pcrecon: An R Package for Nested Principal Component Regression Climate Reconstructions

Laura Smith1\*, Nicholas Nagle1, Stockton Maxwell2, Daniel Hocking^1

*1 University of Tennessee; Knoxville, Tennessee, USA*

*2 Radford University; Radford, Virginia, USA*

*3 NOAA Greater Atlantic Regional Fisheries Office, Gloucester, MA, USA*

*\* Corresponding Author Email:* [*lsmit224@vols.utk.edu*](mailto:lsmit224@vols.utk.edu)

# Abstract

We present pcrecon, a new software package in R for the development of nested principal component regression climate reconstructions from tree ring data in R. This package duplicates and extends the well known Fortran program PCreg. The use of the R statistical programming environment for this application can improve replicability and transparency of climate reconstructions, which are both crucial to the goals of the paleoclimate community.

# Introduction

Climate Reconstructions and Principal Component Regression

Climate reconstructions from tree-ring records are based on the statistical relationship between a target climate variable one or more nearby tree ring chronologies as estimated using linear regression. Model coefficients are calculated for the period of time over which the tree ring data and climate data overlap, usually around 80-100 years, then predictions are made for years that predate the instrumental climate based on the tree rings. Tree-ring records from multiple species and locations within a region are often used to fit the model, particularly for reconstructions in temperate climate regions where the relationships between tree growth and climate is complex (Cook et al. 1999). This can help to parse the local, autocorrelated ecological noise related to forest succession, competition, and disturbance from desired regional climate signal. A common problem with multiple regression methods is that many independent variables can produce models which are overfit or that suffer from issues related to multicollinearity. Principal Component Analysis (PCA) and Principal Component Regression (PCR) are widely used to address these issues by reducing the dimensionality of tree-ring predictor variables(Jacoby and Cook 1981; Briffa et al. 1983; Cook et al. 1999; Williams et al. 2020). PCR has been of particular importance in reconstructing climate of the eastern U.S., where climate-growth relationships are more complex due to the temperate climate and highly disturbed forests (Harley et al. 2017; Maxwell et al. 2011, 2017)

A limitation of PCA is that it requires that data vectors to be of equal length. Because tree-ring records vary significantly in their length,a nesting technique is applied. First, the period of time over which all chronologies overlap is determined, this is called the ‘common period’. PCA eigenvectors and linear regression estimates are calculated for that time period only. The shortest chronology is then dropped from the analysis, and the next common period is determined, and so on, such that a new regression model and predictions are calculated for each common period. These nests are then spliced together to form a long-running reconstruction which has an estimate value for each year that there are tree ring data available.

Data are held out from the model calibration to validate the estimates. Common validation statistics are Reduction of Error (RE) and Coefficient of Efficiency (CE) and compares the model mean square error (MSE) is to the calibration mean and provides a metric of how predictive the model compared to the mean. A value greater than 0 indicates that the model is better than the calibration mean at predicting values. is similar, but compares the MSE to the held out validation period mean, making it a stricter test (*Surface Temperature Reconstructions for the Last 2,000 Years* 2006)

Many researchers rely on specialized FORTRAN based software program called PCreg (<https://www.ldeo.columbia.edu/tree-ring-laboratory/resources/software>) to perform these procedures. This software can be used to filter tree ring data by location, summarize and prewhiten climate data, determine common periods, make estimates, and calculate validation statistics. We introduce an R package, pcrecon, which can perform many of the same functions and numerically replicate results from PCreg. We propose that the R environment improves replicability and transparency in addition to streamlining the workflow of performing these reconstructions.

## Software for Analysis of Paleoclimate Proxy Data

Due to the labor-intensive nature of data collection and the specialized statistical methods used for the analysis of tree-ring data, the paleoclimate community encourages the open sharing of data and software. This is evidenced by the popularity and frequent use of resources such as the International Tree Ring Data Bank (ITRDB, NOAA), the Dendrochronology Program Library (Richard Holmes/University of Arizona), the Lamont-Doherty dendrochronology software repository (<https://www.ldeo.columbia.edu/tree-ring-laboratory/resources/software>), and R packages such as dplR (Bunn 2008) and treeclim (Zang and Biondi 2015).

Incredible efforts have been made over the last 40 years in the development of statistical techniques for tree-ring data analysis and associated software which allow users to standardize tree-ring series, develop chronologies, cross-date, correlate to climate field, estimate climate reconstructions, and more. These software tools revolutionized dendrochronology, and their free availability to the community is a testament to the importance of open communication and collaboration in science. Many of these programs utilize an interactive user interface and are based in FORTRAN, which is a powerful and versatile language well suited to scientific computing. They efficiently perform complex calculations, preventing the user from having to program them.

PcReg has become the standard for many researchers seeking to estimate past climates from tree rings using PCR. Like other programs designed for tree-ring analysis (ARSTAN, COFECHA, EDRM), PcReg handles the legacy fixed-width formats (i.e. Tuscon and chronology formats) that are commonly used to store tree-ring data, but is not otherwise flexible in how the data areformatted. The current publicly available version of PCreg requires users to run each individual nest separately and splice them together. This process is time consuming and creates the potential for human error. The compiled nature of the program does not allow for users to visualize data, perform transformations, check residuals of liner regressions, or other steps that are typically part of the process of statistical analysis. This encourages poor practice, especially among new users or those who are not trained in data analysis.

## R Environment

R is a powerful statistical programming language and software environment which has several advantages over common FORTRAN or MATLAB based programs. Unlike Matlab, R is free and open source, in addition to being extremely versatile in its application for many different types of statistical analyses, data processing, visualization, etc. In comparison to Fortran, R is an interactive programming language, easier to use and develop in, widely taught in graduate programs, and there are large numbers of online learning resources. It has a large global user-base across a wide variety of disciplines, many of whom are actively engaged in various wikis, newsletters, and web forums (such as stack exchange). This create a community of support through active dialogue and problem solving.

In contrast to the commonly used FORTRAN programs for tree-ring analysis, R utilizes the command-line rather than an interactive user interface. On its surface, this may seem like a disadvantage to some users. However, most common tree-ring packages in R are well-documented with vignettes that provide worked examples. There are great workflow advantages in having repeatable scripts that can be adjusted and rerun all within R, without having to open a separate text file, close and reopen a program file, and re-run log files. It’s also simple to reorganize and do any number of manipulations to the data (i.e. filtering, transformations, etc), as well as many types of analyses (tree-ring specific or not), and produce publication quality visualizations, all in one software environment. Thanks to the work of Andy Bunn and others on package dplR (2012), fixed width format files can be read into R, making all manner of analysis on these data more accessible.

The implications for reproducibility in using script-based code are perhaps the most important aspect. Peer-reviewed journals are increasingly requiring data and analysis to be provided as part of the manuscript submission. This is not feasible with the many output files that are rendered in programs like PCreg or ARSTAN, and the outputs are not universally very readable outside of those individuals who work with them frequently. A standard, versatile, common, and cross-disciplinary environment for analysis is important for clear communication and reproducibility of results, as well as collaborations with subject matter experts in other fields.

In the case of PcReg, while the methods are well-documented in a number of publications, there is no documentation for the use of the program. The interactive interface is not intuitive, therefore its use has become something of an oral tradition, where an experienced researchers share how to use the program through workshops or individual interactions. While there is nothing inherently wrong with this approach, the strict documentation requirements for inclusion in the CRAN repository are an additional advantage of working within the R environment.

# Pcrecon Package

## Installation

The current version of pcrecon can be downloaded using the install\_github() function in the devtools packages:

> devtools::install\_github("lgsmith295/pcrecon")

## Citation

When using the nested PCR methods for tree ring reconstructions, please cite Ed Cook’s work on the North American Drought Atlas:

*Cook, E. R., Meko, D. M., Stahle, D. W., & Cleaveland, M. K. (1999). Drought reconstructions for the continental United States. Journal of Climate, 12(4), 1145-1162.*

If using this package to perform those procedures, also cite:

*Laura Smith, Daniel Hocking and Nicholas Nagle (2020). pcrecon: Principal Component Regression for Dendroclimatology. R package version 0.1.0.*

## Functions

There are three main functions in the pcreg package for making reconstruction estimates: load\_clim(), eval\_clim() and pcreg().

**load\_clim()** takes monthly climate data and selects individual months, averages, or sums across selected months (previous year through current year, -12:12). Input is a dataframe or matrix object containing climate data, months of interest, and method for summarizing. Output is an S3 list which includes the climate variable and the ar model if prewhitening is indicated.

**eval\_clim()** selects chronologies for inclusion in the PC regression based on correlation with the climate variable. Inputs are the S3 list as created by the load\_clim() function and a dataframe containing tree-ring chronology data in which each column is a chronology. Output includes a dataframe of correlation coefficients for each chronology and the target climate variable, a dataframe of chronologies selected for inclusion in the model, and a dataframe containing the start and end year of each selected chronology.

**pcreg()** calculates principal component eigenvectors and linear regression model estimates, then makes predictions for each nest and splices them together. Output includes: climate reconstruction, model statistics and validation statistics for each nest, PCA objects for each nest, and lm objects for each nest.

Additional functions are used to retrieve metadata from ITRDB downloaded tree ring files, filter tree rings for inclusion based on spatial location or correlation to target climate, and to upload and summarize climate data from various formats. An example of the workflow using pcrecon is provided below.

# Worked Example

## Data

Tree ring data downloaded from the ITRDB using a bounding box (figure 1). Climate data is USGS Streamflow data for gauge #347432 (Gallinas Creek near Montezuma), Latitude 35°39’07.18“, Longitude 105°19’07.79”, downloaded from waterdata.usgs.gov. These data are included in the pcrecon package. Graphical user interface, map

Description automatically generated

Figure 1: Bounding Box for retrieving ITRDB tree ring chronologies

## Code

Typical workflow for this type of analysis often begins with downloading proxy data from a repository, or uploading the users own data. Tree ring data downloaded from the ITRDB is accompanied by a folder of metadata files in JavaScript Object Notation (JSON) format. The parse\_json() function can be used to creates a table which includes: lat/lon, chronology length, species, elevation, and NOAA ID for each chronology. The user indicates the directory location for the folder containing the .json files using the “dir” argument. When center = TRUE, the mean between the minimum/maximum elevation, northernmost/southernmost latitude, and easternmost/westernmost longitude is reported.

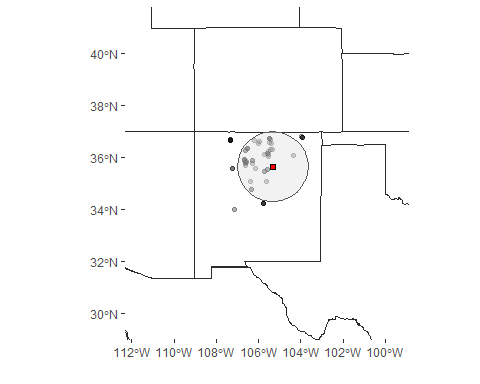
’> metadata <- parse\_json(dir = “extdata/metadata”, center = TRUE)

| ID | NOAADataTableId | earliestYear | mostRecentYear | timeUnit | speciesCode | scientificName | chron\_length | lat | lon | elev |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NM581 | 19237 | 1520 | 2003 | AD | PIED | Pinus edulis Engelm. | 483 | 36.5333 | -106.0167 | 2300 |
| NM582 | 19238 | 1575 | 2003 | AD | PIPO | Pinus ponderosa Douglas ex C. Lawson | 428 | 36.6167 | -105.9833 | 2500 |
| NM588 | 29778 | 620 | 2011 | AD | PIPO | Pinus ponderosa Douglas ex C. Lawson | 1391 | 36.2820 | -106.6450 | 2525 |

Having these metadata in a table allows the user to filter based on variables such as location, species, elevation, etc. The pcrecon package contains two functions for filtering chronologies or other datasets based on spatial location. The first, filter\_rad(), selects chronology IDs that fall within a given radius from a point. The arguments for this function are: dataframe containing site ID, latitude, and longitude, the lat/lon of the center point, and the radius in km. The output of this function is a character vector of chronology IDs that fall within that radius.

Here, we will filter within a 150km radius of the streamgauge, and save the output as an object called select\_crns\_rad.

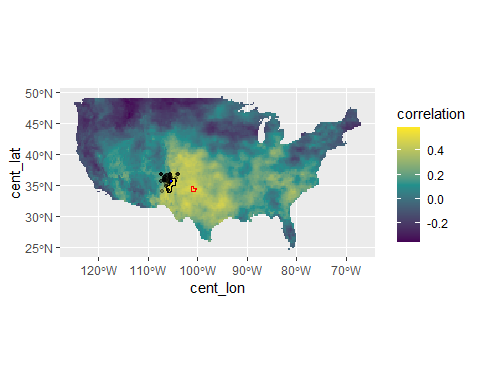
`> select\_crns\_rad <- filter\_rad(x = metadata, cent\_lat = 35.65, cent\_lon = -105.32, radius = 150, plot = TRUE)’



Another option for filtering chronologies spatially is to select those within a given climate footprint using the filter\_foot() function. This method is most often used when reconstructing streamflow. Chronologies are selected within an area where the correlation (r) between streamflow and precipitation or soil moisture meets a given threshhold (Maxwell et al. 2017). This function requires a raster file. In this example, a .nc file created in the KNMI climate explorer (<https://climexp.knmi.nl/start.cgi>) is used, which contains correlation coefficients between PRISM 4km gridcell precipitation data (PRISM Climate Group, 2004) and Gallinas Creek streamflow.

We will select chronologies that are in an area where correlation between streamflow and precipitation is 0.5 or higher (r = 0.5). Often a lower threshold of .3 is used, for the purpose of keeping the example dataset small we use a more strict cutoff.

> footprint <- raster::raster(system.file("extdata/gallinas\_cf.nc", package = "pcrecon", mustWork = TRUE)) > select\_crns\_fp <- filter\_foot(x = metadata, footprint = footprint, r = 0.5, cent\_lat = 35.65, cent\_lon = -105.32, radius = 150, plot = TRUE)



Those tree ring records (circles) that fall within the black polygon are included, anything in the red polygon would not have been included. The radius and center point arguments in this function allow the user to designate a radius that polygons must intersect with (not fall completely within), in order to prevent tree rings in locations with spurious correlations that are outside a reasonable distance to be included in analysis.

filter\_foot and filter\_rad return a character vector containing IDs of the chronologies that should be included in the next step of the analysis. The load\_crns function reads those .crn files from a folder, such as those downloaded from the ITRDB or produced using dplR or ARSTAN. The user can select which measurement and chronology types associated with each ID to read in (see documentation for option). The default is the standard ring width chronology. The output is a dataframe in which columns contain chronologies and rows are ring width observations

> crns\_df <- load\_crns(dir = system.file("extdata/crns", package = "pcrecon", mustWork = TRUE), crns = select\_crns\_fp)

| year | nm014 | nm038 | nm039 | nm040 | nm111 | nm112 | nm118 | nm607 | nm616 | nm617 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 878 | NA | NA | NA | NA | NA | NA | NA | 0.466 | NA | NA |
| 879 | NA | NA | NA | NA | NA | NA | NA | 1.047 | NA | NA |
| 880 | NA | NA | NA | NA | NA | NA | NA | 1.394 | NA | NA |

Climate data can be read in from any file format using base R functions. From here, data can be checked for skewness, transformed, reformatted, checked for missing values and autocorrelation, etc.

The load\_clim function aggregates the climate by calculating the mean or sum over particular months of interest, select individual months. This function requires that these data be in 13 column (year and months 1:12) or long (3 column - “year”, “month”, and “value”) format.

>climate <- load\_clim(clim = climate\_long, mos = c(4:7), method = "mean", prewhiten\_clim = FALSE)

> ar(climate$clim\_small$values)

##   
## Call:  
## ar(x = climate$clim\_small$values)  
##   
##   
## Order selected 0 sigma^2 estimated as 0.8811

eval\_clim() calculates correlations between the climate variable and the chronologies. Here, the user indicates the calibration/validation period for the intended model, the window over which to correlate the chronologies and climate, leads/lags, correlation test type, confidence value, and a correlation cutoff for selecting chronologies. Weighting features will be added later. See function documentation for options.

>data\_obj\_pw <- eval\_clim(crns = crns\_df, lead = 0, prewhiten\_crn = TRUE, climate = climate, calib = 1950:1968, valid = 1927:1949, cor\_window = "calib", type = "pearson", alternative = "two.sided", r = -1, conf = 0.90, print = TRUE, out\_fmt = "csv", out\_dir = "test/", pr\_years = NULL)

The output of eval\_clim() is an S3 object of class “PCreg\_data”. This object contains necessary pieces that feed into the final regression model function. You can view the the components of the object using the “$” operator:

>data\_obj$cors\_table table of all cross correlations

>data\_obj$cors\_table\_small table of selected chronology correlatons

>data\_obj$nests beginning and end year for nests

>data\_obj$clim\_arautoregressive model selected for climate

## [1] "PCreg\_data"

## chronology leads lags correlation p\_value  
## 1 nm014 0 0 0.3485742 0.1435812840  
## 2 nm038 0 0 0.6370008 0.0033556787  
## 3 nm039 0 0 0.5203948 0.0223594935  
## 4 nm040 0 0 0.7004643 0.0008379848  
## 5 nm111 0 0 0.3810636 0.1074696926  
## 6 nm112 0 0 0.6387499 0.0032429250  
## 7 nm118 0 0 0.6548110 0.0023461896  
## 8 nm607 0 0 0.6006589 0.0065375158  
## 9 nm616 0 0 0.5227324 0.0216589846  
## 10 nm617 0 0 0.6721510 0.0016194473

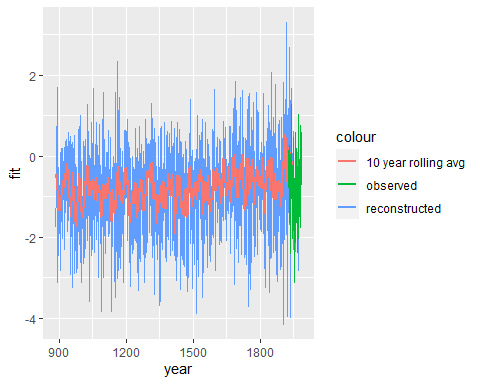
## chronology leads lags correlation p\_value  
## 1 nm038 0 0 0.6370008 0.0033556787  
## 2 nm039 0 0 0.5203948 0.0223594935  
## 3 nm040 0 0 0.7004643 0.0008379848  
## 4 nm112 0 0 0.6387499 0.0032429250  
## 5 nm118 0 0 0.6548110 0.0023461896  
## 6 nm607 0 0 0.6006589 0.0065375158  
## 7 nm616 0 0 0.5227324 0.0216589846  
## 8 nm617 0 0 0.6721510 0.0016194473

## ID startYR endYR  
## 1 nm039 1690 1972  
## 2 nm617 1655 2008  
## 3 nm112 1644 1982  
## 4 nm616 1616 2008  
## 5 nm040 1579 1972  
## 6 nm118 1564 1982  
## 7 nm038 1556 1972  
## 8 nm607 878 1989

## NULL

The pcreg() function takes the information from the previous step ,in the form of the PCreg\_data object, and performs the principal component regression estimates for each nest and splices them together. The user indicates the selection threshold for PCs, the window over which to calculate PCs, and the window over which to scale the variance.

>recon <- pcreg(data = data\_obj, pc\_calc = "calib", select\_pc = "eigenvalue1", scale\_var = "calib", plot = TRUE, weight = NULL, cum\_perc = NULL)



The output of this function is an object of class PCreg\_recon, and you can call the pieces of it as was done above. Output contains reconstruction estimates, validation and calibration period statistics (RE, CE, R2), as well as linear model and Principal Component Analysis objects for each nest and autoregressive model statistics for climate and chronologies if AR prewhitening is used.

>recon$model\_stats linear regression model statistics

>recon$calibration\_stats calibration period statistics`

`>recon$validation\_stats validation period statistics

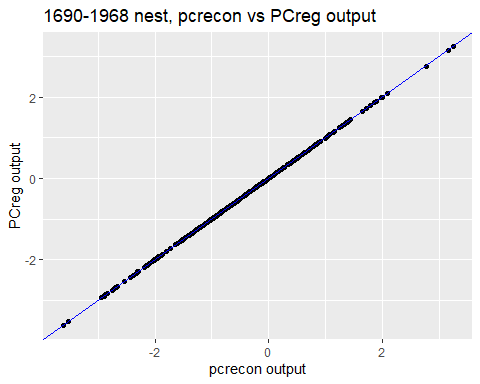
## model sigma r.sq adj.r.sq nest\_lwr  
## 1 PC1 0.6125962 0.6332495 0.6116759 1690  
## 2 PC1 0.6087624 0.6378255 0.6165211 1655  
## 3 PC1 0.6078603 0.6388981 0.6176568 1644  
## 4 PC1 0.6329574 0.6084644 0.5854329 1616  
## 5 PC1 0.6711693 0.5597631 0.5338668 1579  
## 6 PC1 0.7108057 0.5062308 0.4771855 1564  
## 7 PC1 0.7908501 0.3887617 0.3528065 1556  
## 8 PC1 0.8087426 0.3607911 0.3231905 878

## model R2 RE CE  
## 1 1 0.6332495 0.7342048 0.6332495  
## 2 2 0.6378255 0.7375212 0.6378255  
## 3 3 0.6388981 0.7382986 0.6388981  
## 4 4 0.6084644 0.7162424 0.6084644  
## 5 5 0.5597631 0.6809471 0.5597631  
## 6 6 0.5062308 0.6421506 0.5062308  
## 7 7 0.3887617 0.5570172 0.3887617  
## 8 8 0.3607911 0.5367460 0.3607911

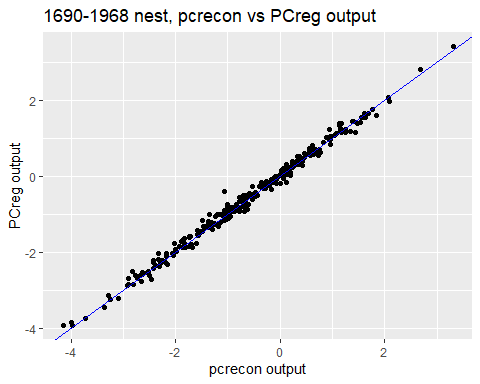
## model R2 RE CE  
## 1 1 0.5893141 0.3061563 -0.055101271  
## 2 2 0.5687449 0.3382536 -0.006292215  
## 3 3 0.5683329 0.3541046 0.017811915  
## 4 4 0.5855548 0.6432338 0.457479447  
## 5 5 0.5563242 0.6970239 0.539275952  
## 6 6 0.4524736 0.6194009 0.421237727  
## 7 7 0.4665442 0.6469050 0.463062020  
## 8 8 0.5133201 0.6441045 0.458803476

# Comparison to PCreg

When autoregressive prewhitening is not used, the pcrecon package perfectly replicates the results of PCreg:



While the process of prewhitening is identical between the two programs, the period of time over which the AR model is estimated varies. PCreg requires the user identify a period within each nest to calculate the autoregressive model coefficients, whereas in pcrecon, we use the entire length of the chronology. This produces very similar results that (= .994) that fall well within the bounds of uncertainty for these types of models.



# Citations

Briffa, K. R., P. D. Jones, T. M. L. Wigley, J. R. Pilcher, and M. G. L. Baillie. 1983. “Climate Reconstruction from Tree Rings: Part 1, Basic Methodology and Preliminary Results for England.” *Journal of Climatology* 3 (3): 233–42. <https://doi.org/10.1002/joc.3370030303>.

Bunn, Andrew G. 2008. “A Dendrochronology Program Library in R (dplR).”

Cook, Edward R., David M. Meko, David W. Stahle, and Malcolm K. Cleaveland. 1999. “Drought Reconstructions for the Continental United States.” *Journal of Climate* 12 (4): 1145–63. [https://doi.org/10.1175/1520-0442(1999)012<1145:drftcu>2.0.co;2](https://doi.org/10.1175/1520-0442(1999)012%3C1145:drftcu%3E2.0.co;2).

Harley, Grant L, Justin T Maxwell, Evan Larson, Henri D Grissino-Mayer, Joseph Henderson, and Jean Huffman. 2017. “Suwannee River Flow Variability 1550–2005 CE Reconstructed from a Multispecies Tree-Ring Network.” *Journal of Hydrology* 544: 438–51. <https://doi.org/10.1016/j.jhydrol.2016.11.020>.

Jacoby, G. C., and E. R. Cook. 1981. “Past Temperature Variations Inferred from a 400-Year Tree-Ring Chronology from Yukon Territory, Canada.” *Arctic and Alpine Research* 13 (4): 409. <https://doi.org/10.2307/1551051>.

Maxwell, R. S., G. L. Harley, J. T. Maxwell, S. A. Rayback, N. Pederson, E. R. Cook, D. J. Barclay, W. Li, and J. A. Rayburn. 2017. “An Interbasin Comparison of Tree-Ring Reconstructed Streamflow in the Eastern United States.” *Hydrological Processes* 31 (13): 2381–94. <https://doi.org/10.1002/hyp.11188>.

Maxwell, R. Stockton, Amy E. Hessl, Edward R. Cook, and Neil Pederson. 2011. “A Multispecies Tree Ring Reconstruction of Potomac River Streamflow (950-2001).” *Water Resources Research* 47 (5): 1–12. <https://doi.org/10.1029/2010WR010019>.

PRISM Climate Group, Oregon State University, <http://prism.oregonstate.edu>, created 4 Feb 2004.

*Surface Temperature Reconstructions for the Last 2,000 Years*. 2006. Washington, D.C.: National Academies Press. <https://doi.org/10.17226/11676>.

Williams, A. Park, Edward R. Cook, Jason E. Smerdon, Benjamin I. Cook, John T. Abatzoglou, Kasey Bolles, Seung H. Baek, Andrew M. Badger, and Ben Livneh. 2020. “Large Contribution from Anthropogenic Warming to an Emerging North American Megadrought.” *Science* 368 (6488): 314–18. <https://doi.org/10.1126/science.aaz9600>.

Zang, C, and Franco Biondi. 2015. “Treeclim: An R Package for the Numerical Calibration of Proxy-Climate Relationships.” *Ecography* 38 (4): 431–36. <https://doi.org/10.1111/ecog.01335.>