Appendix 2. Description of the regression tree methodology, and comparison of linear regression and regression tree results.

To illustrate differences in the predictive power of the linear regressor and the regression tree, we construct regressors for precipitation of the driest quarter using, for simplicity, only the fraction of high crowned taxa (pHYP3) as a covariate. We deliberately chose the precipitation of the driest quarter, because for this variable the relative performance of the linear regressor in comparison with the regression tree is particularly poor (see Table 2 in the article). We chose the fraction of the high crowned taxa as a covariate, because it is the most important factor in predicting the precipitation of driest quarter, as can be seen from the trees 6 and 8 in Appendix 1. The linear regressor is given by the equation

$$\text{Precipitation of the driest quarter (mm)} = 123.1 - 198.7 * \text{pHYP3}$$

Notice that the linear regressor predicts negative precipitation for locations where the fraction of high crowned taxa is at least 0.62, an obviously false value. The corresponding regression tree, where the height of the tree has been restricted to two for simplicity, is shown in Figure 2. The R2 value of this regression tree is 0.360 while the R2 value of the linear regressor is 0.266, that is, the regression tree outperforms the linear regressor. One can read from Figure 2 that the regression tree predicts rainfall of 131.1 mm if pHYP3 is less than 0.041, 295.6 mm if pHYP3 is between 0.041 and 0.118, 50.98 mm if pHYP3 is between 0.118 and 0.380, and 14.98 mm if pHYP3 exceeds 0.380. Note that unlike the linear regressor, the regression tree splits the covariate space into partitions, one partition for each leafnode, and it outputs a constant predictor

within each of these regions (in this example the pHYP3 axis is split into four partitions that correspond to the precipitation values of 131.1 mm, 295.6 mm, 50.98 mm, and 14.98 mm, respectively). The predictions of the linear regressor and regression tree that use the fraction of the high crowned taxa as a covariate are plotted in Figure 1.

The regression tree has an advantage over a linear regression if the response to the covariates is non-linear as is the case, for example, in Figure 1. In this example the prediction of the linear regression decreases linearly with pHYP3 until it incorrectly predicts negative values for the precipitation. The prediction of the regression tree, on the other hand, is always consistent and within range of the values for the given covariates in the grid cell. This is simply because of how the regression tree is constructed: the predicted precipitation in a leaf node is the average precipitation of the grid cells in the training data that are associated with the leaf node; the regression tree would therefore, for example, never predict negative precipitation. Another advantage of the regression trees is that they can identify context-dependent associations among multiple correlated covariate variables; regression trees do not assume that the covariates are independent, as is the case with linear regression. The same covariate variables can occur several times at different levels of the tree. The regression tree is easy to interpret as simple rules (for example, "if pHYP3 exceeds 0.38 the precipitation of the driest quarter is 15 mm").

Figure 1. Precipitation of the driest quarter (mm) as a function of the fraction of high crowned taxa (pHYP3) is shown using grey dots in the background as well as the box plots (the width of the box plots corresponds to the number of grid cells). The solid blue line shows the prediction by regression tree and the dotted red line show the prediction by the linear regressor. Notice the poor performance of linear regression in comparison with the regression tree.

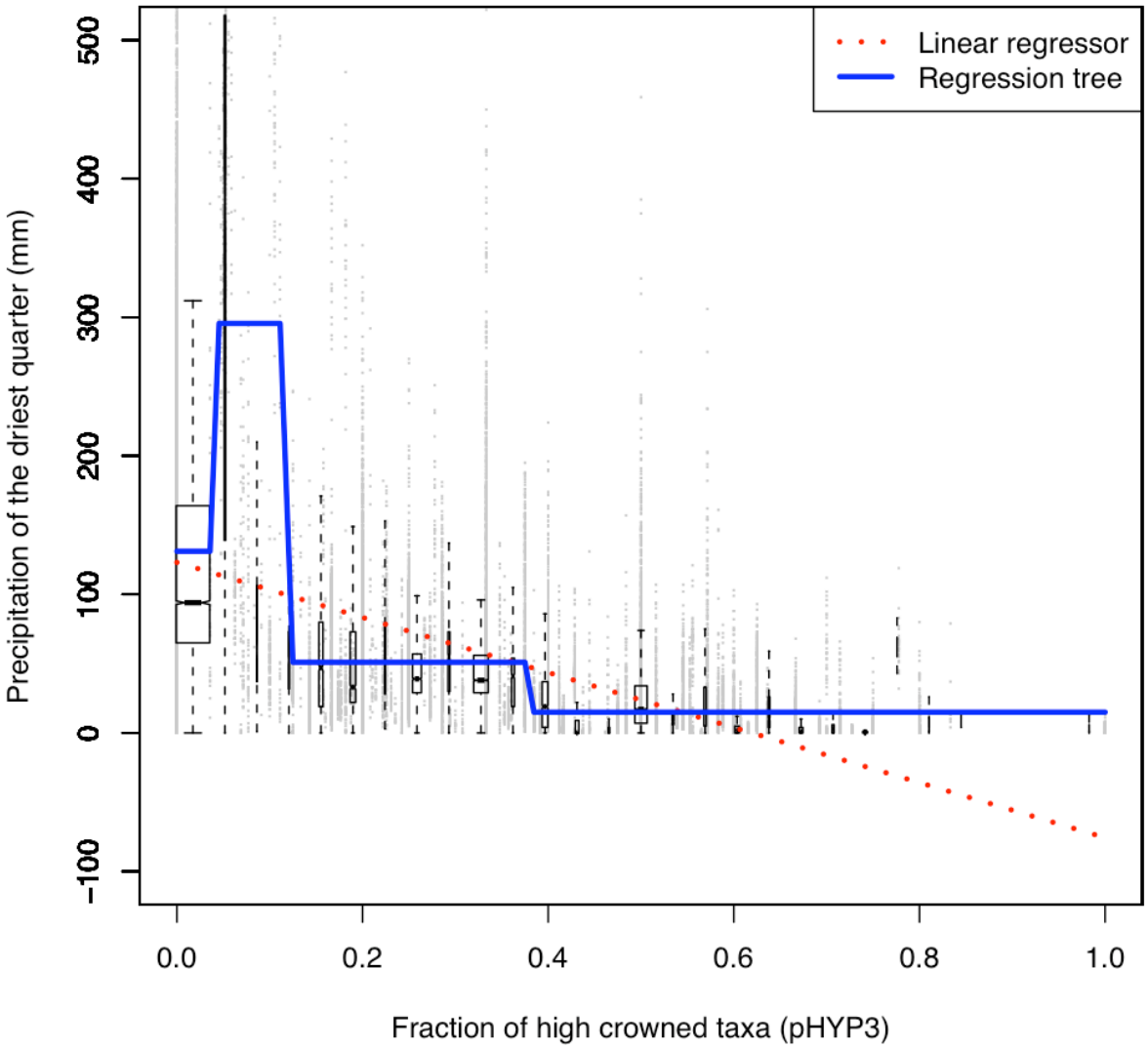

Figure 2: Regression tree that gives the precipitation of the driest quarter (mm) as a function of the fraction of high crowned taxa (pHYP3). See the text for discussion.