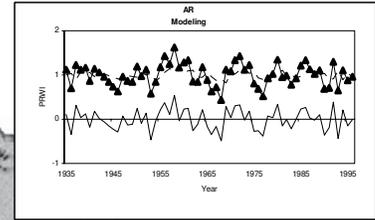
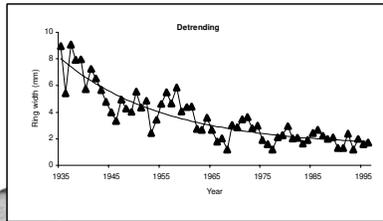
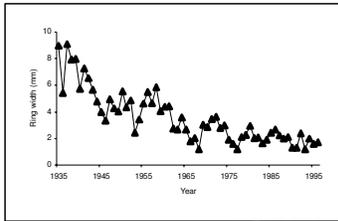




Coping with Multicollinearity: An Example on Application of Principal Components Regression in Dendroecology

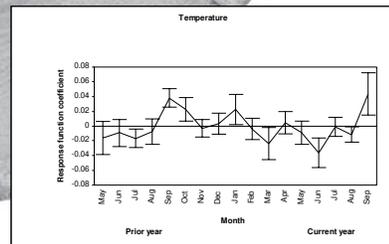
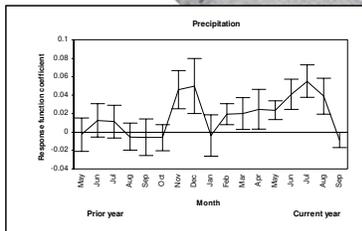
B. Desta Fekedulegn
J.J. Colbert
R.R. Hicks, Jr.
Michael E. Schuckers



$$\text{Radial Growth} = f \left[\text{Climate Variables} \right] + \text{Error}$$

Principal Components Regression

Response Function



Abstract

The theory and application of principal components regression, a method for coping with multicollinearity among independent variables in analyzing ecological data, is exhibited in detail. A concrete example of the complex procedures that must be carried out in developing a diagnostic growth-climate model is provided. We use tree radial increment data taken from breast height as the dependent variable and climatic data from the area as the independent data. Thirty-four monthly temperature and precipitation measurements are used as potential predictors of annual growth. Included are monthly average temperatures and total monthly precipitation for the current and past growing season. The underlying theory and detail illustration of the computational procedures provide the reader with the ability to apply this methodology to other situations where multicollinearity exists. Comparison of the principal component selection rules is shown to significantly influence the regression results. A complete derivation of the method used to estimate standard errors of the principal component estimators is provided. The appropriate test statistic, which does not depend on the selection rule, is discussed. The means to recognize and adjust for autocorrelation in the dependent data is also considered in detail. Appendices and directions to internet-based example data and codes provide the user with the ability to examine the code and example output and produce similar results.

The Authors

B. DESTA FEKEDULEGN is a research associate at Department of Statistics, West Virginia University. He received a M.S. in forest biometrics at University College Dublin and a M.S. in statistics at West Virginia University. His Ph.D. was in forest resource science (dendrochronology) from West Virginia University. His research interests include analytic methods in dendrochronology and forestry, nonlinear modeling, biased estimation, and working on applied statistics.

J.J. COLBERT is a research mathematician with the Northeastern Research Station, USDA Forest Service. He received an M.S. in mathematics from Idaho State University and a Ph.D. in mathematics from Washington State University. His primary research interests include the modeling of forest ecosystem processes and integration of effects of exogenous inputs on forest stand dynamics.

R.R. HICKS, JR. is professor of forestry at West Virginia University. He received his BSF and MS degrees from the University of Georgia and Ph.D from the State University of New York at Syracuse. He was assistant and associate professor of forestry at Stephen F. Austin State University in Texas from 1970 to 1978. Since that time he has been associate professor and professor of forestry at West Virginia University. He currently coordinates the Forest Resources Management program at WVU.

MICHAEL E. SCHUCKERS is currently an assistant professor at the Department of Statistics, West Virginia University. He received an A.M. in statistics from the University of Michigan and a Ph.D. in statistics from Iowa State University. His primary research interests include Bayesian methodology and statistical methods for biometric devices.

Manuscript received for publication 5 May 2002

Published by:
USDA FOREST SERVICE
11 CAMPUS BLVD SUITE 200
NEWTOWN SQUARE PA 19073-3294

September 2002

For additional copies:
USDA Forest Service
Publications Distribution
359 Main Road
Delaware, OH 43015-8640
Fax: (740)368-0152

Contents

Introduction	1
Statistical Method that Accounts for Multicollinearity	1
Interpreting Response Function	1
Developing an Appropriate Measure of Tree Growth	2
Objectives	2
Review of Methodologies	2
The Multiple Regression Model	2
Centering and Scaling	3
Standardizing	4
Principal Components Regression (PCR)	4
The Underlying Concept	4
Computational Technique	4
Elimination of Principal Components	6
Transformation Back to the Original Climatic Variables	7
Tree-ring and Climatic Data	8
Detrending and Autoregressive Modeling	8
Violation of the Two Assumptions on the Response Model	8
Transformations Applied to the Ring Width Measurements	11
Results and Discussion	18
Procedure for Estimating Response Function	18
Response Function Based on Centered and Scaled Climatic Variables	19
Response Function Based on Standardized Climatic Variables	21
Standard Errors of the Principal Component Estimators	22
Inference Techniques	24
Comparison with the Fritts Approach	26
Computational Comparison of Approaches	27
Response Function and Comparison of the Inferential Procedures	29
Sensitivity of the Response Function to Principal Components Selection Rules	31
Summary and Conclusions	34
Acknowledgment	34
Literature Cited	35
Appendix	36
A-SAS Program to Fit the Modified Negative Exponential Model	36
B-SAS Program to Fit an Autoregressive Model	38
C-SAS program to Perform Principal Components Regression	40

Introduction

Many ecological studies include the collection and use of data to investigate the relationship between a response variable and a set of explanatory factors (predictor variables). If the predictor variables are related to one another, then a situation commonly referred to as multicollinearity results. Then results from many analytic procedures (such as linear regression) become less reliable. In this paper we attempt to provide sufficient detail on a method used to alleviate problems associated with dependence or collinearity among predictor variables in ecological studies. These procedures are also applicable to any analysis where there may be reason to have concern for dependencies among continuous independent variables used in a study. In this study, response function analysis was carried out using monthly mean temperatures and monthly precipitation totals as independent variables affecting growth. A response function is a regression model used to diagnose the influence of climatic variables on the annual radial growth of trees. It is rarely used to predict tree growth.

When the independent variables show mild collinearity, coefficients of a response function may be estimated using the classical method of least squares. Because climatic variables are often highly intercorrelated (Guiot et al. 1982), use of ordinary least squares (OLS) to estimate the parameters of the response function results in instability and variability of the regression coefficients (Cook and Jacoby 1977). When the climatic variables exhibit multicollinearity, estimation of the coefficients using OLS may result in regression coefficients much larger than the physical or practical situation would deem reasonable (Draper and Smith 1981); coefficients that wildly fluctuate in sign and magnitude due to a small change in the dependent or independent variables; and coefficients with inflated standard errors that are consequently nonsignificant. More importantly, OLS inflates the percentage of variation in annual radial growth accounted for by climate (R^2_{climate}). Therefore, using ordinary regression procedures under high levels of correlation among the climatic variables affects the four characteristics of the model that are of major interest to dendroecologists: magnitude, sign, and standard error of the coefficients as well as R^2_{climate} .

Statistical Method that Accounts for Multicollinearity

Principal components regression is a technique to handle the problem of multicollinearity and produce stable and meaningful estimates for regression coefficients. Fritts et al. (1971) was the first to introduce the method of principal components regression (PCR) for estimating response functions in dendroecology. The estimators of the parameters in the response function, obtained after performing PCR, are referred to as principal component estimators (Gunst and Mason 1980). Fritts (1976) refers to the values of these estimators as elements of the response function.

The methodology of developing a radial growth response model using PCR as presented by Fritts et al. (1971), Fritts (1976), and Guiot et al. (1982) requires further clarifications and improvements. First, we introduce the distribution of the test statistic used for assessing the significance of the climatic variables. We present the inferential procedure that uses the test statistic given by Gunst and Mason (1980); but the original work was done by Mansfield et al. (1977). This test statistic tests the hypothesis that the parameters are zero using the principal component estimator of the coefficients. Second, we present a complete derivation and provide a formula for estimating standard error of the elements of response function. Third, the various principal component selection rules and their effects on characteristics of the response function are explored.

Interpreting Response Function

Information about the influence of climatic variables on tree radial growth is extracted from the sign, magnitude, and statistical significance of the elements of the response function. The sign indicates the direction of the relationship, the magnitude indicates the degree of influence, and the significance indicates whether the influence was due to chance or not. Detailed review on interpreting response function in dendroecology is given by Fritts (1976).

Developing an Appropriate Measure of Tree Growth

In addition to the problem of multicollinearity among the independent variables, the dependent variable, raw ring width, contains nonclimatic information related to tree size or age. Ring-width data also violate the two assumptions required to fit the proposed response function model: assumptions of independence and homogeneity of variance. For completeness, this study also briefly reviews the methods used to transform ring-width series so that the data satisfies these two assumptions.

Objectives

The objectives of this study are: to present a step-by-step procedure for estimating a response function using principal components regression; to provide a formula for estimating the standard errors of the principal component estimators of the coefficients of the independent variables; to introduce the appropriate test statistic for assessing the significance of the regression coefficients obtained using principal component regression; to explore the effects of the various methods of selecting principal components on characteristics of the response function; and to demonstrate the methods (detrending and autoregressive modeling) used to transform ring-width series to produce a growth measure that reflects the variation in climate.

Review of the Methodologies

The Multiple Regression Model

Consider dendroecological research in which the data consists of a tree-ring chronology (i.e., the response variable y) that spans n years and k climatic variables x_1, x_2, \dots, x_k . Assume that in the region of the x 's defined by the data, y is related approximately linearly to the climatic variables. The aim of response function analysis in dendroecology is to diagnose the influence of variation among input variables on the annual radial growth of trees using a model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (1)$$

where the response variable y is the standard or prewhitened tree-ring chronology, the independent variables x_1, x_2, \dots, x_k are monthly total precipitation and monthly mean temperature, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients to be estimated, n is the number of years, and ε_i is the i^{th} year model error, assumed uncorrelated from observation to observation, with mean zero and constant variance. Here y_i is a measure of tree growth at the i^{th} year, x_{ji} is the i^{th} year reading on the j^{th} climatic variable. In addition, for the purpose of testing hypotheses and calculating confidence intervals, it is assumed that ε_i is normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$. Using matrix notation, the model in Eq. 1 can be written:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdot & \cdot & \cdot & x_{k1} \\ 1 & x_{12} & x_{22} & \cdot & \cdot & \cdot & x_{k2} \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ 1 & x_{1n} & x_{2n} & \cdot & \cdot & \cdot & x_{kn} \end{bmatrix}_{n \times (k+1)} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}_{(k+1) \times 1} \quad \text{and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

The least squares estimator $\mathbf{b} = (b_0 \ b_1 \ b_2 \ \dots \ b_k)'$ of the regression coefficients of the climatic variables is (assuming \mathbf{X} is of full column rank) $\hat{\mathbf{b}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and the variance-covariance matrix of the estimated regression coefficients in vector \mathbf{b} is $\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (Draper and Smith 1981, Myers 1986). Each column of \mathbf{X} represents measurements for a particular climatic variable.

The multiple linear regression model in Equations 1 and 2 can be written in alternative forms by either centering and scaling or standardizing the independent variables. Such transformation of the climatic variables has special merit in dendroecology in that it allows results from different studies to be comparable. These methods are briefly discussed below.

Centering and Scaling. Suppose that the independent variables (each column of \mathbf{X}) are centered and scaled, i.e., x_{ji} , the i^{th} year measurement on the j^{th} climatic variable (x_j) in the natural units, is transformed into x_{ji}^* as follows:

$$x_{ji}^* = \frac{x_{ji} - \bar{x}_j}{s_j} \quad (3)$$

where $s_j = \sqrt{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2}$. The process of centering and scaling allows for an alternative formulation of Eq.1 as follows:

$$y_i = \beta_0^* + \beta_1^* \left[\frac{x_{1i} - \bar{x}_1}{s_1} \right] + \beta_2^* \left[\frac{x_{2i} - \bar{x}_2}{s_2} \right] + \dots + \beta_k^* \left[\frac{x_{ki} - \bar{x}_k}{s_k} \right] + \varepsilon_i \quad (4)$$

Consider the model formulation in Eq. 4. Separating the first column of ones (1) from the \mathbf{X} matrix results in the model form

$$\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \quad (5)$$

where, in this form, $\boldsymbol{\beta}^* = (\beta_1^* \ \beta_2^* \ \dots \ \beta_k^*)'$ is the vector of coefficients, apart from the intercept, and \mathbf{X}^* is then $n \times k$ matrix of centered and scaled independent variables. The notation $\mathbf{1}$ is used to denote an n -vector of ones. Centering and scaling makes $\mathbf{X}^* \mathbf{X}^*$ the $k \times k$ correlation matrix of the independent variables. Let the vector $\mathbf{b}^* = (b_1^* \ b_2^* \ \dots \ b_k^*)'$ be the least squares estimator of $\boldsymbol{\beta}^*$.

If a data set is used to fit the centered and scaled model of Eq. 4, one can obtain the estimated coefficients in the original model of Eq. 1 using the following transformation:

$$b_j = \frac{b_j^*}{s_j} \quad j = 1, 2, \dots, k \quad (6)$$

The estimate of the intercept is obtained by computing

$$b_0 = b_0^* - \frac{b_1^* \bar{x}_1}{s_1} - \frac{b_2^* \bar{x}_2}{s_2} - \dots - \frac{b_k^* \bar{x}_k}{s_k} \quad (7)$$

where b_j^* are estimates from the centered and scaled model of Eq. 4 and $b_0^* = \bar{y}$.

Standardizing. Consider the model in Eq. 1. Suppose the independent variables x_1, x_2, \dots, x_k are standardized as follows: x_{ji} is transformed into x_{ji}^s using

$$x_{ji}^s = \frac{x_{ji} - \bar{x}_j}{S_{x_j}} \quad (8)$$

where S_{x_j} is the standard deviation of the independent variable x_j and the superscript s indicates that the independent variables are standardized. The process of standardizing the independent variables allows for an alternative formulation of Eq. 1 as follows:

$$y_i = \beta_0^s + \beta_1^s \left[\frac{x_{1i} - \bar{x}_1}{S_{x_1}} \right] + \beta_2^s \left[\frac{x_{2i} - \bar{x}_2}{S_{x_2}} \right] + \dots + \beta_k^s \left[\frac{x_{ki} - \bar{x}_k}{S_{x_k}} \right] + \varepsilon_i \quad (9)$$

The model in Eq. 9 can be written in matrix form as:

$$\mathbf{y} = \beta_0^s \mathbf{1} + \mathbf{X}^s \boldsymbol{\beta}^s + \boldsymbol{\varepsilon} \quad (10)$$

where, in this form, $\boldsymbol{\beta}^s = (\beta_1^s \ \beta_2^s \ \dots \ \beta_k^s)'$ is the vector of regression coefficients, apart from the intercept, and \mathbf{X}^s is the $n \times k$ matrix of standardized independent variables.

Let $\mathbf{b}^s = (b_1^s \ b_2^s \ \dots \ b_k^s)'$ be the least squares estimator of $\boldsymbol{\beta}^s$. If a dataset is used to fit the standardized model in Eq. 9, then the estimate of the coefficients of the model of Eq. 1 can be obtained from the estimates of the coefficients for the standardized climatic variables using the following transformations:

$$b_j = \frac{b_j^s}{S_{x_j}}, j=1, 2, \dots, k \quad (11)$$

and

$$b_0 = b_0^s - \frac{b_1^s \bar{x}_1}{S_{x_1}} - \frac{b_2^s \bar{x}_2}{S_{x_2}} - \dots - \frac{b_k^s \bar{x}_k}{S_{x_k}} \quad (12)$$

Note that $S_j = \sqrt{n-1} \times S_{x_j}$, i.e., centering and scaling only differs from standardizing by the constant factor, $\sqrt{n-1}$. The above review indicates that it is always possible to move from one model formulation to another regardless of which model was used for the analysis.

Principal Components Regression (PCR)

The Underlying Concept. Principal components regression (PCR) is a method for combating multicollinearity and results in estimation and prediction better than ordinary least squares when used successfully (Draper and Smith 1981, Myers 1986). With this method, the original k climatic variables are transformed into a new set of orthogonal or uncorrelated variables called principal components of the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure then involves eliminating some of the principal components to effect a reduction in variance. After elimination of the least important principal components, a multiple regression analysis of the response variable against the reduced set of principal components is performed using ordinary least squares estimation (OLS). Because the principal components are orthogonal, they are pair-wise independent and hence OLS is appropriate. Once the regression coefficients for the reduced set of orthogonal variables have been calculated, they are mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables. These new coefficients are principal component estimators (Gunst and Mason 1980). In dendroecological literature, the values of these estimators are known as elements of the response function (Fritts 1976).

Computational Technique. Let \mathbf{X}^* be the centered and scaled $n \times k$ data matrix as given in Eq. 5. The $k \times k$ correlation matrix of the climatic variables is then $\mathbf{C} = \mathbf{X}^* \mathbf{X}^{*\prime}$. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the eigenvalues of the correlation matrix, and $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_k]$ be the $k \times k$ matrix consisting of the normalized eigenvectors associated with each eigenvalue. Note that the eigenvalues are the solutions of the determinant equation $|\mathbf{X}^* \mathbf{X}^{*\prime} - \lambda \mathbf{I}| = 0$ (Draper and Smith 1981), and associated with each eigenvalue, λ_j , is a vector, \mathbf{v}_j , that satisfies the set of homogeneous equations $(\mathbf{X}^* \mathbf{X}^{*\prime} - \lambda_j \mathbf{I}) \mathbf{v}_j = \mathbf{0}$.

The vectors, $\mathbf{v}_j = (v_{1j} \ v_{2j} \ \dots \ v_{kj})'$, are the normalized solutions such that $\mathbf{v}_j' \mathbf{v}_j = 1$ and $\mathbf{v}_j' \mathbf{v}_i = 0$ for $i \neq j$. That is, the eigenvectors have unit length and are orthogonal to one another. Hence the eigenvector matrix \mathbf{V} is orthonormal, i.e., $\mathbf{V} \mathbf{V}' = \mathbf{I}$.

Now consider the model formulation given in Eq. 5. That is, $\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$. Since $\mathbf{V} \mathbf{V}' = \mathbf{I}$ one can write the original regression model (Eq. 5) in the form

$$\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{X}^* \mathbf{V} \mathbf{V}' \boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \quad (13)$$

or

$$\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{Z} \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (14)$$

where $\mathbf{Z} = \mathbf{X}^* \mathbf{V}$ and $\boldsymbol{\alpha} = \mathbf{V}' \boldsymbol{\beta}^*$. \mathbf{Z} is an $n \times k$ matrix of principal components and $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_k)$ is a $k \times 1$ vector of new coefficients. The model formulation in Eq. 14 can be expanded as $y = \beta_0^* + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_k z_k + \varepsilon$, where z_1, z_2, \dots, z_k are the k new variables called principal components of the correlation matrix. Hence, the model formulation in Eq. 14 is nothing more than the regression of the response variable on the principal components, and the transformed data matrix \mathbf{Z} consists of the k principal components.

For the model in Eq. 14 the principal components are computed using:

$$\mathbf{Z} = \mathbf{X}^* \mathbf{V} \quad (15)$$

where \mathbf{X}^* is the $n \times k$ matrix of centered and scaled climatic variables without the column of ones, and \mathbf{V} is the $k \times k$ orthonormal matrix of eigenvectors. The principal components are orthogonal to each other, that is:

$$\mathbf{Z}' \mathbf{Z} = (\mathbf{X}^* \mathbf{V})' (\mathbf{X}^* \mathbf{V}) = \mathbf{V}' \mathbf{X}^{*\prime} \mathbf{X}^* \mathbf{V} = \mathbf{V}' \mathbf{C} \mathbf{V} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k) \quad (16)$$

Equation 16 shows that $\mathbf{z}_j' \mathbf{z}_j = \mathbf{v}_j' \mathbf{C} \mathbf{v}_j = \lambda_j$ and $\mathbf{z}_j' \mathbf{z}_i = 0$, $i \neq j$. From Eq. 15, one can see that the principal components are simply linear functions of the centered and scaled climatic variables and the coefficients of this linear combination are the eigenvectors. For example, the elements of the j^{th} principal component, z_j , are computed as follows:

$$z_j = v_{1j} x_1^* + v_{2j} x_2^* + \dots + v_{kj} x_k^* \quad (17)$$

where $v_{1j}, v_{2j}, \dots, v_{kj}$ are elements of the eigenvector associated with λ_j , and x_j^* 's are the centered and scaled climatic variables obtained using Eq. 3. Note that $\sum_{i=1}^n z_{ji} = 0$ and the sum of squares of the elements of z_j ($\sum_{i=1}^n z_{ji}^2$) is λ_j . Since $\sum_{j=1}^k \lambda_j = k$ then the total sum of squares, $\sum_{j=1}^k \left(\sum_{i=1}^n z_{ji}^2 \right)$, is k . z_j accounts for λ_j of the total variance.

If the response variable (y) is regressed against the k principal components using the model in Eq. 14, then the least squares estimator for the regression coefficients in vector \mathbf{a} is the vector $\hat{\mathbf{a}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ and the variance-covariance matrix of the estimated coefficients in vector $\hat{\mathbf{a}}$ is given by

$$\text{Var}(\hat{\mathbf{a}}) = \hat{\sigma}^2(\mathbf{Z}'\mathbf{Z})^{-1} = \hat{\sigma}^2 \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_k^{-1}) \quad (18)$$

If all of the k principal components are retained in the regression model of Eq. 14, then all that has been accomplished by the transformation is a rotation of the k original climatic variables.

Elimination of Principal Components. Even though the new variables are orthogonal, the same magnitude of variance is retained. But if multicollinearity is severe, there will be at least one small eigenvalue. An elimination of one or more principal components associated with the smallest eigenvalues will reduce the total variance in the model and thus produce an appreciably improved diagnostic or prediction model (Draper and Smith 1981, Myers 1986).

The principal component matrix \mathbf{Z} contains exactly the same information as the original centered and scaled climatic dataset (\mathbf{X}^*), except that the data are arranged into a set of new variables which are completely uncorrelated with one another and which can be ordered or ranked with respect to the magnitude of their eigenvalues (Draper and Smith 1981, Myers 1986). Note that \mathbf{z}_j corresponding to the largest λ_j accounts for the largest portion of the variation in the original data. Further \mathbf{z}_j 's explain smaller and smaller proportions, until all the variation is explained; that is, $\sum_{j=1}^k \lambda_j = k$. Thus, the \mathbf{z}_j are indexed so that $\lambda_1 > \lambda_2 > \dots > \lambda_k > 0$.

In regression model of Eq. 14 one does not use all the \mathbf{z} 's, but follows some selection rule. The property that makes PCR unique and more complex is that there is no universally agreed upon procedure in selecting the \mathbf{z}_j 's to be included in the reduced model of Eq. 14 (Draper and Smith 1981). Methods used to determine which and how many principal components should be removed to gain a substantial reduction in variance include:

- a. The strategy of elimination of principal components should be to begin by discarding the component associated with the smallest eigenvalue. The rationale is that the principal component with smallest eigenvalue is the least informative. Using this procedure, principal components are eliminated until the remaining components explain some pre-selected percentage of the total variance (for example, 85 percent or more). That is, one selects the set of largest r contributors (principal components), which first achieve

$$\frac{\sum_{j=1}^r \lambda_j}{k} > 0.85.$$

- b. Some researchers use the rule that only principal components associated with eigenvalues greater than 1.00 are of interest (Draper and Smith 1981). This method is often referred to as the "Kaiser-Gutman Rule" (Loehlin 1998).
- c. Others use the selection rule that keeps the first principal components whose combined eigenvalue product is greater than 1.00 (Guiot et al. 1982).
- d. A more objective statistical strategy is to treat the principal component reduction as if it were a standard variable screening problem. Since the principal components are orthogonal regressor variables, a reasonable criterion to control the order of reduction are the t -statistics given by

$$t = \frac{\hat{\alpha}_j}{S_{\hat{\alpha}_j}} = \frac{\hat{\alpha}_j \sqrt{\lambda_j}}{S} \quad (19)$$

where $S_{\hat{\alpha}_j}$ is the standard error (*s.e.*) of $\hat{\alpha}_j$ (Myers 1986). Recall that from Eq. 18 $Var(\hat{\alpha}_j) = S^2 \lambda_j^{-1}$, where $S^2 = \hat{\sigma}^2$ and hence, $s.e.(\hat{\alpha}_j) = S(\sqrt{\lambda_j})^{-1}$. In this procedure, t -values should be rank ordered and components should be considered for elimination beginning with the smallest t -value.

Suppose that some such selection rule results in elimination of r principal components, that is, the model in Eq. 14 will now use only $k - r$ components. Let us denote the reduced \mathbf{Z} matrix of Eq. 14 by \mathbf{Z}_{k-r} ($n \times (k - r)$ matrix). Let the reduced vector of coefficients ($\boldsymbol{\alpha}$) be $\boldsymbol{\alpha}_{k-r} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_{k-r})$. The reduced model, after elimination of r principal components, can be written as

$$\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{Z}_{k-r} \boldsymbol{\alpha}_{k-r} + \boldsymbol{\varepsilon}^\circ \quad (20)$$

The \circ symbol on $\boldsymbol{\varepsilon}$ is used simply to differentiate it from $\boldsymbol{\varepsilon}$ in Eq. 14, since they are not the same. But the predicted values and residuals of the model in Eq. 13 or 14 are the same as those in Eq. 1 or 2, 4 or 5, and 9 or 10. Note that:

$$\mathbf{Z}_{k-r} = \mathbf{X}^* \mathbf{V}_{k-r} \quad (21)$$

where $\mathbf{V}_{k-r} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_{k-r}]$ is a $k \times (k - r)$ matrix of eigenvectors associated with the retained eigenvalues or principal components.

The least squares procedure is then used to obtain a diagnostic or prediction equation for the response \mathbf{y} as a function of the selected \mathbf{z} 's; that is, fitting the model in Eq. 20 using ordinary least squares. Once the fitted equation is obtained in terms of the selected \mathbf{z} 's, it can be transformed back into a function of the original \mathbf{x} 's as described in the following sub-section.

Transformation Back to the Original Climatic Variables. Suppose with k variables and hence k principal components, $r < k$ components are eliminated. From Eq. 14, with the retention of all components, $\boldsymbol{\alpha} = \mathbf{V}' \boldsymbol{\beta}^*$, and the coefficients for the centered and scaled climatic variables are obtained as:

$$\boldsymbol{\beta}^* = \mathbf{V} \boldsymbol{\alpha} \quad (22)$$

If one eliminates r components and fits the model given in Eq. 20, the principal component estimators of the regression coefficients, in terms of the centered and scaled climatic variables for all k parameters of the model in Eq. 5, are given by (Gunst and Mason 1980, Myers 1986)

$$\mathbf{b}_{pc}^* = \mathbf{V}_{k-r} \hat{\boldsymbol{\alpha}}_{k-r} \quad (23)$$

$$\begin{bmatrix} b_{1,pc}^* \\ b_{2,pc}^* \\ \cdot \\ \cdot \\ b_{k,pc}^* \end{bmatrix} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdot \ \cdot \ \cdot \ \mathbf{v}_{k-r}] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \cdot \\ \cdot \\ \hat{\alpha}_{k-r} \end{bmatrix}$$

where \mathbf{V}_{k-r} is defined as in Eq. 21, $\hat{\boldsymbol{\alpha}}_{k-r}$ is the vector of estimated coefficients (apart from the intercept) in the model of Eq. 20, and \mathbf{b}_{pc}^* is a vector of estimated coefficients (apart from the intercept) of the parameters in vector $\boldsymbol{\beta}^*$ of Eq. 5. Note that the elements of \mathbf{b}_{pc}^* are principal component estimators of the coefficients of the centered and scaled climatic variables, and subscript

pc is simply used to denote that the estimators are principal component estimators rather than ordinary least squares estimators. Since the \mathbf{x} 's are centered and scaled, the estimate of the constant term (β_0^*) in the model of Eq. 5 and 20 is \bar{y} , that is, $\hat{\beta}_0^* = \bar{y}$.

Transformation to the coefficients of the natural climatic variables is done as follows: the principal component estimator, $\mathbf{b}_{pc} = (b_{0,pc} \quad b_{1,pc} \quad \dots \quad b_{k,pc})$, of β is

$$b_{j,pc} = \frac{b_{j,pc}^*}{s_j}, \quad j = 1, 2, \dots, k \quad (24)$$

and

$$b_{0,pc} = b_{0,pc}^* - \frac{b_{1,pc}^* \bar{x}_1}{s_1} - \frac{b_{2,pc}^* \bar{x}_2}{s_2} - \dots - \frac{b_{k,pc}^* \bar{x}_k}{s_k} \quad (25)$$

Tree-ring and Climatic Data

Tree-ring data from 38 dominant and codominant yellow-poplar (*Liriodendron tulipifera* L.) trees sampled at Coppers Rock Forest, 13 km east of Morgantown, WV (39°39'43" N, 79°45'28" W), were used. Sampled trees were on average 65 years old and the mean diameter at breast height was 38 cm. The climatic variables used to develop the response function were mean monthly temperature and total monthly precipitation for a 17-month period from May of the year preceding to September of the current year, for a total of 34 monthly climatic variables. The monthly climate data for Coopers Rock weather station were obtained from the National Climatic Data Center. When missing data were encountered, extrapolations were made using data from near by weather stations (Fekedulegn 2001).

An examination of the correlation matrix of the 34 variables revealed that there were 48 pairs of significant correlations among the climatic variables (21 of them were just between temperature variables, four between precipitation variables, and 23 were between temperature and precipitation variables). The smallest eigenvalue of the correlation matrix was 0.001. Small eigenvalues suggest problems of multicollinearity among the predictors. Having the tree-ring and climatic data, the main steps toward developing the response function are developing an appropriate measure of tree growth (from tree-ring data) followed by application of PCR.

Detrending and Autoregressive Modeling

Violation of the Two Assumptions on the Response Model. Using the raw ring-width measurements as the response variable (measure of tree growth) in the multiple regression model of Eq. 1 violates the two assumptions of the model: ring-width measurements are independent (uncorrelated), and have a constant (homogeneous) variance independent of time or age of the tree. Ring widths are time-series data that are recorded annually. Radial growth at year $t-1$ has a positive effect on radial growth for year t ; this characteristic violates the assumption of independence. In addition, the variability of ring width is a function of age and decreases with increasing age, a characteristic that violates the assumption of homogeneity of variance.

Figure 1 shows an example of ring-width pattern from selected yellow-poplar trees. The plots indicate that ring width decreases with increasing age or size of the tree, a characteristics of most trees from undisturbed forest environments. To illustrate the linear dependence or autocorrelation within the ring-width measurements, ring width of the current year is plotted against the ring width of the prior year in Figure 2. The scatter plot shows that there is a strong positive linear association between prior and current year's growth (violation of independence). Table 1 shows the first-order (r_1) and second-order (r_2) autocorrelation coefficients of the raw ring-width measurements for selected trees. For the entire 38 yellow-poplar trees sampled, r_1 ranged from 0.56 to 0.89 and r_2 ranged from 0.43 to 0.84. These autocorrelation coefficients were significant at 5 percent level. The sample data in Figure 2 and Table 1 demonstrate the fact that raw-ring widths are highly interdependent in that growth in the

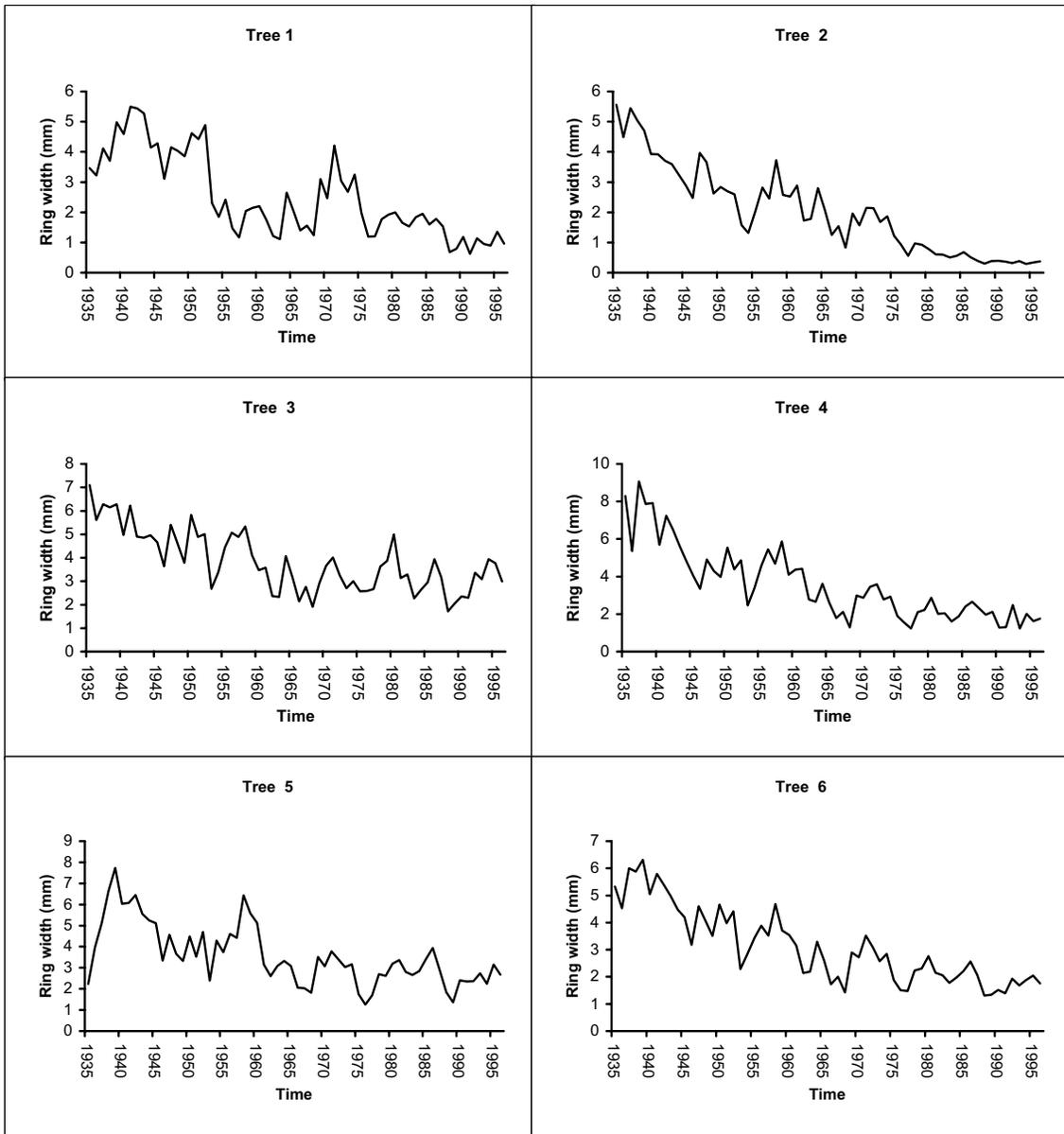


Figure 1.—Time-series plots of ring width showing a decreasing pattern with increasing time or age. This is a typical pattern of ring width of a tree grown in an open, disturbance-free environment.

prior year has a strong positive influence on current growth, a phenomenon commonly understood in dendrochronology.

To illustrate the violation of the assumption of homogeneity of variance, the ring-width series of selected trees was partitioned into 5-year segments. The standard deviation of the 5-year segments was plotted against age in Figure 3. The plots indicate that variability of ring width is a function of age and decreases with increasing tree age. To assess the significance of the decrease a linear trend line was fit and R^2 values were calculated. For all 38 trees analyzed, the values of R^2 varied from 0.01 to 0.89 with 74 percent of the samples having $R^2 \geq 0.5$.

It has been long known that raw ring-width series have sources of variation not related to climate (i.e., tree-size related long-term decreasing trend) and violate the assumptions of independence and homogeneity of variance. Hence, the raw ring width is not an appropriate response variable to be used in the multiple regression model of Eq. 1.

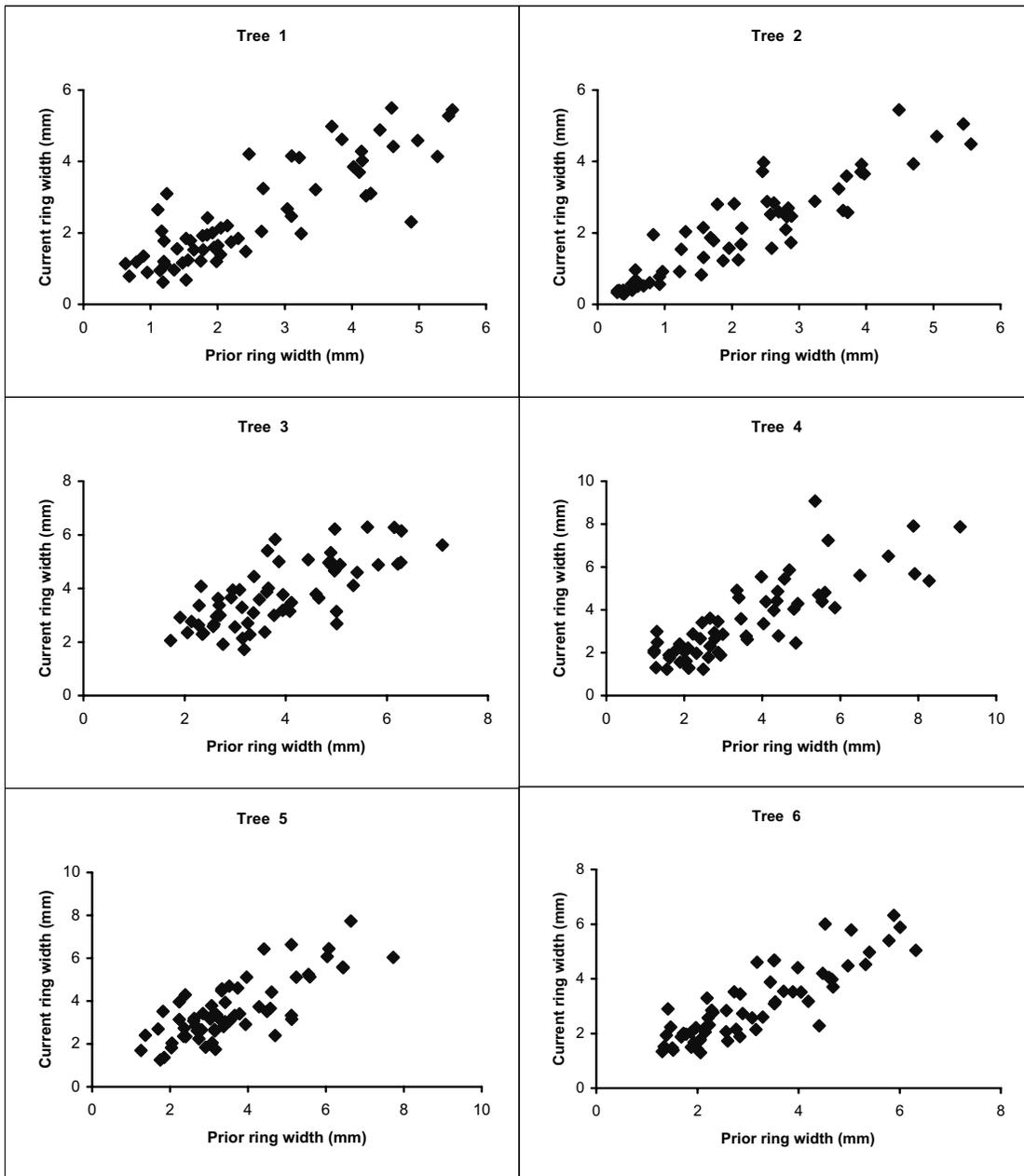


Figure 2.—Scatter plots of current-year ring width against prior-year ring width indicates strong autocorrelation (violation of independence).

Table 1.—First-order (r_1) and second-order (r_2) autocorrelation coefficients of raw ring-width measurements for selected yellow-poplar trees

Tree	r_1	r_2	Mean	Ste. dev.
			----- <i>mm</i> -----	
1	0.83	0.76	2.48	1.39
2	0.85	0.75	1.92	1.35
3	0.67	0.53	3.75	1.21
4	0.78	0.69	3.45	1.84
5	0.77	0.66	3.59	1.44
6	0.85	0.82	3.04	1.33

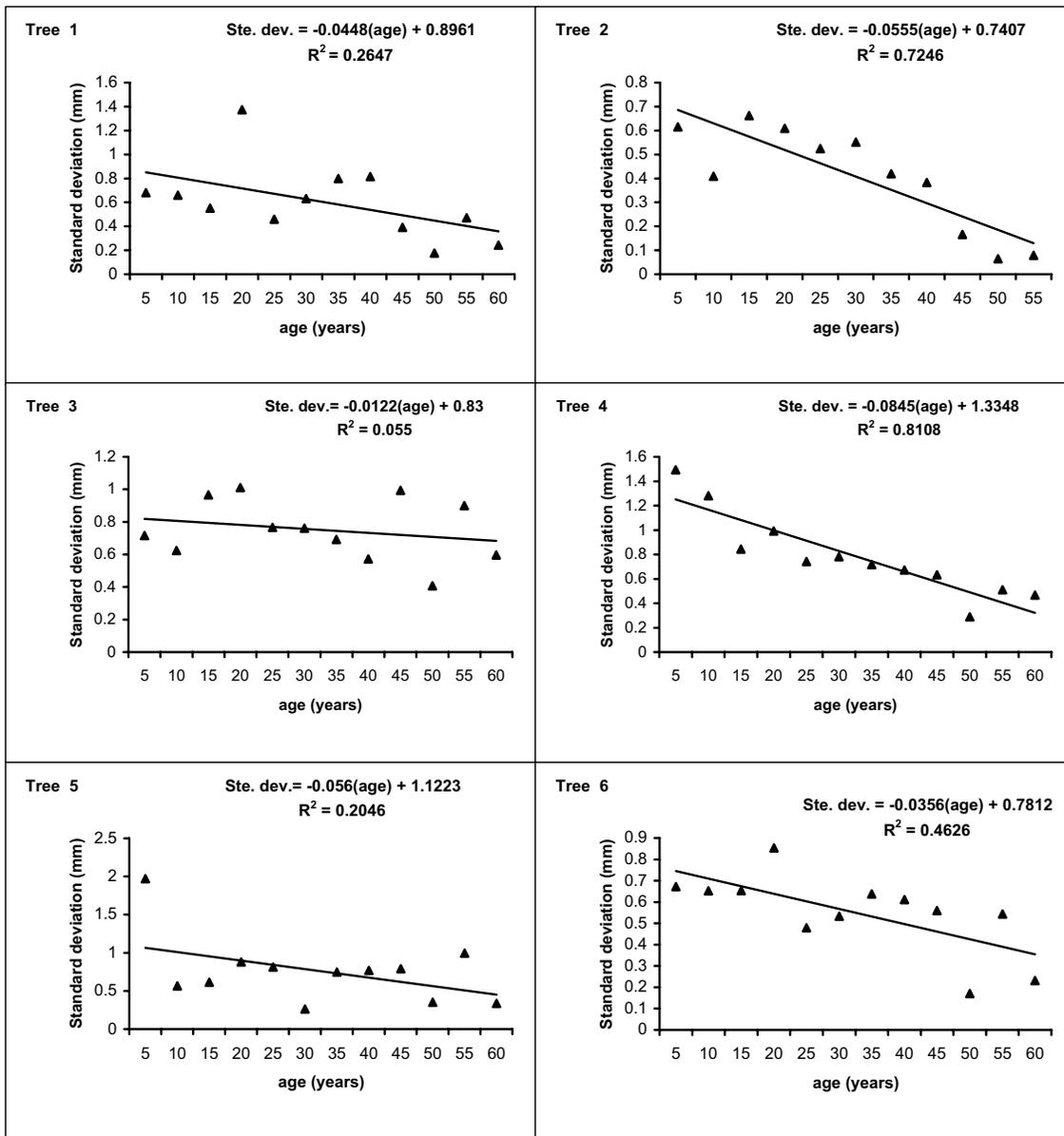


Figure 3.—Patterns of standard deviation of 5-year segments of the ring width series for selected trees. The variability of ring width decreases with increasing time or age (violation of homogeneity of variance).

Transformation Applied to the Raw Ring-Width Measurements

Removing the trend associated with tree-size (detrending). There are several methods for removing the long-term trend from the raw ring-width series (Fritts 1976, Cook 1985, Monserud 1986). But the choice of one detrending model over another depends on study objectives and the actual pattern of the tree-ring series. The choice of a detrending model affects the characteristic of the ring-width index (RWI) and results of growth-climate relations (Fekedulegn 2001).

A model for removing this age-related trend in ring-width series is the modified negative exponential model (Fritts 1976) that has the form

$$G_t = a \exp(-bt) + k \quad (26)$$

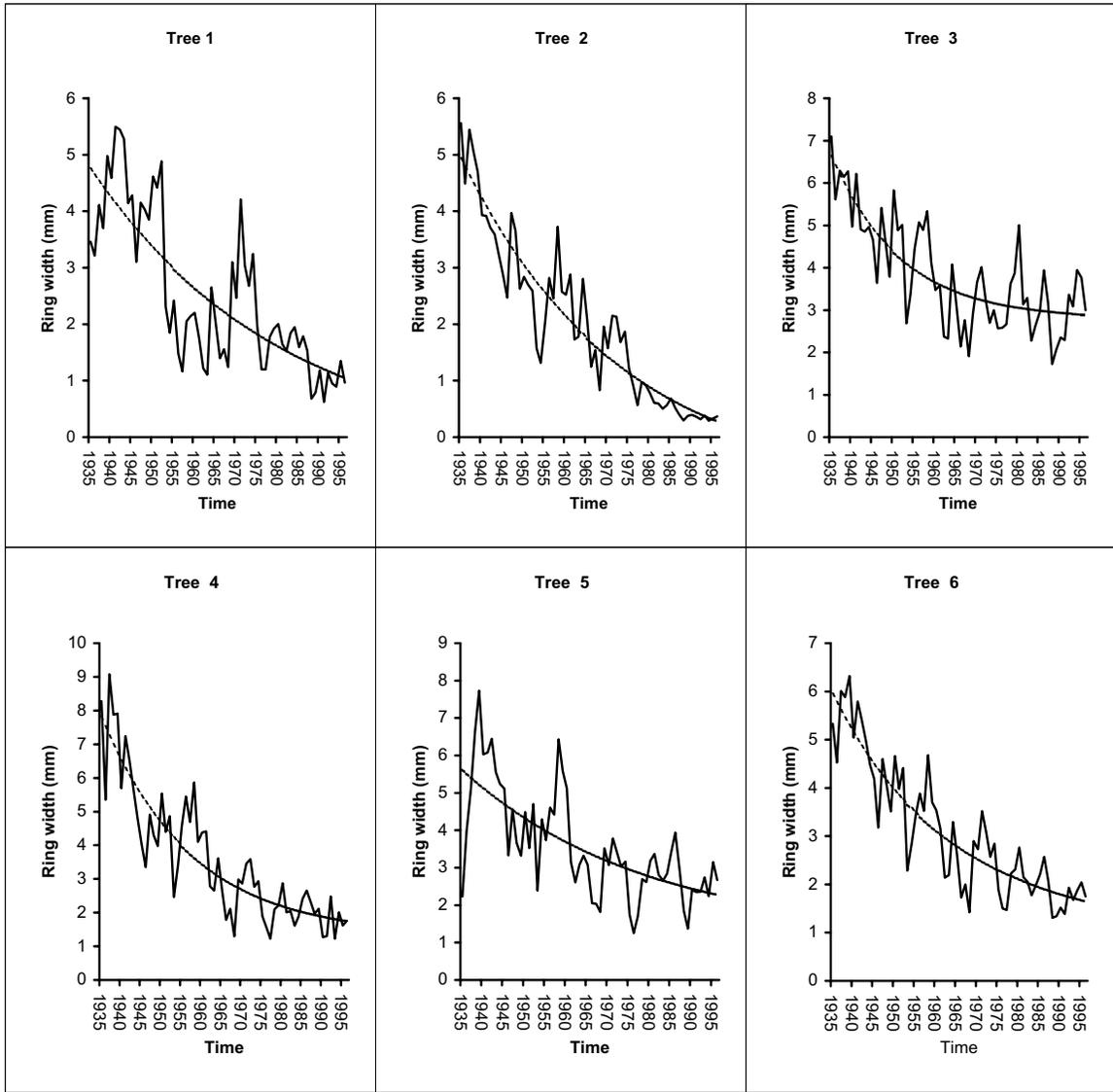


Figure 4.—Plots of the modified negative exponential model fitted to the ring-width series of selected trees.

where a , b , and k are coefficients to be estimated by least squares, t is age in years and G_t is the value of the fitted curve at time t . Detrending is accomplished by fitting the model in Eq. 26 to the raw ring-width series for a tree and calculating the detrended series (RWI) as ratios of actual (R_t) to fitted values (G_t). That is, the RWI (I_t) at time t is

$$I_t = \frac{R_t}{G_t} \quad (27)$$

The modified negative exponential model (MNEM) was fitted to each raw ring-width series for the 38 trees using PROC NLIN in SAS (Appendix A). However, for the few series where the MNEM model did not converge, a cubic smoothing-spline was used. Examples of fitted models are shown in Figure 4. Parameter estimates and the R^2 values are given in Table 2. Figure 5 shows plots of the RWI series. These plots indicate that, unlike the raw ring-width measurements, the RWI series does not exhibit a decreasing trend with increasing age, i.e., the age-related trend is removed. In addition, detrending has the added advantage of stabilizing the variance of RWI over time. To illustrate this

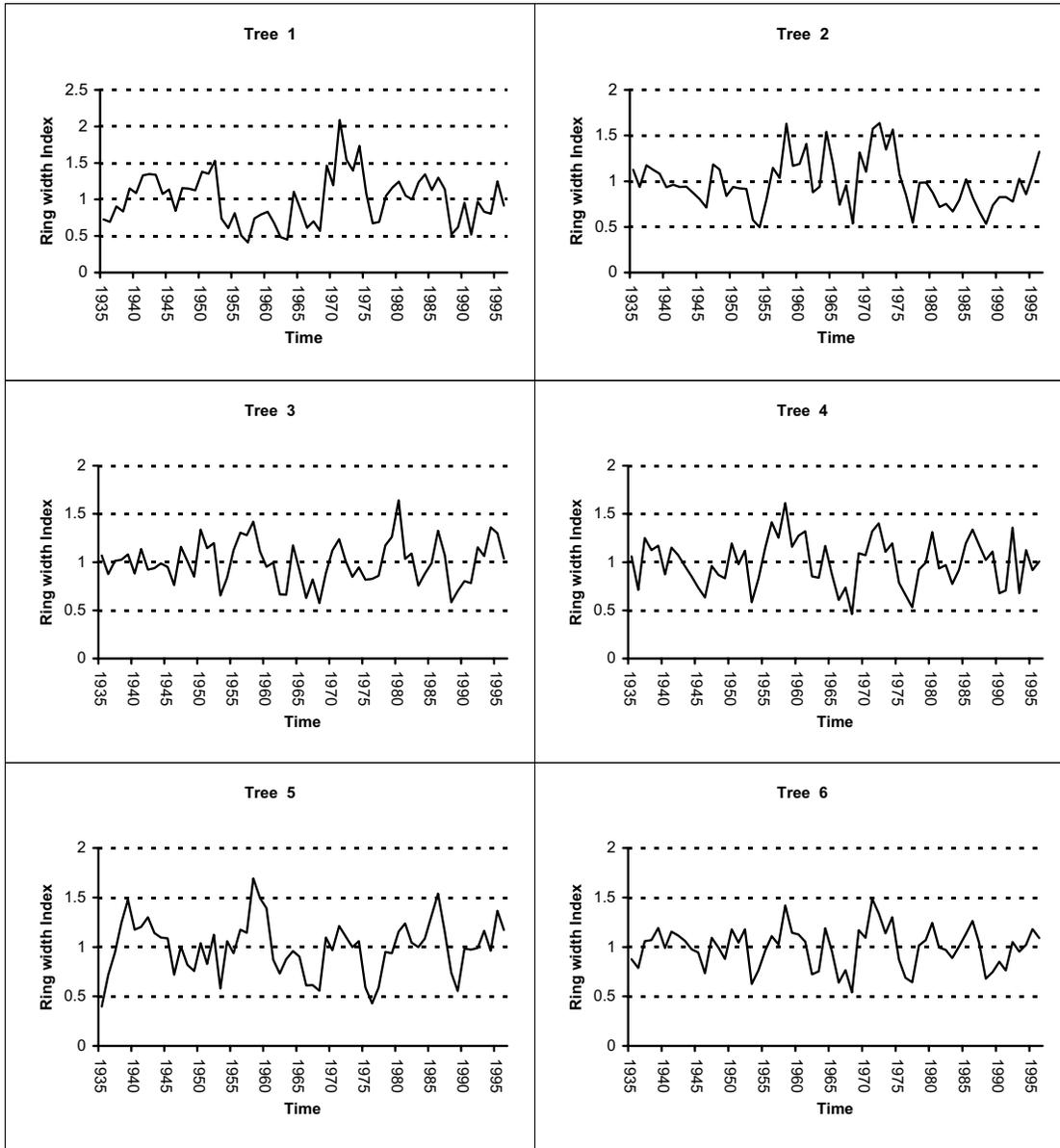


Figure 5.—Patterns of ring-width index (RWI) for selected trees. The plots indicate that detrending has removed the age-related trend. However, the plots also indicate that the values of RWI series are interdependent, i.e., high values follow high values and low values follow low values, a violation of the assumption of independence.

Table 2.—Estimated parameters of modified negative exponential model fitted to raw ring-width series of selected trees and percentage of variability in ring width explained by fitted models

Tree	<i>a</i>	<i>b</i>	<i>k</i>	R ²
				<i>percent</i>
1	5.14	0.0217	-0.286	91
2	5.97	0.0264	-0.878	96
3	4.08	0.0596	2.788	96
4	6.79	0.0449	1.329	95
5	4.53	0.0231	1.198	93
6	4.32	0.0532	2.648	94

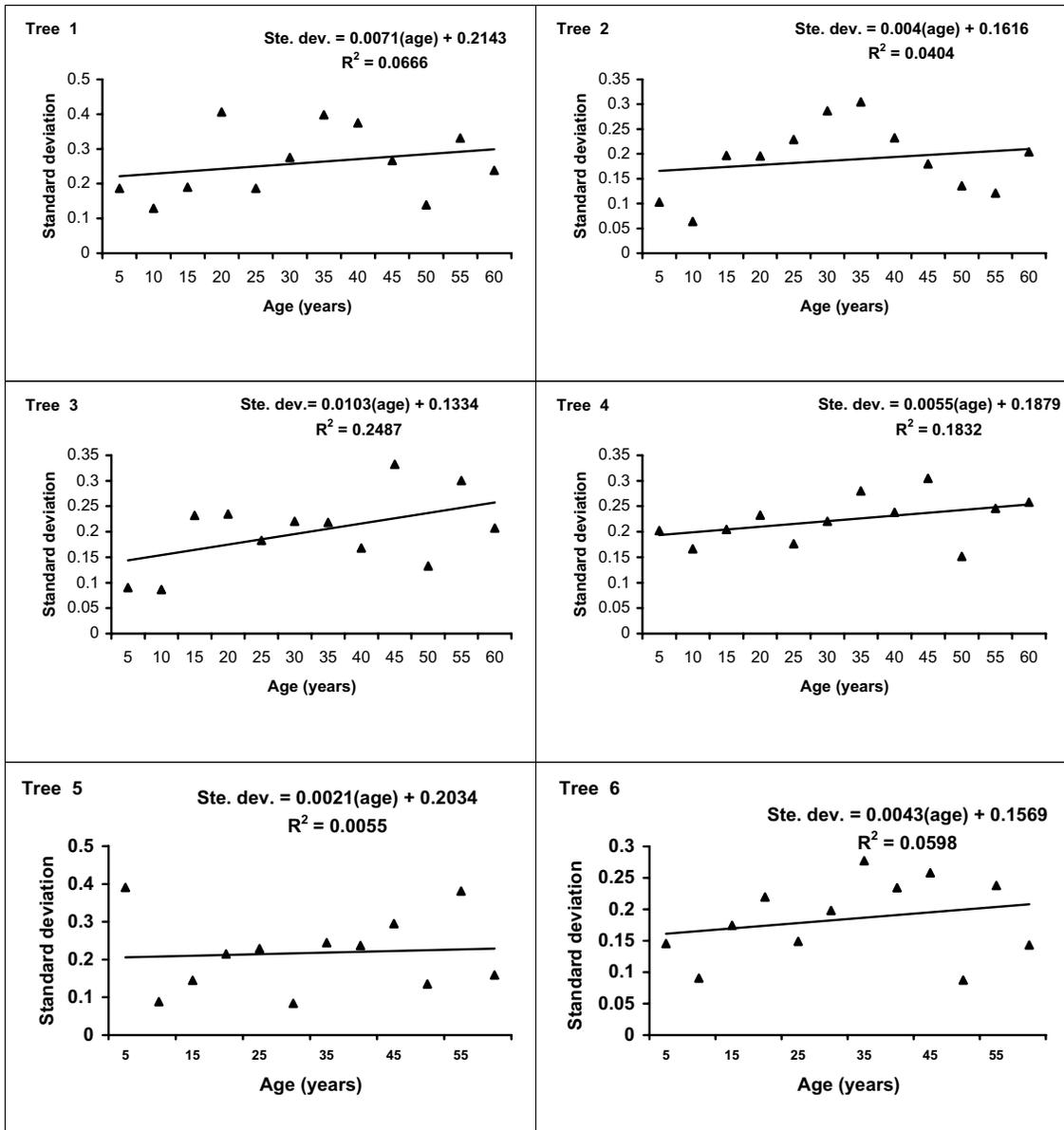


Figure 6.—Patterns of standard deviation of the 5-year segments of the RWI series. The plots indicate that variability of the RWI series is not a function of age and the tree-ring indexes satisfy the assumption of homogeneity of variance.

characteristic, each RWI series was partitioned into 5-year segments. The standard deviation of these 5-year segments was calculated and plotted against age (Fig. 6).

The small values of R^2 and non-significant slopes (age coefficients) in Figure 6 point to the fact that the variability of the detrended series (RWI) does not depend on age. Hence, the RWI series satisfy the assumption of homogeneous variance. However, the values of the RWI series are still interdependent, i.e., low values follow low and high values follow high (violation of the assumption of independence). This characteristic can be seen in Table 3, which shows the first- and second-order autocorrelation coefficients and the standard deviation of the RWI series for selected trees. The tree-ring indexes still exhibit a high degree of serial correlation. The first order autocorrelation coefficients are significant at 5 percent level. Hence the autodependence in the RWI series has to be removed.

Table 3.—First-order (r_1) and second-order (r_2) serial correlation coefficients for ring-width index (RWI) of selected trees

Tree	r_1	r_2	Ste. dev.
			<i>mm</i>
1	0.54	0.36	0.34
2	0.48	0.29	0.27
3	0.38	0.06	0.22
4	0.34	0.22	0.25
5	0.52	0.26	0.26
6	0.37	0.08	0.20

Autoregressive modeling. After removing the age-related component of growth from each series, some positive autocorrelation structure in the ring-width index is apparent. Positive autocorrelation causes underestimates of the error variance in the model of Eq. 1 and this results in narrower confidence intervals and higher test statistics (SAS 1999). Hence, it leads us to conclude that the effect of a particular climatic variable is significant when it is not. Various authors (Cook 1985, Monserud 1986, Visser and Molenaar 1990) have discussed the importance of removing autocorrelation (prewhitening) before climatic models are explored. However, most dendroclimatic studies use the average RWI series (usually called standard chronology) as the response variable in the model of Eq. 1 and to handle the problem of autocorrelation they include two or three lagged variables of the response into the climatic dataset (e.g., Lindholm et al. 2000). It has been long recognized in time-series literature (Granger and Morris 1976) and in some more recent dendroclimatic studies (Visser and Moelanaar 1990) that averaging time series data before removing autocorrelation leads to an averaged series with even higher order autocorrelation and affects proper interpretation.

In this study, autoregressive models (Box and Jenkins 1976, Visser and Moelanaar 1990) were used to remove the autocorrelation from each RWI series before creating an average for the species. The autoregressive model of order p (AR(p)) that was used has the form

$$I_t = \left(\sum_{i=1}^p \phi_i I_{t-i} \right) + a_t \quad (28)$$

where a_t is a purely random process with mean zero and variance σ_a^2 , i.e., “white noise” or prewhitened series, I_t and I_{t-i} represent observations (standardized ring-width index) at time t and $t - i$ respectively, and ϕ_i ’s are the autoregressive parameters. To decide the order of the autoregressive model, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were computed in SAS (Appendix B) and plotted for each series under analysis. An example of these plots is shown in Figure 7. The order of the specific model p was then decided by using these plots in conjunction to Akaike’s information criterion (AIC) and the Mallows’ C_p statistic. After devising the adequate order for each sequence, the autoregressive coefficients were estimated by fitting the model using procedure AUTOREG in SAS (Appendix B).

AR process of low order has proved adequate in modeling the autocorrelation structure of the detrended series. Table 4 shows the estimated autoregressive parameters of the AR(1) models that were fitted to the RWI series of the selected trees. The fitted AR(1) models and the resulting residuals, prewhitened RWI series (PRWI), are displayed in Figure 8.

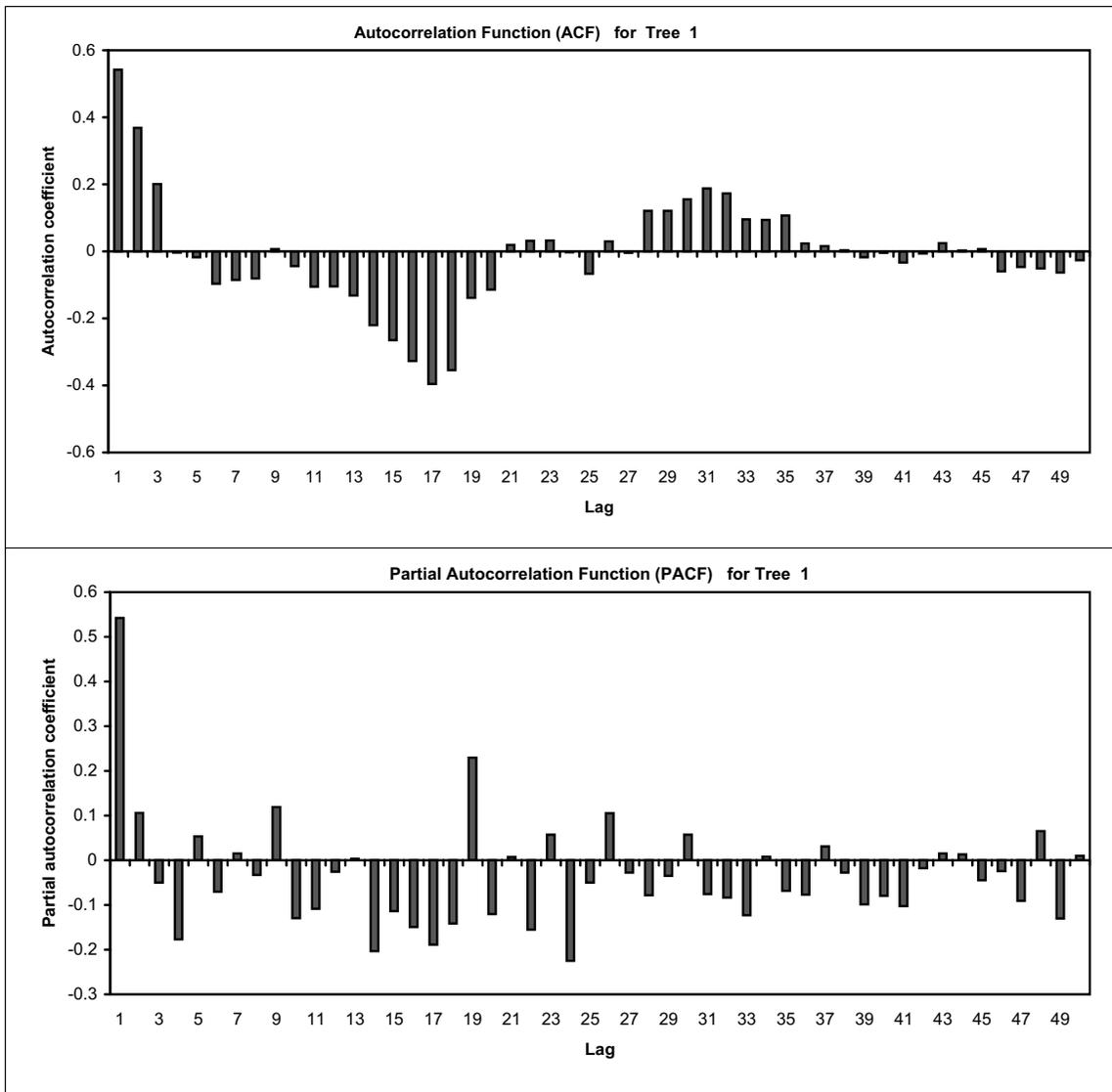


Figure 7.—Autocorrelation (AC) and partial autocorrelation (PAC) functions for the RWI series of tree 1. The higher peak at lag 1 indicates that an autoregressive model of order 1 is sufficient to remove the autocorrelation from the RWI for this tree.

With few exceptions, the ring-width indexes of all sampled yellow-poplar trees showed an autocorrelation structure of order one, meaning that growth of current year is influenced by prior year's growth. The first order autocorrelation coefficients of the residuals from the autoregressive modeling were small and not significant. The significance of the autocorrelation coefficients of the PRWI series was tested using the Durbin-Watson (D-W) statistic. The value of this statistic was about 2 for most sampled trees and this indicates that the values of the *PRWI* are independent.

The mean of the PRWI from each tree produces the prewhitened tree-ring chronology (the PRWC, see plot a of Fig. 9). This chronology is the appropriate response variable to be used in the multiple regression model of Eq. 1. The prewhitened chronology satisfies the assumptions of independence (plot b of Fig. 9) and homogeneity of variance. All growth-climate analysis in this study is based on the prewhitened ring-width chronology as a measure of tree growth.

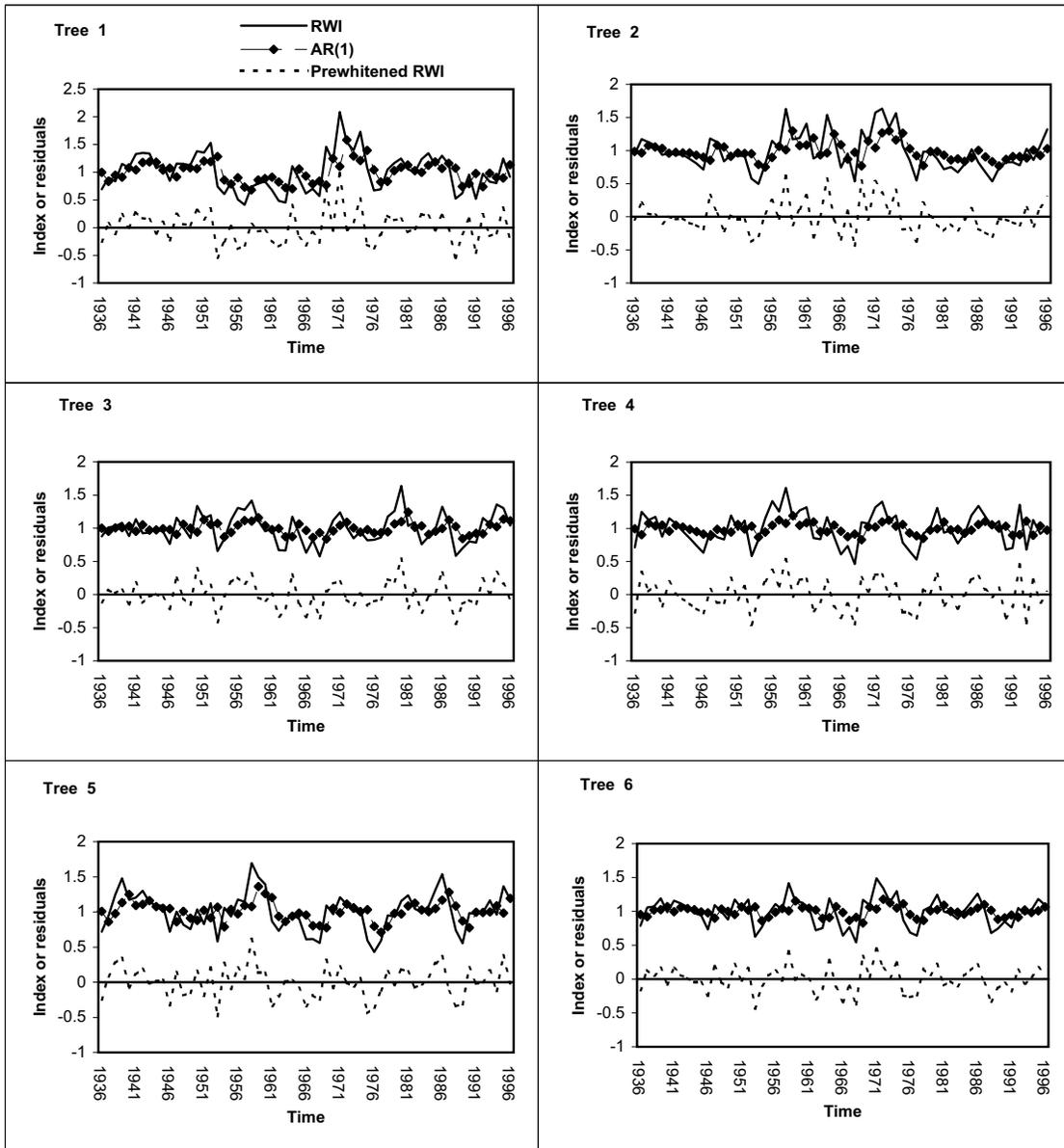


Figure 8.—Plots of the first order autoregressive models (AR(1)) fitted to the RWI series of selected trees and the resulting residuals (or prewhitened tree-ring indexes).

Table 4.—Estimated parameters of first order autoregressive models fitted to RWI series of selected trees and percentage of variability of RWI series explained by autoregressive models. The fourth column (r_1) shows first order autocorrelation coefficient of residuals (prewhitened RWI series) from AR models. The last column shows Durbin-Watson (D-W) test statistic used to detect autocorrelation in prewhitened RWI

Tree	ϕ_1	R ² (%)	r_1	D-W
1	0.54	30	-0.066	2.12
2	0.49	23	-0.040	2.05
3	0.38	14	0.039	1.91
4	0.31	10	-0.032	2.04
5	0.52	29	0.025	1.95
6	0.49	31	0.018	2.01

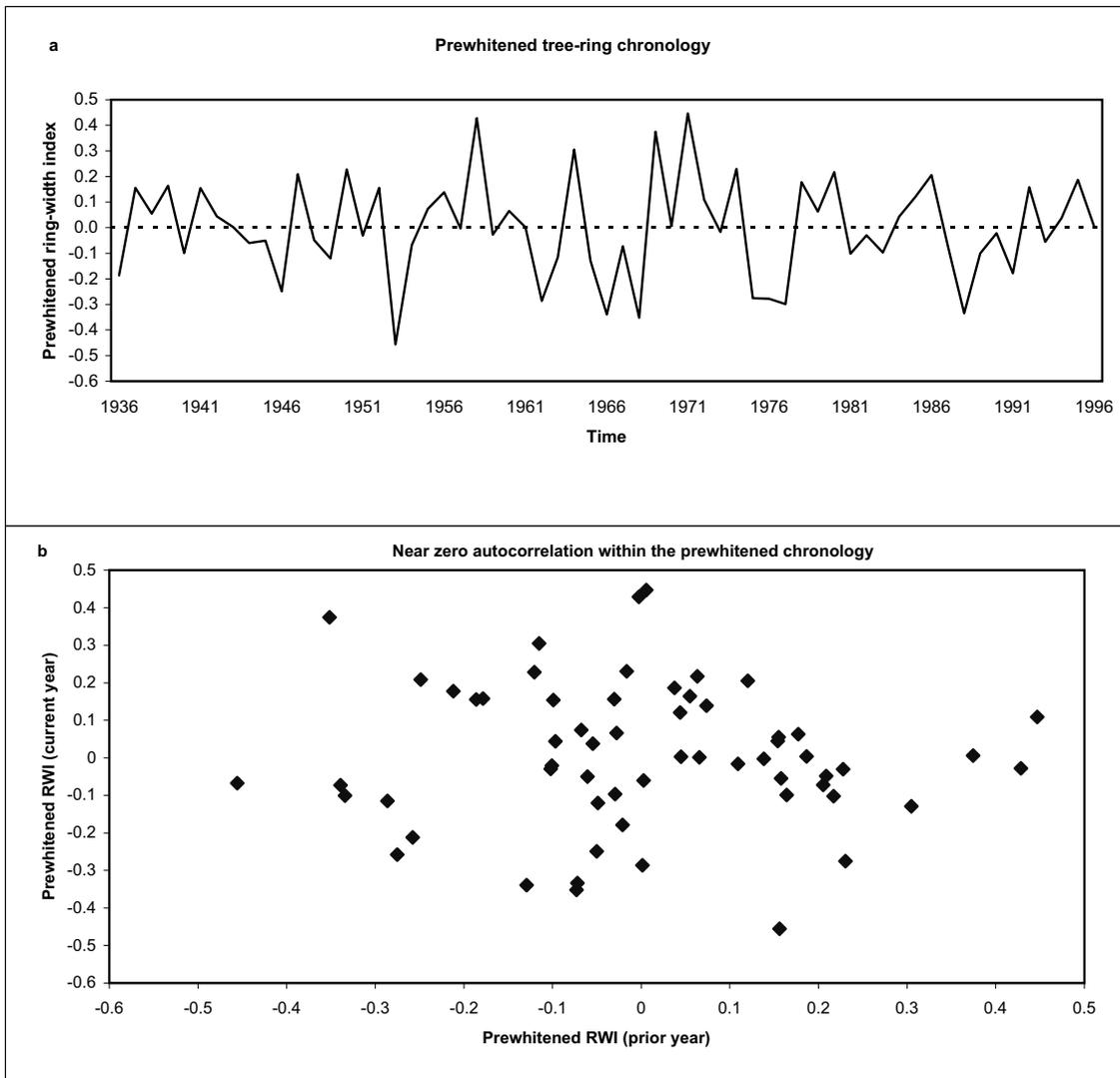


Figure 9.—Prewhitened chronology based on 38 yellow-poplar trees (a). Plot b shows that autoregressive modeling of the tree-ring indexes has removed the serial correlation, making the values of the prewhitened series independent.

Results and Discussion

Procedure for Estimating Response Function

Raw ring-width measurements were transformed in order to remove nonclimatic sources of variation and to account for autocorrelation. The measure of growth that was developed, the PRWC, is assumed to contain growth signals related to variations in climate. The main goal at this point is to relate the prewhitened chronology (y) with a set of 34 climatic variables ($k=34$) using the model given in Eq. 1. However, multicollinearity among the climatic variables necessitates the use of PCR rather than OLS to develop the response function. In addition, there is no well known software specifically designed to perform PCR. Although most statistical software performs the most difficult steps of PCR, none yields the final result and hence a user should understand how to complete the remaining steps. The procedures below show how to compute the principal component estimators of the climatic variables in the model of Eq. 1.

Response Function Based on Centered and Scaled Climatic Variables. Steps 1-4 and 6-8 of the following procedures can be accomplished efficiently in SAS as demonstrated in Appendix C.

1. Compute the $k \times k$ correlation matrix, $\mathbf{C} = \mathbf{X}^* \mathbf{X}^*$, of the climatic variables.
2. Compute the k eigenvalues of the above correlation matrix, ordered largest to smallest, $\lambda_1, \lambda_2, \dots, \lambda_k$.
3. Compute the k eigenvectors, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$, associated with each eigenvalue. Let

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdot & \cdot & \cdot & v_{1k} \\ v_{21} & v_{22} & \cdot & \cdot & \cdot & v_{2k} \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ v_{k1} & v_{k2} & \cdot & \cdot & \cdot & v_{kk} \end{bmatrix} \quad (29)$$

represent the $k \times k$ orthonormal matrix of eigenvectors. The matrix is orthonormal because its columns satisfy the conditions $\mathbf{v}'_j \mathbf{v}_j = 1$ and $\mathbf{v}'_j \mathbf{v}_i = 0$, $j \neq i$.

4. Compute the k principal components (new variables), $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k$, of the correlation matrix. Let \mathbf{Z} , an $n \times k$ matrix, represent the n readings on the k new variables or principal components. Then

$$\begin{aligned} \mathbf{Z} &= \mathbf{X}^* \mathbf{V} \\ &= \begin{bmatrix} x_{11}^* & x_{21}^* & \cdot & \cdot & \cdot & x_{k1}^* \\ x_{12}^* & x_{22}^* & \cdot & \cdot & \cdot & x_{k2}^* \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ x_{1n}^* & x_{2n}^* & \cdot & \cdot & \cdot & x_{kn}^* \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdot & \cdot & \cdot & v_{1k} \\ v_{21} & v_{22} & \cdot & \cdot & \cdot & v_{2k} \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ v_{k1} & v_{k2} & \cdot & \cdot & \cdot & v_{kk} \end{bmatrix} \\ &= \begin{bmatrix} z_{11} & z_{21} & \cdot & \cdot & \cdot & z_{k1} \\ z_{12} & z_{22} & \cdot & \cdot & \cdot & z_{k2} \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ z_{1n} & z_{2n} & \cdot & \cdot & \cdot & z_{kn} \end{bmatrix} \quad (30) \end{aligned}$$

where \mathbf{X}^* is $n \times k$ matrix of centered and scaled climatic variables without the column of ones as given in the model of Eq. 5, and \mathbf{V} is defined in Eq. 29. The elements of matrix \mathbf{Z} are usually called scores or amplitudes of principal components. As described in Eq. 17, the elements of matrix \mathbf{Z} are linear functions of the centered and scaled climatic variables. For example, the first element of the first principal component, i.e., z_{11} , is computed as $z_{11} = v_{11}x_{11}^* + v_{21}x_{21}^* + \dots + v_{k1}x_{k1}^*$ and the first element of the last principal component (z_{k1})

is computed as $z_{k1} = v_{1k}x_{11}^* + v_{2k}x_{21}^* + \dots + v_{kk}x_{k1}^*$. Some of the properties of these new variables or principal components are as follows:

- a. mean of the column vectors, \mathbf{z}_j , is zero, $\bar{z}_j = 0$,
 - b. the sum of squares of elements of \mathbf{z}_j is λ_j , $\mathbf{z}'_j\mathbf{z}_j = \sum_{i=1}^n (z_{ji} - \bar{z}_j)^2 = \sum_{i=1}^n z_{ji}^2 = \lambda_j$,
 - c. the variance of \mathbf{z}_j is hence $\lambda_j/n - 1$, and
 - d. since $\mathbf{z}'_j\mathbf{z}_i = 0$, $i \neq j$, the principal components are independent (orthogonal) of each other, and $\mathbf{Z}'\mathbf{Z} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$.
5. Using one of the principal components selection rules discussed earlier, eliminate some of the principal components. Suppose $r < k$ components are eliminated. These selection methods still leave some principal components that have nonsignificant weight on the dependent variable and hence nonsignificant components can be rejected using the strategy (rule) described in (d, page 6). However, to this date, dendroecological studies (e.g., Fritts 1976, Guiot et al. 1982, Lindholm et al. 2000) use a stepwise regression procedure to eliminate the nonsignificant principal components. This is described below.
 6. Regress the prewhitened tree-ring chronology y against the remaining $k - r$ principal components using linear regression (OLS). That is, estimate the parameters of the model in Eq. 20. Here we use the same decision criteria that is used by default in SAS and other statistical analysis packages. Fifteen percent is the commonly used probability level for entry of a component into a model (Fritts 1976, SAS 1999). Because of changes in the mean squared error from a full model with nonsignificant terms to a reduced model, we recommend initially considering terms with p-values slightly larger than 0.15. These should then be reinvestigated for the reduced model once all clearly nonsignificant terms have been removed.

If one decides to use the stepwise analysis, what is being accomplished can be done using the test statistic in Eq. 19 and it is important to understand that the order of entry of the principal components is irrelevant since they are orthogonal to one another. Once a principal component is added to the model of Eq. 20, its effect is not altered by the components already in the model or by the addition of other components because each principal component has an independent contribution in explaining the variation in the response variable.

To summarize the point, in most dendroecological studies the selection of principal components is accomplished in two stages: eliminate $r < k$ principal components using the cumulative eigenvalue product rule (rule c, page 6), and then further screen the remaining $k - r$ components using a significance level of 15 percent.

Suppose that such a principal components selection rule results in retention of l of the $k - r$ components. The response function will then be computed based on these l principal components (\mathbf{Z}_l^*).

7. Regress the prewhitened tree-ring chronology y against these l principal components. That is, fit the model

$$\mathbf{y} = \beta_0^* \mathbf{1} + \mathbf{Z}_l^* \boldsymbol{\alpha}_l + \boldsymbol{\varepsilon}^{\circ\circ} \quad (31)$$

where $\mathbf{Z}_l^* = \mathbf{X}^* \mathbf{V}_l$ is an $n \times l$ matrix, \mathbf{V}_l is a $k \times l$ matrix of eigenvectors corresponding to these l components, and $\boldsymbol{\alpha}_l$ is $l \times 1$ vector of coefficients associated with the l components. For example, with $k = 34$ climatic variables and hence 34 principal components, suppose that at step 5 the last eight principal components with small eigenvalues are eliminated, that is, $r = 8$, and $k - r = 26$. Further assume that the regression at step 6 eliminates 10 of the 26 principal components, that is, $l = 16$. These 16 components that remained in the model of Eq.

31 are not necessarily the first 16 principal components. The matrix \mathbf{V}_l contains the eigenvectors corresponding to these components.

8. Compute the mean square error (MSE), and standard error of the estimated coefficients in vector $\hat{\boldsymbol{\alpha}}_l$ of the model in Eq. 31. Recall that from Eq. 18 $s.e.(\hat{\alpha}_j) = S(\sqrt{\lambda_j})^{-1} (S^2 = \hat{\sigma}^2)$. Let the estimated standard errors of the estimated coefficients in $\hat{\boldsymbol{\alpha}}_l$ be represented by an $l \times 1$ vector

$$\mathbf{k} = (s.e._{\hat{\alpha}_1} \quad s.e._{\hat{\alpha}_2} \quad \cdot \quad \cdot \quad \cdot \quad s.e._{\hat{\alpha}_l})' \quad (32)$$

These standard errors will be used later for testing the statistical significance of the elements of the response function, i.e., to construct confidence intervals.

9. Obtain the principal component estimators of the coefficients in terms of the centered and scaled climatic variables using Eq. 23. That is, $b_{0,pc}^* = \bar{y}$ and the remaining estimators are obtained as follows:

$$\begin{bmatrix} b_{1,pc}^* \\ b_{2,pc}^* \\ \cdot \\ \cdot \\ \cdot \\ b_{k,pc}^* \end{bmatrix} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdot \quad \cdot \quad \cdot \quad \mathbf{v}_l] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\alpha}_l \end{bmatrix} \quad (33)$$

10. Now transform the coefficients back to the natural climatic variables using Eq. 24 and Eq. 25.

The coefficients obtained at step 10 are the principal component estimators of the regression coefficients of the climatic variables in the model of Eq. 1. The coefficients obtained at step 9 are the principal component estimators of the regression coefficients of the climatic variables in the model of Eq. 5. The principal component estimators at steps 9 and 10 have the same sign and test statistic but different magnitudes and standard errors.

If one decides to report the values of the principal component estimators at step 10 then two difficulties arise: if response functions are calculated by different researchers who use different scales of measurement on the same variables (for example, inches and centimeters for precipitation, degree-Fahrenheit and degree-Centigrade for temperature), the resulting coefficients are not directly comparable; and when comparing the relative importance of several climatic variables in the response function, the climatic variable with the largest magnitude might not be the most influential variable. Its magnitude could be due mainly to the scale in which it was measured. Therefore, to avoid the aforementioned problems, researchers should report the principal component estimates of the centered and scaled climatic variables obtained at step 9.

Response Function Based on Standardized Climatic Variables. Statistical packages such as SAS (1999) compute amplitudes or scores of the principal components as a function of the standardized climatic variables as follows:

$$\mathbf{Z}^s = \mathbf{X}^s \mathbf{V} \quad (34)$$

where \mathbf{X}^s is $n \times k$ matrix of standardized climatic variables without the column of ones as given in Eq. 10, \mathbf{V} is defined in Eq. 26, and \mathbf{Z}^s is $n \times k$ matrix of principal components, $\mathbf{z}_1^s, \mathbf{z}_2^s, \dots, \mathbf{z}_k^s$. The superscript s is used to indicate that the components are computed using the standardized regressors.

Properties of the principal components computed using Eq. 34 are:

- mean of \mathbf{z}_j^s is zero, $\bar{z}_j^s = 0$,
- the variance of \mathbf{z}_j^s is λ_j , and
- the components are orthogonal (independent). That is, $\mathbf{Z}'^s \mathbf{Z}^s$ is a diagonal matrix where the diagonal elements are the sums of squares of the principal components.

If one is interested in computing a response function based on the standardized climatic variables, Eq. 34, the steps outlined above should be followed with the following adjustments (note that steps 1 to 3 are standard computations needed in either approach):

- at step 4 the principal components should be computed using Eq. 34 rather than Eq. 30. That is, replace \mathbf{Z}_l^* by \mathbf{Z}^s and \mathbf{X}^* by \mathbf{X}^s ,
- at step 7 replace \mathbf{Z}_l^* by \mathbf{Z}_l^s , $\boldsymbol{\alpha}_l$ by $\boldsymbol{\alpha}_l^s$, and \mathbf{V}_l by \mathbf{V}_l^s ,
- at step 8 replace $\boldsymbol{\kappa}$ by $\boldsymbol{\kappa}^s$,
- at step 9 in Eq. 33 replace $b_{j,pc}^*$ by $b_{j,pc}^s$. The results obtained at step 9 will be coefficients for the standardized climatic variables (rather than centered and scaled variables). Note that $b_{0,pc}^s = \bar{y}$, and
- the appropriate transformation of the coefficients back to the natural (original) variables at step 10 is accomplished by using Eq. 11 and 12. That is,

$$b_{j,pc} = \frac{b_{j,pc}^s}{S_{x_j}}, j = 1, 2, \dots, k \quad (35)$$

and

$$b_{0,pc} = b_{0,pc}^s - \frac{b_{1,pc}^s \bar{x}_1}{S_{x_1}} - \frac{b_{2,pc}^s \bar{x}_2}{S_{x_2}} - \dots - \frac{b_{k,pc}^s \bar{x}_k}{S_{x_k}} \quad (36)$$

where S_{x_j} is the standard deviation of the j^{th} original climatic variable x_j and $b_{0,pc}^s$, $b_{1,pc}^s$, $b_{2,pc}^s$, ..., $b_{k,pc}^s$ are coefficients of the standardized climatic variables obtained at step 9.

Standard Errors of the Principal Component Estimators

Let $\hat{\boldsymbol{\alpha}}_l = (\hat{\alpha}_1 \quad \hat{\alpha}_2 \quad \dots \quad \hat{\alpha}_l)'$ be the vector of the estimated coefficients in Eq. 31, and $\boldsymbol{\kappa} = (s.e._{\hat{\alpha}_1} \quad s.e._{\hat{\alpha}_2} \quad \dots \quad s.e._{\hat{\alpha}_l})'$ is the vector of the estimated standard errors of the coefficients in vector $\hat{\boldsymbol{\alpha}}_l$. Note that both $\hat{\boldsymbol{\alpha}}_l$ and $\boldsymbol{\kappa}$ are $l \times 1$ column vectors. Let \mathbf{V}_l be the $k \times l$ matrix of eigenvectors.

Now, the prewhitened tree-ring chronology can be statistically predicted from the climatic data using the fitted model of Eq. 31:

$$\hat{\mathbf{y}} = \hat{\beta}_0^* \mathbf{1} + \mathbf{Z}_l \hat{\boldsymbol{\alpha}}_l = \hat{\beta}_0^* \mathbf{1} + (\mathbf{X}^* \mathbf{V}_l) \hat{\boldsymbol{\alpha}}_l = \hat{\beta}_0^* \mathbf{1} + \mathbf{X}^* (\mathbf{V}_l \hat{\boldsymbol{\alpha}}_l) = \hat{\beta}_0^* \mathbf{1} + \mathbf{X}^* \mathbf{b}_{pc}^* \quad (37)$$

Recall that the principal component estimators of the coefficients of the centered and scaled climatic variables, \mathbf{b}_{pc}^* , was given by $\mathbf{b}_{pc}^* = \mathbf{V}_l \hat{\boldsymbol{\alpha}}_l$. From the expression $\mathbf{b}_{pc}^* = \mathbf{V}_l \hat{\boldsymbol{\alpha}}_l$, one can easily recognize that the coefficients in vector \mathbf{b}_{pc}^* are linear combinations of the coefficients of vector $\hat{\boldsymbol{\alpha}}_l$. That is

$$\begin{bmatrix} b_{1,pc}^* \\ b_{2,pc}^* \\ \cdot \\ \cdot \\ b_{k,pc}^* \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \cdot & \cdot & \cdot & v_{1l} \\ v_{21} & v_{22} & \cdot & \cdot & \cdot & v_{2l} \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ v_{k1} & v_{k2} & \cdot & \cdot & \cdot & v_{kl} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \cdot \\ \cdot \\ \hat{\alpha}_l \end{bmatrix} \quad (38)$$

For example, the first coefficient $b_{1,pc}^*$ is computed as $b_{1,pc}^* = v_{11}\hat{\alpha}_1 + v_{12}\hat{\alpha}_2 + \dots + v_{1l}\hat{\alpha}_l$. Hence, $b_{1,pc}^*$ is a linear function of $\hat{\alpha}_1, \hat{\alpha}_2, \dots$, and $\hat{\alpha}_l$ where the coefficients of the linear combination are the eigenvectors. Note that the estimators $\hat{\alpha}_1, \hat{\alpha}_2, \dots$, and $\hat{\alpha}_l$ are independent since they are coefficients of l orthogonal variables (principal components). Mutual independence of $\hat{\alpha}_1, \hat{\alpha}_2, \dots$, and $\hat{\alpha}_l$ facilitates easy computation of the variance (or standard error) of any linear combination of these estimators.

Therefore, the variance and standard error of the coefficients in vector \mathbf{b}_{pc}^* can be computed easily given the variance and standard error of the estimated coefficients in vector $\hat{\boldsymbol{\alpha}}_l$. For example, the variance of $b_{1,pc}^*$ is computed as:

$$\begin{aligned} \text{var}(b_{1,pc}^*) &= \text{var}(v_{11}\hat{\alpha}_1 + v_{12}\hat{\alpha}_2 + \dots + v_{1l}\hat{\alpha}_l) = \text{var}(v_{11}\hat{\alpha}_1) + \text{var}(v_{12}\hat{\alpha}_2) + \dots + \text{var}(v_{1l}\hat{\alpha}_l) \\ \text{var}(b_{1,pc}^*) &= v_{11}^2 \text{var}(\hat{\alpha}_1) + v_{12}^2 \text{var}(\hat{\alpha}_2) + \dots + v_{1l}^2 \text{var}(\hat{\alpha}_l) \end{aligned} \quad (39)$$

To generalize the above formulation using a matrix notation let us label the equations used to calculate the variance of each element of the vector \mathbf{b}_{pc}^* from 1 to k as follows:

$$\text{var}(b_{1,pc}^*) = v_{11}^2 \text{var}(\hat{\alpha}_1) + v_{12}^2 \text{var}(\hat{\alpha}_2) + \dots + v_{1l}^2 \text{var}(\hat{\alpha}_l) \quad [1]$$

$$\text{var}(b_{2,pc}^*) = v_{21}^2 \text{var}(\hat{\alpha}_1) + v_{22}^2 \text{var}(\hat{\alpha}_2) + \dots + v_{2l}^2 \text{var}(\hat{\alpha}_l) \quad [2]$$

·
·
·

$$\text{var}(b_{k,pc}^*) = v_{k1}^2 \text{var}(\hat{\alpha}_1) + v_{k2}^2 \text{var}(\hat{\alpha}_2) + \dots + v_{kl}^2 \text{var}(\hat{\alpha}_l) \quad [k]$$

In matrix notation the expressions from [1] to [k] can be rewritten as follows:

$$\text{Var}(\mathbf{b}_{pc}^*) = \begin{bmatrix} v_{11}^2 & v_{12}^2 & \cdot & \cdot & \cdot & v_{1l}^2 \\ v_{21}^2 & v_{22}^2 & \cdot & \cdot & \cdot & v_{2l}^2 \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ v_{k1}^2 & v_{k2}^2 & \cdot & \cdot & \cdot & v_{kl}^2 \end{bmatrix} \begin{bmatrix} \text{var}(\hat{\alpha}_1) \\ \text{var}(\hat{\alpha}_2) \\ \cdot \\ \cdot \\ \text{var}(\hat{\alpha}_l) \end{bmatrix} \quad (40)$$

The vector $\text{Var}(\mathbf{b}_{pc}^*)$, therefore, gives the variance of the principal component estimators of the coefficients for the centered and scaled climatic variables. The standard deviation of the sampling

distribution of the elements of \mathbf{b}_{pc}^* (also called standard error) is simply the square root of the variance of the coefficients. That is

$$s.e.(\mathbf{b}_{pc}^*) = \left[\text{Var}(\mathbf{b}_{pc}^*) \right]^{\frac{1}{2}} \quad (41)$$

where here the square root is done elementwise for each element of this column vector.

The principal component estimators of the regression coefficients in the model of Eq. 1 are obtained using the relationship $b_{j,pc} = \frac{b_{j,pc}^*}{s_j}$ where s_j is a scale constant defined in Eq. 3. Hence standard error of the principal component estimators of the coefficients of the natural climatic variables are obtained as follows (for the j^{th} principal component estimator):

$$s.e.(b_{j,pc}) = \sqrt{\text{var}(b_{j,pc})} = \sqrt{\text{var}\left(\frac{b_{j,pc}^*}{s_j}\right)} = \sqrt{\frac{1}{s_j^2} \text{var}(b_{j,pc}^*)} = \frac{s.e.(b_{j,pc}^*)}{s_j} \quad (42)$$

where $s.e.(b_{j,pc}^*)$ is the standard error of the principal component estimator of the coefficient associated with the j^{th} centered and scaled climatic variable, or it is the j^{th} element of the vector given in Eq. 41.

If the response function is developed using the standardized climatic variables, then the variance of the principal component estimators of coefficients for the standardized climatic variables is given by

$$\text{Var}(\mathbf{b}_{pc}^s) = \Psi_l^s \mathbf{K}^s \quad (43)$$

where Ψ_l^s contains the squares of the elements of \mathbf{V}_l^s , and \mathbf{K}^s contains the squares of the elements of $\boldsymbol{\kappa}^s$. Paralleling Eq. 41, the corresponding standard errors are given by

$$s.e.(\mathbf{b}_{pc}^s) = \left[\text{Var}(\mathbf{b}_{pc}^s) \right]^{\frac{1}{2}} \quad (44)$$

Recall that the principal component estimator of the coefficients of the original climatic variables are obtained using Eq. 35. Hence the standard error of the principal component estimator associated with the j^{th} original variable is

$$s.e.(b_{j,pc}) = \frac{s.e.(b_{j,pc}^s)}{S_{x_j}} \quad (45)$$

Inference Techniques

To test a hypothesis about the significance of the influence of a climatic variable ($H_0 : \beta_j^* = 0$ vs. $H_a : \beta_j^* \neq 0$) using the principal component estimators, Mansfield et al. (1977) and Gunst and Mason (1980) have shown that the appropriate statistic to use is

$$t = \frac{b_{j,pc}^*}{\left[\text{MSE} \left(\sum_{m=1}^l \lambda_m^{-1} v_{jm}^2 \right) \right]^{\frac{1}{2}}} \quad (46)$$

where $b_{j,pc}^*$ is the principal component estimator of β_j^* , MSE is the mean square error of the l -variable model in Eq. 31, v_{jm} is the j^{th} element of the eigenvector \mathbf{v}_m ($m = 1, 2, \dots, l$), λ_m is the corresponding eigenvalue, and the summation in Eq. 46 is taken over only those components retained at the end of the stepwise analysis. The statistic in Eq. 46 follows the Student's t distribution with $(n - k - 1)$ degrees of freedom under H_0 provided that the true coefficients of the components eliminated at step 5 and 6 on page 20 are zero. Therefore, to test $H_0 : \beta_j^* = 0$ vs. $H_a : \beta_j^* \neq 0$ with significance level of α , reject H_0 if the absolute value of the test statistic in Eq. 46 is greater than or equal to the critical value ($t_{(\alpha/2, n-k-1)}$).

The denominator in Eq. 46 is the standard error of $b_{j,pc}^*$, the j^{th} element of the vector given in Eq. 41. From Eq. 40 one can see that (note: $Var(\hat{\alpha}_j) = \hat{\sigma}^2 / \lambda_j = MSE / \lambda_j$)

$$\begin{aligned} Var(b_{j,pc}^*) &= v_{j1}^2 \text{var}(\hat{\alpha}_1) + v_{j2}^2 \text{var}(\hat{\alpha}_2) + \dots + v_{jl}^2 \text{var}(\hat{\alpha}_l) \\ &= v_{j1}^2 \frac{MSE}{\lambda_1} + v_{j2}^2 \frac{MSE}{\lambda_2} + \dots + v_{jl}^2 \frac{MSE}{\lambda_l} \\ &= MSE \left(\sum_{m=1}^l \frac{v_{jm}^2}{\lambda_m} \right) \end{aligned} \quad (47)$$

Hence, the test statistic in Eq. 46 simplifies to $t = b_{j,pc}^* / s.e.(b_{j,pc}^*)$. However, if the hypothesis to be tested is $H_0 : \beta_j^s = 0$ vs. $H_a : \beta_j^s \neq 0$, then the test statistic becomes

$$t = \frac{b_{j,pc}^s}{\left[\frac{MSE}{n-1} \left(\sum_{m=1}^l \lambda_m^{-1} v_{jm}^2 \right) \right]^{\frac{1}{2}}} \quad (48)$$

The denominator of the test statistic in Eq. 48 is the standard error of $b_{j,pc}^s$, the j^{th} element of the vector given in Eq. 44. Note that if principal component scores are computed using Eq. 34 then $\mathbf{Z}'\mathbf{Z}^s = \text{diag}(\lambda_1(n-1), \lambda_2(n-1), \dots, \lambda_k(n-1))$ and the matrix

$$Var(\hat{\alpha}^s) = MSE(\mathbf{Z}'\mathbf{Z}^s)^{-1} = MSE \times \text{diag} \left(\frac{1}{\lambda_1(n-1)}, \frac{1}{\lambda_2(n-1)}, \dots, \frac{1}{\lambda_k(n-1)} \right)$$

is the variance-

covariance matrix of the estimated coefficients associated with the principal components. Hence,

$$Var(\hat{\alpha}_j^s) = MSE / \lambda_j(n-1). \text{ From Eq. 43 the variance of } b_{j,pc}^s \text{ is then}$$

$$\begin{aligned} Var(b_{j,pc}^s) &= v_{j1}^2 \text{var}(\hat{\alpha}_1^s) + v_{j2}^2 \text{var}(\hat{\alpha}_2^s) + \dots + v_{jl}^2 \text{var}(\hat{\alpha}_l^s) \\ &= v_{j1}^2 \frac{MSE}{\lambda_1(n-1)} + v_{j2}^2 \frac{MSE}{\lambda_2(n-1)} + \dots + v_{jl}^2 \frac{MSE}{\lambda_l(n-1)} \\ &= \frac{MSE}{n-1} \left(\sum_{m=1}^l \frac{v_{jm}^2}{\lambda_m} \right) \end{aligned} \quad (49)$$

Fritts et al. (1971) and Fritts (1976) suggest that the hypothesis $H_0 : \beta_j^* = 0$ vs. $H_a : \beta_j^* \neq 0$ can be tested by constructing a 95 percent confidence interval of the form $b_{j,pc}^* \pm \sqrt{F_{0.05, v_1, v_2}} \times s.e.(b_{j,pc}^*)$, where v_1 is the number of nonzero elements of $\hat{\alpha}_l$ and $v_2 = n - 2 - v_1$. However, Fritts et al. (1971) and Fritts (1976) do not indicate the original source of the above F statistic nor do they give

a theoretical derivation. In dendroecological studies where the chronology length and the number of climatic variables are large, the two inferential procedures can lead to different and possibly contradicting results. Hence we suggest that users should adopt the test statistic developed by Mansfield et al. (1971) for testing significance of the regression coefficients in the model of Eq. 1, Eq. 5 or Eq. 10 when collinearity is present.

Comparison with the Fritts Approach

Here, we contrast the approach presented in this study and that of Fritts et al. (1971). These contrasts relate to the type of response variable used and the relationships between the formulas for estimating standard errors of the elements of the response function.

Fritts et al. (1971) and Fritts (1976) illustrate the concept, method of computation, and interpretation of the response function in dendroecological analysis. The approach in these and other literature uses a standard chronology (average of tree-ring indexes from a set of trees) as a response variable. The regressors are the climatic variables, and additional three-predictor variables that represent ring-width indexes for the 3 most recent years. The purpose of the three additional predictor variables was to account for possible interdependence in ring-width indexes measured by the first-, second-, and third-order autocorrelation.

The approach presented here assumes that the user prewhiten tree-ring indexes from each tree before computing the chronology. Therefore, the use of additional predictor variables that represent ring-width indexes of prior years is not necessary. Averaging tree-ring indexes from a set of trees, where each series exhibit a certain degree of autocorrelation, generally will yield a standard chronology with a higher order of autocorrelation (Granger and Morris 1976, Cook 1985) masking the true autodependence of tree growth. Hence, prewhitening has statistical advantage.

Fritts et al. (1971) and Fritts (1976) compute amplitudes of the principal components of the correlation matrix as a function of the standardized climatic variables. The discussion on the computation of the response function, though not clearly presented in terms of statistical and linear algebra details, is essentially the same as the procedure presented in this study.

Fritts (1976) defines the elements of the response function as the estimated regression coefficients associated with the original (or natural) climatic variables. In this study, these are coefficients obtained at step 10 of the step-by-step procedure, given that the principal components are computed using Eq.34. Fritts et al. (1971) and Fritts (1976) compute standard errors of the elements of the response function, *s.e.*($b_{j,pc}$), from the standard errors of the coefficients in the vector $\hat{\mathbf{a}}_l^s$. Using the notation in this study, the required transformation given in Fritts (1976) is

$$\mathbf{S} = \mathbf{V}_l^s \mathbf{U} \mathbf{U} \mathbf{V}_l^{s'} \quad (50)$$

where \mathbf{V}_l^s is $k \times l$ matrix of eigenvectors, \mathbf{U} is $l \times l$ diagonal matrix whose diagonal elements are standard errors of the coefficients in $\hat{\mathbf{a}}_l^s$, i.e., the diagonal elements of \mathbf{U} are elements of the vector $\boldsymbol{\kappa}^s$ and \mathbf{S} is $k \times k$ symmetric matrix whose diagonal elements are the square of the standard errors of the elements of the response function. However, the diagonal elements of \mathbf{S} are not the square of the standard errors of the elements of the response function, rather they are the variances (square of the standard errors) of the coefficients for the standardized climatic variables. That is, the j^{th} diagonal element of \mathbf{S} is not the same as $\text{var}(b_{j,pc})$ but it is the same as $\text{var}(b_{j,pc}^s)$.

The diagonal elements of the matrix \mathbf{S} from Fritts (1976) are equivalent to elements of the vector $\text{Var}(\mathbf{b}_{pc}^s)$ given in Eq. 43. A user that employs Fritts (1976) equation, i.e., Eq. 50, ends up getting variances of the coefficients for the standardized climatic variables and hence, a straightforward follow-up computation is needed to obtain variance (or standard errors) of the elements of the

response function. The standard error of the j^{th} element of the response function should be computed as

$$s.e.(b_{j,pc}) = \frac{\sqrt{j^{th} \text{ diagonal element of matrix } \mathbf{S}}}{S_{x_j}} = (s_{jj})^{\frac{1}{2}} S_{x_j}^{-1} \quad (51)$$

Computational Comparison of Approaches

In this section a demonstration of the step-by-step procedure for computation of a response function is presented. The eigenvalues, eigenvectors, and the orthogonal principal components of the correlation matrix of the 34 climatic variables were computed in SAS¹ (steps 1-4). Then the cumulative eigenvalue product rule was used to eliminate the last 9 components and the first 25 were kept (step 5). The prewhitened tree-ring chronology was regressed against the 25 orthogonal variables and further screening of these components was carried out by keeping components that were significant at probability level of 20 percent (step 6). This has resulted in retention of four components ($l = 4$) all of which have probability below 15 percent.

For the model relating the response with the four selected components, the estimated regression coefficients (apart from the intercept), $\hat{\boldsymbol{\alpha}}_l^s$, and the matrix of eigenvectors associated with the $l = 4$ retained principal components, \mathbf{V}_l^s are given below: (step 7)

$$\hat{\boldsymbol{\alpha}}_{4 \times 1}^s = \begin{bmatrix} \hat{\alpha}_1^s \\ \hat{\alpha}_2^s \\ \hat{\alpha}_3^s \\ \hat{\alpha}_4^s \end{bmatrix} = \begin{bmatrix} 0.028 \\ -0.055 \\ 0.030 \\ -0.055 \end{bmatrix} \quad \mathbf{V}_{34 \times 4}^s = \begin{bmatrix} -0.31 & 0.21 & 0.11 & 0.23 \\ 0.06 & 0.14 & 0.28 & 0.06 \\ 0.31 & -0.04 & -0.01 & 0.05 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0.14 & -0.20 & 0.05 & 0.20 \end{bmatrix}$$

The estimated standard errors of the coefficients in vector $\hat{\boldsymbol{\alpha}}_{4 \times 1}^s$ (step 8) are given by

$$\boldsymbol{\kappa}^s = [0.01632 \quad 0.01751 \quad 0.01873 \quad 0.02855]'$$

The principal component estimators of the coefficients of the standardized climatic variables were obtained using Eq. 33. That is, $\mathbf{b}_{j,pc}^s = \mathbf{V}_{34 \times 4}^s \hat{\boldsymbol{\alpha}}_{4 \times 1}^s$ (Note: the intercept $b_{0,pc}^s = \bar{y} = 0.00$): (step 9)

$$\begin{bmatrix} b_{1,pc}^s \\ b_{2,pc}^s \\ \cdot \\ \cdot \\ \cdot \\ b_{34,pc}^s \end{bmatrix} = \begin{bmatrix} -0.31 & 0.21 & 0.11 & 0.23 \\ 0.06 & 0.14 & 0.28 & 0.06 \\ 0.31 & -0.04 & -0.01 & 0.05 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0.14 & -0.20 & 0.05 & 0.20 \end{bmatrix} \begin{bmatrix} 0.028 \\ -0.055 \\ 0.030 \\ -0.055 \end{bmatrix} = \begin{bmatrix} -0.02928 \\ -0.00085 \\ 0.008144 \\ \cdot \\ \cdot \\ \cdot \\ 0.004998 \end{bmatrix}$$

¹Except for steps 9 and 10, all computations indicated in the step-by-step procedure can be accomplished in SAS.

Transformation of the coefficients back to the original (or natural) climatic variables was done using Eq. 35 and 36. The estimated intercept using Eq. 36 is 1.003. The coefficients of the original climatic variables were obtained by dividing the coefficients of the standardized variables by the standard deviation of the original variables: (step 10)

$$\begin{bmatrix} b_{1,pc} \\ b_{2,pc} \\ \cdot \\ \cdot \\ \cdot \\ b_{34,pc} \end{bmatrix} = \begin{bmatrix} -0.002928/7.582 \\ -0.00085/7.306 \\ \cdot \\ \cdot \\ \cdot \\ 0.004998/1.915 \end{bmatrix} = \begin{bmatrix} -0.00386 \\ -0.00012 \\ \cdot \\ \cdot \\ \cdot \\ 0.00261 \end{bmatrix}$$

The variance of the principal component estimators of the coefficients for the standardized climatic variables, using Eq. 43, is

$$\text{Var}(\mathbf{b}_{pc}^s) = (\Psi_{34 \times 4}^s)(\mathbf{K}_{4 \times 1}^s)$$

$$\text{var} \begin{bmatrix} b_{1,pc}^s \\ b_{2,pc}^s \\ \cdot \\ \cdot \\ \cdot \\ b_{34,pc}^s \end{bmatrix} = \begin{bmatrix} 0.097 & 0.042 & 0.013 & 0.052 \\ 0.003 & 0.020 & 0.077 & 0.003 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0.019 & 0.040 & 0.002 & 0.042 \end{bmatrix} \begin{bmatrix} 0.000266 \\ 0.000307 \\ 0.000351 \\ 0.000815 \end{bmatrix} = \begin{bmatrix} 0.004047 \\ 0.001932 \\ \cdot \\ \cdot \\ \cdot \\ 0.002237 \end{bmatrix} \quad (52)$$

Notice that the elements of the vector in Eq. 52 can also be obtained by evaluating Eq. 49. The standard error of the principal component estimators of the coefficients of the standardized climatic variables is then

$$\text{s.e.} \begin{bmatrix} b_{1,pc}^s \\ b_{2,pc}^s \\ \cdot \\ \cdot \\ \cdot \\ b_{34,pc}^s \end{bmatrix} = \left[\text{Var}(\mathbf{b}_{pc}^s) \right]^{\frac{1}{2}} = \begin{bmatrix} 0.0636 \\ 0.0440 \\ \cdot \\ \cdot \\ \cdot \\ 0.0473 \end{bmatrix} \quad (53)$$

The standard error of the principal component estimators of the coefficients for the original climatic variables is

$$\text{s.e.} \begin{bmatrix} b_{1,pc} \\ b_{2,pc} \\ \cdot \\ \cdot \\ \cdot \\ b_{34,pc} \end{bmatrix} = \begin{bmatrix} 0.0636/7.582 \\ 0.0440/7.306 \\ \cdot \\ \cdot \\ \cdot \\ 0.0473/1.915 \end{bmatrix} = \begin{bmatrix} 0.008389 \\ 0.006016 \\ \cdot \\ \cdot \\ \cdot \\ 0.024702 \end{bmatrix} \quad (54)$$

Standard error computation using Fritts' (1976) method is performed to compare it with the result obtained in Eq. 54. The diagonal matrix \mathbf{U} is

$$\mathbf{U}_{4 \times 4} = \begin{bmatrix} 0.01632 & 0 & 0 & 0 \\ 0 & 0.01751 & 0 & 0 \\ 0 & 0 & 0.01873 & 0 \\ 0 & 0 & 0 & 0.02855 \end{bmatrix}.$$

Hence,

$$\begin{aligned} \mathbf{S}_{34 \times 34} &= \mathbf{V}_{34 \times 4}^s (\mathbf{U}_{4 \times 4} \mathbf{U}_{4 \times 4}) \mathbf{V}_{4 \times 34}^{s'} \\ &= \begin{bmatrix} -0.31 & 0.21 & 0.11 & 0.23 \\ 0.06 & 0.14 & 0.28 & 0.06 \\ 0.31 & -0.04 & -0.01 & 0.05 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0.14 & -0.20 & 0.05 & 0.20 \end{bmatrix} \begin{bmatrix} 0.000266 & 0 & 0 & 0 \\ 0 & 0.000307 & 0 & 0 \\ 0 & 0 & 0.000351 & 0 \\ 0 & 0 & 0 & 0.000815 \end{bmatrix} \begin{bmatrix} -0.31 & 0.06 & 0.31 & \cdot & \cdot & \cdot & 0.14 \\ 0.21 & 0.14 & -0.04 & \cdot & \cdot & \cdot & -0.20 \\ 0.11 & 0.28 & -0.01 & \cdot & \cdot & \cdot & 0.05 \\ 0.23 & 0.06 & 0.05 & \cdot & \cdot & \cdot & 0.20 \end{bmatrix} \\ &= \begin{bmatrix} 0.004047 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0.001932 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0.002237 \end{bmatrix} \end{aligned}$$

Notice that the diagonal elements of \mathbf{S} are the variances (or the square of standard errors) of the coefficients of the standardized climatic variables. The diagonal elements of \mathbf{S} are the same as the elements of the vector $Var(\mathbf{b}_{pc}^s)$ given in Eq. 52.

Response Function and Comparison of the Inferential Procedures. Response functions developed using two procedures are compared. First, a response function was developed using principal components selected by employing the cumulative eigenvalue product rule followed by further screening of components significant at 15 percent level. The other response function was developed using OLS (this is the same as developing a response function using PCR with all components retained). For the first response function, significance of the estimated coefficients were assessed (at 5 percent level) using the t-statistic given by Mansfield et al. (1977) and the F-statistic given by Fritts (1976). For the response function developed using OLS, test of significance was assessed using the classical test procedure. The results are given in Table 5.

The values of the principal component estimators for the standardized climatic variables and their 95 percent confidence intervals were plotted by month for both temperature and precipitation (Fig. 10). Figure 10 and Table 5 indicate that the two inferential procedures (t and F-statistic) yield nearly similar results but there are some differences. First, the F-statistic tends to give a larger number of significant climatic variables than the t-statistic. Second, there are four variables where the two procedures yield different results in terms of significance. These are precipitation of current September, and temperature of current August, September, and prior May. The critical value of the F-statistic depends on the number of the principal components retained for developing the response function whereas the t-statistic is only a function of the length of the chronology and the number of parameters in the model. Therefore, as the number of principal components varies, the two procedures yield different results. When the response function is estimated using OLS (classical

Table 5.—Sign and statistical significance of climatic variables, precipitation and temperature, by month, identified using the t-statistic given by Gunst and Mason (1980) and Mansfield et al. (1977), and the F-statistic given by Fritts (1976). The last row (classical method) shows significant variables obtained after the model is estimated using OLS. An asterisk indicates climatic variables significant at 5 percent level

Inferential method	No. of significant variables	<u>Prior year</u>												<u>Current year</u>											
		May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep							
t-statistic ^a	9	-	+	+	-	-	-	+	*	+	*	+	*	+	*	+	*	-							
F-statistic ^b	10	-	+	+	-	-	-	+	*	+	*	+	*	+	*	+	*	-							
Classical method	3	-	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-							
		<u>Precipitation</u>																							
t-statistic	7	-	-	-*	-	+	+	+	+	+	+	+	+	+	+	+	+	+							
F-statistic	8	-*	-	-*	-	+	+	+	+	+	+	+	+	+	+	+	+	+							
Classical method	1	-	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+							
		<u>Temperature</u>																							
t-statistic	7	-	-	-*	-	+	+	+	+	+	+	+	+	+	+	+	+	+							
F-statistic	8	-*	-	-*	-	+	+	+	+	+	+	+	+	+	+	+	+	+							
Classical method	1	-	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+							

^aThe response function (final model) was developed using principal components selected according to the selection rule described in method E of Table 6.

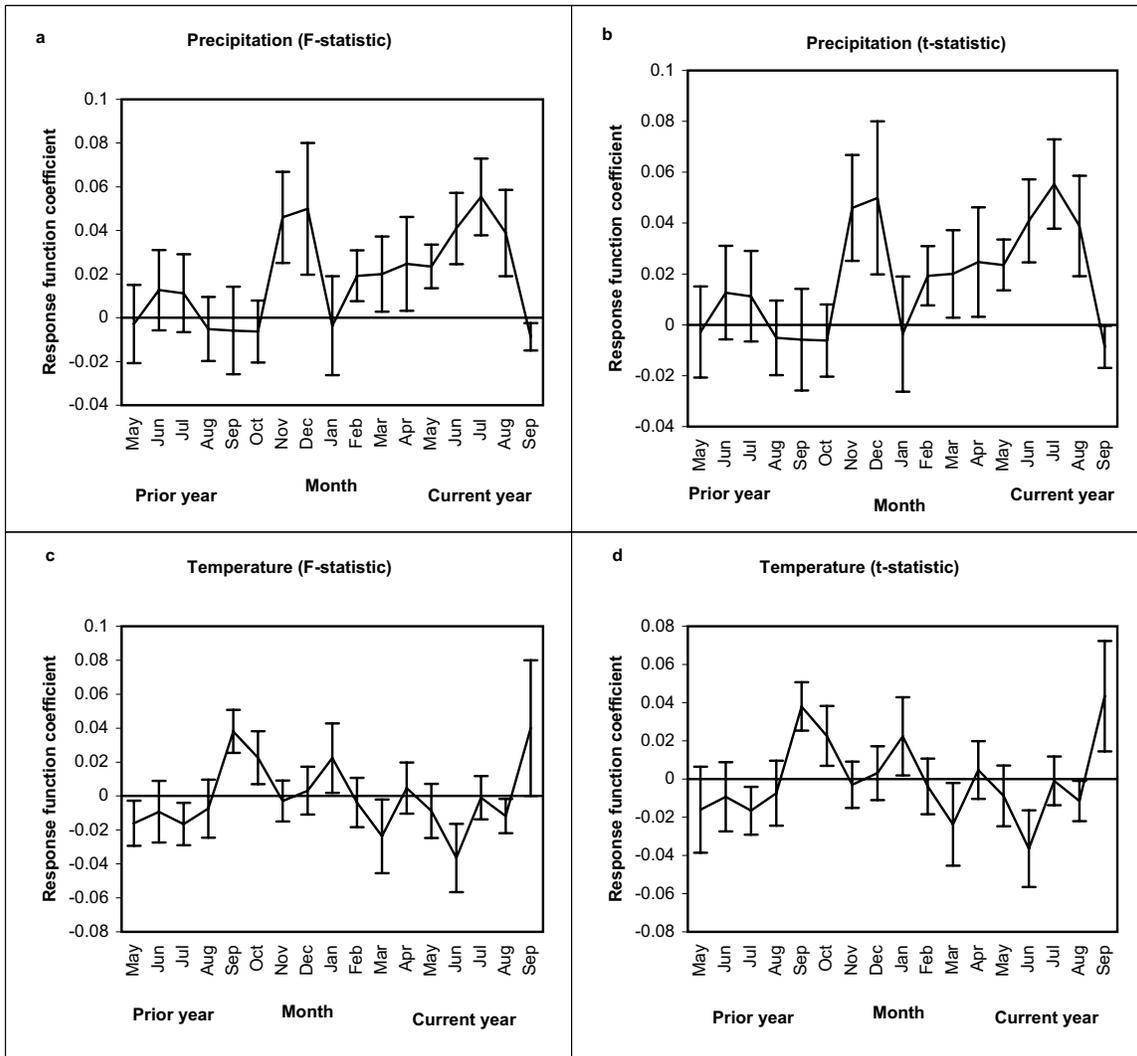


Figure 10.—Response function of yellow-poplar. The statistical significance of the parameters was tested according to the F-statistic (plots a and c) given by Fritts (1976) and the t-statistic (plots b and d) given by Mansfield et al. (1977) and Gunst and Mason (1980). Vertical bars represent the 95 percent confidence interval for each of the response function elements. Significant elements are those for which the confidence interval does not cross the zero line.

method) there are only four significant variables (Table 5). This was the consequence of multicollinearity that inflated the standard error of the estimated coefficients. Besides picking fewer numbers of significant variables, the classical method also suggests some unrealistic results, such as larger radial growth when the temperature of current July is higher than average. There are seven variables where the sign of the predicted coefficient is opposite from that provided by the other two methods. This is one of the main reasons for using PCR than OLS to study radial growth of trees in relation to climate.

Sensitivity of the Response Function to Principal Components Selection Rules. The effect of the various methods of selecting principal components on R^2 of the final model is compared in Table 6. The methods select principal components to be included into the model of Eq. 20 based on five different criteria and are described as follows:

- A. selects the first components which explain 85 percent of the total variance in the original data;

Table 6.—Estimated R² values of response functions developed using five methods of selecting principal components

Selection method	Selection criteria	No. of principal components selected	R ² _{climate} (%) of the final model
A	The first r components that satisfy $\frac{\sum_{j=1}^r \lambda_j}{k} \geq .85$	18	55
B	All components with $\lambda_j > 1$	13	52
C	The first r components that satisfy $\prod_{j=1}^r \lambda_j > 1$	25	57
D	Those components significant at 5 percent level	6	43
E	Apply method (c) first then select those components significant at 15 percent level	11	49
OLS	All components	34	67

B. selects only those components with eigenvalues greater than unity;

C. selects the first components whose combined eigenvalue product is greater than unity;

D. selects only those components significant at 5 percent level; and

E. first applies the eigenvalue product rule C to select the principal components and further screens these selected components using a significance level of 15 percent.

To accomplish the selection in D, the response was first regressed against the 34 principal components. For E, the method requires regressing the response against only those components selected using the product rule. As a reference, R² also was calculated for the final model obtained using all the principal components. This is equivalent to performing OLS.

Following each selection method, the response was regressed against the selected components. Table 6 shows the number of selected components and R² for the fitted models. Selection rules described in D and E select fewer principal components and still have similar measures of fit (R²) as the first three methods that include a larger number of components. Retaining fewer components results in smaller total variance as shown in Table 6. But more importantly, keeping the components with large eigenvalues does not necessarily reduce the model variance. In fact, the last two selection rules, which retain fewer components with smaller eigenvalues, seem to be more effective in reducing the error variance (noise) in the model.

The selection procedure in E of Table 6 is commonly practiced in dendrochronology. Comparison of this method with OLS shows that the two procedures produce noticeable differences, especially in terms of R² (Table 6) and number of significant climatic variables (Table 5). Table 5 also shows that the two methods yield opposite signs for seven of the climatic variables. To compare the magnitude of the two estimates, the absolute value of the principal component estimates was subtracted from the absolute value of the least squares estimates. The difference in magnitude was plotted in Figure 11a. It shows that the OLS estimates are larger in magnitude for 21 of the climatic variables. The standard errors of the least squares and principal component estimators are plotted in Figure 11b. The

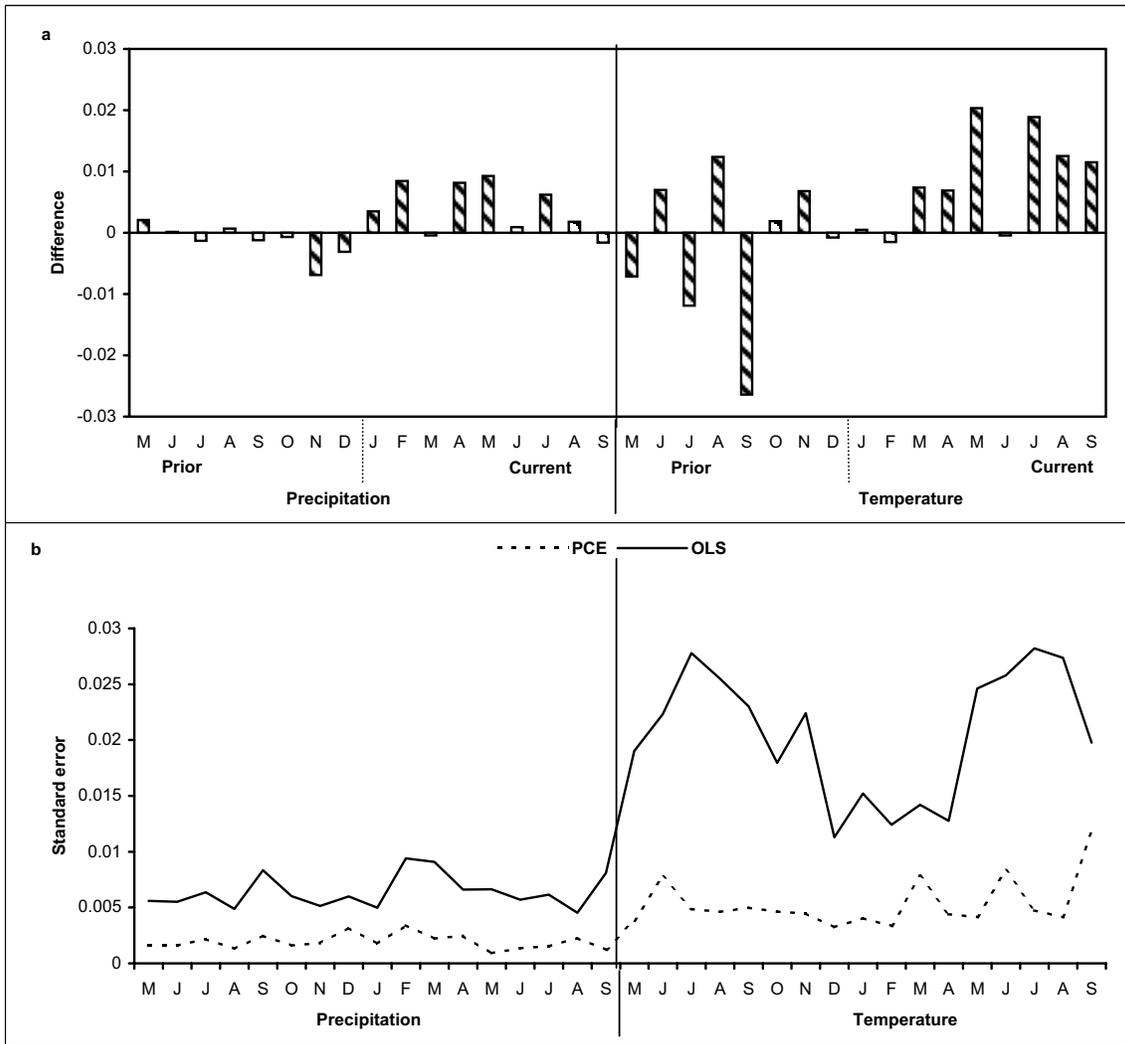


Figure 11.—Comparison of least squares (OLS) and principal component estimators (PCE) of the climatic variables. Plot a shows the difference in magnitude between the two estimates and b compares their standard errors.

principal component estimators have lower standard errors for all climatic variables indicating that the procedure results in more precise and reliable coefficients of the climatic variables. The plot in Figure 11b also shows larger differences in standard errors of temperature variables than precipitation variables. This was because there were more significant correlations among the temperature variables (21 pair) than among precipitation variables (4 pair).

The principal component estimators of the standardized climatic variables are computed and compared for each selection method. The result showed that there are differences in sign and magnitude of the estimated coefficients attributable to the selection method. Comparison of standard errors of the estimated coefficients shows that method D yields estimates with smallest standard errors and method C gives estimates with largest standard errors. With respect to significance of the variables, the first three selection methods yield fewer significant variables than D or E. These two methods tend to provide similar results. Generally, this illustrates that the sign, magnitude, and statistical significance of the estimated coefficients of the response function depends on the method used to select or eliminate the principal components. This issue deserves special consideration in growth-climate studies since studies with similar objective use different selection rules (Pan et al. 1998, Lindholm et al. 2000).

Summary and Conclusions

The theoretical basis, the procedure, and the application of PCR in dendroecology have been illustrated. Selection of principal components is the most important step and makes the procedure unique. Depending on the selection rule, the percentage of variation in annual growth explained by climate varied from 43 to 67 percent. However, dendroecological studies are not consistent with respect to principal components selection rules. For example, Fritts (1976) and Lindholm et al. (2000) use the cumulative eigenvalue product rule followed by screening of components significant at 15 percent level (method E) while Pan et al. (1998) simply choose the first r components that explained certain preselected percentage of the variation in the original climatic data (method A). The method of selecting principal components is shown to affect all the important attributes of the resulting model used for interpreting the effects of climatic variables on radial growth.

While diverse methods for studying annual growth of trees in relation to climate exists, the method of principal components regression has been recognized as good tool for developing response functions. However, differences in the procedure (especially selection of principal components) lead to differences in interpretations regarding tree growth and climatic variables. The accuracy and reliability of the selection procedures has not been fully explored in dendrochronological literature. Based on the analysis made in this study, we draw the following conclusions:

- The difficulty of PCR is to decide how many principal components to introduce and what criterion to use for the selection. The comparison of the various selection procedures showed that important final model characteristics (sign, magnitude, and significance) of the coefficients of the climatic variables and R^2_{climate} vary significantly depending on the selection rule. This demonstrates that consistency in the choice of selection rule is necessary for results of dendroecological studies to be comparable. There is no universally accepted rule for selecting principal components and most current selection rules eliminate principal components that account for the largest variance in the original climatic data. This property is thought to be undesirable. Therefore, this issue deserves further investigation.
- This study has shown a complete derivation of the method used to estimate the standard errors of the principal component estimators for both the natural and transformed climatic variables.
- An appropriate inference procedure (t-statistic) to test the statistical significance of the regression coefficients obtained using PCR is introduced.
- Finally, provided that one has a regression dataset where the predictors exhibit the problem of multicollinearity, the results in this study can be used in any discipline to develop a more stable model, estimate standard errors of the principal component estimators, and test the statistical significance of the individual regressors.

Acknowledgment

The authors thank the following for reviewing an earlier draft of this manuscript: Sara Duke, Northeastern Research Station; William F. Christensen, Department of Statistical Science, Southern Methodist University; Robert A. Monserud, Pacific Northwest Research Station; Edward Cook, Lamont-Doherty Earth Observatory; Harry T. Valentine, Northeastern Research Station, and Mairtin Mac Siurtain, Faculty of Agriculture, University College, Dublin, Ireland.

Literature Cited

- Box, G.E.P.; Jenkins, G.M. 1976. **Time series analysis: forecasting and control**. Rev. ed. San Francisco: Holden Day. 553 p.
- Cook, E.R. 1985. **A time series analysis approach to tree-ring standardization**. Tucson: University of Arizona. 171 p. Ph.D. dissertation.
- Cook, E.R.; Jacoby, G.C. Jr. 1977. **Tree-ring-drought relationships in the Hudson Valley, New York**. Science. 198: 399-401.
- Draper, N.R.; Smith, H. 1981. **Applied regression analysis. 2nd edition**. New York: John Wiley and Sons. 709 p.
- Fekedulegn, Desta. 2001. **Spatial and temporal analysis of radial growth in an Appalachian watershed**. Morgantown, WV: West Virginia University. 270 p. Ph.D. dissertation.
- Fritts, H.C. 1976. **Tree rings and climate**. New York: Academic Press. 567 p.
- Fritts, H.C.; Blasing, T.J.; Hayden, B.P.; Kutzbach, J.E. 1971. **Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate**. Journal of Applied Meteorology. 10: 845-864.
- Granger, C.W.J.; Morris, M.J. 1976. **Time series modeling and interpretation**. Journal of the Royal Statistical Society. 139: 246-257.
- Guiot, J.; Berger, A.L.; Munaut, A.V. 1982. **Response functions**. In: Hughes, M.K.; Kelly, P.M.; Pilcher, J.R.; LaMarche, V.C., eds. Climate from tree rings. Cambridge, UK: Cambridge University Press: 38-45.
- Gunst, R.F.; Mason, R.L. 1980. **Regression analysis and its application: a data-oriented approach**. New York: Marcel Dekker. 402 p.
- Lindholm, M.; Lethonen, H.; Kolstrom, T.; Merilainen, J.; Eronen, M.; Timonen, M. 2000. **Climatic signals extracted from ring-width chronologies of Scots pines from the northern, middle and southern parts of the boreal forest belt in Finland**. Silva Fennica. 34: 317-330.
- Loehlin, J.C. 1992. **Latent variable models: an introduction to factor, path, and structural analysis**. Hillsdale, NJ: Lawrence Erlbaum Associates. 292 p.
- Mansfield, E.R.; Webster, J.T.; Gunst, R.F. 1977. **An analytic variable selection technique for principal components regression**. Applied Statistics. 6: 34-40.
- Monserud, R.A. 1986. **Time series analysis of tree-ring chronologies**. Forest Science. 32: 349-372.
- Myers, R.H. 1986. **Classical and modern regression with applications**. Boston: Duxbury Press. 359 p.
- Pan, C.; Tajchman, S.J.; Kochenderfer, J.N. 1997. **Dendroclimatological analysis of major forest species of the central Appalachians**. Forest Ecology and Management. 98: 77-87.
- SAS Institute, Inc. 1999. **SAS/STAT user's guide, version 6. 4th ed. Vol. 2**. Cary, NC: SAS Institute. 883 p.
- Visser, H.; Molenaar, J. 1990. **Estimating trends in tree-ring data**. Forest Science. 36: 87-100.

Appendix A: SAS codes to fit the modified negative exponential model to a ring-width series.

The following code provides the means to create the ring-width index series from a set of raw ring width measurements using SAS (1999).

```
TITLE1 'Detrending the ring-width series of yellow-poplar';
TITLE2 'Model fit: The Modified Negative exponential model';

DATA ring_width;
  INPUT Age Width ;
  LABEL Age="Age in years";
  LABEL Width="Ring width in mm";
  CARDS;
1      3.46
2      3.21
3      4.11
.      .
.      .
65     0.96
;
  RUN;

PROC PRINT;

PROC MEANS MEAN STD CV NMISS MIN MAX ;

PROC NLIN DATA=ring_width BEST=10 PLOT METHOD=MARQUARDT MAXITER=200;
  PARSMS B0=xxx B1=xxx B2=xxx;
  MODEL Width=B0*EXP(-B1*Age)+B2;
  DER.B0=exp(-B1*Age);
  DER.B1=(-B0*Age)*exp(-B1*Age);
  DER.B2=1;
  OUTPUT OUT=result P=predicted R=residual;

PROC PRINT DATA=result;

PROC PLOT DATA=result;
  PLOT width*age='W' predicted*age='P' /OVERLAY;
  PLOT residual*age / VREF=0;

DATA result_2;
  MERGE ring_width result;
  BY age;
  ring_width_index=width/predicted;
  RUN;

PROC DBLOAD DBMS=XLS DATA=result_2;
  PATH='a:\SAS_OUTPUT_3.xls';
  PUTNAMES=yes;
  LOAD;

PROC PLOT DATA= result_2;
  PLOT ring_width_index*age='I';

PROC PRINT DATA= result_2;

RUN;
QUIT;
```

Explanation of the Code

1. This data step reads a ring-width data and creates a temporary SAS dataset called “ring_width”. This dataset contains two variables, age in years and annual ring widths to the nearest 0.01 mm. Here data are read directly within the data step from card images. It also lists and prints summary statistics for the two variables.
2. This step carries out nonlinear regression and fits the modified negative exponential model to the ring-width series. In this program the user should specify starting values for each of the three parameters prior to executing the program. On the **PARMS** line, BO=xxx indicates that the user must specify a starting value for BO (for example BO =5.5). Reasonable starting values here are BO =7.0, B1=0.05, and B2=1.0 for most data. BO is the value of the detrending line at age 0, B1 is the slope parameter (natural logarithm $\ln[B0/B2] / \text{MaxAge}$, where MaxAge is the length of series), and B2 is the lower limit or asymptote of the detrending line. You also can use the largest ring width for BO and the smallest for B2 starting values. After fitting the model, the outputs from the program (the predicted values, residuals, etc.) are saved in a SAS dataset called “result”. The contents of the “result” dataset are printed and the program also plots ring width and predicted ring width against age in the same graph.
3. This step creates a SAS dataset called “result_2”. This dataset contain all the variables in data sets “ring_width” and “result”. Then the code calculates ring-width index as a ratio of actual ring width to predicted ring width.
4. This procedure saves the dataset “result_2” in an Excel spreadsheet producing five variables: age, actual ring width (measurements), predicted ring width from the fitted model, the residuals, and the ring-width index (RWI). This provides the user with the ability to review and compare data among individual trees, produce graphs, and perform other operations on these data external to SAS.
5. This step creates a plot of ring-width index against age and also prints the contents of the SAS dataset “result_2”.

The SAS code for this and each of the other appendices, along with the example data can be found at: <http://www.fs.fed.us/ne/morgantown/4557/dendrochron>

Appendix B: Modeling to remove autoregressive trend in ring-width index series.

Here is the code for creating the first three partial autoregressive model functions, the autocorrelation coefficients, and following the decision to use a model of order 2, the final data are generated and saved.

```
DATA input; ①
  SET result_2 ;
  time = age+1934;
  LABEL time=" The Calender year";
  rwi = ring_width_index;
  LABEL rwi=" Ring Width Index";
  RUN;

TITLE 'An Intercept Model Fitted';

PROC AUTOREG DATA=input PARTIAL; ②
  MODEL rwi= /;
  RUN;

TITLE 'A First Order AR Model Fitted';

PROC AUTOREG DATA=input PARTIAL; ③
  MODEL rwi= / NLAG=1;
  RUN;

TITLE 'A Second Order AR Model Fitted';

PROC AUTOREG DATA=input PARTIAL; ④
  MODEL rwi= / NLAG=2;
  RUN;

TITLE 'A Third Order AR Model Fitted';

PROC AUTOREG DATA=input PARTIAL; ⑤
  MODEL rwi= / NLAG=3;
  RUN;

TITLE 'The Autocorrelation and partial autocorrelation coefficients';

PROC AUTOREG DATA=input PARTIAL; ⑥
  MODEL rwi= /;
  LAGLIST 1 2 3 4 5 6 7 8 9 10 11 12 13 14
          15 16 17 18 19 20;
  RUN;

TITLE 'Autoregressive Modeling for the Chosen AR Model'; ⑦
TITLE2 'Assuming a Second Order AR Model is Appropriate';

PROC AUTOREG DATA=input PARTIAL;
  MODEL rwi= /NLAG=2 ;
  OUTPUT OUT=result P=PRED R=RESID;
  RUN;

PROC PRINT DATA=result;
  RUN;

PROC DBLOAD DBMS=xls DATA=result; ⑧
  PATH='c:\Data-Auto\out2.xls';
  PUTNAMES=yes;
  LOAD;
  RUN;
```

Explanation of Code

1. The data step reads the data and creates a SAS dataset called “input”. Dataset “input” contains two new variables: the calendar year and the corresponding ring-width index (RWI) for that year, both derived from the output generated in Appendix A.

2. This code fits an intercept model. That is no autoregressive terms are used. The model fitted has the form

$$RWI_t = b_0 + \text{error}$$

where RWI_t is RWI at time t.

3. This code fits a first order autoregressive model

$$RWI_t = b_0 + b_1 (RWI_{t-1}) + \text{error}$$

The output gives the estimated coefficient (b_1) and its significance, the first order autocorrelation coefficient and other summary statistics such as R^2 , the AIC (Akaike Information Criterion) etc. Note that the sign of the autoregressive parameters from PROC AUTOREG are reversed. For example if the output gives $b_1 = -0.386$, then the first order autoregressive parameter is 0.386. User should record summary statistics of the fitted model (R^2 , AIC, etc) from the output with heading “Yule-Walker Estimates”.

4. This code fits a second order autoregressive model

$$RWI_t = b_0 + b_1 (RWI_{t-1}) + b_2 (RWI_{t-2}) + \text{error}$$

RWI_{t-2} is the ring-width index from two years prior. The output gives the estimated coefficients (b_1 , b_2) and their significance, the first and second order autocorrelation and partial autocorrelation coefficients and other summary statistics of the fitted model such as R^2 , AIC, etc. Note that the sign of the autoregressive parameters from PROC AUTOREG are reversed. User should record summary statistics of the fitted model from the output with heading “Yule-Walker Estimates”.

5. This code fits a third order autoregressive model

$$RWI_t = b_0 + b_1 (RWI_{t-1}) + b_2 (RWI_{t-2}) + b_3 (RWI_{t-3}) + \text{error}$$

The output gives the estimated coefficients (b_1 , b_2 , b_3) and their significance, the first, second, and third order autocorrelation and partial autocorrelation coefficients, and other summary statistics of the fitted model such as R^2 and AIC. User again should record summary statistics of the fitted model from the output with heading “Yule-Walker Estimates”.

6. This code calculates autocorrelations and partial autocorrelations for lags 1 to 20. These coefficients should be plotted against lag to help determine the correct AR model for the particular ring-width series. These plots in conjunction with comparison of the three models (from 3 to 5) based on significance of the coefficients, R^2 , AIC should be used to decide the correct autoregressive model for each ring-width index series.

7. Suppose a second order AR model is found to be reasonable for a particular ring-width series. This code fits a second order AR model and saves the predicted ring-width index and the residuals (the prewhitened ring-width index = PRWI) in a SAS dataset named “result” and also prints the contents of this dataset.

8. This code saves the contents of the SAS dataset “result” in an Excel spreadsheet file named “Out2.xls”. The spreadsheet contains the following variables: time, RWI, predicted RWI, and the residuals (PRWI). This is useful for plotting the fitted model and the residuals.

Appendix C: SAS Program To Develop Response Function

The following steps generate the principal components from the raw data:

- a. utilize the eigenvalues to decide how many principal components to retain, i.e., derive the reduced set of eigenvectors and principal components associated with the product selection rule;
- b. screen these to arrive at the final set of vectors;
- c. develop the final model and compute the significance of the original variables as predictors of climate effects on radial growth. First is the annotated SAS code, followed by the notes and related outputs. The numbers inside circles within the code listing refer to the numbered notes that follow the listing.

```
OPTIONS PS=56 linesize=150 nodate nonumber ; /* SYMBOLGEN;
          Add to get values of macro variables placed in log.
          This will cause the log to be longer and less easily read.*/

/* Library definition describes where we want to retrieve or to archive datasets.*/
LIBNAME L 'C:\PCR\8-Appendix';
DATA Response_Function;                                ①
    SET L.climate_data;
    RUN;

TITLE 'Create the principal components from the initial climate data for the 62 years.';
PROC PRINCOMP DATA=Response_Function OUT=L.Result_1 N=34 PREFIX=Z OUTSTAT = L.Result_2;
    VAR May_PP Jun_PP Jul_PP Aug_PP Sep_PP Oct_PP Nov_PP Dec_PP Jan_CP Feb_CP Mar_CP
        Apr_CP May_CP Jun_CP Jul_CP Aug_CP Sep_CP May_PT Jun_PT Jul_PT Aug_PT Sep_PT
        Oct_PT Nov_PT Dec_PT Jan_CT Feb_CT Mar_CT Apr_CT May_CT Jun_CT Jul_CT Aug_CT
        Sep_CT ;
    RUN;

/*Here we find the number of eigenvectors to keep. Using the
product rule, the GLOBAL variable, num, will contain the count of eigenvectors*/

%GLOBAL num ;

DATA temp;                                            ③
    SET L.Result_2;
    IF _TYPE_ = "EIGENVAL" THEN
        DO;
            ARRAY eigen{34} May_PP Jun_PP Jul_PP Aug_PP Sep_PP Oct_PP Nov_PP Dec_PP
                Jan_CP Feb_CP Mar_CP Apr_CP May_CP Jun_CP Jul_CP Aug_CP Sep_CP May_PT
                Jun_PT Jul_PT Aug_PT Sep_PT Oct_PT Nov_PT Dec_PT Jan_CT Feb_CT Mar_CT
                Apr_CT May_CT Jun_CT Jul_CT Aug_CT Sep_CT ;
            prod = 1.0;
            i = 1;
            DO WHILE(prod GE 1.0);
                prod = prod*eigen{i};
                i = i + 1;
            END;
            CALL SYMPUT('num',i - 2); /*num is two less than the exit value of i. */
        END;
    RUN;
```

```

/* Use a macro to build the list Z[i] of the first "num" Z's; call the list "depends." */
%MACRO depends;
  %LOCAL units;
  %DO units=1 %TO &num;
    Z&units
  %END;
%MEND depends;

TITLE "The multiple regression of the prewhitened chronology using &num principal components";

PROC REG DATA=L.Result_1;                                ④
  MODEL chron= %depends;
  RUN;

TITLE 'The multiple regression of the prewhitened chronology with the 4 most important
principal components';
PROC REG DATA=L.Result_1;                                ⑤
  MODEL chron= Z8 Z20 Z6 Z9 ;
  RUN;

TITLE 'A stepwise regression of the prewhitened chronology with the 25 principal components';
PROC STEPWISE data=L.Result_1;                             ⑥
  model chron= %depends;
  RUN;

PROC DBLOAD DBMS=xls DATA=L.Result_1;                    ⑦
  PATH='c:\PCR\8-Appendix\prin.xls';
  PUTNAME=yes;
  LOAD;
  RUN;

PROC DBLOAD DBMS=xls DATA=L.Result_2;
  PATH='C:\PCR\8-Appendix\eigen.xls';
  PUTNAME=yes;
  LOAD;
  RUN;

QUIT;

```

Explanation of the SAS Code

1. This step reads data from the file “climate_data.sas7bdat” which is located in folder path “C:\PCR\8-Appendix” that is designated as the SAS Library “L” on the previous line. Data are read as follows: first column is the year, the second is the prewhitened chronology, and columns 3 to 36 contain the monthly climatic data. In this analysis, monthly total precipitation and monthly mean temperature from May of the prior year to September of the current year are used. The user may choose to compile and read these data in various ways. See the text for details on the weather data source.

2. This step calculates the 34 by 34 correlation matrix of the climatic variables, the 34 eigenvalues and the 34 by 34 eigenvector matrix, and the 34 orthogonal variables called principal components. The program names the principal components using the prefix=Z. That is, the 34 components are named as Z1, Z2, ..., Z34. The 34 principal components as well as all the original variables from the SAS dataset file “climate_data.sas7bdat” are now stored in a SAS dataset called “Result_1.sas7bdat.” The SAS dataset “Result_2.sas7bdat” is saved to the same location on disk (SAS automatically adds the file extension “.sas7bdat”); it contains the correlation matrix, the eigenvalues and the eigenvectors associated with each eigenvalue.

3. Next we calculate the number of eigenvalues, eigenvectors, and associated principal components to retain using the eigenvalue product rule. The global variable “num” is used to store this number.

4. First stage of principal components elimination: Using the eigenvalues provided in step 2 and the number calculated in step 3, we retain the first 25 principal components.

This step further screens the remaining 25 components as follows: Use Proc REG to fit a multiple regression model relating the prewhitened chronology with these 25 components and eliminate those components that are not significant at probability level of 0.15. In other words, examine the t-statistic for each of the 25 coefficients and eliminate those that have the smallest t-statistic. The display below shows the portion of the program output with the 25 principal components parameter fit statistics. In this example we chose to include the Z9 and found that it satisfied the 15 percent criterion.

```

The multiple regression of the prewhitened chronology using 25 principal components

                                The REG Procedure
                                Model: MODEL1
                                Dependent Variable: chron

                                Analysis of Variance

Source                            DF          Sum of
                                Squares          Mean
                                Square          F Value          Pr > F

Model                            25          0.83406          0.03336          0.87          0.6395
Error                            36          1.38353          0.03843
Corrected Total                   61          2.21759

                                Root MSE          0.19604          R-Square          0.3761
                                Dependent Mean    1.00002          Adj R-Sq         -0.0571
                                Coeff Var         19.60363

                                Parameter Estimates

Variable          DF          Parameter
                                Estimate          Standard
                                Error          t Value          Pr > |t|

Intercept        1          1.00002          0.02490          40.17          <.0001
Z1                1          -0.00057485     0.01242          -0.05          0.9634
Z2                1          -0.01055        0.01487          -0.71          0.4827
Z3                1          -0.00336        0.01607          -0.21          0.8357
Z4                1          -0.00266        0.01665          -0.16          0.8741
Z5                1          0.01692         0.01756          0.96          0.3417
Z6                1          0.02824         0.01876          1.51          0.1409
Z7                1          -0.00443        0.01931          -0.23          0.8200
Z8                1          -0.05519        0.02012          -2.74          0.0094
Z9                1          0.03001         0.02153          1.39          0.1719
Z10               1          0.01341         0.02191          0.61          0.5444
Z11               1          0.00317         0.02249          0.14          0.8886
Z12               1          -0.01473        0.02356          -0.62          0.5359
Z13               1          0.01512         0.02486          0.61          0.5468
Z14               1          0.02216         0.02568          0.86          0.3939
Z15               1          -0.01222        0.02618          -0.47          0.6434
Z16               1          0.01487         0.02793          0.53          0.5977
Z17               1          0.00855         0.02817          0.30          0.7631
Z18               1          -0.00029682     0.02970          -0.01          0.9921
Z19               1          -0.00684        0.03016          -0.23          0.8218
Z20               1          -0.05502        0.03281          -1.68          0.1022
Z21               1          0.01438         0.03467          0.41          0.6807
Z22               1          0.03536         0.03651          0.97          0.3393
Z23               1          0.04721         0.03823          1.24          0.2248
Z24               1          0.02500         0.04181          0.60          0.5536
Z25               1          -0.01132        0.04399          -0.26          0.7984

```

One can see from this output that Z8, Z20, Z6, and possibly Z9 should be considered.

5. Second stage of principal components elimination: Further screening of the 25 components (in the step above) resulted in retention of only four components. These four components will be used to develop the response function. The linear regression program (PROC REG) fits a multiple regression model relating the prewhitened chronology with four components. The estimated coefficients of the four components, their standard errors and the mean square error for the fitted model are very important for further analysis. Note that the four components are not necessarily the first four components with largest eigenvalue. These components are strongly related to the response (the prewhitened chronology). The output from this step should be recorded for further analyses. The output includes the estimated coefficients of the four components, which are denoted by $\hat{\boldsymbol{\alpha}}_l$ (see step 8 of the step-by-step procedure in the text), and the estimated standard errors of the four coefficients, which are denoted by $\boldsymbol{\kappa}$. Recall that in step 9 of the step-by-step procedure (see the text), \mathbf{V}_l represents a 34 by l matrix containing eigenvectors corresponding these four retained components derived from the original 34 climate variables. To obtain the elements of this matrix, examine the complete eigenvector matrix in “Result_2.sas7bdat” and select those vectors corresponding the four retained components. The user may choose to carry out the computation in steps 9 and 10 using SAS, a spreadsheet, or hand calculator.

6. (Optional) Historical Comparison: Traditionally, the last screening of principal components in most dendroecological studies was performed by relating the growth response with the principal components using stepwise regression. This step is provided to demonstrate that the same result is obtained in either case. This step is redundant since the principal components are orthogonal and hence the significance of a particular component is not affected by exclusion or inclusion of another component.

7. Finally, the data generated from the study are saved to an Excel spreadsheet format to permit further computations, as described in the text.

Fekedulegn, B. Desta; Colbert, J.J.; Hicks, R.R., Jr.; Schuckers, Michael E. 2002. **Coping with Multicollinearity: An Example on Application of Principal Components Regression in Dendroecology**. Res. Pap. NE-721. Newton Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station. 43p.

The theory and application of principal components regression, a method for coping with multicollinearity among independent variables in analyzing ecological data, is exhibited in detail. A concrete example of the complex procedures that must be carried out in developing a diagnostic growth-climate model is provided. We use tree radial increment data taken from breast height as the dependent variable and climatic data from the area as the independent data. Thirty-four monthly temperature and precipitation measurements are used as potential predictors of annual growth. Included are monthly average temperatures and total monthly precipitation for the current and past growing season. The underlying theory and detail illustration of the computational procedures provide the reader with the ability to apply this methodology to other situations where multicollinearity exists. Comparison of the principal component selection rules is shown to significantly influence the regression results. A complete derivation of the method used to estimate standard errors of the principal component estimators is provided. The appropriate test statistic, which does not depend on the selection rule, is discussed. The means to recognize and adjust for autocorrelation in the dependent data is also considered in detail. Appendices and directions to internet-based example data and codes provide the user with the ability to examine the code and example output and produce similar results.

Keywords: Multicollinearity, PCR, PC selection methods, radial growth, climate





Headquarters of the Northeastern Research Station is in Newtown Square, Pennsylvania. Field laboratories are maintained at:

Amherst, Massachusetts, in cooperation with the University of Massachusetts

Burlington, Vermont, in cooperation with the University of Vermont

Delaware, Ohio

Durham, New Hampshire, in cooperation with the University of New Hampshire

Hamden, Connecticut, in cooperation with Yale University

Morgantown, West Virginia, in cooperation with West Virginia University

Parsons, West Virginia

Princeton, West Virginia

Syracuse, New York, in cooperation with the State University of New York, College of Environmental Sciences and Forestry at Syracuse University

Warren, Pennsylvania

The U. S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, gender, religion, age, disability, political beliefs, sexual orientation, and marital or family status. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (Braille, large print, audiotape, etc.) should contact the USDA's TARGET Center at (202)720-2600 (voice and TDD).

To file a complaint of discrimination, write USDA, Director, Office of Civil Rights, Room 326-W, Whitten Building, 14th and Independence Avenue SW, Washington, DC 20250-9410, or call (202)720-5964 (voice and TDD). USDA is an equal opportunity provider and employer.

“Caring for the Land and Serving People Through Research”