On the Use of Reanalysis Data for Downscaling

S. Brands, J. M. Gutiérrez, and S. Herrera

Instituto de Física de Cantabria, University of Cantabria, Santander, Spain

A. S. Cofiño

Department of Applied Mathematics and Computer Science, University of Cantabria, Santander, Spain

(Manuscript received 5 May 2011, in final form 16 November 2011)

ABSTRACT

In this study, a worldwide overview on the expected sensitivity of downscaling studies to reanalysis choice is provided. To this end, the similarity of middle-tropospheric variables—which are important for the development of both dynamical and statistical downscaling schemes—from 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) and NCEP–NCAR reanalysis data on a daily time scale is assessed. For estimating the distributional similarity, two comparable scores are used: the two-sample Kolmogorov–Smirnov statistic and the probability density function (PDF) score. In addition, the similarity of the day-to-day sequences is evaluated with the Pearson correlation coefficient. As the most important results demonstrated, the PDF score is found to be inappropriate if the underlying data follow a mixed distribution. By providing global similarity maps for each variable under study, regions where reanalysis data should not assumed to be ‘‘perfect’’ are detected. In contrast to the geopotential and temperature, significant distributional dissimilarities for specific humidity are found in almost every region of the world. Moreover, for the latter these differences not only occur in the mean, but also in higher-order moments. However, when considering standardized anomalies, distributional and serial dissimilarities are negligible over most extratropical land areas. Since transformed reanalysis data are not appropriate for regional climate models—in opposition to statistical approaches—their results are expected to be more sensitive to reanalysis choice.

1. Introduction

With over 8200 and 1900 citations respectively, the first National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis (Kalnay et al. 1996) and the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) are the most widely used datasets in modern atmospheric sciences. They have been extensively used in the development of both dynamical and statistical downscaling schemes to either provide the lateral boundary conditions for driving regional climate models (RCMs; e.g., Herrera et al. 2010; Zahn and von Storch 2010) or to calibrate and test statistical transfer functions (Wilby and Wigley 1997; Hanssen-Bauer et al. 2005; Fowler et al. 2007; Maraun et al. 2010 and references therein).

Although it is well known that reanalysis data suffer from inhomogeneities as well as distributional and serial dissimilarities (Kistler et al. 2001; Trenberth et al. 2001; Bengtsson et al. 2004; Sterl 2004; Trenberth and Smith 2005; Trenberth et al. 2005, 2007; Chen et al. 2008a,b; Pitman and Perkins 2009; Screen and Simmonds 2011; Trenberth et al. 2011), which are especially frequent in regions where the observational network is sparse, they are interchangeably used and implicitly assumed to be ‘‘perfect’’ in downscaling studies (Jones et al. 2011). This is a subject of some concern, as the results of both dynamical and statistical downscaling approaches have been shown to be sensitive to the choice of the reanalysis data used for their development (Fernández et al. 2007; Koukidis and Berg 2009; Eum et al. 2011).

In this study, we provide a worldwide overview on the expected sensitivity of downscaling studies to reanalysis choice. To this end, we test the similarity of 1) the

Corresponding author address: S. Brands, Instituto de Física de Cantabria, University of Cantabria, Avenida de los Castros, Santander 39005, Spain.
E-mail: brandssf@unican.es

DOI: 10.1175/JCLI-D-11-00251.1

distributions and 2) the day-to-day sequences of middle-tropospheric circulation, temperature, and humidity variables from ERA-40 and NCEP–NCAR reanalysis data on a daily time scale. Although these variables play a central role in the development of both dynamical and statistical downscaling schemes (Buishand et al. 2004; Abaurrea and Asin 2005; Cavazos and Hewitson 2005; Sauter and Venema 2011) to the best of authors' knowledge, a global survey on their degree of agreement is still missing. We focus on the NCEP–NCAR and ERA-40 reanalyses because, in contrast to more recent reanalysis products (Saha et al. 2010; Dee et al. 2011; Ebita et al. 2011; Rienecker et al. 2011), they have been extensively used by the downscaling community and hence are more relevant within the context of our study.

For assessing distributional similarity, two alternative scores are applied: the two-sample Kolmogorov–Smirnov statistic (KS statistic), defined as the maximum distance between two empirical cumulative distribution functions, and the probability density function (PDF) score (Perkins et al. 2007), to estimate the overlapping probability density area for both series. After detecting which of these scores is preferable for this kind of study, we provide global maps of distributional similarity for both boreal winter (December–February, DJF) and summer (June–August, JJA) and each of the above-mentioned variables. For assessing the correspondence of the day-to-day sequences, global maps of the Pearson correlation coefficient are also shown. These maps should be of general interest to the downscaling community, since they detect those regions of the world where—for a given variable of interest—reanalysis data should not assumed to be "perfect."

Whereas in dynamical approaches the raw reanalysis data have to be applied in order to maintain the internal consistency between different variables used to define the lateral boundary conditions (Laprise 2008), statistical downscaling techniques are able to work with data transformations that potentially correct distributional dissimilarities. Hence, we additionally calculate distributional differences for the anomalies (zero mean) and standardized anomalies (zero mean and unit variance) of the original time series and we map those geographical areas where data transformation is recommended as a precursor step of statistical downscaling approaches.

In addition to showing where downscaling studies are expected to be sensitive to the underlying reanalysis data, the agreement of two distinct reanalyses indicates that they are more constrained by assimilated observations than by internal model variability and thus can reasonably assumed to reflect reality (Sterl 2004). On the contrary, in case of considerable inconsistencies, at least one of the reanalyses is dominated by internal model variability rather than the observations and hence cannot be assumed to reflect reality. There also exists the possibility of both reanalyses being wrong in spite of perfect agreement between them. The probability of such a consistent error is especially high if observations prone to considerable measurement errors are assimilated in both reanalyses, as is the case for moisture data from operational radiosondes (Elliott and Gaffen 1991; Ross and Elliott 2001; Wang et al. 2003). However, reliable global observational datasets for middle-tropospheric variables on daily time scale are not available yet. Thus, even if both reanalyses were mistaken, this could not be verified for these variables. On the basis of these considerations, the disagreement between the two reanalysis datasets is an appropriate first estimator of observational uncertainty.

2. Data

Reanalysis similarity is tested for the 0000 UTC time series of temperature T, geopotential Z, and specific humidity Q at 500 and 850 hPa (e.g., Z500), for both boreal winter (DJF) and summer (JJA). The 21-yr period from 1980 to 2000 is chosen to include satellite data, which considerably improve the quality of the reanalysis products from 1980 onward Sterl (2004).

The ERA-40 data have been downloaded from the ECMWF data server (http://data-portal.ecmwf.int/data/d/era40_daily) and the NCEP–NCAR reanalysis data have also been obtained online (http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/ncep.reanalysis). Both datasets come on an identical regular grid of 2.5°, which is convenient for the present comparison. Note that these publicly available data are the products of internal interpolation processes, which are known to introduce additional errors. Hence, it is advisable to use the native resolution (1.125° for ERA-40 and 1.875° for NCEP–NCAR) to fully analyze a particular reanalysis.

While the satellite data assimilated in NCEP–NCAR consisted of temperature retrievals only (Kalnay et al. 1996), additional moisture-related retrievals were used in ERA-40 (see Uppala et al. 2005, p. 2965, for more details). Hence, inconsistencies in Q are expected to be larger than in Z and T, especially over the ocean areas and ice sheets, where the radiosonde station coverage is sparse.

3. Methods

At each grid box, the agreement between both reanalyses is assessed in terms of 1) distributional similarity and 2) correspondence of the day-to-day sequences (hereafter also referred to as serial similarity). For the first condition, the probability distributions are compared, which accounts for differences in both the mean (bias) and in higher-order moments. To this end, we compare two competing scores: the statistic of the classical two-sample KS statistic (see, e.g., Wilks 2006) and the recently suggested PDF score (Perkins et al. 2007), which has been frequently used for assessing distributional similarity (Maxino et al. 2008; Pitman and Perkins 2009; Mao et al. 2010; Brands et al. 2011a,b; Kjellstrom et al. 2010).

The PDF scores and KS statistics are calculated separately for each season (winter and summer), considering the corresponding daily time series. Moreover, in order to isolate distributional dissimilarities due to errors in the first- and second-order moments, we also consider anomalies and standardized anomalies. In the first case, we remove the seasonal mean, whereas in the second case we additionally divide by the seasonal standard deviation.

As in Pitman and Perkins (2009), the PDF score is used as a metric of agreement between the PDFs of the two reanalysis datasets. Probability densities for both the NCEP–NCAR (f) and ERA-40 (g) time series are estimated at N equally spaced bins $m_1, \ldots, m_N$ spanning the range of the joined sample (in this work we consider $N = 64$). For this purpose, we apply kernel density smoothing with Gaussian kernels, a nonparametric technique for fitting a theoretical distribution to an empirical dataset [see Perkins et al. (2007) for details on the particular Matlab implementation]. Thereafter, the densities are normalized by their sum, and the minimum bin values are aggregated as follows:

$$\text{PDF-score} = \sum_{i=1}^{N} \min\{f(m_i), g(m_i)\}. \tag{1}$$

Thus, the PDF score has an intuitive interpretation as the common overlapping probability density, yielding one for identical distributions and zero for completely disjoint ones.

The KS test is a nonparametric statistical hypothesis test for checking the null hypothesis (H_0) that two candidate datasets come from the same underlying theoretical distribution. It is defined by the statistic

$$\text{KS-statistic} = \max_{i=1}^{2n} |F(z_i) - G(z_i)|, \tag{2}$$

where n is the length of the time series (ranging from 1896 to 1932 days for the DJF and JJA seasons, respectively); F and G are the empirical cumulative frequencies of the NCEP–NCAR and ERA-40 time series, respectively; and z_i denotes the ith data value of the sorted joined sample. This statistic is also bounded between zero and one, but, in contrast to the PDF score, the distributional similarity is indicated by low values.

An advantage of the KS statistic is that its theoretical distribution is known a priori. Consequently, p values for hypothesis testing (H_0: both the ERA-40 and NCEP–NCAR time series come from the same underlying distribution) can be directly estimated (Wilks 2006). For the PDF score, however, no theoretical distribution is available and computationally costly Monte-Carlo methods cannot be circumvented if a statistical inference is to be made (Brands et al. 2011a).

Note that the daily time series used in this study are serially correlated; that is, the number of independent data points in a given time series (the effective sample size n^*) is much lower than the sample size n. Hence, the KS test's assumption of independent data points does not hold and artificially low p values for the KS statistic are obtained, leading to too many type-1 errors (i.e., rejections of the H_0 of equal distributions when it is actually true). Thus, the effective sample size n^* is calculated separately for each NCEP–NCAR and ERA-40 time series before calculating the p value of the KS statistic, assuming that the underlying data follow a first-order autoregressive process (Wilks 2006):

$$n^* = n\frac{1 - p_1}{1 + p_1}, \tag{3}$$

where n^* is the effective sample size and p_1 is the lag-1 autocorrelation coefficient.

In addition to these distribution-oriented scores, the correspondence of the day-to-day sequences is estimated with the Pearson correlation coefficient. Note that both types of differences are important from a downscaling point of view, since they affect the distributional and serial characteristics of the regionalized time series (Charles et al. 2007; Brands et al. 2011b).

4. Results

a. Comparison of KS statistic versus PDF score

To understand which distributional similarity metric is preferable for the present study, we first point out that Q values in the NCEP–NCAR dataset cluster at near-zero values at many grid boxes, which leads to a mixed (discrete–continuous) character for this variable. This is shown in the top and center panels of Fig. 1 for Q500 in DJF and JJA, respectively. These panels map the relative empirical frequency (in percent) of the first of 1000 equally spaced bins and thus illustrate where and to which degree the values for Q cluster near zero. With percentages over 50%, the clustering primarily occurs over Antarctica and Greenland, but as well is relevant over

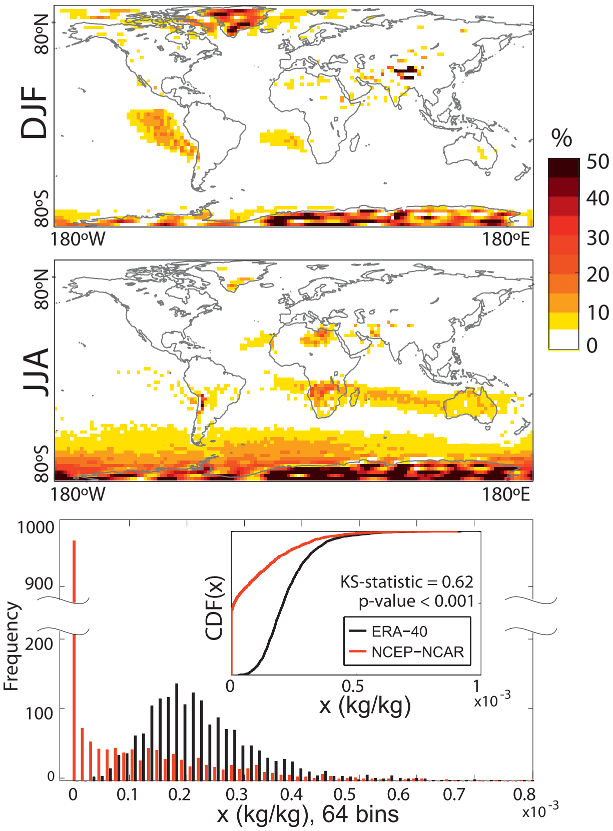

FIG. 1. Percentage of near-zero Q500 values in the NCEP–NCAR data mapped at each grid box for (top) DJF and (middle) JJA. (bottom) The absolute frequencies for a particular illustrative grid box (75°S, 75°E) for DJF. The inset of the histogram shows the empirical CDFs, as well as the resulting KS statistic and its corresponding p value.

the eastern tropical Pacific and the Tibetan Plateau during DJF and over the northeastern Sahara, southern Africa, and the Indian Ocean during JJA. As an illustrative example of this clustering phenomenon, the bottom panel in Fig. 1 shows the histograms [considering 64 bins, as in Eq. (1)] and empirical cumulative distribution functions (CDFs) for a grid box on the East Antarctic Ice Sheet (EAIS) (75°S, 75°E).

Obviously, the clustering of near-zero values for Q in NCEP–NCAR leads to a mixed, precipitation-like distribution (note the scale jump along the ordinate of the histogram in Fig. 1). This poses important limitations on the calculation of the PDF score, since the kernel density smoothing applied in (1) has been found to be inappropriate in this case. In this context, we have found that the calculation of the PDF score may lead to inaccurate results if the percentage of near-zero values exceeds a threshold of 5% (as was the case in the previous example). In those cases, the mixed character of the sample poses numerical problems on the kernel density estimation. In turn, because of its empirical nature, the KS statistic is not affected by this problem and thus is the preferred score when generating global maps of distributional similarities between both reanalyses.

Apart from the above-mentioned problem, our results are generally insensitive to the applied score, with a clear linear relationship between PDF score and KS statistic. This is illustrated in Figs. 2a and 2b, which shows the KS statistics against the PDF scores for Z500 and Q500. In the case of Q500 (see Fig. 2b), those markers departing from the linear relationship (see, e.g., point labeled as 3) correspond to grid boxes where differences between both reanalysis datasets are not in the mean but in higher-order moments (see Figs. 2e and 2h). Two further examples are given for the case of optimal distributional similarity (labeled as 1 and shown in Figs. 2c and 2f) and dissimilarity due to differences in the mean (labeled as 2 and shown in Figs. 2d and 2g).

b. *Maps of distributional similarity*

Figure 3 maps the distributional similarity between NCEP–NCAR and ERA-40 in terms of the KS statistic for the different variables (columns), seasons (rows), and levels (top and bottom panels). The color darkening from yellow to black denotes increasing values for the KS statistic, (i.e., increasing discrepancies in the corresponding distributions). Note that the KS statistic is displayed only in case the distributional dissimilarities are significant at a test level of 5%. Otherwise, the corresponding grid box is whitened, indicating optimal distributional consistency. Results are presented for both the original and anomaly data in DJF and JJA (in different rows in Fig. 3).

The distributional similarity is generally highest for Z, followed by T and then Q. At 500 hPa (see Fig. 3, top), the corresponding spatial patterns can be grouped into two classes: T and Z on the one hand and Q on the other. For Z500, significant distributional dissimilarities are concentrated on the tropical oceans and adjacent land areas like eastern tropical Africa and the Andes, and are of considerable magnitude as measured by the KS statistic. In JJA they additionally cover the Amazon Basin, the eastern Sahel, India, and the Malay Archipelago, while in DJF significant dissimilarities arise over the Subantarctic Belt.

The distributional difference pattern of T500 is similar to that of Z500, but more extensive. In general, more land areas are affected by significant dissimilarities, which is especially evident over the EAIS in DJF. Note that in many regions (e.g. the eastern tropical Pacific, the western tropical to subtropical North Atlantic, and the Amundsen Sea) results are sensitive to seasonality.

Distributional agreement for Q500 is considerably weaker than for Z500 and T500, and exhibits distinct

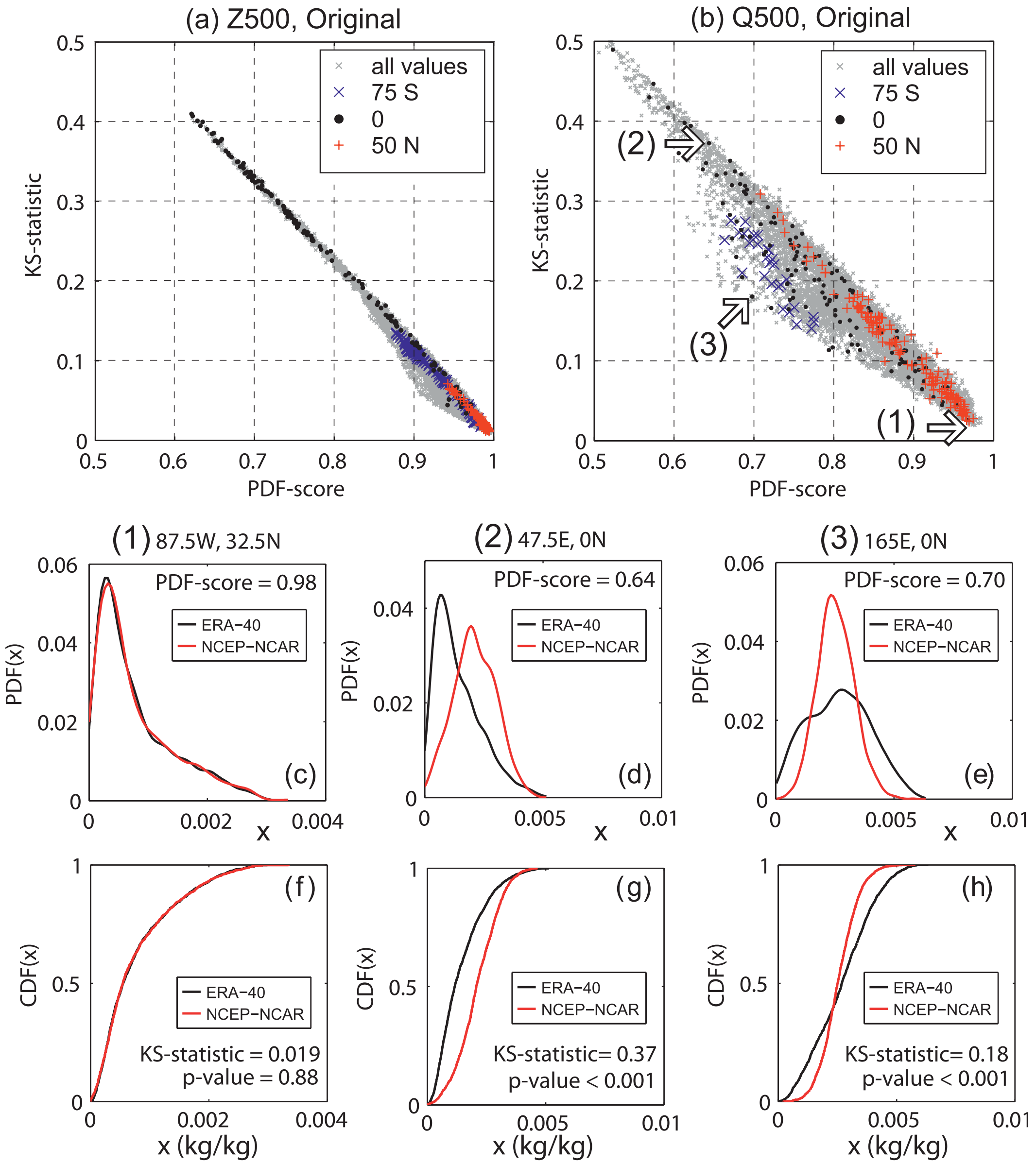

FIG. 2. (a),(b) PDF score vs KS statistic for all grid boxes, except for those where the PDF score could not be calculated (see text for details); grid boxes along 0° and 50°N, 75°S are colored. (c)–(h) PDFs and associated PDF scores, as well as CDFs and their associated KS statistics/p values for the grid boxes labeled '1'–'3' in (b). All results are for the original DJF data.

spatial patterns. With a few exceptions in central to eastern Europe and North America, areas of optimal distributional similarity for both seasons are virtually absent. The lowest consistency is found over Greenland, the tropics, and Antarctica, while the seasonal variations of the results are most pronounced for the Northern Hemisphere midlatitudes.

In contrast to the results obtained at 500 hPa, the distributional difference patterns at 850 hPa (see Fig. 3, bottom) are more similar for T and Q than for Z and T. For Z850, as compared with Z500, considerable distributional dissimilarities are found over the EAIS and Tibet in both DJF and JJA, and over Greenland, the Arabian Peninsula, and the Rocky Mountains in JJA. Except for Patagonia, any region in South America is affected by large dissimilarities in at least one season of the year. The entire African continent is covered by marked inconsistencies, with the exception of northern (southern) Africa in DJF (JJA).

For T850, as compared with T500, distributional differences generally increase over the oceans (with the

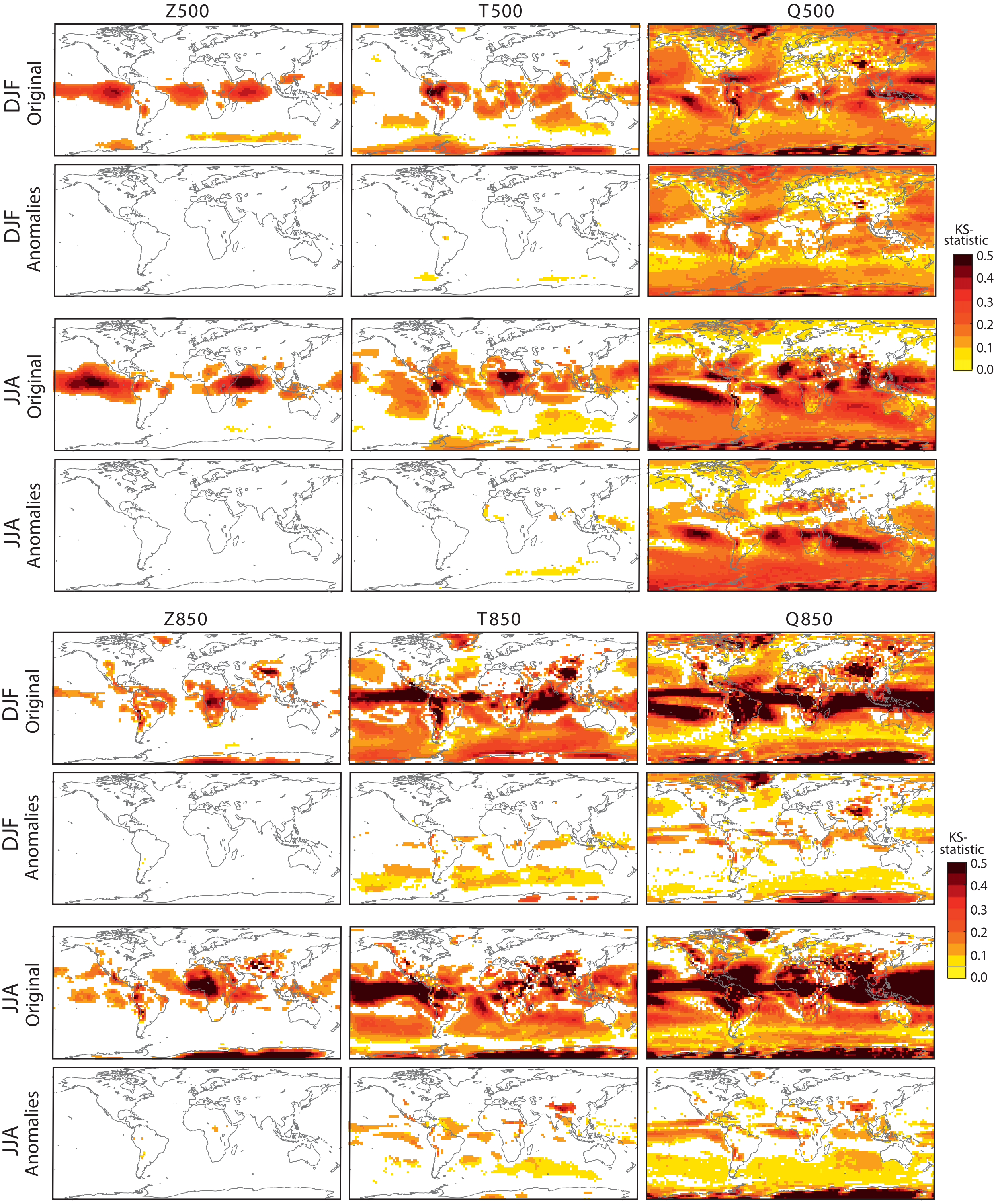

FIG. 3. Maps of distributional similarity for the daily time series of ERA-40 and NCEP–NCAR Z, T, and Q at (top) 500 and (bottom) 850 hPa, as revealed by the KS statistic. Color darkening from yellow to black indicates increasing dissimilarity. If the H_0 values of equal distributions cannot be rejected at a test level of 5%, the grid box is whitened and the distributional similarity is assumed to be optimal. Results are presented for both the original and anomaly data.

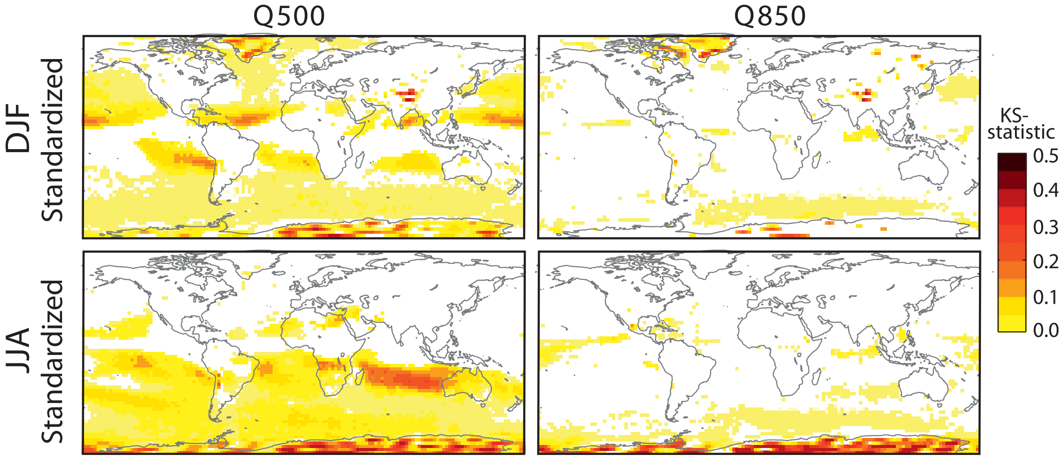

FIG. 4. Maps of distributional similarity for the daily time series of ERA-40 and NCEP–NCAR Q at (left) 500 and (right) 850 hPa, as revealed by the KS statistic. Color darkening from yellow to black indicates increasing dissimilarity. If the H_0 values of equal distributions cannot be rejected at a test level of 5%, the grid box is whitened and the distributional consistency is assumed to be optimal. Results are presented for the standardized anomalies.

exception of the Arctic Ocean), over Greenland, the Antarctic, central Asia, and the Rocky Mountains. For Q850, as compared with Q500, distributional differences are higher in the tropics, but slightly lower in the Southern Hemisphere midlatitudes. A large area of optimal distributional similarity is found over western Eurasia in DJF, as well as over the Arctic (except Greenland) and northern Siberia in JJA.

For Z500, Z850, and T500, the distributional differences are almost exclusively in the mean; that is, they can be corrected by using anomaly time series. This is evident by comparing the KS statistics calculated upon the anomalies with those calculated upon the original data (see Fig. 3).

For T850, distributional differences are limited to the first- and second-order moments. After correcting the mean, significant distributional dissimilarities remain over the tropics and Southern Hemisphere (see Fig. 3, anomalies). However, they completely disappear if both the mean and variance are corrected (i.e., standardized anomalies are compared; not shown).

In contrast to Z and T, large areas of significant distributional differences remain for the anomaly data of Q (see Fig. 3, anomalies), which means that for this variable errors are in higher-order moments rather than in the mean. If standardized anomalies are compared (see Fig. 4, standardized), the distributional similarity for Q850 is optimal over virtually all land areas except Tibet, Greenland, and the EAIS, as well as over most ocean areas of in the extratropical Northern Hemisphere and in the Southern Hemisphere subtropics. For the standardized anomalies of Q500, however, significant distributional dissimilarities persist over a large part of the oceans, the ice sheets, South America, and southern Africa in at least one season of the year, indicating the presence of distributional differences associated with skewness and/or kurtosis.

Note that similar spatial patterns of distributional differences are obtained when applying the same analysis to the 21-yr time series of the presatellite area (1959–78). This indicates that the effect of the major observational changes introduced by assimilating satellite data from 1979 onward is of minor importance for the distributional similarities of both reanalyses.

c. *Correlation maps*

As was the case for the distributional similarity, correlation is generally highest for Z, followed by T and Q (see Fig. 5). Areas of poor correlation are confined to the Antarctica, the tropics and, in case of Q, the subtropics. Relative to the patterns of the distributional differences, areas of poor correlation (below 0.4) are less extensive, indicating that high correlation does not necessarily imply distributional similarity. This is most evident for Q at both height levels, as well as for T850. For these variables, high correlation coefficients are contrasted by considerable distributional differences over the extratropical oceans and Greenland. The same finding can be observed over the EAIS for T850 and Q850 in DJF (cf. Figs. 5 and 3).

Note that the statistical significance of correlation was estimated by first calculating the effective sample size, as described in Kristjansson et al. (2002), and then applying a standard two-sided significance test on the basis of Student's t distribution. This procedure corrects for committing too many type-1 errors in the face of serially

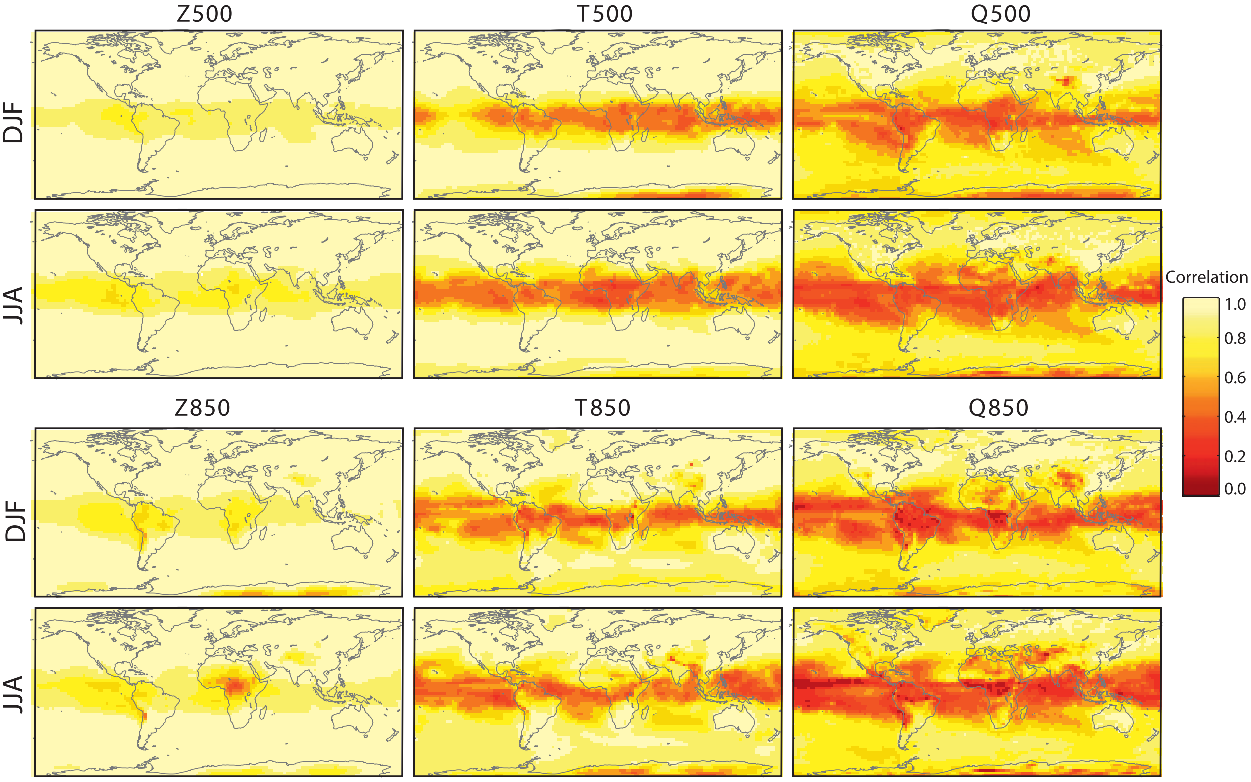

FIG. 5. Maps of consistency for the day-to-day sequence of the daily time series of ERA-40 and NCEP–NCAR Z, T, and Q at (top) 500 and (bottom) 850 hPa, as revealed by the Pearson correlation coefficient. Color darkening from yellow to black indicates increasing dissimilarity.

correlated time series. Virtually all correlations where found to be significant at a test level of 5%, the only exception being Q850, for which spurious (i.e., nonsignificant) correlations where found in only 1.4% of the grid boxes. However, as these coincide with very poor correlation values (below 0.1) in any case, statistical inference provides no relevant practical information at this point.

5. Discussion and conclusions

The agreement of daily NCEP–NCAR and ERA-40 reanalysis data, as defined by the similarity of probability distributions, is generally higher for Z and T than for Q. For the latter, a clustering of near-zero values in the NCEP–NCAR data, which is absent in ERA-40, impedes the applicability of the PDF score, while the KS statistic remains robust due to its empirical nature. These probably erroneous Q values, which have been contributed to the postprocessing of radiosonde data (Elliott and Gaffen 1991), are not restricted to cold and/or dry regions, as has been previously suggested (Elliott and Gaffen 1991; Chen et al. 2008b; Paltridge et al. 2009), but also occur in warm/humid climates, particularly during summer.

In contrast to Pitman and Perkins (2009), who assessed the distributional similarity of reanalysis products for air temperatures at 2 m and 1000 hPa, in our study and in the case of T850 large differences are not only found in the tropics, but also occur at Southern Hemisphere high latitudes, Greenland, central Asia, and the Rocky Mountains. These differences probably occur because Pitman and Perkins (2009) did not compare each season separately, as was done in this study.

Although the variables under study generally suffer from large serial and distributional differences in cases where they lie below ground (e.g., in Tibet and Antarctica), this does not necessarily hold if the underlying observational network is dense, as, for example, is the case for the European Alps.

In accordance with Sterl (2004), who compared Z500 from ERA-40 and NCEP–NCAR on monthly time scale, the similarity of the day-to-day sequences—as measured by the Pearson correlation coefficient—is generally weak in the tropics. High correlation does not necessarily imply high distributional similarity, which is evident from the results of T and Q over the extratropical oceans and Greenland and underlines the added value of assessing distributional similarity.

If compared with our study, the results of earlier global comparisons between NCEP–NCAR and ERA-40, which mostly assessed monthly to annual mean values, are only partly transferable to the daily time scale. To quote an example, significant distributional inconsistencies for atmospheric moisture were not only found over the oceans (Trenberth et al. 2005), but over most land areas as well. This shows that assessing the agreement of reanalysis data on a daily time scale provides added value to doing so on monthly or seasonal mean time scales, which is in accordance with the results of Pitman and Perkins (2009) and Ben Daoud et al. (2009).

The present study should be of general interest to the downscaling community, since it shows that the sensitivity of downscaling applications to reanalysis uncertainty is expected to be significant in most of the regions where the current downscaling efforts are concentrated [e.g., in Africa, one of the target regions of the international Coordinated Regional Climate Downscaling Experiment (CORDEX) initiative (Giorgi et al. 2009; Jones et al. 2011)]. In particular, weighting dynamical models (RCMs) according to their reanalysis-driven performance, or using reanalysis data for perfect prognosis statistical downscaling applications, may be problematic in these regions. Although applying third-generation reanalysis data (Saha et al. 2010; Dee et al. 2011; Ebita et al. 2011; Rienecker et al. 2011) for downscaling is expected to more closely reflect ''reality,'' it is important to recall that validating downscaled time series against in situ observations is not only a measure of model performance, but of reanalysis quality as well.

The final message is that middle-tropospheric variables from reanalysis data should not be uncritically assumed to be ''perfect'' in downscaling studies. This is particularly the case for Q, a variable that not only suffers from differences in the mean but in higher-order moments as well. To alleviate this problem, we recommend researchers work with (standardized) anomalies, which largely reduce distributional differences. For the statistical downscaling approach, reanalysis uncertainty—as defined by the KS statistic and Pearson correlation—can be essentially removed over most extratropical land areas except Greenland and Antarctica by using standardized anomalies. For the dynamical downscaling approach, which has to work with untransformed reanalysis data in order to keep the internal consistency among the boundary variables, we recommend exploring the sensitivity to several driving reanalysis conditions.

Acknowledgments. We are very grateful to the three anonymous referees who helped to considerably improve the former versions of our manuscript, and we thank Jesús Asín and Jesús Abaurrea, University of Zaragoza, for their critical comments on the PDF score. This work has been funded by the strategic action for energy and climate change by the Spanish R+D Program esTcena, Code 200800050084078, and CICYT Project CGL2010-21869.

REFERENCES

Abaurrea, J., and J. Asin, 2005: Forecasting local daily precipitation patterns in a climate change scenario. *Climate Res.,* **28,** 183–197.

Ben Daoud, A., E. Sauquet, M. Lang, C. Obled, and G. Bontron, 2009: Comparison of 850-hPa relative humidity between ERA-40 and NCEP–NCAR re-analyses: Detection of suspicious data in ERA-40. *Atmos. Sci. Lett.,* **10,** 43–47, doi:10.1002/asl.208.

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

Brands, S., S. Herrera, D. San-Martin, and J. M. Gutierrez, 2011a: Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. *Climate Res.,* **48,** 145–161, doi:10.3354/cr00995.

——, J. J. Taboada, A. S. Cofino, T. Sauter, and C. Schneider, 2011b: Statistical downscaling of daily temperatures in the NW Iberian Peninsula from global climate models: Validation and future scenarios. *Climate Res.,* **48,** 163–176, doi:10.3354/cr00906.

Buishand, T., M. Shabalova, and T. Brandsma, 2004: On the choice of the temporal aggregation level for statistical downscaling of precipitation. *J. Climate,* **17,** 1816–1827.

Cavazos, T., and B. Hewitson, 2005: Performance of NCEP–NCAR reanalysis variables in statistical downscaling of daily precipitation. *Climate Res.,* **28,** 95–107.

Charles, S. P., M. A. Bari, A. Kitsios, and B. C. Bates, 2007: Effect of GCM bias on downscaled precipitation and runoff projections for the Serpentine catchment, Western Australia. *Int. J. Climatol.,* **27,** 1673–1690, doi: 10.1002/joc.1508.

Chen, J., A. D. Del Genio, B. E. Carlson, and M. G. Bosilovich, 2008a: The spatiotemporal structure of twentieth-century climate variations in observations and reanalyses. Part I: Long-term trend. *J. Climate,* **21,** 2611–2633.

——, ——, ——, and ——, 2008b: The spatiotemporal structure of twentieth-century climate variations in observations and reanalyses. Part II: Pacific pan-decadal variability. *J. Climate,* **21,** 2634–2650.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.,* **137A,** 553–597, doi: 10.1002/qj.828.

Ebita, A., and Coauthors, 2011: The Japanese 55-year Reanalysis ''JRA-55'': An interim report. *SOLA,* **7,** 149–152, doi:10.2151/sola.2011-038.

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

Eum, H.-I., P. Gachon, R. Laprise, and T. Ouarda, 2011: Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme. *Climate Dyn.,* doi:10.1007/s00382-011-1149-3, in press.

Fernández, J., J. P. Montávez, J. Saénz, J. F. González-Rouco, and E. Zorita, 2007: Sensitivity of the MM5 mesoscale model to physical parameterizations for regional climate studies: Annual cycle. *J. Geophys. Res.,* **112,** D04101, doi:10.1029/2005JD006649.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.,* **27,** 1547–1578, doi: 10.1002/joc.1556.

Giorgi, F., C. Jones, and R. Ghassern, 2009: Addressing climate information needs at the regional level: The CORDEX framework. *WMO Bull.,* **58,** 175–183.

Hanssen-Bauer, I., C. Achberger, R. Benestad, D. Chen, and E. Forland, 2005: Statistical downscaling of climate scenarios over Scandinavia. *Climate Res.,* **29,** 255–268.

Herrera, S., L. Fita, J. Fernandez, and J. M. Gutierrez, 2010: Evaluation of the mean and extreme precipitation regimes from the ENSEMBLES regional climate multimodel simulations over Spain. *J. Geophys. Res.,* **115,** D21117, doi:10.1029/2010JD013936.

Jones, C., F. Giorgi, and G. Asrar, 2011: The Coordinated Regional Downscaling Experiment: CORDEX an international downscaling link to CMIP5. *CLIVAR Exchanges,* No. 56, International CLIVAR Project Office, Southampton, United Kingdom, 34–40.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

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

Kjellstrom, E., F. Boberg, M. Castro, J. H. Christensen, G. Nikulin, and E. Sanchez, 2010: Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. *Climate Res.,* **44,** 135–150, doi:10.3354/cr00932.

Koukidis, E. N., and A. A. Berg, 2009: Sensitivity of the Statistical DownScaling Model (SDSM) to reanalysis products. *Atmos. Ocean,* **47,** 1–18, doi:10.3137/AO924.2009.

Kristjansson, J., A. Staple, J. Kristiansen, and E. Kaas, 2002: A new look at possible connections between solar activity, clouds and climate. *Geophys. Res. Lett.,* **29,** 2107, doi: 10.1029/2002GL015646.

Laprise, R., 2008: Regional climate modelling. *J. Comput. Phys.,* **227,** 3641–3666, doi:10.1016/j.jcp.2006.10.024.

Mao, J., X. Shi, L. Ma, D. P. Kaiser, Q. Li, and P. E. Thornton, 2010: Assessment of reanalysis daily extreme temperatures with China's homogenized historical dataset during 1979–2001 using probability density functions. *J. Climate,* **23,** 6605–6623.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.,* **48,** RG3003, doi:10.1029/2009RG000314.

Maxino, C. C., B. J. McAvaney, A. J. Pitman, and S. E. Perkins, 2008: Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. *Int. J. Climatol.,* **28,** 1097–1112, doi:10.1002/joc.1612.

Paltridge, G., A. Arking, and M. Pook, 2009: Trends in middle- and upper-level tropospheric humidity from NCEP reanalysis data. *Theor. Appl. Climatol.,* **98,** 351–359, doi:10.1007/s00704-009-0117-x.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Climate,* **20,** 4356–4376.

Pitman, A. J., and S. E. Perkins, 2009: Global and regional comparison of daily 2-m and 1000-hPa maximum and minimum temperatures in three global reanalyses. *J. Climate,* **22,** 4667–4681.

Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate,* **24,** 3624–3648.

Ross, R., and W. Elliott, 2001: Radiosonde-based Northern Hemisphere tropospheric water vapor trends. *J. Climate,* **14,** 1602–1612.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.,* **91,** 1015–1057.

Sauter, T., and V. Venema, 2011: Natural three-dimensional predictor domains for statistical precipitation downscaling. *J. Climate,* **24,** 6132–6145.

Screen, J. A., and I. Simmonds, 2011: Erroneous Arctic temperature trends in the ERA-40 reanalysis: A closer look. *J. Climate,* **24,** 2620–2627.

Sterl, A., 2004: On the (in)homogeneity of reanalysis products. *J. Climate,* **17,** 3866–3873.

Trenberth, K., and L. Smith, 2005: The mass of the atmosphere: A constraint on global analyses. *J. Climate,* **18,** 864–875.

——, D. Stepaniak, J. Hurrell, and M. Fiorino, 2001: Quality of reanalyses in the Tropics. *J. Climate,* **14,** 1499–1510.

——, J. Fasullo, and L. Smith, 2005: Trends and variability in column-integrated atmospheric water vapor. *Climate Dyn.,* **24,** 741–758, doi: 10.1007/s00382-005-0017-4.

——, 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.

——, J. T. Fasullo, and J. Mackaro, 2011: Atmospheric moisture transports from ocean to land and global energy flows in reanalyses. *J. Climate,* **24,** 4907–4924.

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

Wang, J., D. Carlson, D. Parsons, T. Hock, D. Lauritsen, H. 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.

Wilby, R., and T. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.,* **21,** 530–548.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Elsevier, 627 pp.

Zahn, M., and H. von Storch, 2010: Decreased frequency of North Atlantic polar lows associated with future climate warming. *Nature,* **467,** 309–312, doi:10.1038/nature09388.