A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles

Subhrendu Gangopadhyay,[1] Benjamin L. Harding,[1] Balaji Rajagopalan,[2,3] Jeffrey J. Lukas,[4] and Terrance J. Fulp[5]

Received 5 June 2008; revised 7 March 2009; accepted 30 March 2009; published 13 June 2009.

[1] As multicentury records of natural hydrologic variability, tree ring reconstructions of streamflow have proven valuable in water resources planning and management. All previous reconstructions have used parametric methods, most often regression, to develop a model relating a set of tree ring data to a target hydrology. In this paper, we present the first development and application of a *K* nearest neighbor (*K*NN) nonparametric method to reconstruct naturalized annual streamflow ensembles from tree ring chronology data in the Upper Colorado River Basin region. The method is developed using tree ring chronologies from the period 1400–2005 and naturalized streamflow from the period 1906–2005 at the important Lees Ferry, Arizona, gauge on the Colorado River to develop annual streamflow ensembles for this gauge for the 1400–1905 period. The proposed *K*NN algorithm was developed and tested using cross validation for the overlap period, i.e., the contemporary observed period for which both the tree ring and streamflow data are available (1906–2005). The cross-validated streamflow reconstructions for the selected contemporary period compare very well with the observed flows and also with published parametric streamflow reconstructions for this gauge. The proposed nonparametric method provides an ensemble of streamflows for each year in the paleohydrologic reconstruction period (1400–1905) and, consequently, a more realistic asymmetric confidence interval than one obtained through most parametric approaches. Also, the *K* nearest neighbors are obtained only from the tree ring chronology data, and thus, the method can be used to reconstruct structured and even nonnumerical data for use in water resources modeling.

Citation: Gangopadhyay, S., B. L. Harding, B. Rajagopalan, J. J. Lukas, and T. J. Fulp (2009), A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles, *Water Resour. Res.*, *45*, W06417, doi:10.1029/2008WR007201.

1. Introduction

[2] Paleohydrologic reconstructions of streamflows are very useful for understanding multidecadal variability and for drought mitigation planning. Nowhere is this more evident than in the western United States, especially in the Upper Colorado River Basin (Figure 1). The utility of the paleohydrologic reconstructions in this basin is underscored by the recent severe and sustained drought. This is illustrated in Figure 2: a paleohydrologic reconstructed streamflow for the period 1490–1997, on the Colorado River at Lees Ferry, Arizona [*Woodhouse et al.*, 2006], a key gauge on the river along with the observed flows. It is evident that the recent 5-year drought of 2000–2004 is unprecedented during the observed period, but the reconstructed streamflows prior to 1906 show severe droughts of 5 years length at least 4 times over the approximately 500-year period, indicating that the recent drought is not unusual.

[3] All previous paleohydrologic reconstructions, including the ones shown in Figure 2, have been developed using a parametric statistical model which fits a set of tree ring chronologies to the historical naturalized streamflows over a calibration period (typically 50–100 years) for which the two sets of records overlap [e.g., *Stockton and Jacoby*, 1976; *Meko et al.*, 1995]. A tree ring chronology is a time series of dimensionless ring width indices derived from a group of trees at one site and corrected for physiological and other biases. The full-length tree ring chronologies (>300 years) are put into that model to estimate streamflows during the precalibration period. In most cases, the model has been derived through a multiple linear regression (MLR) approach, with many variations on this approach. For example, the set of chronologies is sometimes reduced using principal components analysis (PCA), with the leading principle components (PCs) used in calibrating the MLR model [e.g., *Stockton and Jacoby*, 1976; *Hidalgo et al.*, 2000]. The reconstructions in this approach are sensitive to

[1] Earth and Environmental Division, AMEC, Boulder, Colorado, USA.
[2] Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, Colorado, USA.
[3] Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, Colorado, USA.
[4] Institute of Arctic and Alpine Research, University of Colorado at Boulder, Boulder, Colorado, USA.
[5] Lower Colorado Region, U.S. Bureau of Reclamation, Boulder City, Nevada, USA.

0043-1397/09/2008WR007201$09.00

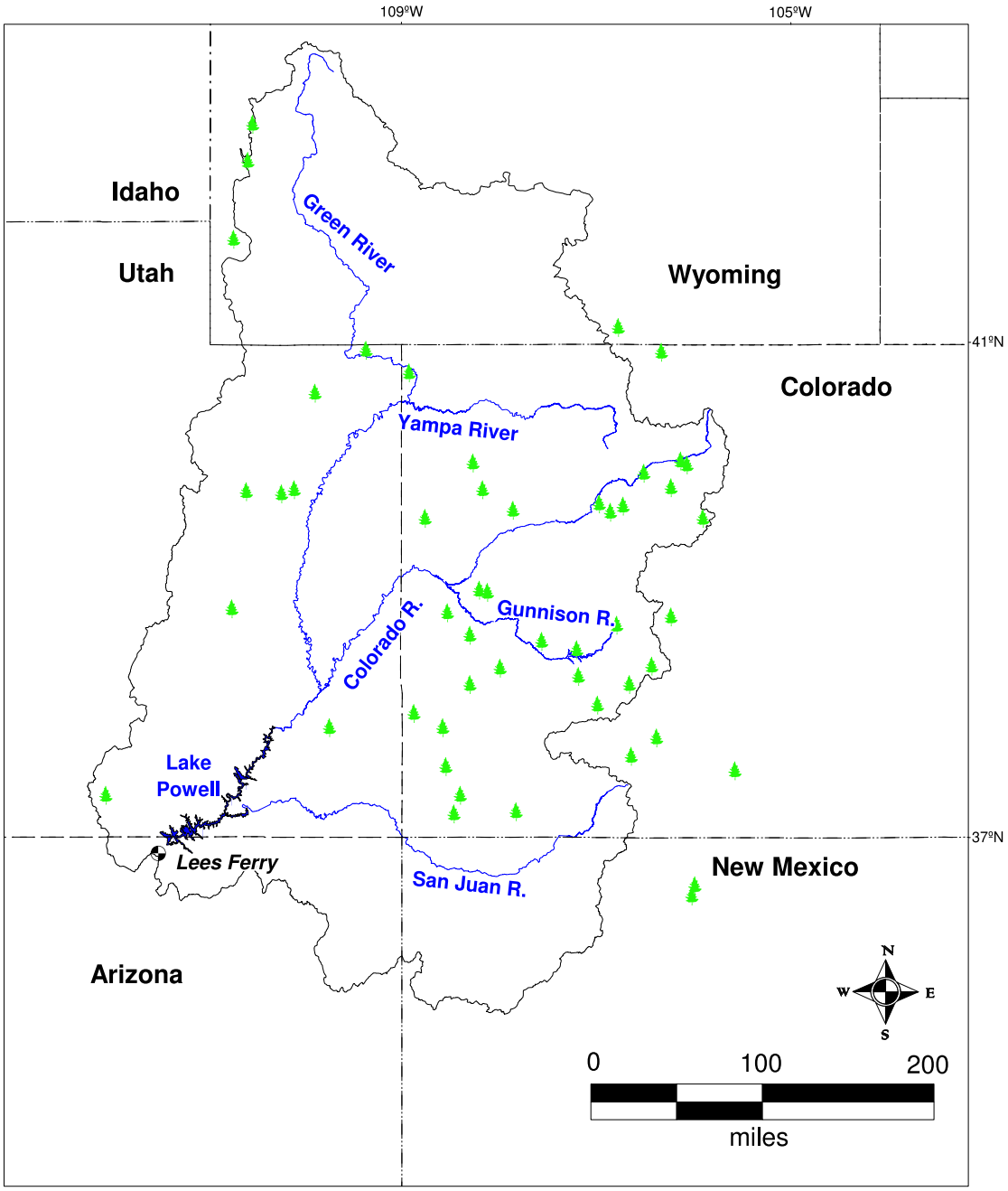

Figure 1. The Upper Colorado River Basin showing the location of chronologies used for this study and the Lees Ferry gauge.

the number of PCs retained as shown by *Hidalgo et al.* [2000].

[4] These parametric reconstruction techniques generally capture the variability of the historical flow very well; however, because a single model is fitted to a limited number of cases and is used to project the precalibration flows, the reconstructed flow magnitudes are sensitive to several aspects of the model-building process. For example, seven different parametric (MLR-based) reconstructions of Lees Ferry flows [*Stockton and Jacoby*, 1976; *Hidalgo et al.*, 2000; *Woodhouse et al.*, 2006] are shown in Figure 3. The divergence of streamflows among the various reconstructions during the preobservation period is due to the use of different calibration techniques, different tree ring data treatment (i.e., prewhitened versus nonprewhitened chronologies and the inclusion of lagged predictors), different sets of tree ring data, and different observed data (both the years used and the hydrologic time series itself) for the calibration. All of these are potential sources of the differences, and these differences should be expected. The fact that these different reconstructions do vary coherently is a testament to the robustness of the hydroclimatic signal in the trees. This is evident in Figure 4, which represents the sequences of annual hydrologic states (wet or dry) on the basis of conditions in a 10-year window. For a particular reconstruction, a wet state (black bar) is defined as when the 10-year mean flow is greater than the mean flow over that entire reconstruction; otherwise it is classified as dry (gray

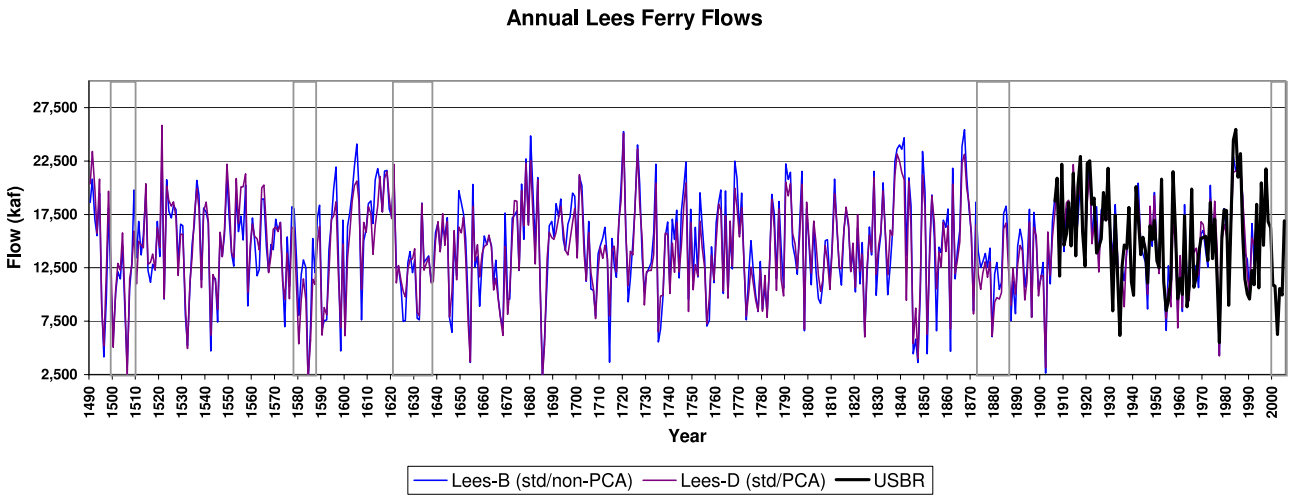

Figure 2. Annual Lees Ferry flows for the period 1490–2005 showing historical (U.S. Bureau of Reclamation) and reconstructed (Lees B and Lees D) data from *Woodhouse et al.* [2006]. Notable droughts are shown with rectangular boxes. Note that the Lees B and Lees D reconstructions are based on standard chronologies (standard chronologies were also used for the NPP reconstructions). Two additional reconstructions, Lees A and Lees C, based on residual chronologies were also developed by *Woodhouse et al.* [2006]. 1 KAF = 1.234 $\times$ 10^6 m^3.

bar). However, because of the divergence of the reconstructions, some water managers have expressed reluctance to use them in their planning efforts. To address this problem, *Prairie et al.* [2008] developed a technique to combine the state information from the reconstructions with the magnitude information of the observed flows to generate a rich variety of streamflow sequences for use in water resources systems analysis.

[5] The MLR technique suffers from four main drawbacks: (1) it assumes that the data are normally distributed, (2) it assumes that there is no correlation between the predictor variables, (3) it produces variance compression of the predictand, and (4) outliers can have an undue influence on the fitted MLR model. To address the first drawback, data have to be tested and, in some cases, transformed to follow a normal distribution via Box-Cox and normal probability swap transformations [*Helsel and Hirsch*, 2002; *Wilks*, 1995; *Deutsch and Journel*, 1992]. In developing a predictive model using MLR, it is assumed that there is no correlation between the predictor variables. However, when several predictor variables are used in MLR, a fairly high level of predictability may exist for one or more predictors from the other predictors. This condition is referred to as multicollinearity and can lead to instability of the regression coefficients [e.g., *Myers*, 1990]. Stepwise regression can largely alleviate multicollinearity, but principal component–based regression can be used to eliminate multicollinearity. To address variance compression in streamflow reconstructions, *Meko et al.* [2001] adopted a noise-added probabilistic approach to interpretation of the reconstructions. In other tree ring studies [e.g., *Cook et al.*, 2004], it was deemed desirable to scale the variance of reconstructions up to that of the observed predictand so that their construction would more realistically

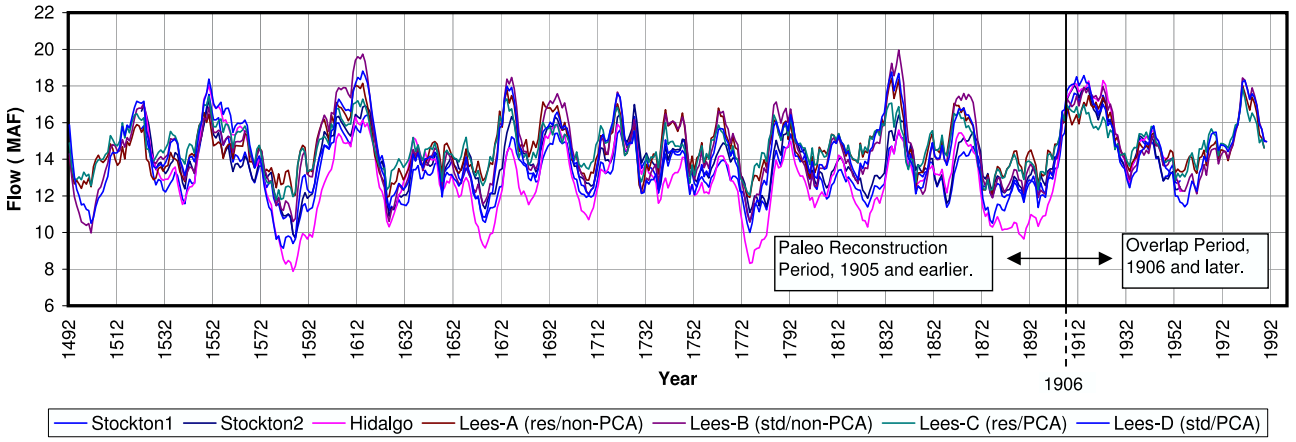

Figure 3. Seven reconstructions of Colorado River annual flow at Lees Ferry. The lines labeled Stockton1 and Stockton2 are from *Stockton and Jacoby* [1976]; the line labeled Hidalgo is from *Hidalgo et al.* [2000]; and the Lees A, Lees B, Lees C, and Lees D traces are from *Woodhouse et al.* [2006].

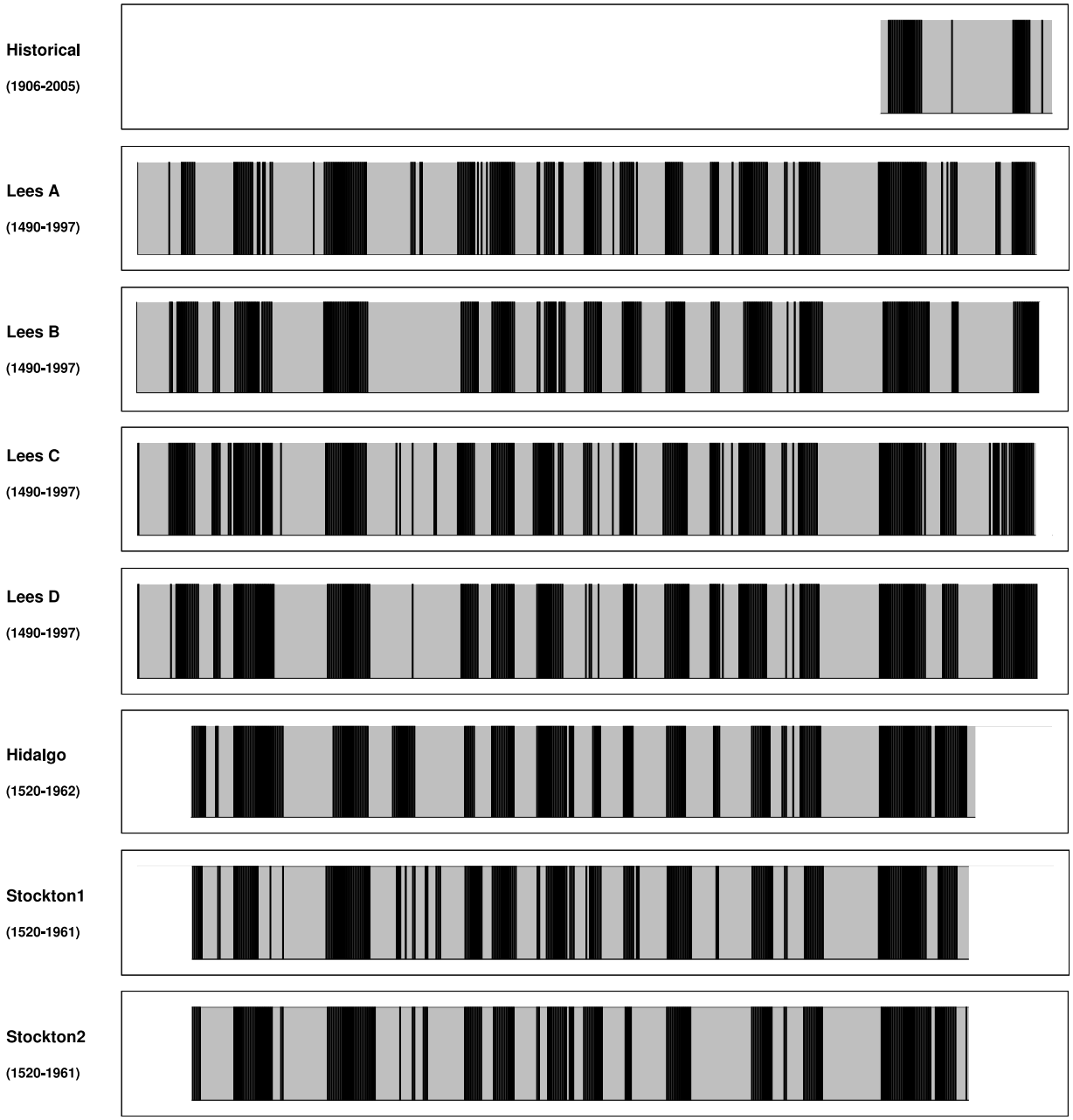

Figure 4. Common signal in tree ring chronologies for Lees Ferry flows. The dark areas represent wet spells, whereas the gray areas represent the dry spells. The plot labeled "Historical" represents the wet-dry spells in the historical period 1906–2005. "Lees A" through "Lees D" are the *Woodhouse et al.* [2006] reconstructions for the 1490–1997 period. The "Hidalgo" plot is the *Hidalgo et al.* [2000] reconstruction (1520–1962), and "Stockton1" and "Stockton2" are the reconstructions by *Stockton and Jacoby* [1976] for the period 1520–1961. Since these reconstructions span different periods, nonoverlapping years are shown as white spaces.

portray extremes. The outlier effect can be addressed only to a limited extent by robust regression techniques [*Helsel and Hirsch*, 2002]. Since the MLR approach involves fitting a single function to the entire observed data, the effect of outliers cannot be fully eliminated. One other shortcoming is that the uncertainty estimates are based on the regression theory and are symmetric and wide. However, several streamflow reconstructions [e.g., *Smith and Stockton*, 1981; *Meko and Graybill*, 1995; *Meko et al.*, 2001] have used regression of log-transformed flow on tree rings. In these cases the error bars around the predicted log 10 flows are symmetric, but after back transforming to original flow units, the error bars are asymmetric. Though such asymmetric uncertainty estimates may be obtained, flow transformation does complicate the interpretation of the regression statistics and the reconstruction.

[6] Thus, the motivation for this research is to develop a simple and flexible technique that alleviates the drawbacks of the traditional MLR approach. To this end, we develop a nonparametric paleohydrologic (NPP) method based on the

K nearest neighbor bootstrap [see, e.g., *Lall and Sharma*, 1996; *Gangopadhyay et al.*, 2005] in empirical orthogonal function (EOF), or principal component (PC) (EOFs are also referred to as PCs), space and apply it to reconstruct annual streamflow ensembles at the Lees Ferry gauge.

[7] This nonparametric approach has important advantages over the standard parametric modeling approach such as the quantification of uncertainty in paleohydrologic streamflow reconstruction using ensembles. Also, in the proposed nonparametric framework, neighbors are obtained only from the tree ring chronology data, so the approach can be used to develop reconstructions of other hydrologic marker variables (e.g., the Palmer drought severity index). These are important contributions to the fields of stochastic hydrology and dendrohydrology.

[8] The paper is organized as follows. A description of the data sets used for the study is given in section 2. The nonparametric algorithm is described in section 3, and the results are discussed in section 4, followed by a summary (section 5).

2. Description of Tree Ring Chronology and Streamflow Data Sets

2.1. Tree Ring Chronology Data Set

[9] This study used 51 tree ring chronologies from across the Upper Colorado River Basin and adjacent basins (Figure 1). The width of annual tree rings in the interior western United States often provides a robust proxy measure of annual streamflow since the same climatic factors (mainly precipitation and evapotranspiration) influence the variability of both tree growth and streamflow. The tree species in the region best suited to capture hydroclimatic variability include ponderosa pine, Douglas fir, and pinyon pine [*Meko et al.*, 1995]. Generally, two core samples are collected from each of 15–30 trees from one species at a given site. The annual rings are cross dated (patterns of wide and narrow rings are matched across samples) to ensure absolute dating to the calendar year, and then these annual ring widths for each sample are measured [*Stokes and Smiley*, 1968; *Fritts*, 1976]. For this study, these raw annual ring width measurements from each sample were then detrended with a fixed-length cubic spline (50% frequency response at a 300-year wavelength) to remove age- and geometry-related growth trends, and then a robust biweight mean was used to calculate an average site ring width index for each year [*Cook and Briffa*, 1990; *Cook et al.*, 1990], with the time series of these annual ring width indices constituting the chronology. Tree ring chronologies typically contain significant low-order autocorrelation, which is either retained in the ''standard'' chronologies or removed using autoregressive modeling (prewhitened) to produce ''residual'' chronologies. For this study, the standard chronologies were used in the analyses and reconstructions.

[10] The 51 chronologies used in this study include all 38 chronologies located within the Upper Colorado River and North Platte River basins that were used in the MLR streamflow reconstructions by *Woodhouse et al.* [2006]. The other 13 chronologies include 9 chronologies more recently collected from the Upper Colorado River Basin and 4 chronologies from the Upper Rio Grande Basin. All 51 chronologies are from species known to be moisture sensitive, and nearly all have significant ($p < 0.05$) relationships with local annual precipitation and streamflow records. In an MLR approach, it would be advisable to reduce the pool of chronologies (i.e., candidate predictors) through some screening or data reduction procedure, but this is not necessary with the NPP approach since there is no parametric model at risk of being over fitted. All 51 chronologies begin in 1604 or earlier and extend through 1997 or later. Since most (32) of the chronologies extend back to at least 1400, for the purposes of the NPP methodology the beginning of the paleohydrologic reconstruction period was set at 1400.

2.2. Lees Ferry Naturalized Streamflow Data Set

[11] The natural streamflow data for the Colorado River Basin are developed by the U.S. Bureau of Reclamation and are updated regularly. Annual updates addressing data changes and additions are typical. Naturalized streamflows are computed by removing anthropogenic impacts (i.e., reservoir regulation, consumptive water use, etc.) from the recorded historic flows. J. Prairie and R. Callejo (Natural flow and salt computation methods: Calendar years 1971–1995, U.S. Bureau of Reclamation report, 2005, available at http://www.usbr.gov/lc/region/g4000/NaturalFlow/FinalMethodsCmptgNatFlow.pdf) present a detailed description of methods and data used for the computation of natural flows in the Colorado River Basin. This study uses the annual water year (October–September) naturalized streamflow at Lees Ferry, Arizona, for the period 1906–2005. Both the tree ring chronology data set and Lees Ferry naturalized streamflow data set are available in the auxiliary material.[1]

3. Methodology

[12] Description of the ensemble streamflow reconstruction algorithm using tree ring chronologies is presented in the following steps.

[13] Step 1. Let [**X**] represent the data matrix of tree ring chronologies for T years (rows) and M sites (columns). Tree ring chronologies for the Upper Colorado River Basin are presently available for 51 sites ($M = 51$). Matrix [**X**] represents data for the entire period of record (paleohydrologic period, 1400–1905, and the overlap period, 1906–2005; $T = 606$). Since the tree ring chronologies evolve over space and time, it should be noted that there would be missing entries in matrix [**X**].

[14] Step 2. Partition matrix [**X**] into two submatrices, [**A**] and [**B**], such that [**A**] is of order $N \times M$ and [**B**] is of order $(T - N) \times M$, where N is the length of the paleohydrologic period (506 years, 1400–1905). Both the paleohydrologic period and the overlap period contain missing values for some sites in some years. This is due to the fact that not all chronologies start and end in the same year. A summary of the numbers of chronologies available over selected periods is given in Table 1.

[15] Step 3. For feature year i, i.e., the year for which reconstruction is sought in the paleohydrologic period, identify the chronology sites m_i ($m_i \leq M$) that are also

[1] Auxiliary materials are available at ftp://ftp.agu.org/apend/wr/2008WR007201.

Table 1. Number of Chronologies Used in Paleohydrologic Reconstruction When the Overlap Period is 1906–1997

Years	Number of Chronologies
1400–1404	32
1405–1436	34
1437–1439	35
1440–1449	36
1450–1453	37
1454–1479	38
1480–1507	39
1508–1510	40
1511–1519	41
1520–1523	42
1524–1535	43
1536–1565	44
1566–1568	46
1569–1570	47
1571–1574	48
1575–1583	49
1584–1603	50
1604–1905	51

present for the overlap period. The m_i chronologies of the feature year i represent the feature vector $\{\mathbf{F}\}$ (row vector of length m_i).

[16] Step 4. Obtain a subset of data matrix [**B**], say, [**S**] of order $n \times m_i$, where $n \leq (T - N)$. In the actual implementation of this algorithm for the paper, 3 end years in the overlap period are considered, 1997, 2002, and 2005, so that n equals 92, 97, and 100, respectively.

[17] Step 5. Estimate correlation matrix [**C**] (order $m_i \times m_i$) from data matrix [**S**].

[18] Step 6. Perform PCA [e.g., *Haan*, 1977; *Wilks*, 1995] using matrix [**C**] to obtain the m_i eigenvalues $\lambda_{(1)}, \ldots, \lambda_{(m_i)}$ and the eigenmatrix (matrix of eigenvectors as columns) [**E**] (order $m_i \times m_i$).

[19] Step 7. Project the feature vector $\{\mathbf{F}\}$ for feature year i onto the eigenvectors in matrix [**E**]. The projected feature vector $\{\mathbf{F}'\}$ is given by

$$\{\mathbf{F}'\}_{1\times m_i} = \{\mathbf{F}\}_{1\times m_i}[\mathbf{E}]_{m_i\times m_i}. \tag{1}$$

[20] Step 8. Calculate the m_i principal components. The principal component matrix [**Z**] is obtained from

$$[\mathbf{Z}]_{n\times m_i} = [\mathbf{S}]_{n\times m_i}[\mathbf{E}]_{m_i\times m_i}. \tag{2}$$

[21] Step 9. For each element t ($t = 1, \ldots, n$) compute the weighted Euclidian distance (d_t) between the projected feature vector $\{\mathbf{F}'\}$ (step 7) and the principal components contained in matrix [**Z**] (step 8):

$$d_t = \left[\sum_{j=1}^{\text{nret}} \frac{\lambda_j}{\sum_{p=1}^{m_i}\lambda_p}\left(f_j' - z_{tj}\right)^2\right]^{1/2}, \tag{3}$$

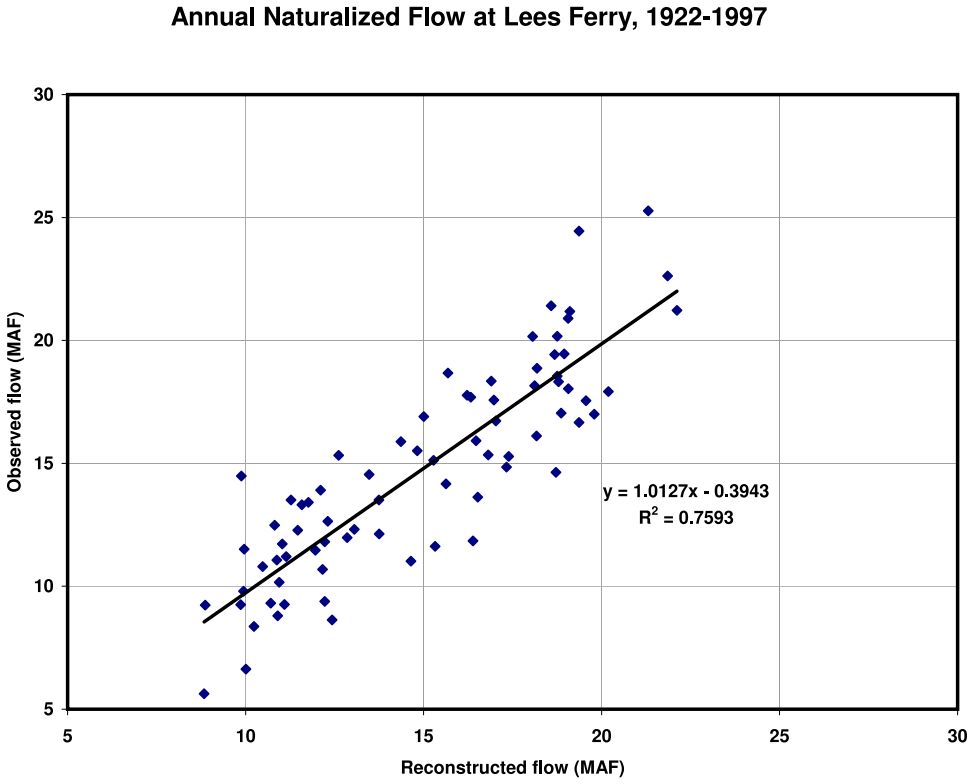

Figure 5. Scatterplot of reconstructed naturalized flow using scheme BSW1997 and observed naturalized flow at Lees Ferry for the period 1922–1997. BSW1997 implies that the reconstruction was done using the bisquare weight function and that chronologies extend up to 1997. The least squares fit line to the scatter data set is also shown in the plot along with the equation and the coefficient of determination R^2.

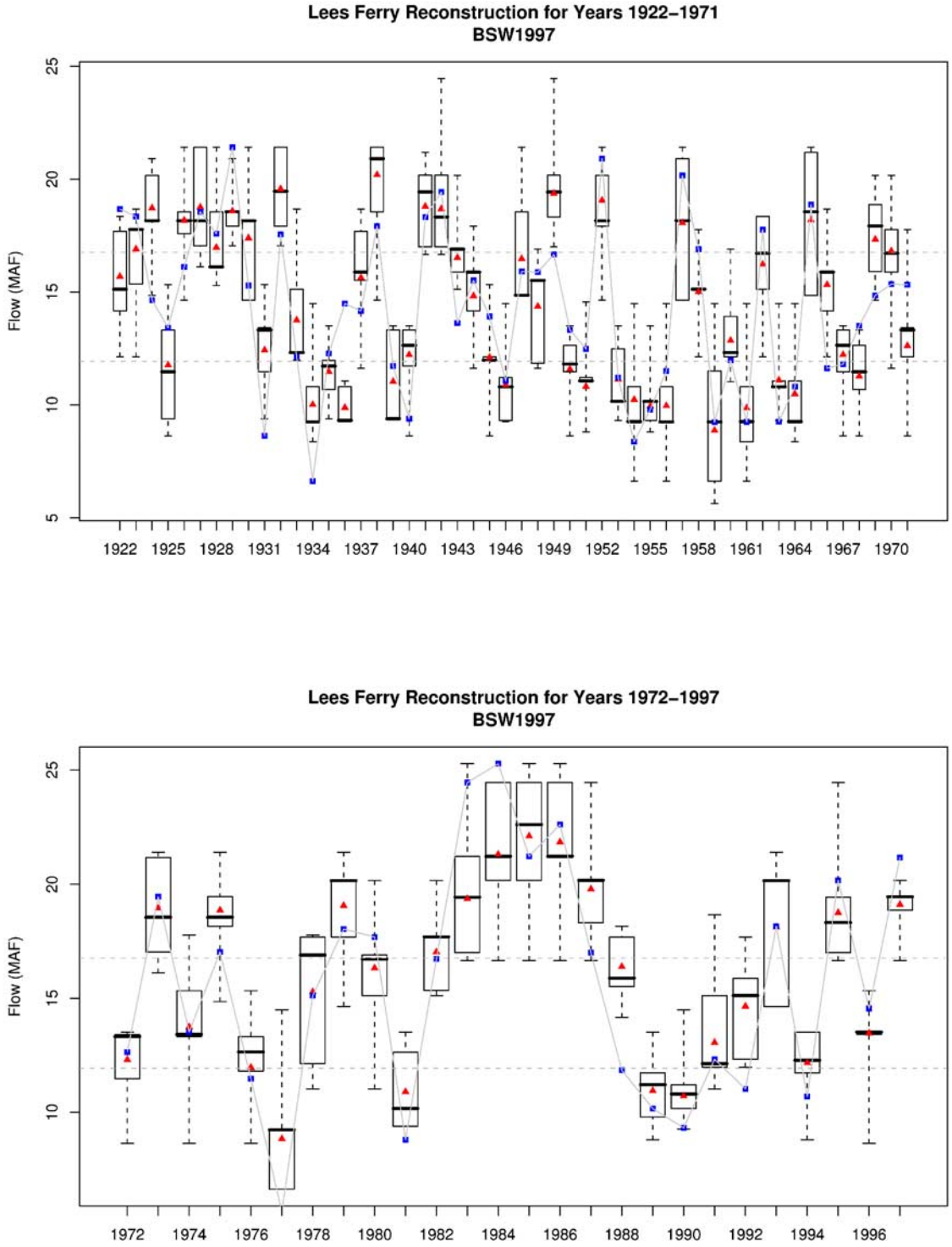

Figure 6. Lees Ferry streamflow reconstructions for years (top) 1922–1971 and (bottom) 1972–1997. The triangle is the expected flow from nonparametric reconstructions, and the squares represent the historical natural flow. Horizontal dashed lines are the terciles calculated from the 1922–1997 flow record.

where nret is the number of principal components retained such that $\sum_{j=1}^{\text{nret}} \lambda_{(j)} \approx 0.90$, z_{tj} are the elements of [**Z**], and f_j' are the elements of the projected feature vector $\{\mathbf{F}'\}$. This gives a set of n distances as possible neighbors from the overlap period to feature year i in the paleohydrologic period.

[22] Step 10. Sort the distances d_t in ascending order and retain only the first K neighbors [*Gangopadhyay et al.*, 2005]. The prescribed choice for K is $\sqrt{n} \approx 9$ in this case. The K nearest neighbors represent the K most similar years from the overlap period to the paleohydrologic feature year i.

Table 2. Coefficient of Determination and Hit Rate of the NPP Algorithm and Seven Existing Reconstructions of Lees Ferry Annual Flows for the 76-Year Verification Period, 1922–1997[a]

Reconstruction	R^2	Hit Rate
NPP	0.76	0.91
Lees A	0.77	0.88
Lees B	0.79	0.92
Lees C	0.70	0.87
Lees D	0.73	0.87
Hidalgo	0.81	0.88
Stockton 1	0.72	0.90
Stockton 2	0.75	0.88

[a]The *Hidalgo et al.* [2000] and the *Stockton and Jacoby* [1976] reconstructions are considered from 1922 to 1961.

[23] Step 11. Select streamflow for each of the K neighbor years from the overlap period, which represents the set of possible streamflow magnitudes for paleohydrologic year i.

[24] Step 12. Assign weights to each of the K streamflow values. Several weighting schemes based either on K [*Lall and Sharma*, 1996; *Rajagopalan and Lall*, 1999] or on distance such as the bisquare weight function [*Gangopadhyay et al.*, 2005] and inverse distance weighting [*Chow et al.*, 1988] are available. We tested our results using these three weighting schemes and found that they produce very similar results. We present results in this paper primarily on the basis of the bisquare weighting (BSW) scheme. The bisquare weight w_k for neighbor k is given by

$$w_k = \frac{\left[1 - \left(\frac{d_{(k)}}{d_{(K)}}\right)^2\right]^2}{\sum_{k=1}^{K}\left[1 - \left(\frac{d_{(k)}}{d_{(K)}}\right)^2\right]^2}. \quad (4)$$

[25] Step 13. Bootstrap [*Venables and Ripley*, 2002] the K streamflow values (step 11) using the weights w_k, $k = 1, \ldots, K$ (step 12) to generate an ensemble of streamflows for year i.

[26] Step 14. For each of the paleohydrologic years 1400–1905 repeat steps 3–13 to obtain an ensemble streamflow reconstruction. In addition, since streamflow quantiles are of significant interest to water managers and decision makers, we summarized our paleohydrologic streamflow ensembles for each of the paleohydrologic reconstruction years from minimum to maximum at a 5% interval.

4. Results and Discussion

[27] The results for the overlap period, 1922–1997, which also serves as verification of the method, are first presented followed by comparisons with parametric reconstructions of *Woodhouse et al.* [2006]. The period 1922–1997 was selected as the verification period because of the uncertainty in the gauge record prior to 1922 (the year the Lees Ferry gauge was installed) and the scarcity of chronologies that extend to 2005.

4.1. Overlap Period Verification

[28] Streamflow ensembles are generated using the nonparametric technique based on tree ring chronologies for the period 1922–1997 in a leave-one-out cross-validation mode. For example, to estimate the flow distribution for the year 1922, data from 1923 to 1997 are used to estimate the EOF projections and loadings, the tree ring data of the feature year (1922) is projected on to the EOFs, and a set of K nearest years are identified. The streamflows from these K years are used to generate an ensemble following the bootstrapping approach using bisquare weights (i.e., the BSW1997 scheme, which is the reconstruction done using the bisquare weight function with chronologies extending up to 1997). We calculate a weighted mean of the generated ensembles to provide a mean value similar to what would be obtained from a regression approach [e.g., *Woodhouse et al.*, 2006].

[29] The mean reconstructed flow and the observed flows are shown as scatterplots in Figure 5 with an $R^2 = 0.76$ (which is significant at the 95% confidence level), implying that the reconstruction captures 76% of the observed flow variance. This cross-validated R^2 is comparable to the fitted R^2 of 0.81, 0.84, 0.72, and 0.77 from MLR for the four reconstructions Lees A, Lees B, Lees C, and Lees D, respectively, given by *Woodhouse et al.* [2006, Table 2]. The standard deviations of the cross-validated mean reconstructed and observed flows are 3.63 and 4.22 MAF (1 MAF = 1.233×10^9 m^3), respectively, which are statistically similar at the 95% confidence level using a nonparametric Fligner-Killeen test [*Conover et al.*, 1981]. The lower variance of our reconstructions is consistent with that of *Woodhouse et al.* [2006]. However, the underlying mechanism of variance reduction is different for the NPP and MLR reconstructions. Observed variance will be reproduced if resampling is done in the same proportion in the NPP approach. But here we are doing a conditional resampling using a weighting scheme, where reconstructed streamflow is resampled from the observed flows on the basis of tree ring information. The variance explained by an MLR model is based on the observed period model fit, i.e., the R^2 of the fitted model. That the cross-validated results from the nonparametric approach are comparable to the estimates from the parametric regression approaches is noteworthy and quite impressive.

[30] Box plots of reconstructed flows along with the observed flows for the 1922–1997 period are shown in Figure 6. The box in a box plot is bounded by the lower quartile (25th percentile) and upper quartile (75th percentile) of the reconstructed flow ensemble. The horizontal line within the box represents the median, and the whiskers are the approximate 5th and 95th percentile confidence bounds. Also, the squares represent the historical natural flow, the triangles are the bisquare weighted mean from the ensembles, and the horizontal dashed lines in the plot (Figure 6) are the lower and upper terciles calculated from the 1922–1997 naturalized streamflow record. The observed flows are mostly within the box (interquartile range of the simulated streamflow ensembles) or close to it. The box plots provide the uncertainty range of the mean estimates. Notice that the boxes are asymmetric around the median (horizontal line within the box), reflective of the nonlinearity and nonnormal features in the data. Whereas, with a regression-

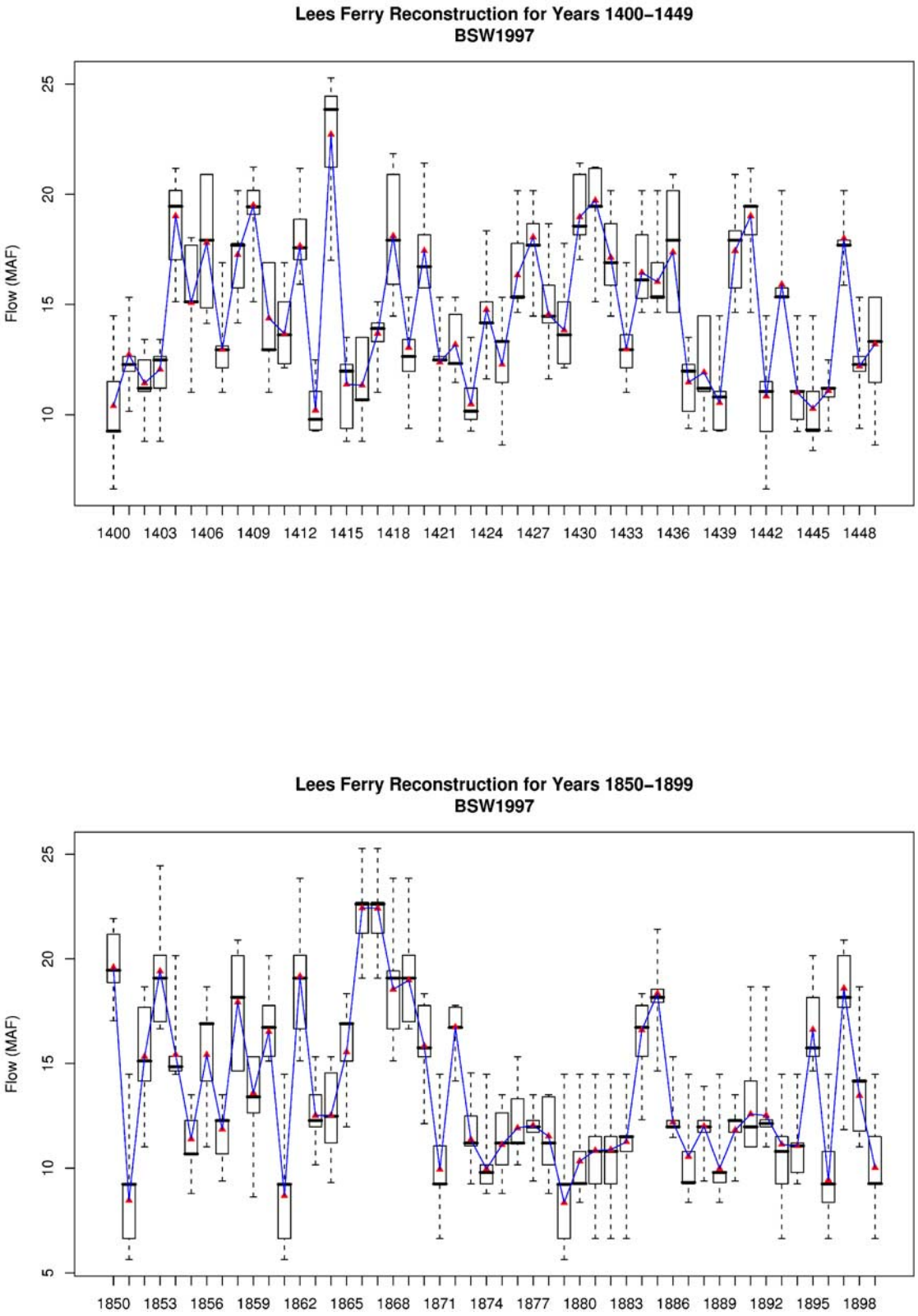

Figure 7. Lees Ferry reconstructions for two selected 50-year periods, (top) 1400–1449 and (bottom) 1850–1899, using the scheme BSW1997. The triangles represent the expected flow.

based parametric model the uncertainty range is generally symmetric and is normally distributed from theory.

[31] It can be seen that the ensembles largely capture the state (wet or dry) of the observed flow very well. To quantify this, we assume two states, wet if the estimated mean flow or the observed flow is above the historical (1922–1997) mean flow of 14.7 MAF and dry if it is below, and compute the hit rate. The hit rate is defined as the sum

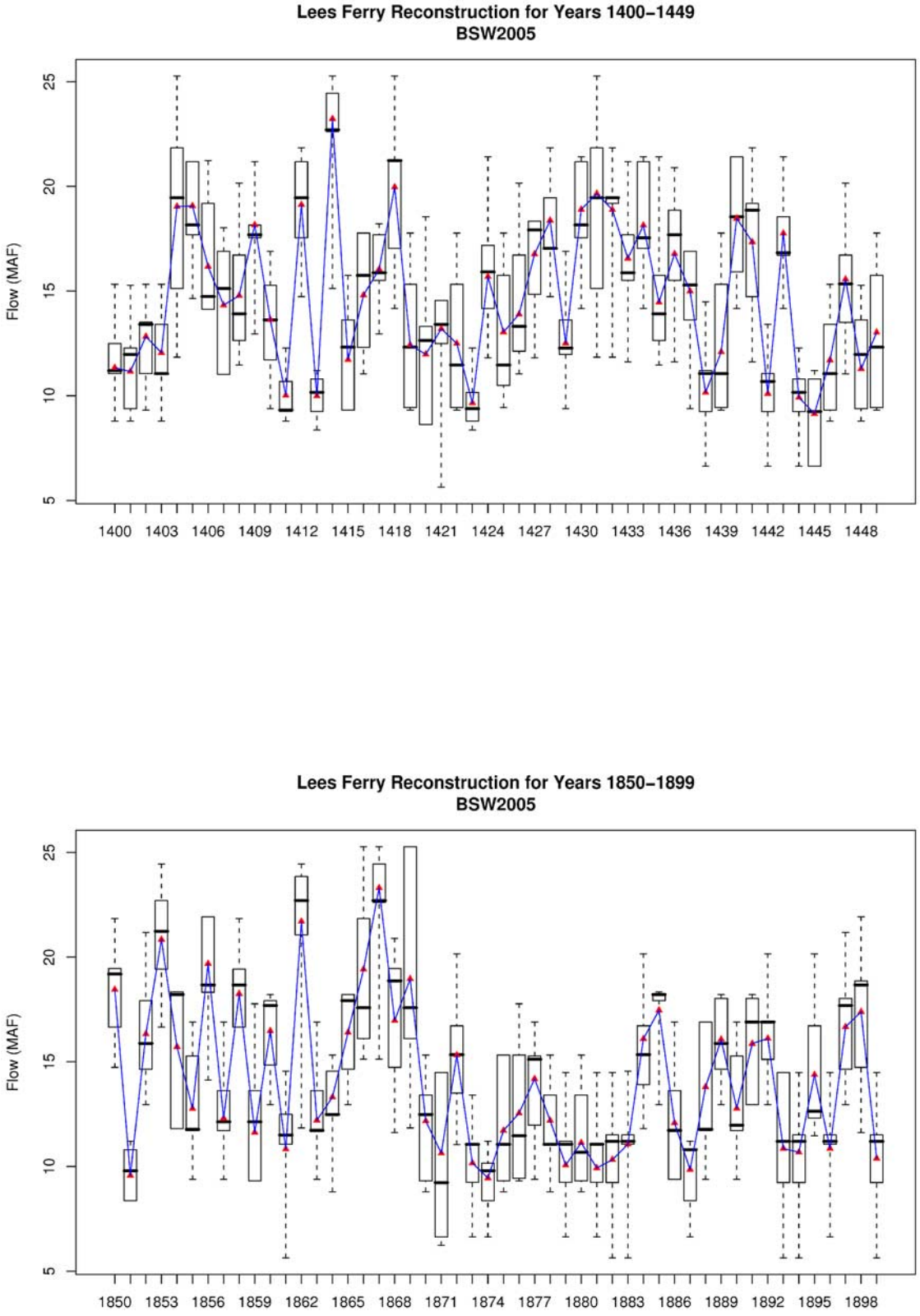

Figure 8. Similar to Figure 7 but using the scheme BSW2005.

of the joint state probability matrix, and for two states (dry and wet) it is the sum of the joint dry and wet probabilities [*Wilks*, 1995]. There are 38 dry and 38 wet years in the 1922–1997 historical record, and the mean flow estimates from the nonparametric method got 34 dry and 35 wet years correctly. The joint probability of dry and wet states is 0.4474 and 0.4605, respectively, resulting in a hit rate of $0.9079 \approx 91\%$; that is, in about 91% of the years (69 out of 76) the nonparametric reconstruction technique correctly simulated the hydrologic state of the system. Similarly, hit

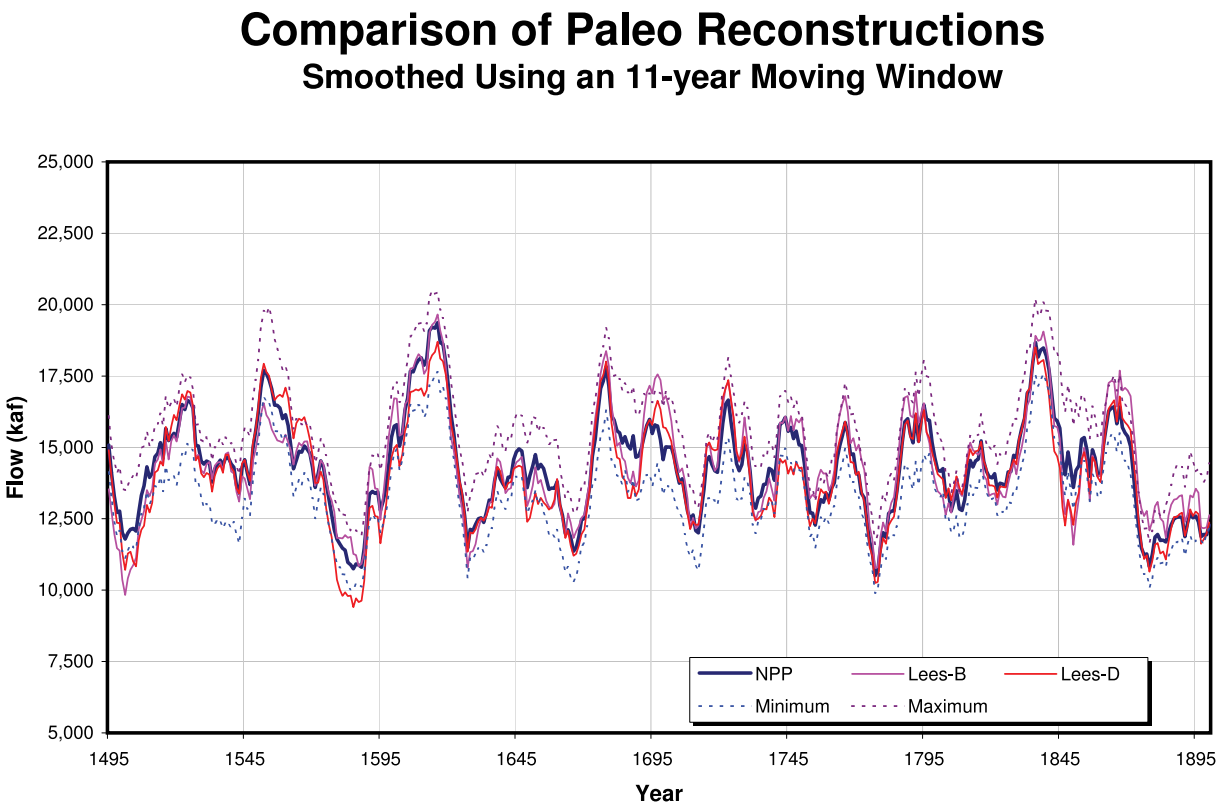

Figure 9. Comparison of paleohydrologic reconstructions, 1495–1900, of Lees Ferry streamflows from the NPP method and the *Woodhouse et al.* [2006] reconstructions, Lees B and Lees D. The streamflow reconstructions were smoothed using an 11-year moving window for this comparison. The minimum and maximum flows from the NPP model provide an estimate of the uncertainty of these reconstructions.

rates and coefficient of determination (R^2) values were computed for the seven existing Lees Ferry reconstructions over the 1922–1997 observed period (the *Hidalgo et al.* [2000] and the *Stockton and Jacoby* [1976] reconstructions are considered from 1922 to 1961) and are given in Table 2. We found that the nonparametric algorithm does extremely well in capturing the state statistic. The ability to capture the hydrologic state is important as it can be coupled with observed streamflows to generate a rich variety of streamflow sequences [*Prairie et al.*, 2006, 2008] useful for system risk and reliability analysis.

4.2. Comparison of Nonparametric and Parametric Reconstructions

[32] Reconstructions for two 50-year periods, 1400–1449, using a variable number of chronologies (refer to Table 1), and 1850–1899, using 51 chronologies, in conjunction with the BSW1997 scheme and the 1906–1997 overlap period, are shown in Figure 7. This shows the ability of the nonparametric approach to easily adapt to a variable sample size, unlike the parametric models, which require new calibration and verification whenever chronologies are either added or dropped. To see the effect of sample size, we also generated reconstructions for the same periods using BSW but with the overlap period of 1906–2005 (Figure 8, scheme BSW2005, which is the reconstruction done using the bisquare weight function with chronologies extending up to 2005). Only seven chronologies were available that span the entire period of 1400–2005; this is mainly because long-lived trees are less and less common as one goes back in time and few chronologies have been updated to 2005. Reconstructions from a smaller sample size (Figure 8) show wider boxes, indicative of sampling variability, relative to those from a bigger sample (Figure 7). This shows the capability of the nonparametric technique to better reflect the uncertainty in the estimates due to sampling variability.

[33] In order to capture the uncertainty arising from sample size and the different weighting schemes (bisquare, inverse distance square, and one over k), we computed the median from the nine possible mean combinations (3 weighting schemes times 3 ending years of the overlap period, 1997, 2002, and 2005, for each of the weighting schemes) as the "reconstructed flow estimate" from the nonparametric method. Reconstructed flow estimates from the nonparametric approach are compared (Figure 9) with two most recent reconstructions from *Woodhouse et al.* [2006], Lees B (based on MLR using standard chronologies) and Lees D (based on PCA and MLR using standard chronologies). For easy visual comparison of the estimates the 11-year running means of the three estimates are shown in Figure 9 along with the maximum and minimum flows from the nonparametric ensemble. All the reconstructions are very similar and fall within the ensemble range from the nonparametric method. To quantify their association, scatterplots of the reconstructed 11-year mean flows from Lees B and Lees D with the nonparametric approach are shown in Figure 10. The nonparametric estimate shows close

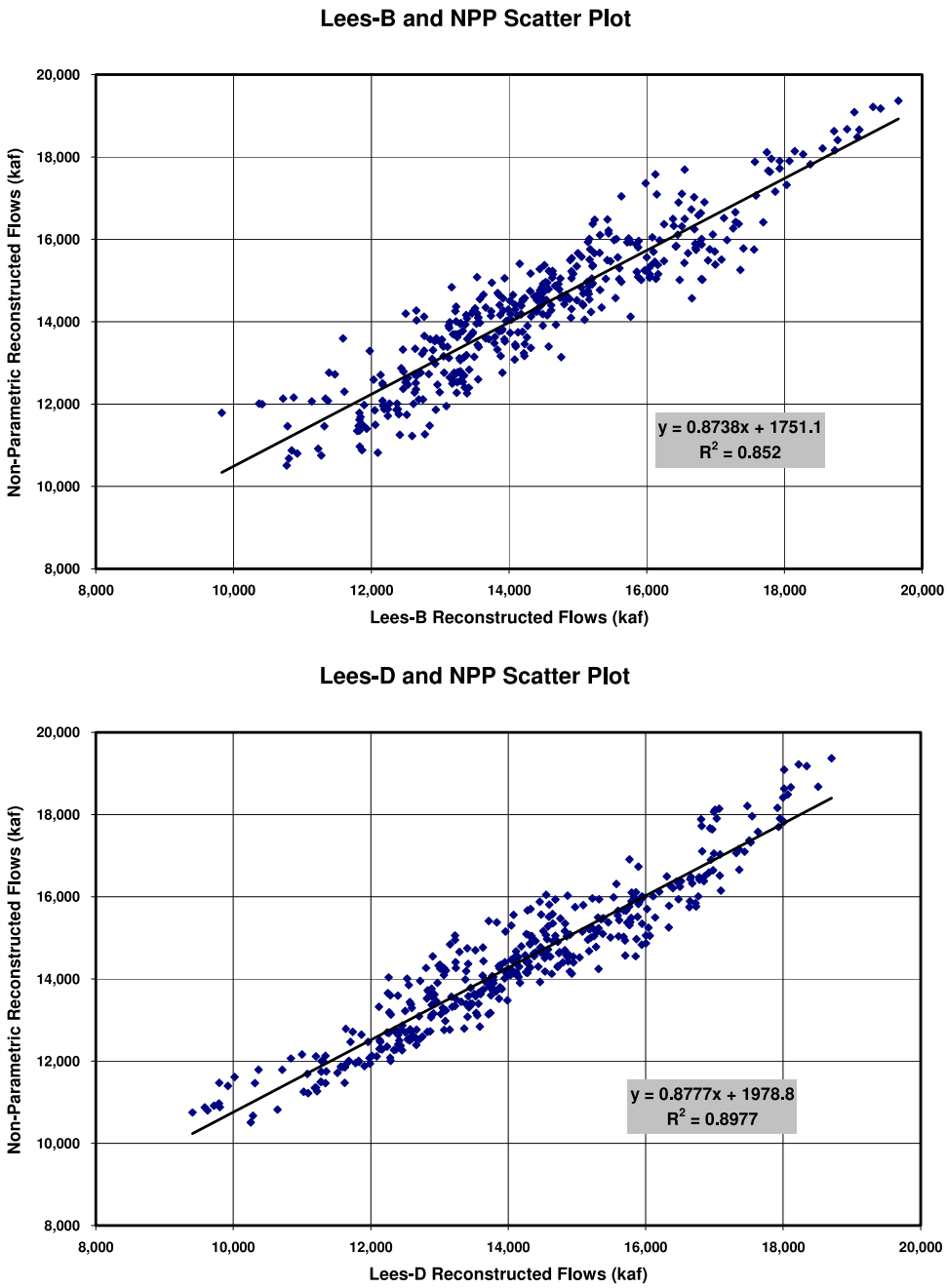

Figure 10. Scatterplots of 11-year moving average values for nonparametrically reconstructed Lees Ferry streamflows and *Woodhouse et al.*'s [2006] (top) Lees B and (bottom) Lees D. The least squares fit line to the scatter data set is also shown in the plots along with the equation and the coefficient of determination R^2.

agreement with the other two methods: R^2 of 0.85 with Lees B and 0.89 with Lees D. The scatterplots indicate that the nonparametric approach tends to generate nominally higher 11-year mean flows than the parametric methods; the box plots of the three reconstructions (Figure 11) show this bias. The box plots are quite similar, but the nonparametric reconstruction has a higher range in the lower flows compared to the parametric reconstructions. The parametric approaches generate lower 11-year mean flows in some periods, such as those centered on 1500, 1585, and 1850, which is the reason for their longer whiskers on the lower-flow side. Regardless, the variance from the three reconstructions is statistically the same at 95% confidence from a nonparametric Fligner-Killeen test [*Conover et al.*, 1981].

[34] The agreement between the nonparametric and parametric reconstructions at annual time scales (not shown) is also very high and is comparable to that for the 11-year reconstructed flows: R^2 of 0.82 with Lees B and 0.89 with Lees D. When examining the annual flows, there is a tendency for the nonparametric method to generate higher estimates of the lowest flows compared to the parametric method. This is because the observed flows which are conditionally resampled provide absolute upper and lower bounds on the reconstructed flows generated by the non-

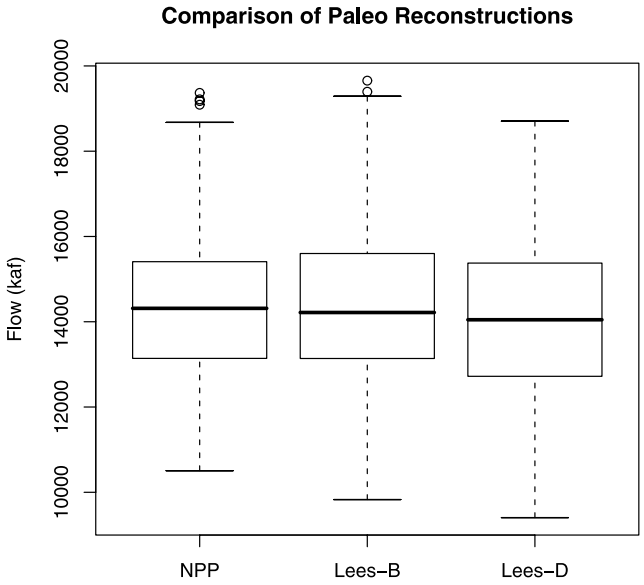

Figure 11. Box plots comparing 11-year moving average values from the NPP method and the parametric *Woodhouse et al.*'s [2006] Lees B and Lees D reconstructions.

parametric method, while the parametric methods are not constrained by the range of the observed flows. In practice, the nonparametric reconstructed flows can approach but will not reach the bounds imposed by the observed record because the estimation from nearest neighbors will tend to pull extreme values toward a central tendency, with that influence dependent on the weighting scheme.

[35] Because of this conditional resampling based on weights, NPP reconstructions have a compressed variance relative to the observed flows, similar to the *Woodhouse et al.* [2006] and other MLR reconstructions. Approaches for circumventing the problem of variance compression have been proposed by *Meko et al.* [2001]. Variance inflation in the NPP approach can be introduced by resampling from a limited number of tree ring chronologies. Also, the fitting of a local polynomial (see section 5) could be used to extend the range of the nonparametric estimates. But it is worth noting that in the context of water management, the very robust dry/wet state information in tree ring data, and the resulting sequences of dry and wet years, is probably of greater relevance than the specific annual flow magnitudes, which will be less certain than the state information regardless of the method used to specify them. However, the NPP reconstructions provide a robust characterization of uncertainty using ensembles and confidence intervals that are asymmetric. Asymmetric confidence intervals are extremely important as water managers use the threshold exceedance to estimate system risk and reliability.

[36] As a last point, it is worthwhile to note that there are many different aspects to the setup of a reconstruction and that differences in results can have sources other than the choice of statistical model. For example, the *Woodhouse et al.* [2006] reconstructions used 62 chronologies, and in this study a subset of 38 of those chronologies, along with 13 other chronologies, were used. That the results of the two reconstruction approaches are so similar is evidence of the robustness of the tree ring data in describing the regional hydrologic condition as well as the utility of the NPP method in developing paleohydrologic reconstructions.

5. Summary and Conclusions

[37] A nonparametric approach that is flexible, simple, and data driven is presented for generating paleohydrologic ensemble streamflow reconstruction on the basis of tree ring chronologies. In this approach, neighbors (i.e., years in the overlap period) for each year in the reconstruction period are identified in a mutually orthogonal eigenspace and are conditionally bootstrapped to provide ensembles. This eliminates the multicollinearity (correlation among variables) that is often a serious issue in traditional MLR. The selection of neighbors is guided only by the tree ring data, not by the observed flows. In essence, the pattern of association between the chronologies in the overlap period and reconstruction period is identified for each reconstruction year and is the basis for the reconstruction. This information is the foundation of the reconstruction process from which analog streamflow values are selected for each of the reconstruction years. Application to streamflow reconstruction at Lees Ferry gauge on the Colorado River shows that the method compares well with the parametric reconstructions. The local estimation of neighbors provides the ability to capture local features (nonlinearities) that might be present which cannot be captured with a single regression equation in the MLR approach. The ensembles provide for asymmetric confidence and uncertainty estimates reflective of the underlying nonnormal features and sampling variability, while the MLR generally results in a symmetric interval based on normal distribution. For MLR, data have to be normally distributed or they have to be transformed before fitting the model. This is obviated in the nonparametric approach. Furthermore, the nonparametric method can provide ensembles easily from varying sample length as seen in the application presented.

[38] The reconstructed ensembles from the nonparametric method are combinations of observed streamflows; thus, the variety is limited to the length of the overlap period. This apparently can be limiting. This limitation can be alleviated by fitting a local polynomial [*Loader*, 1999] to the K nearest neighbors and using it for the mean flow estimation. This approach was developed and demonstrated for streamflow simulation [*Prairie et al.*, 2006] and ensemble streamflow forecast [*Grantz et al.*, 2005].

[39] Given the efficiency and simplicity of the approach, it can be used in the reconstruction of other hydrologic markers such as the Palmer drought severity index and for hydrologic record extensions and can be used to develop ensembles of model inputs for water allocation modeling. This is possible because the NPP algorithm identifies neighbors (i.e., similar years) using only tree ring chronologies, thereby allowing a model or any other hydrologic variable of interest to be resampled from the ensemble of similar years. It is important to note that the tree ring data should be robust to reflect the variability of the water supply, and to that end the trees in the Upper Colorado River Basin are particularly well tuned to the water year. This is mostly because of the snowpack that is important to both the growth of the trees and the water supply for this region. Thus, the performance of the NPP method is implicitly tied to the degree of this relationship.

[40] **Acknowledgments.** This research was funded by the NOAA Climate Program Office through the NOAA-CIRES Western Water Assessment Program and by the U.S. Bureau of Reclamation (USBR). The authors thank Connie Woodhouse, David Meko, and an anonymous reviewer for their comments which helped to improve the manuscript. The authors of this paper would like to dedicate this research work to the memory of USBR staff member Nan Yoder.

References

Chow, V. T., D. R. Maidment, and L. W. Mays (1988), *Applied Hydrology*, 572 pp., McGraw-Hill, New York.

Conover, W. J., M. E. Johnson, and M. M. Johnson (1981), A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, *Technometrics*, *23*, 351–361, doi:10.2307/1268225.

Cook, E. R., and K. Briffa (1990), A comparison of some tree-ring standardization methods, in *Methods of Dendrochronology: Applications in the Environmental Sciences*, edited by E. R. Cook and L. A. Kairiukstis, pp. 153–162, Springer, New York.

Cook, E. R., K. Briffa, S. Shiyatov, and V. Mazepa (1990), Tree-ring standardization and growth-trend estimation, in *Methods of Dendrochronology: Applications in the Environmental Sciences*, edited by E. R. Cook and L. A. Kairiukstis, pp. 104–123, Springer, New York.

Cook, E. R., C. A. Woodhouse, C. M. Eakin, D. M. Meko, and D. W. Stahle (2004), Long-term aridity changes in the western United States, *Science*, *306*, 1015–1018, doi:10.1126/science.1102586.

Deutsch, C. V., and A. G. Journel (1992), *Geostatistical Software Library and User's Guide*, 340 pp., Oxford Univ. Press, New York.

Fritts, H. C. (1976), *Tree Rings and Climate*, 567 pp., Academic, London.

Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using *K*-nearest neighbors, *Water Resour. Res.*, *41*, W02024, doi:10.1029/2004WR003444.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resour. Res.*, *41*, W10410, doi:10.1029/2004WR003467.

Haan, C. T. (1977), *Statistical Methods in Hydrology*, 378 pp., Iowa State Univ. Press, Ames.

Helsel, D. R., and R. M. Hirsch (2002), Statistical methods in water resources, *U.S. Geol. Surv. Tech. Water Resour. Invest., Book 4*, *Chap. A3*, 510 pp.

Hidalgo, H. G., T. C. Piechota, and J. A. Dracup (2000), Alternative principal components regression procedures for dendrohydrologic reconstructions, *Water Resour. Res.*, *36*, 3241–3249, doi:10.1029/2000WR900097.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for time-series resampling, *Water Resour. Res.*, *32*, 679–693, doi:10.1029/95WR02966.

Loader, C. (1999), *Local Regression and Likelihood*, Springer, New York.

Meko, D. M., and D. A. Graybill (1995), Tree-ring reconstruction of Upper Gila River discharge, *Water Resour. Bull.*, *31*, 605–616.

Meko, D. M., C. W. Stockton, and W. R. Boggess (1995), The tree-ring record of severe sustained drought, *Water Resour. Bull.*, *31*, 789–801.

Meko, D. M., M. D. Therrell, C. H. Baisan, and M. K. Hughes (2001), Sacramento River flow reconstructed to A.D. 869 from tree rings, *J. Am. Water Resour. Assoc.*, *37*(4), 1029–1040, doi:10.1111/j.1752-1688.2001.tb05530.x.

Myers, R. H. (1990), *Classical and Modern Regression with Applications*, 488 pp., PWS, Boston, Mass.

Prairie, J., B. Rajagopalan, T. Fulp, and E. Zagona (2006), Modified K-nn model for stochastic streamflow simulation, *J. Hydrol. Eng.*, *11*(4), 371–378, doi:10.1061/(ASCE)1084-0699(2006)11:4(371).

Prairie, J., K. Nowak, B. Rajagopalan, U. Lall, and T. Fulp (2008), A stochastic nonparametric approach for streamflow generation combining observational and paleo reconstructed data, *Water Resour. Res.*, *44*, W06423, doi:10.1029/2007WR006684.

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

Smith, L. P., and C. W. Stockton (1981), Reconstructed streamflow for the Salt and Verde rivers from tree-ring data, *Water Resour. Bull.*, *17*, 939–947.

Stockton, C. W., and G. C. Jacoby (1976), Long-term surface-water supply and streamflow trends in the Upper Colorado River Basin based on tree-ring analyses, *Lake Powell Res. Proj. Bull. 18*, 70 pp., Res. Appl. to Natl. Needs, Natl. Sci. Found., Arlington, Va.

Stokes, M. A., and T. L. Smiley (1968), *An Introduction to Tree-Ring Dating*, Univ. of Chicago Press, Chicago, Ill.

Venables, W. N., and B. D. Ripley (2002), *Modern Applied Statistics with S*, 495 pp., Springer, New York.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, 467 pp., Elsevier, New York.

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

T. J. Fulp, Lower Colorado Region, U.S. Bureau of Reclamation, Box 61470, Boulder City, NV 89006, USA.

S. Gangopadhyay and B. L. Harding, Earth and Environmental Division, AMEC, 1002 Walnut Street, Suite 200, Boulder, CO 80302, USA. (s.gangopadhyay@amec.com)

J. J. Lukas, Institute of Arctic and Alpine Research, University of Colorado at Boulder, Campus Box 450, Boulder, CO 80309, USA.

B. Rajagopalan, Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Campus Box 428, Boulder, CO 80309, USA.