Review of methods and Data – August 25, 2022

Based on the Tree Nob RCS chronology and 10 synthetic chronologies reconstructing Langara SST

**Introduction**

In the dendrochronology community, annual reconstructions of past climate typically consist of “estimates” of a particular climate variable, 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, called the calibration interval. The values of the tree-ring record are transformed into equivalent temperature units, called the reconstructed values, 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 estimates 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.

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. The average difference between the individual reconstruction and climate values in the interval of overlap 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. 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.

**Methods**

Chronologies

Prediction intervals were calculated and tested using xxx chronologies 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.

The Tree Nob geoduck chronology extends continuously from 1725 to 2008, with an average sample depth of xxx. The published reconstruction targets April-November sea surface temperature at Langara Island. The overlap of these records covers 1940-2001, 62 years. 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.

Synthetic Chronologies

In addition to the real chronologies tested, we developed 50 synthetic chronologies for each chronology-target pair. Synthetic chronologies provide additional opportunities to test the prediction interval methods for each climate target. These synthetic chronologies are independent of the real chronologies and therefore also independent of any idiosyncrasies of chronology-target relationships.

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 within each year of the chronology to produce a bootstrapped chronology with sample depths identical to the original chronology. 1000 bootstrapped mean-value chronologies were developed for each bootstrapping method from the robust biweight mean of 1000 individual bootstrapped ring-width index chronologies. 5th and 95th percentile values at each year were retained as chronology confidence intervals from the pool of 1000 mean-value bootstrapped chronologies for each bootstrapping method.

Reconstructions and Verification Error

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. 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 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.

Prediction Interval Testing

All prediction intervals were calculated and tested in the independent 10-year SAI. 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.

**Results**

Langara Sea Surface Temperature

Prediction interval testing results for the reconstruction based on the Tree Nob chronology range from 0.2% by theoretical VE, MEboot regression error, and 90% IC to 6.5% by theoretical VE, traditional bootstrapping, and 90% IC. The average of the 10 synthetic chronologies gave a range of 0.7% for empirical VE, MEboot regression error, and 90% IC to 9.9% for traditionally bootstrapped VE and 50% IC. Empirical VE produced lower average differences between IC and PIC relative to theoretical VE in most trials. MEboot regression error also tended to produce lower average differences between IC and PIC when compared with traditional bootstrapping. The prediction intervals with IC of 90% generally showed greater fidelity that the 50% IC prediction intervals, with the exception of those calculated without regression error.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Table 1. Tree Nob Geoduck - Langara Island Sea Surface Temperature** | | | | | | | | | |  |  |
|  | No Bootstrapping | | | | MEboot | | | | Traditional Bootstrapping | | | |
|  | VEt | | VEe | | VEt | | VEe | | VEt | | VEe | |
| sim50 | 50.8% | 0.8% | 50.8% | 0.8% | 53.6% | 3.6% | 53.6% | 3.6% | 59.9% | 9.9% | 59.9% | 9.9% |
| real50 | 48.5% | 1.5% | 47.8% | 2.2% | 49.2% | 0.8% | 48.4% | 1.6% | 55.0% | 5.0% | 52.4% | 2.4% |
| sim90 | 86.4% | 3.6% | 88.1% | 1.9% | 88.2% | 1.8% | 89.3% | 0.7% | 94.3% | 4.3% | 95.6% | 5.6% |
| real90 | 86.8% | 3.2% | 88.5% | 1.5% | 89.8% | 0.2% | 91.1% | 1.1% | 96.5% | 6.5% | 96.3% | 6.3% |
|  | **2.3%** | | **1.6%** | | **1.6%** | | **1.8%** | | **6.4%** | | **6.1%** | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Table 2. Rogen Blue Light Intensity - JJA Air Temperature** | | | | | | | |  |  |  |  |
|  | No Bootstrapping | | | | MEboot | | | | Traditional Bootstrapping | | | |
|  | VEt | | VEe | | VEt | | VEe | | VEt | | VEe | |
| sim50 |  |  |  |  |  |  |  |  |  |  |  |  |
| real50 | 51.9% | 1.9% | 51.9% | 1.9% | 52.9% | 2.9% | 52.9% | 2.9% | 63.1% | 13.1% | 63.1% | 13.1% |
| sim90 |  |  |  |  |  |  |  |  |  |  |  |  |
| real90 | 87.6% | 2.4% | 90.4% | 0.4% | 93.5% | 3.5% | 94.2% | 4.2% | 88.8% | 1.2% | 91.7% | 1.7% |
|  | **2.2%** | | **1.2%** | | **3.2%** | | **3.6%** | | **7.2%** | | **7.4%** | |

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.