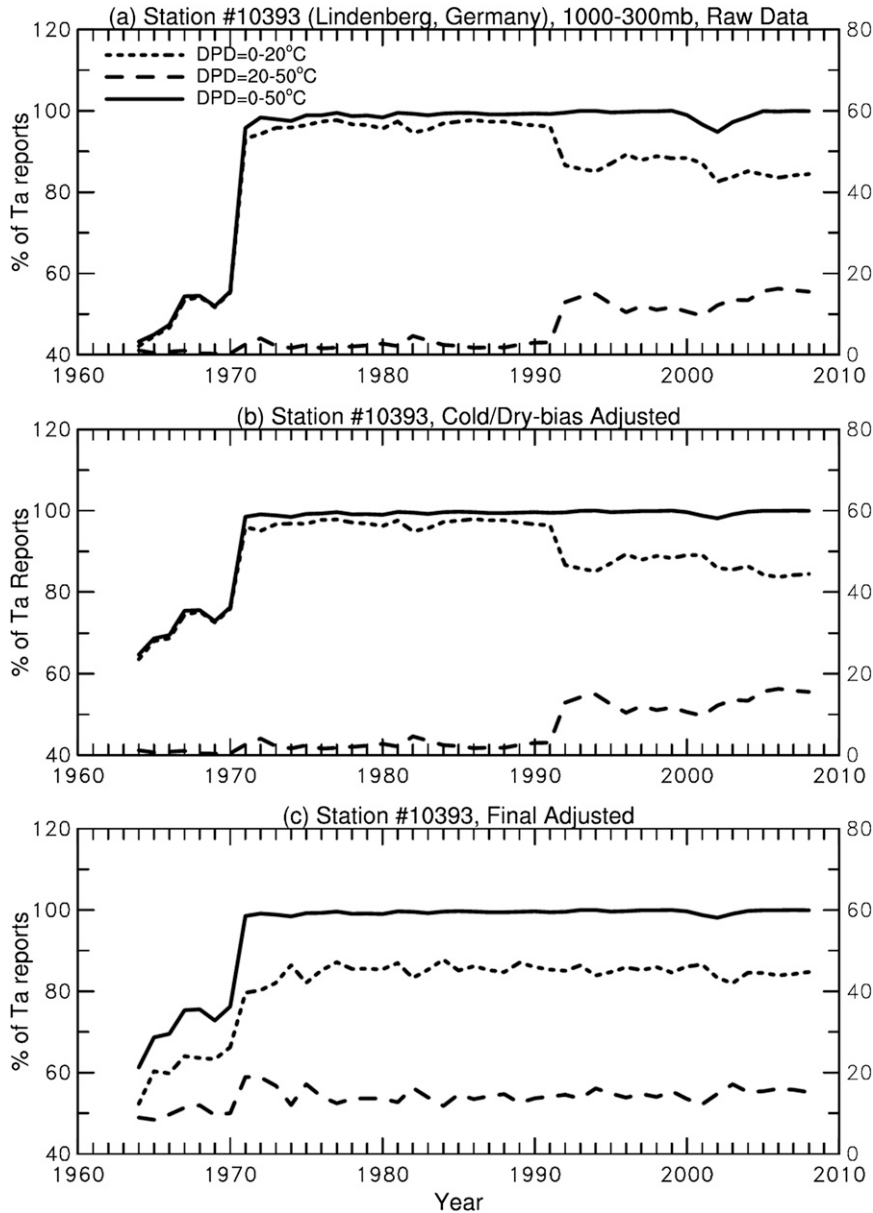
FIG. 9. Time series of annual DPD reports as a percentage of the total number of air temperature reports at the 1000-, 850-, 700-, 500-, 400-, and 300-hPa levels for DPD = 0°–20°, 20°–50°, and 0°–50°C cases in the (a) raw, (b) cold and dry bias-adjusted, and (c) final adjusted DPD data at the Lindenberg station. Note the DPD = 20°–50°C case (long-dashed line) is shown on the right-side ordinate.

5. Adjusting for other discontinuities

Changes in radiosonde types, such as that occurred at the Japanese station in 1982, can induce mean shifts in occurrence frequencies for all DPD categories because of different systematic measurement errors (Fig. 3a). Obviously, the above bias corrections for the dry and cold cases cannot remove discontinuities in all DPD categories (cf. Fig. 3a). Figures 3a, 6, 9, and 10 all show

that additional adjustments are needed to remove the discontinuities associated with the detected changepoints. To further improve the homogeneity of the daily DPD data, we adapted the quantile-matching algorithm recently developed by Wang et al. (2010), briefly described in appendix B, to adjust the daily DPD data to remove the detected shifts. The adjustments were done separately for 0000 and 1200 UTC and for each standard pressure level up to 100 hPa. DPD data at higher levels
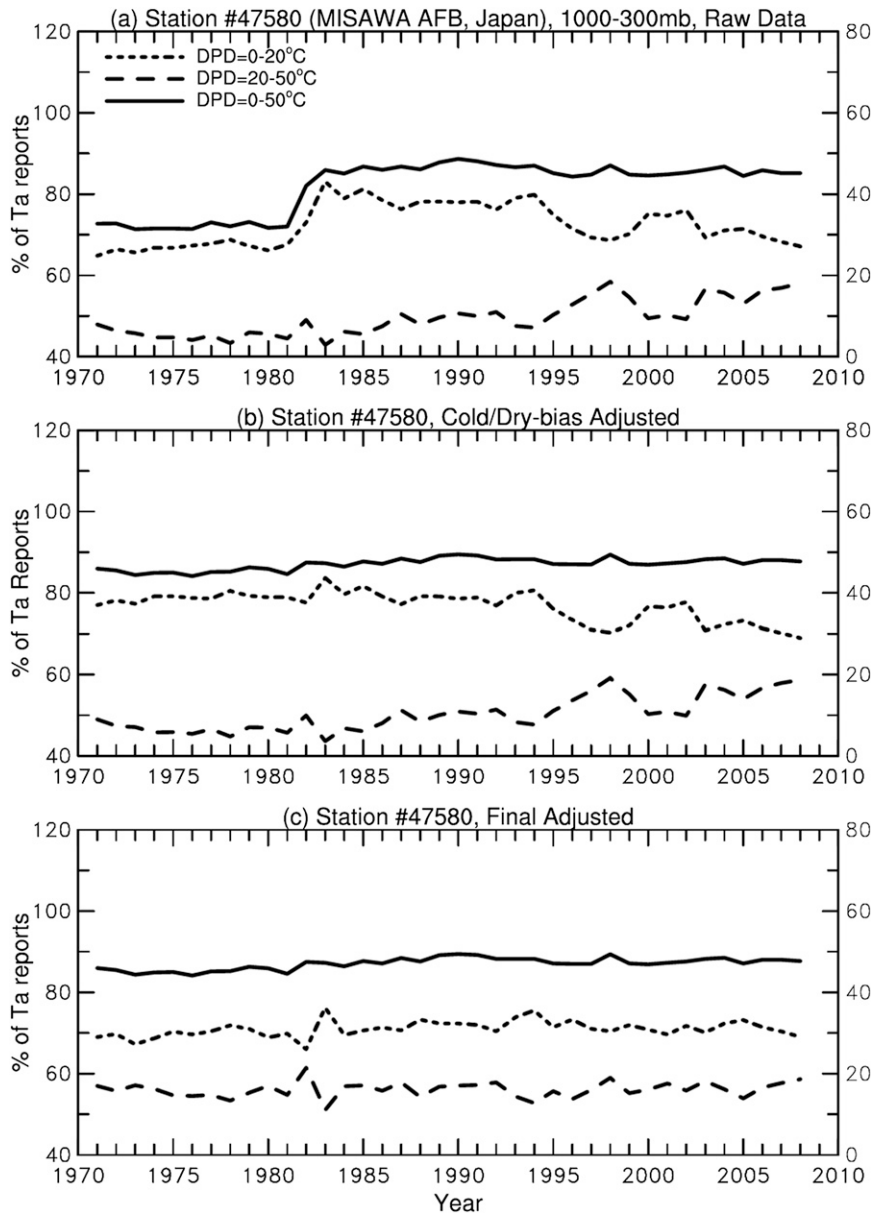
FIG. 10. As in Fig. 9, but for a Japanese station.

often contain many data gaps that prevent reliable adjustments. We used the latest segment as the reference segment based on the fact that hygrometers generally improve over time. We recognize that this may not always be the best choice because the last sonde type may not have the most realistic histogram for some stations (e.g., Fig. 1). In future studies, we plan to identify the reference segment with the most realistic histogram for each of the stations based on bias information for individual sonde types and comparisons to independent humidity measures where these are available (Wang and Zhang 2008).

Figure 11 shows that the adjustment to the individual DPD reports can be very large (10°–15°C) for the top 10 percentiles for some segments. At most stations, the adjustment is largest for the DPD reports in the top quantiles (i.e., dry conditions). This is because DPD histograms from many earlier hygrometers do not have the long tail of large DPDs seen in the recent observations, and the quantile-matching algorithm will try to adjust the top quantiles in the histograms of earlier segments to have such a tail to match that of modern observations. Figure 12 shows one example for the Japanese station. As shown by the sampling frequency series (Fig. 10),
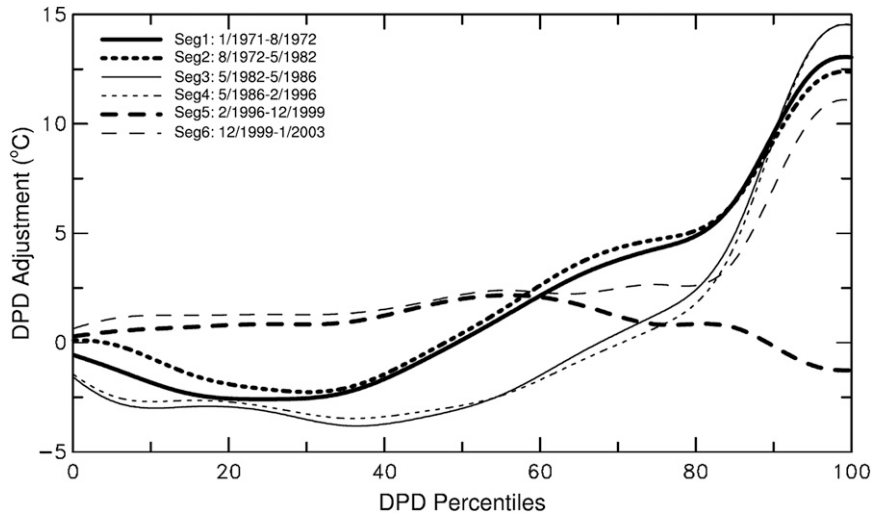
FIG. 11. The amount of DPD adjustment (based on QM) applied to the individual DPD reports at 1200 UTC for the 300-hPa level at a Japanese station (47580) for different segments (represented by different line styles) separated by detected changepoints as a function of the DPD percentiles, with larger DPD values in the higher percentiles.

the lack of the long tail of large DPDs in Fig. 12b is not due to undersampling of conditions with large DPDs. Rather it is because of sensor-dependent biases that caused overreporting of DPD < 20°C cases and underreporting of DPD ≥ 20°C cases, despite the homogeneous sampling of all conditions, as implied by the flat solid line over the data periods in Fig. 10b. The adjustments to the histograms shown in Fig. 12 act to remove the sensor-dependent biases, not to fill in any missed sampling of conditions with large DPDs using samples from other occasions. The corresponding adjustments to RH are shown in Fig. 13, which also shows the adjustment is largest for the driest and most humid conditions. After the adjustment, the DPD series exhibit more homogeneous variations (Figs. 5–7).

The above large adjustments under dry conditions are consistent with our knowledge about RH biases for old hygrometers. Many of these hygrometers have a wet bias because of their long response time and the fact that RH decreases sharply with height in the free troposphere (Held and Soden 2000). For example, the Goldbeater's skin takes 100–200 s to respond to a change in RH when T is below $-20°C$ (WMO 2008), during which time the balloon may rise by 500–1600 m with a speed of 5–8 m s^{-1} (WMO 2008). Thus, the hygrometer would report the RH of the layers several hundred meters below the nominal pressure levels. Furthermore, errors in individual RH measurements under dry (RH = 10%–60%) conditions can be as large as 6%–20% even for $T > -20°C$ (WMO 2008). Another source of error comes from the conversion of measured RH to archived

DPD at individual stations, which can introduce errors more than 1°C under dry conditions (Elliott and Gaffen 1993). Thus, we believe the adjustments shown in Figs. 11–13 are reasonable, even though they are very large for the DPDs under dry conditions for certain segments.

Because we allowed different trends for individual DPD categories in the detrending and used only part of those segments longer than 5 yr in estimating the histogram discontinuity at each changepoint, final adjusted DPDs with the trends included are not constrained to have the exact same histogram as that of the reference segment, as shown in Figs. 12 and 13. This should help prevent overadjustment to the data.

We found that the above adjustments were able to greatly improve the homogeneity of DPD occurrence (Fig. 3c) and sampling (Figs. 9c and 10c) frequencies and remove the major discontinuities associated with changes in radiosonde types (thick solid line in Figs. 5–7). With the adjustments, the histograms of daily DPD over different segments are more comparable, as expected and as shown in Figs. 12 and 13. We emphasize that the QM-based adjustment not only alters the mean for all segments except the reference segment but it also changes the DPD variability of these segment so that the variations are comparable to those of the reference segment, as shown in Fig. 6.

Despite these significant improvements, we recognize that the QM-based adjustments made in this section were based on the assumption that the CDF differences before and after the detected changepoints are entirely
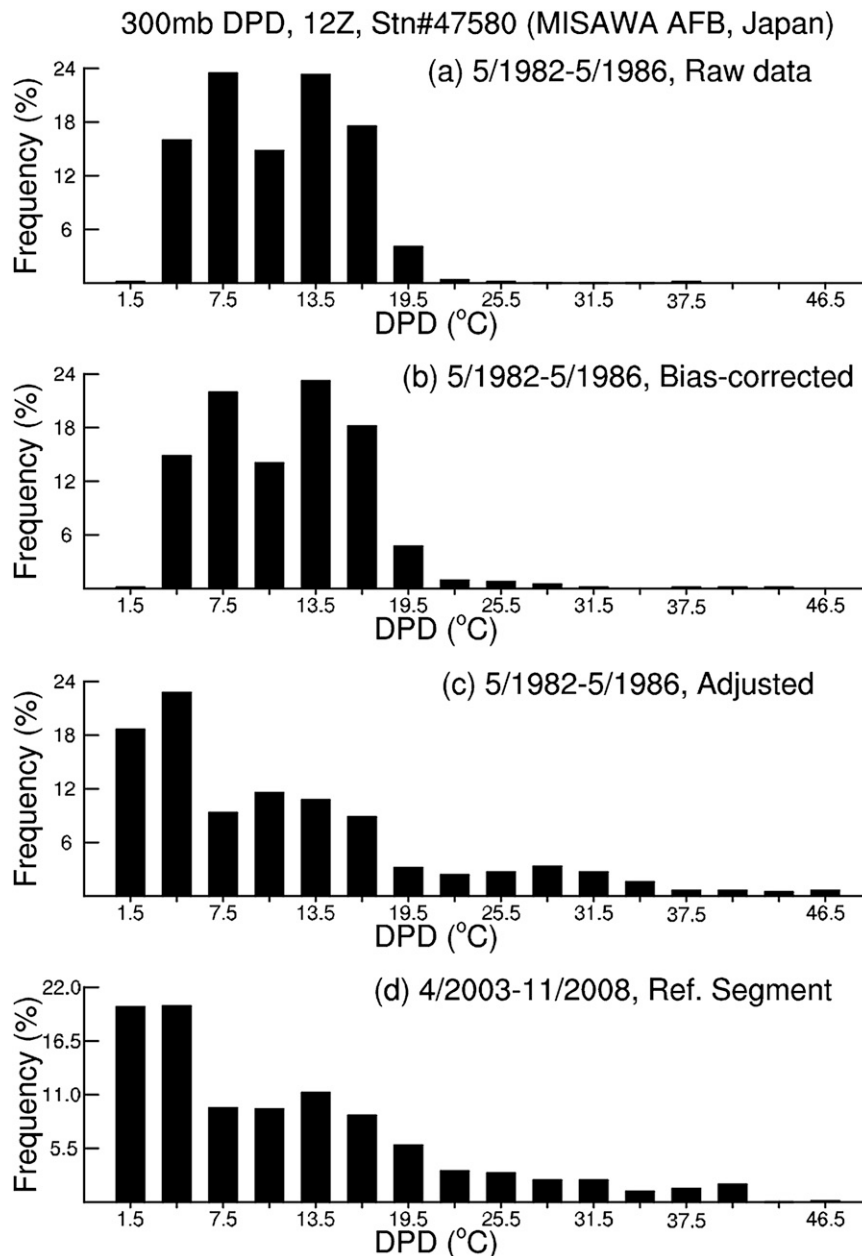
300mb DPD, 12Z, Stn#47580 (MISAWA AFB, Japan)

FIG. 12. Histograms of 300-hPa DPD for 1200 UTC at a Japanese station (47580) from May 1982 to May 1986 from the (a) raw, (b) bias-adjusted, and (c) final adjusted data. (d) Also shown is the histogram of the bias-adjusted DPD of the last segment used as the reference for the adjustment.

due to nonclimatic changes. While this is largely true in most cases for the DPD data, we cannot rule out cases where nonlinear trends and other natural variations may induce changes in the CDFs, thus violating this assumption and causing errors in the adjustments. We also cannot guarantee that the adjusted record is optimal in all cases, as the adjustment is purely statistical and automated and therefore blind to unusual cases where

limited sampling or other influences may require further fine turning.

6. Impacts on long-term trends

To assess the global impacts of our homogenization of the daily DPD data, we computed the annual trend from 1973 to 2008 in the unadjusted, cold and dry bias-corrected
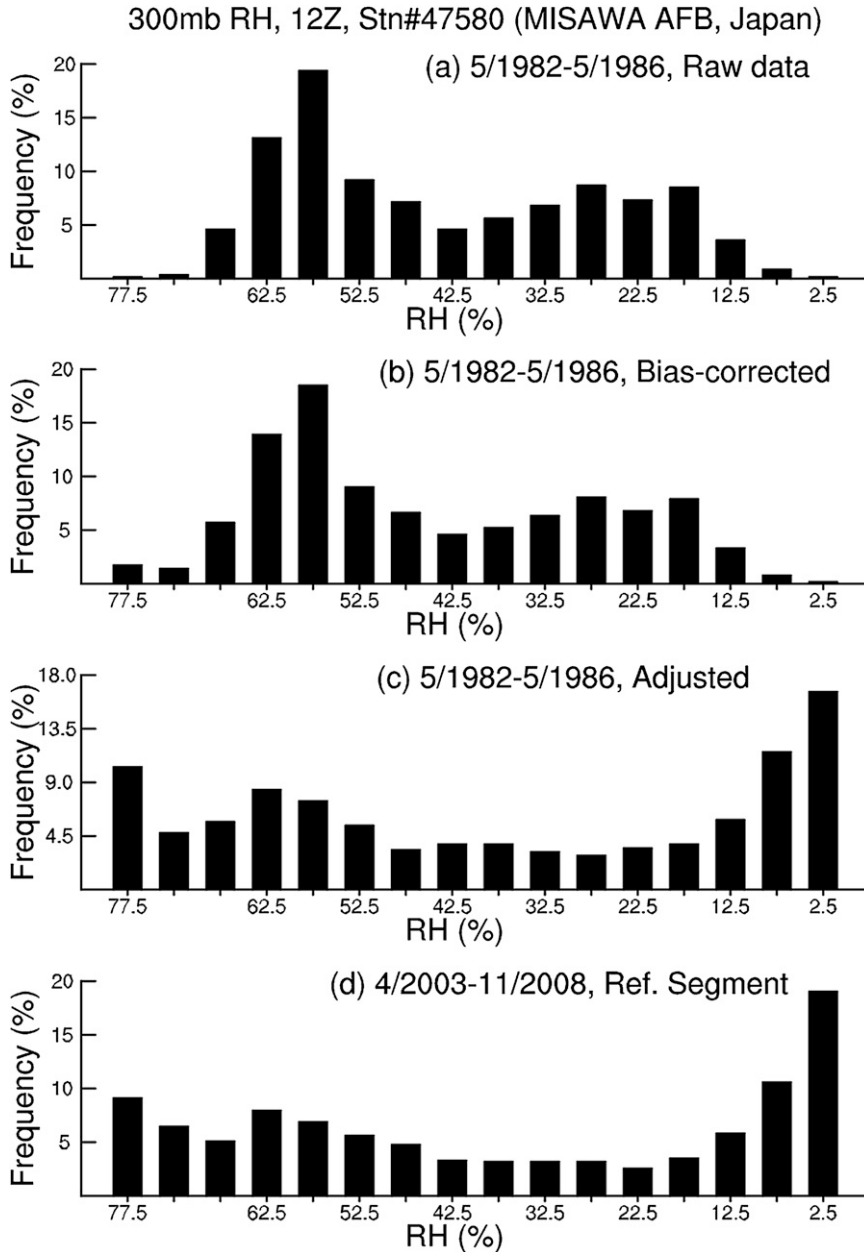
FIG. 13. As in Fig. 12, but for RH.

(section 4), and final homogenized DPD for stations with DPD monthly estimates (requiring ≥10 days of observations per month) for at least 240 months during the 1973–2008 period. We calculated the trends in two ways: (i) estimating the trend slope at each station and then mapping it onto a 2.5° × 2.5° grid for display and (ii) gridding the station DPD anomalies onto a 2.5° × 2.5° grid and then computing the trend for each grid box (requiring ≥240 months of data). We found that the trend maps from the two approaches are very similar; we use the second approach here because it also allows area-weighted regional

averaging of the DPD anomalies. We compared the estimated trends from the simple least squares fit and the median of pairwise slopes (Lanzante 1996) and did not find significant differences. Here we only show the trends from the least squares fit.

Figure 14 shows large, spatially incoherent trends in unadjusted DPD at 700 hPa during 1973–2008. In particular, large negative trends are seen over the continental United States, the Caribbean islands, Hawaii, and other Pacific islands. These downward trends result primarily from the practice of setting DPD to 30°C
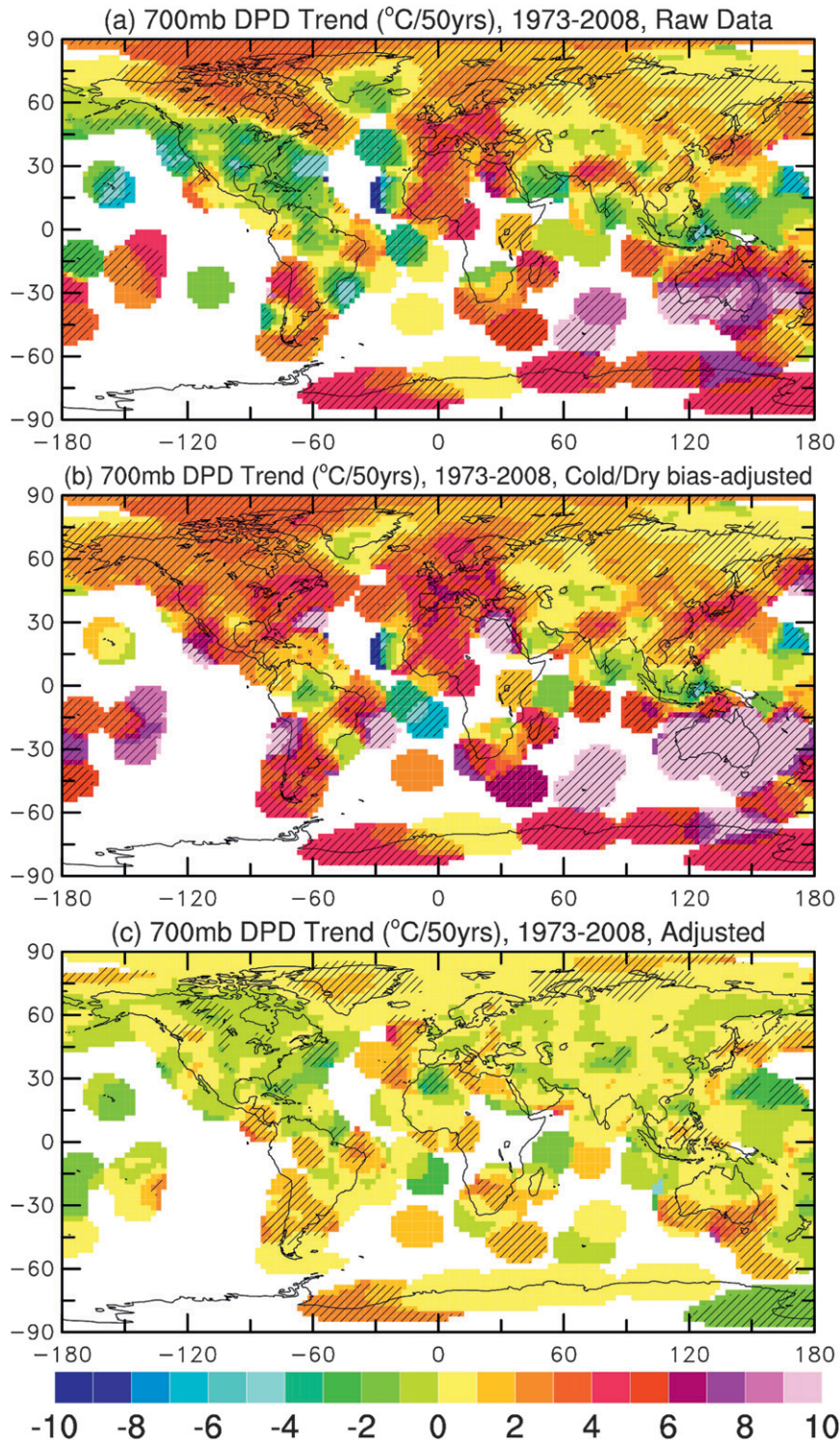
FIG. 14. Annual trends [°C (50 yr)$^{-1}$] in (a) unadjusted, (b) cold and dry bias-adjusted, and (c) final adjusted 700-hPa DPD data from 1973 to 2008. Hatching indicates the trend is statistically significant at the 5% level. A gridding method similar to Dai et al. (1997) with a search radius of 1000 km was used to create the gridded DPD data, from which the trend was computed. This approach was also used for Fig. 15. See Fig. 4a for station locations.

before around October 1993 for dry conditions with RH < 20% at U.S. and some other stations (section 2 and Fig. 1). After the dry and cold bias corrections of section 4, most of the negative trends disappear, with trends over the United States becoming strongly positive as over most other land areas (Fig. 14b). The positive trends are expected because there is an upward trend in the sampling of dry (DPD ≥ 20°C) cases because of improved sensitivity of radiosonde hygrometers to low humidity over the years. This time-varying bias is clearly shown in the DPD histograms (e.g., Figs. 1, 12, and 13) and the frequency time series for large DPD values (e.g., Figs. 3, 9, and 10). In particular, Australian stations had major sonde changes around 1982/83 that resulted in sharp increases in monthly-mean DPD around this time (cf. Fig. 7b) and contributed to the large upward trends over Australia shown in Fig. 14b. Thus, most of the trends shown in Fig. 14b are still not real. After the additional adjustments of section 5, the DPD trends become much smaller [within $\pm 2°C$ (50 yr)$^{-1}$ or 0.4°C (10 yr)$^{-1}$] and statistically insignificant for more than 75% of the sampled locations (Fig. 14c), compared with 32% and 23% for Figs. 14a and 14b, respectively. Further evaluations suggest that the adjusted DPD produces q and PW changes that are more comparable with other independent measurements [e.g., PW from GPS measurements (Wang et al. 2007); T. Zhao et al. 2010, unpublished manuscript; not shown here] and more coherent spatially. This suggests that the adjusted DPD is likely to be more realistic than the unadjusted data, corroborating the known biases associated with changes in sonde types.

The effect on the trend in the derived q from the DPD adjustments is also very large over most of the sampled locations (Fig. 15). The cold and dry bias corrections remove most of the upward trend in q over the continental United States but enhance the downward trend over Australia. The large spatial variations in the q trend in both the unadjusted and cold and dry bias-adjusted q (Figs. 15a and 15b) look spurious. This is because both tropospheric temperature and water vapor content are determined by large-scale atmospheric circulation (Sherwood et al. 2006), and thus their changes should be spatially coherent, as shown by satellite observations of PW (Wentz and Schabel 2000; Trenberth et al. 2005) and temperature (Haimberger et al. 2008). Our adjustments greatly improve the spatial coherence for the q trend (Fig. 15c), as is usually the case for the homogenization of radiosonde temperature (Haimberger et al. 2008). This is also true for PW trends (not shown).

For comparison, Fig. 15d also shows the 700-hPa q trend from 1973 to 2008 derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalyses (ERA; Uppala et al. 2005; Simmons et al. 2007). ERA-Interim used the homogenized radiosonde temperature from Haimberger (2007), while the 40-yr ECMWF Re-Analysis (ERA-40), like most other reanalyses, used the unadjusted radiosonde data. Unhomogenized radiosonde humidity data were used in the ERA-40 and all other atmospheric reanalyses. The ERA-40 q trend patterns over land resemble those of Fig. 15b more than Fig. 15c, and they differ substantially from our adjusted q trends, for example, over northern mid-high latitudes. The large negative trends over the low-latitude land areas in the ERA-40 data also appear to be unrealistic. This suggests that water vapor trends in atmospheric reanalyses are significantly affected by inhomogeneities in radiosonde humidity data and thus cannot be trusted. The large positive trends over the data-sparse oceans in ERA-40 are also unreliable (Trenberth et al. 2005). Use of homogenized radiosonde data could improve the homogeneity of future reanalysis data: this was a primary aim of this research.

Trend maps for RH (not shown) are very similar to those for DPD (Fig. 14), with changes in tropospheric relative humidity from 1973 to 2008 within a few percentage points per 50 yr. More detailed analyses of the long-term changes in RH, q, and PW will be presented in a separate paper.

7. Summary and concluding remarks

Motivated by the need for homogenous radiosonde humidity data in climate analyses and atmospheric reanalyses, we have developed a new approach to homogenize the twice-daily radiosonde reports of dewpoint depression (DPD) from 881 stations included in the IGRA dataset with data gaps filled with data from two other datasets. Primarily because of constant improvements in radiosonde hygrometers, especially in their sensitivity to low relative humidity, response time, and accuracy under cold conditions, large discontinuities exist in the occurrence frequency and histograms of the daily DPD. These discontinuities cause large mean shifts and spatially incoherent trends in monthly time series of DPD and the derived RH, q, and PW. However, the monthly-mean shifts are not suitable for adjusting the daily DPD values because the mean shifts result largely from inhomogeneous sampling of dry and cold conditions and from DPD-dependent biases in individual measurements.

Here we first applied an improved variant of the K–S test to the histograms of daily DPD and the PMFred test to the monthly series of the occurrence frequency for specified bins of DPD. The two tests agreed within 12 months on the timing of most major discontinuities, and each of the merged changepoints was probably associated with a discontinuity according to our visual
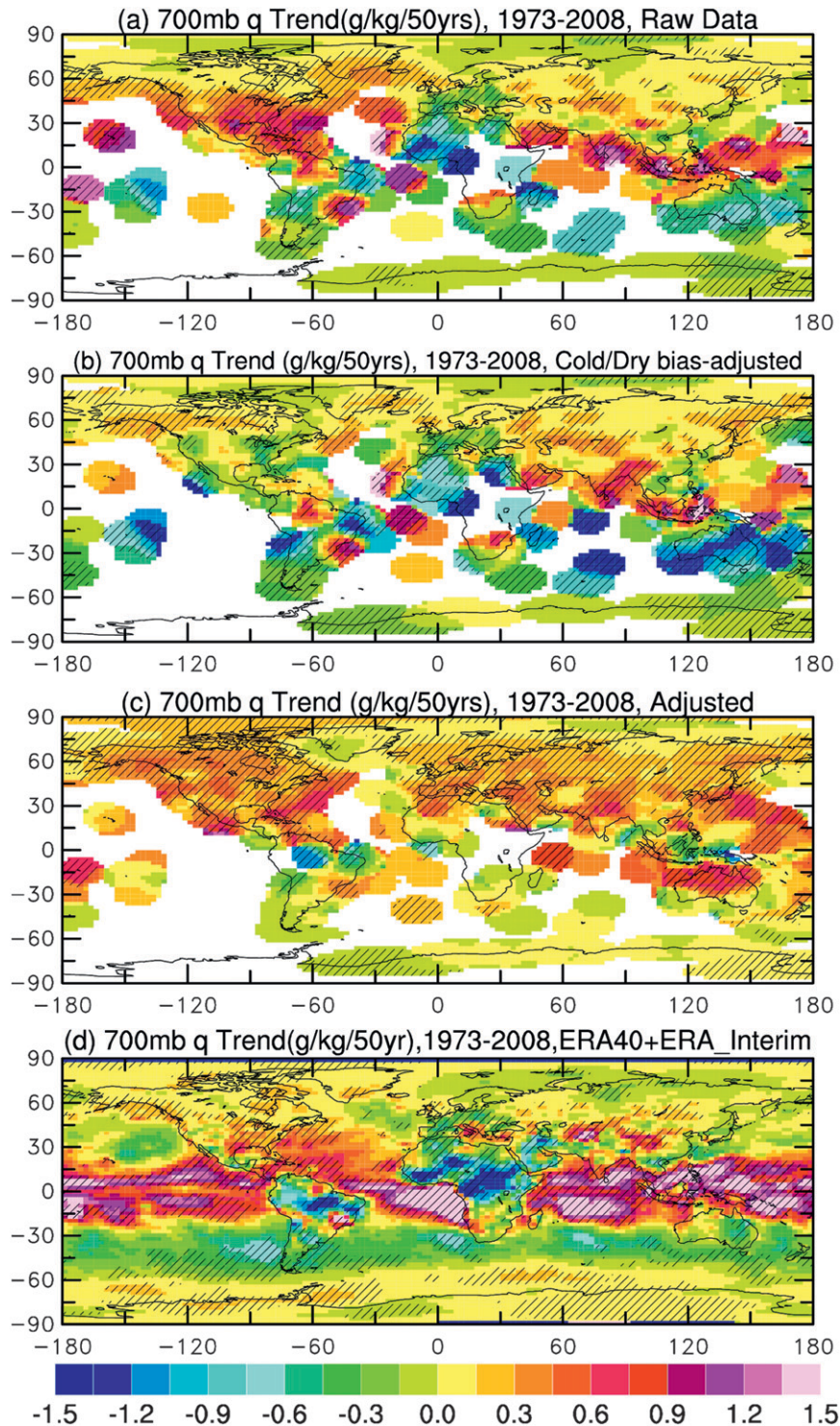
FIG. 15. Annual trends in 700-hPa specific humidity $[\text{g kg}^{-1} (50 \text{ yr})^{-1}]$ derived from (a) unadjusted T and DPD, (b) adjusted T (from Haimberger et al. 2008's RICH v1.4) and cold and dry bias-corrected DPD, (c) adjusted T and final adjusted DPD, and (d) ERA-40 (for 1973–88) and ERA-Interim (for 1989–2008). The ERA-40 data were adjusted by M2 − M1, where M1 and M2 are the ERA-40 and ERA-interim means, respectively, for their overlap period from 1989 to 2001. Hatching indicates the trend is statistically significant at the 5% level.

examination of selected cases and comparison with available metadata. About 45% of the detected changepoints were confirmed by the available and grossly incomplete metadata within 12 months, and about 72% of the changes documented by the metadata (some of which were not expected to cause a discontinuity) were found to be associated with a significant shift by our tests. On average, there were about 8.6 changepoints in each of the analyzed DPD records (~1 changepoint per 5 yr), most of which started in the early 1970s and all ended in March 2009.

After detecting the changepoints, we next improved the DPD sampling for known reporting practice issues by estimating DPD values for the cold cases where DPD was missing and air temperature was below $-30°C$ for all stations and for all years, and for the dry cases where DPD had been set to 30°C for all U.S.-operated stations and some other stations with abnormally high incidence of DPD = 30°C before around the early 1990s. We used local anomaly relationships, similar to the Clausius–Clapeyron equation, between (homogenized) air temperature (T) and vapor pressure derived from radiosonde observations for recent years for cold and dry conditions to estimate DPD from T. The bias-adjusted DPD data show homogeneous sampling in comparisons with radiosonde temperature reports.

We then adapted the quantile-matching algorithm of Wang et al. (2010) to adjust the histograms of the cold and dry bias-adjusted DPD, so that they become comparable (but not identical) to that of a reference segment, which was chosen to be the latest one. This choice was based on the assumption that measurements made by newer hygrometers are generally, but not always, more accurate than older technologies and practices. For Wang et al.'s algorithm to work well for the DPD data, a significant change was made in detrending the data: we allowed each DPD category to have its own trend instead of a common trend for all categories, as in Wang et al. (2010). This change greatly improves the trends in the homogenized data based on comparison with the underlying trends in the original data and the enhanced spatial coherency of the trend. It further allows the individual segments to have slightly different histograms after the adjustments. The adjustment was done separately for 0000 and 1200 UTC observation times and for each standard pressure level up to 100 hPa.

These adjustments remove most of the discontinuities in the time series of occurrence frequency and monthly-mean DPD. They also modify the histograms of the daily DPD for the earlier decades to be comparable to most recent observations. The adjusted DPD shows small [within $±2°C (50 \text{ yr})^{-1}$] and mostly insignificant trends from 1973 to 2008 over the continents and many islands, in contrast to spatially incoherent and obviously spurious large trends [up to $±10°C (50 \text{ yr})^{-1}$] in the unadjusted data. The small DPD trends indicate that middle to lower tropospheric relative humidity has been relatively stable during 1973–2008. The adjusted DPD data at each standard pressure level (up to 100 hPa) were used to derive specific humidity and precipitable water in the troposphere, in combination with homogenized radiosonde temperature based on Haimberger et al. [2008, Radiosonde Innovation Composite Homogenization version 1.4 (RICH v1.4)]. The trends in these derived variables are spatially more coherent and thus likely to be more realistic after the homogenization than in the unadjusted data.

There are two unique features in our homogenization approach: first, it removes major discontinuities not only in the monthly-mean time series but also in the statistical distributions of the daily radiosonde data, which has not been done previously; second, it does not require a reference or background series to remove the natural variations and real trends in the data series to detect changepoints and estimate the amount of adjustment. Instead, synoptic and other natural variations and long-term trends were allowed in the DPD data. We avoided the use of a reference series (often unavailable) by analyzing the histograms and occurrence frequency series instead of the data series itself, in contrast to most previous homogenization efforts (e.g., Haimberger 2007; Durre et al. 2009; McCarthy et al. 2009).

We emphasize that, as in all statistical homogenizations, the adjustments presented here were aimed to remove the overall biases in distributions of the DPD data (not just shifts in the mean). Adjustments to individual DPD values may contain large errors, but overall the adjusted DPDs are more realistic and represent climate variations and long-term changes much better than the raw data. This applies to any homogenization efforts, including those using a reference or background series, because the mean adjustment estimated from a difference series can be applied only in a statistical sense, and it may not be an accurate adjustment for individual data points.

Improvements can be made to our homogenization with more detailed and more complete metadata, better use of the metadata in the changepoint detection, further improvements to the sampling of the DPD data, better detrending and removal of natural variations in the quantile-matching-based adjustments, and quantitative bias information for specific sonde types, such as those presented in Wang and Zhang (2008). With the additional information, we may be able to identify a segment with the most realistic DPD distribution, remove any biases in the DPD data over this segment, and then use it as the reference segment for the quantile-based adjustment for each station instead of using the last segment at every station.

Our homogenized-daily DPD data, together with adjusted air temperature (based on Haimberger et al. 2008) and derived q and RH and other raw sounding data, at standard pressure levels for all 881 stations will be available freely online (at www.cgd.ucar.edu/cas/catalog/). We believe that use of this dataset in climate analyses and atmospheric reanalyses should improve the homogeneity of the upper-air temperature and humidity fields and thus long-term trends in these and other related fields.

APPENDIX A

Details of the Detection of Changepoints in Twice-Daily DPD Data

a. K–S test

As in most statistical tests, the K–S test is usually applied to two given samples for testing whether the samples have similar distributions. However, for detecting unknown changepoints, a search over the entire time series (except some endpoints) is needed to locate the unknown changepoint that produces the largest probability for the two samples separated by this changepoint to have different distributions based on the K–S test. Because of this, the critical value for a given significance level for our K–S test is larger than that used in the standard K–S test applied to two fixed samples as in Press et al. (1992). We had to use a large number of Monte Carlo simulations to generate the empirical critical values (CVs) for our K–S test (see below), as was done for the PMFred test (Wang 2008b). This is an important but often overlooked difference between testing for unknown (or undocumented) and known (or documented) changepoints using any statistical test.

We used Monte Carlo simulations to estimate the CVs for detecting unknown changepoints using Kuiper's statistic, an improved variant of the K–S test (Press et al. 1992, p. 621). For each case, 200 000 simulations were used based on results from test runs that showed very similar CVs using either 100 000 or 200 000 simulations. Synthetic random time series were generated from a Gaussian process $N(0, 1)$ with zero mean and unit standard deviation. A first-order autoregressive model (AR1) was used to create autocorrelated time series from the $N(0, 1)$ random-number generator. Because the K–S test statistic D is based on the maximum difference between the CDFs of the two tested samples and D's distribution is independent of the probability distributions of the sample data (Press et al. 1992), the CVs derived using the Gaussian process is representative for other types of data. This was confirmed in our experiments where the data points were randomly sampled through simple or block bootstrapping of real DPD data with non-Gaussian distributions (cf. Fig. 12d). The block bootstrapping randomly samples a consecutive segment (e.g., five data points) each time, so the autocorrelation within the original data is roughly preserved.

As in the test of the DPD data, we searched only over the middle third of a given time series. In general, the search has been done over a wider section (Wang 2008a,b); however, for our daily DPD data, we had sufficient data to search only the middle year of a moving 3-yr data window (see below) because we required a minimum of 300 data points on either side of the tested point. For example, for a time series with 900 points ($N = 900$), the search for a changepoint was done only over the 301st–600th data points. Because we applied the K–S test each time to data points over a short (3 yr) period, changes induced by long-term trends are likely to be small compared with any nonclimatic shift within the same period.

The values of the test statistic D from the 200 000 tries for each given combination of N and r_1 (lag-1 autocorrelation) were ranked and the values for the 80th, 90th, 95th, 99th, 99.9th and a few other select quantiles were saved as the CVs for use in the K–S test for unknown changepoints. Figure A1 shows that the CVs decrease rapidly as the sample size N increases to around 900; thereafter, the decreasing rate levels off as one would expect. The CVs are also higher for autocorrelated time series and for smaller α values for a given N. We emphasize that the CVs may be different if the search is done differently rather than over the middle third of the series. For example, in the extreme case, the search can be only over the middle point of the series (i.e., testing for distribution difference in two given samples), and the CVs would follow D's analytical distribution presented in Press et al. (1992).

We applied the K–S test to a moving, 3-yr window of twice-daily DPD anomaly series at each station, with the long-term mean for each day of the year and each observation time being removed. The search was done only over the middle year, then the data window moved
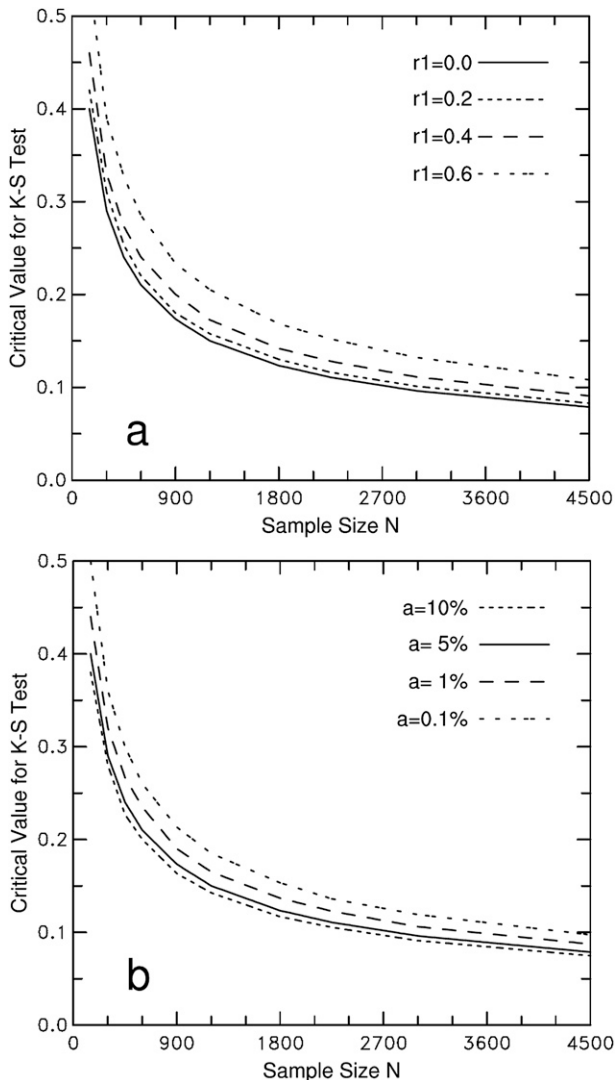
FIG. A1. Critical values determined from Monte Carlo simulations with 200 000 tries for each case as a function of both sample size N and (a) lag-1 autocorrelation (r_1) with a fixed significance level of 5%, and (b) significance level (α) with no autocorrelation for the K–S test (Kuiper's statistic; Press et al. 1992, p. 621). It is assumed that the search for the changepoint is done over the middle third of the time series.

forward by one year. The data window was extended backward (forward) if the first (second) segment in the data window contained fewer than 300 data points. The test was applied separately to surface and 850- and 700-hPa levels, which had the fewest missing data. Upper levels often contain too many data gaps in the earlier years, making the K–S test ineffective. The changepoints from the 850- and 700-hPa levels were merged as follows: any two changepoints within 300 days were considered as the same shift and only the one with the larger test statistic D was kept. We used this combination based on

our observations that changepoints may become obvious at different levels, as the effect of an observational change (e.g., relocation) varies with height. We used the CVs for the 99.9th quantile in our K–S tests to obtain a manageable number of changepoints (e.g., fewer than 15 for most stations) that were roughly comparable to those from the PMFred test. Visual examinations revealed that the changepoints detected by the K–S test captured all the *obvious* shifts in the DPD monthly time series.

b. Application of the PMFred test

The PMFred test of Wang (2008b) was used to detect mean shifts in the time series of monthly-mean occurrence frequency for individual bins of DPD (cf. Fig. 3). Starting from the whole series, the PMFred test finds the most probable (i.e., with the largest probability) and significant changepoint (if exists) that divides the whole series into two segments, it finds the most probable and significant changepoint within each of these two segments, and then it reevaluates the significance for each of the detected changepoints given the existence of the other changepoints and removes any insignificant ones. This process continues until no new changepoints are found or the segments reach a preset minimum length. The PMFred test allows a long-term trend and autocorrelation in the time series and thus a reference series is not required. The occurrence frequency was calculated using the unadjusted DPD data separately for the surface and upper-air levels. Reports from the 1000-, 850-, 700-, 500-, 400-, and 300-hPa levels were combined in computing the upper-air frequency to increase the sample size. The frequency time series for the following DPD bins (in °C, "[" for inclusive) were tested separately with a nominal α value of 5%: [0, 1), [1, 5), [5, 10), [10, 15), [20, 25), [25, 29.99), [29.99, 30.01), and [30.01, 40) for upper-air DPD; and [0, 1), [1, 3), [3, 5), [5, 7), [7, 11), and [11, 15) for surface DPD. Time series with fewer than 120 months of data were skipped. If there were more than one changepoint detected within any 12-month period from the tests of the frequency series for all the DPD bins, then only the changepoint with the largest value of the test statistic was retained.

The PMFred test assumes a common trend in the whole tested series and it does not require a reference series. The DPD, unlike q or PW, is relatively stable and contains only small trends. This should minimize any long-term trends in the DPD sampling frequency and thus helps the PMFred test. In general, we found the mean shifts induced by nonclimatic changes in the frequency series are much larger than trends induced by real long-term changes in the DPD series. Furthermore, short-term variations in the frequency series are relatively small compared with those in the DPD data series

(cf. Figs. 3 and 6). Thus, the use of the DPD sampling frequency series, instead of the DPD data series itself, in the PMFred test should improve the test results.

c. Synthesizing the results of the K–S and PMFred tests

The changepoints detected by the K–S and PMFred tests are usually very close for major discontinuities. For small to moderate discontinuities, they may not agree with each other (see statistics in section 3). To account for uncertainties in the detected locations and to limit the final number of changepoints to a manageable level (e.g., fewer than 15 for most stations), we combined the changepoints as follows: 1) if both the K–S and PMFred tests detect a changepoint within 12 months of each other, then the exact date from the K–S test is used as the location of the shift; 2) if one test detects a changepoint but the other does not within a 12-month range, then this changepoint is retained only if the test statistic is 20% above the critical value. This extra 20% requirement was chosen empirically to limit the number of small to moderate changepoints in the combined final list of changepoints.

Unlike tests of single time series, here we had to make a number of choices in our tests because we were examining multiple levels and multiple frequency time series, and we were using two different tests on two different variables (i.e., DPD distribution and occurrence frequency). An implicit assumption was that an instrumental or observational change in radiosonde observations will cause a discontinuity at the same time for all upper-air levels but with varying magnitudes and properties, so that the discontinuity may be detected more easily at some levels or in one of our tests than in others. Obviously, the combined changepoints will have a false-alarm rate—that is, the percent of the detected changepoints not associated with an actual discontinuity—different from what the nominal significance level (i.e., the α value) implies. This is true in all attempts to combine changepoints from more than tests, in which both the detection rate and false-alarm rate will increase if passing either test constitutes detection, but they decrease if a confirmation is required by more than one test or at numerous levels.

As pointed out in section 3, the final list of changepoints for each station may change if different choices were used in the tests. This will mostly affect the number of changepoints with small to moderate discontinuities because the large ones will be detected with any reasonable choices. Our visual examinations of DPD time series plots from select stations suggest that our changepoints are reasonable and capture all the major discontinuities. While improvements can be made in detecting small to moderate discontinuities with improved metadata and

optimized choices, the effects of the small to moderate discontinuities on estimated long-term trends are much smaller than those from the large ones. We realize that it is subjective regarding how many changepoints one retains from statistical tests. This has a potential consequence: too many can lead to overadjustment whereas too few may cause underadjustment.

APPENDIX B

A Brief Description of the QM Algorithm

The QM algorithm of Wang et al. (2010) includes several steps: 1) detrend the data series to be homogenized, using the linear trend estimated from a multiphase linear regression fit that accounts for the mean shifts at the changepoints in the data series (see Wang 2008b); 2) for each changepoint to be adjusted, divide the data in the chosen time periods (up to 5 yr) immediately before and after the changepoint into Mq (=4–10; see below) quantile categories, and for each quantile, compute the differences between the means of the two periods; 3) in cases of more than one changepoint, accumulate the differences in such a way that the resulting adjustments will lead to adjusting all segments to a chosen reference segment [see Eq. (10) in Wang et al. (2010)]; 4) fit a natural spline to the (accumulated) Mq category-mean differences between the segment to be adjusted and the reference segment (Fig. 11); and 5) use the fitted spline value that corresponds to the empirical cumulative frequency of the datum in the segment as the adjustment amount. The linear trend removed during the detrending is added back to the adjusted data series. See sections 5–7 in Wang et al. (2010) for more details about the QM algorithm and its assumptions, caveats, and ways to relax the assumptions.

As pointed out by Wang et al. (2010), different trends can be estimated for data in different seasons during the detrending step. Similarly, we estimated and removed a linear trend for the DPD data in each quantile category before the QM adjustment, because the DPD data in different quantile categories appear to have different trends (i.e., quantile-dependent trends). That is, we stratified the daily DPD data into categories and fit a multiphase regression model to the series of DPD data in each of the Mq categories separately. Detrending with one common trend for all Mq categories did not work well for the DPD data series (Fig. B1) because it cannot remove the quantile-dependent trends in the DPD data. Note that allowing quantile-dependent trends in the data allows a gradual change over time in the shape of data distribution, although such gradual change is
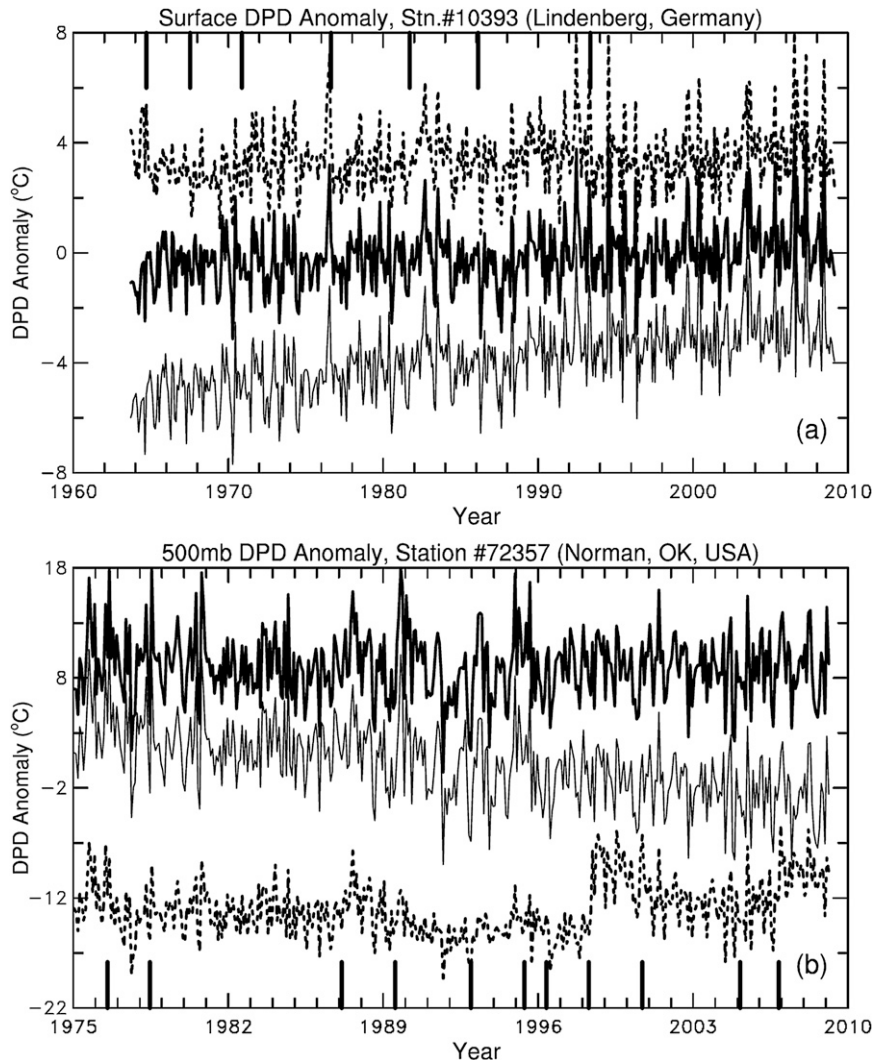
FIG. B1. Monthly anomaly time series from daily DPD data with the cold and dry bias adjustment only (dashed line) and with additional adjustments based on the QM of Wang et al. (2010) with a common trend (thin solid line) for all categories and separate trends for individual categories (dark solid line) for detrending the data before the QM-based adjustment. (top) The surface DPD is from Lindenberg, and (bottom) for 500-hPa DPD from Norman. The vertical bars indicate the locations of the detected changepoints. Note the mean of the time series may be shifted upward or downward to separate the lines.

usually small. Also, the multiphase regression fit only accounts for segment-mean shifts (i.e., one shift value for all data in the same segment), not a value-dependent shift as in the QM adjustment. This inconsistency, however, does not appear to be a problem for the DPD data series. The detrending approach seems to work well for the DPD data, as shown by the dark solid line in Fig. B1 and spatially coherent DPD trend patterns (see section 6). Figure B1 shows that trends in the final adjusted DPD (dark solid lines) are comparable to the apparent underlying trends in the original data series (dashed lines) when the detrending was done separately for each category. In contrast, detrending the whole data series with one common trend leads to an unrealistic trend for both data series (thin solid lines in Fig. B1).

In the QM adjustment of the DPD data, we assumed that the M_q category-mean differences between the two chosen parts of segments separated by a changepoint completely result from nonclimatic changes at the changepoint. This is equivalent to assuming that, except for the linear trends that we estimated and set aside for each of the M_q quantile categories at the beginning of the QM algorithm, there are no low-frequency natural changes in the DPD data over the period covered by the

two chosen parts of segments separated by the change-point that could cause changes in the DPD distribution. Because of this assumption, how one estimates and preserves the deterministic components (trends and natural low-frequency variations) in the data series to be adjusted may have a significant impact on the final adjusted data. In particular, a bias in the linear trend estimate will be passed on to the adjusted data series (e.g., as shown in Fig. B1).

We also tried an additional iteration, in which the adjusted DPD data (with the category-dependent trends added back) were used to reestimate the trends for individual categories for detrending the unadjusted data before estimating the final adjustment. The reason for the iteration is that the first estimates of adjustment remove most of the quantile-dependent shifts, so that the mean-shifts in the second multiphase model fit are small, which could potentially improve the estimate of the trend for each category. However, tests showed that the improvement from the additional iteration was small. Thus, we only show results without the iteration in this paper.

Tests also showed that the histogram of the adjusted DPD data is not very sensitive to the choice of Mq—the number of quantile categories used to estimate the spline for use to derive the QM adjustments. We used Mq = 4 $\sim$ 10, depending on the length of the shortest segment in a given DPD data series (the minimum number of data for estimating a category mean is 20; that is, there must be at least Mq $\times$ 20 data points in a segment). Note that we used a chosen part of segment with up to 1825 data points (over 5 yr if no data gaps but longer if there are missing values) before and after a changepoint to estimate the Mq category differences and then the spline.

REFERENCES

Buck, A. L., 1981: New equations for computing vapor pressure and enhancement factor. *J. Appl. Meteor.,* **20,** 1527–1532.

Clough, C., and M. Padovani, 2008: Upper-air moisture data void discovered and partially rectified. *Bull. Amer. Meteor. Soc.,* **89,** 953–956.

Dai, A., 2006: Recent climatology, variability, and trends in global surface humidity. *J. Climate,* **19,** 3589–3606.

——, I. Y. Fung, and A. D. Del Genio, 1997: Surface observed global land precipitation variations during 1900–88. *J. Climate,* **10,** 2943–2962.

——, T. M. L. Wigley, B. A. Boville, J. T. Kiehl, and L. E. Buja, 2001: Climates of the twentieth and twenty-first centuries simulated by the NCAR climate system model. *J. Climate,* **14,** 485–519.

Della-Marta, P. M., and H. Wanner, 2006: A method of homogenizing the extremes and mean of daily temperature measurements. *J. Climate,* **19,** 4179–4197.

Dessler, A. E., and S. C. Sherwood, 2009: A matter of humidity. *Science,* **323,** 1020–1021.

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

——, C. N. Williams Jr., X. Ying, and R. S. Vose, 2009: Radiosonde-based trends in precipitable water over the Northern Hemisphere: An update. *J. Geophys. Res.,* **114,** D05112, doi:10.1029/2008JD010989.

Elliott, W. P., and D. J. Gaffen, 1991: On the utility of radiosonde humidity archives for climate studies. *Bull. Amer. Meteor. Soc.,* **72,** 1507–1520.

——, and ——, 1993: Effects of conversion algorithms on reported upper-air dewpoint depressions. *Bull. Amer. Meteor. Soc.,* **74,** 1323–1325.

——, R. J. Ross, and B. Schwartz, 1998: Effects on climate records of changes in National Weather Service humidity processing procedures. *J. Climate,* **11,** 2424–2436.

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.

Gaffen, D. J., 1993: Historical changes in radiosonde instruments and practices. Instruments and Observing Methods Rep. 50, WMO/TD-541, 123 pp.

——, 1994: Temporal inhomogeneities in radiosonde temperature records. *J. Geophys. Res.,* **99,** 3667–3676.

——, 1996: A digitized metadata set of global upper-air station histories. NOAA Tech. Memo. ERL ARL-211, 37 pp.

——, T. P. Barnett, and W. P. Elliott, 1991: Space and time scales of global tropospheric moisture. *J. Climate,* **4,** 989–1008.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate,* **20,** 1377–1403.

——, C. Tavolato, and S. Sperka, 2008: Toward elimination of the warm bias in historic radiosonde temperature records—Some new results from a comprehensive intercomparison of upper-air data. *J. Climate,* **21,** 4587–4606.

Held, I. M., and B. J. Soden, 2000: Water vapor feedback and global warming. *Annu. Rev. Energy Environ.,* **25,** 441–475.

Ishihara, M., 2004: Recent tests and comparisons of radiosonde operated by Japan Meteorological Agency. *Extended Abstracts, Commission for Instruments and Methods of Observation (OPAG-Upper-Air): CIMO Expert Team on Upper-Air Systems Intercomparisons and International Organization Committee (IOC) on Upper-Air Systems Intercomparisons,* Geneva, Switzerland, WMO, CIMO/OPAG-UPPER-AIR/ET-UASI-1/IOC-1/Doc. 3.2(3), 8 pp.

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

McCarthy, M. P., 2008: Spatial sampling requirements for monitoring upper-air climate change with radiosondes. *Int. J. Climatol.,* **28,** 985–993.

——, H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2008: Assessing bias and uncertainty in the HadAT-adjusted radiosonde climate record. *J. Climate,* **21,** 817–832.

——, P. W. Thorne, and H. A. Titchner, 2009: An analysis of tropospheric humidity trends from radiosondes. *J. Climate,* **22,** 5820–5838.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 746–845.

Nakamura, H., H. Seko, and Y. Shoji, 2004: Dry biases of humidity measurements from the Vaisala RS80-A and Meisei RS2-91 radiosondes and from ground-based GPS. *J. Meteor. Soc. Japan,* **82,** 277–299.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran 77.* 2nd ed. Cambridge University Press, 933 pp.

Qian, T., A. Dai, K. E. Trenberth, and K. W. Oleson, 2006: Simulation of global land surface conditions from 1948 to 2002. Part I: Forcing data and evaluations. *J. Hydrometeor.,* **7,** 953–975.

Remsberg, E. E., and Coauthors, 2000: Instrumentation and data sets. SPARC assessment of upper tropospheric and stratospheric water vapour, SPARC Rep. 2, WCRP-113, WMO/TD 1043, 33–36.

Ross, R. J., and W. P. Elliott, 1996: Tropospheric water vapor climatology and trends over North America: 1973–93. *J. Climate,* **9,** 3561–3574.

——, and ——, 2001: Radiosonde-based Northern Hemisphere tropospheric water vapor trends. *J. Climate,* **14,** 1602–1611.

Santer, B. D., and Coauthors, 2008: Consistency of modelled and observed temperature trends in the tropical troposphere. *Int. J. Climatol.,* **28,** 1703–1722.

Sapucci, L. F., L. T. Machado, R. B. Da Silveira, G. Fisch, and J. F. G. Monico, 2005: Analysis of relative humidity sensors at the WMO radiosonde intercomparison experiment in Brazil. *J. Atmos. Oceanic Technol.,* **22,** 664–678.

Sherwood, S. C., E. R. Kursinski, and W. G. Read, 2006: A distribution law for free-tropospheric relative humidity. *J. Climate,* **19,** 6267–6277.

Simmons, A. J., S. M. Uppala, D. P. Dee, and S. Kobayashi, 2007: ERA Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter,* No. 110, ECMWF, Reading, United Kingdom, 25–35.

——, K. M. Willett, P. D. Jones, P. W. Thorne, and D. P. Dee, 2010: Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *J. Geophys. Res.,* **115,** D01110, doi:10.1029/2009JD012442.

Solomon, S., K. H. Rosenlof, R. W. Portmann, J. S. Daniel, S. M. Davis, T. J. Sanford, and G.-K. Plattner, 2010: Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science,* **327,** 1219–1223.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005: Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.,* **110,** D18105, doi:10.1029/2004JD005753.

——, J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine, 2011: Tropospheric temperature trends: History of an ongoing controversy. *WIRCC,* **2,** 66–88, doi:10.1002/wcc.80.

Trenberth, K. E., J. Fasullo, and L. Smith, 2005: Trends and variability in column-integrated atmospheric water vapor. *Climate Dyn.,* **24,** 741–758.

——, L. Smith, T. Qian, A. Dai, and J. Fasullo, 2007: Estimates of the global water budget and its annual cycle using observational and model data. *J. Hydrometeor.,* **8,** 758–769.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

U.S. CCSP, 2006: Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences. Synthesis and Assessment Product 1.1, U.S. Climate Change Science Program, 180 pp. [Available online at http://www.climatescience.gov/Library/sap/sap1-1/default.php.]

Wade, C. G., 1994: An evaluation of problems affecting the measurement of low relative humidity on the United States radiosonde. *J. Atmos. Oceanic Technol.,* **11,** 687–700.

Wang, J., and L. Zhang, 2008: Systematic errors in global radiosonde precipitable water data from comparisons with ground-based GPS measurements. *J. Climate,* **21,** 2218–2238.

——, H. L. Cole, and D. J. Carlson, 2001: Water vapor variability in the tropical western Pacific from a 20-year radiosonde data. *Adv. Atmos. Sci.,* **18,** 752–766.

——, ——, ——, E. R. Miller, K. Beierle, A. Paukkunen, and T. K. Laine, 2002: Corrections of humidity measurement errors from the Vaisala RS80 radiosonde—Application to TOGA COARE data. *J. Atmos. Oceanic Technol.,* **19,** 981–1002.

——, D. J. Carlson, D. B. Parsons, T. F. Hock, D. Lauritsen, H. L. Cole, K. Beierle, and E. Chamberlain, 2003: Performance of operational radiosonde humidity sensors in direct comparison with a chilled mirror dew-point hygrometer and its climate implication. *Geophys. Res. Lett.,* **30,** 1860, doi:10.1029/2003GL016985.

——, L. Zhang, A. Dai, T. Van Hove, and J. Van Baelen, 2007: A near-global, 2-hourly data set of atmospheric precipitable water from ground-based GPS measurements. *J. Geophys. Res.,* **112,** D11107, doi:10.1029/2006JD007529.

Wang, X. L., 2008a: Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal *t* or *F* test. *J. Appl. Meteor. Climatol.,* **47,** 2423–2444.

——, 2008b: Penalized maximal *F* test for detecting undocumented mean shift without trend change. *J. Atmos. Oceanic Technol.,* **25,** 368–384.

——, H. Chen, Y. Wu, and Q. Pu, 2010: New techniques for the detection and adjustment of shifts in daily precipitation data series. *J. Appl. Meteor. Climatol.,* **49,** 2416–2436.

Wentz, F. J., and M. Schabel, 2000: Precise climate monitoring using complementary satellite data sets. *Nature,* **403,** 414–416.

Willett, K. M., P. D. Jones, N. P. Gillett, and P. W. Thorne, 2008: Recent changes in surface humidity: Development of the HadCRUH dataset. *J. Climate,* **21,** 5364–5383.

WMO, 2008: Measurement of humidity. WMO guide to meteorological instruments and methods of observation. WMO Rep. WMO-8, 14.1-14.30. [Available from http://www.wmo.int/pages/prog/www/IMOP/publications/CIMO-Guide/CIMO_Guide-7th_Edition-2008.html.]

Zhai, P. M., and R. E. Eskridge, 1996: Analyses of inhomogeneities in radiosonde temperature and humidity time series. *J. Climate,* **9,** 884–894.

——, and ——, 1997: Atmospheric water vapor over China. *J. Climate,* **10,** 2643–2652.