**Title**

Empirical Testing of Prediction Interval Methods used for Climate Reconstruction from Crossdated Archives

**Authors**

David C. Edge1,2, Bryan A. Black1

**Author Addresses**

1Laboratory of Tree Ring Research, University of Arizona, 1215 E Lowell St, Tucson, AZ 85721, USA

2School of Earth and Sustainability, Northern Arizona University, 624 Knoles Dr, Flagstaff, AZ 86011, USA

Corresponding Author: David C. Edge, david.edge@nau.edu

**Abstract**

Paleoclimate reconstructions often include an estimation of uncertainty. Upper and lower bounds are commonly calculated to depict the envelope within which some percentage of the climate target values will be located in times prior to the availability of instrumental climate records. Amongst reconstructions derived from crossdated archives, such as the rings of trees and marine bivalves, various methods are used to calculate uncertainties in the underlying chronologies. Because the representativeness of a chronology value in a given year is variable, this uncertainty must be included in the reconstruction uncertainty. We utilized chronologies from bivalve (*Panopea generosa*) ring widths, mountain hemlock (*Tsuga mertensiana)* maximum latewood density, and blue oak (*Quercus douglasii*) ring widths alongside their respective climate targets to calculate prediction intervals using a variety of methods. For each of the three original chronologies, 200 synthetic chronologies were developed with varying chronology properties. For all 603 chronologies, 90% prediction intervals were calculated by six different methods, and 50% prediction intervals were calculated by three different methods. For each reconstruction, the percentage of target values captured by the prediction intervals was measured in a period independent of the reconstruction calibration and prediction interval calculations. 90% prediction intervals based on maximum entropy bootstrapping (MEboot) and empirical error calculation performed best, with 90% of all trials capturing 90 ± 2.0% of target values. Chronology properties had no impact on the reliability of prediction intervals by this method. Prediction intervals calculated using traditional bootstrapping were consistently wider than necessary. Reliability of these prediction intervals was also impacted by chronology properties, particularly the average correlation between series (rbar).

**1. Introduction**

In dendroclimatology, annual reconstructions of past climate typically consist of “estimates” of a particular climate variable at a particular location, for example, July-September air temperature. These estimates are based on a calibration relationship between the tree-ring record and the measured July-September temperature over an interval when both records are available, often called the calibration interval. The values of the tree-ring record are transformed into equivalent temperature units, called the reconstruction, providing estimates of temperature over the full tree-ring chronology time interval. In each year of the calibration interval, the difference between the measured air temperature and the estimated value can be measured. Reconstructions often display confidence intervals in the calibration period, for example a band around the reconstruction that envelopes 90% of the measured instrumental values. In the time period prior to instrumental measurement data, an analogous band called the reconstruction prediction interval is also often shown. In contrast to the simple nature of the confidence interval, in which the instrumental data values are known, prediction intervals must be estimated based on the relationship between the reconstruction and instrumental target in the interval of data overlap. (Cook and Kairiukstis, 1990)

Accurate prediction intervals are an important and necessary component of a climate reconstruction. A reconstruction that fits a target variable closely in the calibration interval may be a very good proxy of that target. However, it may demonstrate similar character by chance, or the reconstruction model may be overfitted. Prediction intervals purport to show the quality of fit of a reconstruction. A climate reconstruction figure without prediction intervals is not an adequate representation and may mislead an end-user as to the degree of uncertainty.

Many methods have been used to display reconstruction uncertainty outside of the calibration interval, the simplest of which is extending confidence intervals, while more complex approaches employ prediction intervals and regression error. In a common example of the simpler method, the average difference between the individual reconstruction and climate values in the interval of overlap, the root mean squared error (RMSE), is sometimes extended into the pre-instrumental period. This method, however, fails to account for the overfitting of a calibration model to the target values in a particular interval. A valuable lesson can be extended from the practice of testing reconstruction calibrations, in which independent verification intervals are utilized to confirm the validity of calibrations. This testing of climate reconstructions has revealed that over-fitting of models can occur, leading to inaccurate reconstructions and reconstruction error estimates. Because reconstruction uncertainties are based on the reconstruction itself, and the reconstruction is susceptible to calibration-period overfit, we have to measure uncertainties in an independent interval. The independent verification interval used in calibration-verification can be used to measure the reconstruction uncertainty outside of the calibration window, allowing for the calculation of true prediction intervals. This is a significant improvement over merely extending confidence intervals calculated in the calibration interval, however, many variations of this method could be implemented, and the method does not provide any test of the prediction intervals. Here we test many methods of producing prediction intervals by setting aside an additional independent period to validate those intervals. We implement different bootstrapping techniques to account for variability in the climate signal within the chronology as well as different techniques to measure the reconstruction error outside of the calibration interval.

**2. Methods**

2.1 Chronologies

Prediction intervals were calculated and tested using 603 chronologies (3 real, 600 synthetic) and their corresponding targets. All real chronologies were developed from crossdated, replicated, annually resolved growth-increment datasets. We selected chronologies that produced reconstructions of a climate variable by simple regression for simplicity and contained at least 60 years of chronology-target overlap to provide sufficient data for independent intervals for calibration, prediction interval calculation, and prediction interval testing. Chronologies were selected to represent a range of values for sample depth, rbar, first-order autocorrelation (AR1), and correlation to target.

The Tree Nob chronology was developed from Pacific geoduck in coastal British Columbia, Canada, and extends continuously from 1725 to 2008 (Edge et al., 2021). The published reconstruction targets April-November sea surface temperature at Langara Island. The overlap of these records covers 1940-2001, 62 years, with an average sample depth of 15.8 measurement time series. Although the published reconstruction utilizes more complex methods, simple linear regression is used to reconstruct seasonal Langara SST from the published Tree Nob ring-width indices (<https://www.ncei.noaa.gov/access/paleo-search/study/33312>). The original ring widths were first detrended by regional-curve standardization, then log transformed to produce the indices used.

The Rock Springs Ranch chronology is based on the ring widths of Blue Oak in San Benito County, California, USA (https://www.ncei.noaa.gov/access/paleo-search/study/8564?siteId=22851). The chronology spans 1379 to 2003 and is highly sensitive to the local hydroclimate (Stahle et al., 2013; Griffin and Anchukaitis, 2014). Ring widths were detrended with 2/3-length 50% frequency cutoff cubic splines (Cook and Kauriukstus, 1990). Monthly precipitation data from the PRISM analysis 4kM2 was downloaded from KNMI climate explorer (Trouet et al., 2013; https://climexp.knmi.nl). Based on significant correlation values, a target of total January-February precipitation was selected. The chronology-target overlap extends from 1895-2003, 109 years, with an average sample depth of 45.3 measurement time series.

The Arrowsmith Mountain chronology was developed from mountain hemlock on Vancouver Island, British Columbia, Canada and extends from 1629 to 1983. The chronology was originally contributed to the International Tree Ring Database by Briffa, K.R.; Schweingruber, F.H. with no climate reconstruction associated (https://www.ncei.noaa.gov/access/paleo-search/study/2808). A maximum latewood density (MXD) chronology was later developed (Schweingruber, 1988) and subsequently used for climate reconstruction, which is the dataset used in this study (Schweingruber et al., 1991; Briffa et al., 1992; Wiles et al., 1996). The MXD chronology is sensitive to temperature air temperature. In order to define a target season, monthly surface temperature values from the nearest grid box of the HADCRUT5 surface temperature dataset from 1857 to present were downloaded from KNMI climate explorer. Based on significant correlation values, a target of average April-October temperature was selected. The chronology-target overlap extends from 1857-1983, 127 years, with an average sample depth of 28 measurement time series.

2.2 Synthetic Chronologies

In addition to the real chronologies tested, we developed 200 synthetic chronologies for each chronology-target pair. Synthetic chronologies provide additional opportunities to test the prediction interval methods for each climate target and help determine which chronology characteristics most determine the prediction interval fidelity. Key characteristics of the synthetic chronologies were varied to provide information on their contribution to prediction interval reliability. These parameters include correlation to climate target, first-order autocorrelation, sample depth, and rbar, defined as the mean value of all possible pairwise correlations among samples. The sample depths of the three sets of synthetic chronologies were constrained to either half, double, or identical to that of the original chronology. The AR1, rbar, and correlation to the climate target were partially constrained by the character of the original chronology but also given a random component.

2.3 Bootstrapping

All chronologies, real and synthetic, were bootstrapped using both traditional (Efron, 1979)) and MEboot methods (Vinod, 2006; Cook et al., 2013). Traditional bootstrapping was performed by resampling from all values from each year of the chronology to produce a bootstrapped chronology with sample depths identical to the original chronology.

Maximum entropy bootstrapping was performed on each individual timeseries of a chronology us the meboot package in R (Vinod and López-de-Lacalle, 2009). MEboot surrogates are created based on the distances between ordered time series values such that all MEboot surrogate time series, when ordered by value, will have identical sorting order to that of the original time series. A time series 1.1, 1.3, 0.9, 1.7, 1.2 has consecutive distances of 0.2, 0.4, 0.8, 0.5. The trimmed mean of these distances, 0.45, is used to extrapolate the variability for the minimum and maximum values of the series, such that the minimum value for the third value in the time series is 0.9-0.45 and the maximum value at the fourth value in the time series is 1.7 + 0.45. All other ranges are given by the averages of the ordered values, so after sorting the time series by value to 0.9, 1.1, 1.2, 1.3, 1.7, we find midpoint values of 1.0, 1.15, 1.25, 1.5. Therefore, all MEboot ensemble members would be drawn from the uniform distributions: (1.0,1.15), (1,25, 1.5), (0.45, 1), (1.5, 2.15), (1.15, 1.25) (Vinod, 2006).

For each bootstrapping method, 1000 sets of bootstrapped indices were produced from each chronology. 1000 bootstrapped mean-value chronologies were developed for each set of indices based on the annual robust biweight mean. 5th and 95th as well as 25th and 75th percentile values at each year were retained for 90th and 50th percentile chronology confidence intervals from the pool of 1000 mean-value bootstrapped chronologies for each bootstrapping method.

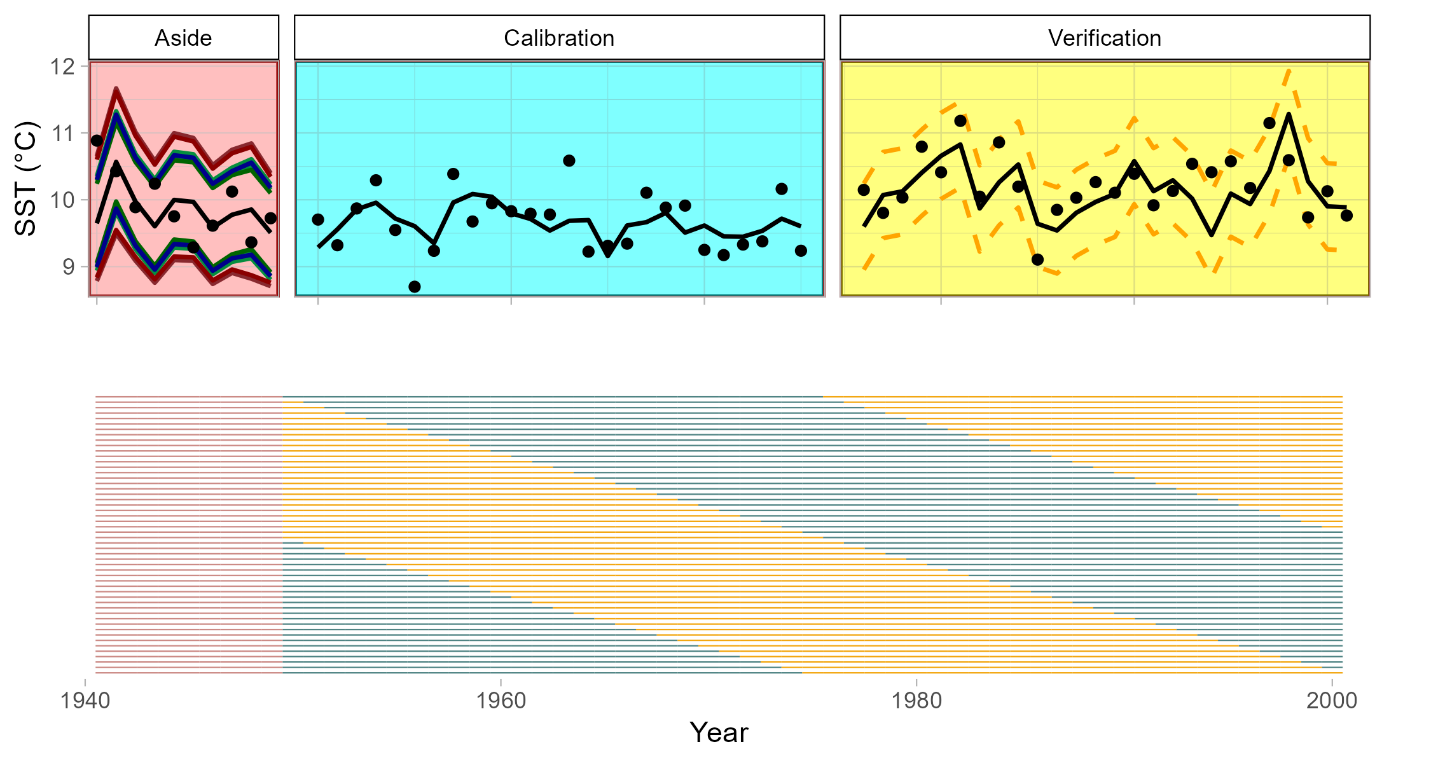
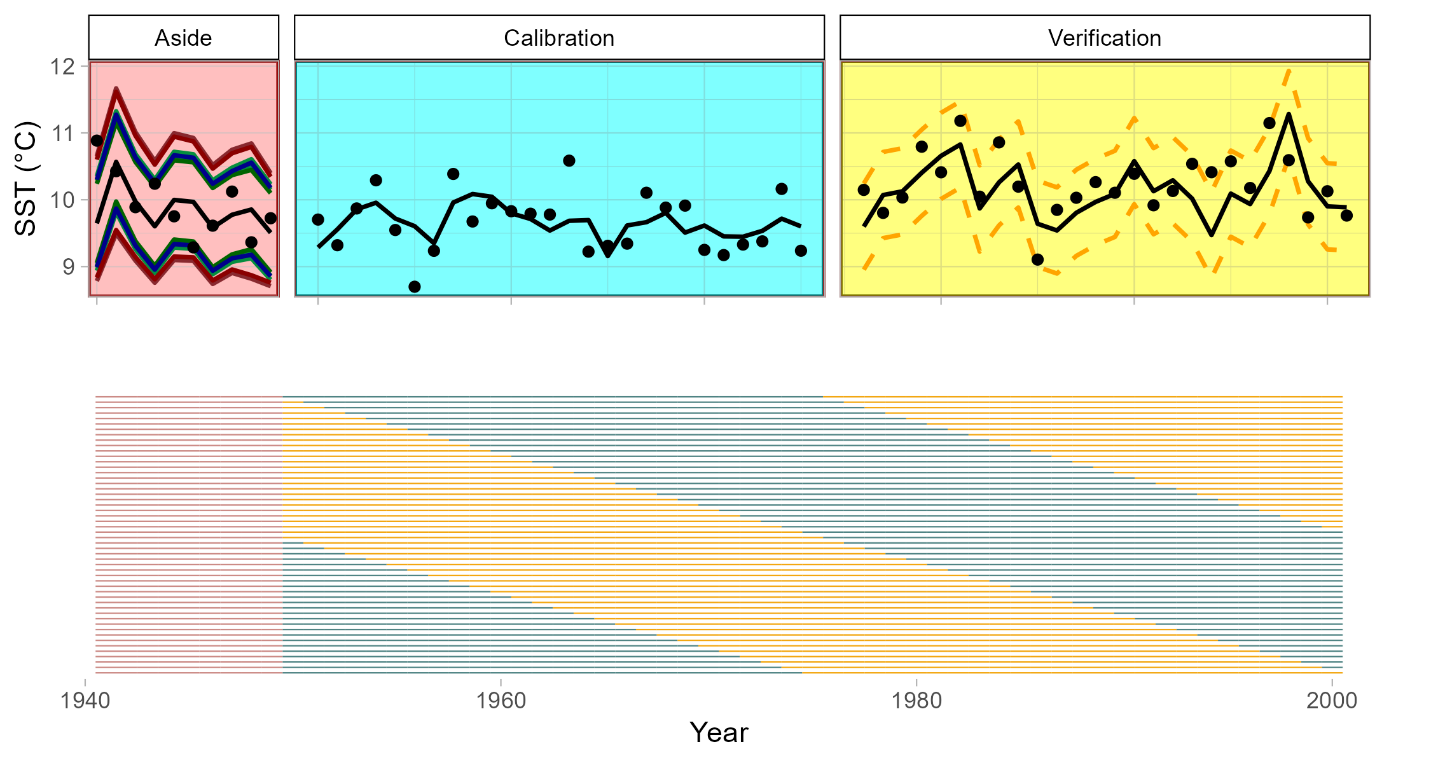
2.4 Reconstructions and Verification Error

The assumptions of regression were tested for each reconstruction. The chronology is regressed onto the target for the full interval of overlap and residuals were tested for autocorrelation, normality, and homoscedasticity. First-order autocorrelation was measured, and chronology residuals of AR1 greater than 0.25 are considered to fail this assumption. Normality of the residuals was tested by the Shapiro-Wilk test, using the stats package in the R programming language, with p-values less than 0.05 considered failing (R Core Team; 2022). Homoscedasticity was tested with the Goldfeld-Quandt test provided by the lmtest package in R with p-values less than 0.05 considered failing (Zeileis and Hothorn 2002).

Prior to calibration of the reconstruction, a ten-year interval was first set aside (called the set-aside interval (SAI)) for independent testing of the prediction intervals (Fig 1). The remaining interval of chronology-target overlap was split in half, the early portion, called the calibration interval, was used for calibrating the regression. The latter portion, called the verification interval, was used to calculate the verification error (VE). Each reconstruction was performed by simple regression between the mean-value chronology and the climate target. The chronology confidence intervals were also regressed on the climate target to capture regression error.

The verification error (VE) set is the group of values given by the absolute value of the difference between the reconstructed and target value in the verification interval for each year. The empirical 50th and 90th percentile verification intervals (VEe50, VEe90) are given by the ascending order 50th and 90th percentile values from the VE set. The theoretical 50th percentile VE (VEt50) is the median value in the VE set, identical to VEe50. Calculation of theoretical VE values assumes the errors are distributed normally and therefore uses values from a t-table based on the percentile error of interest and the degrees of freedom in the distribution. The VEt90 is calculated by adding *x* number of standard deviations to VEt50 where x is the t-table value described above multiplied by the standard deviation of the VE set.

Figure 1 Prediction Interval Testing Method. A total of 3223 calibration and verification calculations and sixty-two set-aside tests are used to define the prediction interval capture for each method for this chronology. Panels a and b show the calculations for one set-aside interval. a) **Calibration**: chronology regressed onto target values, displayed as black points. Reconstruction for Calibration and Verification intervals defined by this regression. **Verification**: orange dashed line: 90-percentile empirical error. Distance from reconstruction to each point measured, empirical and theoretical errors for 50- and 90-percentile defined. **Aside**: black line: reconstruction based on all data in Calibration and verification intervals. colored lines: 90-percentile empirical and theoretical prediction intervals green: no bootstrapping. Blue: MEboot. red: Traditional bootstrapping. Precent capture measured for each method as the percentage of target points falling within the prediction intervals. b) Pink box: Aside interval as in panel a. Blue/yellow lines: calibration/verification intervals, 52 of each. Average verification error values from the 52 possible verification intervals used to build prediction intervals in one set aside interval.



**a**

**b**

All possible continuous intervals were used as calibration and verification intervals (Fig 1) such that a 50-year overlap (after setting aside 10 years for prediction interval testing, 60 years total overlap) produced 50 possible calibration and verification intervals, allowing that the calibration and verification intervals could wrap from the end of the total overlap interval back to the beginning. Thus, for a chronology-target pair with 60 years of total overlap, with 10 years set aside for prediction interval testing, 50 sets of reconstruction regression coefficients, regression error terms, and theoretical/empirical 50th/90th-percentile verification errors are calculated.

2.5 Prediction Interval Testing

All prediction intervals were calculated and tested in the independent 10-year SAI (Fig 1). Prediction interval methods are defined by VE (theoretical, empirical), chronology confidence interval bootstrapping method (none, traditional, MEboot), and intended percent capture (50%, 90%). The possible permutations of these options produce 12 different sets verification intervals, however, because theoretical and empirical VE is identical for 50% intended capture, only 9 unique prediction intervals are produced.

The prediction interval capture (PIC) was calculated for each method by summing the total number of target values captured by the prediction intervals and dividing this total by the length of the independent interval. For example, if 9 of the 10 SAI climate target values fall within the prediction intervals, 90% of the target values were “captured”, in line with the intended capture (IC) rate. This test was repeated for all possible continuous 10-year intervals, such that a chronology-target pair with 60 years of overlap contains 60 possible, overlapping testing intervals. The mean PIC from all trials (PICm) for a given chronology and prediction-interval method was captured for comparison.

The PIC was also compared with the four chronology parameters for all 603 chronologies. Strong correlations between chronology properties and PIC for a given prediction interval method suggest a bias of the method that will produce prediction intervals too wide or narrow based on idiosyncrasies of individual chronologies.

**3. Results**

3.1 Synthetic Chronologies

The synthetic chronologies varied considerably in the four properties parameterized (Fig 2). The rbar, sample depth, AR1, and correlation to target of each of the three original chronologies fell within the spread of the associated synthetic chronologies. The range and distribution of each of these properties differed across the three sets of synthetic chronologies, though generally overlapped one another.

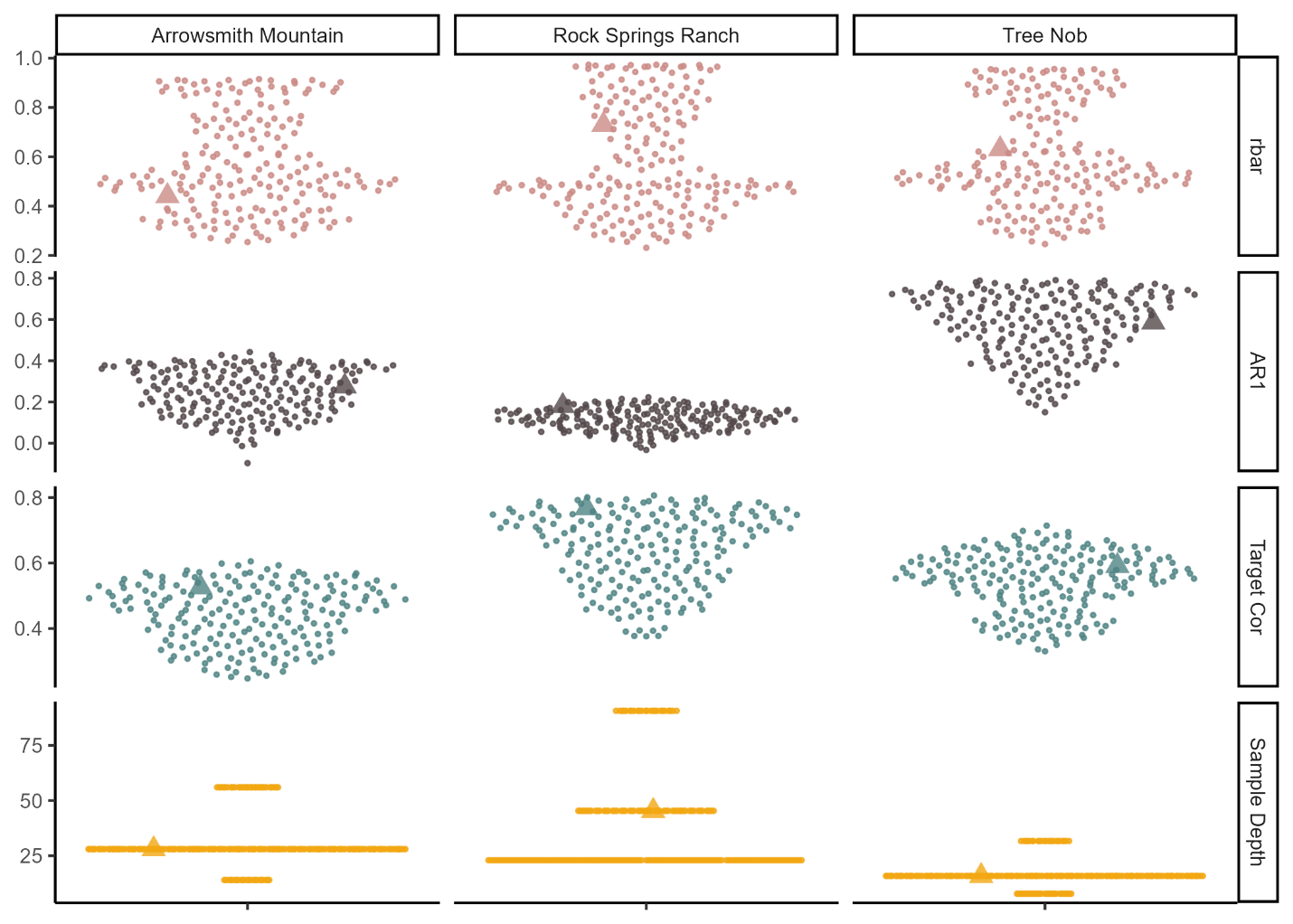
3.2 Bootstrapping

The width of traditional bootstrapped intervals was consistently greater than the MEboot intervals at both 50- and 90-percetile (cite what is now Fig. 4?). Both techniques produced variable chronology error through time, unequal positive and negative errors at any given year, and generally larger error ranges at more extreme chronology values.

3.3 Reconstructions and Verification Error

Many of the chronologies failed one or more of the regression assumptions tested (Fig 3). The regression residuals from the Arrowsmith Mountain chronology and the majority of the associated synthetic chronologies failed the assumption of no autocorrelation. Several of the Arrowsmith Mountain chronologies also failed the test of normality in regression residuals. Several Rock Springs Ranch synthetic chronologies failed tests of normality and homoscedasticity, though all passed the autocorrelation test. Nearly all chronologies from Tree Nob passed all regression assumptions, with only three failing the residual autocorrelation test.

Figure 2 Chronology Properties. **rbar**: average pairwise correlation between individual chronology series. **AR1**: first-order autocorrelation of the chronology. **Target Cor**: correlation between chronology and climate target over the full interval of overlap. **Sample Depth**: Average number of series in the chronology during the interval of target overlap. Small circles represent synthetic chronologies, large triangles represent the originals.



In order to use an independent set-aside interval for testing the prediction intervals, we utilized the longest instrumental records possible. Although there were fewer instrumental climate observations available to produce the gridded dataset at earlier times, all time periods are used to represent both calibration and verification intervals, and no trend in uncertainties became apparent.

3.4 Prediction Interval Testing

The various prediction interval methods showed variable success in capturing the intended percentage of target data (Fig 4). The combination of MEboot, empirical reconstruction errors, and 90-percentile prediction intervals was most effective, and also produced the tightest grouping of outcomes (Fig. 4). Of the 603 reconstructions, all prediction intervals produced captured between 86 and 94% of target values, with the 5th percentile capture rate of 88.1% and the 95th percentile rate of 92.0%. 90-percentile prediction intervals generally captured 90% of the target data with greater consistency that the 50-percentile intervals. Prediction intervals produced with MEboot generally captured closer to the intended interval that those produced without bootstrapping or with Traditional bootstrapping (Fig. 4). Empirical reconstruction error improved the fit of prediction intervals over the theoretical errors.

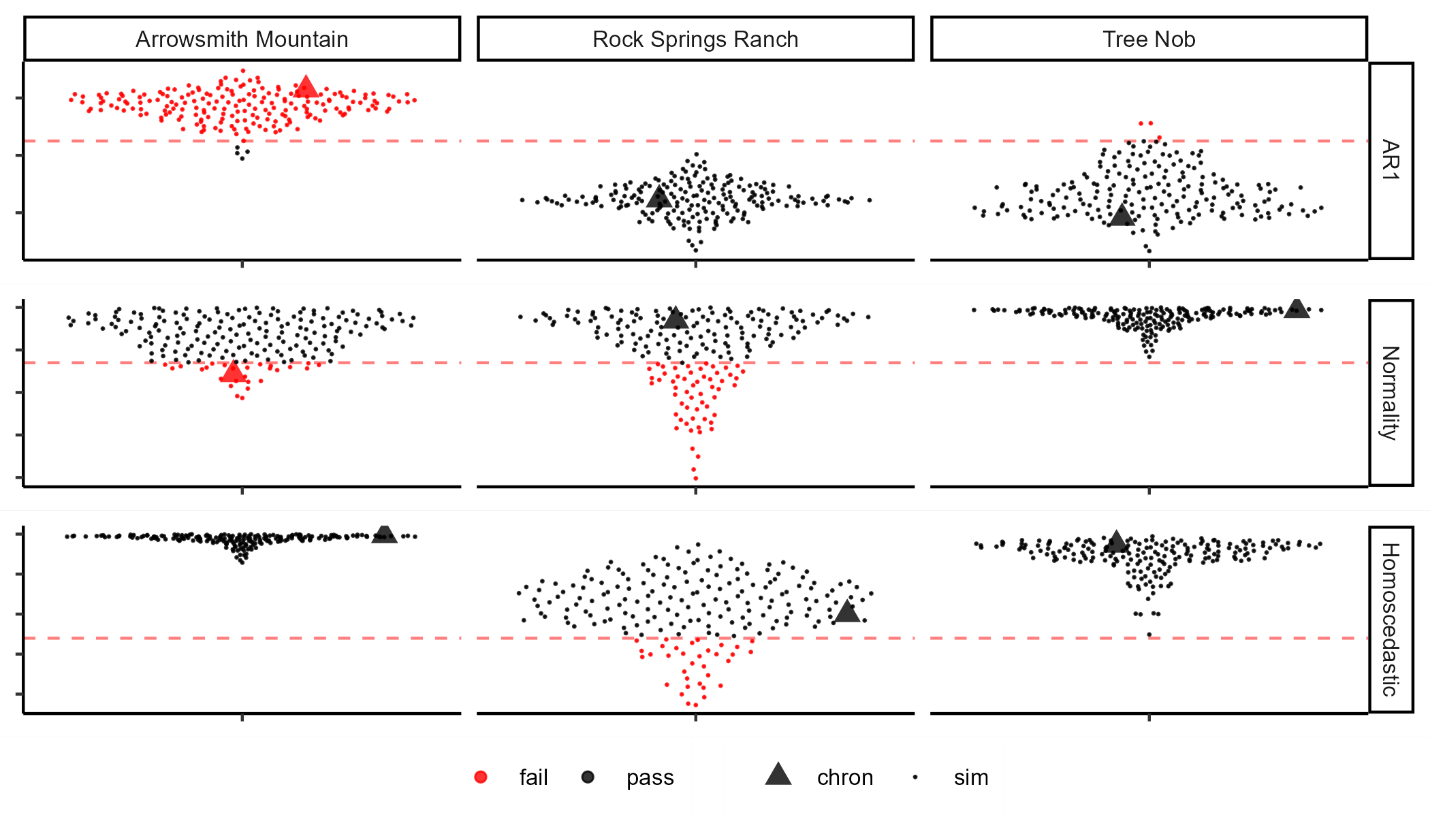


Figure 3 Regression Assumptions Testing. **AR1:** first-orderautocorrelation of residuals; values above 0.25 indicate test failure. **Normality:** normal distribution of residuals at tested by the Shapiro-Wilk normality test, failure below p=.05. **Homoscedasticity**: trend in variance of residuals as tested by the Goldfeld-Quandt test, failure below p=.05. Red points indicate chronologies failing a given regression assumption by the standards listed above. Failure thresholds are demarked by red, dashed lines.

Traditional bootstrapping produced prediction intervals wider than necessary for the intended reconstruction intervals, on average (Fig 4). The prediction interval capture for traditionally bootstrapped chronologies is negatively correlated to the rbar of those chronologies for all methods and chronologies tested (Fig 5). AR1 and sample depth are also negatively correlated to prediction interval capture for some methods and chronology properties. Prediction interval capture based on methods without bootstrapping or by MEboot show little or no consistent relationship to any chronology property.

**4. Discussion**

For producing 90-percent reconstruction intervals, MEboot with empirical reconstruction errors is highly reliable in all conditions tested. These successful reconstruction intervals include results from 281 chronologies that fail one or more regression assumptions and have a wide range of chronology properties.

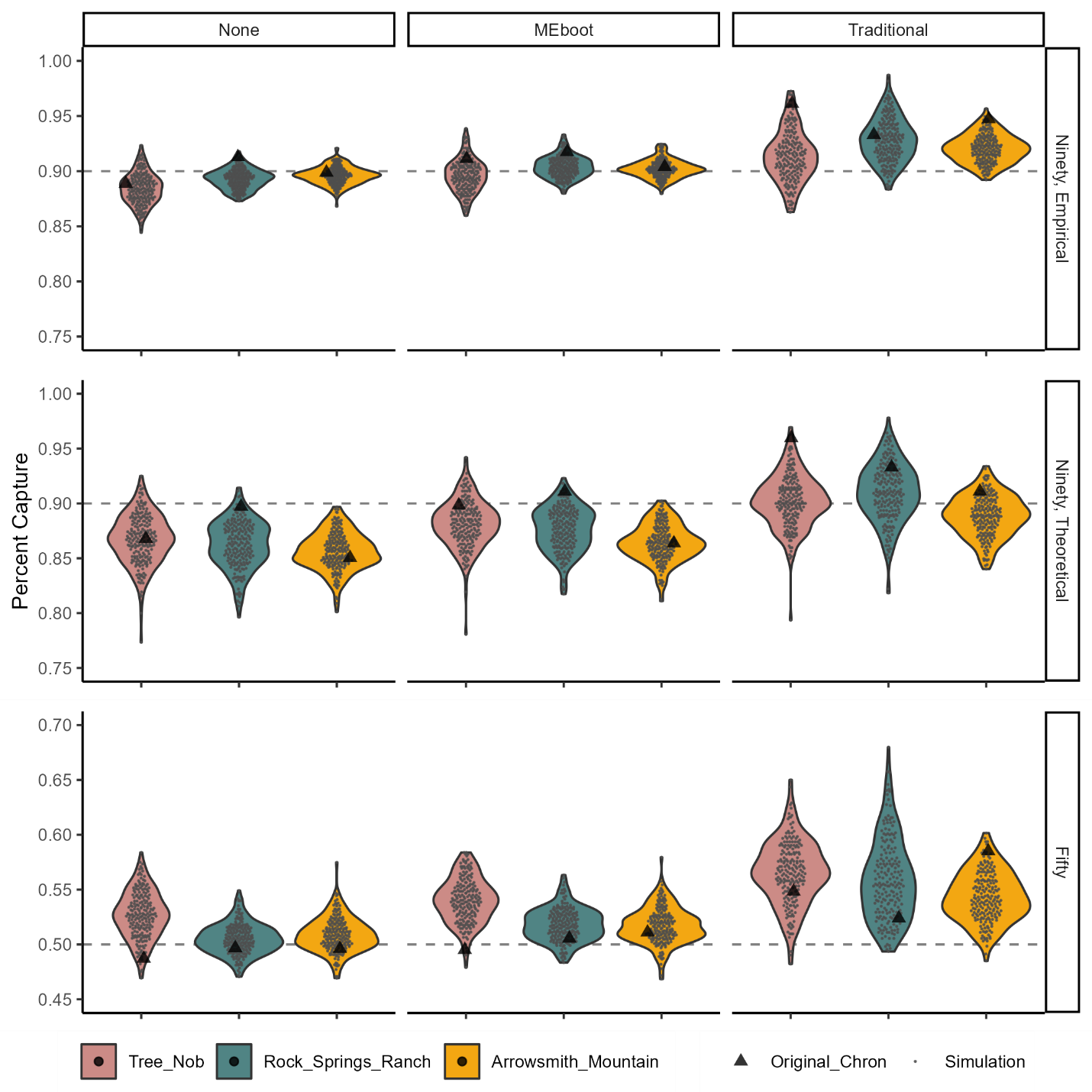
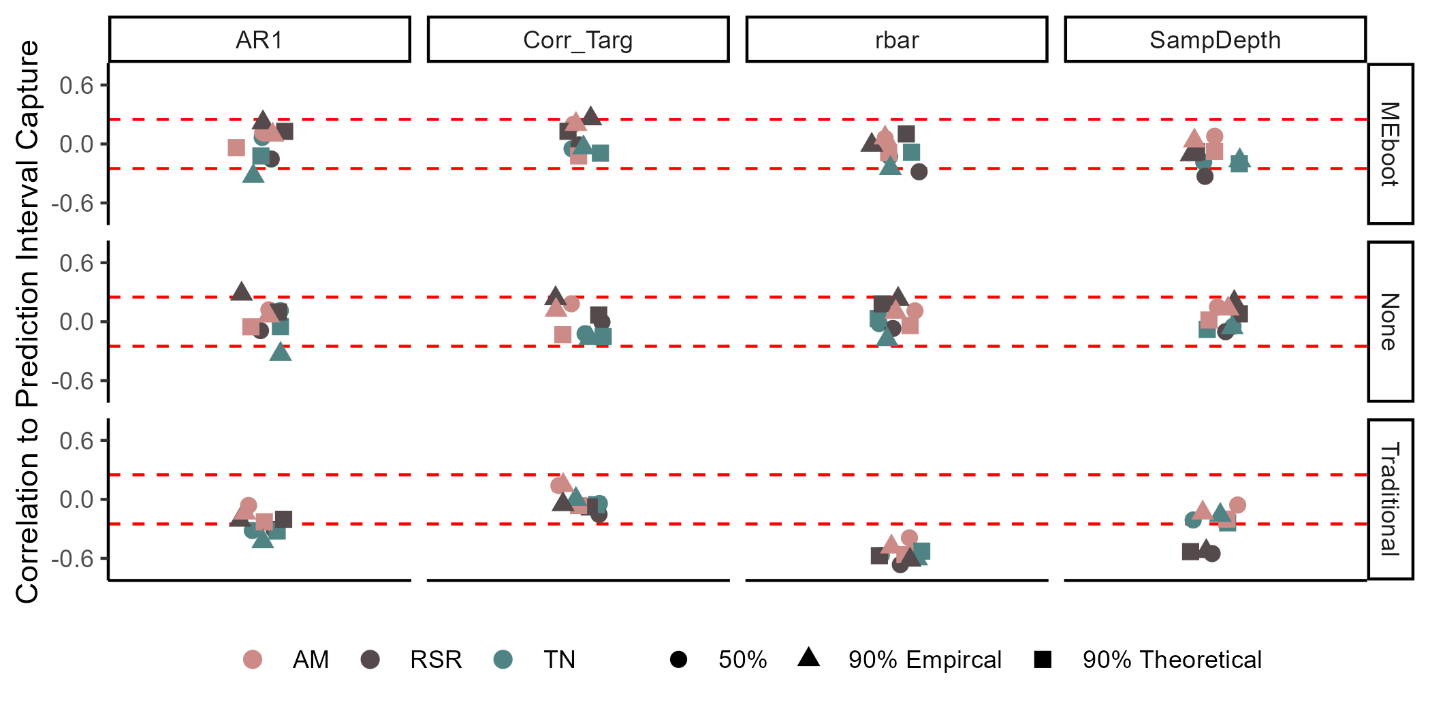


Figure 4 Prediction Interval Performance. **Ninety vs. Fifty**: chronology bootstrapping interval and reconstruction-error interval. **Theoretical vs. Empirical**: method of reconstruction-error calculation. **None vs. MEboot vs. Traditional**: Method of chronology bootstrapping. Dashed lines indicate the intended percent of target values to fall within prediction intervals (intended “percent capture”).

Chronology properties do not have a significant impact on the reliability of reconstruction prediction intervals when constructed by MEboot. Although the chronology simulation algorithm varied each of the chronology properties and affected prediction interval width, the method of bootstrapping had the most significant impact on the reliability of the prediction intervals. In contrast, traditional bootstrapping produces excessively wider prediction intervals, and this is due to the over-sensitivity of this technique to outlier values with biases that are especially pronounced in datasets with low rbar values. The MEboot method assures greater fidelity to the properties of the original series as explained by Cook et al. (2013). MEboot preserves autocorrelation structure while reliably producing median as well as minimum and maximum values of the observed dataset. This is distinct from traditional bootstrapping, which shows a bias toward extreme values when selecting the 5- and 95-percentile representatives.

Figure 5 Correlations between Chronology Properties and Prediction Interval Capture. Individual points represent a correlation between the rate of capture for a prediction interval method at a particular site and the chronology properties of the chronologies utilized (n=201 each). The red dashed lines at r=±0.25 represent a threshold for potentially significant? relationships.



The 90-percent error range is better estimated than 50-percent by the prediction interval methods tested. Generally, smaller prediction intervals are less precise because the target points are more tightly clustered nearer to the reconstruction estimates, assuming the errors are distributed quasi-normally. The average capture of the 50-percent prediction intervals was greater than 50-percent for both bootstrapping methods, suggesting bootstrapping may be unnecessary for capturing errors nearer to the reconstruction estimates. However, the real chronologies show capture rates under 50% for the un-bootstrapped intervals and capture rates of 49-51% for MEboot. So, the chronology simulation algorithm may be introducing chronology characteristics that bias the 50-percent prediction interval capture.

The chronology simulation algorithm was designed to vary important chronology properties in order to test their impacts on the reliability of the prediction intervals. As the synthetic chronologies are based on the original chronologies, they should not be considered independent. It is noteworthy, however, that the prediction interval capture distributions are much more similar across site than across methods at a given site. The intended function of these simulations is displayed in Figure 5, wherein the variability of an important chronology property, rbar, is shown to produce biased prediction intervals when traditional bootstrapping is used. The consistency of this result across the three sites and the strength of the correlation suggest the bias is inherent to the method.

The prediction intervals produced using empirical errors generally outperformed those constructed from theoretical errors. Although normality of regression residuals is an important assumption for this method of reconstruction, strict normality cannot be tested on the small number of errors typical of a calibration. Although the Shapiro Wilk test may be the best test of normality available for small samples, detection is limited to relatively large departures from normality (Ahad et al., 2011). So, error distributions with slightly longer or shorter tails will pass the test, while the 90th percentile error values given by theoretical and empirical approaches will differ. Our results suggest that using the empirical error in prediction interval reconstruction and foregoing the assumption of normal errors, improves the performance of the resulting target capture.

Constructing prediction intervals requires independent time periods in which to calibrate the reconstruction and measure the reconstruction errors. Because many climate proxies derived from crossdated archives contain first-order autocorrelation, consecutive values are not independent. The set-aside intervals could comprise years chosen at random to incorporate the greatest diversity of possible values. Alternatively, the set-aside intervals could only be continuous in order to maintain the persistence structure. The random selection option has the advantage of nearly infinite possible calibration intervals, particularly if resampling is permitted, while a continuous selection limits the number of possibilities. We chose an intermediate solution, wherein all calibration, verification, and set-aside intervals are contiguous with the exception of one break (Fig 1). This allows for a large number of intervals, such that outliers, which are uncommon but can have considerable leverage, are unlikely to produce spurious results. This also allows for testing a calibration interval in a manner fairly analogous to the final reconstruction by maintaining independence of calibration and verification intervals while also accounting for persistence.

The autocorrelation of many of the chronologies also reduced the independence of the set-aside interval, reducing its validity as a test of prediction interval performance. The importance of this phenomenon should be measured by comparing prediction interval performance to AR1, as in Figure 5. The results suggest that the independence of the set-aside interval is not significantly impact by autocorrelation. Tree Nob, the site with the highest AR1 values, does show some negative correlation between PIC and AR1, however, this is not consistent across prediction interval methods.

We have provided evidence for the reliability of prediction intervals constructed from empirical errors and MEboot. However, the simple linear regression method used in the 603 reconstruction here is uncommon in climate reconstructions. Multiple predictors and principal components are commonly employed. We have, though, utilized a diversity of real chronology-target pairs with a range of properties alongside synthetic chronologies with properties of still greater diversity. The methods we have employed can easily be extended to more complex reconstruction methods. Although the robustness of the methods that tested well in these simulations may not hold for all use cases, the simulations do suggest that some methods should not be used. Traditional bootstrapping contains naïve assumptions for use with crossdated archives. This method consistently produced prediction intervals wider than intended in our trials. Furthermore, the consistent negative correlation between prediction interval capture and rbar for traditionally bootstrapped chronologies suggests consistent, predictable bias inherent to this method. We also have not tested all possible bootstrapping methods. Resampling complete individual time series would also maintain the autocorrelation structures of those series while varying the subsets of the sampled population. We did not test this technique here because measurement time series may not overlap to a high degree in tree-ring datasets. Moreover, MEboot provides the benefit of faithfully reproducing data structure while maintaining sample depth characteristics and avoiding issues related to resampling a low number of measurement time series in poorly replicated portions of chronologies.

Estimated error range inherent to any paleoclimate reconstruction is essential to the end-user. Prediction intervals provide at-a-glance information of the relative signal-to-noise ratio of the reconstruction, conveying both what is known and unknown. Reconstruction error is best calculated by measuring errors outside the period of calibration as the errors of interest also lie outside the calibration period. The exact methodology chosen for producing error intervals will impact both the precision and accuracy of the results, and the best methods will likely conform to the appropriateness in relation to the dataset. For many proxies derived from crossdated archives, MEboot is likely to be the most appropriate bootstrapping method. Additionally, utilization of empirical errors rather than assuming a normal distribution of errors may be prudent in many circumstances.

**References**

Ahad, N. A., Yin, T. S., Othman, A. R., & Yaacob, C. R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, *40*(6), 637-641.

Briffa, K. R., Jones, P. D., & Schweingruber, F. H. (1992). Tree-ring density reconstructions of summer temperature patterns across western North America since 1600. *Journal of Climate*, *5*(7), 735-754.

Cook, E. R., & Kairiukstis, L. A. (1990). *Methods of dendrochronology: applications in the environmental sciences*. Springer Science & Business Media.

Cook, E. R., Palmer, J. G., Ahmed, M., Woodhouse, C. A., Fenwick, P., Zafar, M. U., Wahab, M., & Khan, N. (2013). Five centuries of Upper Indus River flow from tree rings. *Journal of hydrology*, *486*, 365-375.

Edge, D. C., Reynolds, D. J., Wanamaker, A. D., Griffin, D., Bureau, D., Outridge, C., Stevick, B. C., Weng, R., & Black, B. A. (2021). A Multicentennial Proxy Record of Northeast Pacific Sea Surface Temperatures From the Annual Growth Increments of Panopea generosa. *Paleoceanography and Paleoclimatology*, *36*(9), e2021PA004291.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer.

Griffin, D., & Anchukaitis, K. J. (2014). How unusual is the 2012–2014 California drought? *Geophysical Research Letters*, *41*(24), 9017-9023.

Schweingruber, F. (1988). A new dendroclimatic network for western North America. *Dendrochronologia*, *6*, 171-180.

Schweingruber, F. H., Briffa, K. R., & Jones, P. (1991). Yearly maps of summer temperatures in Western Europe from AD 1750 to 1975 and Western North America from 1600 to 1982. *Vegetatio*, *92*(1), 5-71.

Stahle, D., Griffin, R., Meko, D., Therrell, M., Edmondson, J., Cleaveland, M., Stahle, L., Burnette, D., Abatzoglou, J., & Redmond, K. (2013). The ancient blue oak woodlands of California: Longevity and hydroclimatic history. *Earth Interactions*, *17*(12), 1-23.

Team, R. C. (2013). R: A language and environment for statistical computing.

Trouet, V., & Van Oldenborgh, G. J. (2013). KNMI Climate Explorer: a web-based research tool for high-resolution paleoclimatology. *Tree-Ring Research*, *69*(1), 3-13.

Vinod, H. D. (2006). Maximum entropy ensembles for time series inference in economics. *Journal of Asian Economics*, *17*(6), 955-978.

Vinod, H. D., & López-de-Lacalle, J. (2009). Maximum entropy bootstrap for time series: the meboot R package. *Journal of Statistical Software*, *29*, 1-19.

Wiles, G. C., D'Arrigo, R. D., & Jacoby, G. C. (1996). Temperature changes along the Gulf of Alaska and the Pacific Northwest coast modeled from coastal tree rings. *Canadian Journal of Forest Research*, *26*(3), 474-481.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships.