**Title**

Estimating Uncertainties in Climate Reconstructions Derived from Crossdated Proxies

**Authors**

David C. Edge1,2, Bryan A. Black1, others…

**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**

Accurate estimates of uncertainty provide a gauge of signal-to-noise in paleoclimate reconstructions, which is critical for the detection and attribution environmental change. The reconstruction uncertainties derived from crossdated archives, including trees and bivalves, represent the envelope within which some percentage of the true climate values will be located in the pre-instrumental interval. The representativeness of a chronology value of the population in a given year is variable and this uncertainty must be included in the reconstruction uncertainty. To address this issue, we calculated pre-instrumental confidence intervals (uncertainties) based on chronologies from marine bivalve (*Panopea generosa*) ring widths, mountain hemlock (*Tsuga mertensiana)* maximum latewood density, and blue oak (*Quercus douglasii*) ring widths and their respective climate targets. For each of the three original chronologies, 100 synthetic chronologies were developed in which chronology sample depth, average correlation between series in a chronology (rbar), first-order autocorrelation (AR1), and correlation to climate target were varied. For all 303 chronologies (300 synthetic, 3 original), 90%- and 50% confidence intervals were calculated using all combinations of three different chronology bootstrapping techniques and two different reconstruction error measures. For each reconstruction, the percentage of target values captured by the confidence intervals was measured in a period independent of the reconstruction calibration and confidence interval calculations. The percentage of climate target values captured by these confidence intervals was also found to covary with the rbar, such that chronologies with lower rbar produced excessively wide confidence intervals. Chronology properties had no impact on the reliability of confidence intervals by maximum entropy bootstrapping (MEboot). The robustness of MEboot confidence intervals across a range of chronology properties and chronology-target relationships in these trials lends a reasonable expectation of accuracy for uncertainties calculated in this way.

**1. Introduction**

In dendroclimatology, annual reconstructions of past climate typically consist of “estimates” of a particular climate variable at a particular location, for example, mean 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, which we call the instrumental overlap interval (IOI). Using this relationship, the values of the tree-ring record are transformed into equivalent temperature units, called the reconstruction, providing estimates of temperature over the full extent of the tree-ring chronology (Fritts, 1976).

A reconstruction that fits a target climate variable closely in the IOI may be a very good proxy for that target. However, it may demonstrate similar character by chance, or the reconstruction model may be overfitted. Several methods have been developed to test the “skill” of a reconstruction, or how the reconstruction compares with a chance relationship. Many reconstructions are evaluated by splitting the IOI into independent calibration and verification intervals. Researchers utilize these intervals to measure the coefficient of determination (R2), Reduction of Error (RE), and the Coefficient of Efficiency (CE) to assess reconstruction skill (see Fritts, 1976; Briffa et al., 1988; Cook and Kairiukstis, 1990 and references therein). The skill/uncertainty can be ascribed to the domains of 1) the biological response to climate as captured by the proxy, 2) the measurement and manipulation of the proxy data, and 3) the transformation of the proxy data into units of climate (Fritts, 1976). In compliment to measures of reconstruction skill, reconstruction confidence intervals (hereafter, confidence intervals) often accompany dendroclimatic reconstructions, for example a band around the reconstruction estimated to envelope 90% of the true climate values (Cook and Kairiukstis, 1990; National Research Council, 2006).

Confidence intervals purport to show the quality of fit of a reconstruction. A recent reconstruction of drought severity translates the uncertainty domains described above into practical distinctions among error sources from detrending, chronology, and calibration (Esper et al., 2007). Detrending error may be the primary source of uncertainty in the low-frequency domain when the climate variable of interest contains sufficient low-frequency variability as extracting the climate signal from biological systems can be quite complex (Cook et al., 1995; Melvin & Briffa, 2008). The percent common signal (Cropper, 1982), Expressed Population Signal (EPS) and Subsample Signal Strength (SSS; Wigley et al., 1984; Buras, 2017) are all used to gauge chronology error. In the drought reconstruction mentioned above, a bootstrapping technique incorporates chronology error while the standard error of the estimate captures calibration error (Esper et al., 2007). There are several bootstrapping and regression error methods used in the dendroclimatic community, but we were motivated to investigate two particular bootstrapping methods and two calibration error methods based on their description in a recent streamflow reconstruction (Cook et al., 2013). Here we test several combinations of these methods using real and synthetic data to measure performance.

**2. Methods**

We calculated confidence intervals and tested their performance using 303 chronologies (3 original, 300 synthetic) and their corresponding target climate variables (hereafter “targets”). We use the term chronology to refer to the assemblage of crossdated indices, which is the primary form in which we utilize these data. We use the term mean-value (MV) chronology to discuss the time series of the annual chronology average, calculated by robust biweight mean (Mosteller and Tukey, 1977).

2.1 Chronologies

We calculated and tested a set of annually resolved chronologies, including two tree-ring chronologies, one bivalve sclerochronology, and 300 synthetic chronologies. All real chronologies were developed from crossdated, replicated, annually resolved datasets. For simplicity, we selected chronologies that produced reconstructions of a climate variable by simple linear regression and contained at least 60 years of IOI to ensure sufficient data for independent intervals for calibration, confidence interval calculation, and confidence interval testing. We selected original, and constructed synthetic, chronologies to represent a range of values for sample depth (i.e., the number of time series representing a single year), rbar, first-order autocorrelation (AR1), and correlation to target.

The Tree Nob chronology consists of crossdated, detrended growth increments of Pacific geoduck from coastal British Columbia, Canada, and extends continuously from 1725 to 2008 (Edge et al., 2021a). We used the chronology as produced for the original study, which were detrended by regional-curve standardization and then log transformed. The Tree Nob sea surface temperature (SST) reconstruction targets April-November SST at Langara Island. The length of the IOI is 62 years, from 1940-2001, with a minimum sample depth of 11 measurement series in this interval.

The Rock Springs Ranch chronology is based on the tree-ring widths of Blue Oak in San Benito County, California, USA (Stahle & Griffin, 2012). The chronology spans 1379 to 2003 and is highly sensitive to the local Jan-Feb rainfall (Stahle et al., 2013; Griffin and Anchukaitis, 2014). We downloaded the raw ring widths and detrended individual ring-width series with 2/3-length, 50% frequency cutoff cubic splines. We then compared the MV chronology to monthly precipitation data from the PRISM analysis 4k (Di Luzio et al., 2008). Based on significant correlation values, we selected a target of total (sum) Jan-Feb precipitation for reconstruction. The IOI extends from 1895-2003, 109 years, with an minimum sample depth of 30 measurement series in this interval.

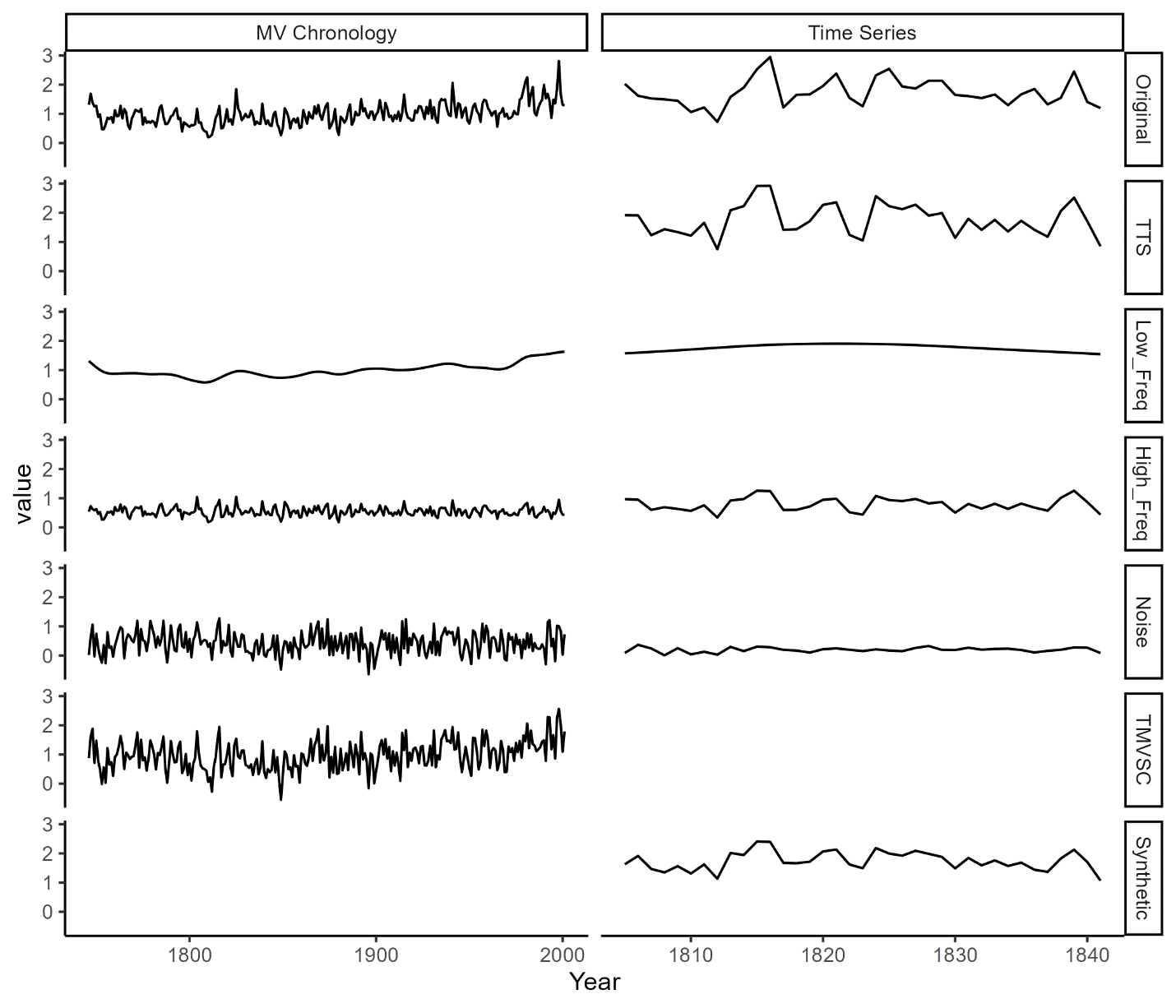
The Arrowsmith Mountain maximum latewood density (MXD) chronology is based on mountain hemlock on Vancouver Island, British Columbia, Canada and extends from 1629 to 1983 (Schweingruber 1988, Schweingruber et al., 1991; Briffa et al., 1992; Wiles et al., 1996; Briffa & Schweingruber, 2002). We combined the raw MXD time series assemblage by robust biweight mean to produce the MV chronology. In order to define a target season for reconstruction, we compared the MXD chronology to monthly air temperature values from the nearest grid box of the HADCRUT5 air temperature dataset (Morice et al., 2021) . Based on significant correlation values, we selected a target of average April-October temperature. Our target is similar to the Apr-Sep target for a reconstruction which utilized this and two additional MXD chronologies in the region (Wiles et al., 1996). The chronology-target overlap extends from 1857-1983, 127 years, with a minimum sample depth of 28 measurement time series in this interval.

2.2 Regression Assumptions

We regressed each chronology onto its target over the full IOI and tested the assumptions of regression (autocorrelation, normality, and homoscedasticity) for the residuals. We utilized the Durbin Watson test of autocorrelation in regression residuals using the lmtest package in r with a threshold of 0.05 based on Pan’s algorithm (Durbin and Watson, 1950; Farebrother, 1980; Zeileis and Hothorn 2002). We tested normality of the residuals with the Shapiro-Wilk test, using the stats package in the R programming language, with p-values less than 0.05 considered to reject the null of normality (R Core Team; 2022). We tested homoscedasticity with the Goldfeld-Quandt test provided by the lmtest package in R, with p-values less than 0.05 considered to reject the null of homoscedasticity (Goldfield and Quandt; 1965; Zeileis and Hothorn 2002).

2.3 Synthetic Chronologies

Figure 1 Synthetic chronology construction. The MV Chronology panel shows the process of building a TMVSC from the original MV chronology from the low/high frequency and noise components. One TMVSC is produced for each synthetic chronology. The Time Series panels shows the development of one synthetic chronology time series from the TMVSC segment, where the Original Time Series is the TMCSV in the interval covered by a particular time series. The TTS is broken into the low/high frequency components. Noise is added to the high frequency before the two components are recombined to form the synthetic time series.



In addition to the three real chronologies, we developed 100 synthetic chronologies for each chronology-target pair using the R programming language (Fig 1). Synthetic chronologies provided additional opportunities to test the confidence interval methods for each climate target and helped determine which chronology properties had the largest influence on the fidelity of the confidence intervals. We varied key properties of the synthetic chronologies in each iteration, including the correlation to target, first-order autocorrelation, sample depth, and rbar, defined as the mean value of all pairwise correlations among samples. We constrained the sample depths of the three sets of synthetic chronologies to either half, double, or identical to that of the original chronology. We constrained the AR1, rbar, and correlation to the climate target by the character of the original chronology but also added a random component.

We designed an algorithm to build synthetic chronologies based on an original chronology, its climate target, and four chronology properties (Fig 1). The synthetic chronology algorithm (hereafter, algorithm) built chronologies in two steps, first building a temporary mean value synthetic chronology (TMVSC), then building component time series based on the TMVSC. The component time series of the new synthetic chronology were built based on the TMVSC, but each of the final synthetic MV chronologies varied slightly from the TMVSC due to noise imparted in the construction of the individual time series. Four perturbation parameters, selected randomly when the synthetic chronology algorithm was initiated, adjusted the sample depth, rbar, AR1, and correlation to target. The algorithm randomly selected an rbar adjustment factor, which adjusted the random noise imparted to individual time series. The sample depth was set to change for 20% of synthetic chronologies, causing those chronologies to halve or double in sample depth with equal probability. An AR1 adjustment factor adjusted the variance of the white noise imparted to the TMVSC, changing the AR1 of the resulting chronology. The algorithm adjusted the correlation to the target, a random value ranging from 50% to 130% of the original chronology-target correlation.

The algorithm constructed the TMVSC from three components that included the low- and high-frequency components of the original MV chronology and white noise. The algorithm first deconstructed the original MV chronology into high- and low-frequency components. The MV chronology was fit with a 50% frequency cubic smoothing spline of a length randomly selected from a uniform distribution 1/10-length to the full length of the MV chronology. The residuals, obtained by dividing the original MV chronology by the fitted-spline values served as the high frequency series. Next, the algorithm generated a white noise series with a standard deviation identical to the high frequency series. The algorithm then combined the high frequency and noise series in a proportion randomly assigned, with each component representing 30-70% of the total variance. The AR1 adjustment factor then determined the scaling of this new series relative to the standard deviation of the original high frequency series, ranging from 10% to 300%. The algorithm then added this scaled series to the low frequency series to create the TMVSC. The algorithm repeated the process of building a TMVSC until the desired correlation to the target was reached.

The time series construction component of the algorithm proceeded similarly to the TMVSC algorithm with high/low frequency and noise components. First, a temporary time series (TTS) was defined as the sum of the TMVSC and white noise, with a variance set by the random rbar adjustment factor. The white noise imparted to each TTS was different, but its variance was set by the rbar adjustment factor, thus imparting a similar amount of noise to each. This TTS was deconstructed into high- and low-frequency components with a cubic smoothing spline as described above, with a spline of the same length as the TTS. The algorithm constructed the final time series from the sum of the spline, the residuals (high frequency series), and a new white noise series. The ratio of original high frequency to white noise was calculated as the square root of the rbar of the original indices.

2.4 Bootstrapping

We bootstrapped all chronologies, real and synthetic, using both traditional (Efron, 1979)) and maximum entropy bootstrapping methods (Vinod, 2006; Cook et al., 2013). We performed traditional bootstrapping by resampling one year at a time, from all possible values for that year of the chronology to produce a bootstrapped chronology with sample depths identical to the original chronology.

We performed MEboot on each individual timeseries of a chronology using the MEboot package in R (Table 1; Vinod and López-de-Lacalle, 2009). The MEboot algorithm creates surrogates based on the distances between ordered time series values such that all MEboot surrogate time series, when ordered by value, will have an identical sorting order to that of the original time series (Vinod, 2006). For instance, a time series 1.1, 1.3, 0.9, 1.7, 1.2 has consecutive distances of 0.2 (1.3-1.1), 0.4 (1.3-0.9), 0.8 (1.7-0.9), 0.5 (1.7-1.2). The 25% trimmed mean, the mean after removing 25% of values at each extreme, of these distances is 0.45. The trimmed mean of the consecutive distances is used to extrapolate the bounds of the ensemble range for the minimum and maximum values of the series, such that the minimum ensemble range for the lowest value in the time series is 0.9-0.45 and the maximum value at the maximum 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). The table shows that any surrogate created from these distributions with have the same sorting order as the original time series.

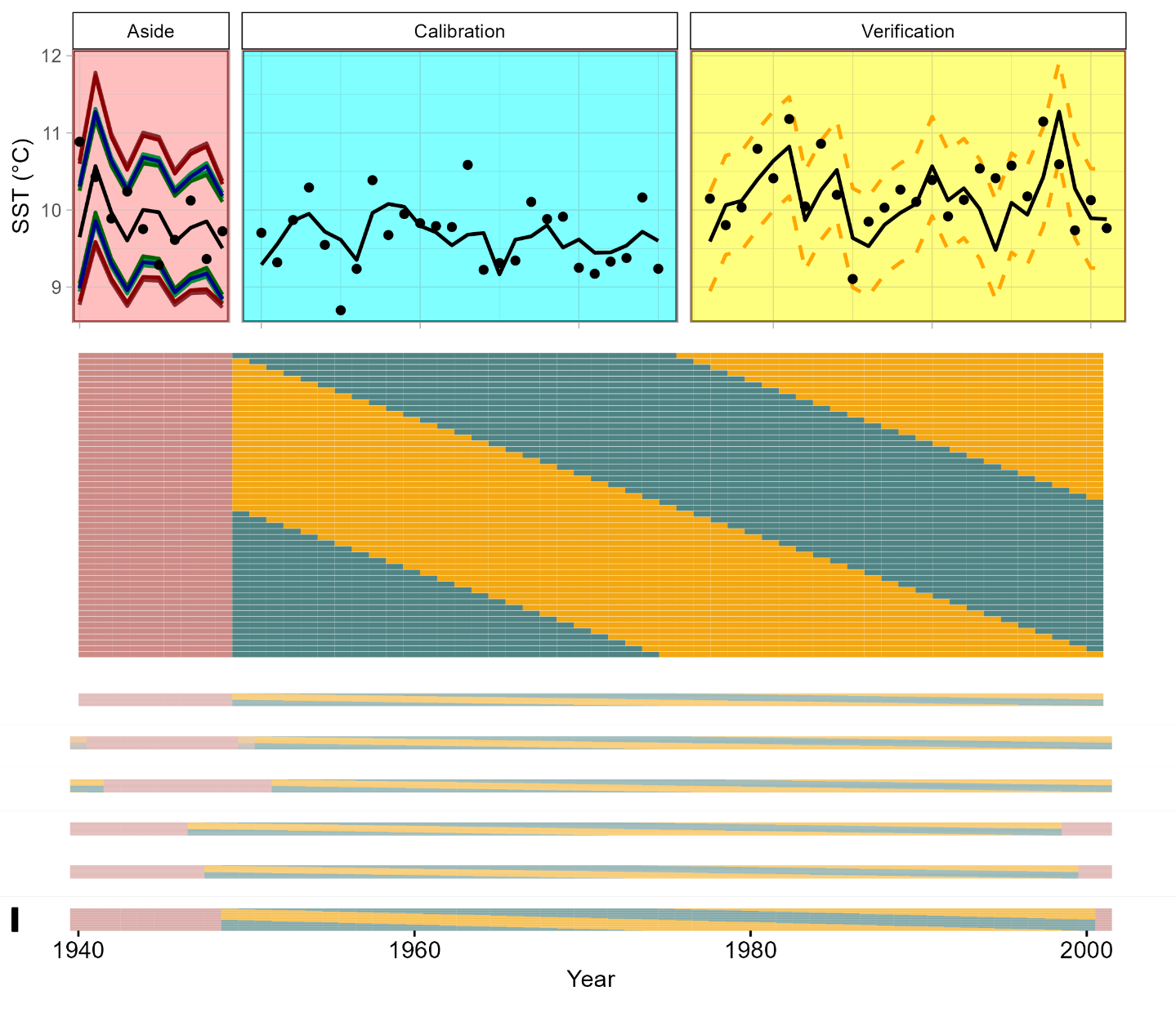
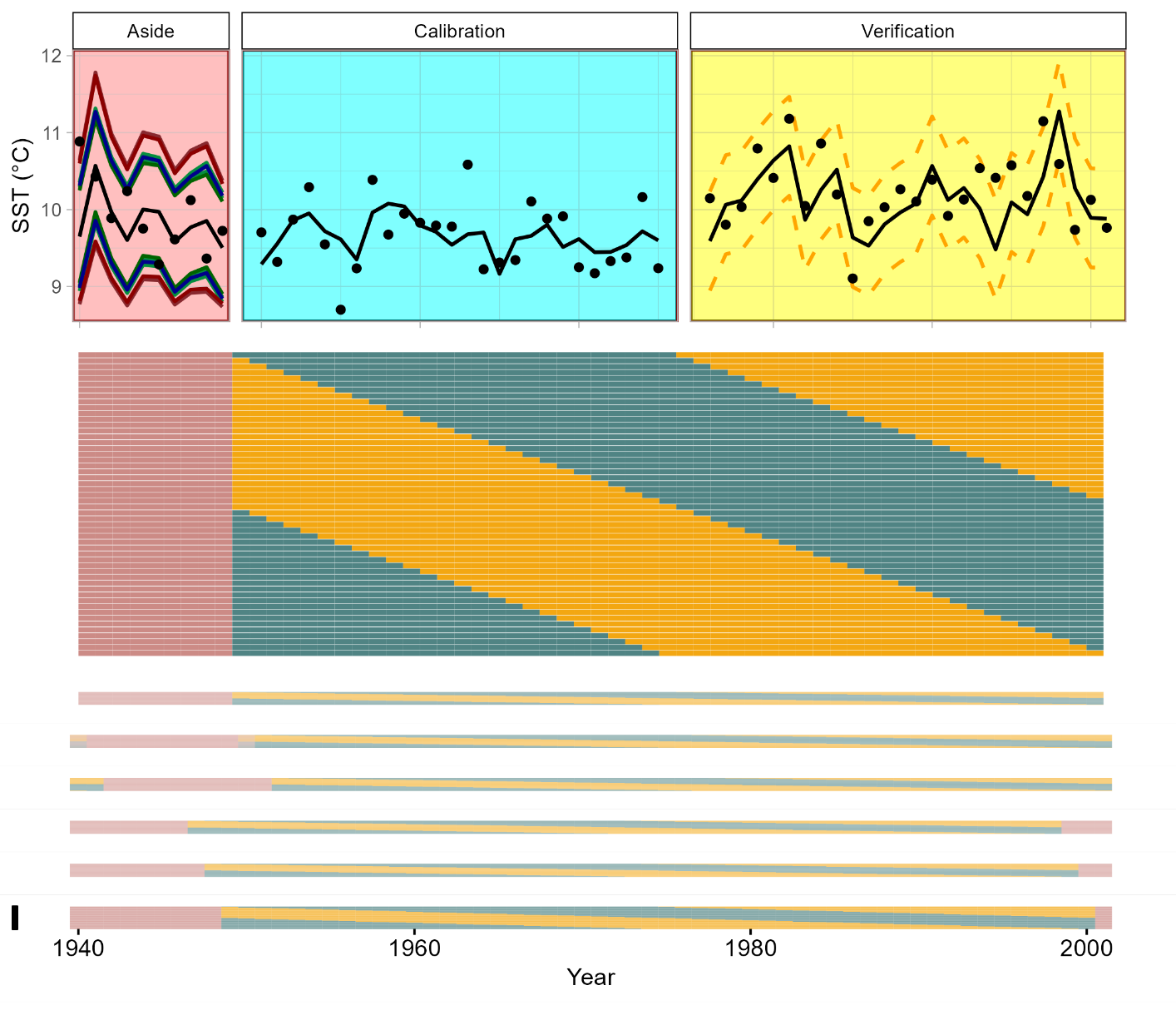
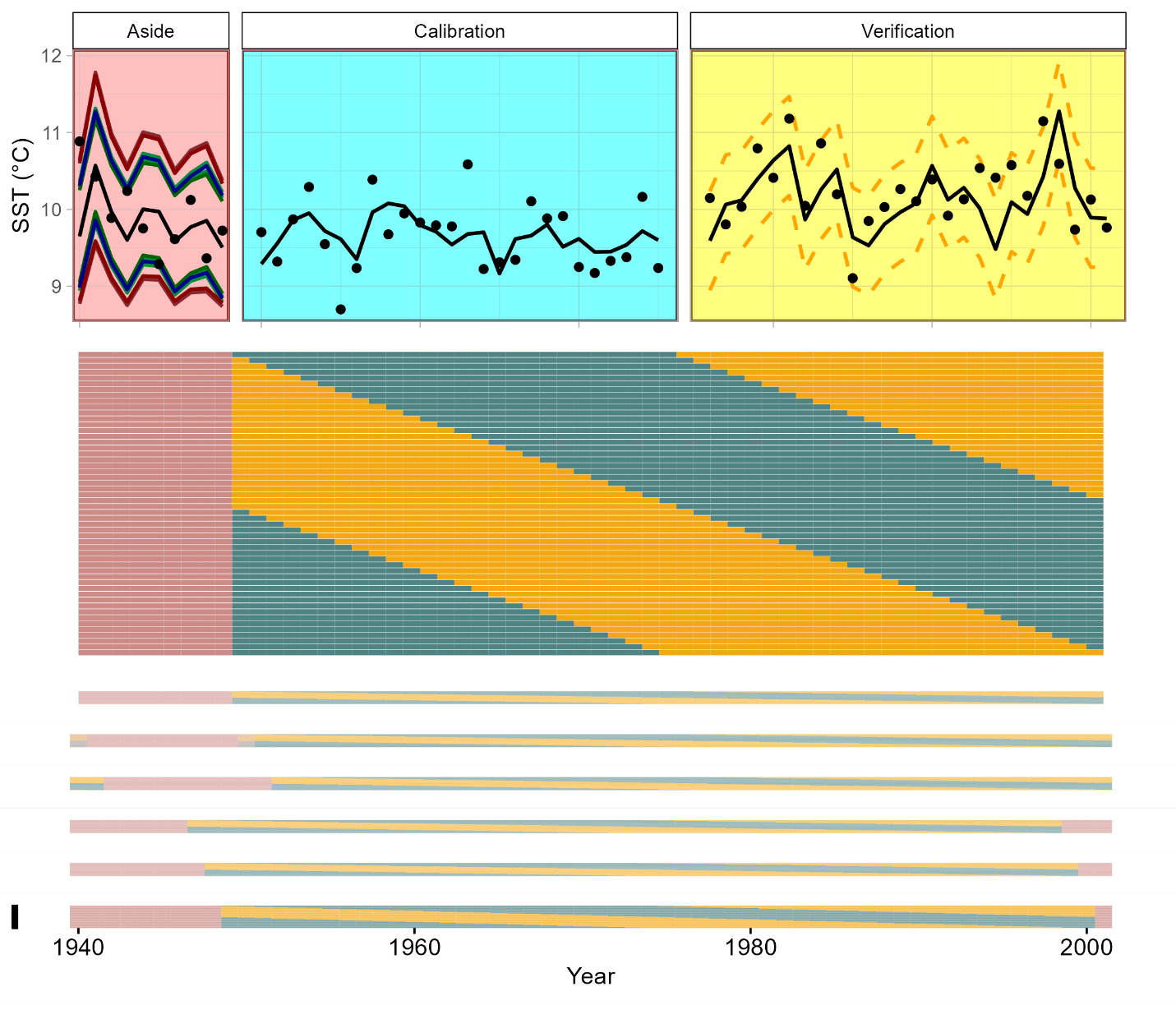
For each bootstrapping method, we produced 1000 sets of bootstrapped indices from each of the 303 chronologies. We developed a bootstrapped MV chronology for each set of indices based on the annual robust biweight mean. We then retained 5th and 95th, as well as 25th and 75th, percentile values at each year for 90th and 50th percentile chronology confidence intervals from the pool of 1000 MV bootstrapped chronologies for each bootstrapping method.

Table 1 MEboot (adapted from Vinod, 2006 Table 1). Maximum entropy bootstrapping development from original time series to uniform distributions. A random draw from the distrubutions produces an example of a bootstrapped series. “abs” indicates absolute difference.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Original Series** | **Sorting Order** | **Consecutive Distances** | **Sorted Series** | **Intermediate Values** | **Ensemble Distributions** | **Example Bootstrapped Series** |
|  |  |  | V**alues** | (0.9-**0.45**)=0.45 |  |  |
| 1.1 | 2 |  | 0.9 |  | (1.0, 1.15) | 1.1137492 |
|  |  | abs(1.1-1.3)=0.2 |  | ((0.9+1.1)/2)=1 |  |  |
| 1.3 | 4 |  | 1.1 |  | (1.25, 1.5) | 1.4529844 |
|  |  | abs(1.3-0.9)=0.4 |  | ((1.1+1.2)/2)=1.15 |  |  |
| 0.9 | 1 |  | 1.2 |  | (0.45, 1) | 0.4946209 |
|  |  | abs(0.9-1.7)=0.8 |  | ((1.2+1.3)/2)=1.25 |  |  |
| 1.7 | 5 |  | 1.3 |  | (1.5, 2.15) | 2.0572953 |
|  |  | abs(1.7-1.2)=0.5 |  | ((1.3+1.7)/2)=1.5 |  |  |
| 1.2 | 3 |  | 1.7 |  | (1.15, 1.25) | 1.2029572 |
|  |  |  |  | (1.7+**0.45**)=2.15 |  |  |
| Consecutive Distances: (0.2, 0.4, 0.8, 0.5), sorted: (0.2, 0.4, 0.5, 0.8), trimmed: (0.4, 0.5), trimmed mean: (**0.45**) | | | | | | |

2.5 Confidence Interval Calculation

Figure 2 Confidence Interval Testing Intervals and Methods. Panels a and b show the calculations for one set-aside interval. Panel c shows six of the sixty-two SAIs for this example. a) Fifty-two calibration and verification intervals were used to calculate the the confidence intervals for each SAI. **Calibration**: black line: chronology regressed onto target values, black points: target values. This regression defines the reconstruction for calibration and verification. **Verification**: orange dashed line: 90-percentile empirical error, black points: target values. VE measured as absolute difference between black line and black points. **Aside**: black line: reconstruction based on regression in full CVI. Colored lines: 90th-percentile empirical and theoretical confidence intervals, green: no bootstrapping. blue: MEboot, red: traditional bootstrapping. CIC is measured for each method as the percentage of target points falling within the confidence 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 confidence intervals in one set aside interval. c) SAI and CVI, the first and last 3 of 62 (for this example) total shown. The mean CIC across all SAI for each confidence interval method was used as the CICm value for the chronology.



**a**

**b**

**c**

Prior to calibration of the reconstruction, we first set aside a ten-year interval (named the set-aside interval (SAI)) of the for independent testing of the confidence intervals (Fig 2). We repeated the following calculations for each SAI in all possible continuous intervals of the IOI (Fig 2c). We used all possible continuous intervals of the remaining period as calibration and verification intervals (CVI), allowing for the calibration and verification intervals to wrap from the last year of the period back to the first year (Fig 2b). We then split the CVI in half and used the early portion, named the calibration interval, for calibrating the regression (Fig 2a). We also regressed the bootstrapped chronology confidence intervals onto the target in the calibration interval. We used the latter portion, named the verification interval, to calculate the verification error (VE), which is the absolute value of the difference between the reconstructed and target value over the verification interval for each year.

For each CVI, we used the set of VE values to calculate the empirical (non-parametric) 50th and 90th percentile verification intervals (VEe50, VEe90) as the ascending order 50th and 90th percentile values from the VE set. To calculate the theoretical (parametric) VE values, we assumed that the differences are normally distributed, and thus that the VE approximate a half-normal distribution with a lower bound of 0. We calculated theoretical 50th and 90th percentile errors (VEt50, VEt90) after adjusting these intervals for θ of the sample, analogous to the sample standard deviation, using the qhalfnorm function in the R package fdrtool (Strimmer, 2008). The VEt50 is typically indistinguishable from the VEe50, so we do not make a distinction in our tests.

We combined the VEe and VEt values with the regression coefficients from both the MV chronology and bootstrapped chronology confidence intervals to calculate the reconstruction confidence intervals in the SAI. We first built a new reconstruction for the SAI by regressing the MV chronology onto the target in the full CVI. We calculated non-bootstrapped confidence intervals by adding the VEe and VEt values to this new reconstruction. We calculated the bootstrapped reconstruction confidence intervals by the same process, substituting the regression coefficients from the bootstrapped chronology confidence intervals. The possible permutations of these bootstrapping (None, MEboot, Traditional) and VE (VEe50, VEt50, VEt90, VEe90) options produce 12 sets of confidence intervals; but with theoretical and empirical VE identical for the 50% intended capture, we produced only 9 unique confidence intervals per SAI.

2.6 Confidence Interval Testing

For each of the 9 confidence intervals per SAI, we calculated the confidence interval capture (CIC), the percentage of target values that fall within a set of confidence intervals in a given period. We calculated CIC in each SAI by dividing the total number of target values captured by the confidence intervals by 10 (the SAI length). For example, if 9 of the 10 SAI climate target values fall within the confidence intervals, 90% of the target values were “captured”, in line with the 90% intended capture (IC) rate. This test was repeated for all possible continuous 10-year intervals, such that a chronology-target pair with a 62-year IOI contains 62 possible, overlapping testing intervals (Fig 2c). We then calculated the mean CIC from all trials (CICm) for a given chronology and confidence-interval method.

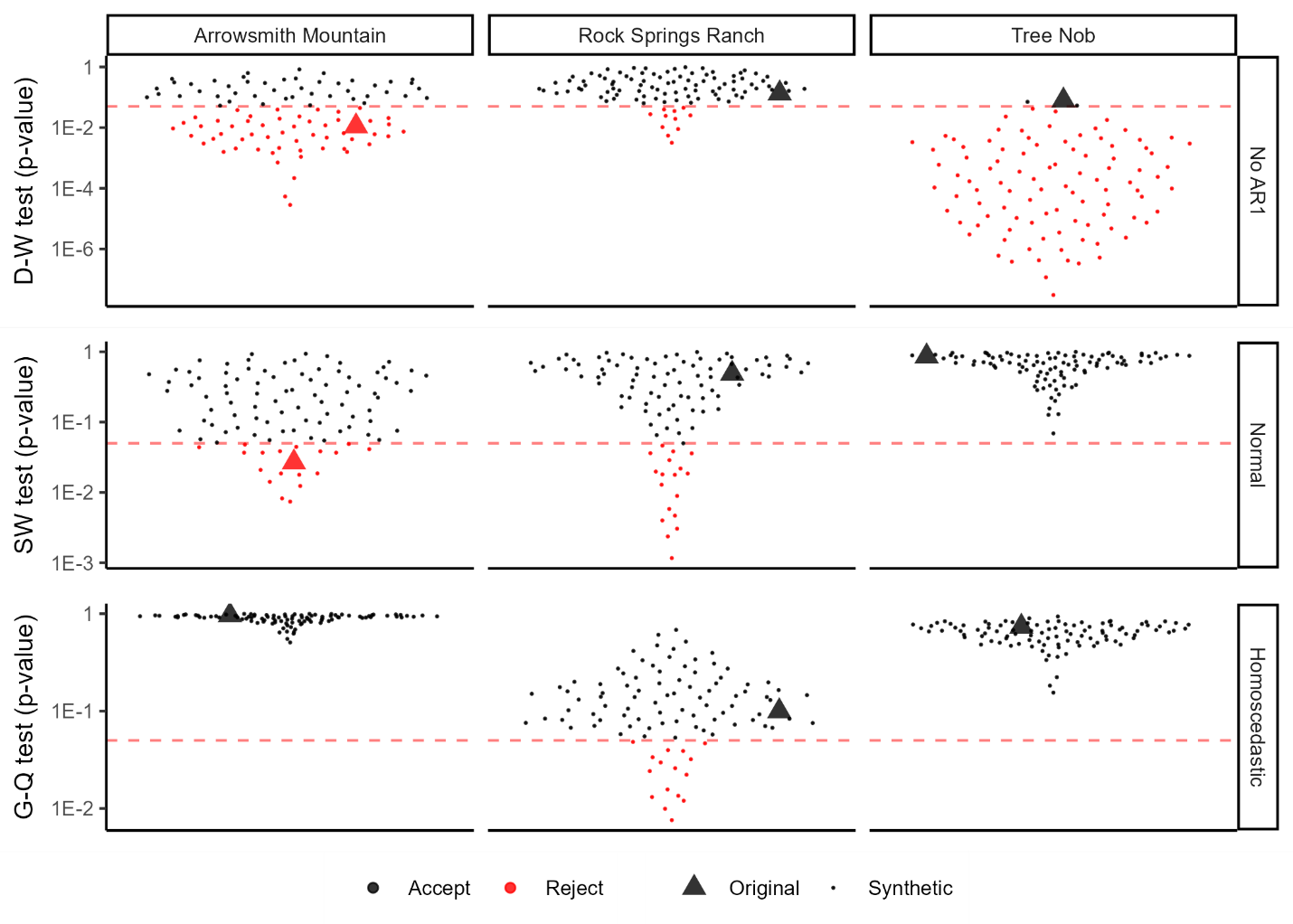
We then compared CICm for the four chronology properties (correlation to climate target, first-order autocorrelation, sample depth, and rbar) for all 303 chronologies. Strong correlations between chronology properties and CIC for a given confidence interval method suggest a bias of the method that will produce confidence intervals too wide or narrow based on idiosyncrasies of individual chronologies.

**3. Results**

3.1 Regression Assumptions

Of the 300 synthetic and three original chronologies, 200 synthetic and one original rejected the null hypothesis of one or more of the regression assumptions tested (Fig 3, red points). The regression residuals from the Arrowsmith Mountain chronology and 54 of 100 of the associated synthetic chronologies rejected the null of no autocorrelation. 19 of the Arrowsmith Mountain chronologies had regression residuals that were not normally distributed. Among the Rock Springs Ranch synthetic chronologies, 11, 16, and 18 rejected the null hypotheses of No AR1, homoscedasticity, and normality respectively. At Tree Nob, only the original and two synthetic chronologies failed to reject the null hypothesis of no AR1.

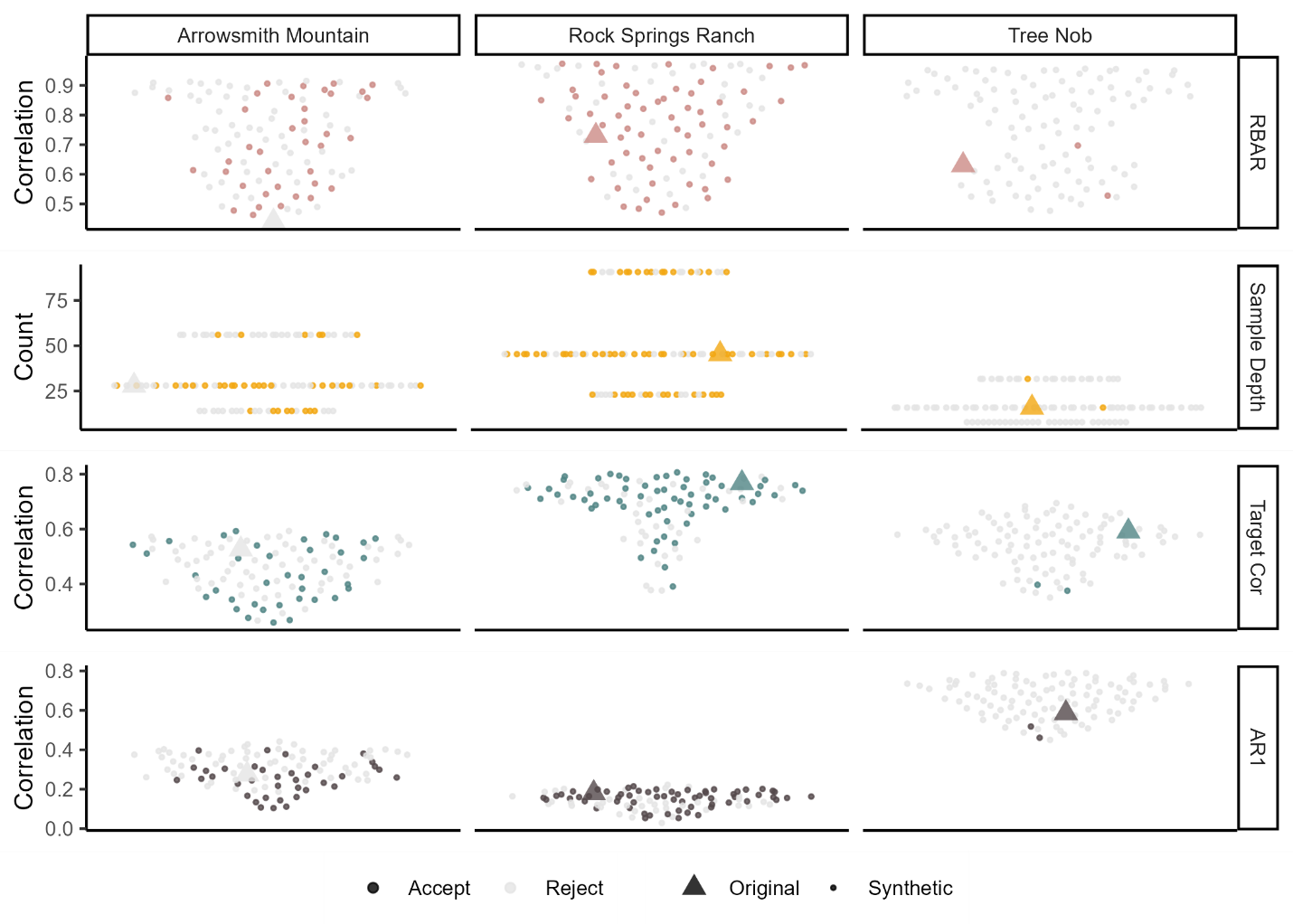
Figure 3 Regression Assumptions Testing. **No AR1:** results of the Durban Watson test, null hypothesis of no AR1 in residuals rejected at p<0.05. **Normality:** normal distribution of residuals tested by the Shapiro-Wilk test, null hypothesis of normal residuals rejected at p<0.05. **Homoscedasticity**: trend in variance of residuals as tested by the Goldfeld-Quandt test, null hypothesis of homoscedastic residuals rejected at p<0.05. Red points indicate chronologies rejecting a given null hypothesis by the standards listed above. Testing thresholds are demarked by red, dashed lines.



3.2 Synthetic Chronologies

The three sets of synthetic chronologies varied considerably in the range and distribution of the four characteristics parameterized (Fig 4), but 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 rbar distributions ranged from less that 0.5 to greater than 0.9 for all three synthetic chronology sets. The AR1 distributions were more constricted by the AR1 of the original chronologies with some Tree Nob synthetic chronologies nearing an AR1 of 0.8, while no Rock Springs Ranch synthetic chronologies rose above 0.3. The upper bounds of the target correlation distributions were constrained by that of the original chronology, such that the original chronologies sat near the tops of the three distributions. The chronologies which rejected the null hypothesis of one of the three regression assumptions were not generally of distinct distributions for any of the chronology properties. However, the three chronologies remaining at Tree Nob sat near the lower portion of the AR1 and rbar distributions.

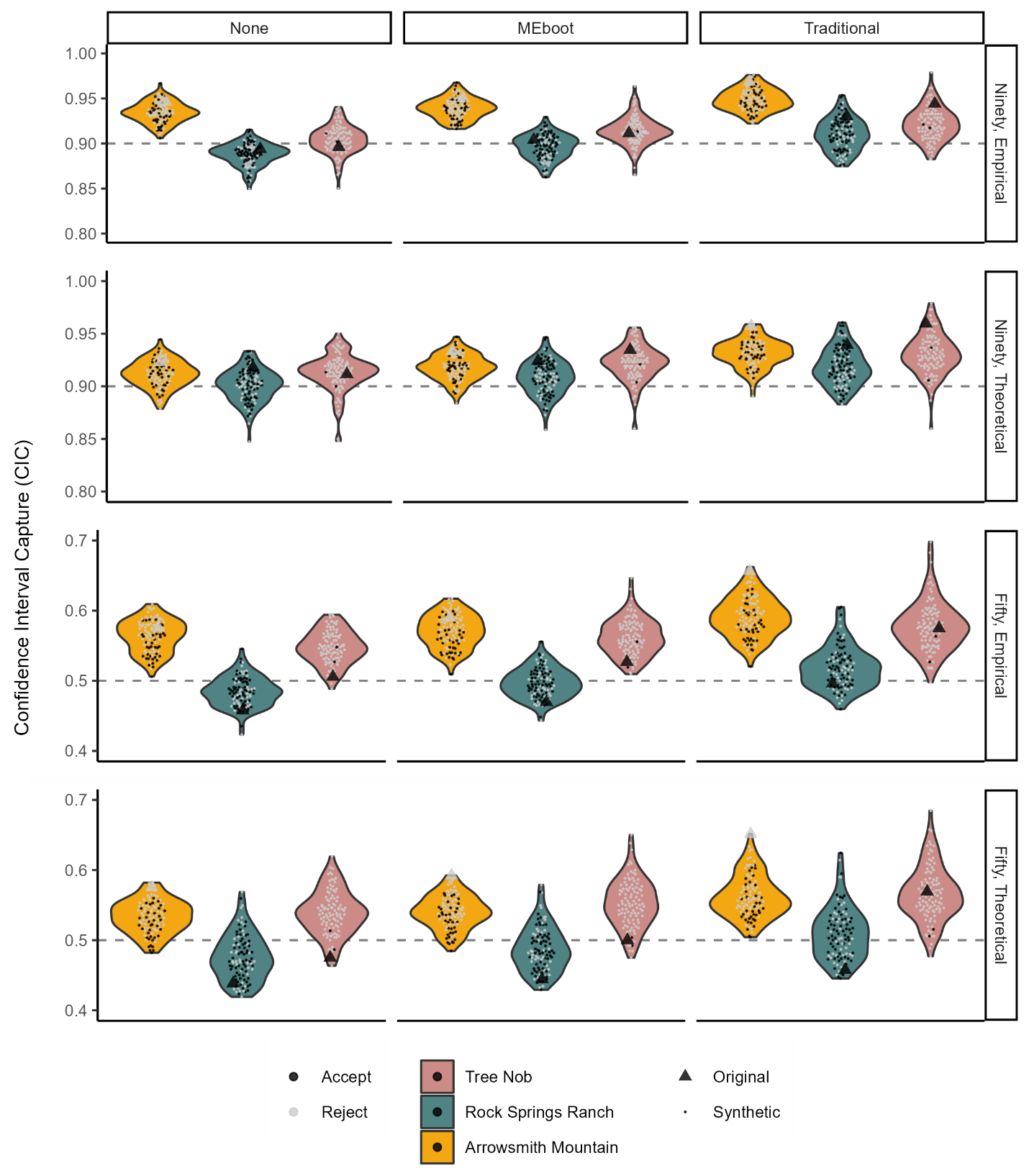
Figure 4 Chronology Properties. 101 points represent the 100 synthetic and one original chronology for each cluster of points which are jittered about the x-axis for visibility. Points shown in grey represent chronologies which rejected the null hypothesis of one of the three regression assumptions, see Fig 3. **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 overlap with the target. Note that all sample depth values for a single chronology site fall on one of three values, the original chronology sample depth, or half or double that value. Small circles represent values for synthetic chronologies, large triangles represent values of the original chronologies.



3.3 Bootstrapping

The width of traditionally bootstrapped intervals was consistently greater than the MEboot intervals at both 50- and 90-percetile (Fig 5). 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.

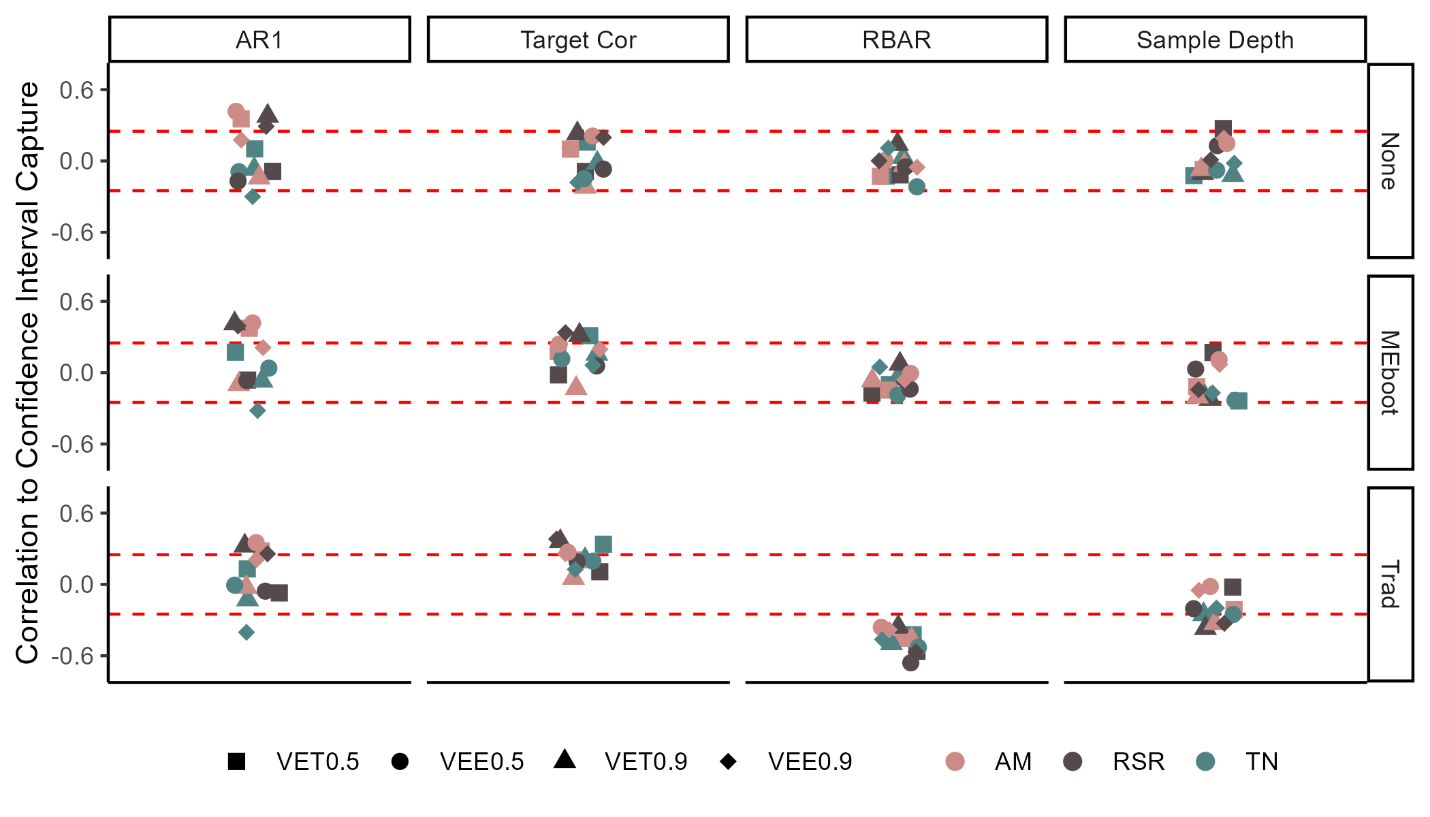
Figure 5 Confidence Interval Performance. Each point represents the CICm of a chronology. Each color represents by a real chronology with associated synthetic chronologies, with nine violin plots corresponding to the nine confidence interval methods. **Ninety vs. Fifty**: intended capture (IC). **Theoretical vs. Empirical**: method of VE calculation. **None vs. MEboot vs. Traditional**: Method of chronology bootstrapping. Dashed lines indicate the intended percent of target values to fall within confidence intervals (IC).



3.4 Confidence Interval Testing

The various confidence interval methods showed variable success in capturing the IC (e.g., 0.9 for 90th percentile confidence intervals; see dashed lines, Fig 5). Of the 12 methods implemented, the Non-bootstrapped VEE 90th percentile confidence intervals produced an average CICm closest to the IC at 90.6%, with and interquartile range of 3.7%. If only the chronologies that failed to reject the null hypotheses of all regression assumptions are considered (ignoring the grey points in Fig 5), the Meboot VET 50th percentile confidence interval performed best, with an average CICm of 49.9%, IQR = 4.6%. The CICm values for Arrowsmith Mountain chronologies were generally higher than the IC and higher than the other two sites across all methods. The difference between the average CICm values across sites is reduced with VET as opposed to VEE, while the range of CICm values for a given site is reduced with VEE relative to VET. The 90th percentile theoretical errors (VEt90) produced confidence intervals with CICm slightly above the IC, but the CICm distributions were consistent across the three sites. The combination of MEboot, empirical reconstruction errors, and 90-percentile confidence intervals was most effective at minimizing CICm - IC for Tree Nob and Rock Springs Ranch chronologies and produced the tightest grouping of CICm (Fig. 5). CICm was successively wider from no bootstrapping, to MEboot, to traditional bootstrapping, with traditional bootstrapping CICm generally greater than IC. Individual CICm distributions were generally wider for traditional bootstrapping compared with no bootstrapping and MEboot.

Figure 6 Correlations between Chronology Properties and Confidence Interval Capture. Individual points represent a correlation between the rate of capture for a confidence 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 relationships of interest.



The CICm for traditionally bootstrapped chronologies was negatively correlated to the rbar of all chronologies for all methods (Fig 6). the correlation to target and sample depth values were positively and negatively correlated to CIC, respectively, for some methods and chronology properties. The CICm, based on methods without bootstrapping or by MEboot, show little or no consistent relationship to any chronology property.

**4. Discussion**

Although the chronology simulation algorithm produced chronologies with a wide range of properties and affected confidence interval width, the methods of bootstrapping and error measurement had the most significant impact on the reliability of the confidence intervals. Traditional bootstrapping produces excessively wide confidence 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 (Fig 6). The MEboot method assures greater fidelity to the properties of the original series (Cook et al. 2013). MEboot preserves autocorrelation structure while reliably producing median, as well as minimum and maximum, values of the dataset. This is distinct from traditional bootstrapping, which shows a bias toward extreme values when selecting the 5- and 95-percentile representatives (Cook et al. 2013).

We designed the chronology simulation algorithm to vary important chronology properties in order to measure their impacts on the reliability of the confidence intervals. The resulting synthetic chronologies are based on the original chronologies and are not independent. The CIC distributions of each site, for a given method, are distinctive. However, the trends across methods are also noteworthy (Fig. 5). The intended function of the simulations is demonstrated in Figure 6 wherein the variability of an important chronology property, rbar, produces biased confidence 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 direction of the bias is also consistent with a conceptual understanding of the method wherein the 90th-percentile values in the bootstrapped ensemble necessarily overrepresent chronology indices with more extreme values. In a chronology with low rbar, the distribution of values in a typical year is, by definition, wider, exacerbating the problem inherent to the method. Conversely, a chronology with a high rbar, in which index values of a typical year are tightly grouped, traditional bootstrapping produces 90th-percentile intervals much closer to the central tendency of all values.

Some of the CICm distributions which best relate to their respective IC are not bootstrapped. When considering sources of reconstruction uncertainty, we made a distinction between chronology error and calibration error. We intended to capture chronology error with bootstrapping and calibration error with either VEE or VET. The methods for calculating these errors were chosen based on the principle that calibration ought to be independent from the measurement of skill, such as in the calculation of the CE metric. Our testing method suggests that the VEE and VET alone produce sufficiently wide confidence intervals without the benefit of bootstrapping. However, our testing method does not account for substantial changes in sample depth and makeup within a chronology. It is notable, therefore, that when the Arrowsmith Mountain chronology, which contains no change in sample makeup for the entire IOI, is removed from the analysis, MEboot with VEE produces the best outcomes (average CICm = 89.9%, 49.7%; IQR = 1.8%, 2.7%).

Constructing confidence 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. For this purpose, the SAIs could either comprise years chosen at random to incorporate the greatest diversity of possible values or 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 and verification intervals are continuous with the exception of up to two breaks, and SAIs are continuous with the exception of up to one break (Fig 2b,c). 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 (Rousseeuw & Leroy, 2005). This also allows for testing a calibration interval in a manner analogous to the final reconstruction, in which the full IOI is utilized, 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 reconstruction and confidence intervals in the SAI, which reduces the validity as a test of confidence interval performance when the associated targets also have high AR1. We measured the importance of this phenomenon by comparing confidence interval performance to MV chronology AR1 (Figure 6) and found no consistent relationship, suggesting that the independence of the SAI is not significantly impacted by autocorrelation. At Tree Nob, the site with the highest AR1 values, CIC and AR1 are negatively correlated, but this relationship is not consistent across confidence interval methods.

We have provided evidence for the unreliability of confidence intervals constructed from traditional bootstrapping and some evidence to support MEboot. However, the simple linear regression method we used to develop the 303 reconstructions utilized is an uncommon approach in the development of climate reconstructions. Multiple predictors and distributed lag regression models as well as principal components are commonly employed (National Research Council, 2006). Nevertheless, we have utilized a diversity of real chronology-target pairs with a range of properties alongside synthetic chronologies with properties of still greater diversity. The approach we have employed can easily be extended to more complex reconstruction methods. Although the robustness of the methods that performed well in these simulations may not hold for all reconstruction methods, the simulations suggest that some methods, including traditional bootstrapping, should not be used. Traditional bootstrapping contains assumptions inappropriate for pairing with crossdated archives. This method consistently produced confidence intervals wider than intended in our trials (Fig 5). Furthermore, the consistent negative correlation between CIC and rbar for traditionally bootstrapped chronologies (Fig 6) suggests consistent, predictable bias inherent to this method.

We did not exhaustively test all possible bootstrapping methods. Resampling complete individual time series would also maintain the autocorrelation structures of the 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.

5. Conclusion

A reliable estimated error range inherent to any paleoclimate reconstruction is essential to the end-user. Confidence intervals provide at-a-glance information of the relative signal-to-noise ratio of the reconstruction, conveying both what is known and unknown. Confidence intervals are 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 confidence intervals will impact both the precision and accuracy of the results, and the best methods will likely conform to the method’s appropriateness in relation to the dataset. We find that 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. While future work is needed to extend this approach to reconstructions with multiple predictors and other complex reconstruction methods, our results suggest that MEboot paired with VEE or VET provide the most robust confidence intervals for crossdated paleoclimate proxies. Furthermore, we have introduced a new approach to synthetic chronology construction will allow for the testing of many methods with similar simulation experiments.

**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., Jones, P., Pilcher, J., & Hughes, M. (1988). Reconstructing summer temperatures in northern Fennoscandinavia back to AD 1700 using tree-ring data from Scots pine. Arctic and Alpine Research, 20(4), 385-394.

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.

Briffa, K.R.; Schweingruber, F.H. (2002). NOAA/WDS Paleoclimatology - Briffa - Arrowsmith Mountain - TSME - ITRDB CANA113. NOAA National Centers for Environmental Information. https://doi.org/10.25921/r7fr-m938. Accessed 21 October 2022.

Buras, A. (2017). A comment on the expressed population signal. Dendrochronologia, 44, 130-132.

Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A., & Funkhouser, G. (1995). The'segment length curse'in long tree-ring chronology development for palaeoclimatic studies. The Holocene, 5(2), 229-237.

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.

Di Luzio, M., Johnson, G. L., Daly, C., Eischeid, J. K., & Arnold, J. G. (2008). Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. Journal of Applied Meteorology and Climatology, 47(2), 475-497.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. Biometrika, 37(3/4), 409-428.

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.

Edge, D.; Reynolds, D.J.; Wanamaker, A.D.; Griffin, D.; Bureau, D.; Outridge, C.; Stevick, B.C.; Weng, R.; Black, B.A. (2021). NOAA/WDS Paleoclimatology - British Columbia 2,900 Year Bivalve Sclerochronology and Sea Surface Temperature Reconstruction. NOAA National Centers for Environmental Information. https://doi.org/10.25921/ms94-wy29. Accessed 21 October 2022.

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

Farebrother, R. W. (1980). Algorithm AS 153: Pan's procedure for the tail probabilities of the Durbin-Watson statistic. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(2), 224-227.

Fritts, H. (1976). Tree rings and climate. Academic Press. New York.

Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. Journal of the American statistical Association, 60(310), 539-547.

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

Melvin, T. M., & Briffa, K. R. (2008). A “signal-free” approach to dendroclimatic standardisation. Dendrochronologia, 26(2), 71-86.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... & Simpson, I. R. (2021). An updated assessment of near‐surface temperature change from 1850: the HadCRUT5 data set. Journal of Geophysical Research: Atmospheres, 126(3), e2019JD032361.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression. A second course in statistics. Addison-Wesley series in behavioral science: quantitative methods.

National Research Council. (2006). Surface temperature reconstructions for the last 2,000 years. National Academies Press.

Rousseeuw, P. J., & Leroy, A. M. (2005). Robust regression and outlier detection. John wiley & sons.

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.W.; Griffin, R.D. (2012). NOAA/WDS Paleoclimatology - Stahle - Rock Springs Ranch - QUDG - ITRDB CA646. NOAA National Centers for Environmental Information. https://doi.org/10.25921/mz3p-1721. Accessed 21 October 2022.

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.

Strimmer, K. (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics, 24(12), 1461-1462.

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.

Wigley, T. M., Briffa, K. R., & Jones, P. D. (1984). On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. Journal of Applied Meteorology and Climatology, 23(2), 201-213.

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.