

FOR 496 / 796 Introduction to Dendrochronology

Lab exercise #4: Tree-ring Reconstruction of Precipitation

Adapted from an exercise developed by
M.K. Cleaveland and David W. Stahle, Tree-Ring Laboratory, University of Arkansas.

Introduction

This exercise introduces students to the quantitative reconstruction of climate from tree-ring data, using correlation and regression analysis. You are supplied with a tree-ring chronology and seasonalized climate data from California (Figure 1). Only 10 years of data are supplied so you can actually complete all the computations involved by hand, with the assistance of a calculator or, of course, the use of a computer spreadsheet program.

Tree-Ring Data (the *predictor*):

In this exercise, the tree-ring predictor is actually an average of four ring-width chronologies of California blue oak (*Quercus douglasii*). Standard ring-width chronologies were produced by first detrending and standardizing the individual ring-width measurement series. This transforms each measurement series to a dimensionless index series with a mean and variance that are relatively stable over time, and a new overall mean of 1.0, which is the same for all series. These standardized ring width indices are then averaged together on a year-by-year basis to complete the standard chronology. The so-called residual ring-width chronology is actually used in this exercise, where the low order autocorrelation (persistence) has been removed with autoregressive (AR) modeling procedures [see Fritts (1976), Cook (1985; 1987), or Cook and Kairiukstis (1990)].

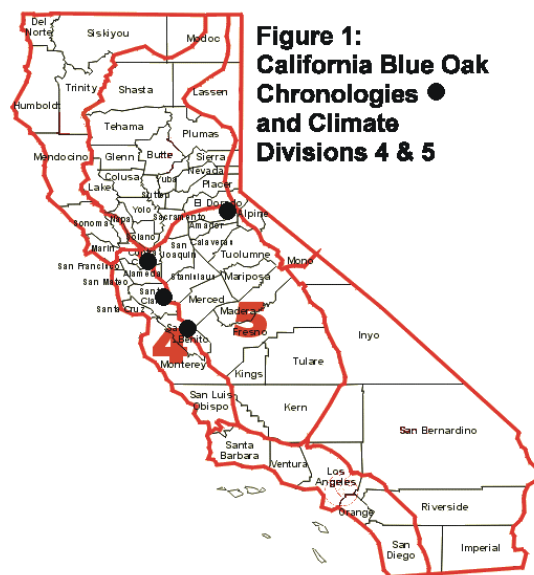


Figure 1:
California Blue Oak
Chronologies ●
and Climate
Divisions 4 & 5

Climate Data (the *predictand*):

The climate data, or *predictand*, used for this exercise are total precipitation (in mm) for the period December of the previous year ($t-1$) through April of the year concurrent with growth (t) averaged for the two nearest climate divisions in California (i.e., divisions 4 and 5, Karl et al. 1983). This December-April season was chosen because correlation analyses with the monthly data indicated that precipitation during this period has the greatest influence on blue oak growth. The precipitation data were modeled as an AR-0 process, meaning that successive December-April precipitation totals are not correlated (so-called white noise). For this reason, we use the residual tree-ring chronology in this exercise, because it is also an AR-0 process and matches the time series structure of the precipitation data.

Correlation:

Correlation is a basic statistical tool in dendrochronology, widely used to identify the monthly or seasonal climate variables related to ring width growth. Some of the computations involved in correlation analysis are also used to calculate the slope parameter of regression models, so we begin with a correlation exercise. We first provide the formula used to compute the Pearson correlation coefficient (r). We then present an exercise in Table 1, where you actually compute the correlation coefficient for ten years of paired tree-ring and precipitation data. We then test the statistical significance of the calculated correlation coefficient, and find the amount of variance shared by climate and ring-width.

The correlation coefficient varies between +1.0 and -1.0. A correlation of +1.0 means that the two series vary perfectly in phase, that is, when one series changes (positive or negative anomaly from its mean), the other changes with an anomaly of the same sign, and the changes are always exactly proportional. A correlation of -1.0 means that the two series are exactly out of phase and varying proportionally. Of course, real world data rarely achieve perfect correlation.

The Pearson correlation coefficient (r) is computed as follows:

$$r = \frac{\sum_{t=1}^n (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^n (x_t - \bar{x})^2 \sum_{t=1}^n (y_t - \bar{y})^2}}$$

where x_t is the tree-ring index value for year t , y_t is the climate variable (i.e., precipitation) for year t , \bar{x} and \bar{y} are the means on the tree-ring and climate values, respectively, and the subscript t varies from 1 to n (n equals the number of years of observation, $n=10$ in our simple example).

You can calculate r by filling in the cells in Table 2, or by using any spreadsheet or statistical software of your choosing.

Table 1: Sample data

Year	Tree-ring index	Precipitation (mm)
1941	1.76	716.7
1942	1.37	513.2
1943	0.98	468.8
1944	1.00	374.0
1945	0.92	342.0
1946	0.78	321.6
1947	0.73	211.7
1948	0.68	303.7
1949	0.91	345.8
1950	0.88	360.6

The sample data above produce a $r =$ _____.

Significance of the Correlation Coefficient

We can test the significance of our derived correlation coefficient in Table 3. Note that we have 10 observations, but only 9 degrees of freedom [degrees of freedom (df) = n - number of predictor variables (which always equals 1 in the case of correlation)]. So in Table 3, read the column for one independent variable and the row for nine degrees of freedom (under the column "Error df "). Note the magnitude of the correlation coefficient (r) needed to achieve statistical significance at the $\alpha = 0.05$ and $\alpha = 0.01$ probability levels. These levels are the probability of

finding a higher correlation coefficient between two unrelated series purely by chance. Note that if you use the $\alpha = 0.05$, there is a one in twenty chance that the correlation between two unrelated variables will be significant. If you use the more stringent $\alpha = 0.01$, only one correlation in 100 would be expected to be significant just by random chance.

So, is the correlation between the California ring-width index and precipitation series significant? (Y / N) _____
At what level? _____

Variance explained

The square of the Pearson correlation coefficient, known as the coefficient of determination (R^2), is used to estimate the proportion of the variance in one variable is explained by the covariance between the two correlated variables. So, in our 10-year sample, how much of the variance in ring-width is explained by the variance in precipitation.

$R^2 =$ _____

The percentage variance explained is $R^2 * 100$. Percent explained = _____?

Regression Analysis

We can use linear regression analysis to estimate or predict values of a dependent variable (y_t) from the corresponding value of an independent variable (x_t), with the subscript t referring to observations of x and y from the same year. This process is called *calibration*. To calibrate a tree-ring chronology with a climate time series, we simply make the climate series the dependent variable (y), and use the tree-ring chronology as the independent variable. Of course, climate is not really "dependent" on tree growth, quite the opposite is true. But, we suspend logic in this case and simply use the regression equation to obtain the quantitative estimates of precipitation from the tree-ring chronology. This regression-based calibration model is also referred to as a *transfer function*.

The best-fitting, straight-line equation or *model* for the linear regression is :

$$\hat{y}_t = \beta_0 + \beta_1 x_t$$

where β_0 is the intercept of the best-fitting straight line, β_1 is the slope of the best-fitting straight line, \hat{y}_t is our estimate or predicted value of the dependent variable (i.e., precipitation) in a given year t , based on the corresponding value of the independent tree-ring variable x_t in the same year t . In regression analysis, the line of best fit is the one which minimizes the sum (i.e., total) of the squared differences between each observed and predicted dependent variable [i.e., $\sum (y_t - \hat{y}_t)^2$].

If you are using a calculator, the formula to compute the estimate of the slope of the regression line is:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

The formula to compute the intercept (β_0) of the regression line is simply:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

You may, if you wish, estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ using a computer spreadsheet or statistics program.

Based on the data provided in Table 1, the best fitting, linear equation between seasonal precipitation and tree-ring indices is:

$$\hat{y}_t = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} x_t$$

Evaluation of the regression model

There are four types of analysis that are routinely performed to evaluate or test the strength of a regression model (equation) linking two variables, x and y . These include:

1. Graphical analysis I: scatterplot
2. Graphical analysis II: time series plot
3. Coefficient of determination (R^2)
4. Testing the statistical significance of the slope

In this exercise, we will evaluate our 10-year regression model with these techniques. Create a table similar to Table 4 (either on the page provided, or in a spreadsheet program) to assist with the evaluation.

1. Graphical analysis: Scatterplot

Using the graph paper provided (Figure 2), or using a computer graphics package, prepare an x-y scatterplot by plotting the intersections of the tree-ring values (x_t) as the abscissa (x -axis) and the climate values (y_t) as the ordinate (y -axis) taken from Table 1. Next, add the regression line to the scatterplot. The predicted values of precipitation (\hat{y}_t) fall exactly on the regression line. The regression line goes through the intersection of the means of x and y , so plot that point of Figure 2. Then, if you set x_t to zero in the regression model, the predicted value for \hat{y}_t equals the intercept (β_0). So plot the intercept on Figure 2. Use a straight edge to draw the regression line passing through those two points (intersection of the \bar{x} and \bar{y} , and the intercept). This regression line should pass roughly through the middle of your scatter of data points.

2. Graphical analysis: Time series plot

Using the graph paper provided (Figure 3), or using a computer graphics package, draw a time series plot of the observed (y_t) and reconstructed precipitation data (\hat{y}_t). The x -axis will be years used in the analysis (1941-1950) and the y -axis will be precipitation (in mm). The observed and reconstructed precipitation data have the same mean over this period, so the two time series should superimpose pretty nicely on the time series plot. Use a dashed line to connect the observed values and a solid line to connect the reconstructed values (or use different colours).

3. Coefficient of determination

Now we want to determine the amount of precipitation variance is explained by the ring-width data. For this, we determine the coefficient of determination (R^2) from the regression estimates. To compute R^2 , we calculate the *residuals from regression* and two *sums of squares* in Table 4. The residuals are also referred to as *error* since they represent the difference between our predicted tree-ring estimates of precipitation and the actual observed measurements of precipitation. These residuals should show up clearly in Figure 2 as the largest departures from

the fitted regression line, and in Figure 3 as the largest differences between the observed and reconstructed values. The regression residuals are calculated as:

$$\text{Residual} = (y_t - \hat{y}_t)$$

Place these residuals in column 5 of Table 4, then square each residual (column 6) and compute the *sum of the squares about the regression line, or Error (SSE)* at the bottom of column 6.

$$SSE = \sum (y_t - \hat{y}_t)^2$$

Next, compute the *sum of the squares due to regression (SSR)* using columns 7 & 8 of Table 4.

$$SSR = \sum (\hat{y}_t - \bar{y})^2$$

Now calculate the total corrected sum of squares (SST) in Table 4 as:

$$SST = SSE + SSR$$

This gives us the sums needed to calculate the coefficient of determination squared (R^2), the best measure of the relationship between the dependent and independent variables in a regression model. Using the values computed in Table 4 and above, we calculate R^2 as:

$$R^2 = SSR / SST$$

$$R^2 = \underline{\hspace{2cm}} / \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$$

This R^2 statistic tells us the proportion of variance in the dependent variable (y) accounted for, or explained by, the variance in the independent variable (x). The percent variance explained is simply $R^2 * 100$.

4. Testing the statistical significance of the slope

Testing the null hypothesis $H_0: \beta_1 = 0$ is another way to measure the strength of a regression model. If the slope is not significantly different than zero, then the R^2 is also probably low, and there may be no real relationship between the two variables. The sign of the slope is also telling us about the nature of the physical relationship between the two variables. With tree growth as the dependent variable (y), a positive slope with precipitation as the independent variable (x) simply indicates that the greater the precipitation, the greater the tree growth. But a regression between tree growth and temperature should be negative on moisture stressed sites, meaning that the higher the temperature, the lower the growth.

To test the significance of the slope (β_1), we form the null and alternate hypotheses:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0 \text{ for precipitation}$$

The test statistic is: $t = \frac{\hat{\beta}_1 - 0}{se_{\beta_1}}$

where se_{β_1} = standard error of $\beta_1 = \sqrt{\frac{SSE/(n-2)}{\sum (x_i - \bar{x})^2}}$

Therefore,

$$t = \frac{\text{_____}}{\text{_____}} = \text{_____}$$

Lookup the critical values of $t_{(n-2, \alpha)}$ in a t-table, where α is the level of significance for the analysis. Determine if we can reject the null hypothesis that the slope of the relationship could be equal to zero at an $\alpha = 0.05$.

Verification of the Reconstruction

Even if the regression model passes all the tests we have discussed above, it is imperative to check the reconstructed values against independent climate data outside of the calibration period used to develop the regression model. The reason is simple. The regression procedure finds the best model possible, and the mean of the observed and reconstructed series are forced to be the same [i.e., the means are removed from x and y in the computation of the slope, and the predicted values are scaled in terms of the dependent variable by the intercept]. But how well the tree-ring reconstruction actually tracks changes in the mean and variation of climate outside the calibration period is a major question, and can be evaluated to some degree with verification analysis using independent climate data.

One solution to the verification problem is to use only part of the climate data for calibration and the rest for verification (Snee 1977). Because the regression model tends to improve as the sample size increases, using part of the climate data just for verification may reduce the accuracy of the calibration. A better solution may be to conduct two calibration experiments on two halves of the climate/tree-ring time series, with each half used to develop a reconstruction. Then each reconstruction can be verified against the climate data not used in its calibration. If the regression coefficients in the two calibrations are not significantly different and the verification statistics are satisfactory, a third calibration using all the climate data can be used for the final reconstruction.

The analyses and statistics commonly used to test how closely the reconstruction resembles the independent climate data include:

1. Graphical analysis of the observed and reconstructed time series
2. Correlation analysis

In this exercise we will use the additional, independent precipitation data from 1931-1940 to check the accuracy of our tree-ring reconstruction during the verification

1. Graphical analysis of the observed and reconstructed time series

Use Table 3 to first compute the tree ring reconstruction of precipitation for 1931-1940, then draw a time series plot of both the observed and reconstructed precipitation data from 1931-1940 in Figure 5. These two time series should match very well if your regression model or “*transfer function*” is valid.

2. Correlation Analysis

The correlation coefficient (r) can be used to measure how much the observed and reconstructed time series covary during the verification period (usually a bit lower than during the calibration period). Use Table 5 to compute r between the reconstructed or predicted (\hat{y}_i) and observed (y_i) precipitation during the 1931-1940 verification period.

Critical Thinking Questions

1. What assumptions are automatically implied in the climate reconstruction using tree-ring information (hint: go back and review some of the principles of dendrochronology)?
2. Based on the four analytical tests used to *evaluate* the climate reconstruction, comment on the strength of the observed relationship between seasonal precipitation and tree-ring indices.
3. Comment on the strengths and weaknesses of using correlation and regression procedures for climate reconstruction using tree-ring information.

Table 2: Calculation of the Pearson correlation coefficient (r)

x_t = ring-width index

y_t = precipitation (mm)

$$r = \frac{\sum_{t=1}^n (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^n (x_t - \bar{x})^2 \sum_{t=1}^n (y_t - \bar{y})^2}} = \frac{\sum_{t=1}^n col(8)}{\sqrt{\sum_{t=1}^n col(4) \sum_{t=1}^n col(7)}}$$

(1)	(2)	(3)	(4)=(3)*(3)	(5)	(6)	(7)=(6)*(6)	(8)=(3)*(6)
Year	x_t	$(x_t - \bar{x})$	$(x_t - \bar{x})^2$	y_t	$(y_t - \bar{y})$	$(y_t - \bar{y})^2$	$(x_t - \bar{x})(y_t - \bar{y})$
1941	1.76			716.7			
1942	1.37			513.2			
1943	0.98			468.8			
1944	1.00			374.0			
1945	0.92			342.0			
1946	0.78			321.6			
1947	0.73			211.7			
1948	0.68			303.7			
1949	0.91			345.8			
1950	0.88			360.6			
Sum (Σ) =							
Mean =							

$$r = \frac{\text{Sum (8)}}{\sqrt{\text{Sum (4)} \times \text{Sum (7)}}}$$

$$= \frac{\text{Sum (8)}}{\sqrt{\text{Sum (4)} \times \text{Sum (7)}}}$$

$$r = \frac{\text{Sum (8)}}{\sqrt{\text{Sum (4)} \times \text{Sum (7)}}}$$

Table 3: Significant values of r

df Error	Probability, p		
	0.05	0.01	0.001
1	0.997	1.000	1.000
2	0.950	0.990	0.999
3	0.878	0.959	0.991
4	0.811	0.917	0.974
5	0.755	0.875	0.951
6	0.707	0.834	0.925
7	0.666	0.798	0.898
8	0.632	0.765	0.872
9	0.602	0.735	0.847
10	0.576	0.708	0.823
11	0.553	0.684	0.801
12	0.532	0.661	0.780
13	0.514	0.641	0.760
14	0.497	0.623	0.742
15	0.482	0.606	0.725
16	0.468	0.590	0.708
17	0.456	0.575	0.693
18	0.444	0.561	0.679
19	0.433	0.549	0.665
20	0.423	0.457	0.652
25	0.381	0.487	0.597
30	0.349	0.449	0.554
35	0.325	0.418	0.519
40	0.304	0.393	0.490
45	0.288	0.372	0.465
50	0.273	0.354	0.443
60	0.250	0.325	0.408
70	0.232	0.302	0.380
80	0.217	0.283	0.357
90	0.205	0.267	0.338
100	0.195	0.254	0.321

Table 4: Reconstruction of precipitation to evaluate the regression model

Regression model: $\hat{y}_t = \beta_0 + \beta_1 x_t$

$$\hat{y}_t = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} x_t$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year	x_t	y_t	\hat{y}_t	Residual ($y_t - \hat{y}_t$)	$(y_t - \hat{y}_t)^2$	$(\hat{y}_t - \bar{y})$	$(\hat{y}_t - \bar{y})^2$
1941	1.76	716.7					
1942	1.37	513.2					
1943	0.98	468.8					
1944	1.00	374.0					
1945	0.92	342.0					
1946	0.78	321.6					
1947	0.73	211.7					
1948	0.68	303.7					
1949	0.91	345.8					
1950	0.88	360.6					
Sum of Squares:				SSE =		SSR =	

Table 5: Verification of precipitation reconstruction using independent precipitation data from 1931-1940.

Regression model: $\hat{y}_t = \beta_0 + \beta_1 x_t$

$$\hat{y}_t = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} x_t$$

$$r = \frac{\sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}})^2 \sum_{t=1}^n (y_t - \bar{y})^2}} = \frac{\sum_{t=1}^n \text{col}(9)}{\sqrt{\sum_{t=1}^n \text{col}(5) \sum_{t=1}^n \text{col}(8)}}$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Year	x_t	\hat{y}_t	$(\hat{y}_t - \bar{\hat{y}})$	$(\hat{y}_t - \bar{\hat{y}})^2$	y_t	$(y_t - \bar{y})$	$(y_t - \bar{y})^2$	$(\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})$
1931	0.54				194.4			
1932	1.24				425.3			
1933	0.81				286.1			
1934	0.64				243.1			
1935	1.34				444.5			
1936	1.44				440.9			
1937	1.27				551.7			
1938	1.22				663.6			
1939	0.62				256.0			
1940	1.38				559.7			
Sum (Σ) =								
Mean =								

$$r = \frac{\underline{\hspace{2cm}}}{\sqrt{(\underline{\hspace{2cm}})(\underline{\hspace{2cm}})}}$$

$$= \underline{\hspace{2cm}}$$

$$r = \underline{\hspace{2cm}}$$

Figure 2: Scatterplot

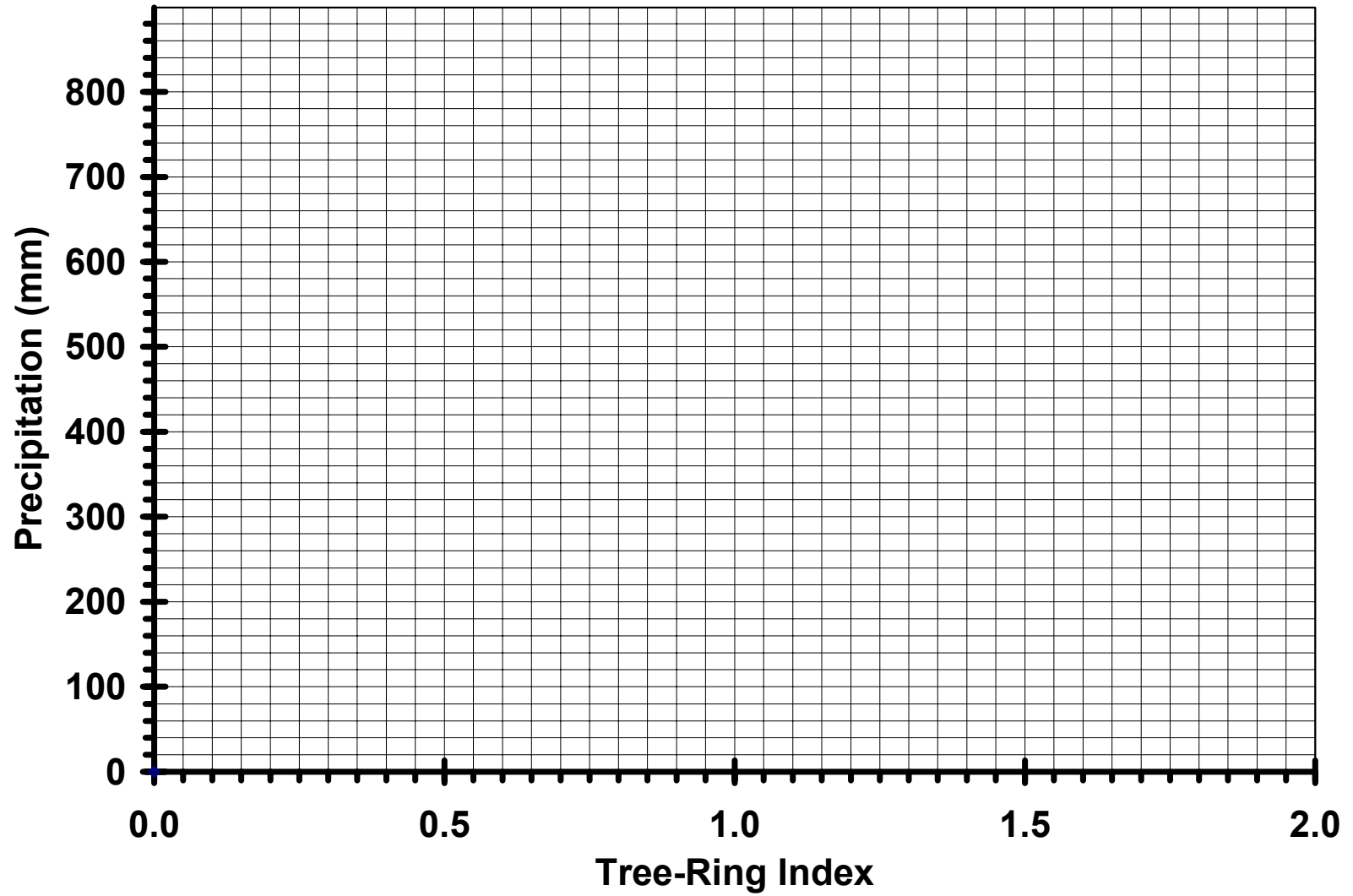


Figure 3: Time series plot

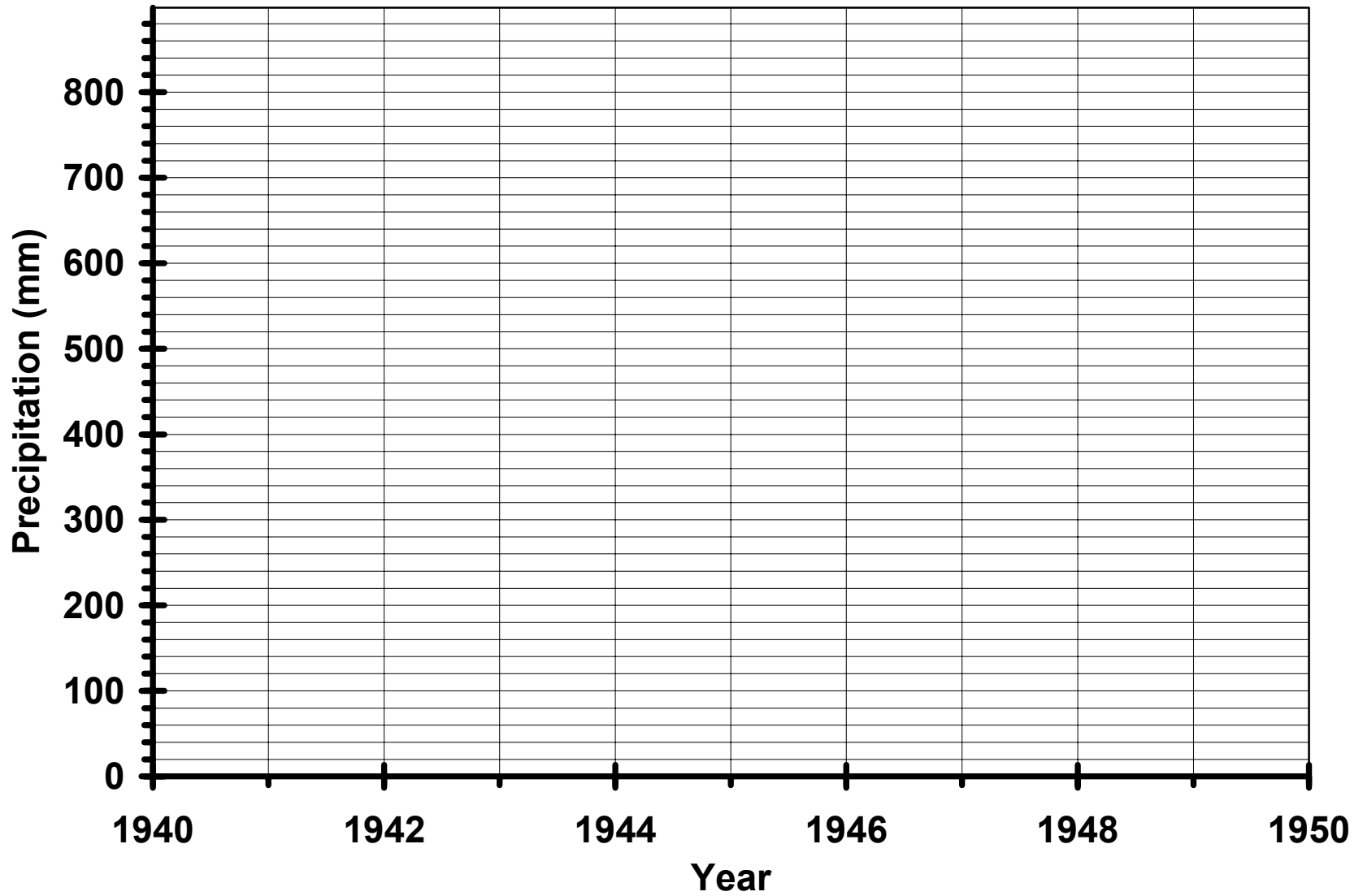


Figure 4: Time series plot for verification period