AGU100 ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL086689

Bias Correction of Paleoclimatic Reconstructions: A New Look at 1,200+ Years of Upper Colorado River Flow

Scott M. Robeson[1], Justin T. Maxwell[1], and Darren L. Ficklin[1]

[1]Department of Geography, Indiana University, Bloomington, IN, USA

Key Points:
- Bias correction makes paleoclimatic models more comparable to observations
- After bias correction, the 1100s Colorado River megadrought became more extreme
- The early 1600s Colorado River pluvial rivals that of the early twentieth century after bias correction

Supporting Information:
- Supporting Information S1

Correspondence to:
S. M. Robeson,
srobeson@indiana.edu

Citation:
Robeson, S. M., Maxwell, J. T., & Ficklin, D. L. (2020). Bias correction of paleoclimatic reconstructions: A new look at 1,200+ years of Upper Colorado River flow. *Geophysical Research Letters*, *47*, e2019GL086689. https://doi.org/10.1029/2019GL086689

Received 16 DEC 2019
Accepted 20 DEC 2019
Accepted article online 3 JAN 2020

Abstract Bias correction, while widely used with climate-model output, is not typically applied to paleoclimatic reconstructions. While many reconstruction models have low average error, they still may contain bias, especially in the tails of distributions. Bias correction, used cautiously, can be a valuable procedure that alters interpretations of past events. Analyzing the iconic tree-ring reconstruction of Upper Colorado River flow, we find that its probability distribution is markedly different from that of observed flow. Using quantile mapping to bias correct the reconstruction, we analyze the full reconstructed record and two events in particular: the 1100s megadrought and the early 1600s pluvial. Overall, bias correction made the 1100s megadrought, the largest in the 1,200+year record, even more extreme. After bias correction, the early 1600s pluvial marginally exceeds the early twentieth century pluvial in magnitude but not in duration. Overall, bias correction should be considered whenever paleoclimatic reconstructions are compared directly to observations.

Plain Language Summary To study past climates, scientists use indicators such as tree-ring widths that are related to temperature, precipitation, or streamflow. While these indicators usually are very reliable, they sometimes do not perform as well with extreme events such as intense droughts and wet periods. Here, we adopt a method that can correct for these limitations and apply it to a 1,200+year record of streamflow for the Upper Colorado River, a critical source of water for much of the southwestern United States. After using our method, we find that several extreme events from the tree-ring record of streamflow were even more intense than formerly thought. In particular, the largest drought in the record that occurred during the 1100s was drier and longer lasting after our correction. During the 56-year duration of the 1100s drought, our correction makes the flow in the river lower by nearly 52×10^9 m^3 of water, which is the equivalent of 1.45 times the capacity of Lake Mead (the largest reservoir in the United States). And, while it was known that the early 1900s was among the wettest periods in the last 1,200+ years, we identify a period in the early 1600s that matches it.

1. Introduction

Reconstructions using paleoclimatic proxies are among the most important tools for studying past climates. Information from tree rings, in particular, has been used to reliably reconstruct climate variables such as temperature, precipitation, soil moisture, and streamflow on centennial to millennial scales (Cook et al., 1999; Esper et al., 2002; Oliver et al., 2019; Stahle & Cleaveland, 1992; Wilson et al., 2016; Wise, 2010; Woodhouse & Overpeck, 1998). Dendroclimatic reconstructions provide information at higher temporal resolution than most other proxies, allowing modern climatic events to be placed into a longer-term context (Cook et al., 1999; Griffin & Anchukaitis, 2014; Routson et al., 2011; Stahle et al., 2007). The reliability of reconstruction models, therefore, is of great interest and errors in reconstruction models have been carefully studied (Buras, 2017; Wigley et al., 1984). Potential biases in reconstruction models, however, are not as routinely analyzed (Robeson, 2015).

An important tool in climate modeling and downscaling, bias correction is used to adjust climate model output so that biases estimated during the control period are removed from future projections (Abatzoglou & Brown, 2012; Berg et al., 2012; Ficklin et al., 2016; Maraun, 2013; Teutschbein & Seibert, 2012). For instance, a climate model that systematically runs too hot or too cold during the observed period would be expected to have the same biases when projecting future climate conditions. While adjustments to probability distribution parameters (e.g., mean and standard deviation) may address some aspects of bias in reconstruction models, more sophisticated methods such as quantile mapping often are needed

(Gudmundsson et al., 2012; Robeson, 2015). Given that extreme events such as droughts and pluvials often are the focus of paleoclimatic reconstructions (Cook, Seager et al., 2015; Griffin & Anchukaitis, 2014; Pederson et al., 2013; Routson et al., 2011; St. George & Nielsen, 2003; Woodhouse et al., 2005) and that reconstruction models have a tendency to underestimate extremes (Esper et al., 2005; Meko, 1997; Meko et al., 2007), quantile mapping is likely to produce reconstructions that more faithfully reproduce these events.

Here, we reanalyze the reconstruction of streamflow in the Upper Colorado River at Lee's Ferry that was originally developed by Stockton and Jacoby (1976) and then substantially updated by Woodhouse et al. (2006) and Meko et al. (2007). The reconstruction shows much more clearly than observations alone that the 1922 Colorado River Compact was negotiated during an exceptionally wet period (Stockton & Jacoby, 1976), resulting in overallocation of regional water resources during much of the twentieth century (MacDonnell et al., 1995; Pulwarty et al., 2005). As water resources in the region will almost certainly be further stressed in the 21st century when warmer and drier conditions (Cook, Ault et al., 2015; Seager et al., 2007) are expected to reduce Colorado River streamflow substantially (Ficklin et al., 2013; McCabe et al., 2017; Udall & Overpeck, 2017), further analysis of this important reconstruction may provide additional insight. In addition to its impact on regional research and policy on water resources, the legacy of the set of reconstructions also includes inspiring a generation of researchers to explore past streamflow variability and extremes in a wide array of other environments (e.g., Axelson et al., 2009; Coulthard et al., 2016; Graumlich et al., 2003; Harley et al., 2017; Margolis et al., 2011; Maxwell et al., 2017; Wise, 2010). In our analysis of the Meko et al. (2007) Upper Colorado River reconstruction, we use a quantile-mapping procedure that corrects the probability distribution of the reconstruction model to better match that of the observed data. We then use the full bias-corrected reconstruction to assess the overall behavior of the streamflow time series as well as a pair of important extreme events in the reconstructed record: the 1100s megadrought and the early 1600s pluvial.

2. Methods

The approach used here is adapted from how bias correction is used with regional or global climate-model output (Christensen et al., 2008; Gudmundsson et al., 2012; Li et al., 2010; Teutschbein & Seibert, 2012), except that a paleoclimatic reconstruction is the model of interest. When evaluating bias, the goal is to estimate how faithfully the cumulative distribution function (cdf) of modeled values reproduces the cdf of observed values during the time period where they overlap. When empirical quantiles of the observed and modeled values are used to determine the cdfs and their differences form the basis for adjusting the modeled values, the procedure is referred to as "quantile mapping" (Chen et al., 2013; Gudmundsson et al., 2012; Teutschbein & Seibert, 2012). This name is apt because the paired quantiles (the differences in the inverse cdfs) form the "mapping" from unadjusted to adjusted data. While quantile mapping has been widely used with climate-model output, the tendency of statistical paleoclimatic models to underestimate variance has previously been quantified and evaluated using additive noise models (e.g., Biondi & Meko, 2019; Meko et al., 2001) and ensemble approaches (Gangopadhyay et al., 2009, 2015).

The specific quantile mapping approach used here is the "RQUANT" method implemented in the *qmap* package (Gudmundsson et al., 2012) in R (R Core Team, 2018). RQUANT uses local, linear regressions on sequences of regularly spaced quantiles to approximate the quantile-quantile relationship. Within the paired quantile-quantile space, the bias for a particular quantile in the reconstruction is estimated using the difference between the local regression and the same (interpolated) quantile from the observations. That bias estimate, positive or negative, can then be added to the quantile from the reconstruction. This method is easy to implement and produces a close match between the observed and modeled cdf while not preserving every individual difference in the quantiles, so the potential for overfitting is reduced. When using bias correction with values whose range extends beyond those of the reconstruction during the observational period, care must be taken to ensure that the extrapolated bias correction does not produce values that are unrealistic or overinterpreted (Boé et al., 2007). Similar to other statistical approaches, quantile mapping typically assumes temporal stationarity of the relationship (Maraun, 2013; Maraun et al., 2010). In this case, the stationarity assumption means that biases in the reconstruction during the preobservational period are expected to be similar to those during the observed period.

As quantile mapping estimates the cdf empirically, it is a nonparametric approach. Parametric bias correction, in contrast, adjusts the reconstruction to have the same probability distribution parameters as the observed data (e.g., Griffin & Anchukaitis, 2014; Maxwell et al., 2011; Maxwell et al., 2015; Pederson et al., 2012). Parametric approaches implicitly assume an underlying distribution (e.g., Gaussian) and, therefore, are not as robust as quantile mapping. In addition, parametric approaches are relatively inflexible to variations in bias across the empirical cdf. A two-parameter bias correction, for instance, would assume that the bias correction is well modeled by using a single linear transfer function for the entire cdf. By using a substantially larger number of estimates—the empirical quantiles—quantile mapping allows for a more reliable bias correction to be developed (Gudmundsson et al., 2012).

Given its widespread use in dendroclimatology, but despite its limitations (Willmott et al., 2015), we use the Nash-Sutcliffe coefficient of efficiency (*CE*) to summarize overall model error before and after bias correction:

$$CE = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$$

where P_i and O_i are the predicted (reconstructed) and observed values, $\overline{O}$ is the observed mean, and n is the sample size during the observational period (here, 1906–2004, so $n = 99$). We also use the refined index of agreement (d_r) that, unlike *CE*, is based on absolute values and has a finite range from −1 to +1 (Willmott et al., 2012):

$$d_r = 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{2\sum_{i=1}^{n}|O_i - \overline{O}|}, \quad \text{when } \sum_{i=1}^{n}|P_i - O_i| \le 2\sum_{i=1}^{n}|O_i - \overline{O}|$$

$$d_r = \frac{2\sum_{i=1}^{n}|O_i - \overline{O}|}{\sum_{i=1}^{n}|P_i - O_i|} - 1, \quad \text{when } \sum_{i=1}^{n}|P_i - O_i| > 2\sum_{i=1}^{n}|O_i - \overline{O}|$$

To diagnose how much of the total error is represented as bias, the ratio of systematic mean squared error (MSE_s) to total mean squared error (*MSE*) also is used:

$$\frac{MSE_s}{MSE} = \frac{\sum_{i=1}^{n}(\hat{P}_i - O_i)^2}{\sum_{i=1}^{n}(P_i - O_i)^2}$$

where $\hat{P}_i$ is the ordinary least squares regression estimate of P_i on O_i. Developed by Willmott (1981), this approach provides quantitative insight into bias that indices of overall error, such as *CE* and d_r, cannot.

3. Data

We use data from the most recent version (Meko et al., 2007) of the well-known reconstruction of streamflow for the Upper Colorado River at Lee's Ferry, Arizona. As over 90% of the flow in the Colorado River originates upstream of Lee's Ferry (Christensen & Lettenmaier, 2007), the observed and reconstructed records there serve as important indicators for most locations that rely on Colorado River water. Originally developed by Stockton and Jacoby (1976) and then substantially updated by Woodhouse et al. (2006), the Meko et al. (2007) reconstruction extended estimates of streamflow for each water year (1 October to 30 September, with the yearly designation referring to the latter calendar year) back to 762 CE. Over the 1,200+year tree-ring record, Meko et al. (2007) developed 11 site-level streamflow reconstructions and then subsets of these reconstructions were combined using principal components analysis. They then regressed observed streamflow at Lee's Ferry from 1906 to 2004 on the scores of the first principal component of the subsetted site-level reconstructions. As water-year totals for a large river, all observed and

reconstructed flow are reported in billions of cubic meters per year. The observed data and, therefore, the reconstructed streamflow are for "naturalized" streamflow, which are estimates of flow after removing the impacts of human activity such as reservoirs, irrigation-water extractions, and other depletions. Naturalized streamflow is used because it represents interannual climate variability rather than human use of river water.

Many paleoclimatic reconstructions, including that of Meko et al. (2007), rely on "nested" models that are calibrated using different sample data through time (because the full set of tree-ring chronologies are not available over the entire reconstruction period). It is rare, however, to have access to output from all of the nested regression models during the calibration period. Without that information, it would be necessary to make the assumption that the bias of the "full" model (that was used to reconstruct the calibration period) is representative of the other nested models. Given that nested models from earlier periods typically use fewer predictors and are less accurate than the full model, they are likely to have more bias in their cdfs and be less able to reproduce the tails of the distribution. As a result, if it is necessary to use only the full model for bias correction, it is likely to be a conservative estimate of bias for the earlier time periods. Here, Dr. David Meko shared the full time series for each of the four nested models used by Meko et al. (2007). As a result, we were able to develop a separate bias correction for each nested model and then composite the four reconstructions in the same way as the original (i.e., we used Model 1 for 762–1181 and 2003; Model 2 for 1182–1364; Model 3 for 1365–2002; and Model 4 for 2004–2005).

4. Results and Discussion

4.1. Bias Assessment

An examination of the Meko et al. (2007) reconstruction over the calibration period (1906–2004) shows a very close match with observed streamflow (Figure 1; $CE = 0.754$ and $d_r = 0.762$). There is evidence, however, that extremes are not as well modeled as other parts of the distribution (Figure 1a). Twelve of the 13 water years from 1906 to 2004 that had observed flow over 25×10^9 m^3 (approximately the 85th percentile) have reconstructed streamflow that is lower than the observed flow, with 9 of those 12 years being underestimates of more than 2.5×10^9 m^3. So, while the reconstruction of lower Colorado River flow tracks observations remarkably well overall, it appears to consistently underestimate high flows. More surprisingly given the tendency of regression models, the two years with the lowest observed flow (1977 and 2002) have reconstructed values that are 2×10^9 m^3 lower than the observed values (i.e., more extreme). Kernel-density estimates of the probability density functions clearly show how the empirical distribution of the reconstruction differs from that of the observations (Figure 1b). Applying a parametric bias correction that simply adjusts the mean and standard deviation of the reconstruction to match those of the observations improves on the distribution somewhat in the middle and upper tail, but makes the mismatch larger in the lower tail (Figure 1b). The different types of bias that occur in the tails and the middle of the distribution are features that quantile mapping is designed to remedy.

After applying quantile mapping to each of the nested models separately, the bias-corrected reconstruction better matches the observed data during extremely low and high flow events (Figure 2a). A paired scatterplot shows the lack of bias after quantile mapping in that the error becomes much more symmetric about the 1:1 line (Figure 2b). In this instance, overall error, as measured by both the CE and d_r values, is virtually unchanged by the bias-correction procedure ($CE = 0.754$ and $d_r = 0.762$ in the original reconstruction and $CE = 0.768$ and $d_r = 0.770$ after quantile mapping). In general, bias correction has the potential to either substantially increase or decrease unsystematic (random) error, so the result here of virtually unchanging overall error will not occur in all applications of quantile mapping. Bias correction, however, should always reduce systematic error. A quantile-quantile (q-q) plot makes this abundantly clear by showing the bias in the original reconstruction and the similarity of the two distributions after quantile mapping (Figure 2c). Using the ratio of systematic error to total error (MSE_s/MSE) to quantify the overall bias reveals that 25.3% of MSE was systematic in the original reconstruction while just 7.8% of MSE is systematic after quantile mapping.

Quantile mapping estimates bias in a reconstruction model during the calibration period. Here, however, four nested models are composited to produce the full reconstructed time series. As a result, each of those

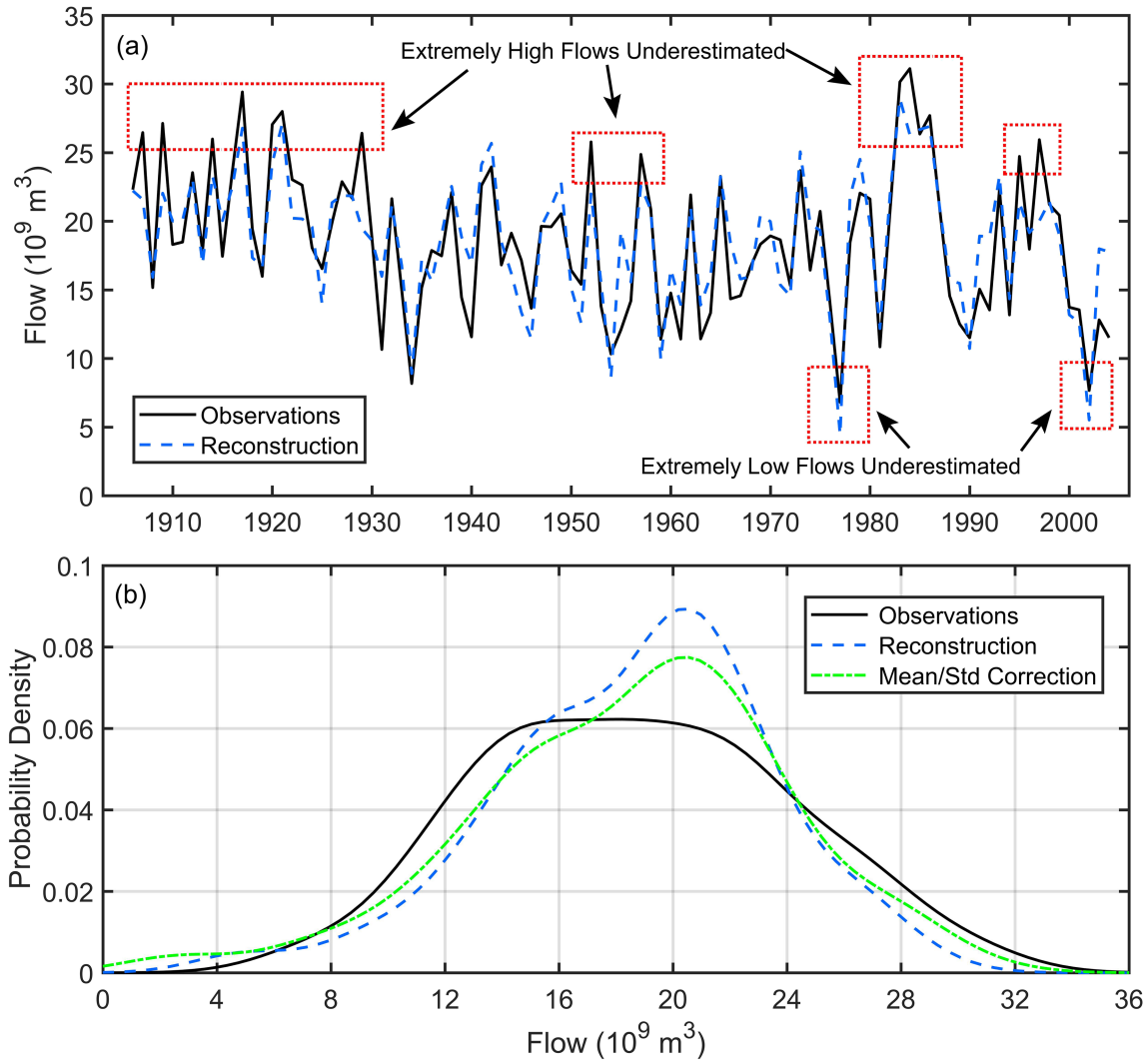

Figure 1. Observed and reconstructed Upper Colorado River flow for 1906–2004 (data from Meko et al., 2007): (a) time series, showing that the reconstruction has high fidelity to the observations but also some potential bias in the most extreme observations and (b) kernel-density estimates of the probability density functions for the observed and reconstructed flow, showing how the probability distribution of the uncorrected reconstruction differs from that of the observations. The common parametric correction of matching the mean and standard deviation of the reconstructed flow to that of the observed flow (green line) also is unable to reproduce the observed distribution.

models will produce different estimates of bias. In this instance, all four nested models have similar bias responses (Figure S1 in the Supporting Information). In particular, quantile mapping adjusted the extremely high flows upward for all of the nested models while adjusting the lower to middle portion of the distributions downward. The one difference between the models is that three of the four models were biased low in the extreme lower tail of the distribution, but Model 1, which is used for the earliest portion of the reconstruction, had much less bias in the extreme lower tail. The weaker response of tree-ring models to extremely wet conditions was not unexpected (e.g., Maxwell et al., 2016), but quantile mapping allows any bias that may be produced by this diminished response to be estimated and accounted for in the reconstruction.

4.2. Impacts on the Full Reconstruction

Moving beyond the impacts of bias correction during the observational period, we evaluate how quantile mapping alters the interpretation of flow during the full reconstruction period from 762 to 2005. To help identify long-term droughts and pluvials and adopting the same strategy as Meko et al. (2007), we use a

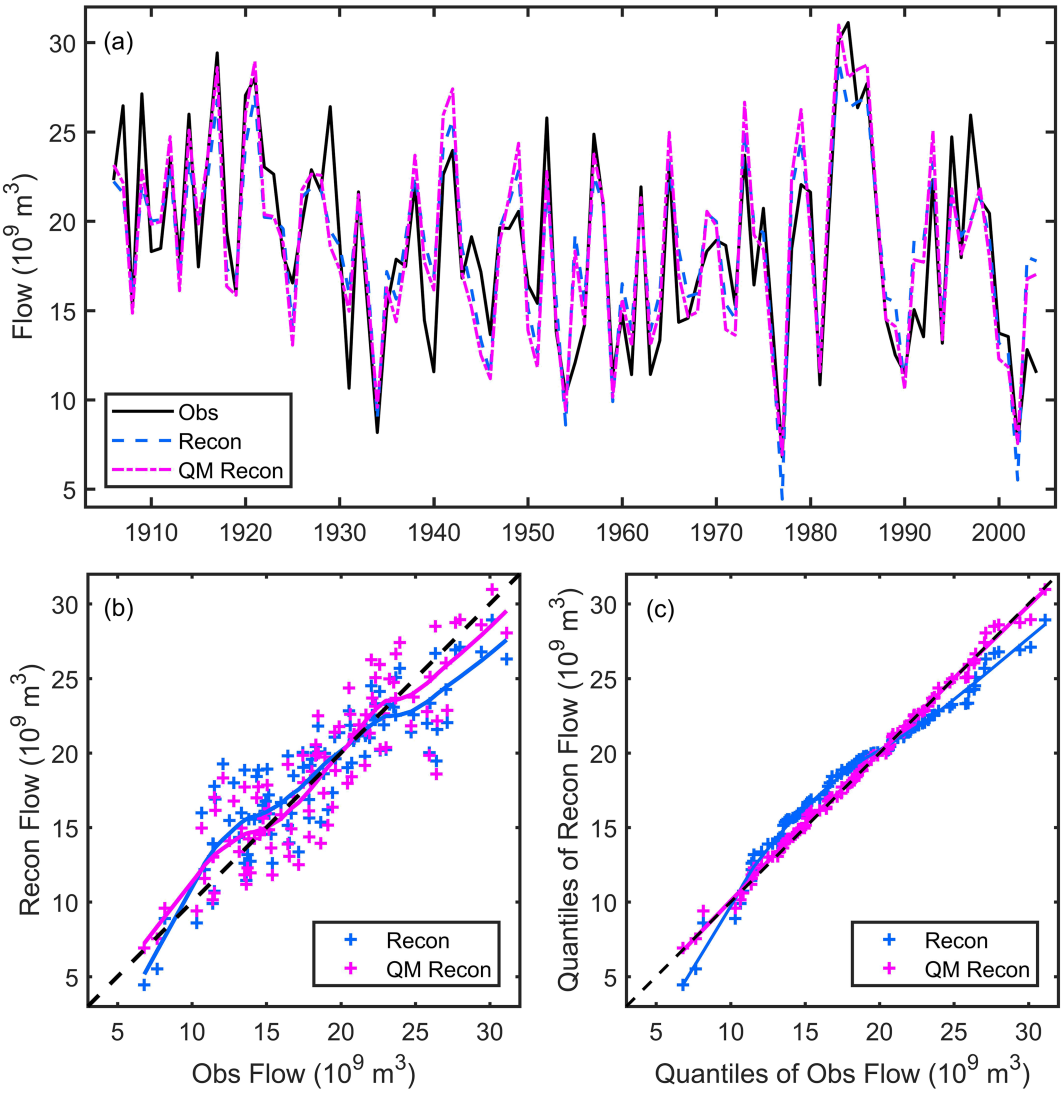

Figure 2. Observed and reconstructed Upper Colorado River flow during the calibration period (1906–2004) before and after quantile mapping (QM): (a) time series, (b) 1:1 scatterplot, and (c) quantile-quantile plot. Locally weighted scatterplot smoothing (LOWESS) curves are used to visualize the degree to which the data in the scatterplot and q-q plot follow the 1:1 line.

centered 25-year moving average on both the original and bias-corrected reconstruction. Like Meko et al. (2007), we give the lowest 25-year average of streamflow during the calibration period (1953 to 1977) as a point of reference. The annual time series for the full 762–2005 reconstruction after bias correction (Figure S2) has nearly the same mean as the Meko et al. (2007) reconstruction (18.07×10^9 m^3 before and 17.91×10^9 m^3 after bias correction). Like those given in Stockton and Jacoby (1976) and Woodhouse et al. (2006), these values are lower than the observed mean during the observed period (18.53×10^9 m^3 for 1906–2004), all of which are well below the annualized 20.35×10^9 m^3 (17.5 million acre feet) that was allocated via the 1922 Colorado River Compact and 1944 Mexican Water Treaty (MacDonnell et al., 1995).

Given that the bias-correction procedure in most of the nested models adjusted the extremely low flows upward (because they were biased low during the calibration period), we expected extreme droughts in the past, such as the megadrought during the 1100s, to become less extreme. Surprisingly, the 1100s drought, which was the most extreme and longest lasting drought in the Meko et al. (2007) reconstruction, became even more severe after bias correction (Figure 3). The most extreme 25-year drought in the original reconstruction (1130 to 1154) decreased substantially from 84.0% of the 1906–2004 average in the original

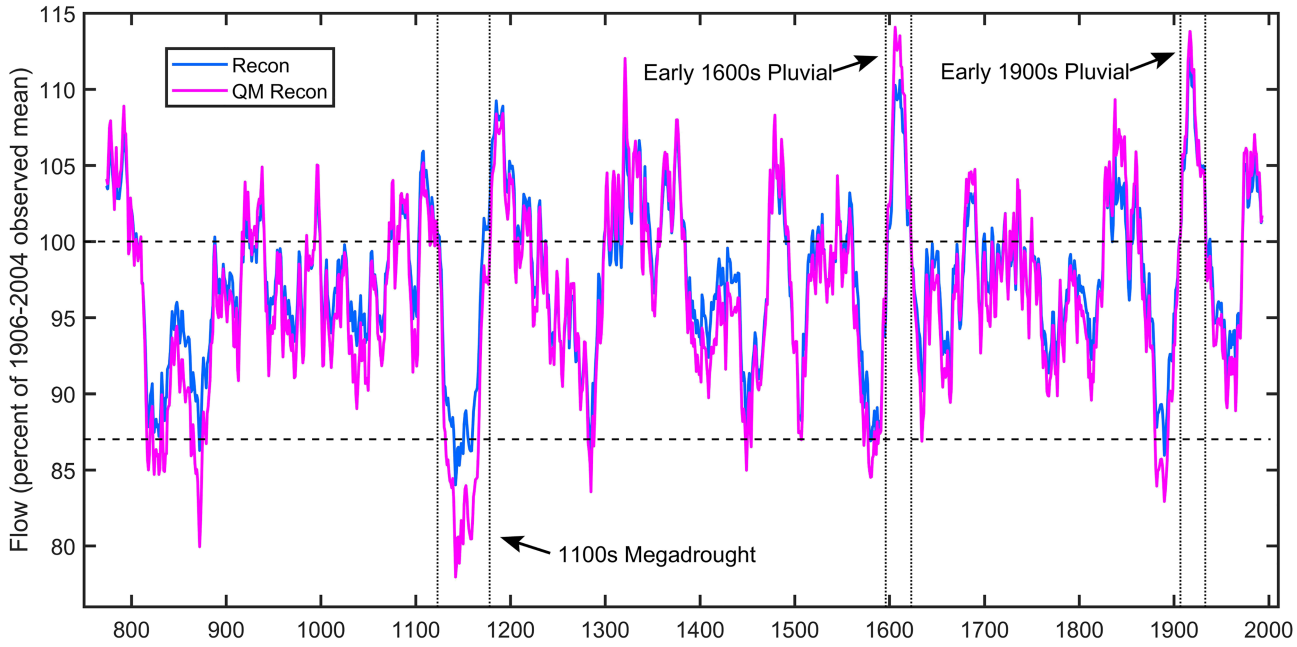

Figure 3. Twenty-five-year moving averages of Upper Colorado River flow over the entire reconstruction period, expressed as a percentage of the 1906–2004 mean (18.53×10^9 m^3). The blue line is the original reconstruction from Meko et al. (2007), whereas the magenta line is the reconstruction after bias correction (quantile mapping). The dashed line at 100% indicates the 1906–2004 mean, while the dashed line at 87% shows the lowest 25-year average during the 1906–2004 period, which is the average of the observed flow from 1953 to 1977.

reconstruction to 77.9% of average in the bias-corrected reconstruction. Recognizing the potential need for bias correction, Meko et al. (2007) specifically comment that they expected this event to be even more extreme than estimated: "[b]ecause regression biases the reconstructed flows toward the calibration-period mean, flows in the mid-1100s were quite possibly lower than indicated by the reconstruction." After bias correction, the 1100s drought also lasted longer, with the 25-year moving average staying below the 1906–2004 average for a decade longer than in the original reconstruction (Figure 3).

Our new estimate of the 1130 to 1154 minimum within the 1100s megadrought barely extends beyond the confidence intervals given in Meko et al. (2007), but the magnitude of the difference in hydrological terms is profound. The before-and-after bias correction difference of 6.1% represents a cumulative 25-year difference of 28.1×10^9 m^3, which is 1.52 times the annual mean flow at Lee's Ferry and represents 78.7% of the capacity of Lake Mead (the largest reservoir on the Colorado River and in the United States). Similarly, the cumulative difference between the original and bias-corrected estimates over the entire 56-year megadrought (the duration of time that the 25-year moving average remains below the 1906–2004 mean) is 51.8×10^9 m^3, which is 2.8 times the annual mean flow or the equivalent of 1.45 Lake Meads (the Bureau of Reclamation gives the capacity of Lake Mead to be 28.9 million acre-feet or 35.7×10^9 m^3). Beyond the mid-1100s megadrought, several other droughts that exceeded the largest 25-year drought during the calibration period (1953 to 1977) were identified in the bias-corrected reconstruction, notably events in the 800s, late 1200s, mid 1400s, late 1500s, and late 1800s all being more extreme than the 1953 to 1977 drought in the bias-corrected reconstruction.

Given that the 1100s megadrought is the most severe in the record and that it became even more extreme in the bias-corrected reconstruction, it warrants a closer look. The original reconstruction shows this megadrought as a period of 46 years with 25-year means that are below average while the bias-corrected reconstruction extends that duration to 56 years (Figure 4a). The reason that the 1100s megadrought is more extreme and lasted longer in the bias-corrected time series is that it was composed of a long period of lower-than-average but not extremely low flow. Specifically, the 1100s megadrought consistently had annual flows in the range of 11 to 18×10^9 m^3, a range where both nested models used to reconstruct the 1100s were biased high (i.e., Models 1 and 2 in Figure S1). So the original reconstruction is mostly

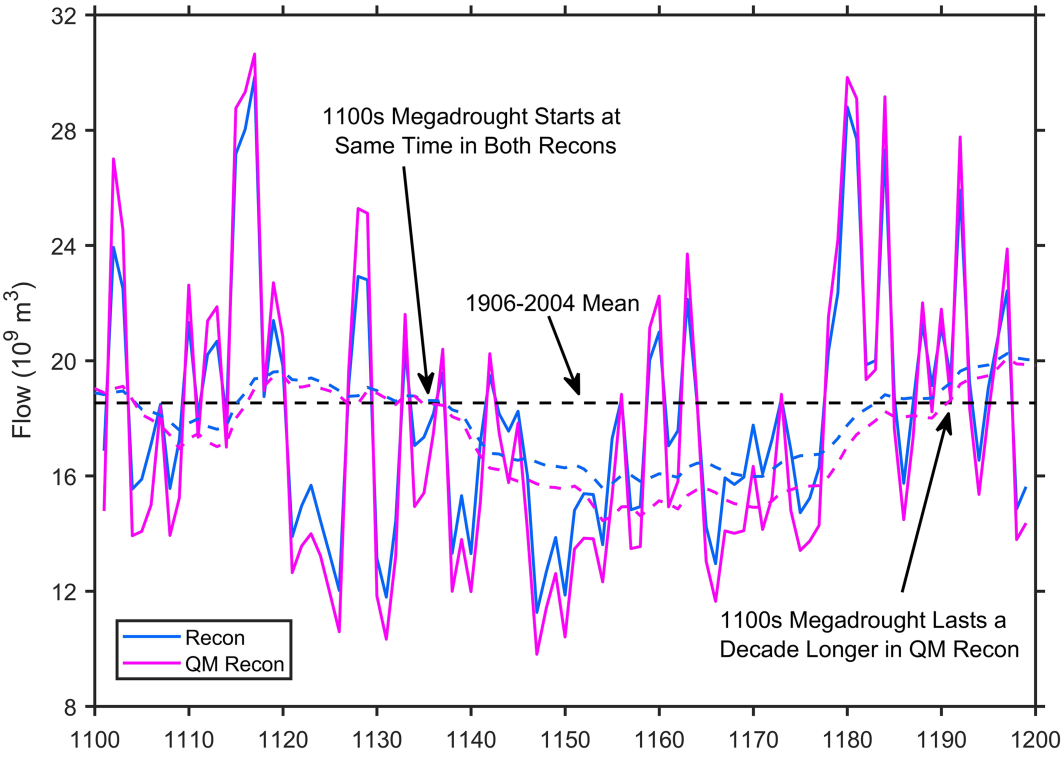

Figure 4. A closer inspection of the 1100s megadrought. Time series of reconstructed and bias-corrected reconstructed Upper Colorado River annual flow during the 1100s. Blue lines are the original reconstruction from Meko et al. (2007) whereas the magenta lines are the reconstruction after bias correction (QM: quantile mapping). Dashed blue and magenta lines are trailing 25-year moving average of the two reconstructions.

adjusted down during the 1100s megadrought. Cumulative cdfs of the observations and models show these features more clearly (Figure S3), where from about the 0.06 to the 0.60 quantiles of the observed flow distribution, all four nested models are biased high. It is worth noting that, while the 1100s had the most extreme multidecadal drought in the 1,200+year record, 3 years during the twentieth century (1934, 1977, and 2002) had annual observed flow that were lower than any individual year during the bias-corrected 1100s megadrought.

Based on the reconstructions from Lee's Ferry, the early 1900s pluvial often has been viewed as the wettest extended period in the last millennium. However, the 25-year moving averages (Figure 3) suggest that, after bias correction, the early 1600s pluvial rivals the well-known early 1900s one, so we provide a direct comparison of the two events before and after bias correction (Figure 5). With a pluvial defined as a 25-year period of above-average flow, the early 1900s pluvial was modestly changed by bias correction, with its cumulative anomaly increasing by only 8.8% because it was composed of a mix of biased-low and biased-high annual flows (Figure S4a). The cumulative anomaly of the early 1600s pluvial, on the other hand, increased by 42.2% after bias correction because it was composed of annual flows that were more consistently biased-low than the early 1900s pluvial and therefore received larger positive bias corrections (Figure S4b). Over the duration of the two pluvials, the early 1600s event had a cumulative flow anomaly of 36.0×10^9 m^3 over the 1906–2004 mean, while the early 1900s event produced a cumulative flow anomaly of 35.2×10^9 m^3 over the 1906–2004 mean. So after the reconstructed record is bias corrected, the peak of the early 1600s pluvial (in the 25-year moving averages) is slightly larger than that of the early 1900s pluvial, but the early 1900s pluvial is 2 years longer in duration (cf. the two magenta lines in Figure 5). As a result, the two events are very similar in magnitude, duration, and intensity, with both cumulative anomalies being approximately equal to twice the annual mean flow or the capacity of Lake Mead.

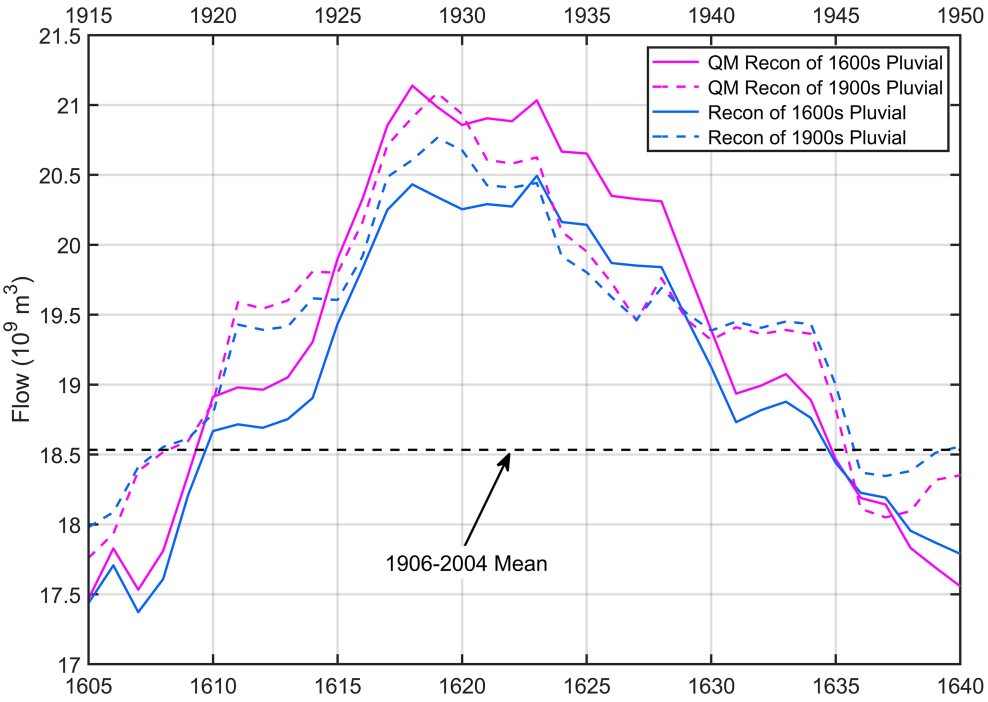

Figure 5. Direct comparison of pluvials during the early 1600s (solid lines) and early 1900s (dashed lines) in Upper Colorado River flow before (blue) and after (magenta) bias correction (note different *x* axes at top and bottom). For each, the trailing 25-year moving average is shown (e.g., on the solid lines, the values for 1625 are the averages for 1601–1625). After bias correction, the early 1600s pluvial is slightly larger in intensity and cumulative magnitude, but the early 1900s pluvial is longer in duration by 2 years (as defined by the number of 25-year averages that are above the 1906–2004 mean).

5. Conclusions

The record of reconstructed flow in the Upper Colorado River is one of the most important paleoclimatic data sets available. Here, we have shown that, while the reconstruction has low overall error, it does not match the probability distribution of observed flow. We show that, after bias correction, the early 1100s megadrought was even more extreme than previously thought and that the early 1600s had a pluvial that rivaled the well-known pluvial of the early 1900s. The early 1900s pluvial was a particularly important event, as it formed the baseline for apportioning Colorado River water to U.S. and Mexican states and has previously been thought to be the wettest period in the past 1,200+ years. In the bias-corrected reconstruction of flow, it still stands out as uncharacteristically wet, but the early 1600s pluvial appears to be at least as large in magnitude and nearly as long in duration.

While we think that bias correction has an important role to play in the development of paleoclimatic data sets and the interpretation of past events, it is not a panacea. Without careful controls, it is possible for bias correction to increase model error or to produce physically implausible values. The type of correction and how closely it matches observations, especially if observations contain errors, should be chosen carefully. Variable-specific bias-correction procedures have been developed, with some ensuring that all adjusted values are nonnegative or that large extrapolations do not occur when the magnitude of model values greatly exceeds the observed minimum or maximum. In some cases, multivariate bias correction may be needed to ensure that the covariance between multiple reconstructed variables is preserved (Cannon, 2018). It also is important to recognize that the minimum-error properties of a statistical reconstruction model are affected by bias correction. Even so, the goal of the bias-correction model is fundamentally different from that of the model used for reconstruction, as the response variable is no longer paired with the predictor (or even the observation from the same time) and the goal is to minimize bias not prediction error. The potential for bias correction to amplify noise should be weighed against its ability to accurately reproduce the full range of observed variability.

Similar to climate-model projections, those who would like to bias correct reconstruction models must be wary of the implications of assuming stationarity, especially when reconstruction models use short instrumental records or have poor fit. Further, bias correction of a model with poor fit is likely not worthwhile, but we currently do not have good guidance on when it becomes inappropriate. Bias corrections of paleoclimatic reconstructions also are limited by the representativeness of the instrumental record, especially for events that occur rarely over millennial timescales. Overall, bias-corrected results should be inspected as carefully as any other model output. To support different bias corrections using nested paleoclimatic models, we urge researchers to make available the calibration-period estimates of all nested reconstruction models. In addition to recommending that bias correction be added cautiously to the paleoclimatology toolbox, we also think that quantile-quantile and kernel-density plots are overlooked but essential components of paleoclimatic model evaluation. The degree to which a model is able to reproduce the observed probability distribution is often as important as overall model performance as measured by an index such as CE or d_r.

Bias correcting a paleoclimatic reconstruction makes it better able reproduce the observed probability distribution. Extending this argument, we suggest that, unless a reconstruction is bias corrected or closely matches the observed probability distribution, it is inappropriate to compare reconstructed values directly to observed values. This is especially important for extreme events such as droughts or pluvials, where reconstructed values often are more muted than observations. If a reconstruction is biased—and is not biased corrected—then reconstructed values from the past should only be compared with reconstructed values, including those during the observational period. For instance, because the original Upper Colorado River reconstruction deviates from observations in the upper tail of the distribution, it is not appropriate to compare the observed early 1900s pluvial to other reconstructed pluvials unless they are bias corrected. Similarly, as new observations become available, such as recent years of lower-than-average Colorado River flow, they should only be compared to other observations or to a bias-corrected reconstruction. Bias correction thus allows for an "apples-to-apples" comparison of observations and reconstructed values over a much wider span of time.

Acknowledgments
The data for the original reconstruction as well as the bias-corrected version developed here are available from NOAA's Paleoclimatology Data site (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data). An R script for generating the bias-corrected reconstruction also is available at the NOAA Paleoclimatology Data site. Dr. David Meko generously provided the full time series for each of the four nested reconstruction models. We appreciate the insightful comments of the three reviewers and the contributions of all of the researchers who helped to produce the Upper Colorado River reconstructions used here.

References

Abatzoglou, J. T., & Brown, T. J. (2012). A comparison of statistical downscaling methods suited for wildfire applications. *International Journal of Climatology*, *32*(5), 772–780.

Axelson, J. N., Sauchyn, D. J., & Barichivich, J. (2009). New reconstructions of streamflow variability in the South Saskatchewan River Basin from a network of tree ring chronologies, Alberta, Canada. *Water Resources Research*, *45*, W09422. https://doi.org/10.1029/2008WR007639

Berg, P., Feldmann, H., & Panitz, H. J. (2012). Bias correction of high resolution regional climate model data. *Journal of Hydrology*, *448*, 80–92.

Biondi, F., & Meko, D. M. (2019). Long-term hydroclimatic patterns in the Truckee-Carson Basin of the eastern Sierra Nevada, USA. *Water Resources Research*, *55*, 5559–5574. https://doi.org/10.1029/2019WR024735

Boé, J., Terray, L., Habets, F., & Martin, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *International Journal of Climatology*, *27*(12), 1643–1655.

Buras, A. (2017). A comment on the expressed population signal. *Dendrochronologia*, *44*, 130–132.

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, *50*(1-2), 31–49.

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

Christensen, J. H., Boberg, F., Christensen, O. B., & Lucas-Picher, P. (2008). On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, *35*, L20709. https://doi.org/10.1029/2008GL035694

Christensen, N., & Lettenmaier, D. P. (2007). A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River basin. *Hydrology and Earth System Sciences*, *11*, 1417–1434.

Cook, B. I., Ault, T. R., & Smerdon, J. E. (2015). Unprecedented 21st century drought risk in the American Southwest and Central Plains. *Science Advances*, *1*(1), e1400082. https://doi.org/10.1126/sciadv.1400082

Cook, E. R., Meko, D. M., Stahle, D. W., & Cleaveland, M. K. (1999). Drought reconstructions for the continental United States. *Journal of Climate*, *12*(4), 1145–1162.

Cook, E. R., Seager, R., Kushnir, Y., Briffa, K. R., Büntgen, U., Frank, D., et al. (2015). Old World megadroughts and pluvials during the Common Era. *Science Advances*, *1*(10), e1500561. https://doi.org/10.1126/sciadv.1500561

Coulthard, B., Smith, D. J., & Meko, D. M. (2016). Is worst-case scenario streamflow drought underestimated in British Columbia? A multi-century perspective for the south coast, derived from tree-rings. *Journal of Hydrology*, *534*, 205–218.

Esper, J., Cook, E. R., & Schweingruber, F. H. (2002). Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science*, *295*(5563), 2250–2253.

Esper, J., Frank, D. C., Wilson, R. J., & Briffa, K. R. (2005). Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. *Geophysical Research Letters*, *32*, L07711. https://doi.org/10.1029/2004GL021236

Ficklin, D. L., Abatzoglou, J. T., Robeson, S. M., & Dufficy, A. (2016). The influence of climate model biases on projections of aridity and drought. *Journal of Climate*, *29*(4), 1269–1285.

Ficklin, D. L., Stewart, I. T., & Maurer, E. P. (2013). Climate change impacts on streamflow and subbasin-scale hydrology in the Upper Colorado River Basin. *PloS ONE*, *8*(8), e71297. https://doi.org/10.1371/journal.pone.0071297

Gangopadhyay, S., Harding, B. L., Rajagopalan, B., Lukas, J. J., & Fulp, T. J. (2009). A non-parametric approach for paleo reconstruction of annual streamflow ensembles. *Water Resources Research*, *45*, W06417. https://doi.org/10.1029/2008WR00,720

Gangopadhyay, S., McCabe, G. J., & Woodhouse, C. A. (2015). Beyond annual stream-flow reconstructions for the Upper Colorado River Basin: A paleo-water-balance approach. *Water Resources Research*, *51*, 9763–9774. https://doi.org/10.1002/2015WR017283

Graumlich, L. J., Pisaric, M. F. J., Waggoner, L. A., Littell, J. S., & King, J. C. (2003). Upper Yellowstone River flow and teleconnections with Pacific Basin climate variability during the past three centuries. *Climatic Change*, *59*, 245–262.

Griffin, D., & Anchukaitis, K. J. (2014). How unusual is the 2012–2014 California drought? *Geophysical Research Letters*, *41*, 9017–9023. https://doi.org/10.1002/2014GL062433

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling RCM precipitation to the station scale using statistical transformations—A comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390.

Harley, G. L., Maxwell, J. T., Larson, E., Grissino-Mayer, H. D., Henderson, J., & Huffman, J. (2017). Suwannee River flow variability 1550–2005 CE reconstructed from a multispecies tree-ring network. *Journal of Hydrology*, *544*, 438–451. https://doi.org/10.1016/j.jhydrol.2016.11.020

Li, H., Sheffield, J., & Wood, E. F. (2010). Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research*, *115*, D10101. https://doi.org/10.1029/2009JD012882

MacDonnell, L. J., Getches, D. H., & Hugenberg, W. C. Jr. (1995). The law of the Colorado River: Coping with severe sustained drought. *JAWRA Journal of the American Water Resources Association*, *31*(5), 825–836.

Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, *26*(6), 2137–2143.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*, RG3003. https://doi.org/10.1029/2009RG000314

Margolis, E. Q., Meko, D. M., & Touchan, R. (2011). A tree-ring reconstruction of streamflow in the Santa Fe River, New Mexico. *Journal of Hydrology*, *397*(1-2), 118–127.

Maxwell, J. T., Harley, G. L., & Matheus, T. J. (2015). Dendroclimatic reconstructions from multiple co-occurring species: A case study from an old-growth deciduous forest in Indiana, USA. *International Journal of Climatology*, *35*(6), 860–870.

Maxwell, J. T., Harley, G. L., & Robeson, S. M. (2016). On the declining relationship between tree growth and climate in the Midwest United States: the fading drought signal. *Climatic Change*, *138*(1-2), 127–142.

Maxwell, R. S., Harley, G. L., Maxwell, J. T., Rayback, S. A., Pederson, N., Cook, E. R., et al. (2017). An interbasin comparison of tree-ring reconstructed streamflow in the eastern United States. *Hydrological Processes*, *31*(13), 2381–2394.

Maxwell, R. S., Hessl, A. E., Cook, E. R., & Pederson, N. (2011). A multispecies tree ring reconstruction of Potomac River streamflow (950–2001). *Water Resources Research*, *47*, W05512. https://doi.org/10.1029/2010WR010019

McCabe, G. J., Wolock, D. M., Pederson, G. T., Woodhouse, C. A., & McAfee, S. (2017). Evidence that recent warming is reducing upper Colorado River flows. *Earth Interactions*, *21*(10), 1–14.

Meko, D. (1997). Dendroclimatic reconstruction with time varying predictor subsets of tree indices. *Journal of Climate*, *10*(4), 687–696.

Meko, D., Woodhouse, C. A., Baisan, C. A., Knight, T., Lukas, J. J., Hughes, M. K., & Salzer, M. W. (2007). Medieval drought in the upper Colorado River Basin. *Geophysical Research Letters*, *34*, L10705. https://doi.org/10.1029/2007GL029988

Meko, D. M., Therrell, M. D., Baisan, C. H., & Hughes, M. K. (2001). Sacramento River flow reconstructed to A.D. 869 from tree rings. *Journal of the American Water Resources Association*, *37*(4), 1029–1040.

Oliver, J. S., Harley, G. L., & Maxwell, J. T. (2019). 2,500 years of hydroclimate variability in New Mexico, USA. *Geophysical Research Letters*, *46*, 4432–4440. https://doi.org/10.1029/2019GL082649

Pederson, N., Bell, A. R., Cook, E. R., Lall, U., Devineni, N., Seager, R., et al. (2013). Is an epic pluvial masking the water insecurity of the greater New York City region? *Journal of Climate*, *26*(4), 1339–1354.

Pederson, N., Bell, A. R., Knight, T. A., Leland, C., Malcomb, N., Anchukaitis, K. J., et al. (2012). A long-term perspective on a modern drought in the American Southeast. *Environmental Research Letters*, *7*(1).

Pulwarty, R. S., Jacobs, K. L., & Dole, R. M. (2005). The hardest working river: drought and critical water problems in the Colorado River Basin. In D. A. Wilhite (Ed.), *Drought and water crises: Science, technology, and management issues* (pp. 249–285). Boca Raton: CRC Press.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Robeson, S. M. (2015). Revisiting the recent California drought as an extreme value. *Geophysical Research Letters*, *42*, 6771–6779. https://doi.org/10.1002/2015GL064593

Routson, C. C., Woodhouse, C. A., & Overpeck, J. T. (2011). Second century megadrought in the Rio Grande headwaters, Colorado: How unusual was medieval drought? *Geophysical Research Letters*, *38*, L22703. https://doi.org/10.1029/2011GL050015

Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., et al. (2007). Model projections of an imminent transition to a more arid climate in southwestern North America. *Science*, *316*(5828), 1181–1184. https://doi.org/10.1126/science.1139601

St. George, S., & Nielsen, E. (2003). Palaeoflood records for the Red River, Manitoba, Canada, derived from anatomical tree-ring signatures. *The Holocene*, *13*(4), 547–555.

Stahle, D. W., & Cleaveland, M. K. (1992). Reconstruction and analysis of spring rainfall over the southeastern US for the past 1000 years. *Bulletin of the American Meteorological Society*, *73*(12), 1947–1961.

Stahle, D. W., Fye, F. K., Cook, E. R., & Griffin, R. D. (2007). Tree-ring reconstructed megadroughts over North America since AD 1300. *Climatic Change*, *83*(1-2), 133.

Stockton, C. W., and G. C. Jacoby (1976), Long-term surface-water supply and streamflow trends in the Upper Colorado River Basin, Lake Powell Res. Proj. Bull. 18, Natl. Sci. Found., Arlington, Va.

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456*, 12–29.

Udall, B., & Overpeck, J. (2017). The twenty-first century Colorado River hot drought and implications for the future. *Water Resources Research*, *53*, 2404–2418. https://doi.org/10.1002/2016WR019638

Wigley, T. M., Briffa, K. R., & Jones, P. D. (1984). On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. *Journal of climate and Applied Meteorology*, *23*(2), 201–213.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088–2094.

Willmott, C. J., Robeson, S. M., Matsuura, K., & Ficklin, D. L. (2015). Assessment of three dimensionless measures of model performance. *Environmental Modelling & Software*, *73*, 167–174.

Wilson, R., Anchukaitis, K., Briffa, K. R., Büntgen, U., Cook, E., D'Arrigo, R., et al. (2016). Last millennium northern hemisphere summer temperatures from tree rings: Part I: The long term context. *Quaternary Science Reviews*, *134*, 1–18. https://doi.org/10.1016/j.quascirev.2015.12.005

Wise, E. K. (2010). Tree ring record of streamflow and drought in the upper Snake River. *Water Resources Research*, *46*, W11529. https://doi.org/10.1029/2010WR009282

Woodhouse, C. A., Gray, S. T., & Meko, D. M. (2006). Updated streamflow reconstructions for the Upper Colorado River Basin. *Water Resources Research*, *42*, W05415. https://doi.org/10.1029/2005WR004455

Woodhouse, C. A., Kunkel, K. E., Easterling, D. R., & Cook, E. R. (2005). The twentieth-century pluvial in the western United States. *Geophysical Research Letters*, *32*, L07701. https://doi.org/10.1029/2005GL022413

Woodhouse, C. A., & Overpeck, J. T. (1998). 2000 years of drought variability in the central United States. *Bulletin of the American Meteorological Society*, *79*(12), 2693–2714.