AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015559

**Key Points:**
- The fidelity of four common downscaling methods is assessed in current climate
- Some methods have problems with wet days, wet/dry spells, and extreme events
- Most methods have problems with spatial scaling and interannual variability

**Supporting Information:**
- Readme
- Supplemental figures

**Correspondence to:**
E. Gutmann,
gutmann@ucar.edu

**Citation:**
Gutmann, E., T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. M. Rasmussen (2014), An intercomparison of statistical downscaling methods used for water resource assessments in the United States, *Water Resour. Res.*, *50*, 7167–7186, doi:10.1002/2014WR015559.

Received 13 MAR 2014
Accepted 19 AUG 2014
Accepted article online 23 AUG 2014
Published online 9 SEP 2014

# An intercomparison of statistical downscaling methods used for water resource assessments in the United States

**Ethan Gutmann[1], Tom Pruitt[2], Martyn P. Clark[1], Levi Brekke[2], Jeffrey R. Arnold[3], David A. Raff[4], and Roy M. Rasmussen[1]**

[1]National Center for Atmospheric Research, Boulder, Colorado, USA, [2]United States Bureau of Reclamation, Denver, Colorado, USA, [3]United States Army Corps of Engineers, Seattle, Washington, USA, [4]United States Army Corps of Engineers, Alexandria, Virginia, USA

**Abstract** Information relevant for most hydrologic applications cannot be obtained directly from the native-scale outputs of climate models. As a result the climate model output must be downscaled, often using statistical methods. The plethora of statistical downscaling methods requires end-users to make a selection. This work is intended to provide end-users with aid in making an informed selection. We assess four commonly used statistical downscaling methods: daily and monthly disaggregated-to-daily Bias Corrected Spatial Disaggregation (BCSDd, BCSDm), Asynchronous Regression (AR), and Bias Corrected Constructed Analog (BCCA) as applied to a continental-scale domain and a regional domain (BCCAr). These methods are applied to the NCEP/NCAR Reanalysis, as a surrogate for a climate model, to downscale precipitation to a 12 km gridded observation data set. Skill is evaluated by comparing precipitation at daily, monthly, and annual temporal resolutions at individual grid cells and at aggregated scales. BCSDd and the BCCA methods overestimate wet day fraction, and underestimate extreme events. The AR method reproduces extreme events and wet day fraction well at the grid-cell scale, but over (under) estimates extreme events (wet day fraction) at aggregated scales. BCSDm reproduces extreme events and wet day fractions well at all space and time scales, but is limited to rescaling current weather patterns. In addition, we analyze the choice of calibration data set by looking at both a 12 km and a 6 km observational data set; the 6 km observed data set has more wet days and smaller extreme events than the 12 km product, the opposite of expected scaling.

## 1. Introduction

Many of the impacts of climate change on society are going to be felt at the local level, and even regional-scale problems, such as water resource planning, require detailed spatial information for hydrologic model input. However, currently available climate models (i.e., coupled models of the climate system, including land, atmosphere, ocean, and sea ice) are not able to perform long-term simulations with outputs at spatial resolutions sufficient for use in many impact assessments. As a result, many methods, including statistical and dynamical models, have been proposed to downscale coarse-resolution climate model results to locally relevant information.

There is a large range of statistical downscaling methods available [see *Fowler et al*., 2007; *Maraun et al*., 2010; *Schoof*, 2013], but many recent studies of the impacts of climate change in the water sector use rather basic statistical downscaling methods, some of which are primarily bias correction methods. In particular, many impact assessments are based on statistical downscaling methods that simply rescale the coarse-scale precipitation output from climate models to the finer spatial scales necessary for hydrologic modeling [e.g., see *Brekke et al*., 2009; *Bureau of Reclamation*, 2012; *Brown et al*., 2012; *Hanson et al*., 2012; *Miller et al*., 2013; *Hay et al*., 2014]. It is important to carefully evaluate the suitability of these downscaling methods for water resource and other applications [*Barsugli et al*., 2013].

Here we present an analysis of four statistical downscaling methods and assess their fidelity in reproducing precipitation in current climate using an array of metrics. We focus specifically on methods that are used to rescale precipitation from climate models. However, in this work we analyze these statistical methods as applied to a coarse-resolution reanalysis (NCEP/NCAR Reanalysis [*Kalnay et al*., 1996]) as a surrogate for a

**Table 1.** Sampling of the Availability of Data Sets Developed With the Methods Examined in the Current Study

| Method | Online Availability |
|---|---|
| BCSD, BCCA | http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/dcpInterface.html |
| BCSD, AR | http://cida.usgs.gov/climate/gdp/ |
| BCSD | http://www.engr.scu.edu/~emaurer/global_data/ |
| BCSD, BCCA, AR | http://earthsystemcog.org/projects/ncpp/downscportals |
| BCSD | https://portal.nccs.nasa.gov/portal_home/published/NEX.html |

climate model, as is commonly done in the statistical downscaling literature [e.g., *Wilby et al*., 2000; *Schmidli et al*., 2006; *Benestad et al*., 2007; *Wetterhall et al*., 2007; *Huth et al*., 2008; *Maurer and Hidalgo*, 2008; *Fasbender and Ouarda*, 2010; *Abatzoglou and Brown*, 2011; *Nicholas and Battisti*, 2012]. Because reanalyses are essentially a series of short-term weather forecasts [*Zhang et al*., 2011], this approach does not consider problems with climate model simulations of regional-scale precipitation, and instead focuses directly on differences in statistical downscaling methods. Additional errors associated with problems in simulating regional-scale precipitation may be present when these methods are applied to climate models, and future work will look at that application. Our approach also does not address the effect these methods have on the climate change signal of a climate model, this too will be addressed in future work. In addition, fidelity in simulating current climate is a necessary, but not sufficient condition to ensure skill in simulating future climate.

The four statistical methods are (1) the Bias Corrected Spatial Disaggregation approach (BCSD) [*Wood et al*., 2004], (2) BCSD applied directly at a daily time step (BCSDd) [*Thrasher et al*., 2012], (3) the Constructed Analog (CA) [*Hidalgo et al*., 2008] approach modified with a bias correction (BCCA) [*Maurer et al*., 2010] as well as a regional application of BCCA (BCCAr), and (4) the Asynchronous Regression approach (AR) [*Dettinger et al*., 2004; *Stoner et al*., 2012]. To distinguish the more typical BCSD approach, which is applied at a monthly time step and disaggregated-to-daily values, from the direct daily approach, we will refer to the typical BCSD as BCSDm. Each method is used to compute daily and 12 km fields of precipitation from the 1.9° reanalysis precipitation data in the NCEP/NCAR Reanalysis. It should be noted that these methods may be referred to as bias correction methods, with statistical downscaling being reserved for methods that relate other climate model fields, e.g., wind speed, humidity, and pressure, to precipitation. However, within the communities that develop and use these methods, they are referred to statistical downscaling methods [e.g., *Wood et al*., 2004; *Maurer et al*., 2010; *Stoner et al*., 2012; *Bureau of Reclamation*, 2012; *Yoon et al*., 2012; *Hwang and Graham*, 2013], as such, we retain that nomenclature here. Within the classification scheme of *Wilby et al*. [2004], BCSDd and AR could be considered transfer functions; BCSDm is a combination of a transfer function on the monthly time scale and a delta approach on the daily time scale; BCCA is an analog scheme. Though *Wilby et al*. [2004] also note that downscaling should rely on variables that are well simulated by the climate model, and one could argue that precipitation is not.

These methods are selected because they are widely used and products are available from a variety of websites (Table 1), and thus there is a need for a review of these methods. For example, *Hay et al*. [2014] used the AR method to look at hydrology and stream temperature changes in future climate. *Brown et al*. [2012] used the BCSD data set produced by *Maurer et al*. [2007] when looking at decision making for water resources. Similarly, *Miller et al*. [2013] used the *Maurer et al*. [2007] BCSD data set when looking at changes in streamflow. *Hanson et al*. [2012] used a modified constructed analog when looking at interactions between surface water and groundwater usage scenarios. Finally, *Brekke et al*. [2009] used the *Maurer et al*. [2007] BCSD data set when evaluating climate change impacts on water resources, as did *Bureau of Reclamation* [2012].

Our primary analyses cover the Contiguous United States (CONUS), but to compare the BCCA and BCCAr approaches, we perform additional analyses over three subdomains (Figure 1). These regional foci are analogous to assessing performance for a water resources study [e.g., *Bureau of Reclamation*, 2012]. The Central Rockies domain can also be compared to a 4 km dynamically downscaled data set [*Rasmussen et al*., 2014], which only exists for this region due to computational constraints.

This paper builds on the literature on evaluation and comparison of downscaling methods. Many previous studies focus on only a single statistical method, often one developed in the same study [e.g., *Salathe*, 2005; *Katz and Parlange*, 1995; *Fatichi et al*., 2011; *Jarosch et al*., 2012; *Ning et al*., 2012; *Pandey et al*., 2000; *Vrac et al*., 2007]. A smaller number of studies have compared multiple approaches, sometimes including dynamical downscaling [*Wood et al*., 2004; *Dibike and Coulibaly*, 2005; *Fowler et al*., 2007; *Maurer and Hidalgo*, 2008;

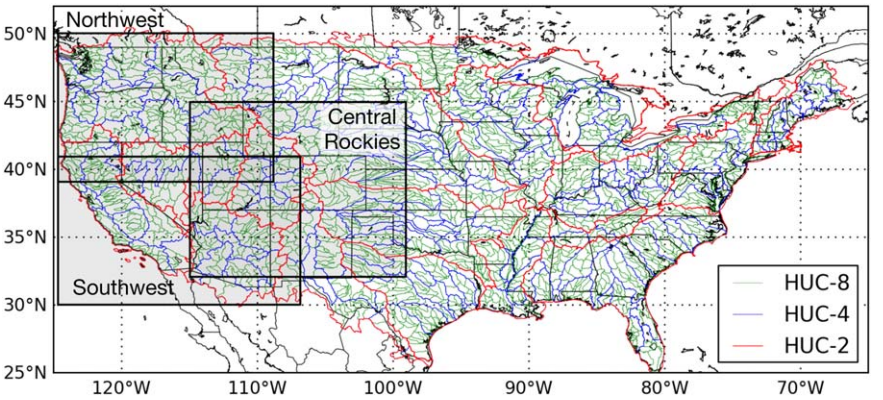


**Figure 1.** Hydrologic Unit Code (HUC) region outlines over the CONUS. Boxes indicate the subdomains used to test the sensitivity of the BCCA method to domain size.

*Chen et al*., 2012; *Gutmann et al*., 2012; *Hwang and Graham*, 2013; *Chen et al*., 2013], though often with only a few statistical validation metrics, e.g., the bias at annual and monthly time scales and, in a few instances, the cumulative distribution function (CDF). Additional metrics such as wet day fraction, extreme event size, or dry spell lengths are often used to validate weather generator type approaches [e.g., *Rajagopalan and Lall*, 1999; *Fatichi et al*., 2011; *Vrac et al*., 2007; *Wilks*, 1998; *Wilby*, 1994; *Mehrotra et al*., 2006], but are rarely used in comparisons of spatially distributed statistical downscaling methods. Particularly unique about the present study is our investigation of the spatial scaling characteristics that are important to hydrology.

This paper is organized as follows: section 2 describes the data sets used in this study; section 3 describes the statistical downscaling methods; section 4 describes the metrics used to assess the output; section 5 describes these results and discusses why each method performs as it does; and section 6 summarizes our findings and provides some suggestions for future research.

## 2. Data Sets

Statistical downscaling methods need a reference data set to calibrate the method, and coarse-resolution model output, which is used in both the calibration and application of the method. The reference data set is typically a fine-resolution gridded observation data set when gridded output is desired, though weather stations are often used for point downscaling. The coarse-resolution model output used is typically from a climate model, but in this case we use a reanalysis data set as a surrogate for climate model simulations that matches historical weather patterns.

### 2.1. Gridded Observational Data Sets

We use the gridded observation precipitation data set of *Maurer et al*. [2002] to calibrate the downscaling methods. To provide additional guidance for the use of various downscaled products, we explore differences between two possible calibration data sets, *Maurer et al*. [2002] and the gridded product of *Livneh et al*. [2013]. Both of these data sets use spatial interpolations from point observations with a topographic adjustment to precipitation similar to that of PRISM (Parameter Regression on Independent Slopes Model) [*Daly et al*., 1994]. Both data sets were designed for climate studies for water resources and incorporate only stations that exist for at least 20 years to minimize homogeneity problems through the record. The Maurer et al. data are on a 1/8° grid (∼12 km on a side at midlatitudes) and the *Livneh et al*. [2013] data are on a 1/16° grid (∼6 km). Both data sets provide daily values of 24 h accumulated precipitation. Both data sets cover the period from 1949 to 2010, and both data sets primarily cover the CONUS domain used in the present study. Both data sets used approximately 20,000 stations in total, although a maximum of approximately 12,000 stations, which occurred in 1970, were used at any given time [*Livneh et al*., 2013]. This corresponds to, at best, one station for every four grid cells in the *Maurer et al*. [2002] data set, and one station for every 16 grid cells in *Livneh et al*. [2013]. Unless otherwise noted, any mention of "observations" will refer to the *Maurer et al*. [2002] data set. Gridded data sets inherently have some limitations due to the necessary spatial interpolations; however, for distributed applications, gridded data sets are typically required.

### 2.2. NCEP/NCAR Reanalysis

We use the coarse-resolution National Center for Environmental Prediction and National Center for Atmospheric Research reanalysis (NCEP/NCAR) [*Kalnay et al*., 1996], henceforth referred to as NCEP product. This product is based on a coarse-resolution atmospheric model similar to that used in a climate model, but it assimilates observations both in the atmosphere and on the surface, and so is representative of current weather. Such reanalyses have been likened to a series of short-term weather forecasts [*Zhang et al*., 2011]. The NCEP data are produced on a 1.9° Gaussian grid (∼210 km on a side) comparable to the resolution of many current climate models. This product provides subdaily output, here aggregated to 24 h totals. While the NCEP data cover the period from 1949 to present, we focus on the period containing satellite microwave and infrared atmospheric soundings (1979 on). The NCEP precipitation data are not assimilated, rather precipitation is modeled by the atmospheric model, as it is in a climate model.

## 3. Downscaling Methods

Below we present a brief description of the downscaling methods evaluated with only the critical elements of the methods presented; the reader is referred to the primary relevant literature for details. These methods were calibrated in the period 1979–1999 and applied to both the calibration period and a validation period (2001–2008). All downscaling methods were used to generate daily data, calibrated to the gridded observations.

### 3.1. Bias Corrected Spatial Disaggregation (BCSDm)

The BCSDm method was developed by *Wood et al*. [2004] and is applied in two steps. First, monthly biases in the coarse model are corrected using quantile mapping [*Panofsky and Brier*, 1968] based on observations regridded to the coarse model resolution. Second, the bias-corrected output is spatially disaggregated to the fine-resolution grid by bilinearly interpolating, and then applying a fine-resolution spatial anomaly pattern derived from the observations. The anomaly is calculated as the difference between the fine-resolution observations and the coarsened observations bilinearly interpolated to the fine-resolution grid. These steps are performed at a monthly time step, and daily disaggregation is performed by selecting a semirandom month from the historical period and scaling it to match the monthly total at each grid cell. To select a historical month, the historical months are grouped into the wettest and driest months, and a month is then selected randomly from either the wet or dry months depending on the BCSDm derived monthly total. Variations of this step are possible as discussed in *Raff et al*. [2009]. The BCSDm method is calibrated independently for each month of the year. To prevent anomalously large extreme events, any events that exceed 150% of the observed maximum daily precipitation for a given grid cell are limited to 150% of the observed maximum, and the excess is spread evenly across the rest of that month to preserve the monthly total.

### 3.2. BCSD—Direct Daily (BCSDd)

Recently, the BCSD method has also been performed directly on a daily time step [*Thrasher et al*., 2012]. In this method, the bias correction and spatial disaggregation are both applied to daily coarse-resolution precipitation. This method requires no further temporal disaggregation, and maintains the daily spatial and temporal structure from the coarse-resolution climate model. As a result, this method allows greater modification to the daily precipitation occurrence and intensity distributions in a future climate than the BCSDm method does; however, it also corrects fewer errors in the climate model's spatial representation of precipitation. For example, the bias correction step forces daily climate model wet day fraction to match a corresponding aggregated coarse-scale wet day fraction from the observations, but every fine-scale grid cell within a coarse grid cell will have the same wet day fraction. There is no method to disaggregate wet day occurrence within a coarse grid cell, as a result, if there is precipitation in the coarse grid cell, then there is precipitation at all fine grid cells within that coarse grid cell simultaneously.

### 3.3. Bias Corrected Constructed Analog (BCCA, BCCAr)

The constructed analog method [*Hidalgo et al*., 2008; *Maurer and Hidalgo*, 2008] is derived from traditional analog techniques [e.g., *Glahn et al*., 1972]; however, it constructs a new analog from a linear combination of past dates. To downscale a given date, this method selects 30 analog days from coarsened historical observations that best match the coarse-resolution model spatial pattern of precipitation. Prior to this comparison, we applied a bias correction step to the coarse model output using quantile mapping [*Maurer*

*et al*., 2010]. The BCCA method takes these 30 best analog days, calculates the linear combination that would best match the coarse model for that day, then applies that same linear combination to the fine-resolution observations from those 30 days to construct a downscaled analog.

In the BCCA technique, the results are dependent on the entire domain for the analog selection process. As a result, the technique may perform better or worse over regions of different sizes. To test this sensitivity, we apply the BCCA method independently to both the CONUS, as well as to three subdomains (Figure 1, BCCAr) in the Northwest (NW), Southwest (SW), and Central Rockies (CR). In addition, because this method is sensitive to the coarse model representation of weather patterns over a large area, it may have additional problems not examined here when applied to a climate model, which may not simulate the correct spatial distribution of precipitation for individual weather events even on a continental scale.

### 3.4. Asynchronous Regression (AR)

The AR method was first applied to climate data by *Dettinger et al*. [2004] and recently refined by *Stoner et al*. [2012]. In this method, the coarse-resolution model output is bilinearly interpolated to the fine-resolution grid. Next, the time series from the each interpolated grid point, and from the fine-resolution observation data set are each sorted independently from low to high values. Once sorted, the resulting ordered arrays are used as input to a linear regression. *Stoner et al*. [2012] used a piecewise linear regression, where the ordered arrays are subdivided and the regression is performed on each subdivision. We follow that method with six segments spread evenly across the distribution. The result is akin to a quantile mapping performed directly on the fine-resolution grid, but with fewer degrees of freedom. As a result, this method will be similar to BCSD if the order of the BC and SD steps is reversed as in *Hwang and Graham* [2013].

The method is performed independently for each month, with 2 weeks on either side included to increase the effective sample size and allow for shifting seasonality in a future climate as in *Stoner et al*. [2012]. While this may aid in downscaling cooler months in a warmer climate, it is unlikely to help the warmest month. Precipitation regressions are calculated after applying a log transformation to the data. As in *Stoner et al*. [2012], we correct the tails of distribution by capping precipitation with a maximum value, here 120% of the maximum observed value in each grid cell. One hundred twenty percent was selected as intermediate value between the 150% used in the BCSD code, and the published value of *Stoner et al*. [2012], which used 2% above the observed maximum after adding 250 mm to the observed precipitation. Excess precipitation is discarded. While this removal of precipitation does not conservation of mass, it should be noted that none of the downscaling procedures do so; they all rescale precipitation to match observed mean values, and this rescaling both adds mass in some locations and removes it in others with no guarantee of balancing the two. In contrast, conservation of mass and related internal consistency is one of the intrinsic benefits of a dynamical downscaling, it plays no role in the statistical methods evaluated here. For the bottom of the distribution, we do not force the regression of the first segment through zero; however, if the coarse model had no precipitation, the downscaled precipitation was set to zero.

## 4. Evaluation Metrics

Relying only on statistics such as the bias can be misleading in isolation because they potentially mask important spatial and temporal errors in the data. To present a more complete assessment of these statistical products, we present the following additional metrics compared to observations: wet day fraction, wet spell length, dry spell length, interannual variation, extreme precipitation values, and spatiotemporal statistics. Where relevant, our metrics are related to the Climate Variability and Predictability (CLIVAR) Expert Team on Climate Change Detection and Indices (ETCCDI) metrics [*Zhang et al*., 2011], for example, mean annual precipitation is the same as the CLIVAR PRCPTOT metric. Wet day fraction is defined as the fraction of days with precipitation greater than some minimum threshold, similar to the CLIVAR Rnnmm metric, which is the number of days per year in excess of a number (nn) of millimeters of precipitation. Wet (dry) spell lengths are the mean number of days between dry (wet) periods, defined as one or more days with precipitation less (greater) than some threshold, this is similar to the CLIVAR cdd and cwd metrics, which are maximum wet and dry spell lengths. Given the shorter validation time period, we felt that the mean would be a more robust statistic than the maximum. Interannual variation here is simply the standard

deviation of annual values. The extreme precipitation and spatiotemporal statistics used are described below. Statistics are calculated at multiple space and time scales as described below.

We assess the impact of methods on extreme events by looking at 1, 2, 3, 4, and 5 day precipitation totals and calculating the 2, 10, 50, and 100 year return interval values. Though 2 and 10 year return interval storms may not be classified as extreme events, they are important for planning purposes. This metric is similar to the CLIVAR Rx1d and Rx5d metrics, which are maximum 1 and 5 day precipitation totals. However, for the relatively short validation period, the maximum values might not be robust. Instead, we calculate the extreme statistics by fitting a gamma distribution to the data [*Katz*, 1999]. We only use the periods in which daily precipitation was greater than 2.54 mm for each grid cell because some of the downscaling methods have excess drizzle, as discussed in section 5.4, and we did not want that to contaminate the fitted distribution. This metric was also calculated using an exponential and a Weibull distribution; while the absolute magnitude of the values differed depending on the type of distribution used, the conclusions reported here were insensitive to the chosen distribution and are not reported separately.

To assess changes in the spatiotemporal patterns between the observations and the downscaled data sets, we use a geostatistical metric, a spatially lagged temporal autocorrelation for each grid cell. The correlation is calculated between the time series of an individual grid cell and that of its neighbors over spatial lags from 12 to 600 km away (1–50 grid cells). This results in fewer samples for coastal grid cells, but they retain the same weight in the final analysis because correlations for a given grid cell are averaged first, next correlations for a given lag distance are averaged across CONUS to produce a correlogram. Differences in space and time with this metric could be significant; however, best geostatistical practices require a domain size that is twice the largest lag distance [*Journel and Huijbregts* 1978], limiting the degree to which we can subdivide the domain.

Hydrologic responses are particularly dependent on issues related to spatial and temporal scales. Responses depend on basin total values in addition to individual grid cell values; hence, we calculate all statistics described above after aggregating precipitation data to hydrologic regions defined by the Hydrologic Unit Code (HUC) [*Seaber et al*., 1987] scales for HUCs 2, 4, and 8 (Figure 1). The HUC-8 scale is only slightly coarser than the resolution of the observations, while the HUC-2 scale is on the order of one quarter of the subdomain. HUC-6 is not used because it is very similar to HUC-4. HUCs 2, 4, and 8 provide approximately an order of magnitude shift in scale between each; there are approximately 50,000 12 km grid cells, 2000 HUC-8s, 200 HUC-4s, and 20 HUC-2s. We use the most recently available HUC data set from the Watershed Boundary Data set (WBD). The WBD is available online at http://datagateway.nrcs.usda.gov [accessed 8 January 2013]. Many applications are also sensitive to the seasonality of precipitation, for this reason, we compute the above statistics for each month as well as for annual values.

As described previously, we apply each of the four methods to the NCEP data sets in both the calibration and validation period. Because wet day occurrence can be important for multiple thresholds, we calculate related statistics using both a 1 mm threshold (as in CLIVAR statistics) and a 0 mm threshold, which is important for some applications. In particular, a 0 mm threshold is used by the MTCLIM microclimate simulator [*Hungerford et al*., 1989] to calculate solar radiation and humidity, which are then used by hydrologic models. While we mention all of these statistics combinations for completeness, we will only present results from the combinations that showed significant differences, additional figures are available in an online supporting information S1.

## 5. Results and Discussion

We focus our results on the validation period, and a summary of common statistics is presented in Table 2. Additional results from the calibration period will be discussed where relevant.

### 5.1. Precipitation Bias

Maps of the mean annual precipitation agree qualitatively with the observations except for the BCCA method (Figure 2). However, substantial spatial biases ($>$400 mm/yr) are present in some areas (Figure 3) and are consistent with changes in NCEP between the calibration and validation periods (Figure 4). Most methods also match the basic seasonal cycle of mean monthly precipitation (Figure 5). With the exception of BCCA, all methods have more precipitation in May and June and less in January and February. Annually,

**Table 2.** Summary Statistics for Each Downscaling Method[a]

| | Mean Annual Precipitation (mm/yr) | Interannual Variation (mm/yr) | 50 yr Return Interval (mm/d) | Wet Day Fraction (0, 1 mm Threshold) | Wet Spell (Days) | Dry Spell (Days) |
|---|---|---|---|---|---|---|
| BCSDm | 805 | 132 | 145 | 0.43, 0.26 | 2.4 | 7.7 |
| BCSDd | 850 | 139 | 109 | 0.88, 0.36 | 4.5 | 5.8 |
| AR | 817 | 161 | 149 | 0.34, 0.24 | 2.1 | 8.1 |
| BCCA | 579 | 101 | 85 | 0.79, 0.27 | 2.0 | 7.4 |
| Observed | 776 | 142 | 140 | 0.39, 0.24 | 2.1 | 7.6 |

[a]All statistics calculated in the validation period (2001–2008) on the individual grid cell level on an annual basis and averaged across the entire (CONUS) domain.

the BCCA method has a substantial dry bias (−197 mm/yr), while all other models are essentially unbiased or have a slightly wet bias (BCSDm: 29 mm/yr, AR: 41 mm/yr, BCSDd: 74 mm/yr). All biases are statistically significant ($p < 0.01$). The largest variation between methods occurs in the late summer months (July–September), when convection is most active. Convection is known to be a very difficult process to parameterize at coarse resolutions [*Randall et al*., 2003; *Weisman et al*., 2008; *Holloway et al*., 2012].

Attributing these biases to the procedures within the downscaling method or to potential discrepancies between the NCEP and gridded observation data sets is difficult. Reanalysis products are not necessarily stationary in time because the observations they incorporate can themselves change over time [*Trenberth et al*., 2011; *Zhang et al*., 2011]. Satellites come and go, as do surface and upper air observations. Of particular relevance, the Advanced Microwave Sounding Unit was first launched on board the NOAA-15 satellite in 1998, prior to 1998 the Microwave Sounding Unit was used to provide global atmospheric sounding input.

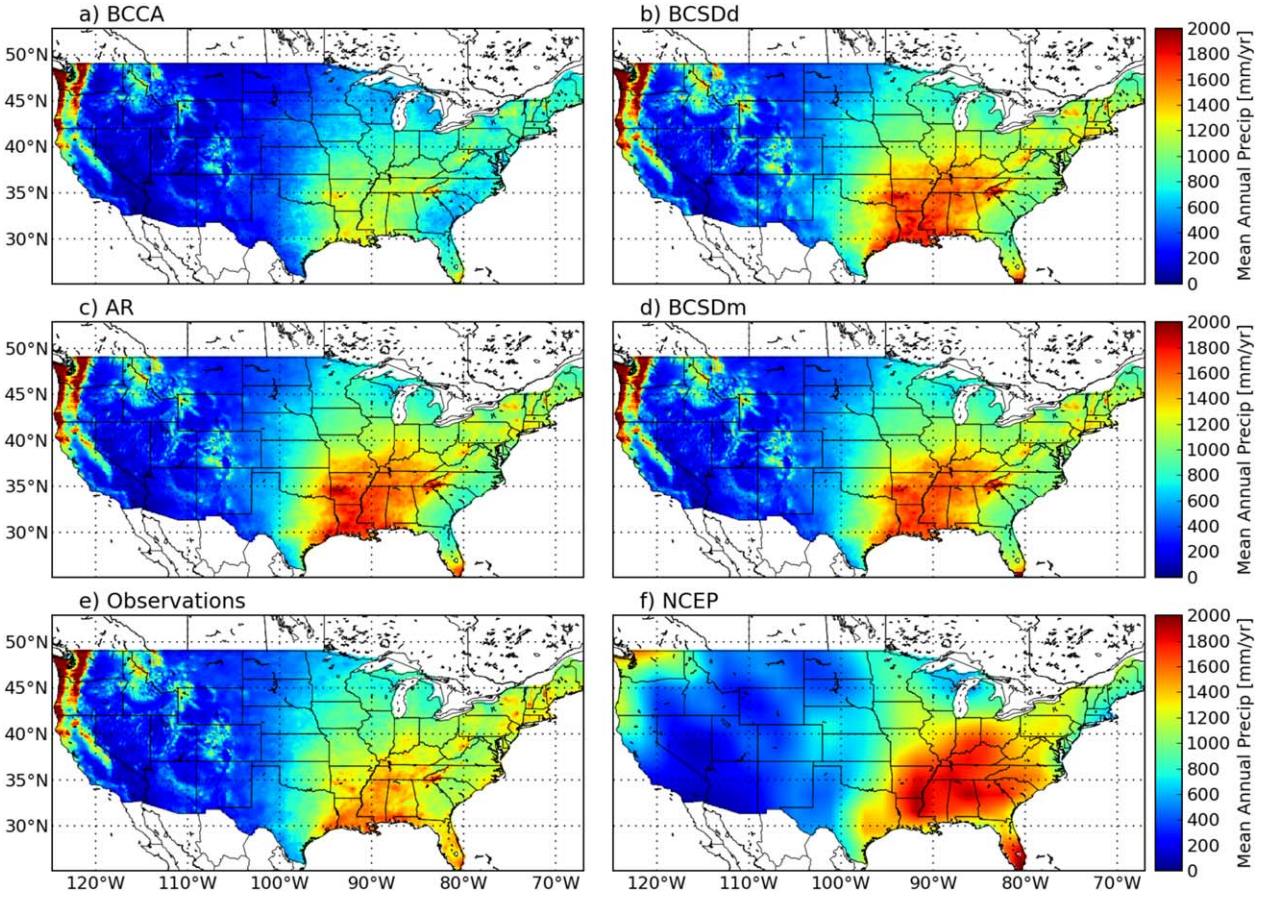


**Figure 2.** (a–d) Maps of mean annual precipitation (mm) simulated by each of the statistical methods, (e) the observations, and (f) NCEP for the period 2001–2008 on a 12 km grid.

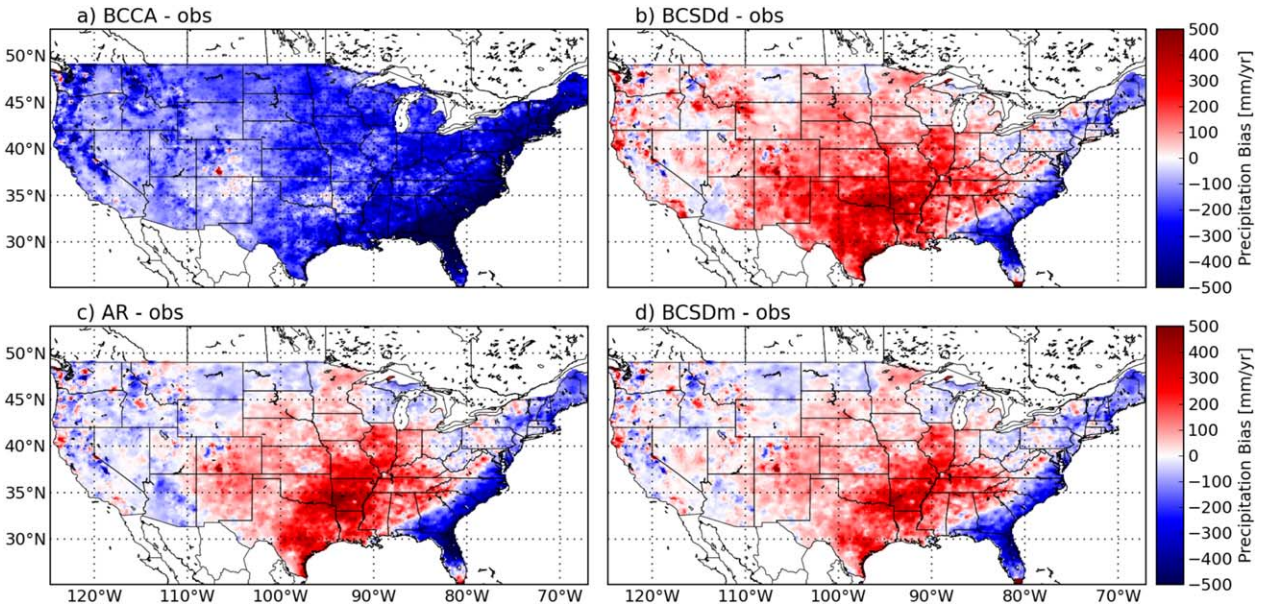


**Figure 3.** Maps of bias in mean annual precipitation (mm) in the validation period (2001–2008) for (a) BCCA, (b) BCSDd, (c) AR, and (d) BCSDm.

This provided a step change in quality of important global atmospheric observations between our calibration and validation periods. Nevertheless, the BCCA method is biased dry even during the calibration period, while other methods are not, which suggests there is a procedural step that results in this bias. BCCAr does not exhibit this bias, instead having a slightly wet bias in each subdomain (NW: 47 mm, SW: 153 mm, CR: 34 mm). *Hwang and Graham* [2013] also found a dry bias in the BCCA method when applied to a smaller region. It is possible that, particularly when fitting a large area, BCCA is likely to select analog days with smoother spatial patterns, resulting in a decrease of larger precipitation events, as discussed in section 5.3, and these larger events will affect mean annual totals substantially.

Comparing maps of bias in the statistically downscaled products over CONUS (Figure 3) with maps of changes in precipitation in the observations and in NCEP (Figure 4) helps suggest an attribution for the bias in some methods. While CONUS-wide biases are relatively minor, the local biases are as large as 400 mm/yr or more and have clear spatial patterns. Because these broad patterns of bias are nearly identical in all methods and match the change in NCEP between calibration and validation periods, it appears that all methods directly inherit mean changes in precipitation from the driving model, and do not correct those changes substantially. It is not clear if those changes are real changes in the mean precipitation between these time periods that is simply represented differently by NCEP and the observations, or if the temporal instability in NCEP data due to variations in ingested data are responsible for the bias.

Similarly, small spatial-scale patterns, for example, the small dry (blue) spot in south central Colorado (37°N, 107°W) in the observation change map (Figure 4), are inversely correlated with the bias maps. While it could

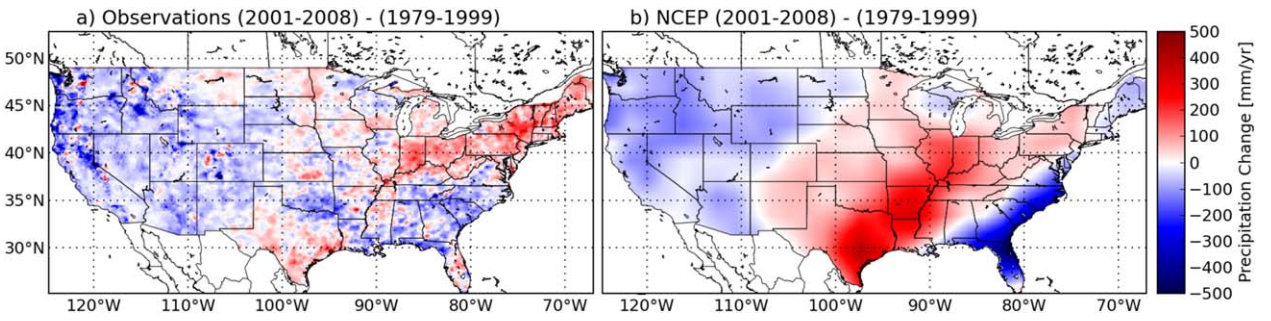


**Figure 4.** (a) Maps of differences in precipitation between the validation and the calibration period for the observations and (b) NCEP.

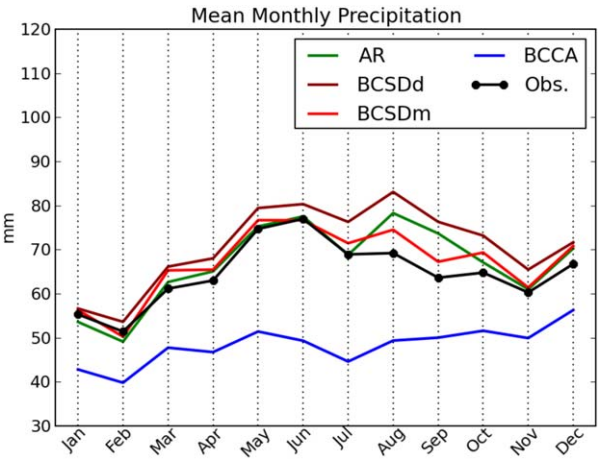


**Figure 5.** Mean monthly precipitation for all methods (colors) compared to observations (black) as averaged over CONUS in the validation period (2001–2008).

be that the local climate has indeed changed over this time period, it seems more likely, given the spatial structure that the observations have changed due to a change in station existence, location, or reporting habits. Lacking any confirmation, this remains a hypothesis in this study; however, observers are human and prone to errors; observer bias in the National Weather Service Cooperative Observer Program network is a known problem, but is difficult to address [*Daly et al*., 2007].

Finally, the larger biases in the AR and BCSDd methods in August and September (Figure 5) suggest that they may be more sensitive to aspects of the driving model's representation of precipitation. Convective precipitation in particular is the dominant form of precipitation in the summer months in many regions. The chaotic nature of convection and the parameterization required in coarse-resolution models means that convective precipitation may not be well represented in coarse-resolution models. Because the BCSDd and AR methods directly rescale daily coarse model precipitation, they are sensitive to changes in the distribution of precipitation within an individual grid cell. Convection in a coarse model will result in many days with a small amount of precipitation, while the fine-scale observations will have fewer wet days with more intense precipitation. This feature leads to a steep slope in the BCSDd quantile mapping and the AR regression; as a result, minor changes in the coarse models simulation of convection can result in large changes in the downscaled precipitation. The BCSDm method only works with monthly coarse model precipitation directly, while the two BCCA methods are not as sensitive to individual grid cells because they fit an analog over a larger region.

### 5.2. Interannual Variability

Maps of the interannual variability are presented in Figure 6. These maps show that all downscaling methods improve interannual variability spatial patterns in the western United States, where large mountain ranges play a key role in the spatial variability of precipitation. However, in the south central and eastern United States, the spatial patterns of the downscaled interannual variability are correlated better with spatial patterns in NCEP than in the observations. In addition, the BCCA method is biased low (101 mm/yr) compared to the observations (142 mm/yr). The BCSD methods are largely unbiased (BCSDm: 132 mm/yr, BCSDd: 139 mm/yr), while the AR method is biased slightly high (166 mm/yr). The seasonal cycle of interannual variability is roughly captured by all methods (Figure 7), although the BCCA has too little seasonality; the AR has increased variability in July, August, and September; and the two BCSD methods are biased low for individual months even though their variability on an annual scale is unbiased. Interannual variation in monthly precipitation does not necessarily aggregate to interannual variation in annual totals because wet and dry months may offset each other in a given year.

Interannual variation is not explicitly corrected by any of the downscaling techniques, and so is essentially inherited directly from the driving coarse-resolution model and is only substantially modified by any scaling provided by the downscaling for mean precipitation. This is particularly evident in the patterns of spatial variability (Figure 6). For example, the mountains in the western United States show increased interannual variability in the downscaled products relative to NCEP, this can easily be explained by the fact that total precipitation in the mountains has to be scaled up considerably. The increase in interannual variability in the AR method in the summer months (Figure 7) appears to be similar to the increased errors in the bias during this time period. We hypothesize that this is again due to issues with scaling summer convection. BCCA is biased low simply because its mean annual precipitation is biased dry, when interannual variability is normalized by mean annual precipitation BCCA is relatively unbiased.

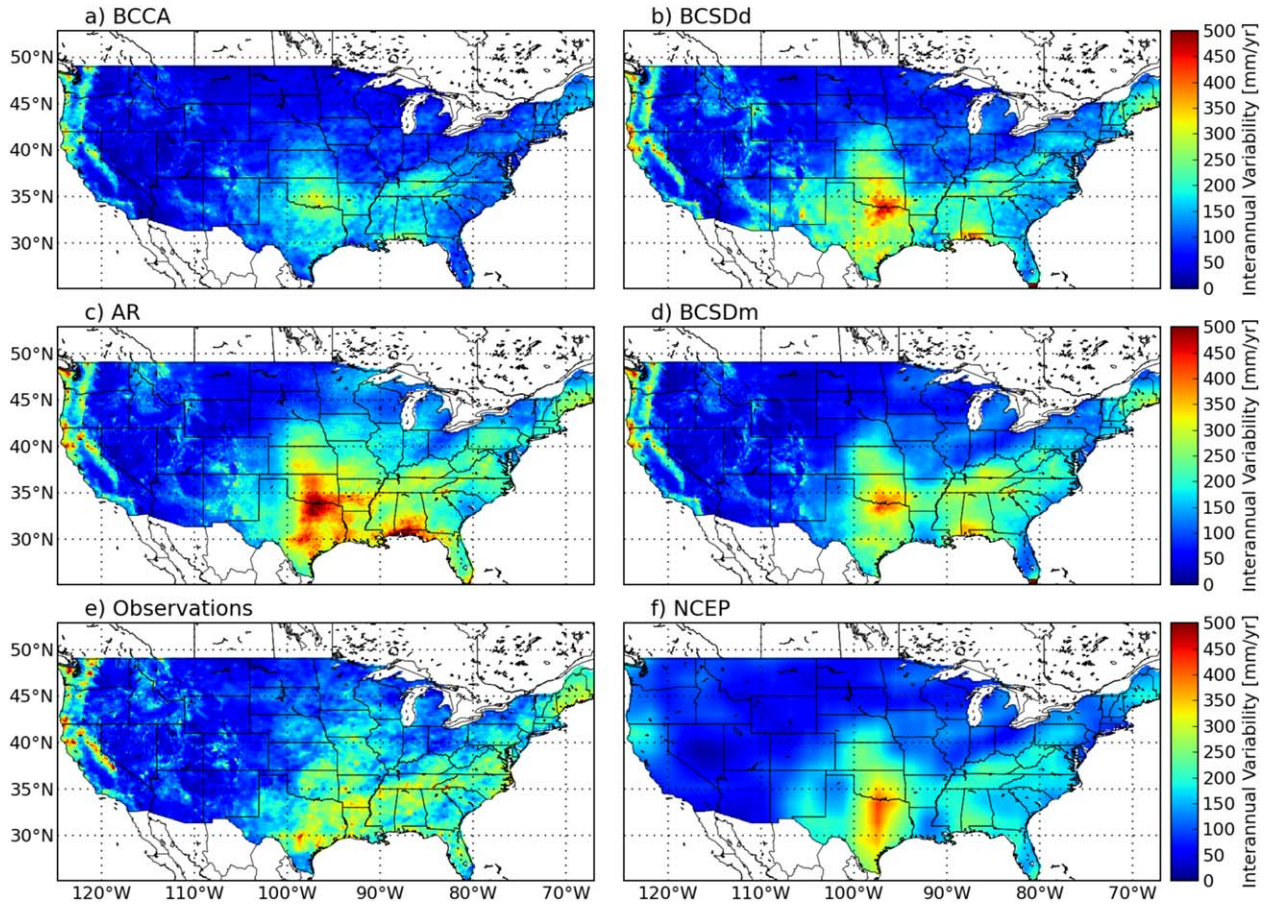


**Figure 6.** (a–d) Maps interannual variability during the validation period (2001–2008) as represented by the standard deviation of annual precipitation for all methods, (e) observations, and (f) NCEP.

### 5.3. Extreme Events

We assess the representation of extreme events by presenting the 50 yr return interval for a 1 day event across spatial scales (Figure 8). Results are similar for the 2, 10, and 100 yr return interval storms and for 2–5 day event totals (supporting information S1). The BCCA method substantially underestimates the more extreme precipitation values (70 mm/d versus 136 mm/d for observations) especially at the grid cell scale. The BCSDm method is relatively unbiased (140 mm/d) and changes properly as a function of scale. The AR method is biased slightly high (149 mm/d) at the individual grid-cell scale; however, it does not change correctly when aggregated to coarser scales and is biased higher at all larger scales. The BCSDd method is biased low at the grid cell scale (90 mm/d). All values are significantly different from observations ($p < 0.01$). In addition, BCSDd does not change appropriately as a function of scale. As a result, BCSDd is unbiased at the coarsest scale. The same general pattern is evident over the subdomains (Figure 8), but here we can compare BCCAr as well. BCCAr is exhibits different biases in each subdomain (NW:

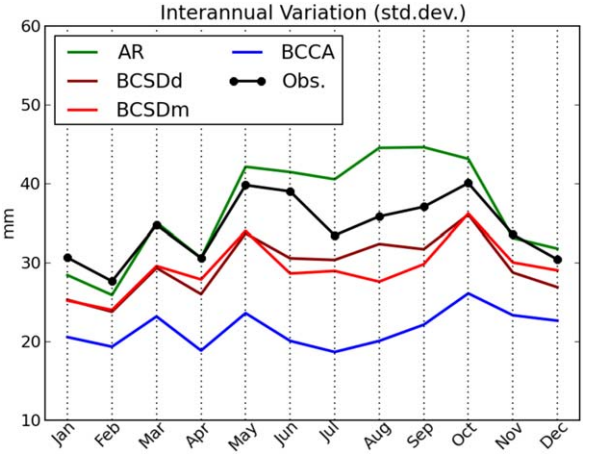


**Figure 7.** Seasonality of interannual variability of monthly totals for all methods (colors) compared to observations (black) as averaged over CONUS in the validation period (2001–2008).

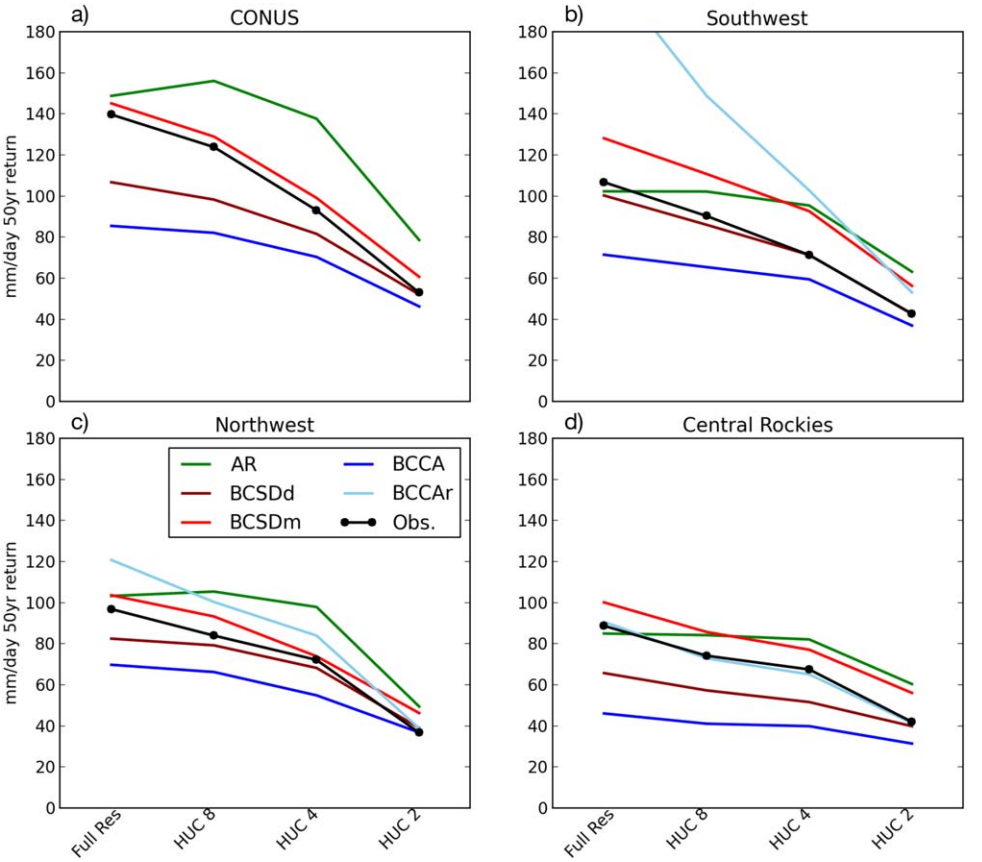


**Figure 8.** (a) Change in representation of extreme events across scales for all methods (color) and observations (black) averaged over CONUS and (b–d) the subdomains in the validation period (2001–2008).

24 mm/d, SW: 105 mm/d, CR: 2 mm/d) at the grid cell scale though it scales more appropriately than BCCA in all regions. BCCA is biased low in all subdomains. Interestingly, BCSDd exhibits smaller biases in the Northwest (−14 mm/d) and Southwest (−6 mm/d) subdomains, while BCSDm is biased high (21 mm/d) in the Southwest subdomain and, to a lesser extent, in the Central Rockies subdomain (11 mm/d).

The BCSDd and AR methods both directly scale the coarse-resolution precipitation, thus, when an extreme event occurs in the coarse-resolution model grid cell, an extreme event will occur at every downscaled grid cell within that coarse-resolution grid cell. In reality, the largest storm basin average will be spatially heterogeneous within the basin. This creates the spatial scaling problems in these methods. In contrast, the BCSDm method selects an arbitrary day in the past, which will have a realistic pattern of spatial variability. However, when rescaling that precipitation, it can lead to local biases as seen when averaging over smaller areas in the subdomains in BCSDm, while AR is more consistent across subdomains.

Similarly, the BCCAr method changes scale more reliably than AR (Figure 8) because it uses real spatial patterns from the past. The dry bias in BCCA occurs because fitting a precipitation map over a large area will tend to select smoother fields as analog days. Compounding this problem, the BCCA method averages multiple analog days together to construct a new analog; as a result, it will inherently smooth out spikes, which might be present in some analog days, or may be in different (subgrid) locations in each of the analogs. However, over small regions, BCCAr can have nearly as many coarse model grid cells (∼50) to fit as it has degrees of freedom (30 analogs), as such it can easily overfit the coarse model output and end up with unrealistic scaling artifacts. For example, on 1 day during the calibration period for the Southwest domain, BCCAr had a weight of −42 applied to one analog and a compensatory weight of +44 applied to another analog, analog weights are expected to sum to a value near 1 and are typically between 0 and 1. At the coarse resolution, these weights result in a near perfect fit to the coarse model, but these two analogs did not have precipitation on the same fine-scale grid cells, as a result precipitation magnitudes in excess of 1000 mm/d were generated, negative values were discarded.

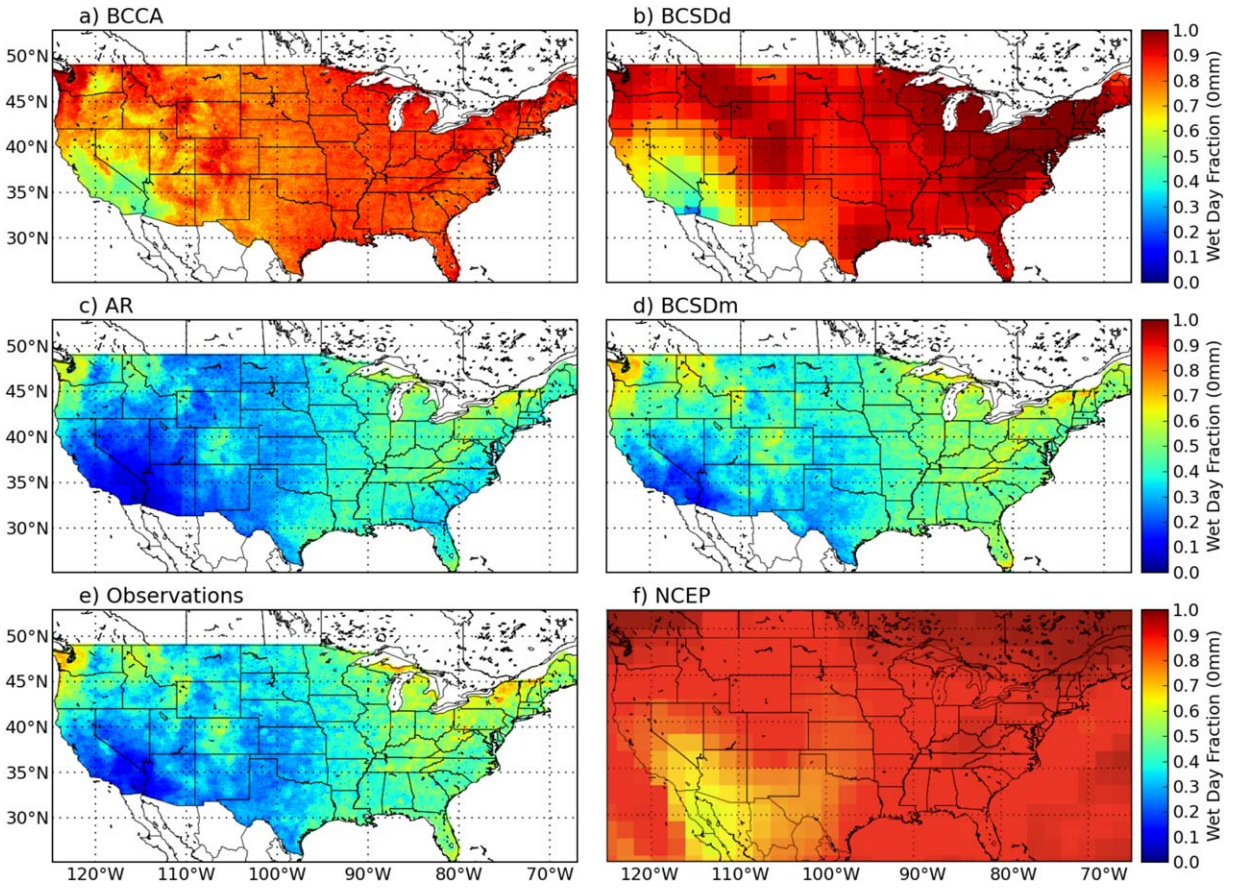


**Figure 9.** (a–d) Wet day fraction (0 mm threshold) maps from all methods (e) observations and directly from (f) NCEP in the validation period (2001–2008).

The BCSDd method tends to produce extreme values that are too small at the individual grid cell level because the most extreme days are only bias corrected at the coarse model grid-cell scale. When these coarse-resolution data are disaggregated to each individual grid cell, the BCSDd method is unable to recreate local hot spots of precipitation necessary to recreate extreme events. However, as the spatial averaging domain is increased to a scale comparable to the coarse grid cell, that evenly distributed event aggregates into a fairly extreme event over the entire basin. Of note, in some subdomains these local hot spots are less important, as a result, BCSDd performs better.

### 5.4. Wet Day Fraction

Wet day frequency is important when downscaled data are used as input to hydrologic models. Precipitation occurrence is used as a predictor in hydrologic models that use an algorithm similar to MTCLIM [*Hungerford et al*., 1989] to calculate solar radiation as in the Variable Infiltration Capacity (VIC) model. Specifically, solar radiation is decreased to 75% on wet days. Precipitation occurrence is also used when calculating humidity with MTCLIM. As a result, biases in wet day fraction can severely affect evapotranspiration calculations in such models. Here we show results both from the 0 mm threshold used in MTCLIM and the 1 mm threshold defined by CLIVAR as a wet day.

To illustrate methodological performance at reproducing wet day fractions, we present maps for all methods with a 0 mm threshold (Figure 9), as well as a plot of mean wet day fraction as a function of scale for both a 0 and 1 mm threshold (Figure 10). Large biases exist in the BCCA and BCSDd methods representation of wet day fraction particularly for the 0 mm threshold, and our initial discussion focuses on the 0 mm threshold. The observed wet day fraction is 0.39, while the BCCA and BCSDd methods are biased high (0.79 and 0.88, respectively). The BCSDm method is biased slightly high (0.43) while the AR method is biased slightly low (0.34). BCCAr improves slightly compared to BCCA. All values are significantly different from observations ($p < 0.01$).

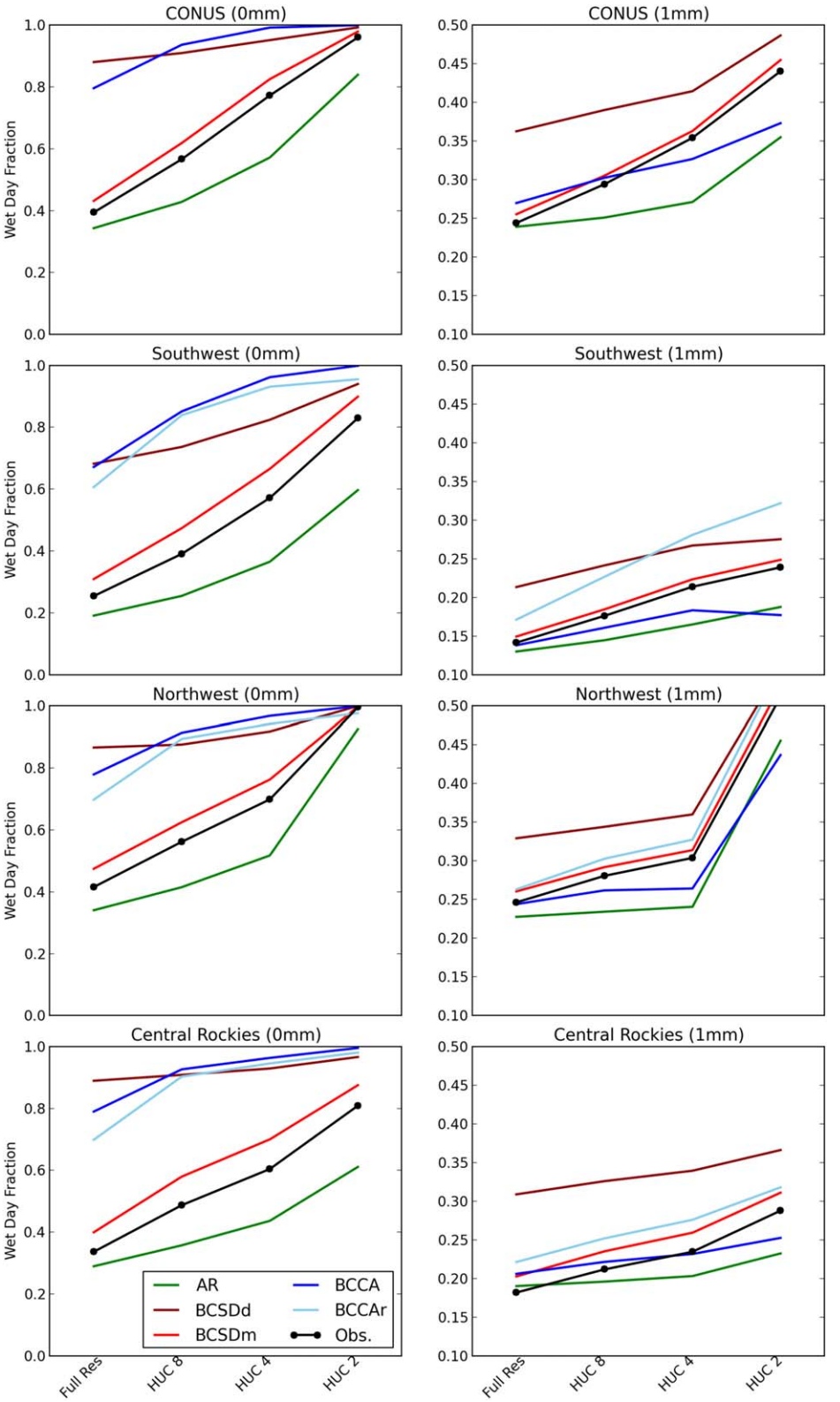


**Figure 10.** (top) Wet day fraction as a function of spatial scale over the CONUS and (bottom) subdomain for each of the five statistical methods downscaled from NCEP (colors) and the 12 km observed product (black) in the validation period (2001–2008). (left) Calculated using a 0 mm threshold, and (right) a 1 mm threshold.

Maps of wet day fraction show substantial spatial structure, most notably, the BCSDd method retains the original NCEP grid structure with only very slight modifications. This grid structure remains because the BCSDd method primarily changes wet day occurrence at the coarse model grid scale, while local values are simply scaled from that coarse model grid cell. In contrast, the BCCA and BCSDm methods resample the historical record and thus obtain spatially variable wet day occurrence. The AR method shows a hint of the NCEP grid structure, for example, there is a vertical line step change at 35°N, 110.5°W, but because it can modify wet day occurrence at the fine grid cell level, it is able to produce realistic variability.

Both BCSDd and the BCCA methods substantially overestimate the wet day fraction at all spatial scales with a 0 mm wet day threshold (Figure 10). For the same reasons that BCCA methods do not produce large extreme events (smoothing of the precipitation field), they also produce too much drizzle. The BCSDd method has two features leading to its increased wet day fraction. First, performing the bias correction step at the coarse model grid scale leads to more wet days than any fine-resolution grid cell would have on its own. This happens because when any fine-resolution grid cell has precipitation, the aggregated coarse grid cell will be forced to have precipitation. When the observations are aggregated to the NCEP grid, many locations have wet day fractions $>$0.9. Second, the bilinear interpolation step increases drizzle because when any of the four surrounding coarse model grid cells have precipitation, every fine grid cell between them will have some amount of precipitation and the spatial disaggregation step will not substantially decrease wet days in fine-resolution grid cells. The AR method avoids similar problems by fitting the precipitation intensity distribution for each fine-resolution grid cell independently. However, it does not properly scale wet day fraction (Figure 10). When precipitation does occur, the AR method typically spreads wet and dry days across an entire coarse grid cell at once, and thus it does not substantially increase the wet day fraction when aggregated to coarser scales.

The BCSDm method is largely immune to problems reproducing wet day frequency because it resamples data from the historical record and only scales the results. However, the BCSDm method will add new wet days to mitigate problems with extreme events. If one of the rescaled precipitation values exceeds 150% of the observed maximum precipitation, the excess precipitation is spread evenly across the remaining days in the month. This preserves the monthly total precipitation and prevents artificially large extreme events from being created, but it leads to a small increase in the wet day fraction. This resampling of the historical record also means that BCSDm cannot change the wet day fraction in future climate even if climate models suggest it should.

The use of a 1 mm threshold, instead of 0 mm, to define a wet day improves the results for the BCCA and BCSDd methods (Figure 10). The observed wet day fraction is 0.24; BCSDd is still biased high (0.36), while the AR method is unbiased (0.24) and the BCCA and BCSDm methods are biased slightly high (0.27 and 0.26, respectively). All differences remain statistically significant ($p < 0.01$). The AR method and the BCSDm method both behave the same as the observations to changes in the threshold used because they match the histogram of the observations (supporting information S1). The scaling behavior of all of these methods is also unchanged when using the 1 mm threshold (Figure 10), but because BCCA has less bias at the grid-cell scale, it ends up biased low at scales greater than HUC-8 because it still does not change scale correctly. BCCAr exhibits a better scaling relationship when using a 1 mm threshold, because it is no longer close to the physical upper bound and thus can increase with scale (Figure 10).

### 5.5. Wet/Dry Spell Length

Hydrologically, increased wet spell lengths can result in wetter soils during rain events and more runoff, increased dry spell lengths can result in drier soils, plant stress, and more demand for irrigation. Here we use a 1 mm threshold to determine wet and dry spell lengths because the MTCLIM threshold is not relevant. The observed average wet spell length is 2.1 days; the BCCA (2.0), BCSDm (2.4), and AR (2.1) methods are relatively unbiased, while the BCSDd method has a substantial high bias (4.5 days). The observed dry spell length is 7.6 days, BCCA (7.4), and BCSDm (7.7) have very little bias, while BCSDd (5.8) is biased low, and AR (8.1) is biased high.

Reproducing the seasonal cycle of wet and dry spell lengths is more challenging for these methods (Figure 11). Wet spell length is nearly unchanged over the year in the observations with a slight peak in May, but the BCCA and AR methods have a maximum in August. The BCSDd method has far too strong a seasonality

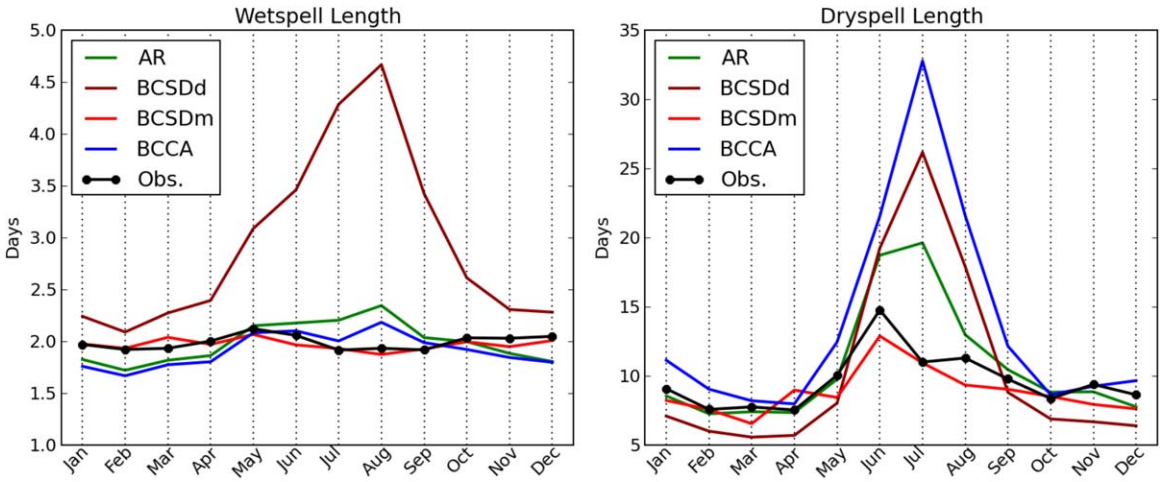


**Figure 11.** (left) Seasonality of wet spell length and (right) dry spell length for each of the statistical methods downscaled from NCEP (colors) and the 12 km observed data set (black) as averaged over CONUS in the validation period (2001–2008).

with a large peak in August, while the BCSDm method exhibits approximately the correct seasonality. The biases in BCCA and AR are unlikely to substantially affect most applications; however, the BCSDd high bias from May to September may have consequences for hydrologic or agricultural applications.

There is a stronger seasonal cycle in the observed dry spell length, and BCSDm is the only method that roughly matches it. The observed cycle shows an increase in May and June that slowly decreases through the year. BCCA, BCSDd, and to a lesser extent AR, all overestimate the seasonal cycle. The AR method overestimates the peak dry spell length in June and July, but is close to the observations for the remainder of the year. We hypothesize that this change in June, July, and August is likely due to problems with the convective parameterization as discussed in the section on bias in mean annual precipitation. The skill of BCSDm in both these metrics may signify a problem in future climate simulations, where it is unable to increase or decrease wet and dry spell lengths, although that may be preferable to potentially tripling the mean dry spell length by using BCCA in July.

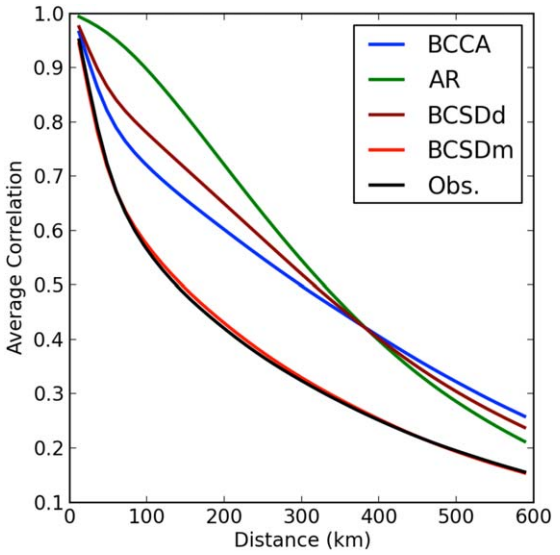


**Figure 12.** Spatiotemporal autocorrelations for all methods (colors) and observations (black) as computed over CONUS in the validation period (2001–2008).

### 5.6. Spatiotemporal Statistics

The geostatistical properties of precipitation are important hydrologically because runoff accumulates over a basin. Spatiotemporal autocorrelograms (Figure 12) reveal that the BCSDm method simulates the geostatistical features of the observations better than other methods, with correlations decreasing sharply from 0.95 to 0.55 over 100 km of separation. All other methods over estimate spatiotemporal correlations for all lag distances. When applied to NCEP, at a lag of 100 km, the AR method has the strongest correlation (>0.9), followed by BCSDd (0.8) and BCCA (>0.7). For lags longer than 400 km, approximately the width of two NCEP grid cells, this relationship is reversed, with the BCCA method having larger correlations and the AR method having lower correlations; all three remain higher than the observations or the BCSDm method. Because the BCSDm method does not attempt to match daily patterns from the driving coarse model, it

**Table 3.** Comparison of Summary Statistics Between Observation Resolutions[a]

| | 50 Year Return Interval (mm/d) | Wet Day Fraction | Wet Spell (Days) | Dry Spell (Days) |
|---|---|---|---|---|
| 6 km | 107 | 0.49 | 4.4 | 9.2 |
| 12 km | 140 | 0.39 | 2.1 | 7.6 |

[a]All statistics calculated in the validation period (2001–2008) on the individual grid cell level on an annual basis and averaged across the entire (CONUS) domain, wet day fraction calculated with a 0 mm threshold.

maintains the correct geostatistical properties across all scales.

### 5.7. Observation Data Set Resolution

We assessed the effect of the resolution of the observation data set using the same metrics discussed previously, and there are significant differences between the 6 km and the 12 km observation data sets (Table 3). Although the mean annual precipitation totals were roughly the same, the 6 km data set had a higher wet day fraction (0.49) compared to the 12 km product (0.39), longer wet spell lengths (6 km: 4.4 days, 12 km: 2.1 days), but also slightly longer dry spell lengths (6 km: 9.2 days, 12 km: 7.6 days). Similarly, the 6 km data set had smaller extreme events; the average 50 yr return interval single day storm totals for the 6 km data set was 102 mm/d, and for 12 km it was 140 mm/d.

These differences are notable because they do not correspond with expected changes. In particular, finer resolution precipitation data sets should be expected to have lower wet day fractions with corresponding shorter wet spell lengths because the probability of precipitation occurrence will be lower for smaller areas, while in a large grid cell there is a higher probability of precipitation occurring somewhere within that grid cell. Finer resolution data sets should also have larger extreme events because these events have not been averaged across a larger area that would be likely to have some areas with less precipitation. While both of these statistics change in the opposite direction from what they should, the changes are not surprising given the methods used to generate these data sets. These data sets are generated by interpolating data between point observations. By interpolating over more grid cells, as is required in the 6 km data set, wet days will typically be added. Similarly, extreme events will typically be strongest in the grid cell containing the observation, and decrease in grid cells where it is interpolated to lower values between observations. Similar issues in gridded data sets were noted in *Gervais et al*. [2014].

Such interpolation artifacts are clearly evident in maps of wet day fraction (0 mm threshold) for the two data sets (Figure 13). These maps have a polk-a-dot appearance, where the dots are located on grid cells that contain station data. This polk-a-dot appearance is strongest in the 6 km product because there is a higher ratio of grid cells to observation points. Figure 13 shows the subdomain to make it easier to see spatial features, but the same patterns are seen across the CONUS. The artifacts described in this section are carried through to any downscaled products based on these data sets. For end-users, these artifacts must be balanced against the added utility of a finer resolution product that is able to better represent important topographically controlled spatial heterogeneity, e.g., colder temperatures and more precipitation at higher elevations.

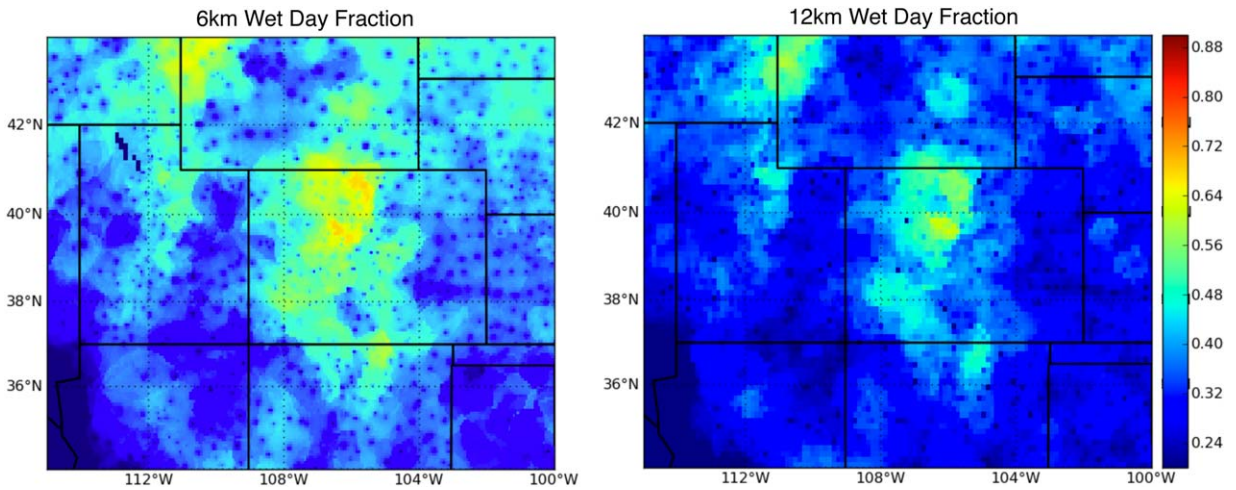


**Figure 13.** (left) Subdomain maps of wet day fraction (0 mm threshold) for the 6 km and the (right) 12 km (right) observed data sets from 1979 to 2008.

## 6. Conclusions

We have presented a comprehensive overview of the performance of four common statistical downscaling methods for the CONUS using a suite of hydrologically relevant measures. These methods were compared to observations in both a validation and calibration period, as applied to the NCEP/NCAR Reanalysis product. Methods were compared by analyzing the bias in means, extreme events, wet day fraction, and wet/dry spell lengths. These comparisons were performed for annual and monthly totals across spatial scales ranging from individual 12 km grid cells to Hydrologic Unit Code (HUC) 8, 4, and 2 regions. This increases our knowledge of one of the major sources of uncertainty in climate simulations identified by *Vano et al*. [2014].

Some important distinctions can be drawn to aid researchers and water resource managers. The BCCA technique applied at a continental scale has some serious deficiencies with a dry bias, decreased extreme precipitation values, and substantial increases in drizzle as was found in *Hwang and Graham* [2013]. Performance on all of these metrics is improved when it is applied on a regional basis (BCCAr), though in some regions BCCAr overestimates extreme events substantially. BCSDd is essentially unbiased, but has problems with increased drizzle and smaller magnitude extreme events. The AR method does well in most statistics at the individual grid-cell scale; however, it produces extreme events that are too large, and too few wet days when aggregated to larger scales as expected by *Maraun* [2013]. The BCSDm technique introduces the fewest artifacts in current climate, but is limited in how much it can change in a future climate compared to other methods because it resamples a month of historical weather at a time, thus limiting its ability to reproduce changes in storm frequency or type in the future. This may also result in discontinuities in the weather sequence between months, which could have negative consequences for some applications. Finally, no method substantially modifies interannual variability; all methods inherit the coarse model interannual variability and only scale it slightly as a side effect of scaling precipitation totals. While temporal instability in the reanalysis and observational products are noted and affect the estimates of bias in all but BCCA, all of the other metrics produce the same results in both the 10 year validation and the 20 year calibration periods, showing that these instabilities do not affect our conclusions, nor does the length of the testing period.

In addition, none of these methods will capture changes in spatial patterns as illustrated in *Gutmann et al*. [2012]. All of the current methods contain assumptions of stationarity in the fine-scale spatial patterns, and thus may have other problems when applied to a climate change scenario. This assumption is likely stronger in methods such as BCSDm, which relies historical weather patterns and sequences directly, than in methods such as BCCA, which construct new spatial patterns. Stationarity assumptions are difficult to assess, but recent work by *Dixon et al*. [2013] provides one approach to the problem.

One additional result of this study is the finding that the 6 km observed precipitation product introduces artifacts when compared to the 12 km product. The additional interpolation required leads to a larger number of wet days, and smaller magnitude of extreme events. Future work needs to improve finer resolution observational products, possibly incorporating spatial variations from radar data or a weather model, but doing so without introducing homogeneity problems in the record is difficult.

The paper is deliberately limited in scope to assess statistical downscaling methods used in some cases to support studies for long-term water resource planning [*Brekke et al*., 2011]—the methods examined in this paper are all based, in one way or another, on rescaling coarse model precipitation, and we do not consider the broader class of statistical downscaling methods, see *Wilby et al*. [1998] or *Fowler et al*. [2007] for a review. The downscaling approaches reviewed here may actually violate a fundamental tenet of statistical downscaling, that is, to use variables reliably simulated by the climate model in a statistical model to provide information at local scales [*Benestad et al*., 2008]. However, these methods require evaluation because of their widespread use [*Barsugli et al*., 2013]. Ongoing work considers methods that make extensive use of information on atmospheric circulation patterns [e.g., *Clark and Hay*, 2004; *Bárdossy and Pegram*, 2011], and computationally efficient precipitation models as in *Jarosch et al*. [2012]. Future work will also look at how different methods modify the climate change signal produced by a climate model; fidelity in current climate does not guarantee a good representation of future climate.

**Acknowledgments**
NCEP Reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/. The *Maurer et al*. [2002] gridded observations of precipitation are available online at http://hydro.engr.scu.edu/files/gridded_obs/daily/ncfiles/. The *Livneh et al*. [2013] gridded observations of precipitation are available online at ftp://ftp.hydro.washington.edu/pub/blivneh/CONUS/. This research has been funded by a cooperative agreement with the United States Bureau of Reclamation (USBR), a contract with the United States Army Corps of Engineers (USACE), and the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation (NSF AGS-0753581). Bridget Thrasher provided code for the BCSD and BCCA downscaling methods. We are grateful to Patrick Laux and three anonymous reviewers for their contributions in making this a stronger manuscript.

## References

Abatzoglou, J. T., and T. J. Brown (2011), A comparison of statistical downscaling methods suited for wildfire applications, *Int. J. Climatol.*, *32*, 772–780, doi:10.1002/joc.2312.

Bárdossy, A., and G. Pegram (2011), Downscaling precipitation using regional climate models and circulation patterns toward hydrology, *Water Resour. Res.*, *47*, W04505, doi:10.1029/2010WR009689.

Barsugli, J. J., et al. (2013), The Practitioner's Dilemma: How to assess the credibility of downscaled climate projections, *Eos Trans. AGU*, *94*(46), 424, doi:10.1002/2013EO460005.

Benestad, R. E., I. Hanssen-Bauer, and E. J. Førland (2007), An evaluation of statistical models for downscaling precipitation and their ability to capture long-term trends, *Int. J. Climatol.*, *27*, 649–665, doi:10.1002/joc.1421.

Benestad, R. E., I. Hanssen-Bauer, and D. Chen (2008), *Empirical-Statistical Downscaling*, 228 pp., World Sci., Singapore.

Brekke, L. D., J. E. Kiang, J. R. Olsen, R. S. Pulwarty, D. A. Raff, D. P. Turnipseed, R. S. Webb, and K. D. White (2009), Climate change and water resources management—A federal perspective, *U.S. Geol. Surv. Circ.*, *1331*, 65 pp. [Available at http://pubs.usgs.gov/circ/1331/.]

Brekke, L., et al. (2011), *Addressing Climate Change in Long-Term Water Resources Management: User Needs for Improving Tools and Information, Civil Works Tech. Ser. CWTS-10-02*, 161 pp., U.S. Army Corps of Eng., Springfield, Va. [Available at http://www.ccawwg.us/index.php/activities/addressing-climate-change-in-long-term-water-resources-planning-and-management.]

Brown, C., Y. Ghile, M. Laverty, and K. Li (2012), Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector, *Water Resour. Res.*, *48*, W09537, doi:10.1029/2011WR011212.

Chen, J., F. P. Brissette, and R. Leconte (2012), Coupling statistical and dynamical methods for spatial downscaling of precipitation, *Clim. Change*, *114*, 509–526, doi:10.1007/s10584-012-0452-2.

Chen, J., F. P. Brissette, D. Chaumont, and M. Braun (2013), Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America, *Water Resour. Res.*, *49*, 4187–4205, doi:10.1002/wrcr.20331.

Clark, M. P., and L. E. Hay (2004), Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, *J. Hydrometeorol.*, *5*, 15–32, doi:10.1175/1525-7541(2004)005<0015:UOMNWP>2.0.CO;2.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical-topographic model for mapping climatological precipitation over mountainous Terrain, *J. Appl. Meteorol.*, *33*, 140–158, doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Daly, C., W. P. Gibson, G. H. Taylor, M. K. Doggett, and J. I. Smith (2007), Observer bias in daily precipitation measurements at United States cooperative network stations, *Bull. Am. Meteorol. Soc.*, *88*, 899–912, doi:10.1175/BAMS-88-6-899.

Dettinger, M. D., D. R. Cayan, M. K. Meyer, and A. E. Jeton (2004), Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American River basins, Sierra Nevada, California, 1900–2099, *Clim. Change*, *62*, 283–317, doi:10.1023/B:CLIM.0000013683.13346.4f.

Dibike, Y. B., and P. Coulibaly (2005), Hydrologic impact of climate change in the Saguenay watershed: Comparison of downscaling methods and hydrologic models, *J. Hydrol.*, *307*, 145–163, doi:10.1016/j.jhydrol.2004.10.012.

Dixon, K. W., K. Hayhoe, J. Lanzante, A. M. K. Stoner, and A. Radhakrishnan (2013), Examining the stationarity assumption in statistical downscaling of climate projections: Is past performance an indication of future results?, paper presented at 93rd American Meteorological Society Annual Meeting (Presentation: TJ15.4 AMS 2013), Am. Meterol. Soc., Austin, Tex., 6–10 Jan.

Fasbender, D., and T. B. M. J. Ouarda (2010), Spatial Bayesian model for statistical downscaling of AOGCM to minimum and maximum daily temperatures, *J. Clim.*, *23*, 5222–5242, doi:10.1175/2010JCLI3415.1.

Fatichi, S., V. Y. Ivanov, and E. Caporali (2011), Simulation of future climate scenarios with a weather generator, *Adv. Water Res.*, *34*, 448–467, doi:10.1016/j.advwatres.2010.12.013.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007), Review: Linking climate change modeling to impacts studies: Recent advances in downscaling techniques for hydrological modeling, *Int. J. Climatol.*, *27*, 147–178.

Gervais, M., L. B. Tremblay, J. R. Gyakum, and E. Atallah (2014), Representing extremes in a daily gridded precipitation analysis over the United States: Impacts of station density, resolution, and gridding methods, *J. Clim.*, *27*, 5201–5218, doi:10.1175/JCLI-D-13-00319.1.

Glahn, H. R., and D. A. Lowry (1972), The use of model output statistics (MOS) in objective weather forecasting, *J. Appl. Meteorol.*, *11*, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gutmann, E. D., R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson (2012), A comparison of statistical and dynamical downscaling of winter precipitation over complex Terrain, *J. Clim.*, *25*, 262–281, doi:10.1175/2011JCLI4109.1.

Hanson, R. T., L. E. Flint, A. L. Flint, M. D. Dettinger, C. C. Faunt, D. Cayan, and W. Schmid (2012), A method for physically based model analysis of conjunctive use in response to potential climate changes, *Water Resour. Res.*, *48*, W00L08, doi:10.1029/2011WR010774.

Hay, L., J. LaFontaine, and S. Markstrom (2014), Evaluation of statistically downscaled GCM output as input for hydrological and stream temperature simulation in the Apalachicola-Chattahoochee-Flint River Basin (1961–1999), *Earth Interact.*, *18*, 1–32, doi:10.1175/2013EI000554.1.

Hidalgo, H., M. Dettinger, and D. Cayan (2008), Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States, *Rep. CEC-500-2007-123*, Calif. Energy Comm., PIER Energy-Related Environ. Res., Sacramento, Calif.

Holloway, C. E., S. J. Woolnough, and G. M. S. Lister (2012), Precipitation distributions for explicit versus parametrized convection in a large-domain high-resolution tropical case study, *Q. J. R. Meteorol. Soc.*, *138*, 1692–1708, doi:10.1002/qj.1903.

Hungerford, R. D., R. R. Nemani, S. W. Running, and J. C. Coughlan (1989), MTCLIM: A mountain microclimate simulation model, *Res. Pap. INT-414*, U. S. Dep. of Agric. For. Serv. Intermountain Res. Stn., Ogden, Utah.

Huth, R., S. Kliegrová, and L. Metelka (2008), Non-linearity in statistical downscaling: Does it bring an improvement for daily temperature in Europe?, *Int. J. Climatol.*, *28*, 465–477, doi:10.1002/joc.1545.

Hwang, S., and W. D. Graham (2013), Development and comparative evaluation of a stochastic analog method to downscale daily GCM precipitation, *Hydrol. Earth Syst. Sci. Discuss.*, *10*, 2141–2181, doi:10.5194/hessd-10-2141-2013.

Jarosch, A. H., F. S. Anslow, and G. K. C. Clarke (2012), High-resolution precipitation and temperature downscaling for glacier models, *Clim. Dyn.*, *38*, 391–409, doi:10.1007/s00382-010-0949-1.

Journel, A. G., and Ch. J. Huijbregts (1978), *Mining Geostatistics*, Academic Press, San Francisco, Calif.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-Year Reanalysis Project, *Bull. Amer. Meteorol. Soc.*, *77*, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Katz, R., and M. Parlange (1995), Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*, 1331–1341, doi:10.1029/94WR03152.

Katz, R. W. (1999), Extreme value theory for precipitation: sensitivity analysis for climate change, *Adv. Water Resour.*, *23*(2), 133–139, doi:10.1016/S0309-1708(99)00017-2.

Livneh, B., E. A. Rosenberg, C. Lin, V. Mishra, K. Andreadis, E. P. Maurer, and D. P. Lettenmaier (2013), A long-term hydrologically based data set of land surface fluxes and states for the conterminous U.S.: Update and extensions, *J. Clim.*, *26*, 9384–9392, doi:10.1175/JCLI-D-12-00508.1.

Maraun, D. (2013), Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue, *J. Clim.*, *26*, 2137–2143, doi:10.1175/JCLI-D-12-00821.1.

Maraun, D., et al. (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.

Maurer, E., A. Wood, J. Adam, D. Lettenmaier, and B. Nijssen (2002), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, *J. Clim.*, *15*(22), 3237–3251, doi:10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2.

Maurer, E. P., and H. G. Hidalgo (2008), Utility of daily vs. monthly large-scale climate data: An intercomparison of two statistical downscaling methods, *Hydrol. Earth Syst. Sci.*, *12*(2), 551–563, doi:10.5194/hess-12-551-2008.

Maurer, E. P., L. Brekke, T. Pruitt, and P. B. Duffy (2007), Fine-resolution climate projections enhance regional climate change impact studies, *Eos Trans. AGU*, *88*(47), 504, doi:10.1029/2007EO470006.

Maurer, E. P., H. G. Hidalgo, and T. Das (2010), The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California, *Hydrol. Earth Syst. Sci.*, *14*, 1125–1138, doi:10.5194/hess-14-1125-2010.

Mehrotra, R., R. Srikanthan, and A. Sharma (2006), A comparison of three stochastic multi-site precipitation occurrence generators, *J. Hydrol.*, *331*(1–2), 280–292, doi:10.1016/j.jhydrol.2006.05.016.

Miller, W. P., G. M. DeRosa, S. Gangopadhyay, and J. B. Valdes (2013), Predicting regime shifts in flow of the Gunnison River under changing climate conditions, *Water Resour. Res.*, *49*, 2966–2974, doi:10.1002/wrcr.20215.

Nicholas, R. E., and D. S. Battisti (2012), Empirical downscaling of high-resolution regional precipitation from large-scale reanalysis fields, *J. Appl. Meteorol. Climatol.*, *51*, 100–114, doi:10.1175/JAMC-D-11-04.1.

Ning, L., M. E. Mann, R. Crane, and T. Wagener (2012), Probabilistic projections of climate change for the Mid-Atlantic Region of the United States: Validation of precipitation downscaling during the historical era, *J. Clim.*, *25*, 509–526, doi:10.1175/2011JCLI4091.1.

Pandey, G. R., D. R. Cayan, M. D. Dettinger, and K. P. Georgakakos (2000), A hybrid orographic plus statistical model for downscaling daily precipitation in Northern California, *J. Hydrometeorol.*, *1*, 491–506, doi:10.1175/1525-7541(2000)001<0491:AHOPSM>2.0.CO;2.

Panofsky, H. W., and G. W. Brier (1968), *Some Applications of Statistics to Meteorology*, 224 pp., Pennsylvania State Univ. Press, Philadelphia, Pa.

Raff, D. A., T. Pruitt, and L. D. Brekke (2009), A framework for assessing flood frequency based on climate projection information, *Hydrol. Earth Syst. Sci.*, *13*, 2119–2136, doi:10.5194/hess-13-2119-2009.

Rajagopalan, B., and U. Lall (1999), A k-nearest neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*, 3089–3101, doi:10.1029/1999WR900028.

Randall, D., M. Khairoutdinov, A. Arakawa, and W. Grabowski (2003), Breaking the cloud parameterization deadlock, *Bull. Am. Meteorol. Soc.*, *84*, 1547–1564, doi:10.1175/BAMS-84-11-1547.

Rasmussen, R. M., et al., (2014), Climate change impacts on the water balance of the Colorado Headwaters: High-resolution regional climate model simulations, *J. Hydrometeorol.*, *15*, 1091–1116, doi:10.1175/JHM-D-13-0118.1.

Reclamation (2011), Colorado River Basin Water Supply and Demand Study, Interim Report No. 1, Status Report, Bureau of Reclamation, US Department of the Interior, Boulder City, Nev., doi:10.5194/hess-16-3989-2012.

Salathe, E. P. (2005), Downscaling simulations of future global climate with application to hydrologic modeling, *Int. J. Climatol.*, *25*, 419–436, doi:10.1002/joc.1125.

Schmidli, J., C. Frei, and P. L. Vidale (2006), Downscaling from GCM precipitation: A benchmark for dynamical and statistical downscaling methods, *Int. J. Climatol.*, *26*, 679–689, doi:10.1002/joc.1287.

Schoof, J. T. (2013), Statistical downscaling in climatology, *Geogr. Compass*, *7*, 249–265, doi:10.1111/gec3.12036.

Seaber, P. R., F. P. Kapinos, and G. L. Knapp (1987), Hydrologic unit maps, *U.S. Geol. Surv. Water Supply Pap.*, *2294*, 63 pp.

Stoner, A., K. Hayhoe, and X. Yang (2012), An asynchronous regional regression model for statistical downscaling of daily climate variables, *Int. J. Climatol.*, *33*, 2473–2494, doi:10.1002/joc.3603.

Thrasher, B., E. P. Maurer, C. McKellar, and P. Duffy (2012), Technical Note: Bias correcting climate model simulated daily temperature extremes with quantile mapping, *Hydrol. Earth. Syst. Sci.*, *16*, 3309–3314, doi:10.5194/hess-16-3309-2012.

Trenberth, K. E., J. T. Fasullo, and J. Mackaro (2011), Atmospheric moisture transports from ocean to land and global energy flows in reanalyses, *J. Clim.*, *24*, 4907–4924, doi:10.1175/2011JCLI4171.1.

Vano, J. A., et al. (2014), Understanding uncertainties in Future Colorado river streamflow, *Bull. Am. Meteorol. Soc.*, *95*, 59–78, doi:10.1175/BAMS-D-12-00228.1.

Vrac, M., M. Stein, and K. Hayhoe (2007), Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, *Clim. Res.*, *34*(3), 169–184, doi:10.3354/cr00696.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp (2008), Experiences with 0–36-h Explicit Convective Forecasts with the WRF-ARW Model, *Weather Forecasting*, *23*, 407–437, doi:10.1175/2007WAF2007005.1.

Wetterhall, F., S. Halldin, and C.-Y. Xu (2007), Seasonality properties of four statistical-downscaling methods in central Sweden, *Theor. Appl. Climatol.*, *87*(1–4), 123–137, doi:10.1007/s00704-005-0223-3.

Wilby, R., T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks (1998), Statistical downscaling of general circulation model output: A comparison of methods, *Water Resour. Res.*, *34*, 2995–3008, doi:10.1029/98WR02577.

Wilby, R. L. (1994), Stochastic weather type simulation for regional climate change impact assessment, *Water Resour. Res.*, *30*, 3395–3403, doi:10.1029/94WR01840.

Wilby, R. L. (1998), Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices, *Clim. Res.*, *10*, 163–178.

Wilby, R. L., L. E. Hay, W. J. Gutowski, R. W. Arritt, E. S. Takle, Z. Pan, G. H. Leavesley, and M. P. Clark (2000), Hydrological responses to dynamically and statistically downscaled climate model output, *Geophys. Res. Lett.*, *27*(8), 1199–1202, doi:10.1029/1999GL006078.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns (2004), Guidelines for use of climate scenarios developed from statistical downscaling methods, technical report, *Intergovt. Panel on Clim. Change*, Geneva, Switzerland.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191, doi:10.1016/S0022-1694(98)00186-3.

Wood, A., L. Leung, V. Sridhar, and D. Lettenmaier (2004), Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs, *Clim. Change*, *62*, 189–216.

Yoon, J.-H., L. Ruby Leung, and J. Correia Jr. (2012), Comparison of dynamically and statistically downscaled seasonal climate forecasts for the cold season over the United States, *J. Geophys. Res.*, *117*, D21109, doi:10.1029/2012JD017650.

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers (2011), Indices for monitoring changes in extremes based on daily temperature and precipitation data, *WIREs Clim. Change*, *2*, 851–870, doi:10.1002/wcc.147.