Homogeneity of Gridded Precipitation Datasets for the Colorado River Basin

GALINA GUENTCHEV
Postdocs Applying Climate Expertise Fellowship Program, Climate Variability and Predictability Project, University Corporation for Atmospheric Research,* Boulder, Colorado

JOSEPH J. BARSUGLI
NOAA/Cooperative Institute for Research in Environmental Sciences Western Water Assessment, University of Colorado, Boulder, Colorado

JON EISCHEID
NOAA/Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

(Manuscript received 27 January 2010, in final form 8 June 2010)

ABSTRACT

Inhomogeneity in gridded meteorological data may arise from the inclusion of inhomogeneous station data or from aspects of the gridding procedure itself. However, the homogeneity of gridded datasets is rarely questioned, even though an analysis of trends or variability that uses inhomogeneous data could be misleading or even erroneous. Three gridded precipitation datasets that have been used in studies of the Upper Colorado River basin were tested for homogeneity in this study: that of Maurer et al., that of Beyene and Lettenmaier, and the Parameter–Elevation Regressions on Independent Slopes Model (PRISM) dataset of Daly et al. Four absolute homogeneity tests were applied to annual precipitation amounts on a grid cell and on a hydrologic subregion spatial scale for the periods 1950–99 and 1916–2006. The analysis detects breakpoints in 1977 and 1978 at many locations in all three datasets that may be due to an anomalously rapid shift in the Pacific decadal oscillation. One dataset showed breakpoints in the 1940s that might be due to the widespread change in the number of available observing stations used as input for that dataset. The results also indicated that the time series from the three datasets are sufficiently homogeneous for variability analysis during the 1950–99 period when aggregated on a subregional scale.

1. Introduction

This research is a part of a larger project focused on the historical characteristics and projected future changes of precipitation variability in the Colorado River basin, which includes a quantitative comparison between the observed precipitation variability and that simulated by general circulation models (GCMs) for the historical period 1950–99. Such a long-term trend or variability analysis requires the homogeneity of the datasets to be evaluated. Homogeneous climate data reflect only the variations in weather and climate (Conrad and Pollack 1950). Inhomogeneities in the data may distort or even hide the climate signal (Tuomenvirta 2001), resulting in erroneous analyses (Peterson et al. 1998). Data homogeneity can be affected by changes in station location, instrumentation, observation practices, and data processing methods. Inhomogeneity in gridded data could result from the use in the gridding process of data from stations with an inhomogeneous record, or it could be introduced during the gridding process itself. In the latter occasion, the inclusion in the gridding process of different numbers of stations that have become operational at different years, or the use of different sets of stations throughout the period of interest, could be sources of inhomogeneity in the final gridded product.

* The University Corporation for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Galina Guentchev, NOAA/ESRL PSD, 1D-513, 325 Broadway, Boulder, CO 80305. E-mail: galina.guentchev@noaa.gov

DOI: 10.1175/2010JAMC2484.1

© 2010 American Meteorological Society
The two major approaches to homogeneity testing are the direct methods and the indirect methods. Direct methods include, for example, use of metadata and studies of the effects of instrumental changes on the homogeneity of a given time series (Peterson et al. 1998). Indirect methods include statistical or graphical approaches for homogeneity testing of the time series of interest. The indirect methods can be applied to a single station time series to test its homogeneity characteristics (absolute tests) or can be used to compare two or more series to assess the homogeneity characteristics of one or more of the time series (relative tests). In the second instance, one or more reference series are used in the comparison with the series of interest. The reasoning behind this approach is that the differences of correlated time series (the series of interest and a reference series) have lower variance than the original time series, which allows for an easier detection of inhomogeneities (Menne and Williams 2009). The reference series could be tested for homogeneity in advance, or its homogeneity characteristics could be unknown at the time of the analysis. One drawback of the use of reference series from neighboring stations is that if the inhomogeneities are caused by changes occurring simultaneously within a region or a country (e.g., simultaneous introduction of a new instrument at many stations), these changes will not be detected by the relative tests (Peterson et al. 1998; Karabörk et al. 2007). On the other hand, absolute homogeneity tests might flag natural climate fluctuations as inhomogeneities (Aguilar et al. 2003), particularly if these fluctuations are anomalously large, are abrupt, and occur rarely during the period of analysis. In mountainous regions, such as the Colorado River basin, it is almost impossible to identify appropriate stations to use as reference series in homogeneity testing. For this reason we opted for the use of absolute homogeneity tests in this study.

For the historical comparison of the GCM-simulated precipitation variability and the observed precipitation variability, which is an important stage of the larger project, the use of a gridded precipitation dataset based on observations was essential. A gridded dataset allows for a consistent spatial representation of the precipitation distribution over the region of interest comparable in coverage to the GCM output. In addition, such gridded datasets are also being used routinely to force hydrological models (e.g., Christensen and Lettenmaier 2007) and to downscale GCM output (Maurer 2007; Brekke et al. 2007). In this project, three gridded datasets based on observed precipitation were considered: the dataset developed by Maurer et al. (2002), the dataset based on the Maurer et al. (2002) method but with an extended period of record before 1950 and after 1999 created by Beyene and Lettenmaier (referred to hereinafter as BL), and the Daly et al. (1994, 1997) Parameter–Elevation Regressions on Independent Slopes Model (referred to hereinafter as PRISM) dataset.

Testing the homogeneity of gridded datasets has rarely been considered. Moberg and Alexandersson (1997) investigated the homogeneity on a gridcell scale of a subset of the Jones (1994) global gridded temperature dataset covering the Scandinavian region. This dataset has a 5° latitude × 5° longitude resolution. Using the standard normal homogeneity test (SNHT; Alexandersson 1984, 1986), they found that all of the individual gridcell series within the Scandinavian region were inhomogeneous. The authors calculated that the 95% confidence interval for the error in estimates of regional temperature changes, resulting from the use of nonhomogeneous station records in the gridding process, is ±0.5°C for a region covered by about 5 grid cells. In addition, Moberg and Alexandersson (1997) found that the long-term temperature changes calculated using the averaged-over-the-region inhomogeneous gridcell series were no different from the mean regional long-term temperature changes calculated using a separate homogeneous gridded dataset. They attributed this result to the cancellation of the various inhomogeneities present in the Jones (1994) dataset.

These results lead us to hypothesize that, even if the individual gridcell data series of the datasets used here might turn out to be inhomogeneous, a time series averaged over a larger-scale area could be homogeneous enough for variability analysis. Hence, the results from the homogeneity testing of the datasets of interest will indicate not only the usefulness of the data for variability analysis but will inform also the choice of a spatial scale for the larger project focused on precipitation variability in the Colorado River basin.

2. Datasets and methods

a. Datasets

Three gridded monthly precipitation datasets at various spatial resolutions were used in this analysis. The Maurer dataset (Maurer et al. 2002) was developed for the period 1950–99 using daily precipitation totals from the National Oceanic and Atmospheric Administration/National Weather Service Cooperative stations (hereinafter referred to as COOP stations). The precipitation data were gridded to the 1/8° (approximately 12 km) resolution using the synergraphic mapping system (SYMAP) algorithm of Shepard (1984). The gridded daily precipitation data were then scaled to match the long-term monthly averages of the PRISM precipitation dataset (Daly et al. 1994, 1997).

The BL gridded dataset with an approximately 12-km resolution was created also by using daily precipitation
data from COOP stations and applying the Maurer et al. (2002) method. A different set of stations was used in the gridding process relative to those used in the Maurer et al. (2002) dataset (A. Hamlet 2010, personal communication). The final gridded data extend from 1915 to 2006. The dataset was obtained online (http://www.hydro.washington.edu/Lettenmaier/Data/gridded/index_tazebe.html).

The PRISM dataset with a 4-km resolution (Daly et al. 1997) is developed using a digital elevation model (DEM) to estimate the “orographic” elevations of the COOP stations used in the procedure. The stations are grouped onto similar topographic facets. The precipitation at a DEM grid cell is estimated through a regression of precipitation versus DEM elevation established from the stations on the cell’s topographic facet. PRISM continually adjusts its frame of reference by using localized stations on the cell’s topographic facet. PRISM continues its precipitation–DEM elevation relationships (Daly et al. 1994) and also incorporates local adjustments in the method determined by expert judgment.

b. Homogeneity tests

Following Wijngaard et al. (2003), four homogeneity tests were used: the Alexandersson SNHT for a single breakpoint (Alexandersson 1986), the Buishand range test (Buishand 1982), the Pettit test (Pettit 1979), and the Von Neumann ratio test (Von Neumann 1941). These are all absolute tests, relying only on the information from the station (or grid cell) of interest. All of the tests assume that the testing variables are independent and identically distributed. More details about the tests are given in Table 1.

A breakpoint is defined as the starting point of inhomogeneity (Aguilar et al. 2003). After detection of breakpoints in the time series, a comparison with the available station metadata is needed to accept or disregard these breakpoints and to be able to determine potential contributors to the identified inhomogeneities. If a breakpoint is confirmed, one has a choice: analyze the homogeneous portions of the time series separately, adjust the data for inhomogeneities, or estimate the amount of error incorporated in the results by using inhomogeneous data. Even if the detected breakpoint is due to a natural climate fluctuation, one must exercise care in trend and variability analysis. Such a breakpoint might occur in a short time series that sampled only one change in sign that is part of a longer-term climate oscillation.

The homogeneity tests were applied initially to the individual gridcell time series from the three datasets. Because the monthly precipitation amounts exhibit greater variability that might obscure breaks in the time series, the annual precipitation amounts were used instead as a testing variable in the homogeneity analysis, similar to the approaches of other authors (Karabörk et al. 2007; Machiwal and Jha 2008; Beaulieu et al. 2008). The annual sums were accumulated on a water-year basis for two periods: October 1949–September 1999 and October 1915–September 2006. The shorter period was chosen for the original larger study and was restricted to 1950–99 because the available GCM-downscaled data from the World Climate Research Programme Coupled Model Intercomparison Project, phase 3, (CMIP3) experiment models span only the 1950–99 period. The longer period was used in a study focused on “reconciling Colorado River flows” for which it was important to test the homogeneity of the datasets for the period 1916–2006. In addition, the results of this longer period analysis enhance our understanding of the sources of inhomogeneities indicated during the shorter 1950–99 period. Although the annual precipitation gridcell time series were very close to normality, we proceeded with the application of a cube root transformation following the World Meteorological Organization recommendations because some of the tests require the testing variable to be normally distributed.

The homogeneity testing procedure was subsequently applied to the time series of the mean annual precipitation amounts per subregion [the 4-digit hydrological unit boundary code (HUC4) regions] defined according to the U.S. Geological Survey (USGS) hydrological unit classification (Table 2, Fig. 1). This larger spatial scale was considered to test our hypothesis regarding the homogeneity of a time series averaged over a larger-scale area in comparison with the homogeneity of a gridcell time series and to find out an appropriate spatial scale for the subsequent precipitation variability analyses. The results from the homogeneity tests were used to classify the data series in terms of data quality. The classification developed by Wijngaard et al. (2003) was used to determine the applicability of the tested series to trend and variability analysis. According to this approach, time series can be classified into three categories: 1) A

<table>
<thead>
<tr>
<th>Test/details</th>
<th>Normality of testing variable</th>
<th>Year of breakpoint</th>
<th>Sensitive to breakpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buishand</td>
<td>Assumes normality</td>
<td>Provides</td>
<td>In middle of period</td>
</tr>
<tr>
<td>SNHT</td>
<td>Assumes normality</td>
<td>Provides</td>
<td>In beginning and end of period</td>
</tr>
<tr>
<td>Pettit</td>
<td>Rank test</td>
<td>Provides</td>
<td>In middle of period</td>
</tr>
<tr>
<td>Von Neumann</td>
<td>Assumes randomly distributed</td>
<td>Does not provide</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

TABLE 1. Details about the homogeneity tests.
time series is classified as “useful” if one or no tests reject the null hypothesis at the 1% significance level. No clear signal of an inhomogeneity in the series is apparent. The series is sufficiently homogeneous for trend and variability analysis. 2) A times series is classified as “doubtful” if two tests reject the null hypothesis at the 1% level. The results of trend and variability analysis should be interpreted critically. 3) A time series is classified as “suspect” if three or four tests reject the null hypothesis at the 1% level. Suspect series should not be used in trend and variability analyses unless a clear climatic mechanism for the breakpoint is identified, and even then should be used only with proper care in interpretation. In the remainder of the paper, the use of the terms useful, doubtful, and suspect refers to the above-mentioned classification categories, although the quotation marks are dropped hereinafter.

3. Results

a. Results of the homogeneity analysis for the 1950–99 period

1) GRIDCELL HOMOGENEITY

After application of the four homogeneity tests, the results of the analysis indicated that the data series for more than 75% of the grid cells in the three datasets were classified as useful (Fig. 2). Whereas the number of series classified as doubtful is similar in the three datasets, the series indicated as suspect are about 2 times as great in number in the BL dataset relative to the PRISM dataset. In general, the geographical distribution and the quantity of grid cells with time series classified as doubtful and suspect are very similar for the Maurer and the BL datasets, although the Maurer dataset fares better in these analyses. Large clusters of grid cells with time series classified as doubtful or suspect are evident in several subregions—for example, the Colorado head waters (1401), the Lower Colorado–Lake Mead (1501), the Little Colorado (1502), and the Upper Gila (1504). The Little Colorado and, to an extent, the Upper Gila subregions are characterized in all three datasets by large areas with grid cells classified as having doubtful and suspect time series.

Three of the four homogeneity tests used in the analysis give indication of a possible year of break in the time series. The tests are considered to agree if at least two of them specify the same year. For more than 80% of the grid cells, the tests agree on a single year of break, whereas for the rest of the grid cells, the tests do not agree and indicate two, or even three in some cases, different years of break. The results for the grid cells classified as doubtful or suspect from all three datasets show that when the tests agree they most often specify
1977 or 1978 as a year of break in the time series (Fig. 3). Because the two years are consecutive, the break is considered to be single—that is, one and the same break. This break in the series is seen clearly in most of the annual precipitation time series for suspect or doubtful grid cells (e.g., Fig. 4). Highest frequency in the double-break year results have the years 1956 and 1978 (plots not shown).

2) SUBREGIONS HOMOGENEITY

As a next step, the tests were applied to the mean annual precipitation time series for the subregions. The data for 14 of the 16 subregions from the BL dataset were classified as useful. The precipitation series for two subregions, Little Colorado (1502) and Sonora (1508), were classified as doubtful based on the test results (figure not shown; refer to Fig. 1). When the PRISM data were aggregated by subregion, by calculating the mean annual amount for each of the 16 subregions, the results for the 50-yr period October 1949–September 1999 pointed only to the time series of subregion Upper Gila (1504) as being doubtful (figure not shown; refer to Fig. 1). The data series for the rest of the subregions were found to be useful. The results for the Maurer dataset indicated that all of the subregionally aggregated series were useful for variability analysis.

The breakpoints, as confirmed by at least two location-specific tests (where “location specific” means indicating the location of the breakpoint within the time series of interest) that reject the null hypothesis, are 1978 for subregion 1502 and 1977 for subregion 1508 of the BL dataset. The years of break for subregion 1504 from the PRISM dataset as specified by the results are 1956 (indicated by SNHT) and 1978 (indicated by the Pettit test). Given these results, the time series of the subperiods determined by the years of break were subjected to additional testing to find out whether the subperiods denoted by the breakpoints were homogeneous. This step was needed because the homogeneity tests used in this analysis indicate only one breakpoint at a time in a given time series. None of the homogeneity tests rejected the null hypothesis at 1% significance level, and the subperiod series for subregions 1502 and 1508 from the BL dataset as well as the subperiod series for subregion 1504 from the PRISM dataset were classified as useful for variability and trend analysis.

b. Results of the homogeneity analysis for the 1916–2006 period

1) GRIDCELL HOMOGENEITY

The entire available gridcell and subregional time series (October 1915–September 2006) were also tested for homogeneity to search for possible explanations for the inhomogeneities found on different scales. The period 1916–2006 was dictated by the availability of the BL dataset. Given that the Maurer dataset does not extend before 1950, the homogeneity of the time series for the longer period was tested only for the BL and PRISM datasets. The interest in the longer period was also based on the need to establish the applicability of these longer
time series for use in hydrological research focused on the Colorado River basin, as already mentioned in section 2.

The results of the four homogeneity tests applied to the gridded data for the overall period (1916–2006) indicated that the grid cells within the Colorado River basin from the BL dataset were split into the three classes as follows: 58.8% were useful, 8.9% were doubtful, and 32.3% were suspect (Fig. 5). Whereas little change is evident in the number of grid cells classified as doubtful when compared with the period 1950–99, the number of the grid cells classified as suspect has more than doubled. The results for the PRISM dataset indicate that 96% of the analyzed gridcell data series were found to be useful, 2.3% were found to be doubtful, and 1.7% were found to be suspect (Fig. 5). The number of data series classified as doubtful and suspect has decreased dramatically, up to 4 times, for the longer 1916–2006 period for the PRISM dataset.

The possible years of break in the gridcell time series for the period 1916–2006 indicated by the test results cluster in the 1940s (1949 and 1950 were specified most often) for the BL dataset and 1977 and 1978 for the PRISM dataset for the single–break year results (Fig. 6). The breakpoint in 1949 is clearly evident in an example of a gridcell series from the BL dataset (Fig. 7a), and so is the break in 1977 for series from the PRISM dataset (Fig. 7b). For a large number of grid cells (11.5%) from the BL dataset, the tests specify as a break year 1941, which was an unusually wet year. The years most frequently specified by the different tests for the two-years-of-break results are 1923 and 1941 for the BL dataset and 1941, 1976, and 1977 for the PRISM data.

2) SUBREGIONS HOMOGENEITY

The tests were also applied to the annual precipitation time series of the subregions in the Colorado River basin for the entire October 1915–September 2006 period. The data for 12 of the 16 subregions were classified as useful for the BL dataset. The data series for one subregion, Lower Green (1406), was classified as doubtful and the annual precipitation time series for three subregions—Great Divide–Upper Green (1404), Upper Colorado–Dirty Devil (1407), and Lower Colorado–Lake Mead (1501)—were classified as suspect (figure not shown; refer to Fig. 1). Note that some of these subregions—for example, Great Divide–Upper Green (1404) and Lower Green (1406)—are among the main contributors (with 14.4% and 9.9%; Fig. 8) to the Colorado River streamflow at Lee’s Ferry. The application of the homogeneity procedure to the mean annual precipitation amounts for each of the 16 subregions from the PRISM dataset showed results that contrast with those obtained for the BL dataset. All 16 subregional data series were classified as useful (figure not shown).
The plots of the maxima/minima of the test statistics indicated the break years for the subregions from the BL dataset that were not classified as useful. For two of the subregions (1407 and 1501) the year of break was almost the same, 1942 and 1941, respectively, and for subregion 1404 the tests specified 1950 as the year of break; for subregion 1406 the tests disagreed and indicated 1929 or 1930 as break years of the time series. The subperiods indicated by the break years for the four subregions were in turn tested for homogeneity. The results classify all of the subperiod series as useful.

4. Discussion

a. Potential contributors to detected inhomogeneities

Possible contributors to the detected inhomogeneities could include 1) changes in the number or the set of available stations involved in the gridding process throughout the period, 2) inclusion of inhomogeneous stations series in the gridding process, or 3) potential for absolute homogeneity tests to identify shifts in the series resulting from natural variability (Aguilar et al. 2003). To address some of these sources, the number of available COOP stations throughout the Colorado River basin was plotted against time (Fig. 9). It is evident that up to 1965 the number of stations available to be included in the gridding process steadily increased. In addition, a larger jump is seen in 1948. These differences in the number and location, respectively, of the available stations could be one of the sources of inhomogeneities detected in this analysis.

To investigate further the causes leading to classification of the data series for subregion 1502 from the BL dataset as doubtful for the 1950–99 period (the main dataset and period of interest used in the larger precipitation variability study), station data that might have been used in the gridding process were identified. Out of the total 102 COOP stations available in the 1502 subregion, 34 stations with periods of existence spanning the 1950–99 period were chosen for homogeneity analysis. A common characteristic of the station time series was the existence of months with missing data. The minima/maxima of the test statistics for the stations classified as doubtful and suspect clustered mostly around the same year of break (1978) as found for the subregion. Further, the metadata for all of the stations subjected to homogeneity analysis were also inspected to compare the homogeneity results with the available information about the history of the stations. The metadata did not indicate any changes in the location, instrumentation, or time of
measurement that coincide with the year of break for the doubtful and suspect stations. Subregion 1508 was not the focus of this analysis because of its limited areal extent and minimum contribution to the precipitation characteristics of the Colorado River basin, resulting from its elevation and geographic location.

In addition to these considerations of potential sources of inhomogeneity, we also acknowledge that the inclusion of snowpack telemetry (SNOTEL) data, which became available in the late 1970s in many areas, may have affected some of the homogeneity analysis results—specifically, for the PRISM dataset. The Maurer and BL datasets were not created by inclusion of SNOTEL data. These datasets only included an adjustment in relation to the long-term PRISM climatological means, which we do not consider as a potential source of inhomogeneity, given that there were no historical sequences of precipitation values involved in these adjustments but rather just long-term mean precipitation amounts. We cannot rule out completely the influence that the inclusion of SNOTEL data in the late 1970s could have on the homogeneity of the PRISM gridded dataset. However, the fact that the geographical patterns of the grid cells classified as doubtful or suspect are very similar for the 1950–99 period in all three datasets leads us to be more inclined not to consider the SNOTEL data inclusion as a potential source of inhomogeneities in the PRISM dataset.

A further look into the potential contributors to inhomogeneities shows that it is possible that the influence of the Pacific decadal oscillation (PDO) on the precipitation time series—that is, the shift toward positive PDO in 1977 (Mantua et al. 1997; Gedalof and Smith 2001; Mantua and Hare 2002)—is reflected in the breaks detected in 1977 and 1978. To investigate whether differences exist between the subperiods suggested by the years of break indicated by the homogeneity analysis of the grid cells and the subregions, average precipitation amounts by grid cell were calculated for the 1916–49, 1950–77, 1978–99, and 1978–2006 periods for the BL and PRISM datasets (Fig. 10, showing only the PRISM data maps; the results were the same, but more pronounced, for the BL dataset). There is an increase in the annual precipitation amounts of about 50–150 mm yr$^{-1}$ throughout the Colorado River basin in the 1978–2006 period relative to the 1950–77 period. Even greater is the increase if the shorter 1978–99 period is used for comparison. These larger precipitation amounts that characterize the 1978–99 (or 2006) periods relative to the 1950–77 period have been linked to the change in sign of the PDO (Mantua and Hare 2002). On the other hand, there are no large differences in the annual precipitation amounts by grid cell (varying in some areas only between −50 and +50 mm yr$^{-1}$) between the 1950–77 and 1916–49 periods to indicate a significant shift as evident in the BL time series in 1949 and 1950. Although a 1946/47 shift of the PDO toward a negative phase has been established (Mantua et al. 1997; Hidalgo and Dracup 2003), no effects of this shift were evident in the hydroclimatic time series in the Upper Colorado River basin (Hidalgo and Dracup 2003).

Following the recommendations of one of the reviewers, we decided to test more rigorously the potential
relationship between breakpoints in 1977 and 1978 and the time series of the PDO. Using a method described in Vincent (1998), we performed a linear regression analysis with the PDO as an independent variable and the time series of subregion 1502, which clearly showed a breakpoint in 1977, as the dependent variable. The results from the analysis indicated that a significant relationship existed, given that the adjusted $R^2$ was 0.29 and the F-test results as well as the $p$ value of the regression coefficient were very low. Furthermore, we plotted the residuals of the regression against time, and the fact that the residuals were randomly distributed along the 0 line clearly showed that this model fully explained the inhomogeneity of interest in the tested subregional time series. We performed similar analyses for the longer period (1916–2006) using the PDO and the number of operational stations in the Colorado River basin as independent variables. There was no statistically significant relationship between the PDO and the subregional time series with a break in 1950 that we tested. Hence, we concluded that the PDO is not an explanatory variable for the breakpoint seen in 1950 for the longer period for this subregion. On the contrary, in this case, the different number of available stations was found to explain the existing year of break.

b. Hydrological implications

Most of the runoff in the Upper Colorado River is generated from a small fraction of the area that lies at high elevations along the rim of the basin (Fig. 8). Most of the grid cells identified as suspect during the 1950–99 period were fortunately not located in these productive areas, with the exception of the Colorado River headwaters in the Front Range. In the Lower Colorado River basin, there were considerable suspect areas in the headwaters of the Gila and Little Colorado rivers. As noted above, this may reflect a particularly sensitive response to the PDO in that area. Over the longer period, the BL dataset shows large suspect areas in the headwaters of the Green River, even when aggregated to the subregional scale. These are not present in the PRISM dataset. This would imply that the choice of precipitation dataset has implications for the hydrological modeling of the Green River and hence the Colorado River flow at Lee’s Ferry.

5. Conclusions

Below, the major results of the gridcell homogeneity analysis are given:

- Almost all of the gridcell time series from the PRISM dataset and most of the gridcell time series from the Maurer and the BL datasets within the subregions that generate the most runoff in the Colorado River basin are classified as useful for the 1950–99 period.
- The gridcell series for the period 1950–99 for all three datasets indicate a 1977 or 1978 discontinuity.
- The longer period (1916–2006) is characterized by a greater gridcell time series homogeneity for the PRISM dataset (96% of the gridcell series were classified as useful) in comparison with the BL dataset.
- The breakpoints for the longer period are 1977 for the PRISM dataset and 1949 or 1950 for the BL dataset.
Because the 1940s breakpoints do not appear in both (BL and PRISM) datasets, the detected inhomogeneities in the BL dataset could be due to the gridding method or the choice of stations to include. The BL dataset does show inhomogeneities in high-precipitation areas in the 1940s.

Given that the 1977–78 breakpoints appear in the three datasets, no systematic problems indicated by the metadata were found, and because the upward jump in precipitation is a part of a basinwide pattern it cannot be ruled out that the 1977–78 inhomogeneity is due to natural variability.

The major results of the hydrological subregions homogeneity analysis are given here:

- The subregional homogeneity testing indicates that the time series are useful for most of the subregions from the BL and PRISM datasets and for all of the subregions from the Maurer dataset for 1950–99.
- For the longer period, the PRISM subregional data series were classified as useful, and several subregions from the BL dataset along the western boundary of the Colorado River basin were classified as doubtful or suspect.
- The three datasets are sufficiently homogeneous during the 1950–99 period for trend and variability analysis when averaged on a hydrological-subregion level.

The good quality of the data series aggregated by subregion for all of the datasets as indicated by the homogeneity analysis after 1950, and the fact that the GCMs cannot resolve very well processes at smaller scales, lead to the decision to use the subregional aggregated time series for the subsequent analyses of precipitation variability. Only series classified as useful, or homogeneous portions of series classified as doubtful, were considered for further analyses in the subsequent research.

The results from these analyses have informed several studies focused on the Colorado River water supply as well as the design of new gridded datasets such as the index station–method gridded dataset under development by D. Lettenmaier and A. Wood at the University of Washington. Our preliminary analysis shows much improved homogeneity characteristics of this new dataset, comparable to those of the PRISM data over the 1916–2006 and 1950–99 periods.

In conclusion, we share some thoughts about the development of gridded datasets. The introduction to Hamlet and Lettenmaier (2005) includes several important steps to be considered when creating gridded meteorological datasets. We agree with the set of measures they indicate as essential to compile a credible gridded observed dataset. In addition, we stress the importance of the initial testing of the homogeneity of the stations to be used in the gridding process. Having secured the inclusion of homogeneous stations in the gridding process, it would be ideal to require that the set of stations be the same from the beginning to the end of

FIG. 10. Plots of the differences by grid cells in the means between historical periods (as labeled at tops of panels) for the PRISM dataset.
the period of interest. This may not always be possible and may throw out useful data. We recommend that pure spatial interpolation be avoided when the station mix changes appreciably and that homogenization method be explicitly included in the gridding process. It should be explicitly noted that homogeneity may not apply to all aspects of a dataset. For example, the gridding process for the Hamlet and Lettenmaier (2005) dataset, by construction, is supposed to produce homogeneous data on the 3-month time scale (although we have not tested this assumption) but not necessarily on shorter time scales. So the data could be appropriate for trend analysis on seasonal means but not necessarily on a monthly or daily scale. Therefore, if a dataset is provided online, documentation should state which aspects of the data are likely to be homogeneous and which are not in a manner that is easy for the user of the data to interpret. In addition, we recommend that a station inventory be made available with each dataset. Because the gridding technique or the method of spatial interpolation itself could also introduce inhomogeneities in the final gridded product, we point out the importance of testing various techniques. The topic of the influences of different gridding/spatial-interpolation techniques on the homogeneity characteristics of the produced gridded datasets could be an interesting topic for future research. Last, we would be remiss if we did not emphasize the value of performing stringent testing of the characteristics of the final gridded product from a homogeneity standpoint and also from the standpoint of the representation of the actual meteorological and climatological characteristics of an area/region.

As a final note, we bring to the attention of the potential users of any gridded dataset that it may have been created for a different purpose or tested to different specifications than is appropriate for their analysis. It is important not to trust the usefulness of a gridded dataset blindly but rather to explicitly evaluate its quality for a given application.

Acknowledgments. This research was supported by the Postdocs Applying Climate Expertise (PACE) Fellowship Program, administered by the University Corporation for Atmospheric Research, jointly funded by NOAA, the Bureau of Reclamation, the Western Water Assessment, and the Southern Nevada Water Authority. We acknowledge the invaluable advice and expertise of Klaus Wolter, Robin Webb, and Martin Hoerling, who helped to improve this manuscript. We also extend our gratitude to Dennis Lettenmaier and his group at the University of Washington for making available the Beyene and Lettenmaier dataset for our analyses. Last, we kindly express our gratitude to the editor and to the anonymous reviewers. Their valuable comments and suggestions were greatly appreciated and increased the quality of this paper.

REFERENCES


Maurer, E. P., 2007: Uncertainty in hydrologic impacts of climate
cchange in the Sierra Nevada, California, under two emissions
——, A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen,
2002: A long-term hydrologically based dataset of land surface
fluxes and states for the conterminous United States. *J. Climate*, 15,
3237–3251.
Menne, M. J., and C. N. Williams, 2009: Homogenization of tem-
perature series via pairwise comparisons. *J. Climate*, 22,
1700–1717.
Moberg, A., and H. Alexandersson, 1997: Homogenization of
Swedish temperature data. Part II: Homogenized gridded air
temperature compared with a subset of global gridded air
temperature since 1861. *Int. J. Climatol.*, 17, 35–54.
Peterson, T. C., and Coauthors, 1998: Homogeneity adjustments of
in situ atmospheric climate data: A review. *Int. J. Climatol.*, 18,
1493–1517.
Pettit, A. N., 1979: A non-parametric approach to the change-point
Shepard, D. S., 1984: Computer mapping: The SYMAP inter-
Tuomenvirta, H., 2001: Homogeneity adjustments of temperature
and precipitation series—Finnish and Nordic data. *Int. J.
Climatol.*, 21, 495–506.
Vincent, L., 1998: A technique for the identification of inhomoge-
neities in Canadian temperature series. *J. Climate*, 11,
1094–1104.
Von Neumann, J., 1941: Distribution of the ratio of the mean
square successive difference to the variance. *Ann. Math. Stat.*, 13,
367–395.
Wijngaard, J. B., A. M. G. Klein Tank, and G. P. Können, 2003:
Homogeneity of 20th century European daily temperature