model structures sometimes yielded high-quality reconstructions with distinctly different statistics, suggesting that different tree-ring characteristics may have been calibrated by the models (see App. 1, Sec. F.2). In some cases the reconstructions from models of different structures were quite different. This result was thought to reflect competing tree-growth responses between different species and contrasting sites that gave rise to the spatial variability noted earlier in the tree-ring data set. The reconstructions of two or three models showing different responses, but with high-ranking statistics, were averaged in an attempt to make use of some of the diversity in response.

Canonical regression finds the best-fitting line through the cluster of points representing the predictand data. The success of the regression is determined by calculating the percentage of the actual data variance that is mimicked by the estimate. These results are often expressed as a fraction of the total variance reduced by the reconstruction (see App. 1, Secs. B and C, for examples and more information on calibration procedures).

C. Verification

The results of transfer-function relationships modeled in the dependent period are assumed to be the same as those modeled during the independent period (Webb and Clarke 1977; Bryson 1985). The coefficients of the transfer function should be stable; that is, they should be the same no matter what data are used as the dependent data. A successful model does not simply mimic the dependent data, but also expresses a universal property regarding the tree-growth response to variations in both present and past climate. Nevertheless, the process used to optimize the coefficients of the transfer function virtually ensures that the model will be more accurate for the dependent data than for any other body of data to which it may be applied (Larson 1931; Wherry 1931; Anderson et al. 1972; Stone 1974). Thus, the predictive power of a regression model must decrease when the model is applied to independent data. This deterioration in accuracy should be measured whenever possible and the measurements used either to evaluate the performance of the model or to provide the proper perspective with which to view the climate reconstructions.

Any procedure that is used to assess the reliability of a reconstruction on independent data is a *verification procedure*. Reliability can be measured by *verification statistics* that assess the degree of association between the estimates of climate independent of the calibration data and the corresponding instrumental data they are supposed to mimic.

Not only must a successful reconstruction have significant calibration statistics, but its verification statistics must demonstrate that the independent estimates continue to be accurate and that the accuracy measurement is not likely to be the result of chance. The following discussion is an abbreviated version of an unpublished report on the topic by Gordon (1980) (more details are given in App. 1, Sec. D).

D. Verification Statistics

Statistical verification involves the comparison of independent predictand estimates with corresponding instrumental data and the calculation of a score or statistic that measures their similarity. Various statistics can be used. Some, called *parametric statistics*, involve assumptions about the underlying distributions of the data and may be sensitive to violations of those assumptions (Graumlich 1985). Others, called *nonparametric statistics*, do not involve such assumptions but are generally less sensitive measures of agreement between the predictand estimates and instrumental data. Both kinds of statistics are used in this study, and a variety of statistical tests are applied in different ways to assess different attributes of similarity.

The well-known *product moment correlation* coefficient, r (Clark 1975), measures the relative variation (covariance) that is common to two data sets and reflects the entire spectrum of variation, including both high and low frequencies. The correlation coefficient can be affected markedly by any trends in the two time series that are being compared. It is totally insensitive to differences in the scale and to differences in the mean between the two data sets. The effect of trends can be eliminated by calculating a new correlation coefficient from the first differences of the two data sets. Correlations calculated from the first differences, r_d, measure only the high-frequency variation in common expressed by the year-to-year differences.

The *sign test* is a nonparametric statistic involving a count of the number of times that departures from the sample means agree or disagree. The number of signs is significant whenever it exceeds the number expected from random numbers. The test measures the associations between two series at all frequencies but does not assume that the data are normally distributed. A similar test is made for the first differences, and the test of significance is the same as the test of departures. The sign of the first difference measures the associations at high frequencies.

The *product means* (*PM*) test (Fritts 1976) accounts for both the signs and the magnitudes of the similarities in two data sets. It emphasizes the larger deviations from the mean over the smaller deviations by collecting the products of the deviations in two separate groups based on their signs. The means of these two groups are calculated, and the difference between the absolute values of the two means is tested for significance.

The *reduction-of-error* (*RE*) statistic provides a sensitive measure of reliability and has useful diagnostic capabilities (Gordon 1980). It is similar in some respects to the explained variance statistic obtained with the calibration of the dependent data (Lorenz 1956, 1977). The value of *RE* can range from negative infinity to a maximum value of 1.0, which indicates perfect estimation. The theoretical distribution of the *RE* statistic has not been determined adequately, so its significance cannot be tested. Any positive value of *RE* indicates that the regression model, on the average, has some skill and that the reconstruction made with the particular model is of some value. The errors are unbounded, however, so that one extreme error value in what was otherwise a nearly correct set of estimates could cause the *RE* statistic to be negative.

The *RE* can be partitioned into three component parts—the *RISK*, *BIAS*, and *COVAR* terms—which express various attributes of the relationship (Gordon and LeDuc 1981; Gordon 1980). These components can be extremely useful as diagnostic tools for analyzing sources of error affecting a particular climatic reconstruction.

The *RISK* term is always negative; its absolute magnitude is a comparative measure of the variability of both the estimates and the actual observations used in testing. This term represents the risk that the model takes in making the independent estimates. Ideally, the *RISK* term should be -1.0 (see App. 1, Sec. D.4). Estimates with a small explained variance usually have *RISK* values between -0.5 and 0.0, and reconstructions that have a larger variance than the actual data will have values that are less than -1.0. This overspecification of the variance can occur when an excessive number of predictors is included in the transfer function.

The *BIAS* term is positive when the mean of the estimates is on the same side of the calibration mean as the actual independent climatic data used for the verification testing. Usually, shifts in the mean are insignificant, but for a small sample the *BIAS* can be an important *RE* component. The covariation term, *COVAR*, reflects the strength of the correlation between the reconstruction and the instrumental data and measures the similarity of the temporal patterns in these two data sets. To obtain a positive *RE*, the *RISK* term must be offset by the accuracy of the estimates as indicated by the *BIAS* and *COVAR* terms.

The partitioned *RE* components can be used in the following ways to diagnose reconstruction characteristics. Some reconstructions can successfully duplicate the temporal patterns of variation in the actual observations but contain no appreciable amount of variability. The correlation coefficient would not differentiate such a reconstruction from one with more variability, but the *RISK* term would clearly reveal this difference. Cases frequently arise in which regression estimates have a negative *RE* statistic and yet still pass a majority of other verification tests, especially the correlation statistics. In this situation, a *RISK* term that is less than -0.1 may reveal that the model has overestimated the instrumental data variance. A negative *BIAS* term may indicate differences in the reconstructed and instrumental mean values. Of course, a small *COVAR* term would occur only if there is little correlation.

It is possible that nonlinear relationships could exist between the two samples which cannot be properly evaluated by using the aforementioned verification statistics. Therefore, a *contingency analysis* (Beyer 1968) was used to test for a relationship without making any assumption about linearity. A chi-square statistic is then used to test whether the relationship is sufficiently strong to be significant.

E. Strategy Imposed by Data Availability

All available temperature and precipitation data from the same stations used for calibration but reported for years before 1901 were used for the verification of temperature and precipitation. Each statistic was calculated only for stations with at least seven independent observations of climate. A different calibration and verification strategy, called *subsample replication* (Mosteller and Tukey 1968, 1977; Stone 1974; McCarthy 1976; Gordon 1980), was used to obtain the independent data for verification of sea-level pressure (see App. 1, Sec. G.4, for more information).

All of the verification tests applied to temperature and precipitation estimates at individual stations could be applied to the gridded sea-level pressure estimates. It was more efficient, however, to apply them to the PCs of sea-level pressure, and this method avoided problems with spatial correlation. These series diminished in variance as the order of the PC increased, though, and it was not appropriate to normalize the PC values before calculating the statistics. As a result, the statistics requiring normalized estimates such as the contingency analysis could