**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, 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 knowledge imparted.

Many methods have been used to display reconstruction uncertainty outside of the calibration interval, ranging from the simple extension of confidence intervals to the more complex prediction intervals with regression error. The average difference between the individual reconstruction and climate values in the interval of overlap, the root mean sqared error (RMSE) is sometimes extended into the pre-instrumental period. This method however fails to account for the idiosyncratic nature of a calibration performed in a particular interval, and the practice of independent calibration-verification testing of climate reconstructions gives credence to the issue of over-fitting models. Another common method of producing prediction intervals, making use of an independent verification interval, one can measure the difference between the reconstructed and target values only in this independent interval. Although this is a significant improvement over merely extending confidence intervals calculated in the calibration interval, 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 interval to validate the prediction 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. 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, average correlation between individual chronology indices (rbar), first-order autocorrelation, 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. Although the published reconstruction utilizes more complex reconstruction 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 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 a 2/3 spline (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.

The Arrowsmith Mountain chronology is based on Mountain Hemlock from 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 (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.

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. Important characteristics of the synthetic chronologies are 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. 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 some random component.

2.3 Bootstrapping

All chronologies, real and synthetic, were bootstrapped using both traditional (Efron, 1979)) and maximum entropy bootstrapping methods (MEboot; 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. 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 0.2, 0.4, 0.8, 0.5. The trimmed mean of these distances, 0.45, is used to extrapolate the variability for the min/max values of the series, such that the minimum value for T3 = 0.9-0.45 and the maximum value at T4 = 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 are tested for each reconstruction. The chronology is regressed onto the target for the full interval of overlap and residuals are tested for autocorrelation, normality, and homoscedasticity. First-order autocorrelation is measured, and chronology residuals of AR1 greater that 0.25 are considered to fail this assumption. Normality of the residuals is 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 is tested with the Goldfeld-Quandt test provided by the lmtest package in R with p-values less than 0.05 considered failing (Zeileis & Hothorn 2002).

Prior to calibration of the reconstruction, a ten-year interval is 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 is split in half, the early portion, called the calibration interval, is used for calibrating the regression and the latter portion, called the verification interval, is used to calculate the verification error (VE). Each reconstruction is performed by simple regression between the mean-value chronology and the climate target. The chronology confidence intervals are also regressed on the climate target to capture regression error.

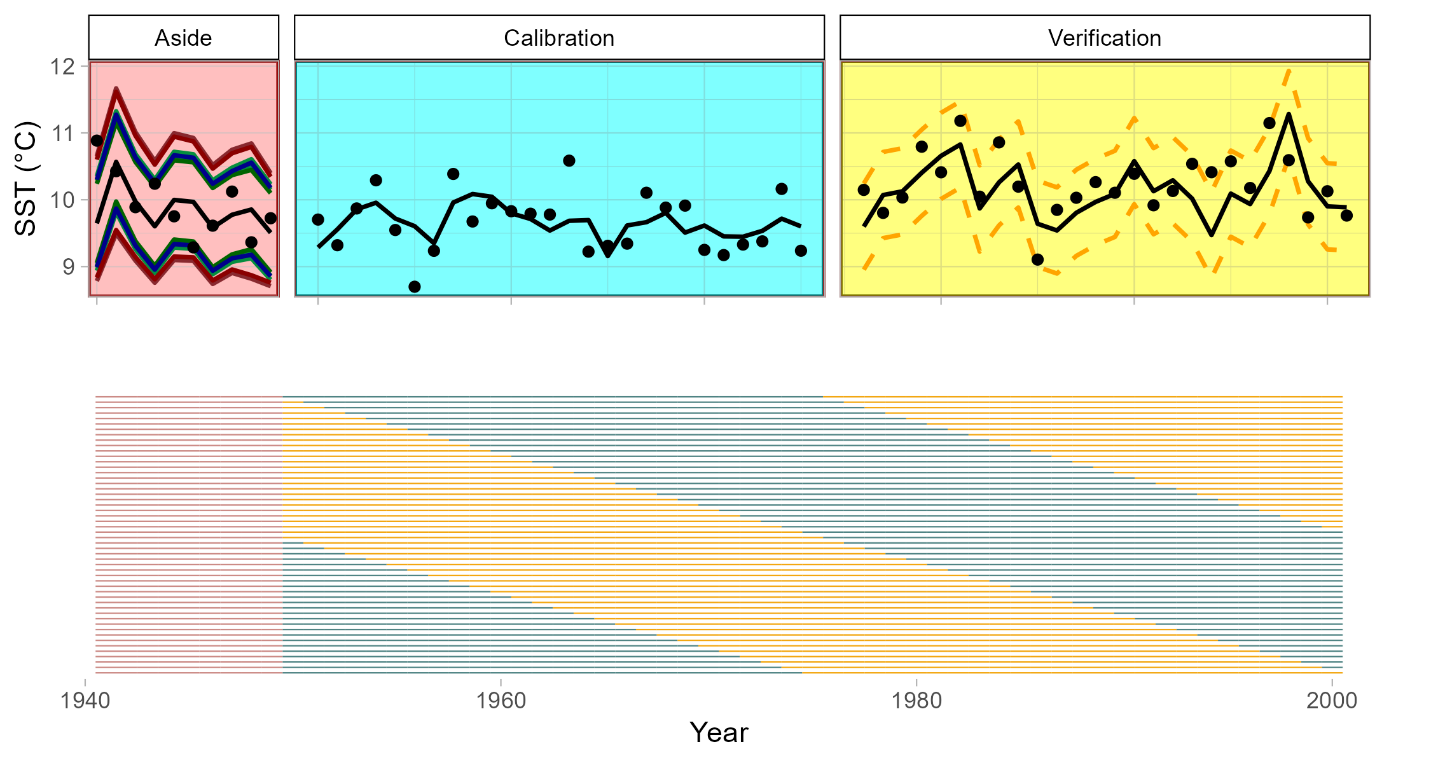
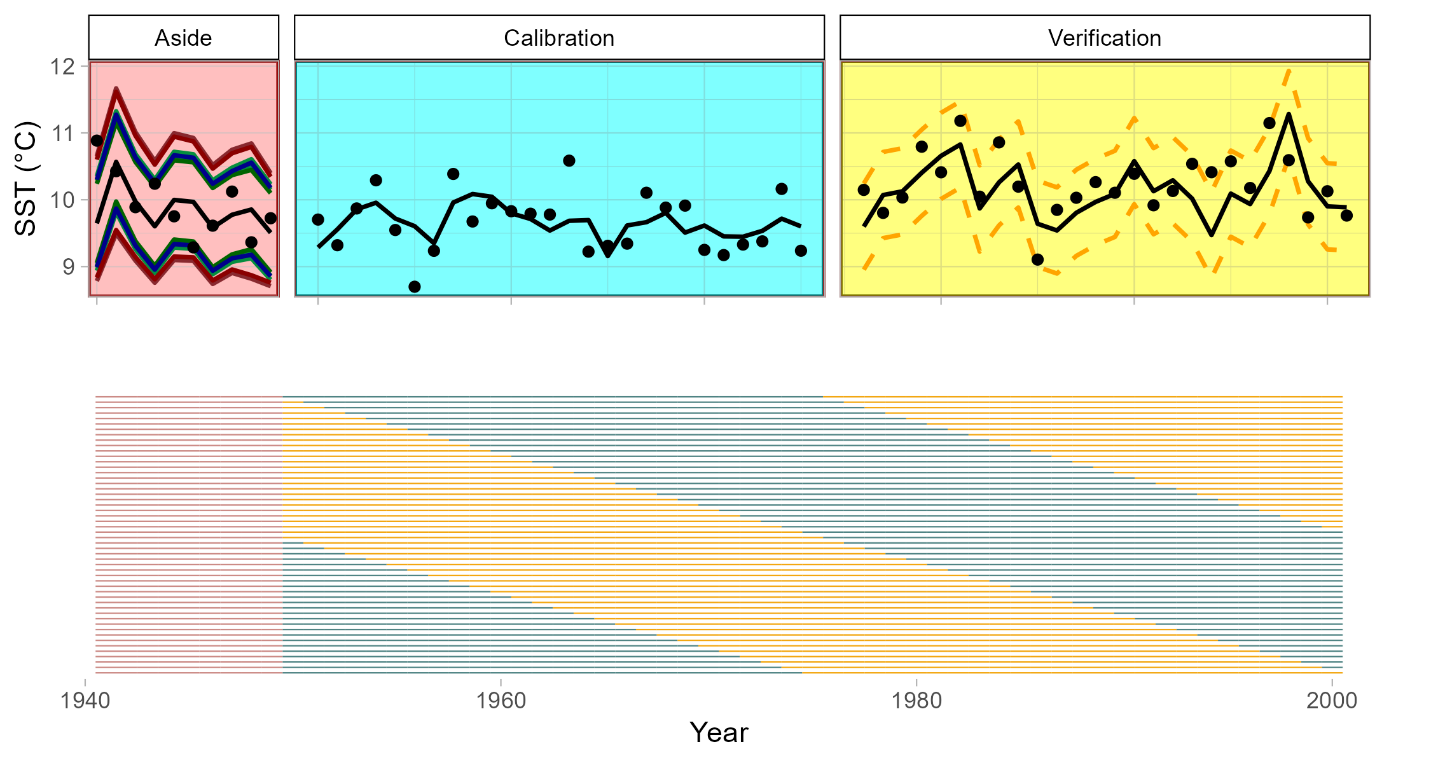
The verification error (VE) set is the group of values given by the absolute value of the difference between the reconstructed and target value in the verification interval for each year. The empirical 50th and 90th percentile verification intervals (VEe50, VEe90) are given by the ascending order 50th and 90th percentile values from the VE set. The theoretical 50th percentile VE (VEt50) is simply 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.

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.

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: 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**

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.

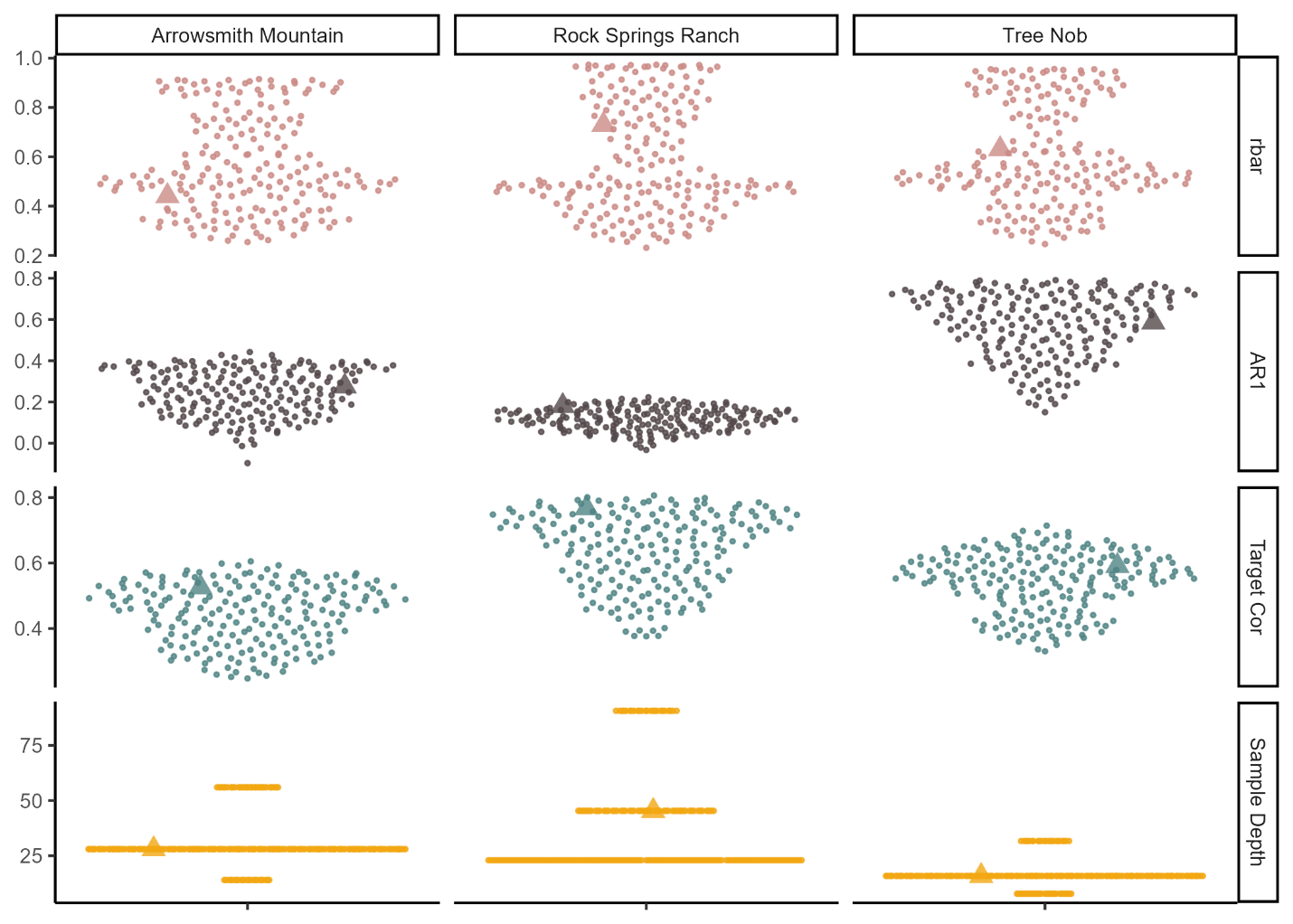
The PIC was also compared with the four chronology parameters for all 603 chronologies. Strong correlations between chronology properties and PIC for a given prediction interval method suggest a bias of the method that will produce prediction intervals too wide/norrow based on idiosynchasies of individual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

**3. Results**

3.1 Synthetic Chronologies

The synthetic chronologies vary considerably in the four properties parameterized (Fig 2). The rbar, sample depth, AR1, and correlation to target of each of the three original chronologies falls within the spread of the associated synthetic chronologies. The range and distribution of each of these properties differs across the three sets of synthetic chronologies, however, the distributions all overlap one another.

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



3.2 Bootstrapping

The width of traditional bootstrapped intervals was consistently greater than the MEboot intervals at both 50- and 90-percetile. Both techniques produced variable chronology error through time, unequal positive/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.

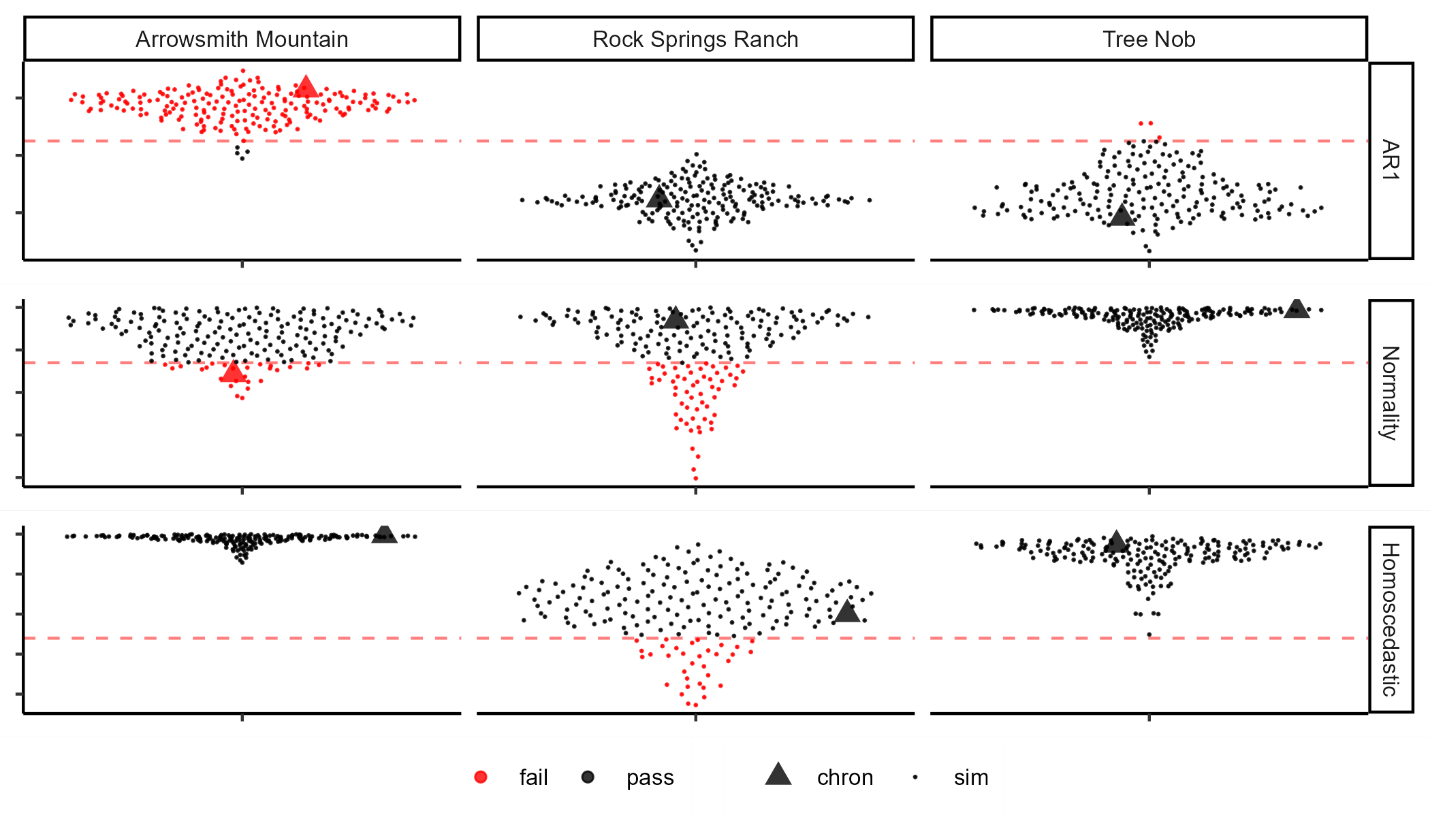


Figure 3 Regression Assumptions Testing. **AR1:** first-orderAutocorrelation of residuals, test failure above 0.25. **Normality:** normal distribution of residuals at tested by the Shapiro-Wilk normality test, failure below p=.05. **Homoscedasticity**: trend in variance of residuals as tested by the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

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 produced both the average measured capture nearest that intended as well as the tightest grouping of outcomes. Of the 603 reconstructions, all prediction intervals produced by this method 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 capture closer to the intended interval that those produced without bootstrapping or with Traditional bootstrapping. Empirical reconstruction error improved the fidelity of prediction intervals over the theoretical errors.

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.

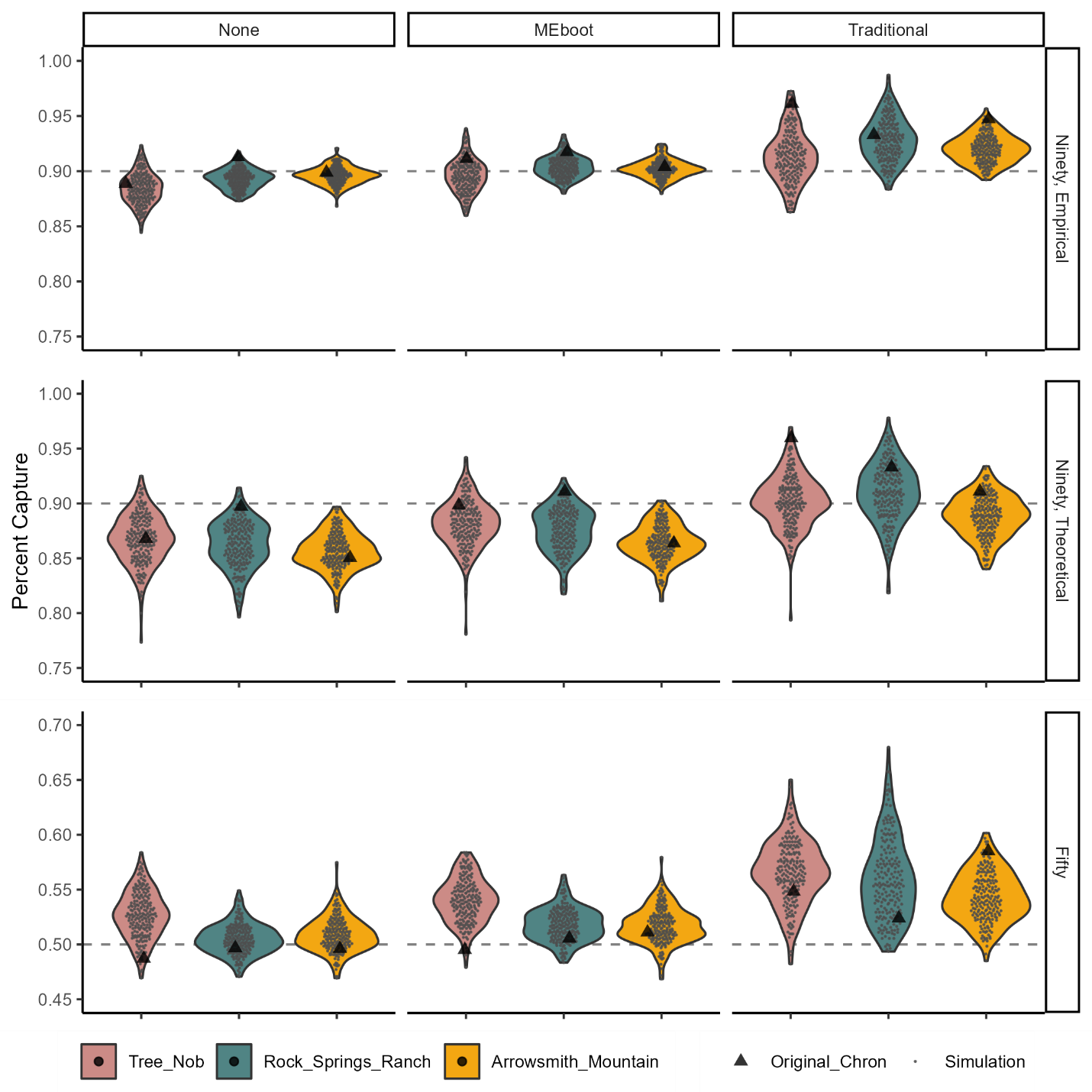
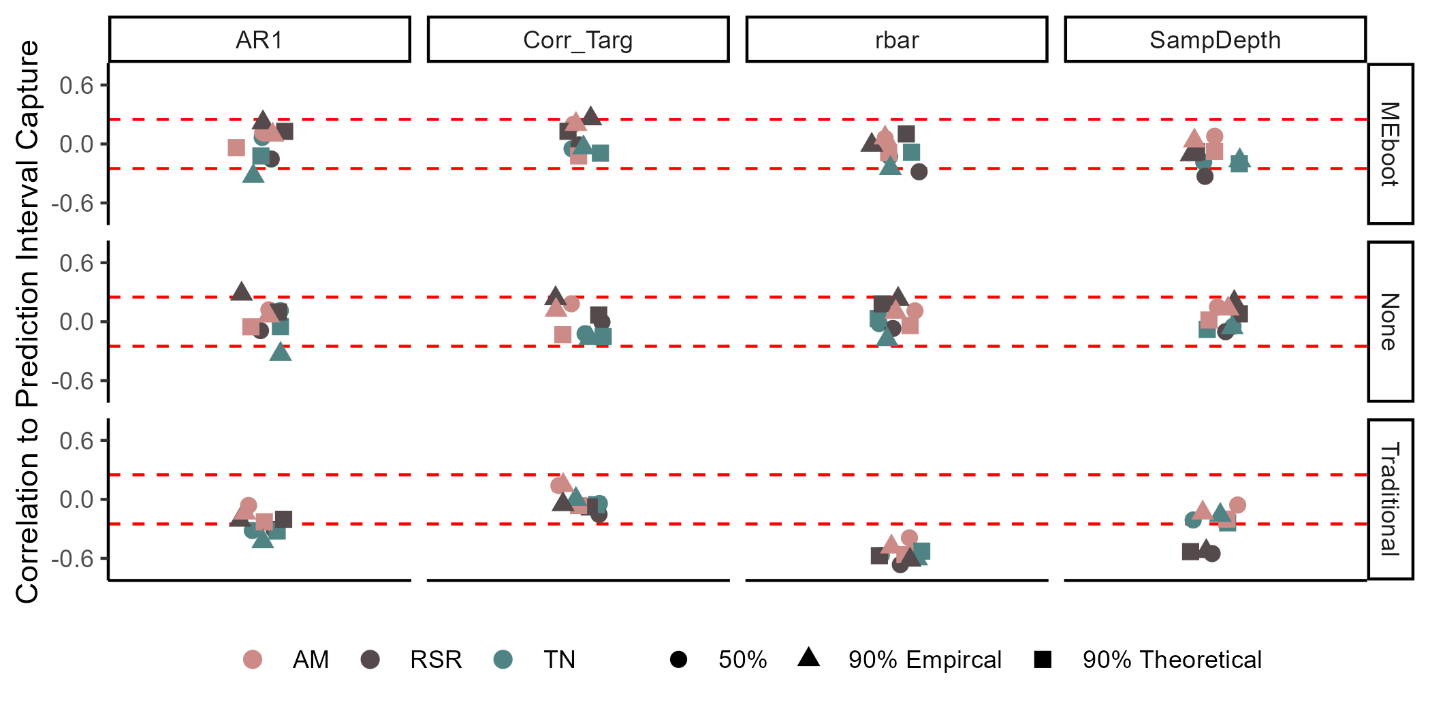


Figure 4 Prediction Interval Performance. **Ninety/Fifty**: chronology bootstrapping interval and reconstruction-error interval. **Theoretical/Empirical**: method of reconstruction-error calculation. **None/MEboot/Traditional**: Method of chronology bootstrapping. Dashed lines indicate the intended percent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

**4. Discussion**

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 important relationships.



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.

Chronology properties do not have a significant impact on the reliability of reconstruction prediction intervals tested when constructed by MEboot. Although each of the chronology properties varied by the chronology simulation algorithm changes the resulting prediction interval width, the method of bootstrapping has the most significant impact on the reliability of the prediction intervals. Traditional bootstrapping shows much greater range in prediction interval capture, and this is due to the over-sensitivity of this bootstrapping technique, particularly to 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 the overall shape (i.e., the temporal order) of the data… MEBoot is also unique in the way it preserves the persistence structure and overall properties of any arbitrary stochastic process, including those that are non-stationary and heteroscedastic.” Thus, MEboot results in little variability for surrogate values near the median value of a time series and greater variability for values near the time series min and max. This is similar to traditional bootstrapping, however, utilizing the input of every time series in each chronology bootstrap helps to avoid a bias toward extreme values when selecting the 5- and 95-percentile representatives.

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 the impacts of those properties 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 for a given method 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. Therefore, the set-aside intervals could comprise years chosen at random to incorporate the greatest diversity of possible values or only continuous intervals in order to maintain the persistence structure and capture potential biases of limited-window calibration. The former option has the advantage of nearly infinite possible calibration intervals, particularly if resampling is permitted, while the latter allows for only a handful. 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 substantial number of possible intervals, such that unique outliers are unlikely to produce spurious results. This also allows for testing a calibration interval in a manner fairly analogous to the final reconstruction in terms of the independence of intervals when 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 factor 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 at Tree Nob.

We have not provided exhaustive evidence for the reliability of prediction intervals constructed from empirical errors and MEboot. 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, however, 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 individual time series would also maintain the autocorrelation structures of those series while sampling different possible subsets of the population. We did not test this technique here because MEboot provides these benefits while also maintaining sample depth characteristics of the original chronology and avoiding issues related to resampling 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.

Appendix 1 (Methods in outline form)

1. Load chronology rwi and target data
2. Bootstrap chronology rwi values, building 1000 replicate chronologies
3. Build mean-value chronologies from the replicate chronologies
4. Measure confidence interval of chronology by sorting mean value chronology data at each year and selecting corresponding 5th and 95th percentile values (90% chronology confidence intervals)
5. Find interval of proxy-target overlap
6. Set aside 10 years of proxy-target overlap for independent testing of the prediction intervals
7. Perform reconstruction calibration by split calibration-verification
   1. Calibrate reconstruction
      1. Capture regression coefficients in calibration interval
      2. Capture regression coefficients for upper and lower chronology confidence intervals to capture “regression error”
   2. Use coefficients to build reconstruction into verification interval
      1. Measure the difference between target and reconstruction values in the verification interval
8. Repeat step 7 for all possible continuous calibration-verification intervals
9. Build prediction intervals in independent set-aside interval
   1. Build prediction intervals based on empirical 90th percentile error
   2. Build prediction intervals based on median error + (standard deviation of error x z-score corresponding to 90th percentile)
   3. Build prediction intervals based on 90th percentile regression error + empirical reconstruction error (7.c.i. above)
   4. Build prediction intervals based on 90th percentile regression error + theoretical reconstruction error (7.c.ii. above)
10. Test prediction intervals in independent set-aside interval
    1. Measure the number of climate target values captured by each set of prediction intervals
    2. Save the captured total divided by the length of the set aside interval (e.g. 9/10 is the expected capture for 90% prediction intervals)
11. Repeat steps 6-10 for all possible set-aside intervals
12. Repeat steps 2-11 with MEboot method
13. Repeat steps 2-12 with 50% chronology confidence intervals and reconstruction predition intervals
14. Build 100 synthetic chronologies and repeat steps 1-13 for all
    1. Build 1,000,000 Ebisuzaki surrogates of the climate target in the interval of hronology-target overlap
       1. Select the surrogate with the correlation most similar to the correlation between the chronology and target
       2. This time series serves as a synthetic mean-value chronology – a basis for building synthetic ring-width indices
    2. Build 50 random length Ebisuzaki surrogates as ring-width indices
       1. A ring-width index is first modelled by 1000 Ebisuzaki surrogates
       2. The surrogates are rank-ordered by correlation to the synthetic mean-value chronology
       3. A synthetic ring-width index time series is randomly selected from the subset of 95th- to 99th-percentile surrogates

\*Note that prediction intervals produced in step 9a and b are identical regardless of bootstrapping methods.