Accepted Manuscript

Title: Hierarchical regression models for dendroclimatic standardization and climate reconstruction

Authors: Scott Steinschneider, Edward R. Cook, Keith R. Briffa, Upmanu Lall

PII: S1125-7865(17)30004-8
DOI: http://dx.doi.org/doi:10.1016/j.dendro.2017.05.003
Reference: DENDRO 25447

To appear in:

Received date: 21-1-2017
Revised date: 3-5-2017
Accepted date: 24-5-2017

Please cite this article as: {http://dx.doi.org/

Hierarchical Regression Models for Dendroclimatic Standardization and Climate Reconstruction

Scott Steinschneider[1,*] ss3378@cornell.edu , Edward R. Cook[2] drdendro@ldeo.columbia.edu , Keith R. Briffa[3] k.briffa@uea.ac.uk , Upmanu Lall[4] ula2@columbia.edu

[1]Assistant Professor, Riley-Robb Hall, 111 Wing Dr., Department of Environmental and Biological Engineering, Cornell University, Ithaca, NY, 14853. Email: ss3378@cornell.edu, Tel: 201-913-1964 (Corresponding Author)

[2]Research Professor, 61 Route 9W, Tree Ring Lab, Lamont-Doherty Earth Observatory, Palisades, NY, 10964. Email: drdendro@ldeo.columbia.edu, Tel: 845-365-8618

[3]Emeritus Professor, School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, UK, 10964. Email: k.briffa@uea.ac.uk

[4]Professor, 500 West 120th Street, Department of Earth & Environmental Engineering, Columbia University, New York, NY, 10027. Email: ula2@columbia.edu, Tel: 212-854-7081

[*]Corresponding author

Abstract

Tree-ring based paleoclimate reconstructions entail several sequential estimation or processing steps. Consequently, it can be difficult to isolate climatic from non-climatic variability in the raw ring width measurements, estimate the uncertainty associated with a reconstruction, and directly infer how specific techniques used to sequentially fit growth curves or to reconstruct climate influence the final estimates. This paper explores the use of hierarchical regression models to address these problems. The proposed models simultaneously model the entire reconstruction process in a way that is consistent with the existing step-by-step estimation framework, but allow for uncertainty estimation and propagation across steps, which can help determine how best to improve a candidate model. The utility of hierarchical models is tested for an example, the reconstruction of summertime temperatures in northern Sweden in a cross-validated framework relative to 1) a sequential process of growth curve fitting followed by chronology development, 3) an iterative, "signal-free" approach, and 2) a signal-free regional curve standardization (RCS-SF). Further, an exploration of different structures within the unifying hierarchical framework is provided to illustrate how one could easily test a variety of choices of model design. We focus on a subset of choices relevant to recent dendroclimatic studies using hierarchical methods and related to 1) data transformation, 2) the benefits of biological detrending and climate reconstruction in a single step 3) partial pooling of the age model across trees, 4) the homogeneity of variance across tree-ring residuals, 5) the structural form of the age model, and 6) the inclusion of autoregressive processes for the tree-ring residuals. The work described here represents part of a series of ongoing explorations of potential advances over current dendroclimatic reconstruction approaches

and commonly implemented ways in which they have and are specifically implemented. The results show that hierarchical modeling appears to offer improved climate reconstructions over the standardization techniques explored in this exercise, substantially so for the non-RCS sequential and iterative methods.

Introduction

Paleoclimate reconstructions from tree rings have proven enormously useful for understanding past climate variability prior to instrumental or historical records. The development of these reconstructions requires that variability in tree-ring width measurements (or other growth-related data) related to external climate forcing be isolated from other variability in the tree-ring measurements associated with internal growth processes, such as biological age-related trends. These trends emerge as the stem expands over the life of the tree and subsequently radial ring widths slowly decline.

In dendroclimatology, methodologies to separate climatic from non-climatic variability in the raw ring width measurements are referred to as standardization techniques (Fritts, 1976). These techniques generally follow a three-step, sequential procedure in which 1) age-related growth trends are estimated and removed from each tree-ring series, 2) trend-adjusted series are averaged across trees to develop a single chronology, and 3) a target climate series of interest is modeled as a function of the chronology to develop the reconstruction. The possible removal of part of the climate signal with the biological age-related trend is a common problem that arises in the first two stages of this procedure. This problem, known as the 'segment length curse' (Cook et al., 1995; Briffa et al., 1996), arises because the age-related growth trend is fit to the length of each tree-ring

series using deterministic (e.g., monotonic decreasing linear, modified negative exponential growth) or flexible (e.g., smoothing splines) curves that, by construct, assume trends across the length of the data series consistent with the growth model are associated with biological and not climatic variability. Thus, decadal to centennial scale climate variability present in the tree rings but with period longer than the length of the tree-ring series is subsumed into the biological trend model and removed from the chronology and subsequent climate reconstructions.

A variety of approaches have been proposed to mitigate the loss of climate information when fitting and removing age-related trends. Regional curve standardization (RCS) is an empirical curve fitting technique that assumes a homogenous growth rate across all trees of the same age (or age class) and estimates that rate based on the average ring width for all tree rings in a given age class, with post-average smoothing (Briffa et al., 1992). The RCS approach assumes that the distribution of age classes is sufficiently random across trees in any given time period so that climate-related variance in that time period is averaged out in the calculation of age-related growth for each age class. Because the biological growth curve is estimated using all tree-ring series, it is not constrained by the length of any one series and the resulting chronology can exhibit variability on long timescales up to the length of the full chronology (Esper et al., 2002; Peters et al., 2015). However, the assumptions made in the RCS procedure, namely that a single, homogenous growth curve can be applied to all trees in a stand, are often violated due to variations in local conditions (e.g., soil, competition, microclimate, etc.) experienced by individual trees (Briffa and Melvin, 2011).

To circumvent these challenges and minimize the effects of the segment-length curse, Melvin and Briffa (2008) proposed the "signal-free" method of standardization. In this approach, biological age-related trends are estimated and removed for individual trees and a chronology then estimated, similar to a traditional standardization. However, the chronology is then removed from each tree and the individual age models re-estimated. A new chronology is developed and the entire procedure iterated until the chronology converges to a sufficiently fixed time series. Through this iteration, the signal-free approach removes the influence of common, climate-forced signal in individual tree-ring width series prior to biological trend estimation, thus improving the chances that the biological trend does not subsume the climate signal while still allowing for heterogeneity in biological trends. The signal-free method has also been extended to the RCS approach (i.e., RCS-SF standardization) to better manage situations where only a few older trees with common germination dates are available to estimate the climate series from early parts of the chronology (Briffa and Melvin, 2011; Melvin and Briffa, 2014a).

While the signal-free approach improves the retention of external climate forcing in the final chronology, some amount of climate signal may still be lost in the early iterations of the procedure. Recently, hierarchical regression models have been proposed as an alternative approach for isolating climate and non-climate variance in tree-ring series. In hierarchical models, the biological age-related trend and the shared climate signal across trees are estimated jointly and simultaneously in a single-step modeling procedure. To the

authors' knowledge, only a handful of studies have utilized hierarchical regression models for ring width detrending and chronology development. Concurrently, Duncan et al. (2010) and Bontemps et al. (2010) were the first to propose such an approach. Duncan et al. (2010) found that cross-validated temperature reconstructions in New Zealand were substantially improved over a reconstruction based on a sequential procedure that utilized individual smoothing splines for detrending. The model proposed in Bontemps et al. (2010) was compared against an RCS procedure and found to produce similar chronologies (Bontemps and Esper, 2011), although they did not present a comparative, cross-validated assessment of reconstructed climate. Schofield et al. (2016) adopted a Bayesian hierarchical approach and proposed a novel framework in which the model linking the chronology to the climate series targeted for reconstruction was calibrated simultaneously with the models of biological trend for each tree-ring series. In that study, a variety of model variants were developed to test different underlying assumptions in model structure, and these different model versions were compared to both standard and RCS procedures. While Schofield et al. (2016) did present a novel framework and a thorough discussion of hierarchical model development and inter-comparison, they were unable to show substantive improvements in cross-validated reconstructions of Scandinavian summer temperature over other standardization techniques. Through our work we find that this was primarily due to the length of the temperature series used in the analysis. Schofield et al. (2016) also did not compare their results to signal-free approaches designed to better separate age- and climate-related variability in the ring width series. Our results show that a RCS-SF approach is quite robust and has comparable out-of-sample performance to the hierarchical models, although the two

approaches do lead to different chronologies and reconstructions prior to the instrumental record.

This study builds directly from the work presented in Schofield et al. (2016) and further explores the use of hierarchical regression models for dendroclimatic standardization and climate reconstruction and how they compare to existing approaches. Similar to Schofield et al. (2016), we adopt a hierarchical Bayesian framework, although this is not necessary to implement the hierarchical construct. Our work differs from the original study presented in Schofield et al. (2016) in three primary ways. First, we consider a variety of additional model choices not explicitly assessed in the original study and test their implications for the fidelity of climate reconstructions. These choices include 1) the type of data transformation, 2) biological detrending and climate reconstruction in a single modeling step, 3) partial pooling of the age model across trees, 4) the homogeneity of error variance across tree-ring residuals, 5) the structural form of the age model, and 6) the inclusion of autoregressive processes for the tree-ring residuals. Second, we compare the hierarchical models to signal-free approaches for standard and RCS detrending, which are better designed to avoid subsuming the climate signal into the biological trend. Finally, we use a substantially longer instrumental temperature record to better differentiate the reconstruction skill of different hierarchical and conventional standardization approaches.

The remainder of the paper will introduce the hierarchical modeling framework considered in this work, develop model variants that represent alternative hypotheses of

the underlying data generating process, detail the estimation and cross-validation frameworks used to assess the fidelity of different model-based reconstructions, and present the results of the comparison.

Data

To motivate the model developments presented in this work and compare them against the results of Schofield et al. (2016), we use the same tree-ring data set composed of annual growth increments of 247 living and subfossil Scots pine (*Pinus sylvestris*) growing near the latitudinal tree-line in Torneträsk, northern Sweden (Grudd et al., 2002; Briffa et al., 2008). After cross-dating, the earliest ring widths in this dataset extend back to 1497 and the most recent rings end in 1997. All series have at least 25 annual increments, with the average and maximum series length equal to 179 and 484 years, respectively. Figure 1 shows the distribution of tree-ring data across years.

Schofield et al. (2016) developed their methods based on an 83-year (1913-1995) record of Torneträsk summertime (JJA) temperatures recorded at the Abisko weather station. We also test our standardization approaches against a slightly longer Abisko record (1913-1997) to facilitate a direct comparison against the results in Schofield et al. (2016). However, we focus our attention primarily on tests using a 182-year record of summer temperature from 1816-1997 at Tornedalen, Sweden (Klingbjer and Moberg, 2003). Though there may be some degradation in the signal between the tree rings and temperature at the farther Tornedalen site, the use of this longer record enables each model variant to be tested against a much longer out-of-sample period in the cross-

validation framework. Importantly, the Tornedalen record exhibits a substantial shift in temperature around 1912 (see Figure 1), thus providing a challenging testing dataset that can be used to help distinguish between the performance of different models.

The Tornedalen temperature record is a composite of four different temperature series that required considerable adjustments to produce a merged homogeneous temperature record. See Klingbjer and Moberg (2003) for details. Even so, we conducted independent homogeneity tests of the summer target temperature season by comparing the Tornedalen record against the long homogeneous Stockholm (http://bolin.su.se/data/stockholm/air_temperature.php) and St. Petersburg (Phil Jones, pers. comm.) temperature records. In neither comparison was any sign of inhomogeneity found in the Tornedalen summer temperature series.

Hierarchical Regression Models for Standardization and Reconstruction

Consider that we have *M* tree-ring series of length n_i, where $y_{i,t}$ is the radial increment for the i^{th} tree in year *t*. Here we assume that all tree-ring series have already been cross-dated, each series contains data within a range of years $T_i = \{t_1^i, \dots, t_{n_i}^i\}$, and together the *M* series span a total period of *T* years. Each series *y* contains a biological age-related trend, $B(age_{i,t}|\theta_i)$, that is a function of the cambial age of the tree, $age_{i,t}$, and parameters θ_i for each tree, as well as a common signal of external climate forcing, η_t, that is shared amongst all of the trees. The goal is to develop a reconstruction of a target climate series, x_t, that is related to η_t, and simultaneously an estimate of the biological growth curves. We assume that the instrumental climate series x_t is available from some

time t_0 to T and has been standardized by removing its mean and dividing by its standard deviation.

Conventional, Non-Hierarchical Models

In a conventional three-stage standardization approach (hereafter TS standardization), the function $B(age_{i,t}|\theta_i)$ is selected from a list of possible options (linear, modified negative exponential, Hugershoff, smoothing splines, etc.) and estimated separately for each tree. Residuals $\widehat{\varepsilon_{i,t}}$ are then estimated as the difference between $y_{i,t}$ and $B(age_{i,t}|\widehat{\theta_i})$. These residuals are averaged across trees for each time period t to develop an estimated chronology $\widehat{\eta_t}$, and the target climate series x_t in the instrumental period is modeled as a function of $\widehat{\eta_t}$ in order to develop the reconstruction.

The signal-free approach of standardization (hereafter SF standardization) begins similar to TS standardization, where $B(age_{i,t}|\widehat{\theta_i})$ is first estimated for each tree, and then $\widehat{\eta_t}$ is estimated as the average $\langle y_{i,t} - B(age_{i,t}|\widehat{\theta_i})\rangle$ across trees. However, in the signal-free approach, this process is iterated with an adjusted estimate of the age model $B(age_{i,t}|\widetilde{\theta_i})$ based on the adjusted series $\widetilde{y_{i,t}} = y_{i,t} - \widehat{\eta_t}$, followed by a new estimate $\widetilde{\eta_t}$ based on $\langle y_{i,t} - B(age_{i,t}|\widetilde{\theta_i})\rangle$. These iterations continue until the chronology $\widetilde{\eta_t}$ converges to a nearly fixed time series.

A basic RCS approach follows the same general procedure as the TS approach, but $B(age_{i,t})$ is developed as a single growth curve by averaging $y_{i,t}$ for different age classes (e.g., $0 < age_{i,t} < 10$, $10 < age_{i,t} < 20$, etc.) and smoothing the resulting curve

using the selected function $B(age_{i,t})$. Ideally, pith offset estimates (the number of inner rings missing to pith in the actual ring-width measurement series used) should be included in the estimation of the RCS curve (Kershaw, 2007; Briffa and Melvin, 2011). However, this was not done in the Schofield et al. (2016) study, and so it is not done here to allow for a more direct comparison using the same data as in that study. However, in terms of methodology, Schofield et al. (2016) used the original (simplified) RCS approach described in Briffa et al. (1992). Improved methods for implementing the RCS concept have been developed (Helama et al., 2016), such as the signal-free RCS (RCS-SF) approach (Melvin and Briffa, 2014a,b). The RCS-SF is particularly useful in situations where limited subfossil samples can bias the original RCS detrending. We compare our hierarchical models to the RCS-SF approach because this method is better suited for separating biological and climate related signals, particularly for our dataset using limited ring width series from subfossil Scots pine, and so provides a more robust benchmark against which to compare the hierarchical models.

The TS, SF, and RCS methodologies have been discussed at length in the literature (Helama et al., 2004; Peters et al., 2015 and sources within) and the reader is directed to these references for additional detail. We also recognize that numerous other variants of these standardization approaches have been proposed (Briffa et al., 2001; Bontemps and Esper, 2011; Björklund et al., 2013; Briffa et al., 2013; Matskovsky and Helama, 2014; Linderholm et al., 2015; Helama, et al., 2016), but for brevity of exposition we focus on the three strategies discussed above (TS, SF, RCS-SF) and their comparison to a basic hierarchical model to highlight the major methodological and practical differences.

Basic Hierarchical Model

In a basic hierarchical modeling approach, the age models and chronology are integrated into a single model, expressed as:

$$y_{i,t} = B\left(age_{i,t}|\theta_i\right) + \eta_t + \varepsilon_{i,t} \quad (1)$$

$$\eta_t \sim Normal\left(0, \sigma_\eta^2\right) \quad (2)$$

$$\varepsilon_{i,t} \sim Normal(0, \sigma_i^2) \quad (3)$$

Here, η_t is a zero-mean deviation common to all trees, analogous to the chronologies developed in standard dendroclimatic practice. The moniker "hierarchical" comes from the fact that the parameters vary at different levels in the model, most importantly at the "lower" level that describes the actual tree ring width series (θ_i), and at "upper" levels that influence the values taken by terms in the lower level $\left(\sigma_\eta^2, \sigma_i^2\right)$. The key to the hierarchical framework is that the parameters and unknown quantities in Eqs. 1-3 are estimated jointly, instead of in stages. In the joint estimation, the upper level parameters $\left(\sigma_\eta^2, \sigma_i^2\right)$ inform the estimates of lower level parameters, (θ_i), and simultaneously, information at the lower level informs upper level estimation. Information about the common signal η_t that is potentially subsumed by the age model in the sequential process of the TS approach or the early iterations of the SF or RCS-SF procedures is preserved in the joint estimation procedure, and this is a primary benefit of the hierarchical model. Once the hierarchical model is fit, the chronology η_t can be used to reconstruct target climate series of interest in a similar manner to conventional standardization techniques.

The joint estimation procedure for hierarchical modeling can proceed by maximizing the likelihood function of all of the data. Alternatively, a Bayesian approach can be adopted (as in this study), whereby other sources of information (e.g., expert opinion, results from past studies) can be included in the estimation through the use of prior distributions for model parameters. These prior distributions provide a flexible way to impart additional structure onto the parameters, which could improve model stability and prediction. For instance, we could require that all age-related parameters be drawn from a common prior distribution, which will have the effect of pulling the parameters for each tree towards the mean of the prior while still allowing for heterogeneity between trees. This formulation draws from both the TS and RCS approaches to biological trend modeling. The prior distributions for parameters can also be made sufficiently vague (e.g., uniform distributions, normal distributions with very large variances) so that the data dominate the estimation.

To complete the model in Eqs. 1-3, we require a specific formulation for the biological age-related trend. Based on an initial assessment of the Torneträsk tree-ring data, we adopt a simple linear model after first transforming the original ring widths using a Box-Cox transformation:

$$B\left(age_{i,t}|\beta_{0,i},\beta_{1,i}\right) = \beta_{0,i} + \beta_{1,i}age_{i,t} \quad (4)$$

This age model is also used for two of the conventional standardization techniques (TS and SF). For the RCS-SF standardization, however, we used an age-dependent spline to maximize the flexibility of the homogenous age curve used across tree ring series.

Exploring Model Structure in a Hierarchical Framework

Hereafter, the hierarchical model described in Eqs. 1-4 is considered the reference model and denoted M0. This model is very similar to that of Duncan et al., 2010. A variety of other model formulations will be developed from the reference model to demonstrate the flexibility of the hierarchical modeling structure and provide guidance for future dendroclimatic studies seeking to use hierarchical models for standardization. These model formulations are selected to build from the work presented in Schofield et al. (2016) and provide insight into their utility for climate reconstruction. Table 1 summarizes all of the models considered in this work, which are described further below. Additional detail on the Bayesian estimation framework is provided afterwards.

Data Transformation

The model for radial growth increments in Eq. 1-4 was expressed as an additive model under a flexible Box-Cox transformation of the original ring widths. Multiplicative models are also commonly employed for tree-ring analyses, where annual growth increments are modeled as the product of expected growth, $B\left(age_{i,t}|\theta_i\right)$, a common deviation across trees (i.e., the chronology, η_t), and an error term (Melvin and Briffa, 2008):

$$y_{i,t} = B\left(age_{i,t}|\theta_i\right) \times \eta_t \times \varepsilon_{i,t} \quad (5)$$

However, assuming a negative exponential growth curve, the multiplicative model is equivalent to a linear additive model under a logarithmic transformation. In both Duncan et al. (2010) and Schofield et al. (2016), the tree-ring widths were assumed to follow such a model. However, this assumption may result in a biased estimation if the logarithmic transformation over-corrects for the skew commonly found in strictly positive ring width measurements (Helama et al., 2016). In such cases a less extreme and more flexible transformation, such as the Box-Cox transform, may be more appropriate (see Cook and Peters (1997), Helama et al. (2004), and supporting material). Schofield et al. (2016) mentioned briefly in their discussion that their results did not vary much between a logarithmic and square root transformation of the ring widths, but our results indicated more significant differences when comparing logarithmic and Box-Cox transformations. Therefore, we include a comparison between a linear standardization model after logarithmic transformation (hereafter M1) and the additive model after a Box-Cox transformation used in the reference model.

Joint Estimation for Target Climate Series Reconstruction

The primary hierarchical model formulation presented in Schofield et al. (2016) linked the target climate series directly to the chronology within the model framework. Their proposed formulation, which we adopt here, can be expressed by expanding the reference model as follows:

$$\eta_t \sim Normal\left(\beta_2 x_t, \sigma_\eta^2\right) \quad (6)$$

$$x_t \sim Normal(\mu_x, \sigma_x^2) \quad (7)$$

Here, the mean function for the chronology is modeled as a linear function of the climate series to be reconstructed, whose distribution is also modeled. The argument for such an approach is that during the instrumental period, $t \epsilon (t_0, T)$, the model will be able to "see" variations in the climate series that correspond to common variations in many of the tree-ring series when fitting the age models for each tree, and so the model is less likely to subsume the climate forcing into the age model and be more prone to incorporate it into the chronology, η_t, through the time-varying term $\beta_2 x_t$. For time periods prior to t_0 when x_t is unavailable, the model will consider these values of the climate series as unknown quantities that require posterior estimation. This estimation will combine information from the prior distribution in Eq. 7 with the posterior chronology η_t for $t < t_0$. We discuss this further in the section entitled *Climate Reconstruction*.

While innovative, the study presented in Schofield et al. (2016) did not test the improvements in climate reconstruction afforded by this approach against a suitable control that could isolate the effects of the explicit link between climate and chronology. We develop such a controlled experiment here by comparing the model formulation in Eqs. 6-7 (hereafter M2) against the reference model, where the calibrated values of the zero-mean η_t are regressed against the climate series outside of the hierarchical standardization model. We also note that the formulation in M2 may render the fitted

chronology η_t inappropriate for reconstructions of other climate data besides x_t, which may detract from its value as a generalized tool for chronology development.

Partial Pooling

The parameters of the hierarchical model are allowed to exhibit additional structure that could improve model stability and prediction, depending on the amount of tree-ring data available for the reconstruction. For instance, parameters for the biological age-related model of each tree can be linked through a parent distribution in the prior. For the linear age model considered in this study, these priors can be specified as:

$$\beta_{0,i} \sim N\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right) \tag{8}$$

$$\beta_{1,i} \sim N\left(\mu_{\beta_1}, \sigma^2_{\beta_1}\right) \quad \text{T}(-\infty,0) \tag{9}$$

Here, $\mu_{\beta_0}, \sigma^2_{\beta_0}, \mu_{\beta_1}, \sigma^2_{\beta_1}$ are hyperparameters also calibrated in the model and T(-∞,0) indicates that the prior for β_1 is truncated to be negative to ensure a decreasing growth curve. The effect of this additional structure will be to pull, or shrink, the age parameter estimates towards their mean value across all trees. The partial pooling of information for the age model across trees somewhat resembles the RCS approach, but rather than requiring a homogenous age model for all trees, heterogeneous models are permitted with the additional constraint that they share some amount of information dependent on the uncertainty associated with the data from each tree and the regional prior. Both Duncan et al. (2010) and Schofield et al. (2016) partially pool information across trees, but neither assessed whether such an approach provides benefits for the climate reconstruction.

Partial pooling affords the most benefits when the data are corrupted by outlier values or there is a paucity of data for individual trees in the model. However, since tree-ring series often contain dozens to hundreds of rings for each tree, it is unclear whether partial pooling of age models will provide any benefit to the reconstruction or will just degrade the age models for each tree. While the answer to this question is likely case study specific and dependent on the data available for each tree series, we provide some insight based on the Torneträsk data by comparing the reference model to a similar model with the priors given in Eqs. 8-9 (hereafter M3).

Homogenous or Heterogeneous Variance

One seemingly innocuous model choice includes whether to include a common or tree-specific error variance. In both Duncan et al. (2010) and Schofield et al. (2016), the variance parameter σ_i^2 is assumed the same for all trees and set equal to σ^2. However, if the variability in the residuals for each tree series varies significantly from tree to tree, this assumption may give too much weight to the information from certain trees and overly discount others. We test this underlying assumption by altering the reference model to include only a single, constant variance for all tree-ring residuals (hereafter M4). We also note that a compromise is possible in the hierarchical model by partially pooling these variance parameters across trees, although no such model is tested here.

Choice of Biological Trend Model

In the models specified above, simple linear age models were adopted following a Box-Cox or logarithmic transform. However, a variety of other biological age-related trend

models are possible, including other deterministic functions and flexible smoothing splines. We do not attempt a thorough review of all of these approaches here. Rather, an examination of the data suggests that a modified Hugershoff curve (Warren, 1980) is potentially more appropriate for the Box-Cox transformed data than a simple linear model (see supporting material and Briffa and Melvin, 2011). The modified Hugershoff curve is given as:

$$B(age_{i,t}|\theta_i) = a_i \times age_{i,t}{}^{b_i} exp(-c_i \times age_{i,t}) + d_i \quad (10)$$

We first tested several formulations for the Hugershoff curve, including one where parameters are constant across all trees (similar to a RCS standardization), one where separate parameter sets are allowed for each tree, and a compromise where the parameters a_i and d_i are allowed to vary by tree but parameters b and c (which control curvature) are common to all trees. This final formulation performed equal or better than the other formulations under out-of-sample cross validation (see supporting information) and is therefore considered in the broader model comparison (hereafter M5).

Modeling Autocorrelation

After accounting for the expected growth of each tree series and any common climate forcing, the residuals of the tree-rings, $\varepsilon_{i,t}$, can often still be autocorrelated in time (Macias-Fauria et al., 2012). Conventionally, it is considered good practice to represent residual autocorrelation directly in a model, as this will improve predictions of the

modeled data and more accurately estimate the uncertainty of other regression parameters. However, in dendroclimatology we are less interested in accurate and precise estimates of the tree rings themselves, but rather are more concerned with accurate and precise estimates of out-of-sample climate reconstructions. By modeling the autocorrelation of the tree-ring residuals, there is a risk that part of the climate signal will be subsumed into the parameters for residual autocorrelation, although such a result is not intuitive prior to model fitting and testing. Therefore, we augment the reference model by allowing the residuals to follow an AR(1) process and include this variant in the inter-model comparison (hereafter M6). Before selecting the AR(1) process, we first examined the residuals of the fitted reference model and tested a variety of AR(p) formulations (see supporting material). The AR(1) model performed as well or better than other variants under cross-validation and therefore was chosen for the broader model comparison.

Model Fitting and Prediction

Bayesian Inference via Markov Chain Monte Carlo Sampling

In the Bayesian approach taken in this study, the estimation process involves the evaluation of the joint posterior distribution of all model parameters, given by Bayes Theorem:

$$p(\Theta|Y) = \frac{p(Y|\Theta)p(\Theta)}{\int_{\Theta} p(Y|\Theta)p(\Theta)d\Theta} \quad (11)$$

Here, Θ is the vector of all parameters, $p(Y|\Theta)$ is the likelihood function of the data, and $p(\Theta)$ is the prior distribution for the model parameters. The integral in the denominator is

a constant of proportionality required to ensure that the right hand side is a well-defined probability density function.

For the purposes of exposition, we develop the components of the posterior distribution for the reference model, M0. The joint prior distribution $p(\Theta)$, which summarizes our knowledge of the parameters prior to model fitting against the data, can often be more conveniently partitioned into a series of conditional and marginal distributions:

$$p(\Theta) = p\left(\eta_{1:T}|\sigma_\eta^2\right) \times p\left(\sigma_\eta^2\right) \times \left[\prod_{i=1}^{M} p(\sigma_i^2) \times p\left(\beta_{0,i}\right) \times p\left(\beta_{1,i}\right)\right] \quad (12)$$

with the conditional prior distribution for the chronology, η, given by:

$$p\left(\eta_{1:T}|\sigma_\eta^2\right) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{{\eta_t}^2}{2\sigma_\eta^2}\right) \quad (13)$$

All other prior distributions are set to non-informative distributions (see supporting material), with the exception of $p\left(\beta_{1,i}\right)$, which is uniformly distributed from -5 to 0 to ensure a decreasing growth curve with age. The likelihood function for the tree-ring series in M0 is given by:

$$p(Y|\Theta) = \left[\prod_{i=1}^{M} \prod_{j \in T_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(-\frac{\left(y_{i,j}-(\beta_{0,i}+\beta_{1,i}age_{i,j}+\eta_j)\right)^2}{2\sigma_i^2}\right)\right] \quad (14)$$

The posterior distribution in Eq. 11, and in particular the integral in the denominator, is often too complex to be solved using analytical methods. This challenge has been largely ameliorated with computational advances that enable the generation of samples from the posterior distribution that can be used to empirically summarize any of its features. Markov chain Monte Carlo (MCMC) sampling provides a straightforward way to generate these samples; we provide a brief summary here and direct the reader to Gelman et al. (2013) for more detail. MCMC sampling uses sampling chains to simulate a random process that has the posterior distribution as its equilibrium distribution. In one common algorithm to simulate these chains (the Metropolis-Hastings procedure), parameters are sampled from a proposal distribution and the numerator of the posterior density in Eq. 11 is evaluated under both the new and previous parameter samples to determine whether the chain should move towards the newer sample. This process is iterated over all parameters, enabling each level of the hierarchical model to inform the estimation of all the other levels. If multiple MCMC chains are initiated with very different parameter values but converge to the same region in the parameter space, the MCMC algorithm is said to have converged on the posterior distribution. This can be assessed using the Gelman and Rubin convergence criterion [Gelman and Rubin, 1992]. In this study, the posterior distribution is explored using the MCMC sampler in the software package JAGS [Plummer, 2011] with 20,000 samples for ‘burn-in’ and 20,000 samples to develop the posterior. We note that these Bayesian methods are slower to fit compared to conventional methods. As a benchmark, the reference model took approximately 1 hour to run on a MacBook Pro laptop with a 2.6 GHz Intel Core i7 processor and 16GB of Random Access Memory (RAM).

Climate Reconstruction

The predicted reconstruction of climate prior to the instrumental record is the primary variable of interest. For all hierarchical models except M2, these predictions are generated by regressing the climate variable x_t against the fitted chronology, taken as the posterior median values, $\widehat{\eta_t}$. This can be summarized as:

$$x_t \sim Normal(\beta_x \widehat{\eta_t}, \sigma_x^2), \qquad t=1,\ldots,T \tag{15}$$

This approach, while simple, ignores the uncertainty in η_t (which will depend on the data availability for each year *t*), and this can lead to a biased estimate of the regression coefficient, $\widehat{\beta_x}$, when estimated using least squares (Fuller, 1987). However, this bias will be small if the uncertainty in η_t for each year *t* is small compared to the variance of the median chronology over the fitting period. This situation is likely since the fitting period is almost always coincident with the modern instrumental period of climate records, and this is the time period when most tree-ring series are available to constrain the uncertainty in η_t. For instance, under the reference model M0, bias in $\widehat{\beta_x}$ would be on the order of 1% of its estimated value. Therefore, we maintain the simple yet slightly biased approach of ignoring the uncertainty in η_t when estimating Eq. 15. However, a full Bayesian regression model could easily be developed that accounts for the uncertainty in η_t by coupling Eq. 15 (i.e., the likelihood function) with priors for each time period, $\eta_t \sim Normal\left(\widehat{\eta_t}, \widehat{\sigma_{\eta_t}^2}\right)$, where the values $\widehat{\eta_t}, \widehat{\sigma_{\eta_t}^2}$ are estimated from the posterior distribution of $\eta_{1:T}$.

For M2, the climate series is modeled directly, and so no secondary regression is necessary. In this model, the MCMC algorithm will automatically provide posterior distributions for the out-of-sample temperature values, $x_{1:(t_0-1)}$, that include uncertainty from the remaining model parameters. These are the posterior predictions presented in Schofield et al. (2016). However, no estimates will be provided for the within-sample climate $x_{t_0:T}$ because the model sees these values as data and not unknown quantities, precluding the generation of within-sample performance statistics. This can be resolved by deriving the maximum likelihood estimates of $x_{1:T}$ for the whole period conditional on the fitted values $\widehat{\sigma_\eta^2}$, $\widehat{\mu_x}$, $\widehat{\sigma_x^2}$, $\widehat{\beta_2}$, and $\widehat{\eta_{1:T}}$:

$$x_t \sim Normal\left(\mu_{x,t}^*, {\sigma_x^2}^*\right) \tag{16}$$

$$\mu_{x,t}^* = \frac{\left(\widehat{\sigma_\eta^2}\widehat{\mu_x} + \widehat{\sigma_x^2}\widehat{\beta_2}\widehat{\eta_t}\right)}{\widehat{\sigma_\eta^2} + \widehat{\beta_2}^2\widehat{\sigma_x^2}} \tag{17}$$

$${\sigma_x^2}^* = \left(\frac{1}{\widehat{\sigma_x^2}} + \frac{\widehat{\beta_2}^2}{\widehat{\sigma_\eta^2}}\right)^{-1} \tag{18}$$

Here, we assume that the Bayesian inference described above has already been performed. Given the formulation of M2, Eqs. 17-18 describe the analytical solutions to the conditional posteriors for the mean and variance of the climate series. The time-varying mean $\mu_{x,t}^*$ is the term used for the mean climate reconstruction, whereas the full distribution in Eq. 16 contains the uncertainty around those mean estimates. Therefore, the final climate reconstruction from M2 can be developed by developing point estimates of model parameters from their fitted posterior distributions $\left\{\widehat{\sigma_\eta^2}, \widehat{\mu_x}, \widehat{\sigma_x^2}, \widehat{\beta_2}, \widehat{\eta_{1:T}}\right\}$ and

inserting these point estimates into Eqs. 17-18 to estimate a time series of mean climate $\mu^*_{x,t}$, as well as the uncertainty around the mean, $\sigma_x^{2^*}$. We use the posterior median of each conditioning parameter in the estimation above, but the uncertainty of these parameters can also be propagated into the reconstructed values of $\mu^*_{x,t}$ and $\sigma_x^{2^*}$ by sampling from their joint posterior distribution and passing those samples into Eqs. 16-18.

Cross-Validation Framework

Temperature reconstructions from the seven hierarchical models listed in Table 1 are compared against each other and reconstructions from the three conventional, non-hierarchical approaches (TS, SF, RCS-SF) to determine if and how the hierarchical framework provides advantages for climate reconstruction. We adopt a split-sample cross-validation framework in which half of the temperature data is included for model fitting and the other is reserved for testing. The cross-validation is performed twice for both the Tornedalen and Abisko temperature series, with fitting and testing periods reversed. For the Tornedalen series, we use the 1816-1912 and 1913-1997 periods, while for the Abisko series we use the 1913-1955 and 1956-1997 periods.

Four performance metrics are used to assess model performance. Three of the metrics are adopted from Cook et al. (2007): the square of the Pearson correlation (*RSQ*), the reduction of error (*RE*), and the coefficient of efficiency (*CE*). The metrics are each calculated for both the calibration period (subscript c) and the validation period (subscript v) as follows:

$$RSQ_c = \frac{\left[\sum(x_t - \overline{x_c})(\widehat{x_t} - \overline{\widehat{x_c}})\right]^2}{\sum(x_t - \overline{x_c})^2 \sum(\widehat{x_t} - \overline{\widehat{x_c}})^2} \quad , \quad RSQ_v = \frac{\left[\sum(x_t - \overline{x_v})(\widehat{x_t} - \overline{\widehat{x_v}})\right]^2}{\sum(x_t - \overline{x_v})^2 \sum(\widehat{x_t} - \overline{\widehat{x_v}})^2} \tag{18}$$

$$RE_c = 1 - \frac{\sum(x_t - \widehat{x_t})^2}{\sum(x_t - \overline{x_v})^2} \quad , \quad RE_v = 1 - \frac{\sum(x_t - \widehat{x_t})^2}{\sum(x_t - \overline{x_c})^2} \tag{19}$$

$$CE_c = 1 - \frac{\sum(x_t - \widehat{x_t})^2}{\sum(x_t - \overline{x_c})^2} \quad , \quad CE_v = 1 - \frac{\sum(x_t - \widehat{x_t})^2}{\sum(x_t - \overline{x_v})^2} \tag{20}$$

The *RSQ* metric is a measure of covariance between the reconstructed and observed climate series without any consideration of bias. The *RE* includes a bias component, but always considers the bias with respect to the mean of the observations in the period not being evaluated. Conversely, the *CE* considers bias with respect to the observed mean for the period being evaluated. High values of CE_v are the most difficult to achieve, because in order to do so the model must predict any epoch-scale shifts in the mean of the observations from the calibration to the validation period.

The three metrics above summarize the skill of the mean reconstructed climate series, but they do not capture the ability to appropriately model the uncertainty of predicted values. Therefore, the continuous rank probability skill score (*CRPSS*) is also considered. The *CRPSS* is a relative score that is only defined by the comparison of two models:

$$CRPSS = 1 - \frac{CRPS}{CRPS_{ref}} \tag{21}$$

Here, *CRPS* and $CRPS_{ref}$ are the continuous rank probability scores of the current model under consideration and the reference model, respectively. The *CRPS* is the average of

integrated square differences between the cdf of the predicted values, $F_{\hat{x}_t}$, and the cdf of the observation, F_{x_t} (defined here for the calibration period):

$$CRPS_c = \frac{1}{n_{T_c}} \sum_{t \in T_c} \int_{-\infty}^{\infty} \left(F_{\hat{x}_t}(z) - F_{x_t}(z) \right)^2 dz \quad (22)$$

The cdf of the predicted values is based on the predictive uncertainty from the regression between x_t and $\widehat{\eta_t}$ (or the pdf given in Eq. 16 for M2), while the cdf of the observed value is just the heaviside step function at the value x_t. Here, T_c and n_{T_c} are the range of years and number of years in the calibration period. $CRPS_v$ and $CRPSS_v$ can be defined similarly for the validation period. By including the entire cdf of predicted climate values, and not just the mean estimate, the *CRPS* (and thus *CRPSS*) can assess both the accuracy and precision of the reconstructed temperature series, with lower values of the *CRPS* indicating a more accurate and precise prediction.

Finally, for the Tornedalen series we also compare how the reconstructions differ when fitting on the 1816-1912 period versus the 1913-1997 period. Small differences in the two temperature reconstructions suggest that a model is robust to sampling variability in the tree-ring and climate data.

Results

Model Diagnostics

Prior to comparing temperature reconstructions across hierarchical model variants and other standardization techniques, we first examine whether basic modeling assumptions

are being met under the hierarchical models. First, we calculate the time series of normalized residuals for each tree-ring series under each model variant. Normalized residuals are defined as the residuals adjusted for any autocorrelation and divided by their standard deviation, $\zeta_{i,t} = \left(\left(\varepsilon_{i,t} - \alpha \varepsilon_{i,t-1} \right) \Big/ \sigma_i \right)$. Note that $\alpha = 0$ for all models but M6 and $\sigma_i = \sigma$ in M4. The normality of normalized tree-ring residuals is tested using a Shapiro-Wilk test (Shapiro and Wilk, 1965) and homoscedasticity (i.e., constant variance) is tested with a Breusch–Pagan test (Breusch and Pagan, 1979). We also examine the autocorrelation of the normalized residuals and the mean squared error of the original residuals.

Model diagnostics are presented in Figure 2. Somewhat surprisingly, the residual series for many trees under all models fail the test for normality – between 40-45% of trees for most models have p-values from the Shapiro-Wilk test below 0.05. This number increases to 67% for M1. For most models, the lack of normality stems from both positive and negative outlier values that stretch the tails of the distribution for tree-ring residuals beyond the kurtosis of a normal distribution. This can be effectively solved by modeling the residuals using Student's t-distribution and fitting the degrees of freedom for each tree series, but the temperature reconstructions from such an approach are nearly identical to those from a model with normal residuals (not shown). This suggests that the influence of residual outliers for individual tree-ring series is damped when a sufficient number of trees are included in the analysis, and therefore the Student's t model was not considered further in this work. We note that the assumptions of normality fail more often under M1 mainly because the logarithmic transformation applied in that model

over-corrects for the skew in the rings and imparts a negative skew to most series (see supporting material). This cannot be solved using a Student's t-distribution, and the negative skew may impart errors in the temperature reconstruction – this is addressed later.

The Breusch–Pagan test fails to reject the assumption of homoscedastic residuals for most ring width series and all models, suggesting that after the rings are transformed using a Box-Cox or logarithmic function, the variance of the rings does not vary substantially with their magnitude. All models have substantial autocorrelation in the residuals of tree-ring series, except M6, where this autocorrelation was explicitly modeled. Correspondingly, M6 also has substantially lower mean squared error for the residuals, since the modeled autocorrelation explains a substantial portion of the variability in tree-ring width series. We examine whether these improvements translate into improved temperature reconstructions below.

Performance of Temperature Reconstructions

Figure 3 shows the *RSQ*, *RE*, *CE*, and *CRPSS* for all models and the Tornedalen series calculated over the calibration and validation periods for both combinations of fitting and testing periods. In addition to the suite of hierarchical models, we also include the three other standardization approaches for comparison. Several insights emerge from Figure 3. The difference in performance is most apparent when comparing the TS and SF models to all other models. For out-of-sample performance, the TS method consistently performs the worst of all the models. The SF approach provides improvements over the TS

method, suggesting that the iterative approach reduces the amount of climate information subsumed by the age models. However, the SF method based on the negative exponential curve is far inferior to the RCS-SF method, which performs more like the hierarchical models than the other conventional approaches. This is because conventional SF is still highly susceptible to the segment-length curse, which inevitably leads to the loss of low-frequency signal compared to RCS-SF. These results indicate that for this case study, the heterogeneity in growth among trees is less important compared to the risk of climate information being subsumed into the age model. We also note that Schofield et al. (2016) found poor out-of-sample performance for a simple RCS standardization approach, despite previous studies that determined the RCS approach was well-suited for the study region (Briffa et al., 1992; Melvin et al., 2012). Our results indicate that a RCS-SF standardization leads to relatively accurate cross-validated temperature reconstructions, as compared to other models.

The range of *RSQ* values for all the models considered here is similar to those developed for other summer temperature series in the region (Grudd et al., 2002). We note that the differences between models are much more stark when comparing performance metrics that account for bias (*RE*, *CE*, and *CRPSS*) than for the *RSQ*. This indicates that much of the climate information that gets subsumed into the age models is related to the long-term mean and not year-to-year fluctuations, which is not particularly surprising given the smooth form of the age models. We also note that the differences in performance for TS and SF are much less apparent when examining within-sample performance, which highlights the importance of a robust cross-validation framework for model evaluation.

Differences in performance amongst the other hierarchical models and the RCS-SF approach are subtler, initially suggesting that the assumptions underlying the different model variants are less critical for temperature reconstructions, at least in this study. There are some modest differences to note. First, M3, M4, M5, and especially M6 generally underperform M0, M1, and M2 for out-of-sample periods. Therefore, it would appear that the use of partial pooling for the age model, homogenous variance, and the Hugershoff age curve do not substantively improve model performance, while auto-correlative structure for tree-ring series residuals significantly degrades performance. These results are modest (except for M6) and likely sensitive to sampling variability in the cross-validated skill statistics. Still, they may provide some guidance for future model development, especially since previous hierarchical modeling studies have not discussed the implications of modeling residual autocorrelation and have defaulted to the use of partial pooling and homogeneity of error variance (Duncan et al., 2010; Schofield et al., 2016).

Importantly, some of the major differences between models seen using the Tornedalen data are not apparent if the cross-validation testing is performed on the shorter Abisko data. Table 2 shows the verification *CE* values for all models for both the Abisko and Tornedalen temperature series. Using the Abisko data, it is very difficult to distinguish any of the models. The largest differences are seen for M4 and M1 in the 1913-1955 verification period, suggesting some degradation in performance when using a model with homogenous variance and a logarithmic transformation. However, most *CE* values

for the Abisko series are extremely close and likely indistinguishable from sampling variability, including those for the conventional TS and SF models. The cross-validated skill scores presented in Schofield et al. (2016) were similarly unable to differentiate a TS-type model from their primary hierarchical model (the same as M2 in this study). The differences between the TS and SF approaches and all other models are much clearer when using the longer Tornedalen series. Here, the *CE* values for the TS and SF models, as well as M6, are significantly lower than the other models, suggesting that these models are poor candidates to produce a final temperature reconstruction.

While the longer Tornedalen series highlights clear improvements when using some models over others, not all models are easily distinguished, even with the longer cross-validation period. For instance, the similarities in performance between M0, M1, and M2 would initially suggest that data transformation and the inclusion of climate into the standardization model have little effect on the climate reconstruction. Also, the similar cross-validated performance of the RCS-SF approach raises doubts about whether the hierarchical models provide substantially different reconstructions compared to a more conventional method, at least in this case study. However, these comparisons only include data for reconstructed temperatures during the period of instrumental data (1816-1997); a thorough comparison requires the examination of the entire 500-year period.

Figures 4 and 5 show the reconstructed Tornedalen temperature series for 1497-1997 based on the fitting periods 1816-1912 and 1913-1997, respectively. The mean reconstruction and 95% confidence bounds are shown along the main diagonal, and

differences between mean reconstructions of model pairs are shown in off-diagonal positions. For brevity, we only show and discuss the reconstructions for M0, M1, M2, and RCS-SF, since these three hierarchical models were particularly difficult to distinguish using performance metrics over the instrumental period, and the RCS-SF approach provided similar performance using a conventional standardization technique. However, we show comparisons between the reconstructed temperature series of all hierarchical models in the supporting material, which are generally very similar, in line with their similar performance statistics. The exceptions are M1 and M6, but we only focus on M1 in Figures 4 and 5 because M6 was already shown to have poor predictive skill.

When comparing the full reconstructions, the differences between M0, M1, M2, and RCS-SF are better resolved. First, the reconstructions of M0 and M2 are essentially identical for both fitting periods, suggesting that there really isn't much difference in the reconstruction if the climate series is modeled simultaneously with the tree rings. The same result is seen when the models are fit to the shorter Abisko data (not shown). This result is somewhat contrary to the thesis presented in Schofield et al., (2016), which argued strongly for simultaneous standardization and reconstruction. However, the result is perhaps not too surprising given that M0 and M2 have identical priors and extremely similar likelihood functions, differing only by the addition of the temperature data in M2 (~ 90 data points, or ~0.2% of the total data in the model). If multiple climate series were included in M2 and not just a single temperature series, then it is possible that there

would be greater differences between these two models. However, this is beyond the scope of this study and left for future work.

Figures 4 and 5 also show that the temperature reconstruction differs substantially under M1 compared to M0 and M2, although these differences are mainly seen prior to 1800. In M1, the cold anomaly centered around 1600 and extending from 1500 to 1750 is deeper than in M0 and M2. We argue that the anomalously cold reconstruction in M1 is actually a spurious artifact of the logarithmic transformation of tree rings that over-corrects for the positive skew in the data. After logarithmic transformation, many of the ring series have negative skew, and this skew imparts downward bias on either the age models or the regression linking the chronology to the temperature series. The lack of fit is also seen in Figure 2, where M1 exhibits the most egregious failure of the normality test and the highest MSE for individual tree-ring series (as compared to M0 and M2). We also note that some of the models in Schofield et al. (2016) produced implausibly cold growing season temperatures (<4°C) (Körner and Paulsen, 2004; Körner, 2008) in their reconstructions for the Abisko site, which we also found when using M1 for the Abisko data, but which was resolved using M0 (see supporting information). This further suggests that the logarithmic transformation imposes an artificial cold bias in the reconstruction.

The difference in the temperature reconstruction between the RCS-SF approach and both M0 and M2 is reversed as compared to the differences seen for M1. That is, the RCS-SF approach simulates a milder cold anomaly centered around 1600 as compared to M0 and

M2, but has a similar reconstruction after 1700. We note that the hierarchical model that used a Hugershoff curve (M5) utilizes a common curvature across tree-ring series (common parameters *b* and *c* across series), which is similar to an RCS-style approach that shares the same age model across trees. Therefore, the differences that originate between RCS-SF and M0 and M2 (but not between M5 and M0 and M2) are unlikely associated with the homogeneity of the age model in the RCS-SF approach. Rather, we speculate that the differences between the RCS-SF approach and both M0 and M2 are most likely linked to a signal in the ring widths early in the record (< 1700) that is attributed to the age model in early iterations of the signal-free fitting process for the RCS-SF approach, but is attributed to the chronology in M0 and M2. It is difficult to determine whether this signal is biological or climate related and which model is correct in its attribution. However, it is significant that the RCS-SF and hierarchical models differ substantially in their reconstruction, since both model types perform similarly under-cross validation and are constructed to avoid subsuming the climate signal into the age related trend.

Finally, we note that the differences between the models are more apparent for the 1913-1997 fitting period (Figure 5) compared to the 1816-1912 fitting period (Figure 4). To explore this further, Table 3 shows the average and mean square of year-by-year differences in mean temperature reconstruction based on the two fitting periods for all of the models considered. If the mean square differences between reconstructions for a model are small and the average difference is near zero, this indicates that the model is relatively insensitive to the sampling variability of the two different fitting periods. We

argue that a model that is highly sensitive to the fitting period used is vulnerable to over-fitting and may lead to degraded climate predictions in the pre-instrumental period. Table 3 shows that reconstructions based on the two fitting periods are most similar for RCS-SF, M0, and M2, although the RCS-SF approach has a slight mean bias between the two periods. Models M3 and M5 also have very similar reconstructions, and to a lesser extent M4 as well. Differences become more pronounced for M1 and M6, suggesting that the logarithmic transformation and auto-correlative structure are either altogether inappropriate for the data being considered (M1) or are too uncertain to be estimated accurately without interfering with the climate reconstruction (M6). Importantly, the largest differences between reconstructions occur for two of the conventional standardization techniques (TS and SF), indicating that these methods are the most sensitive to sampling variability. Overall, the results in Table 3, taken in tandem with the previous comparisons, suggest that the reconstructions from M0, M2, and RCS-SF provide robust, albeit somewhat competing, representations of temperature variability in northern Sweden over the last 500 years.

Discussion and Conclusion

This study has presented an exploration of hierarchical regression models for use in tree-ring standardization and chronology development. A series of hierarchical regression models were proposed to illustrate how one could easily test a variety of choices of model design. These models were tested against three more conventional standardization approaches in a cross-validated framework for the reconstruction of summertime temperatures in northern Sweden. Compared to the TS and SF approaches, the results of

the study show that hierarchical models are better suited to isolate climate-related variability from non-climatic variability in an ensemble of tree-ring series, leading to improved out-of-sample climate reconstructions that are more stable under sampling variability. The RCS-SF approach used here performed much better than both TS and SF techniques under cross-validation, and similarly to the hierarchical models. However, the RCS-SF approach led to substantially different temperature reconstructions compared to the hierarchical models early in the pre-instrumental period.

The RCS-SF approach, while promising in this application, may be limited for settings that exhibit more heterogeneity in growth across trees (Esper et al., 2002; Briffa and Melvin, 2011). The hierarchical models presented here are flexible enough to mitigate issues of heterogeneity, similar to TS and SF approaches, while also circumventing the segment length curse, which is the primary advantage of RCS procedures over individual curve fitting approaches. Additional work is needed to compare hierarchical models to the RCS-SF approach in heterogeneous growth areas and also more recent advances in RCS curve fitting, e.g., multiple RCS curves fit using signal free methods (Helma et al., 2016). We also suggest further testing of the hierarchical models on additional hydroclimate series that exhibit stronger relationships with the chronologies than those seen in this study to determine whether a similar model ranking holds in such cases. These matters will be investigated in future publications. Nevertheless, solely on the evidence of the results presented here we believe that the dendroclimatic community should consider further exploration of hierarchical models as a possible standard method

for the development of tree-ring-based chronologies for environmental, more specifically, climatological interpretation.

The hierarchical models presented in this work were developed in a Bayesian framework to enable greater flexibility in model design and promote uncertainty propagation across all model parameters during the model fitting process. However, Bayesian methods are slower to fit compared to conventional methods, as indicated earlier. The long calibration time could be a constraint for some practitioners interested in developing chronologies from different subsets of trees using a hierarchical modeling approach. More generally, the hierarchical models presented in this work may be difficult for practitioners to adopt if they are less familiar with Bayesian methods. We note that a Bayesian approach is not necessary for hierarchical modeling and direct the reader to Duncan et al. (2010) for details on the faster mixed-effects modeling framework used in that study. For those readers interested in the Bayesian models presented in this work, code written in the *R* statistical modeling environment is available in the supplemental material to support model development and testing.

The comparison between hierarchical models also provides insight and guidance for future model development. The results of the study show that certain modeling choices, such as partial pooling of the age model, homogeneity of variance across trees, flexible age models, residual autocorrelation, and overly corrective data transformations, can alter climate reconstructions to varying degrees, sometimes leading to significant degradation. In some cases, as for the logarithmic transform, this degradation is not obvious from an

out-of-sample cross-validation and can only be uncovered if the analyst checks for specific violations of underlying assumptions (in this case, normality). Conversely, other modeling choices that appear appropriate for the tree-ring series, such as autoregressive modeling, can have severe repercussions for the fidelity of climate reconstructions. These lessons suggest that modeling assumptions need to be tested against simpler models and adopted only if there is evidence to support the additional complexity, for instead through a vigorous cross-validation framework.

Still, an assessment of model assumptions and a comparison of skill statistics under cross-validation may not definitively identify which models most accurately recover the true climate signal. These methods are useful for screening out poorly performing models, but it can be difficult if not impossible to discriminate between similarly performing models given the degree of sampling variability in the cross-validated statistics. This was the case for several of the models in this study using the longer Tornedalen data, and essentially all models using the Abisko data. In such cases, parsimony is a useful guide for model selection. Numerical experiments could also be used to test how well different families of models recover an underlying, true climate signal in synthetic examples that span a range of data generation processes. We leave such experiments for future work.

The comparison of M0 and M2 showed that explicitly modeling the climate series during standardization did not substantively improve the reconstructions, contrary to the conclusions of Schofield et al. (2016). However, the climate data were dominated by an

extensive tree-ring dataset in this study, and therefore the addition of the climate information in M2 likely had little influence on the likelihood function during model fitting. This may not be the case for data collected in regions with sparser tree cover, or if multiple temperature series are included in the model, since temperature data would constitute a larger percentage of data points in the likelihood function. In this case, M2 could be adapted to include separate regression coefficients linking the chronology to each temperature series. These coefficients could be partially pooled, which would be particularly beneficial if certain series are relatively short and temperature is reasonably homogenous across sites. The joint distribution of all the temperature data could also be modeled explicitly, accounting for the often high cross-correlation in temperature across sites. We suggest a more thorough comparison of M0 and M2 for these situations as the basis of future research investigating hierarchical models for dendroclimatic standardization and climate reconstruction.

Acknowledgements

We would like to acknowledge Tom Melvin for his feedback and helpful suggestions on this work. Phil Jones likewise contributed suggestions and comments on testing the Tornedalen temperature record for homogeneity. KRB acknowledges the ongoing medical support of Mr G. Kapur. Lamont-Doherty Earth Observatory contribution number xxxx.

References

Björklund, J.A., Gunnarson, B.E., Krusic, P.J., et al.,;1; 2013. Advances towards improved low-frequency tree-ring reconstructions, using an updated *Pinus sylvestris* L. MXD network from the Scandinavian Mountains. Theoretical and Applied Climatology 113: 697–710.

Bontemps, J.-D., Esper, J.,;1; 2011. Statistical modeling and RCS detrending methods provide similar estimates of long-term trend in radial growth of common beech in north-eastern France. Dendrochronologia 29, 99-107.

Bontemps, J.-D., Hervé, J.-C., Dhôte, J.-F.,;1; 2010. Dominant radial and height growth reveal comparable historical variations for common beech in north-eastern France. Forest Ecology and Management 259, 1455-1463.

Briffa, K. R., Jones, P. D., Bartholin, T. S., Eckstein, D., Schweingruber, F. H., Karlen, W., Zetterberg, P., Eronen, M.,;1; 1992. Fennoscandian summers from AD 500: temperature changes on short and long timescales. Climate Dynamics 7, 111–119.

Briffa, K.R., Jones, P.D., Schweingruber, F.H., Karle´ n, W., Shiyatov, S.G.,;1; 1996. Tree-ring variables as proxy-climate indicators: problems with low frequency signals. In: Jones, P.D., Bradley, R.S., Jouzel, J. (Eds.), Climatic Variations and Forcing Mechanisms of the Last 2000 Years. Springer, Berlin, pp. 9–41.

Briffa, K.R., Melvin, T.M.,;1; 2011. A closer look at Regional Curve Standardization of tree-ring records: Justification of the need, a warning of some pitfalls, and suggested improvements of its application. In: Hughes MK, Diaz HF and Swetnam TW (eds) Dendroclimatology: Progress and Prospects. Berlin: Springer Verlag, pp. 113–145.

Briffa, K.R., Melvin, T.M., Osborn, T.J. et al.,;1; 2013. Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, northwest Siberia. Quaternary Science Reviews 72, 83–107.

Briffa, K.R., Osborn, T.J., Schweingruber, F.H. et al.,;1; 2001. Low- frequency temperature variations from a northern tree ring density network. Journal of Geophysical Research 106, 2929–2941.

Briffa, K. R., Shishov, V. V., Melvin, T. M., Vaganov, E. A., Grudd, H., Hantemirov, R. M., Eronen, M., Naurzbaev, M. M.,;1; 2008. Trends in recent temperature and radial tree growth spanning 2000 years across northwest Eurasia. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 2269–2282.

Breusch, T. S., Pagan, A. R., ;1;1979. A simple test for heteroskedasticity and random coefficient variation. Econometrica 47 (5), 1287–1294.

Cook, E.R., Briffa, K.R., Meko, D.M., Graybill, D.A., Funkhouser, G.,;1; 1995. The segment length curse in long tree-ring chronology development for paleoclimatic studies. Holocene 5, 229–237.

Cook, E.R., Peters, K.,;1; 1997. Calculating unbiased tree-ring indices for the study of climatic and environmental change. Holocene 7, 359–368

Cook, E.R., Seager, R., Cane, M.A., Stahle, D.W.,;1; 2007. North American drought: Reconstructions, causes, and consequences. Earth Science Reviews 81, 93-134.

Duncan, R.P., Fenwick, P., Palmer, J.G., McGlone, M.S., Turney, C.S.M.,;1; 2010. Non-uniform interhemishperic temperature trends over the past 550 years. Climate Dynamics 35, 1429-1438.

Esper, J., Cook, E.R., Schweingruber, F.H.,;1; 2002. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. Science 295, 2250–2252.

Fritts, H. C.,;1; 1976. Tree Rings and Climate, Academic Press: London.

Fuller, W.A.,;1; 1987. Measurement Error Models, Wiley: New York.

Gelman, A., Rubin, D. B.,;1; 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7(4), 457–511.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin D. B.,;1; 2013. Bayesian Data Analysis, third edition. London: CRC Press.

Grudd, H., Briffa, K. R., Karlen, W., Bartholin, T. S., Jones, P. D., Kromer, B.,;1; 2002. A 7400-year tree-ring chronology in northern Swedish Lapland: natural climatic variability expressed on annual to millennial timescales. The Holocene 12, 657–665.

Helama, S., Lindholm, M., Timonen, M., and Eronen, M.,;1; 2004. Detection of climate signal in dendrochronological data analysis: a comparison of tree-ring standardization methods. Theoretical and Applied Climatology 79, 239-254.

Helama, S., Melvin, T.M., and Briffa, K.R.,;1; 2016. Regional curve standardization: State of the art. The Holocene, doi:10.1177/0959683616652709.

Kershaw, Z.L.;1; 2007. The Torneträsk tree-ring chronology: exploring the potential for biased climatereconstructions in recent millennia. Unpublished BSc thesis, University of East Anglia, Norwich, UK

Klingbjer, P., Moberg, A.,;1; 2003. A composite monthly temperature record from Tornedalen in northern Sweden, 1802-2002. International Journal of Climatology 23, 1465-1494.

Körner, C.,;1; 2008. Winter Crop Growth at Low Temperature May Hold the Answer for Alpine Treeline Formation, Plant Ecology & Diversity, 1, 3– 11.

Körner, C., and Paulsen, J.,;1; 2004. A World-Wide Study of High Altitude Treeline Temperatures, Journal of Biogeography, 31, 713– 732.

Linderholm, H.W., Björklund, J., Seftigen, K. et al.,;1; 2015. Fennoscandia revisited: A spatially improved tree-ring reconstruction of summer temperatures for the last 900 years. Climate Dynamics 45, 933–947.

Macias-Fauria, M., Grinsted, A., Helama, S., and Holopainen, J.,;1; 2012. Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia 30, 179-187.

Matskovsky, V.V., Helama, S.,;1; 2014. Testing long-term summer temperature reconstruction based on maximum density chronologies obtained by reanalysis of tree-ring data sets from northernmost Sweden and Finland. Climate of the Past 10, 1473–1487.

Melvin, T.M., Briffa, K.R.,;1; 2008. A “signal-free” approach to dendroclimatic standardization. Dendrochronologia 26, 71-86.

Melvin, T.M., Grudd, H., and Briffa, K.R., ;1;2012, Potential bias in ‘updating’ tree-ring chronologies using regional curve standardization: Re-processing 1500 years of Tornetrask density and ring-width data, The Holocene, 23(3), 364-373.

Melvin, T.M., Briffa, K.R.,;1; 2014a. CRUST: Software for the implementation of Regional Chronology Standardisation: Part 1. Signal-Free RCS. Dendrochronologia 32, 7-20.

Melvin, T.M., Briffa, K.R., ;1;2014b. CRUST: Software for the implementation of Regional Chronology Standardisation: Part 2. Further RCS options and recommendations. Dendrochronologia 32, 343–356.

Peters R.L., Groenendijk P., Vlam M., Zuidema, P.A.;1;, 2015. Detecting long- term growth trends using tree rings: A critical evaluation of methods. Global Change Biology 21, 2040–2054.

Plummer, M.,;1; 2011. Rjags: Bayesian graphical models using MCMC. Rpackage version 2.2.0-4, [Available at http://CRAN.R-project.org/package=rjags], RFoundation for Statistical Computing, Vienna, Austria.

Schofield, M.R., Barker, R.J., Gelman A., Cook, E.R., Briffa, K.R.,;1; 2016. A model-based approach to climate reconstruction using tree-ring data. Journal of the American Statistical Association, 111(513), 93-106.

Shapiro, S. S., Wilk, M. B.,;1; 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3–4), 591–611.

Warren, W. G.,;1; 1980. On removing the growth trend from dendrochronological data. Tree-Ring Bulletin, 40, 35–44.

Figures

Figure 1. a) Time periods of available ring width data for the 247 trees used in this study. b) Tornedalen (red) and Abisko (blue) temperature series, with mean levels for the Tornedalen series in the 1816-1912 and 1913-1997 periods shown by horizontal bars.

Figure 2. Diagnostics for hierarchical models, including tests for normality (Shapiro-Wilks) and heteroscedasticity (Breusch-Pagan), autocorrelation of normalized residuals, and mean squared error of the original residuals.

Figure 3. Cross-validation performance statistics for all models and each of the four metrics. Within-sample (out-of-sample) performance is shown on the main (off) diagonal in each subplot.

Figure 4. Temperature reconstructions for models M0, M1, M2, and RCS-SF based on the 1816-1912 fitting period. The mean reconstruction, 95% confidence bounds, and instrumental temperature data are shown along the diagonal. Differences between the mean reconstructions of different model pairs are shown in the off-diagonal positions. The average difference is indicated by a dashed red line.

Figure 5. Same as Figure 4, but for the 1913-1997 fitting period.

Tables

Table 1. Description and formulation for all hierarchical models considered in this study. For all models besides M0, the formulation highlights the differences from the reference model. Priors for all models can be found in the supporting material. The ^ annotation indicates the posterior median.

Description	Model Label	Formulation	Climate Reconstruction
Reference Model	M0	$y_{i,t} = \beta_{0,i} + \beta_{1,i} age_{i,t} + \eta_t + \varepsilon_{i,t}$ $\eta_t \sim N(0, \sigma_\eta^2)$ $\varepsilon_{i,t} \sim N(0, \sigma_i^2)$ $y_{i,t} = \text{BoxCox}(ring\ widths_{i,t})$	$x_t \sim N(\beta_x \hat{\eta_t}, \sigma_x^2)$ β_x and σ_x^2 estimated via OLS
Transformation	M1	$y_{i,t} = \log(ring\ widths_{i,t})$	Same as M0
Climate Link	M2	$\eta_t \sim N(\beta_2 x_t, \sigma_\eta^2)$ $x_t \sim N(\mu_x, \sigma_x^2)$	$x_t \sim N(\mu_{x,t}^*, \sigma_x^{2^*})$ $\mu_{x,t}^* = \frac{(\widehat{\sigma_\eta^2}\widehat{\mu_x} + \widehat{\sigma_x^2}\widehat{\beta_2}\hat{\eta}_t)}{\widehat{\sigma_\eta^2} + \widehat{\beta_2}^2 \widehat{\sigma_x^2}}$ $\sigma_x^{2^*} = \left(\frac{1}{\widehat{\sigma_x^2}} + \frac{\widehat{\beta_2}^2}{\widehat{\sigma_\eta^2}}\right)^{-1}$
Partial Pooling	M3	$\beta_{0,i} \sim N(\mu_{\beta_0}, \sigma_{\beta_0}^2)$ $\beta_{1,i} \sim N(\mu_{\beta_1}, \sigma_{\beta_1}^2)$ T(-∞,0)	Same as M0
Homogenous Variance	M4	$\varepsilon_{i,t} \sim N(0, \sigma^2)$	Same as M0
Hugershoff	M5	$y_{i,t} = a_i \times age_{i,t}^{b_i} e^{-c_i \times age_{i,t}} + d_i + \eta_t + \varepsilon_{i,t}$	Same as M0
Residual Autocorrelation	M6	$\varepsilon_{i,t} = \alpha \varepsilon_{i,t-1} + \zeta_{i,t}$ $\zeta_{i,t} \sim N(0, \sigma_i^2)$	Same as M0

Table 2. The verification period coefficient of efficiency (CE) for each model fit to the Abisko and Tornedalen series.

	Abisko		Tornedalen	
Model	1913-1955	1956-1997	1816-1912	1913-1997
M0	0.25	0.16	0.15	0.08
M1	0.2	0.13	0.13	0.1
M2	0.25	0.17	0.15	0.09
M3	0.26	0.17	0.13	0.05
M4	0.17	0.15	0.12	0.05
M5	0.26	0.17	0.14	0.06
M6	0.25	0.19	0.06	-0.03
TS	0.24	0.19	-0.14	-0.22
SF	0.23	0.2	-0.04	-0.13
RCS-SF	0.21	0.17	0.12	0.06

Table 3. The average and mean square of year-by-year differences between mean temperature reconstructions for the Tornedalen series based on the 1816-1912 and 1913-1997 fitting periods for each model.

Model	Average	Mean Square
M0	-0.01	31.70
M1	0.21	89.49
M2	-0.05	28.29
M3	-0.13	35.54
M4	0.03	42.73
M5	-0.09	33.18
M6	-0.31	75.03
TS	-0.58	196.13
SF	-0.46	141.15
RCS-SF	-0.15	27.15