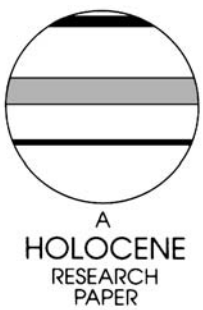
The Holocene 14,4 (2004) pp. 502–516

Alternative methods of proxy-based climate field reconstruction: application to summer drought over the conterminous United States back to AD 1700 from tree-ring data

Zhihua Zhang,[1]* Michael E. Mann[1] and Edward R. Cook[2]

([1]Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia, 22903, USA; [2]Tree-Ring Laboratory, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, 10964, USA)

Received 1 April 2003; revised manuscript accepted 10 June 2003



Abstract: We describe an alternative method of climate field reconstruction and test it against an existing set of dendroclimatic reconstructions of summer drought patterns over the conterminous US back to AD 1700. The new reconstructions are based on a set of 483 drought-sensitive tree-ring chronologies available across the continental US. In contrast with the 'point-by-point' (PPR) local regression technique used previously, the tree-ring data were calibrated against the instrumental record of summer drought (June–August Palmer Drought Severity Index (PDSI)) based on application of the 'Regularized Expectation Maximization' ('RegEM') algorithm to relate large-scale patterns of variation in proxy and instrumental data over a common (twentieth century) interval. A screening procedure was first used to select an optimal subset of candidate tree-ring drought predictors, and the predictors (tree-ring data) and predictand (instrumental PDSI) were prewhitened prior to calibration, with serial correlation added back into the reconstruction at the end of the procedure. The PDSI field was separated into eight relatively homogenous regions of summer drought through a cluster analysis, and three distinct calibration schemes were investigated: (i) 'global' (i.e., entire conterminous US domain) proxy data calibrated against 'global' PDSI; (ii) regional proxy data calibrated against regional PDSI; and (iii) global proxy data calibrated against regional PDSI. The greatest cross-validated skill was evident for case (iii), suggesting the existence of useful non-local information in the tree-ring predictor set. Cross-validation results based on withheld late nineteenth/early twentieth-century instrumental data, as well as a regionally limited extension of cross-validation results back to the mid-nineteenth century based on long available instrumental series, indicate a modest improvement in reconstructive skill over the PPR approach. At the continental scale, the 1930s 'Dust Bowl' remains the most severe drought event since 1700 within the context of the estimated uncertainties, but more severe episodes may have occurred at regional scales in past centuries.

Key words: Drought, tree rings, dendroclimatology, climate reconstruction, regression, regularized expectation maximization, conterminous United States.

Introduction

Droughts and floods can be among the most devastating of climate-related natural hazards facing the United States. The 'Dust Bowl' droughts of the 1930s and 1950s, which, at their most severe, covered 70% of the conterminous US and persisted for 5–7 years at a time, incurred an estimated cost of $39 billion, including losses in energy, water resources, ecosystems and agriculture (Riebsame *et al.*, 1991). It is thus of critical importance to determine the likelihood of such

*Author for correspondence (e-mail: zz9t@virginia.edu)

10.1191/0959683604hl727rp

droughts occurring in the future. More generally, it is important to estimate the natural frequency and intensity of extended drought, and the possible impacts of climate change on patterns of continental drought. Interannual and decadal variability in drought patterns over the US is subject not only to the influence of patterns of large-scale climatic variability associated with the El Niño-Southern Oscillation (ENSO), and North Pacific (NPO) and North Atlantic (NAO) oscillations (e.g., Rajagopalan *et al.*, 2000), but also possible natural external forcing (Mitchell *et al.*, 1979; Cook *et al.*, 1997). Climate change, moreover, may have altered patterns of hydroclimatic variability in the United States during the late twentieth century (e.g., Karl and Knight, 1998; Karl *et al.*, 1996). It is thus unlikely that the temporal and spatial patterns of drought recorded in the relatively short instrumental record of the past 100 years are adequate to characterize the full potential range of drought (e.g., Woodhouse and Jonathan, 1998). It is therefore essential, in placing modern drought in an appropriate long-term context, to make use of the more limited information available to reconstruct drought patterns in past centuries.

The most promising reconstructions of past continental drought have employed proxy climate data, and, in particular, tree-ring or 'dendroclimatic' indicators which offer the hydrological sensitivity, spatial availability and annual resolution necessary to reconstruct large-scale patterns of drought (see, for example, D'Arrigo and Jacoby, 1991; Fritts, 1991; Hughes and Brown, 1992; Stahle and Cleaveland, 1992; Graumlich, 1993; Meko *et al.*, 1993; Hughes and Graumlich, 1996; Cook *et al.*, 1999). Employing 425 potentially drought-sensitive tree-ring chronologies across the United States as candidate predictors, Cook *et al.* (1999) demonstrated the ability to skilfully reconstruct the Palmer Drought Severity Index (PDSI) over the conterminous United States back to 1700. The method of reconstruction used, point-by-point regression (PPR), explicitly assumes that tree-ring chronologies proximal to a given PDSI gridpoint are most likely to provide successful predictors of drought for that gridpoint. This approach seems generally appropriate for characterizing continental drought, which is often more regional in nature than other purely climate fields, owing, for example, to the greater influence of land surface heterogeneity and topography. Given the importance of large-scale phenomena such as ENSO on large-scale patterns of drought, it is nonetheless likely that the appropriate use of teleconnected relationships between predictors and predictand can provide increased skill in reconstructing past drought patterns.

Methods of climate field reconstruction (CFR) that make use of nonlocal relationships through the use of large-scale covariance information (e.g., Smith *et al.*, 1996; Kaplan *et al.*, 1997) in the calibration of instrumental data against networks of proxy climate indicators have successfully been applied to the reconstruction of past large-scale surface temperature (Mann *et al.*, 1998; 1999; Evans *et al.*, 2002) and atmospheric circulation (Fritts *et al.*, 1971; Luterbacher *et al.*, 2002) patterns. In some cases, different methods of CFR yield quite similar results. For example, Cook *et al.* (1994) reviewed and compared two alternative spatial regression approaches to palaeoclimate reconstruction, orthogonal spatial regression (OSR) and canonical regression (CR), and concluded that the differences were, in practice, minor. Differences in reconstructions resulting from application of different methods can be more significant, however, when the calibration period is short, and the methods make different use of large-scale covariance information within and between predictor fields (e.g., Schneider, 2001).

Particularly in the limit of a relatively short (i.e., less than one century) calibration period, weak or unstable climate teleconnections between proxy and/or instrumental data at distant spatial scales can potentially limit the utility of methods of CFR that exploit large-scale correlation structure. The utility of covariance-based CFR approaches represents a delicate compromise between the incorporation of both potentially physical and potentially spurious distant statistical relationships in the calibration process. When the calibration period is short, it may be difficult to separate out real large-scale relationships between predictor and predictand fields from spurious ones. Cole and Cook (1998), for example, observed significant decadal variability in the apparent influence of ENSO on drought patterns in the United States back through the late nineteenth century. Indeed, owing to the relatively local correlation structure of drought and the relatively dense and homogeneous nature of the continental drought-sensitive dendroclimatic network such as that used by Cook *et al.* (1999), the problem of continental drought reconstruction from such a network presents a useful challenge for establishing the relative strengths and weaknesses of covariance-based CFR approaches relative to simpler approaches.

It is thus of considerable interest to see how results based on the application of covariance-based CFR methods, which exploit large-scale teleconnected variability, compare in their apparent level of skill to methods such as PPR which do not explicitly account for distantly teleconnected variability within and between predictor and predictand data. It should be noted that it is indeed possible to generalize the PPR approach through the incorporation of a regionally adaptive search radius (see the discussion in Cook *et al.*, 1999) to better accommodate larger-scale correlation structure, and such modifications do appear to yield an improvement in reconstructive skill (Cook, unpublished data). It can nonetheless be argued that the use of empirically determined basis functions, as in covariance-based CFR techniques, is the most natural means of identifying and making use of large-scale coherent structure in CFR.

In this study, we employed such a recently proposed method of CFR (Schneider, 2001) to the problem of reconstructing PDSI patterns over the conterminous US from a similar dendroclimatic network to that used by Cook *et al.* (1999). The method is based on a regularized expectation maximization algorithm (RegEM), which offers some theoretical advantages over previous methods of CFR. This approach calibrates the proxy data set against the instrumental record by treating the reconstruction as initially missing data in the combined proxy/instrumental data matrix, and optimally estimating the mean and covariance of the combined data matrix through an iterative procedure which yields a reconstruction of the PDSI field with minimal error variance (Schneider, 2001; Rutherford *et al.*, 2003; Mann and Rutherford, 2002). We first describe the instrumental and proxy data used in the study. Secondly, we discuss the various strategies for drought reconstruction taken in this study and the statistical reconstruction methodologies employed, including data preprocessing and details of the 'RegEM' technique. Thirdly, we measure the relative skill of the drought reconstructions in the context of previous reconstructions based on cross-validation results using withheld late nineteenth/early twentieth-century data, and fourthly we discuss details of the long-term reconstructions.

Data

We used a version of the 155-point grid (2° lat $\times$ 3° long) of instrumental PDSI data developed by Cook *et al.* (1999) for their map correlation and congruency analyses. This gridded

instrumental data set is based on monthly PDSI records estimated from state climate division temperature and precipitation data over the period 1895–1995. It differs from the gridded PDSI data used by Cook *et al.* (1996; 1999) to reconstruct past drought from tree rings, which was based on 1036 single-station monthly PDSI records estimated from the Historical Climatology Network (HCN) (Karl *et al.*, 1990). While this difference in predictand data sets does not allow for exact comparisons of PPR and RegEM performance, calibration and verification tests using the climate division-based PDSI data as predictands in PPR were extremely similar to those described in Cook *et al.* (1999). The PDSI data used here also have the advantage of beginning in 1895 at all 155 gridpoints. In contrast, only 97 of the 155 gridpoints in the PDSI grid based on single-station records extended back to 1895 (Cook *et al.*, 1999). As discussed later, it is also possible to extend the instrumental PDSI data set back to 1870 over a modest subset of gridpoints based on multivariate regression of the more recent PDSI data (1895–1995) on long monthly surface temperature and precipitation data available at a smaller number of stations in the HCN network.

The Palmer Drought Severity Index (PDSI) was developed as an integrated measure of moisture balance which closely approximates the societally relevant notion of drought (Palmer, 1965). Monthly PDSI, which is related to soil and runoff conditions, as well as integrated precipitative input, depends both on current and past monthly precipitation and temperature values, with varying weight. Because similar factors dominate the seasonal growth of trees in many regions, summer (June–August) PDSI patterns appear to be particularly amenable to reconstruction from tree-ring information (see Cook *et al.*, 1996; 1999).

For purposes described later, we subsequently divided the PDSI field into a small number of apparently homogeneous regions of drought (Figure 1) based on application of a cluster analysis to the 1895–1978 PDSI gridpoint data used for calibration of the dendroclimatic indicators. An additional classification based on the 1928–1978 interval used for calibration in the cross-validation exercises yields essentially the same regions, with only three marginal gridpoints observed to shift regions. An independent analysis based on patterns of correlation between neighbouring gridpoints as an estimate of boundaries between core regions yielded essentially the same boundaries between core regions. Our classification into (eight) homogeneous regions, moreover, agrees closely with the summer drought varimax factors for the US as identified by Cook *et al.* (1999) over the interval 1913–1978, with region 1 corresponding well with varimax factor 6, region 2 with factor 8, region 3 with factor 4, region 4 with factor 7, region 5 with factor 1, region 6 with factor 2, region 7 with factor 5 and region 8 with factor 3 (varimax factor 9 in Cook *et al.*, Figure 9, indicates a domainwide drought pattern, and thus has no counterpart in the cluster analysis). These comparisons establish the robustness of our classification of regions of homogenous drought variability over the conterminous US.

We used a network of 483 tree-ring chronologies available over the conterminous United States as candidate predictors in the subsequent dendroclimatic PDSI reconstructions. This network (Figure 2) represents a modest expansion of the network of 425 chronologies used by Cook *et al.* (1999), but subsequent analyses show little detectable difference in cross-validation statistics upon inclusion of the additional 58 chronologies. This latter finding is consistent with the expectation that the few (i.e., eight or so as estimated above) effective spatial degrees of freedom in conterminous US drought variability are reasonably saturated by a modest uniformly distributed set of candidate predictors. All chronologies are available back to at least 1700. As evident in Figure 2, the network is sparse in the centre of the continent due to the lack of forest land and the limits of natural forest communities (see Cook *et al.*, 1996).

As described below, an additional screening procedure was used to filter the entire network of indicators for those chronologies most likely to be useful in palaeoclimate reconstruction.

Methods

Calibration schemes

Three alternative calibration schemes making use of the RegEM method (see below) were tested in this study. The three schemes differ in how the selected dendroclimatic drought indicators (see below) were calibrated against the large-scale PDSI field, allowing for a variable representation of local and large-scale relationships between and within the predictor (proxy) and predictand (gridded instrumental PDSI field) data. As discussed above, the instrumental PDSI field was classified into eight distinct roughly homogenous regions of drought, establishing a natural means of characterizing both 'regional' (within one of the eight core regions) and 'global' (anywhere within the conterminous US) domains.

Figure 1 Map of the continental United States showing locations of the PDSI gridpoints used for drought reconstructions in this study. The grid spacing is 2° latitude × 3° longitude. The eight homogeneous regions of discussed in the text are labelled.

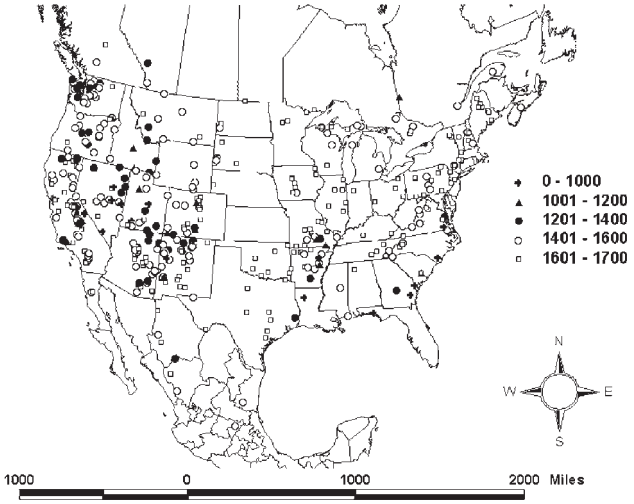


Figure 2 Map of locations of the 483 annual tree-ring chronologies used in this study. All chronologies are available back to 1700, with some available significantly farther back in time as indicated.

The first of the three schemes (i) involves the calibration of 'global' proxy data against the 'global' PDSI field by the simultaneous calibration of the entire proxy and entire instrumental PDSI network. The second scheme (ii) involves the calibration of 'regional' proxy data against the 'regional' PSDI, through separate application of this process repeatedly for each of the eight domains, restricting both predictors and predictand to a given domain during each of eight distinct calibrations. The third scheme (iii) involves the calibration of 'global' proxy data against 'regional' PDSI. The interpretation of these three distinct schemes is clear; the first scheme allows for large-scale relationships both within and between predictor and predictand data, and should produce the most skilful results if both the proxy data and the PDSI field itself both contain nonlocal, large-scale correlation information. The second scheme, by contrast, presumes no large-scale relationships either between or within the predictor and predictand data, and should produce the greatest skill if there is no true large-scale information between or within the predictor and predictand data. The third scheme allows for large-scale relationships within the predictor data and between predictor and predictand data, but assumes that there is no large-scale information within the predictand data itself. This scheme should produce the greatest skill if drought patterns themselves are regional in nature with no robust large-scale structure, but the proxy data (through their more complex statistical dependence on climate) themselves contain nonlocal climatic information. An example of how such relationships might be important is that proxies which are sensitive to winter ENSO influences in the southwestern US (e.g., Stahle *et al*., 1998) might contain significant information regarding summer drought in other regions which experience warm-season ENSO influences simply through their ability to specify the phase of the ENSO signal. Changing drought teleconnections over the twentieth century (e.g., Cole and Cook, 1998), which imply a potential instability of covariance estimates based on a short (e.g., twentieth century) period, present a greater problem for the 'global versus global' approach than for the 'global versus regional' approach. The latter approach does not make of use of the large-scale covariance structures of the short instrumental record, but does make use of the large-scale covariance information in the considerably longer-term proxy data, which can be more robustly estimated. This 'global versus regional' approach is thus more likely to establish the true underlying spatial relationships between predictor and predictand fields in the face of interdecadal variability in teleconnection patterns. The relative performance of these three distinct schemes can help clarify the importance of making use of teleconnected variability both in the tree-ring network and the PDSI field in reconstructing past drought patterns.

Candidate predictor selection

Similar to Cook *et al*. (1999), we employed a screening process to prefilter the full network of candidate predictors (483 or 425 depending on whether the full network or network equivalent to that used by Cook *et al*. was used) for those series most likely to be useful in climate reconstruction. It was impossible to employ an identical screening procedure to that used by Cook *et al*. (1999) owing to the nonlocal nature of the ('RegEM') CFR method, which contrasts with the local nature of the 'PPR' method used by Cook *et al*. (1999). Thus, a variety of alterative screening criteria were explored for comparison.

For the 'global versus global' scheme (i) described above, a single threshold screening correlation (|Rc|) was employed for including a chronology, based on the requirement that a chronology exhibit a 95% significant two-sided correlation with at least one of the PDSI gridpoints in the global domain (this corresponded to $|Rc| = .276$ for the 51-year (1928–78) calibration interval used). A more flexible criterion was also employed in which the threshold value of |Rc| was increased or decreased (equivalent to increasing or decreasing the required level for significance from 95%) in such a way as to maximize the global cross-validated resolved variance. This procedure arguably improves the reconstruction by employing a more selective subset of indicators. However, the dependence on the cross-validation results removes the objectivity of the cross-validation procedure, requiring additional independent cross-validation exercises (as discussed below) to independently establish statistical skill (alternatively, one could make use of experiments using synthetic networks derived from coupled model simulations to objectively tune the RegEM procedure – Rutherford *et al*., unpublished data). The correlation thresholds used here for selecting candidate predictors differ from the one fixed value ultimately selected by Cook *et al*. (1999) based on experimenting with a number of thresholds. In their case, Cook *et al*. (1999) found that a 90% significance level threshold, corresponding to $|Rc| = .240$, represented a near-global optimum.

For the 'regional versus regional' and 'global versus regional' schemes ('ii' and 'iii' respectively) both globally fixed and regionally variable values of |Rc| were employed. In the first case, a fixed threshold of $|Rc| = .276$ was used to insure a 95% significant correlation of a given candidate tree-ring series with at least one of the PDSI gridpoints *within the selected region*, while in the second case a regionally variable |Rc| was selected such that the cross-validated resolved variance was maximized on a regional basis.

Prewhitening procedure

As the PDSI, by construction, represents a seasonally integrated representation of hydrological balance, PDSI time series exhibit considerably greater persistence than time series of other typical climatic variables (e.g., surface air temperature or sea-level pressure). It is particularly important, therefore, to take the serial persistence structure of the time series explicitly into account. As in Cook *et al*. (1999), prior to calibration, both the predictors (tree-ring data) and predictand (instrumental PDSI) were prewhitened. This procedure allows the potentially differing levels of serial correlation between instrumental drought data and drought-sensitive tree-ring chronologies (the latter exhibiting temporal autocorrelation due both to the serial correlation in drought, and nonclimatic serial correlation associated with stand dynamics, other nonclimatic influences on tree growth, and internal physiologically based feedbacks) to be removed during the calibration process. The estimated autocorrelation as modelled and removed from the instrumental PDSI data over the calibration interval is added back to the reconstructions of the prewhitened PDSI, restoring the estimated climatic serial correlation of the calibration period to the final PDSI reconstructions (see Cook *et al*., 1999). It is important to keep in mind the potential limitations of such a procedure. Without further, more involved statistical modelling considerations, such a procedure implicitly 'builds' the serial persistence structure of the modern instrumental data into the entire reconstruction if the time period selected for estimating the tree-ring prewhitening coefficients predates that of the instrumental PDSI data. See Meko (1981) and Appendix A in Cook *et al*. (1999) for details. In contrast to calibration of nonprewhitened predictors and predictand, the persistence structure of the reconstruction is not allowed to change over time. While a significantly greater share of the calibration

period variance may be resolved and well verified using prewhitened data, possible past changes in the serial persistence structure of drought cannot be modelled without more work.

Regularized EM ('RegEM') algorithm

The regularized EM (RegEM) method employed in this study (see Schneider, 2001, and references therein) is an iterative method for estimating missing data through the estimation of means and covariances from an incomplete data field to impute missing values in a manner that makes optimal use of the spatial and temporal information in the data set. When a reconstruction is sought from proxy data based on calibration against modern instrumental measurements, the combined (proxy-plus-instrumental) data set can be viewed as an incomplete data matrix, which contains both instrumental data (PDSI gridpoint values arranged with rows representing the years and columns representing gridpoints) and proxy data (tree-ring indices with rows representing the years and columns representing tree sites). Missing values in this matrix represent the unknown preinstrumental PDSI gridpoint series, and are considered as values to be imputed through an iterative infilling of the data matrix which makes use of the covariance information between all available (instrumental and proxy) data. In analogy with conventional palaeoclimate reconstruction approaches (see, for example, Rutherford *et al*., 2003), an effective 'calibration' interval can be defined as the time interval over which the proxy and instrumental data overlap, while a 'verification' interval is defined by additional cross-validation experiments in which an appropriate subset of the available instrumental data are withheld from the process (e.g., through their specification as missing values in the initial matrix). Schneider (2001) provides a detailed description of the regularized EM algorithm, including a comparison with conventional methods such as principal components regression, and application to the infilling of missing values in climate field data, while Rutherford *et al*. (2003) discuss specific applications to palaeoclimate reconstruction. Here, we summarize the primary features of the methodology relevant to the current analysis.

The RegEM method is analagous to other methods of CFR in which missing spatial data are estimated from sparse early data (or proxy data) through relating the patterns evident in the sparse longer data to the patterns defined by the empirical eigenvectors (EOFS) estimated from a shorter, data-rich interval during which a nearly complete version of the field of interest (e.g., surface temperature) is available (e.g., Smith *et al*., 1996; Kaplan *et al*., 1997; Mann *et al*., 1998). In contrast, however, with purely EOF-based methods in which the available eigenvectors basis set is truncated above some determined cutoff, higher-order patterns are retained, but are diminished in their contributions through the use of a 'regularization parameter' which effectively smoothes out increasingly heterogeneous covariance structures. This procedure thus allows more complete use of the available spatiotemporal information in the reconstruction process. The statistics of the data set are estimated from all available data, including proxy data outside the 'calibration' interval during which proxy and instrumental data overlap, based on an iterative (and thus nonlinear) approach to estimating the complete data matrix.

One concern that arises with this, and other similar approaches, to CFR is that the reconstructions of past anomalies may be biased by nonstationarity in the covariance information, particularly in the face of possible recent anthropogenic influences on patterns of climate present in the recent, more data-rich interval. Rutherford *et al*. (2003) tested the RegEM method with both instrumental data and control and forced model surface temperature fields, and found that if radiative forcing is relatively stationary over a data-sparse period (e.g., an older interval with no instrumental data) and increases rapidly over a data-rich period (e.g., the more recent interval containing both instrumental data and proxy data) the imputed anomalies over the data-sparse period remain essentially unbiased as long as an adequate length (multidecade) calibration interval is available. It thus appears that use of the data-rich twentieth-century instrumental record (which may contain trends that that are, at least in part, associated with the effects of anthropogenic climate forcing) in the calibration process does not significantly bias reconstructions of climate in previous centuries. In fact, such considerations are probably less important in reconstructions of continental drought, as there is no evidence for nonstationary behaviour in the mean (although, as discussed above, there is reasonable evidence of nonstationary patterns of drought response to tropical ENSO forcing). Additional experiments have been performed with RegEM making use of synthetic proxy data networks ('pseudoproxies') to establish the level of skilful resolved variance that might reasonably be expected in climate field reconstructions based on proxy data networks of varying size and statistical quality (Mann and Rutherford, 2002).

The regularized EM algorithm is similar to the conventional EM ('Expectation Maximization') algorithm for estimating the means and covariances of a data matrix. The estimation problem is 'regularized', however, in that a 'ridge parameter' is used to inflate the diagonal elements of the covariance matrix so as to avoid the problem of estimating the eigenstructure of a rank-deficient matrix. The algorithm starts with initial estimates of the mean and data covariance matrix and iteratively refines these estimates until they approach asymptotic values.

The regularization parameter h in the RegEM algorithm effectively plays a similar role to the choice of how many eigenvectors to retain in EOF-based reconstruction methods (e.g., Kaplan *et al*., 1997; Mann *et al*., 1998), controlling the tradeoff between retained variance and the degree of smoothing and spatiotemporal noise suppression. Thus, the selection of h is a key decision in the RegEM approach. If the value of h is too small, then the imputed values will be increasingly compromised by sampling error, while if h is too large the imputed values will tend towards the data mean values, leading to an underestimate of the data variance (and thus, an increased regularization error). The optimal value of h should consequently minimize the total imputation error (the sum of the sampling error and the regularization error), and can be estimated by generalized cross-validation (GCV). In practice, a well-defined global minimum in the GCV function is often difficult to obtain, and a number of additional practical constraints on the selection of h must be employed (see Schneider, 2001, and Rutherford *et al*., 2003, for a more detailed discussion).

In our procedure, the missing (prewhitened, as discussed above) PDSI estimates were initially assigned the mean values of the associated PDSI gridpoint series. As in Rutherford *et al*. (2003) we used, as stopping criterion, the requirement of a root-mean-square change in imputed values between iterations of less than 0.5%. An 'inflation factor' adjusts the residual covariance matrix for the underestimation of the imputation error due to the regularization (see Schneider, 2001), which must be taken into account in estimating uncertainties in the imputed values (for details, see Schneider, 2001; Rutherford *et al*., 2003). Owing to the added complication of estimating the uncertainty in the full reconstructions from the imputation error in the prewhitened data, we have adopted the alternative approach of estimating self-consistent uncertainties in the reconstructions based on the distribution of verification period residuals (see, for example, Mann *et al*., 1998). The resulting uncertainty estimates include contributions from both the

reconstruction of the prewhitened field and from the serial persistence contribution to the full reconstruction.

Cross-validation results

A cross-validation (or 'verification') procedure was used to estimate the fidelity of the PDSI reconstructions. In this procedure, the instrumental PDSI records from a restricted 1928 to 1978 interval were used in the calibration of the tree-ring predictor data, while earlier PDSI data available across the full domain from 1895 to 1927 were withheld to independently test the skilfulness of the imputed values. The associated calibration interval corresponds exactly to that used by Cook *et al.* (1999). However, as already noted, the gridded instrumental PDSI data sets used by Cook *et al.* (1999) and this study differ somewhat. Therefore, while the calibration/verification period cross-validation statistics can be compared between studies, they will differ by a small unknown amount due to differences in the predictand data fields themselves. Unlike other traditional methods of palaeoclimate reconstruction, an independent estimate of *calibrated* variance is not strictly possible in the RegEM procedure (see Rutherford *et al.*, 2003). However, as calibration resolved variance, prone to statistical overfitting, is typically an overestimate of actual resolved variance, cross-validated resolved variance is in any case a more rigorous metric of true reconstructive skill, and a more meaningful basis for comparison of reconstructions.

Domain-wide cross-validation results (AD 1895–1927)

Employing, as in Cook *et al.* (1999) a 1928–78 calibration interval, and a fixed 1895–1927 verification interval here, we evaluated the fidelity of the various PDSI reconstructions using standard measures of cross-validated reconstructive skill (e.g., Cook *et al.*, 1999; Mann and Rutherford, 2002), including the 'Reduction of Error' (RE) statistic ('β' in the terminology of Mann *et al.*, 1998) and the squared Pearson correlation coefficient, $r(m)^2$. The latter measures the level of covariation between two variables, but ignores the possible differences between the two variables in their mean and variance. The former is arguably a more rigorous measure of skill, measuring the correspondence not only in terms of the relative departures from mean values but also in terms of the means and absolute variance of the two series. These statistics can be calculated for individual gridpoints, means over particular regions or the entire global domain, or the multivariate field. The multivariate versions of the statistics can be defined as a simultaneous sum over time (during the verification interval) and gridpoint,

$$RE = 1 - \sum_i \sum_j (X_{ij} - \hat{X}_{ij})^2 / \sum_i \sum_j X_{ij}^2 \qquad (1)$$

$$r(m)^2 = \left[\sum_i \sum_j (X_{ij} - \overline{X})(\hat{X}_{ij} - \overline{\hat{X}}) \right]^2 / \left[\sum_i \sum_j (X_{ij} - \overline{X})^2 \sum_i \sum_j (\hat{X}_{ij} - \overline{\hat{X}})^2 \right] \qquad (2)$$

where i and j represent year and gridpoint, respectively, X_{ij} and $\hat{X}_{ij}$ are the actual and reconstructed PDSIs, and $\overline{X}$ and $\overline{\hat{X}}$ are the mean values of X_{ij} and $\hat{X}_{ij}$. $RE = 0$ represents the lower limit for a statistically 'skilful' reconstruction in the sense that the nominal skill associated with reconstructing the climatological mean is matched or exceeded (note, however, that $RE < 0$ may in some cases also be argued to exhibit skill in the sense that the verification period mean is not *a priori known*).

The results of the cross-validation exercises (as summarized in Table 1) indicate that RegEM calibration scheme (iii) ('regional' drought calibrated against 'global' tree-ring predictors) based on a regionally variable screening threshold, provides the best apparent skill, with a multivariate value of $RE = 0.41$ and $r^2 = 0.42$, and with $RE = 0.71$ and $r^2 = 0.70$ for the domain or 'global' mean series (Table 1d), suggesting that the associated PDSI reconstruction skilfully resolves nearly half of the variance in the full PDSI field and 70% of the variance in the 'global' mean series (note that an analysis using the restricted set of 425 chronologies used by Cook *et al.*, 1999, rather than the full 483 available chronologies, yields no significant differences in skill, with multivariate $RE = 0.39$ and global mean series $RE = 0.71$ in comparison; Table 1d). Case (ii) ('regional' drought calibrated against 'regional' tree-ring predictors) and case (i) ('global' drought calibrated against 'global' tree-ring predictors) both exhibit moderately lower levels of skill for both multivariate and global mean diagnostics (Table 1a, 1b, 1c), with case (i) employing a 95% significance criterion in screening giving the poorest performance (Table 1b), though even in this case a quite skilful reconstruction is evident at the gridpoint level ($RE = 0.29$).The results of the 'regional versus regional' experiments with fixed selection criterion (not shown) are quite similar to those shown in Figure 3c, with verification scores observed to be slightly lower (Table 1c).

The skill evident in the Cook *et al.* (1999) reconstructions based on PPR (Table 1f) is moderately below the 'optimal' RegEM results achieved in scheme (iii) with a regionally variable screening threshold (Table 1d), and similar to that achieved in both RegEM scheme (iii) with fixed screening threshold (Table 1e), and case (ii) with regionally variable screening threshold (Table 1c). The comparisons thus suggest that a modest gain in reconstructive skill can be accomplished through the explicit incorporation of nonlocal information in the tree-ring predictor network as is permitted in scheme (iii). A similar conclusion has been independently based on use of the PPR method in conjunction with a regionally variable search radius (Cook, unpublished data).

The verification scores show sizeable differences at the regional scale (both for the multivariate field and regional domain means). The cross-validation skill estimates for the 'optimal' RegEM results as defined above generally remain superior to those evident in all other (RegEM and PPR) reconstructions at the regional scale (Table 1). The exceptions are that the estimated skill is slightly exceeded by case (ii) in region 3 and by PPR in regions 7 and 8 for the multivariate skill measures, and are exceeded by PPR in regions 7 and 8 for the regional domain mean series.

While the diagnostics discussed above support the fidelity of the reconstructions at global or regional scales, further work is necessary to evaluate the pattern of skill in greater spatial detail. We calculated the verification skill diagnostics for all available gridpoints for the various RegEM reconstruction schemes and PPR. Given the length (33 years) of the 1895 to 1927 verification interval, any $r^2 > 0.10$ is statistically significant at the $\alpha = 0.05$ level based on a one-sided test. Based on this criterion, only five of the 155 gridpoint reconstructions failed the significance test for the case (i), 11 failed for case (i) with fixed screening threshold of $|\mathrm{Rc}| = 0.276$, 11 failed for case (ii), two failed for case (iii) (with variable screening threshold) and two failed for case (iii) with fixed screening threshold of $|\mathrm{Rc}| = 0.276$, and seven failed for PPR (Cook *et al.*, 1999).

Table 1

(a) Verification scores for the case of global proxy calibrated against global PDSI

	RE (a)	$r(a)^2$	RE(m)	$r(m)^2$	|Rc|	num
region-1	0.48	0.48	0.20	0.24		
region-2	0.68	0.64	0.49	0.40		
region-3	0.49	0.47	0.17	0.20		
region-4	0.50	0.57	0.08	0.19		
region-5	0.62	0.62	0.44	0.44		
region-6	0.53	0.54	0.38	0.38		
region-7	0.31	0.33	0.08	0.12		
region-8	0.36	0.40	0.16	0.18		
global	0.70	0.69	0.32	0.32	0.40	352

(b) Verification scores for the case of global proxy calibrated against global PDSI with fixed |Rc|

	β(a)	$r(a)^2$	β(m)	$r(m)^2$	|Rc|	num
region-1	0.47	0.49	0.21	0.22		
region-2	0.63	0.63	0.45	0.39		
region-3	0.43	0.42	0.14	0.18		
region-4	0.52	0.54	0.09	0.17		
region-5	0.60	0.60	0.43	0.42		
region-6	0.50	0.51	0.34	0.35		
region-7	0.28	0.28	0.06	0.09		
region-8	0.28	0.36	0.12	0.17		
global	0.67	0.65	0.29	0.29	0.276	467

(c) Verification scores for the case of regional proxy calibrated against regional PDSI

	RE(a)	$r(a)^2$	RE(m)	$r(m)^2$	|Rc|	num
region-1	0.36	0.36	0.21	0.22	0.57	13
region-2	0.70	0.59	0.53	0.42	0.60	11
region-3	0.59	0.61	0.29	0.32	0.51	12
region-4	0.49	0.69	0.15	0.29	0.53	23
region-5	0.56	0.61	0.38	0.45	0.27	137
region-6	0.66	0.73	0.49	0.51	0.57	11
region-7	0.53	0.56	0.30	0.30	0.42	10
region-8	0.32	0.54	0.09	0.20	0.31	18
global	0.68	0.67	0.36	0.38		

(d) Verification scores for the case of global proxy calibrated against regional PDSI

	RE(a)	$r(a)^2$	RE(m)	$r(m)^2$	|Rc|	num
region-1	0.61	0.61	0.30	0.33	0.37	135
region-2	0.77	0.70	0.56	0.46	0.38	200
region-3	0.61	0.61	0.27	0.30	0.58	14
region-4	0.56	0.68	0.27	0.35	0.34	122
region-5	0.62	0.63	0.45	0.45	0.21	361
region-6	0.71	0.76	0.53	0.54	0.57	16
region-7	0.57	0.62	0.32	0.33	0.48	11
region-8	0.41	0.52	0.17	0.23	0.41	16
global	0.71	0.70	0.41	0.42		

(e) Verification scores for the case of global proxy calibrated against regional PDSI with fixed |Rc|

	RE(a)	$r(a)^2$	RE(m)	$r(m)^2$	|Rc|	num
region-1	0.53	0.53	0.24	0.27		276
region-2	0.71	0.66	0.53	0.45		319
region-3	0.54	0.53	0.22	0.26		296
region-4	0.54	0.62	0.21	0.29		211
region-5	0.59	0.59	0.43	0.44		275
region-6	0.60	0.62	0.42	0.43		225
region-7	0.55	0.54	0.23	0.26		136
region-8	0.20	0.39	0.08	0.22		131
global	0.70	0.71	0.36	0.37	.276	

(f) Verification scores for PPR

	RE(a)	$r(a)^2$	RE(m)	$r(m)^2$
region-1	0.55	0.56	0.26	0.31
region-2	0.66	0.68	0.45	0.38
region-3	0.52	0.49	0.17	0.24
region-4	0.52	0.63	0.20	0.33
region-5	0.60	0.63	0.42	0.42
region-6	0.67	0.69	0.46	0.46
region-7	0.58	0.58	0.33	0.32
region-8	0.58	0.62	0.24	0.28
global	0.69	0.69	0.35	0.36

(g) Verification scores for the case of global proxy calibrated against global PDSI based on 425 chronologies

	RE(a)	$r(a)^2$	RE(m)	$r(m)^2$	|Rc|	num
region-1	0.64	0.64	0.29	0.34	.37	117
region-2	0.75	0.66	0.52	0.42	.38	179
region-3	0.61	0.61	0.27	0.30	.58	14
region-4	0.56	0.66	0.26	0.35	.34	116
region-5	0.59	0.60	0.42	0.42	.20	336
region-6	0.69	0.73	0.51	0.52	.56	16
region-7	0.51	0.61	0.29	0.31	.48	10
region-8	0.42	0.55	0.18	0.24	.41	14
global	0.71	0.70	0.39	0.40		

*RE(a) is the RE value for mean fields, RE(m) is the RE value for multi-gridpoints (multivariate value), $r(a)^2$ is the squared correlation coefficient between actual PDSI and reconstructed PDSI for mean fields, $r(m)^2$ is the squared correlation coefficient between actual PDSI and reconstructed PDSI for multi-gridpoints, |Rc| is the optimum correlation coefficient obtained from screening method for a, c, d and g and a 95% significant criteria (fixed) for b and e, and num is the numbers of proxy data (subset) selected for reconstruction.

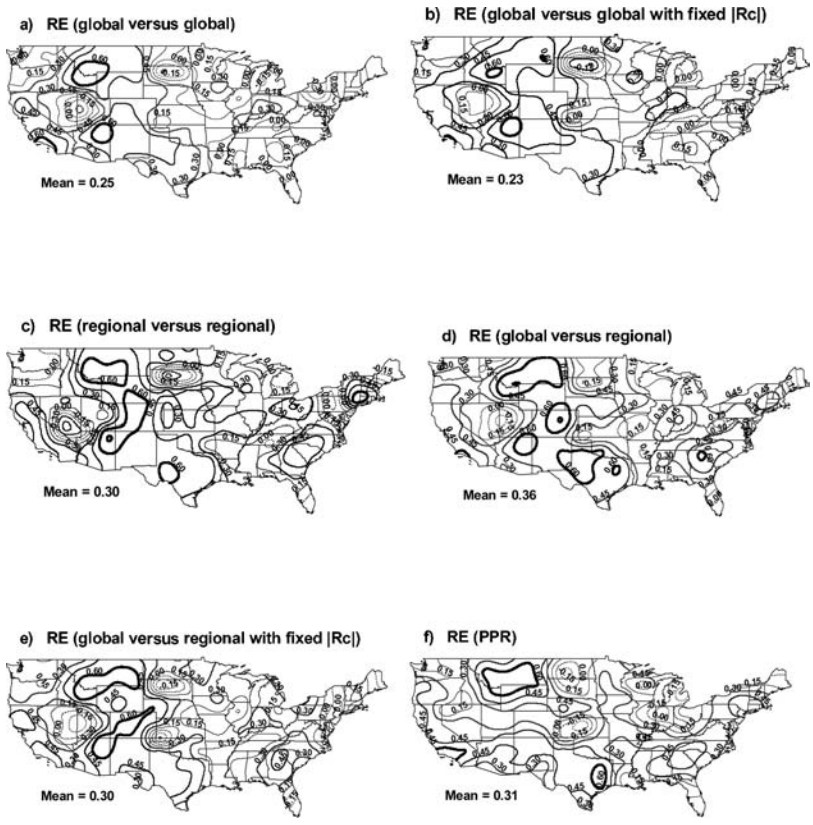


Figure 3 Maps of the verification *RE* statistic for the various RegEM calibration schemes and PPR method. Note that $RE > 0$ represents the threshold value for skilful reconstruction.

Of the 155 gridpoints, 18 gridpoints failed to pass the $RE = 0$ criterion for case (i), 15 failed for case (i) with fixed $|\mathrm{Rc}| = 0.276$, 20 failed for case (ii), four failed for case (iii) and 12 failed for case (iii) with fixed $|\mathrm{Rc}| = 0.276$, and 10 failed for PPR. The optimal RegEM reconstruction (case (iii) with regionally variable screening threshold) thus provides the greatest evident skill, with only two gridpoints failing both the RE and r^2 significance tests. These two gridpoints (154 and 155 in Figure 1), both in the far northeastern United States, and neighbouring the Atlantic ocean, may be located in regions where tree growth exhibits a particularly low sensitivity to drought owing to the plentiful nature of summer rainfall and relatively low summer temperatures. The other two gridpoints that failed to pass the $RE = 0$ test are located in Nevada, a semi-arid region in which tree-ring sensitivity to drought is, by contrast, expected to be quite high. Because other gridpoints neighbouring these gridpoints all exhibit relatively high *RE* values, we suspect that the failure of verification here results, instead, from data quality problems with the early instrumental PDSI gridpoint series. Since the problem in this case appears in the *RE* statistic and not the r^2 statistic, it may indicate an artificial change in the mean or variance of the instrumental data contributing to the PDSI gridpoint estimate. Indeed, Cook *et al.* (1999) demonstrated degradation of correlation of PDSI data from one weather station in Nevada with estimates of nearby stations over time. Such examples underscore the possibility that the cross-validation statistics may sometimes actually underestimate the true skill of the reconstruction.

Figure 3 compares the spatial patterns of *RE* verification scores for the three RegEM schemes and PPR. Case (iii) with regionally variable screening threshold exhibits the most homogeneous distribution of *RE* value and highest map mean score $RE = 0.36$ (Figure 3d). This mean score is notably higher than for case (i) (Figure 3a, $RE = 0.25$), case (i) with fixed $|\mathrm{Rc}| = 0.276$ (Figure 3b, $RE = 0.23$), case (ii) (Figure 3c, $RE = 0.30$), case (iii) with fixed $|\mathrm{Rc}| = 0.276$ (Figure 3e, $RE = 0.30$), and PPR (Figure 3f, $RE = 0.31$). In all cases, the *RE* scores in the western United States indicate a high fidelity of reconstructions in that region (isolated low *RE* scores are found near the border of Nevada and Utah for all RegEM cases; however, the fact that these low *RE* scores are not observed for the PPR reconstruction is somewhat enigmatic, suggesting the possibility that this feature represents a locally specific failure in the regEM reconstructions). It is interesting that the case (i) RegEM ('global versus global') reconstructions exhibit particularly low verification scores in most of the eastern US. This probably results from the presence of unstable drought teleconnection patterns that are isolated in an analysis of the covariance of the entire instrumental PDSI field, whereby the skilful information present in the western United States is inappropriately communicated to the eastern portions of the global domain.

It is useful to focus in detail on the comparison between the 'optimal' RegEM reconstruction defined by case (iii) with regionally variable screening threshold, and the PPR reconstruction of Cook *et al.* (1999). Detailed comparison of the *RE* score patterns (Figure 3, d and f), suggest similar fidelity of reconstructions in the eastern United ($RE = 0.30$ in both case). Reconstructions in the western US show higher *RE* scores (0.30 to 0.60), with the RegEM reconstructions exhibiting greater apparent skill in most regions, with the glaring exception of the border of Nevada and Utah as discussed earlier. The most profound improvement in indicated skill of RegEM over PPR is found in North Dakota, South Dakota, Kansas and the Great Lakes region, wherein the tree-ring predictor data is quite sparse. Remote tree-ring chronologies, which contain nonlocal information relevant to climate, are made use of in the RegEM approach, providing a potential advantage over the PPR approach,

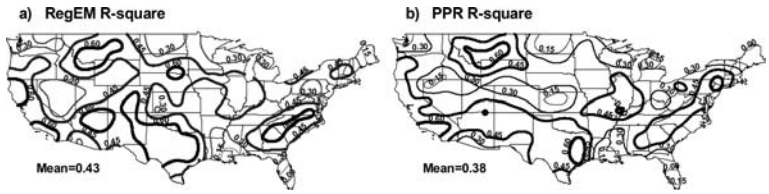


Figure 4 Maps of the verification r^2 statistic for the various RegEM calibration schemes and PPR method. Note that $r^2 > 0.10$ represents the threshold value for statistical significant at the $\alpha = 0.05$ level.

which is forced to make use of a relatively restricted set of candidate predictors in such data-sparse regions. Such comparisons suggest that use of large-scale covariance information, such as in RegEM, may be of greatest value when dealing with sparse predictors (e.g., as is the case with currently available global multiproxy data as in Mann *et al*., 1998; 1999) and of less utility when dealing with predictor-rich regions (as is the case in many regions with the Cook *et al*., 1999, network). Improvement in the fidelity of reconstructions afforded by RegEM is also evident in the comparison of spatial patterns of the r^2 verification statistic across the domain (Figure 4), with map mean $r^2 = 0.43$ for RegEM and 0.38 for PPR. A comparison of Figures 3 and 4 suggests similar patterns of the *RE* and r^2 statistics for a given reconstruction, with the notable exception of the isolated low values of *RE* in the Nevada/Utah border region noted earlier. In the case of PPR, the low values of *RE* in the Great Lakes Region have no obvious counterpart in the r^2 statistic, suggesting that the disagreement arises from a mismatch in the mean values and/or variance.

Spatial variability in the pattern of variance skilfully resolved by the reconstructions can potentially compromise the interpretation of spatial patterns of variability in the reconstructions. Figure 5 shows the contoured maps of gridpoint standard deviation for the actual PDSIs and estimated PDSIs during the verification interval based on both RegEM and PPR methods. The loss of variance in the reconstructed PDSI patterns is obvious for both RegEM and PPR, but the greater amplitude variability of drought in the western half of the US is captured in both cases. The verification period map correlation between the patterns of the standard deviation in the observed and reconstructed PDSI are $r = 0.66$ for RegEM and $r = 0.48$ for PPR, suggesting that the RegEM reconstructions may capture the actual spatial patterns of drought variation over the US somewhat more faithfully (note that PPR *calibration period* values are somewhat higher than the indicated PPR verification period value; Cook *et al*., 1999).

Extension of cross-validation results (AD 1870–94)

To further investigate the fidelity of the reconstructions, we compared the reconstructed PDSI field with a more spatially restricted, but temporally extended, set of instrumental PDSI gridpoint estimates available back to 1870. This extended verification data set is statistically independent of any screening optimization procedures described earlier, thus providing a truly independent measure of statistical reconstructive skill. The extended PDSI gridpoint series were calculated based on a multivariate regression of the gridded instrumental PDSI gridpoint data used here against longer available seasonal station precipitation and temperature records from the HCN network. Unfortunately, very few long such records exist in the western United States, so that the extended records are largely limited to the eastern half of the US, limiting the spatial extent of longer-term cross-validation possible. We selected 62 stations that have monthly temperature records and 70 stations that have monthly precipitation records back to 1870 (Table 2a). Gridding these data onto the same PDSI grid used by Cook *et al*. (1999) provided 16 gridpoints with monthly precipitation and temperature information back to 1870 (an attempted extension further back in time yielded very few useful gridpoint records, as the individual instrumental records available become far more sparse, and contain far more missing data. Therefore, the use of an even longer interval was likely to significantly impair the reliability of the comparison.)

We estimated the summer PDSI gridpoint series from 1870 to 1894 for these 16 gridpoints based on a simple multivariate statistical model whose predictors included current (summer) and antecedent (spring and winter) precipitation, and current (summer) temperature available during the 1895–1978 overlap interval between the PDSI gridpoint series and station instrumental data. The reliability of the PDSI reconstruction

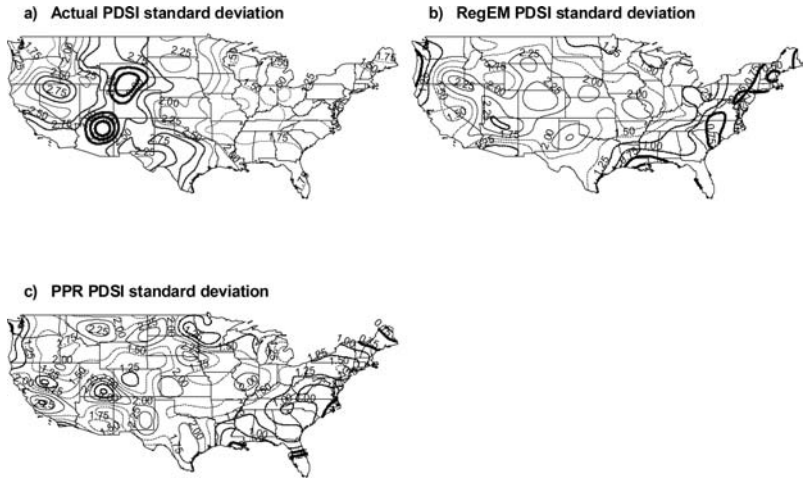


Figure 5 Maps of the standard deviation for actual and reconstructed PDSI in the verification period (1896–1927). Reconstructions based on both RegEM and PPR show at least modest loss of variance relative to the observed data.

Table 2
(a) Regression statistics and cross-validation statistics of the instrumental PDSI

longitude	− 83.5	− 95.5	− 89.5	− 86.5	− 83.5	− 95.5	− 89.5	− 83.5	− 74.5	− 71.5	− 89.5	− 83.5	− 77.5	− 92.5	− 74.5	− 68.5
latitude	33	39	39	39	39	41	41	41	41	41	43	43	43	45	45	45
stations (tem.)	1	5	4	3	2	1	5	6	5	2	3	4	13	1	5	2
stations (ppt.)	2	6	5	6	3	1	5	4	7	3	4	4	13	1	4	2
Regression statistics for multiregression models 1 (overlap interval is 1928−78)																
R^2	**0.72**	**0.76**	**0.78**	**0.75**	**0.67**	**0.62**	**0.66**	**0.65**	**0.69**	**0.73**	**0.55**	**0.66**	**0.74**	**0.50**	**0.63**	**0.65**
F-value	30.1	36.4	39.6	34.9	23.0	18.8	22.6	20.9	25.1	31.1	14.0	22.6	31.8	11.3	19.8	21.2
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Regression statistics for multiregression models 2 (overlap interval is 1895−1978)																
R^2	**0.68**	**0.74**	**0.77**	**0.78**	**0.64**	**0.59**	**0.69**	**0.63**	**0.57**	**0.65**	**0.57**	**0.54**	**0.64**	**0.48**	**0.49**	**0.61**
F-value	42.4	56.9	65.3	68.2	34.6	28.4	42.9	33.0	25.8	37.1	26.3	23.1	35.1	18.0	19.0	30.7
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Verification scores (1895−1927) between actual and extended PDSIs based on multiregression models 1																
RE	**0.62**	**0.73**	**0.79**	**0.85**	**0.46**	− 0.22	**0.79**	**0.34**	**0.00**	**0.12**	**0.46**	− 0.09	**0.19**	**0.18**	− 0.27	− 0.77
r	**0.80**	**0.86**	**0.89**	**0.92**	**0.84**	**0.66**	**0.89**	**0.74**	**0.63**	**0.80**	**0.81**	**0.75**	**0.80**	**0.65**	**0.48**	**0.76**
Multivariate verification scores: *RE* (actual) = **0.36**, *r*(actual) = **0.70**																

(b) Cross-validation of the reconstructions (RegEM and PPR) based on the extended instrumental series

Verification scores (1870−94) between reconstructions and extended PDSIs based on multiregression models 2																
longitude	− 83.5	− 95.5	-89.5	− 86.5	− 83.5	− 95.5	− 89.5	− 83.5	− 74.5	− 71.5	− 89.5	− 83.5	− 77.5	− 92.5	− 74.5	− 68.5
latitude	33	39	39	39	39	41	41	41	41	41	43	43	43	45	45	45
RE (RegEM)	**0.11**	**0.24**	**0.20**	**0.22**	− 0.06	**0.35**	**0.52**	**0.47**	− 0.31	− 0.62	**0.52**	**0.33**	− 0.15	− 0.25	**0.21**	− 0.03
RE (PPR)	− 0.00	**0.32**	− 0.01	**0.03**	− 0.22	**0.30**	**0.35**	**0.40**	**0.18**	− 0.46	**0.39**	− 0.07	**0.58**	− 1.51	**0.04**	− 0.23
r-(RegEM)	**0.41**	**0.50**	**0.53**	**0.50**	0.26	**0.59**	**0.72**	**0.69**	− 0.14	0.01	**0.76**	**0.62**	0.13	0.27	**0.58**	0.10
r-(PPR)	0.37	**0.58**	**0.45**	**0.44**	− 0.01	**0.41**	**0.76**	**0.56**	**0.42**	0.27	**0.70**	**0.44**	**0.75**	0.11	0.33	− 0.33
Multivariate verification scores: *RE* (RegEM) = **0.13**, *RE* (PPR) = **0.03**, *r*(RegEM) = **0.42**, *r*(PPR) = **0.39**																

*The bold values for reduction of error (*RE*), correlation (*r*) and explained variance of multi-regression model (R^2) have passed a certain significant level.

was estimated based on R^2, and *F* and *p* values from the multivariate regression. To further test the reliability of the resulting extended PDSI estimates, we performed a cross-validation exercise employing a restricted 1928−78 training interval, using the 1895−1927 interval for verification. The results of these analyses are described in Table 2a. The cross-validation results indicate reasonable skill for the instrumental-based extensions of the PDSI overall with a multivariate $RE = 0.36$ for the instrumental PDSI reconstructions during the 1895−1927 verification period. Certain gridpoints, however, exhibit much greater levels of skill ($RE = 0.7$ to 0.9) and are particularly useful for extending the cross-validation exercise. Four of the 16 gridpoints did not pass the $RE = 0$ test, probably due to an offset in mean from the true PDSI series.

These instrumental-based PDSI reconstructions provide a useful basis for extending the cross-validation results described above. Table 2b shows the verification scores in the interval of 1870 to 1894 between the extended PDSI gridpoint series (as estimated from the multivariate regression model) and the reconstructed PDSI field from both RegEM (optimal) and PPR methods. For the total multivariate field, the RegEM reconstructions yield a modestly skilful verification $RE = 0.13$, while the PPR results exhibit a more marginally skilful $RE = 0.03$. If the skill estimates are restricted, however, to the four extended instrumental PDSI gridpoint series which resolve greater than 70% of the variance in cross-validation with the true instrumental PDSI series (and are therefore most likely to represent reliable extensions of the actual PDSI series), the respective numbers are considerably better: $RE = 0.30$ (RegEM) and $RE = 0.17$ (PPR). Of the 16 gridpoints, six fail the $RE = 0$ test for RegEM, while seven fail for PPR; both RegEM and PPR have six gridpoints that failed to pass the 95% significant level for r^2. Most of the gridpoints that did not pass *RE* or r^2 significance tests are located along the east coast of the US. These extended cross-validation exercises nonetheless independently substantiate the skilfulness in both the RegEM and PPR reconstructions, as well as the modest improvement in skill in the RegEM reconstruction. Moreover, the threshold selection criteria for selecting candidate predictors described earlier is entirely independent of these cross-validation results in all cases.

Yearly map verification scores

We are furthermore interested how well the spatial patterns for particular years are replicated by the tree-ring-based reconstructions. To test the temporal homogeneity between actual and reconstructed PDSI maps, we used the following spatial diagnostics: spatial *RE*, Pearson correlation coefficient (r), and congruence coefficient (c),

$$RE_i = \sum_j (X_{ij} - \hat{X}_{ij})^2 / \sum_j X_{ij}^2 \tag{3}$$

$$r_i = \sum_j (p_{ij} - \bar{p}_i)(q_{ij} - \bar{q}_i) / \left[\sum_j (p_{ij} - \bar{p}_i)^2 \sum_j (q_{ij} - \bar{q}_i)^2 \right]^{1/2} \tag{4}$$

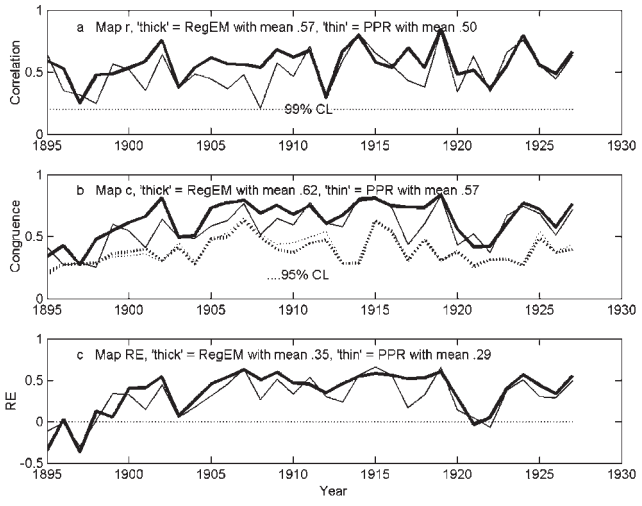


Figure 6 Map comparison statistics (*RE*, *r* and *c* statistics as described in text) quantifying the degree of similarity between actual and reconstructed PDSI patterns over time for RegEM (thick) and PPR (thin) methods. Significance limits are represented by dotted lines.

$$c_i = \sum_j p_{ij} q_{ij} \Big/ \left[\sum_j {p_{ij}}^2 \sum_j {q_{ij}}^2 \right]^{1/2} \qquad (5)$$

where i and j represent year and gridpoint, respectively, X_{ij} and $\hat{X}_{ij}$ are the actual and reconstructed PDSIs, respectively, p_{ij} and q_{ij} are the normalized actual and reconstructed PDSIs, respectively, and $\bar{p}_i$ and $\bar{q}_i$ are the mean fields for the year i.

As in Cook *et al.* (1999), we normalized actual and reconstructed PDSI gridpoint values for each gridpoint based on its calibration period mean and standard deviation prior to estimating the *r* and *c* statistics. This latter step is performed to avoid any regional bias in map correlation estimates owing to spatial variations in the PDSI standard deviation.

The congruence (*c*) test was originally developed as a measure of the similarity between two factor patterns in multivariate research (Richman, 1986; Broadbrooks and Elmore, 1987), and penalizes the difference between the two mean estimates, unlike the correlation (*r*) test. *c* arguably thus provides a more complete measure of the similarity between the two fields. There is no theoretical null distribution for *c* owing to its partial dependence on the random variables $\bar{p}_i$ and $\bar{q}_i$. We thus used a Monte Carlo procedure (employing 10 000 realizations) to estimate an empirical null distribution and significance levels for *c* (e.g., Broadbrooks and Elmore, 1987).

Figure 6 shows the time-dependent verification measures provided by *r*, *c* and *RE* as defined above. Both the RegEM and PPR reconstructions pass the 99% significance level for the *r* test for all years, with means of 0.57 (RegEM) and 0.50 (PPR), though PPR nearly fails for 1908. For the *c* statistic, RegEM fails the 95% significance test for one year (1897) as does PPR (1898). The mean values of *c* are 0.62 and 0.57 for RegEM and PPR respectively (note that *c* tends to be biased towards 1.0 relative to *r*; Richman, 1986). The *RE* test, arguably the most rigorous, indicates three years (1895, 1897 and 1921) that fail to exceed the $RE = 0$ skill threshold for RegEM, and four years (1895, 1896, 1897 and 1922) that fail for PPR, with mean values of 0.35 (RegEM) and 0.29 (PPR). While RegEM tends to outperform PPR for all three spatial skill measures, the overall pattern over time of the yearly skill measures is similar for both RegEM and PPR for all three diagnostics, suggesting a modest decrease back in time (particularly prior to the twentieth century) in each case. It is tempting to conclude that this trend might arise, at least in part, from diminished instrumental data quality in the earliest years, owing for example to a reduced number of stations contributing to the gridpoint averages. The relatively better performance of the *r* test in the earliest years suggests that that the degradation in the *c* and *RE* skill measures may result from a bias in estimates of the mean in either the actual or reconstructed PDSI field.

PDSI reconstructions

As the verification results described in the previous section indicate that an 'optimal' RegEM results from a calibration scheme involving regional PDSI and global tree-ring predictors with regionally variable screening threshold, all further PDSI reconstructions are based on this choice of methodology. Furthermore, having withheld the 1895–1927 instrumental PDSI data for statistical model validation, all available

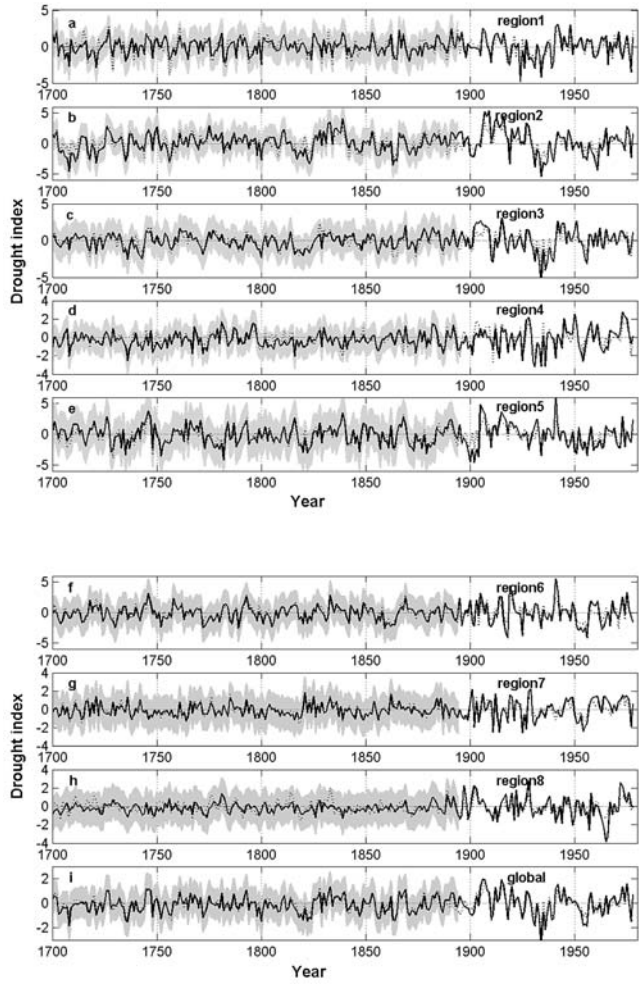


Figure 7 Time series of regional and global mean drought back to 1700. Shown are optimal RegEM reconstruction from 1700 to 1894 (solid line), PPR reconstruction from 1700 to 1978 (dotted line) and instrumental PDSI value from 1895 to 1978 (solid line). The two standard error uncertainties in the RegEM reconstructions are indicated by the grey shading. Note that the 1930s 'Dust Bowl' drought is the largest continental-scale drought in the reconstructed record, exceeding the two standard error limits for the 'global' reconstruction.

(1895–1978) instrumental PDSI data are subsequently used in calibration to produce the final PDSI reconstructions. The reconstructions are performed back to 1700, coinciding with the shortest tree-ring chronologies in the candidate predictor set of 483. The final reconstructions employed the same subset's candidate predictors indicated in the cross-validation experiments (Table 1).

Regional and 'global' mean PDSI reconstructions

Figure 7 shows the RegEM reconstructed regional and global mean summer drought reconstructions (and self-consistent two standard error uncertainties) from 1700 to 1894 along with the PPR reconstructions from 1700 to 1978, and the instrumental data from 1895 to 1978. The comparisons show a high degree of similarity between RegEM and PPR reconstructions for the global (domain) mean as well as the individual regions (with the exception, to some extent, of region 8). The correlation coefficients between the two reconstructions over the interval 1700–1894 are $r = 0.80$ for region 1, 0.84 for region 2, 0.81 for region 3, 0.86 for region 4, 0.91 for region 5, 0.92 for region 6, 0.92 for region 7, 0.49 for region 8 and 0.92 for the 'global' mean over the continental United States. The similarity between the PPR and RegEM reconstructions at this scale underscore the robustness of regional and global drought estimates derived from the tree-ring predictor network of Cook *et al.* (1999).

We subsequently consider the indicated history of drought and wet episodes. The global mean PDSI series (Figure 7) indicates the Dust Bowl drought of the 1930s to be the most severe drought at this spatial scale to have occurred in the US since 1700, exceeding by more than two standard errors any other indicated drought periods in the reconstruction. As noted by Cook *et al.* (1999), other particularly notable drought periods occur during the 1820s and 1860s, and a prominent wet period is observed over the interval 1825–40. This latter wet period is comparable to the wet interval of 1900–20 recorded in the instrumental record. The regional drought series (Figure 7) show some significant differences in both amplitude, and detailed features of the chronology. Consistent with the observation of greater drought variability in the western US (e.g., Figure 5), regions 1, 2, 3, 5 and 6 exhibit greater amplitude variability than 4, 7 and 8, with regions 5 and 2 exhibiting the greatest amplitude variability. Regions 2 and 5 (and to a lesser extent 6), in particular, show considerable multidecadal variability in drought. At the regional scale, it is less obvious that twentieth-century episodes (e.g., the Dust Bowl droughts) are unusual in a longer-term context. For example, the RegEM mean drought series for region 2 indicates two pronounced drought periods occurring around 1710–20 and 1820–30 that are similar in magnitude and duration to the Dust Bowl episode, within the indicated uncertainties. The Dust Bowl episode is barely evident in region 5, and certainly is less

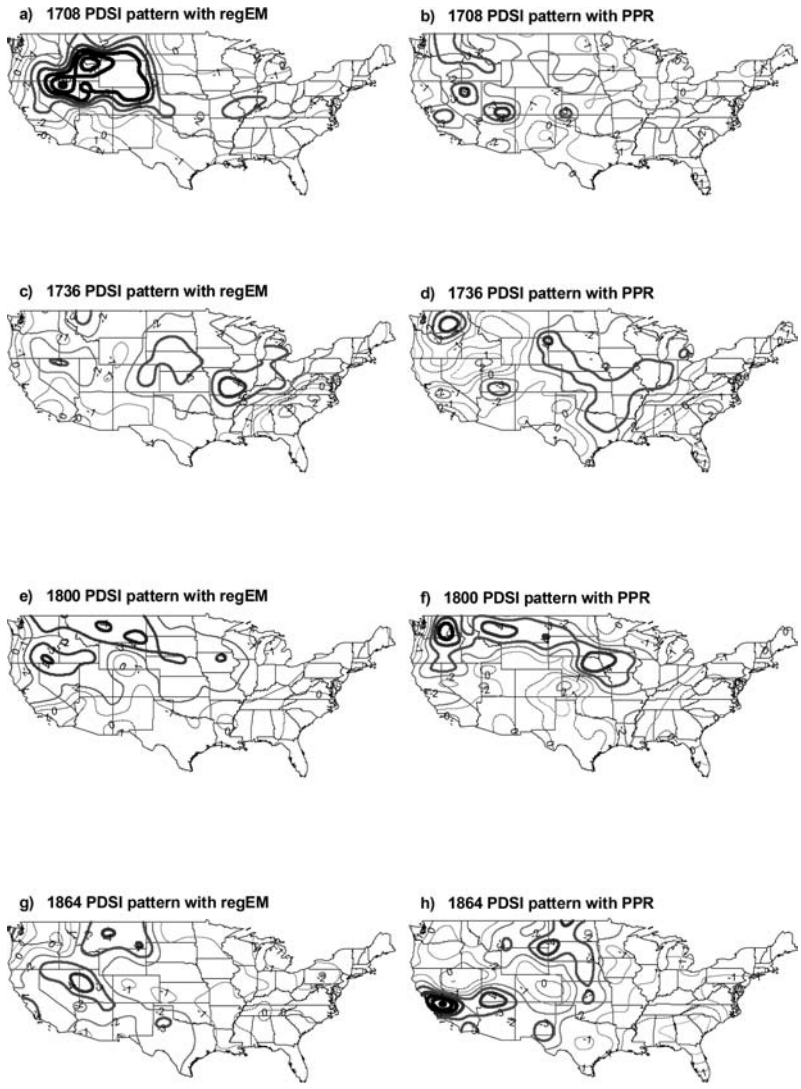


Figure 8 The spatial patterns of four significant drought years based on RegEM and PPR.

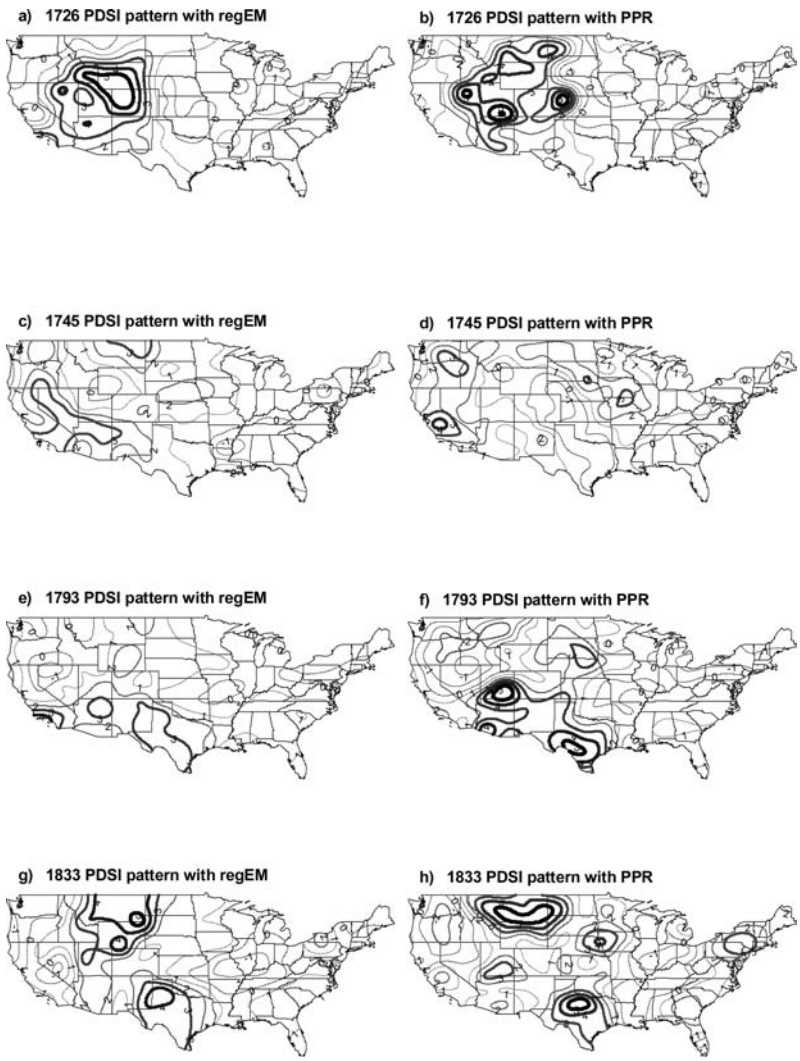


Figure 9 The spatial patterns of four significant wet years based on RegEM and PPR, respectively.

prominent than drought events occurring during the 1730s, 1750s, 1780s, 1820s and 1880s. For region 6, a large drought occurred in the 1950s, but both RegEM and PPR reconstructions show similarly prominent drought periods in 1750s, 1770s, 1820s and 1860s. Region 8 is the only area where reconstructions based on RegEM and PPR indicate relatively large differences. Both reconstructions show considerable loss of variance relative to the instrumental PDSI series, which, as discussed earlier, may be due to the relatively lower sensitivity of species in this region to drought.

Spatial patterns for significant drought and wet years

Figure 8 shows the reconstructed PDSI patterns of four significant continental drought years based on both RegEM and PPR methods. The pattern for 1708 indicates a nearly continental-scale pattern of drought, with the western US exhibiting the most severe drought. RegEM shows a particularly severe drought over the northern mountain states. Both RegEM and PPR indicate similar patterns of drought for 1736, with moderate wetness in the southeastern US, and pronounced drought in the central and western US. Moderate differences are observed in the northern border region between Nevada and Utah where PPR indicates wet conditions, while RegEM indicates a mild drought. Both RegEM and PPR reconstructions indicate a prominent drought in the northwestern and north central regions of the US in 1800, with the RegEM drought pattern slightly broader and less regionally intense than the PPR pattern. In the year 1864, drought is evident in the western half of the US for both RegEM and PPR reconstructions. Within the large drought area, PPR shows some small-scale normal climate regions.

Figure 9 indicates the patterns of four significant continental wet years. Both RegEM and PPR show a prominent wet event in the mountain states for 1726, with the pattern more localized in the RegEM case. For the year 1745, wet conditions are observed in both the far west and midwestern US, with RegEM indicating a more longitudinally extended pattern, and PPR indicating a more latitudinally extended pattern. In 1793, both RegEM and PPR indicate wet conditions in the south central US, with PPR also indicating mild drought conditions in the northeastern and northwestern US. For the year 1833, much of the central and mountain region exhibit wet conditions with both RegEM and PPR reconstructions, with the RegEM pattern more localized, and the PPR pattern organized into a number of disjoint wet regions. It is worthy of note that most of the significant continental drought and wet years are associated with anomalies in the central and west regions, suggesting that these regions dominate continental-scale drought and wet episodes. The RegEM reconstructions are typically more spatially homogenous than the PPR reconstructions. This is consistent with the greater homogeneity of the retained spatial variance (i.e., Figure 5), although it could also

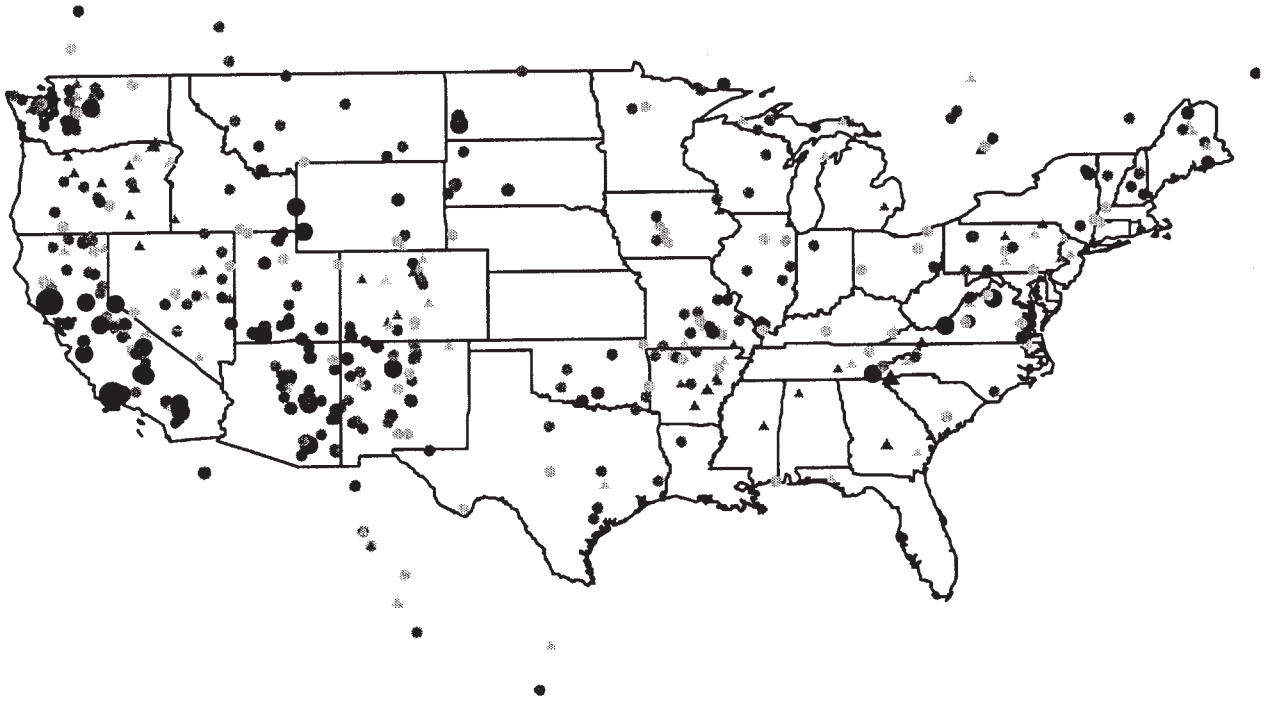


Figure 10 Tree-ring width anomalies for the year 1864. Circular/triangular dots represent negative/positive values with larger and heavier dots indicating the magnitude of anomalies.

result, in part, from the regional spatial smoothing implicit in the RegEM approach.

It is useful to investigate differences between the patterns of drought reconstructed by the PPR and RegEM approaches, for a specific example, the year 1864 (see Figure 8, lower right and left panels), based on comparison with the pattern of raw tree-ring width anomalies (Figure 10). The PPR reconstruction is seen to follow closely the local pattern of tree-ring width anomalies, indicating, for example, locally strong regions of drought in south central California/western Nevada, along the western Utah/Arizona border, and in the north central states centred near the border of Nebraska and the Dakotas. Each of these areas exhibits dense pockets of sizeable positive width anomalies. By contrast, the RegEM reconstruction shows a far less close local relationship between reconstructed drought and tree-ring width anomalies, making use of more intricate statistical information in the candidate predictor network. The RegEM reconstruction exhibits less prominent drought in southern California, a more western-shifted pattern of drought in the north central US (centred closer to Montana than the Dakotas), and a more prominent centre of drought in the Nevada/Utah border.

Conclusions

The RegEM algorithm employed in this study, in contrast with previous approaches, makes use of large-scale and nonlocal covariance information in relating predictors and predictand in reconstructing patterns of continental drought from tree-ring proxy data. The appropriate use of this larger-scale and nonlocal information appears to lead on average to modest improvements over the drought reconstructions based on more localized regression approaches such as PPR, as measured by cross-validation statistics. The optimal RegEM drought reconstruction appears to be achieved when 'global' (i.e., conterminous US domainwide) candidate predictors are used to reconstruct patterns of drought on a region-by-region basis using predictor variable screening. We infer from this observation the existence of significant nonlocal information within the long-term tree-ring predictor network that is useful in the reconstruction of regional drought. The fact that the use of global information in the predictand (instrumental PSDI) data diminishes, rather than improves, the reconstructive skill, on the other hand, suggests that the fundamental patterns of large-scale drought variability cannot adequately be captured through evaluating the spatial covariance information in the relatively short (less than one century) instrumental calibration data set. This limitation probably results from the apparent somewhat unstable nature of large-scale drought teleconnection patterns in the US over the twentieth century, and the greater regional character of the drought field as compared to other (e.g., surface temperature) climatic fields.

The RegEM reconstruction appears to yield a modest improvement over previous conterminous US summer drought reconstructions in general, as confirmed by a variety of metrics of reconstructive fidelity. The most obvious improvements in reconstructive skill seem to be found in tree-ring data-sparse regions (e.g., the Dakotas and Kansas) where the RegEM method makes use of nonlocal information, while the PPR method is highly limited by data availability within the selected search radius. The relatively low topographic relief of the Great Plains would also allow for larger and more homogeneous fields of drought variability, thus extending the useful correlation-decay distance between local drought and more remote tree-ring chronologies. As commented earlier, PPR with an adaptive search radius presents a useful alternative strategy for dealing with such limitations.

Despite the modest improvements, and some differences that are evident in the precise pattern of reconstructed drought for particular years, it is quite encouraging that two very different methodologies (RegEM and PPR) for assimilating tree-ring information into a reconstruction of past drought patterns give, in general, such similar results (e.g., with respect to global and regional average past drought histories). This similarity seems to underscore the fundamental quality of the

underlying predictor network of continental US drought-sensitive tree-ring chronologies, and the consequent robustness of PDSI reconstructions from this network based on the application of different statistical reconstructions methodologies. The current results reaffirm the key conclusions of Cook *et al*. (1999). At the continental scale, the 1930s 'Dust Bowl' remains the most severe drought event since 1700 within the context of the estimated uncertainties. More severe episodes may have occurred at regional scales in past centuries.

Acknowledgements

This work was supported by the NSF- and NOAA-funded 'Earth Systems History' program (M.E.M. and Z.Z.-NA16GP2913; E.R.C.-NA06GP0450). Lamont-Doherty Earth Observatory Contribution Number LDEO 6581.

References

Broadbrooks, W.J. and **Elmore, P.B.** 1987: A Monte Carlo study of the sampling distribution of the congruence coefficient. *Educational and Psychological Measurement* 47, 1–11.

Cole, J.E. and **Cook, E.R.** 1998: The changing relationship between ENSO variability and moisture balance in the continental United States. *Geophysical Research Letters* 25(24), 4529–32.

Cook, E.R., Briffa, K.R. and **Jones, P.D.** 1994: Spatial regression methods in dendroclimatology: a review and comparison of two techniques. *International Journal of Climatology* 14, 379–402.

Cook, E.R., Meko, D.M., Stahle, D.W. and **Cleaveland, M.K.** 1996: Tree-ring reconstructions of past drought across the coterminous United States: tests of a regression method and calibration/verification results. In Dean, J.S., Meko, D.M. and Swetnam, T.W., editors, *Tree rings, environment, and humanity*, Radiocarbon, 155–69.

——1999: Drought reconstructions for the continental United States. *Journal of Climate* 12, 1145–62.

Cook, E.R., Meko, D.M. and **Stockton, C.W.** 1997: A new assessment of possible solar and lunar forcing of the bidecadal drought rhythm in the western United States. *Journal of Climate* 10, 1343–56.

D'Arrigo, R.D. and **Jacoby, G.C.** 1991: A 1000-year record of winter precipitation from northwestern New Mexico, USA: a reconstruction from tree-rings and its relation to El Niño and the Southern Oscillation. *The Holocene* 1, 95–101.

Evans, M.N., Kaplan, A. and **Cane, M.A.** 2002: Pacific sea surface temperature field reconstruction from coral δ18O data using reduced space objective analysis. *Paleoceanography* 17, 7-1–7-13.

Fritts, H.C. 1991: *Reconstructing large-scale climatic patterns from tree-ring data.* Tucson and London: The University of Arizona Press, 286 pp.

Fritts, H.C., Blasing, T.J., Hayden, B.P. and **Kutzbach, J.E.** 1971: Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate. *Journal of Applied Meteorology* 10(5), 845–64.

Graumlich, L.J. 1993: A 1000-year record of temperature and precipitation in the Sierra Nevada. *Quaternary Research* 39, 249–55.

Hughes, M.K. and **Brown, P.M.** 1992: Drought frequency in central California since 101 B.C. recorded in giant sequoia tree rings. *Climate Dynamics* 6, 161–67.

Hughes, M.K. and **Graumlich, L.J.** 1996: Multimillennial dendroclimatic records from western North America. In Bradley, R.S., Jones, P.D. and Jouzel, J., editors, *Climatic variations and forcing mechanisms of the last 2000 years*, Berlin: Springer, 109–24.

Kaplan, A., Kushni, Y., Cane, M.A. and **Blumenthal, M.B.** 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *Journal of Geophysical Research* 102(C13), 27,835–60.

Karl, T.R. and **Knight, R.W.** 1998: Secular trends of precipitation amount, frequency, and intensity in the United States. *Bulletin of the American Meteorology Society* 79, 1107–19.

Karl, T.R., Knight, R.W., Easterling, D.R. and **Quayle, R.G.** 1996: Indices of climate change for the United States. *Bulletin of the American Meteorology Society* 77, 279–91.

Karl, T.R., Williams, C.N. Jr and **Quinlan, F.T.** 1990: *United States Historical Climatology Network (HCN) serial temperature and precipitation data*. Environmental Science Division, Publication no. 3404, Oak Ridge, TN: Carbon Dioxide Information and Analysis Center, 371 pp.

Luterbacher, J., Xoplaki, E., Dietrich, D., Rickli, R., Jacobeit, J., Beck, C., Gyalistras, D., Schmutz, C. and **Wanner, H.** 2002: Reconstruction of sea level pressure fields over the eastern North Atlantic and Europe back to 1500. *Climate Dynamics* 18, 545–61.

Mann, M.E. and **Rutherford, S.** 2002: Climate reconstruction using 'pseudoproxies'. *Geophysical Research Letters* 29, 139-1–139-4.

Mann, M.E., Bradley, R.S. and **Hughes, M.K.** 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* 392, 779–87.

——1999: Northern Hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophysical Research Letters* 26(6), 759–62.

Meko, D.M. 1981: Applications of Box-Jenkins methods of time series analysis to the reconstruction of drought from tree rings. Unpublished PhD dissertation, University of Arizona, Tucson, 149 pp.

Meko, D.M., Cook, E.R., Stahle, D.W., Stockton, C.W. and **Hughes, M.K.** 1993: Spatial patterns of tree-growth anomalies in the United States and southeastern Canada. *Journal of Climate* 6, 1773–86.

Mitchell, J.M. Jr, **Stockton, C.W.** and **Meko, D.M.** 1979: Evidence of a 22-year rhythm of drought in the western United States related to the Hale solar cycle since the 17th century. In McCormac, B.M. and Seliga, T.A., editors, *Solar-terrestrial influences on weather and climate*, Dordrecht: D. Reidel, 125–44.

Palmer, W.C. 1965: *Meteorological drought*. Research Paper no. 45, Washington, DC: US Department of Commerce Weather Bureau.

Preisendorfer, R.W. 1988: *Principal component analysis in meteorology and oceanography.* New York: Elsevier Science, 425 pp.

Rajagopalan, B., Cook, E.R., Lall, U. and **Bonnie, K.R.** 2000: Spatiaotemporal variability of ENSO and SST teleconnections to summer drought over the United States during the twentieth century. *Journal of Climate* 13, 4244–55.

Richman, M.B. 1986: Rotation of principal components. *Journal of Climate* 6, 293–335.

Riebsame, W.E., Changnon, S.A. and **Karl, T.R.** 1991: *Drought and natural resources management in the United States: impacts and implications of the 1987–89 drought.* Boulder, CO: Westview Press, 11–92.

Rutherford, S., Mann, M.E., Delworth, T.L. and **Stouffer, R.** 2003: Climate field reconstruction under stationary and nonstationary forcing. *Journal of Climate* 16, 462–79.

Schneider, T. 2001: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14, 853–71.

Smith, T.M., Reynolds, R.W., Livezey, R.E. and **Stokes, D.C.** 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climate* 9, 1403–20.

Stahle, D.W. and **Cleaveland, M.K.** 1992: Reconstruction and analysis of spring rainfall over the southeastern US for the past 1000 years. *Bulletin of the American Meteorology Society* 73(12), 1947–61.

Stahle, D.W., D'Arrigo, R.D., Krusic, P.J., Cleaveland, M.K., Cook, E.R., Allan, R.J., Cole, J.E., Dunbar, R.B., Therrell, M.D., Gay, D.A., Moore, M.D., Stokes, M.A., Burns, B.T., Villanueva-Diaz, J. and **Thompson, L.G.** 1998: Experimental dendroclimatic reconstruction of the Southern Oscillation. *Bulletin of the American Meteorological Society* 79(10), 2137–52.

Woodhouse, C.A. and **Jonathan, T.O.** 1998: 2000 years of drought variability in the central United States. *Bulletin of the American Meteorology Society* 79, 2693–714.