Master's Thesis

Analyzing the Effects of Climate Change on Vegetation Dynamics using Functional Data Analysis

Department of Statistics
Ludwig Maximilian University Munich

Theresa Meier

Supervisors: Prof. Dr. Thomas Nagler, Jana Gauss

Project Partners: Prof. Dr. Anja Rammig, Lucia Layritz

September 24, 2024

Abstract

As one of the largest carbon sinks on the planet, the boreal forest is particularly vulnerable to disturbances such as wildfires and storms. The frequency, intensity and magnitude of these disturbances – the so-called disturbance regime – is altering, probably as a result of changes in climate conditions, with potentially drastic effects on the boreal ecosystem. This thesis aims to identify patterns and key drivers of vegetation composition during post-disturbance recovery at different locations scattered across the boreal circumpolar belt in North America, Asia and Europe under varying climatic conditions. Using the dynamic vegetation model LPJ-GUESS, simulations under four different climate scenarios – pre-industrial, SSP1-RCP2.6, SSP3-RCP7.0 and SSP5-RCP8.5 – covering time series of aboveground carbon in terms of five different tree species groups serve as the basis for statistical analysis. Post-disturbance recovery trajectories are examined by combining building blocks of functional data analysis, including multivariate functional principal component analysis and multivariate functional linear regression. Clustering of principal component scores indicates that vegetation recovery follows three distinct dominance patterns: pioneering broadleaf, temperate broadleaf and needleleafed trees. In addition, results from the modelling approach indicate that more extreme climate scenarios favour the establishment of deciduous species, while soil properties such as clay, sand and silt content significantly influence the ability of species to establish during recovery. These findings are critical for understanding the long-term consequences of climate change on the boreal forest, with implications for global carbon storage and water and energy cycles.

Table of Contents

1 Introduction

> "Holding the increase in the global average temperature to well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change"
>
> *Paris Agreement, Article 1(a), United Nations (2015)*

The Earth's climate has been warming since the dramatic increase in carbon dioxide emissions associated with industrialisation in the 19th century. This change in climatic conditions is causing, for example, shifts in vegetation composition, particularly in boreal and temperate forests, where rising temperatures favour species that are more tolerant of heat and drought (N. G. McDowell et al., 2020; Mack et al., 2021). These changes in the state and functioning of ecosystems, and many other consequences for humans and nature, have prompted policymakers to take action: to combat climate change together and to intensify efforts towards a sustainable low-carbon future, on 12 December 2015, 196 parties signed the so-called Paris agreement. The central objective in Article 1(a) of the Paris agreement is to keep the global average temperature increase well below 2°C above pre-industrial levels, i.e., the climate around 1850, and to pursue efforts to limit the increase to 1.5°C. To assess whether or not this goal is still achievable, climate scenarios are essential tools for understanding possible future climate conditions based on different levels of greenhouse gas emissions and policies. They highlight the long-term effects of different emission trajectories, enabling policymakers to choose actions that can limit global warming and to develop adaptation strategies. These actions and adaptations are necessary as the impact of global climate change on people and nature are already being experienced. Forests in particular are suffering from increased tree mortality, pest outbreaks and more frequent and severe wildfires, all driven by rising temperatures and prolonged droughts (Allen et al., 2010).

The boreal forest is one of the three major forest biomes on the planet and serves as a large reserve of terrestrial carbon (Malhi, Baldocchi, and Jarvis, 1999). According to Gauthier et al. (2015), the vegetation in boreal forests, which is at present dominated by conifers, is facing drastic changes due to climate change: As the climate warms, the composition of tree species may change, with some species migrating northwards and others unable to survive the changing conditions (Baltzer et al., 2021; Mack et al., 2021). These vegetation shifts not only reflect a fundamental reorganization of forest ecosystems but also have broader implications for carbon storage, hydrological cycles, and energy exchange, which could even amplify the effects of climate change (Mack et al., 2021; Bonan, 2008). For example, carbon storage can be reduced by planting new tree species adapted to milder climate zones, which release carbon back into the atmosphere more quickly than previously established tree species (Mack et al., 2021). This process is further accelerated by forest disturbances such as windstorms and wildfires, and changes in the frequency or intensity of such disturbances, i.e., the disturbance regimes, can have a major impact on the vegetation of the biome (Pugh et al., 2019; Allen et al., 2010). As disturbance events are key drivers of ecosystem processes, tree species have evolved over time to withstand and recover from these periodic

disturbances, thus maintaining a dynamic balance within the ecosystem (Pfadenhauer and Klötzli, 2020; Ilisson and H. Y. Chen, 2009). However, recent studies suggest that as climate change alters these disturbance regimes, this balance is increasingly at risk, with certain coniferous species failing to regenerate after disturbances (Ilisson and H. Y. Chen, 2009). Instead, these regions are experiencing transitions to different vegetation types, such as deciduous forests, or even non-forested landscapes (Baltzer et al., 2021; Mack et al., 2021). Investigating the effects of forest disturbance on vegetation recovery is therefore crucial for better understanding the consequences of changing climatic conditions and for assessing the associated risks and necessary policy measures.

Statistical approaches offer powerful tools for quantifying changes in vegetation, understanding recovery patterns, and predicting future ecosystem responses. Both simulated and observed vegetation data are often high dimensional, and as a consequence, methods such as principal component analysis are widely used to reduce dimensionality. For instance, Radovan Hladky and Stych (2020) and Smith-Tripp et al. (2024) decrease the complexity of satellite data as a precursor to further processing the data. To explore how vegetation might change under future climatic conditions, several statistical methods are used in practice. Ma et al. (2022) use standard regression techniques, while Ovenden et al. (2021) make use of time series models. Recently, also machine learning techniques are applied to vegetation data, like for instance random forests by Coop (2023) and Senf and Seidl (2022), and support vector machines by Zhu et al. (2020) to e.g., infer impacts of climate change on post-disturbance forest recovery in China. An approach to using functional data tools for vegetation recovery is proposed by Serra-Burriel, Delicado, and Cucchietti (2021), which examines the effect of wildfires on vegetation recovery in California using observed remote sensing data.

The aim of this thesis is to investigate the effects of different climatic conditions on the recovery of the boreal forest after disturbances. The analysis is based on time series data from the dynamic vegetation model LPJ-GUESS (Smith, Prentice, and Martin T Sykes, 2001; Smith, Wårlind, et al., 2014), which can simulate vegetation after disturbances on a global scale for different climate scenarios. In this analysis, a pre-industrial scenario is chosen as the baseline and three warming scenarios, namely SSP1-RCP2.6, SSP3-RCP7.0 and SSP5-RCP8.5, depict increasing radiative forcing. The simulation model provides a number of different variables for the description of the forest vegetation. Besides aboveground carbon, which serves as a measure of the distribution of tree species, variables such as soil properties and temperature and precipitation curves are available. From this data, recovery trajectories are derived by focusing on forest disturbances in a given time interval and considering the proportions of different tree species in the first hundred years of recovery. Here, the focus lies on the predominant vegetation type and the mutual displacement of tree species during the recovery period. In this thesis, the perspective on the data is changed: instead of a time series, the recovery trajectories are considered as functions. Therefore, to adequately address the functional format of the data, this thesis establishes a framework combining different techniques used in functional data analysis, such as multivariate functional principal component analysis as well as multivariate functional additive regression techniques, to identify patterns and key drivers of vegetation composition during recovery from disturbance. This approach is novel in terms of both methodology and application, as there is a lack of existing

literature on the use of multivariate functional data analysis for climate-induced recovery dynamics.

In order to find these patterns and drivers, this thesis will first introduce the necessary theoretical background to functional data analysis methods in Section 2, before delving deeper into the data, i.e., the time series data generated by LPJ-GUESS, provided in Section 3, including details of the simulation process, the climate scenarios considered in this project, and the variables available for statistical analysis. In Section 4, after a brief descriptive analysis to get a first impression of the data set, the goal of finding patterns within recovery trajectories is addressed by clustering scores derived by performing functional principal component analyses. A novel modelling approach based on principal component scores and multivariate functional regression is then introduced and evaluated in Section 5 to derive drivers of vegetation composition after disturbance. The final results of this analysis are summarised in Section 6, before concluding with a discussion of limitations and future opportunities in Section 7. Appendix A includes detailed descriptions of spatial dependencies and clustered trajectories, before this thesis closes with technical details on programming in Appendix B.

2 Theoretical Background

The overall goal of this research is to explore the composition of vegetation after disturbances in the boreal zone. Given that the variable of interest - the shares of aboveground carbon over years after disturbance for different types of vegetation - can be considered as functional, an analysis based on functional data is appropriate for the data at hand.

In this chapter, the necessary theory for the methods used in this analysis is presented. This includes background information on functional data analysis and univariate and multivariate functional principal component analysis, as well as an overview of k-means clustering.

If not stated otherwise, the following theory is based on J. O. Ramsay and B. W. Silverman (2005), J. O. Ramsay, Hooker, and Graves (2009) and Young (2014).

2.1 Functional Data Analysis

When measurements are taken repeatedly over a recurring time period, like daily temperature measurements over several years, or space, like different weather stations spread over a region, it can be useful for later analyses to change the perspective on the available data. **Functional Data Analysis (FDA)** is a statistical framework for analysing data that can be represented by functions, often over time or space. The fundamental concept behind FDA is to treat each function as a singular data point, rather than focusing solely on the individual data points that define these functions. This approach is particularly useful for dealing with complex data structures that are not adequately addressed by traditional multivariate methods. Moreover, FDA seamlessly integrates with many established statistical techniques, often requiring only minor adjustments for adaptation. This adaptability ensures that analytical tools and methods commonly used in non-functional statistics remain applicable within the FDA framework.

In order to get a first impression how multiple data points can be expressed as a smooth function, Figure 1 shows how data points (black) can be approximated by splines of order 6 (blue). The spline serves as one data point in following analyses. The derivation of this functional fit is usually the first step in FDA and is explained in more detail in the section below.

2.1.1 Basis Representation

Building functions in the framework of FDA is based on the concept of *basis functions*. Basis functions are a set of predefined, simple functions that can be combined in various ways to approximate more complex functions. Thus, in order to approximate multiple data points by a functional fit, two steps are necessary:

1. Define basis functions ϕ_k which serve as functional building blocks.

2. Define target functions as linear combinations of the basis functions and basis coefficients c_k.

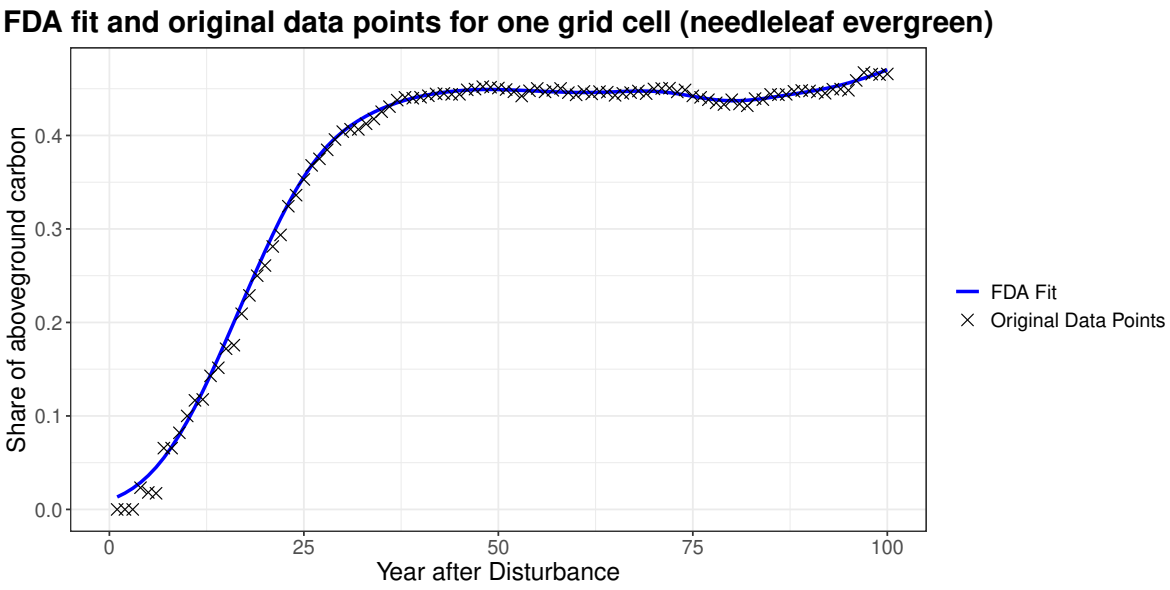


Figure 1: Smooth functional fit and original data points for one example grid cell.

Mathematically speaking, let $x : \mathcal{T} \to \mathbb{R}$ be a (possibly complex) function defined on the set $\mathcal{T}$. The *basis function expansion* of function $x(t)$ is given by

$$x(t) = \sum_{k=1}^{\infty} c_k \phi_k(t) \approx \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}^T \boldsymbol{\phi}(t),$$

where $\boldsymbol{\phi}(t) = (\phi_1(t), ..., \phi_K(t))^T$ represents basis functions with corresponding basis coefficients $\mathbf{c} = (c_1, ..., c_K)^T$. Using a finite basis representation for functional data comes with several advantages. Although it is a way of simplifying complex functional data, basis functions can capture a wide variety of functional shapes and patterns by adjusting the basis coefficients. In addition, working with a finite number of basis functions instead of potentially infinite functional data makes computations more efficient and may serve as an initial smoothing by keeping the number K of basis functions ϕ_k small relative to the amount of data. Moreover, instead of saving all data points that underlie the approximated function, only the coefficients c_k of the basis expansions are stored. This results in an initial dimension reduction.

The choice of the *basis*, i.e., a set of basis functions ϕ_k which may or may not be orthogonal to each other, is crucial for an appropriate fit to the underlying data. There are two types of basis functions ϕ_k which are most commonly used in practice:

- **Fourier basis:** This type of basis is usually used for periodic, or nearly periodic, data, e.g., annual temperature data. The orthogonal basis functions are defined as follows:

  $$\phi_0(t) = 1, \phi_{2r-1}(t) = \sin(r\omega t), \phi_{2r}(t) = \cos(r\omega t),$$

  where ω determines the period $2\pi/\omega$, which is equal to the length of the interval $\mathcal{T}$.

- **B-Spline basis:** B-Splines are used for non-periodic locally smooth data and are more flexible, but therefore more complex than finite Fourier series. They are functions that are defined piecewise by polynomial functions of *degree* k, and they are smooth at the *knots*, i.e., the points where the polynomials join. For knots t_0, t_1, ..., t_{n-k+1} that divide the interval into $n+1$ sub-intervals, they are defined as:

$$\phi_{i,0}(t) = \begin{cases} 1 & \text{if } t_i \leq t < t_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } k = 0,$$

$$\phi_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i}\phi_{i,k-1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}}\phi_{i+1,k-1}(t), \quad \text{for } k \geq 1.$$

  Note that B-spline basis functions are not orthogonal to each other.

Other bases like polynomial or Wavelet bases are possible as well, but less frequently used. For details, see Chapter 3.2 in J. O. Ramsay and B. W. Silverman (2005).

2.1.2 Regularization

The basis representation obtained in the previous section has two major drawbacks: first, it does not adequately account for measurement errors in the data and is therefore prone to overfitting. Secondly, the approach presented tends to produce wiggly fits, which is contrary to the aim of obtaining a smooth fit to the data. Consequently, **Regularization** is applied.

The most intuitive approach is based on the standard regression analysis model

$$y_i = x(t_i) + \varepsilon_i = \mathbf{c}^T\boldsymbol{\phi}(t_i) + \varepsilon_i = \boldsymbol{\phi}^T(t_i)\mathbf{c} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Here, y_i are the observed values at time points t_i for $i = 1, ..., n$, $x(t_i)$ is the true function value at t_i, $\boldsymbol{\phi}(t_i)$ are the basis functions evaluated at t_i, $\mathbf{c}$ is the vector of coefficients, and ε_i represents the measurement error. In the example curve in Figure 1, the observed values y_i, $i = 1, ..., 100$ correspond to the 100 original data points, while the smooth fit in blue is represented by the function $x(t_i)$ at time points $t_1 = 1, ..., t_{100} = 100$. Fitting an appropriate function $x(t)$ is then defined as the minimization of the sum of squared errors of the residuals:

$$SSE(\mathbf{c}) = \sum_{i=1}^{n}[y_i - \sum_{k=1}^{K} c_k\phi_k(t_i)]^2 = \sum_{i=1}^{n}[y_i - \boldsymbol{\phi}(t_i)^T\mathbf{c}]^2$$

Equivalently to standard regression, the least square estimate of the coefficient vector $\mathbf{c}$ can be derived as:

$$\hat{\mathbf{c}} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y},$$

where $\mathbf{y}$ is the vector of n values to be fitted and Φ is a $n \times K$ matrix containing the basis function values $\phi_k(t_i)$. Approaches where spline curves are fitted by regression analysis are called *regression splines*. However, they suffer from one major disadvantage: their derivative estimates tend to be unstable at the boundaries of the intervals because the spline functions have fewer neighboring points to anchor the fit at the edges of the data range. This instability can lead to large oscillations or artifacts near the boundaries as indicated in Chapter 5.1

of B. Silverman and J. Ramsay (2002), which do not represent the true underlying function.

To address these issues, a regularization term is added to the regression criterion. The idea is to use a large number of basis functions, possibly even $n > K$, and to ensure smoothness by penalizing the function complexity. The penalized sum of squared errors (PENSSE) criterion is given by:

$$\begin{aligned} PENSSE(\boldsymbol{c}) &= \sum_{i=1}^{n} [y_i - x(t_i)]^2 + \lambda \int_{\mathcal{T}} [Lx(t)]^2 \, dt \\ &= \sum_{i=1}^{n} [y_i - \boldsymbol{\phi}(t_i)^T \boldsymbol{c}]^2 + \lambda \boldsymbol{c}^T \left[\int_{\mathcal{T}} [L\boldsymbol{\phi}(t)]^2 \, dt \right] \boldsymbol{c}, \end{aligned}$$

where λ is a smoothing parameter that controls the trade-off between fit and smoothness, and L is a *differential operator*. This operator may capture the *roughness* of a curve expressed by e.g., the m-th derivative $L = D^m$ or by a (weighted) combination of derivatives. Note that when choosing to penalize a derivative of the chosen basis $\boldsymbol{\phi}s$, the number of available derivatives must be at least the order being penalized. In this penalized version, the least squares estimate of coefficient vector $\mathbf{c}$ can be derived as:

$$\hat{\mathbf{c}} = \left(\Phi^T \Phi + \lambda \int_{\mathcal{T}} \boldsymbol{\phi}(t) \boldsymbol{\phi}(t)^T \, dt \right)^{-1} \Phi^T \mathbf{y},$$

The smoothing parameter λ can be determined by using cross validation or generalized cross validation. Details on these methods are described in Chapter 7.5 of J. O. Ramsay and B. W. Silverman (2005).

In this analysis, the functions are mappings from $\mathcal{T} = [0, 100]$ to $[0, 1]$, i.e., the codomain comprises proportions. This means in particular, that the linear combination of basis functions should result in values between 0 and 1. In order to address this constraint, the smoothing problem described in Equation 1 is transformed (see Chapter 5.4.1 in B. Silverman and J. Ramsay (2002)):

$$y_i = \exp\left(\boldsymbol{\phi}(t_j)^T \mathbf{c}\right) + \varepsilon_i.$$

This ensures positivity of the resulting fit. However, the fit is not forced to maintain values below one.

2.1.3 Summary Statistics for Functional Data

Since the data points are now functional objects, there is a need for adaptation of common definitions of mean and covariance to the functional context. Assume that a sample of k curves $x_1(t), x_2(t), ..., x_k(t)$ is given, where each curve corresponds to a realization of a random function $x(t)$. Under the assumption that the observations $x_1(t), ..., x_k(t)$ are independent and have the same distribution as x, the *sample mean function* is defined as:

$$\hat{\mu}(t) = \frac{1}{k} \sum_{i=1}^{k} x_i(t). \tag{2}$$

This means in particular, that the mean is the average of the functions pointwise across replications. The *sample variance function* is defined in the same pointwise manner:

$$\hat{\sigma}^2(t) = \frac{1}{k-1}\sum_{i=1}^{k}[x_i(t) - \hat{\mu}(t)]^2.$$

The *sample standard deviation* is the square root of the sample variance function. For the *sample covariance function* holds:

$$\hat{v}(t,s) = \frac{1}{k}\sum_{i=1}^{k}(x_i(t) - \hat{\mu}(t))(x_i(s) - \hat{\mu}(s)), \tag{3}$$

and the associated *sample correlation function* is given by

$$\hat{r}(t,s) = \frac{\hat{v}(t,s)}{\sqrt{\hat{\sigma}^2(t)\hat{\sigma}^2(s)}}.$$

2.2 Functional Principal Components Analysis

The data smoothing procedure introduced in Section 2.1.2 is based on variation within observations. This section focuses on variation between different data points, i.e., different functions in the context of FDA. **Principal Components Analysis (PCA)** is a fundamental technique in multivariate and functional statistics used for dimensionality reduction, data compression and feature extraction. It transforms the original variables of a data set into a new set of uncorrelated variables, known as *principal components (PCs)*, which capture the maximum variance in the data. By projecting the data onto these PCs, the dimensionality of the data set is reduced while retaining most of its variability. This is particularly useful for visualization, noise reduction, and simplifying further analysis, especially in the framework of infinite dimensional functional data.

First, the general concept of PCA in the multivariate setting is derived in detail, before it is transferred to the functional and multivariate functional context.

2.2.1 Multivariate Principal Components Analysis

To give a first impression of how PCA works in general, this section considers multivariate data $\mathbf{x}_1, ..., \mathbf{x}_n$, with $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$, $i = 1, ..., n$. Note that in Section 2.1 the index n represented the number of time points used to fit the function, whereas here n denotes the number of individual curves considered. Similar to Section 2.1.1, PCA is based on linear combinations of variable values. The goal is to weight the observed values x_{ij} such that the dominant types of variation in the data are highlighted the most. The derivation of PCs comprises the following step-wise procedure:

1. To appropriately maximize the sample variance, demean or standardize the variables.

2. **First PC:** Find the weight vector $\boldsymbol{\xi}_1 = (\xi_{11}, ..., \xi_{p1})^T$ for which the linear combination

$$y_{i1} := \sum_{j=1}^{p} \xi_{j1} x_{ij}$$

has the largest possible mean square $\frac{1}{n}\sum_{i=1}^{n} y_{i1}^2$ subject to the constraint

$$\sum_{j=1}^{p} \xi_{j1}^2 = \|\boldsymbol{\xi}_1\|^2 = 1.$$

This constraint ensures that the problem is well defined.

3. **Second and subsequent PCs:** For the m-th PC, derive weight vector $\boldsymbol{\xi}_m$ and values $y_{im} = \sum_{j=1}^{p} \xi_{jm} x_{ij}$ with maximum mean square error, subject to m constraints:

   (a) $\sum_{j=1}^{p} \xi_{jm}^2 = \|\boldsymbol{\xi}_m\|^2 = 1$

   (b) $\sum_{j=1}^{p} \xi_{jk}\xi_{jm} = 0$ for $k < m$.

   The second set of $m-1$ constraints enforces orthogonality of the weights derived in previous steps.

The orthogonality constraint ensures that each PC finds variation in directions that are unexplored at the time of defining the next PC. Note that the resulting weight vectors are not uniquely defined, since the signs in any vector $\boldsymbol{\xi}_m$ might be changed without affecting the amount of variance that this PC accounts for.

An alternative characterization of the previously derived PCs $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m$ is explained in Kauermann, Küchenhoff, and Heumann (2021). It is based on the eigenanalysis of the covariance matrix Σ of the centered data matrix $X \in \mathbb{R}^{n \times p}$, where each row represents a demeaned observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$:

1. Compute the sample covariance matrix $\hat{\Sigma}$ of X by calculating:

$$\hat{\Sigma} = \frac{1}{n-1} X^T X.$$

2. Determine the eigenvectors $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_p$ and corresponding eigenvalues $\lambda_1, ..., \lambda_p$ of the covariance matrix $\hat{\Sigma}$ by solving the eigenvalue problem for $\hat{\Sigma}$:

$$\hat{\Sigma}\,\boldsymbol{\xi}_i = \lambda_i\,\boldsymbol{\xi}_i, \quad i = 1, ..., p. \tag{4}$$

   This can be done computationally efficiently by using a spectral decomposition of $\hat{\Sigma}$.

3. Order the eigenvalues in descending order and arrange the corresponding eigenvectors accordingly.

4. The eigenvectors form the PCs. The first PC $\boldsymbol{\xi}_1$ corresponds to the eigenvector with the largest eigenvalue, the second PC $\boldsymbol{\xi}_2$ to the second largest eigenvalue, and so on.

The eigenvectors, i.e., the PCs, $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_p$, are orthogonal directions that capture the most dominant variations in the data, with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ indicating the amount of variance each PC is accounting for. In fact, the PCs build an orthonormal basis system. For more background information on eigenanalysis, see Chapter 6 of Strang (2023).

The derived PCs can now be used for dimensionality reduction of the initial data set X. Despite the availability of p eigenvectors, in many practical applications only a few PCs capture nearly all the variation in the data. As a result, considering $h \ll p$ PCs for reconstructing X is sufficient. First, the original data is projected onto the eigenvectors to derive the *principal component scores*:

$$\mathbf{s}_i = X\,\boldsymbol{\xi}_i, \quad i = 1, ..., p,$$

or in matrix notation with $U = (\boldsymbol{\xi}_i, ..., \boldsymbol{\xi}_p)$:

$$S = X\,U.$$

The initial data set X is then reconstructed by

$$\hat{X} = S_h\,U_h^T,$$

where the subscript h corresponds to the first h columns of the respective matrix. If the original data set X was demeaned or standardized, this transformation needs to be reversed for the reconstructed data $\hat{X}$. Note that instead of storing a $n \times p$ matrix, the PCA results in storing two matrices with dimensions $n \times h$ for the scores and $p \times h$ for the PCs. Depending on the sizes of n, p and h, this dimensionality reduction may lead to a substantial decrease of necessary memory and run time when further processing the data.

The choice of an appropriate number of PCs h is crucial to ensure that the reduced dimensionality representation captures essential features of the data while discarding noise and redundant information. However, there is no standardised technique for its derivation. One method is based on *scree plots*, which show the proportion of total variance explained by each PC for a chosen number of PCs. The total variance explained for each PC can be derived by the eigenvalues λ_i, $i = 1, ..., p$:

$$Var_{explained_i} = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k}.$$

Then, h could be chosen such that the cumulative variance explained exceeds a certain threshold, say 90%.

2.2.2 Functional Principal Components Analysis

Performing a **Functional Principal Components Analysis (FPCA)** implies that the data are in a functional form, i.e., $x_1(t), ..., x_n(t)$, where $x_i : \mathcal{T} \to \mathbb{R}$ for all $i = 1, ..., n$. Similar to multivariate PCA, the goal of deriving PCs is to analyse the variation in the data and to find a lower dimensional representation of the infinitely dimensional functional data in an optimal way. There are two major differences from the PCA derived in the previous section:

first, the PC vectors are now PC functions, and second, most of the theory behind PCA is easily transferred to the functional framework by replacing sums with integrals. Again, there are several perspectives on how to motivate FPCA. First, consider the derivation based on maximizing the mean squares of linear combinations of centered data points and PCs. The derivation of the PCs is straightforward:

1. **First PC:** Find the weight *function* ξ_1 for which the integral

$$\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2 := \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathcal{T}} \xi_1(s)x_i(s)\,ds\right)^2$$

   is maximized subject to the constraint

$$\|\xi_1\|^2 = \int_{\mathcal{T}} \xi_1(s)^2\,ds = 1.$$

2. **Second and subsequent PCs:** For the m-th PC, derive weight function $\xi_m(t)$ by maximizing $\frac{1}{n}\sum_{i=1}^{n} f_{im}^2$ subject to m constraints:
   (a) $\|\xi_m\|^2 = \int_{\mathcal{T}} \xi_m(s)^2\,ds = 1$
   (b) $\int_{\mathcal{T}} \xi_k(s)\xi_m(s)\,ds = 0$ for $k < m$.

Note again that the orthogonal weight functions $\xi_1(t), ..., \xi_p(t)$ are only defined up to a change of sign. Rather than components, in FDA the weight functions are referred to as *harmonics*.

Similar to the multivariate case discussed in Section 2.2.1, FPCA can also be motivated in terms of eigenanalysis, here considering the sample covariance function instead of the sample covariance matrix. The eigenproblem in Equation 4 leads in its functional form to the eigenequation

$$\int_{\mathcal{T}} \hat{v}(s,t)\,\xi_i(s)\,ds = \lambda_i\,\xi_i(t), \quad t \in \mathcal{T},\ i = 1, 2, ..., \tag{5}$$

with the covariance function $\hat{v}$ as defined in Equation 3. The PC scores are again the projection of the centered functional data onto the eigenfunctions:

$$\boldsymbol{s}_{ik} = \int_{\mathcal{T}} x_i(s)\xi_k(s)\,ds, \tag{6}$$

where the index i represents the i-th function and the index k the k-th harmonic.

In multivariate analysis, the number of available PCs is equal to the number of variables. The counterpart in the functional setting is the number of function values, which is infinity. That is, in the case of linearly independent functions $x_i(t)$, the covariance operator V will have rank $n-1$ which results in $n-1$ non-zero eigenvalues.

The concepts introduced for the multivariate PCA apply also in the functional setting: the harmonics are ordered according to the associated eigenvalues which indicate the amount of

captured variance in the data. Again, the number of PCs h may be determined using a scree plot. Equivalently to the multivariate setting, in the functional framework, the leading h harmonics define an orthonormal basis system that can be used to approximate the sample functions $x_i(t)$. In fact, this basis is the most efficient possible of size h as it minimizes the total error sum of squares with coefficients $\boldsymbol{c}$ chosen accordingly:

$$\sum_{i=1}^{n} \int_{\mathcal{T}} [x_i(t) - \boldsymbol{c}_i^T \xi(t)]^2 \, dt = \sum_{j=h+1}^{n-1} \lambda_j. \tag{7}$$

However, this result does not imply that there are no other orthonormal bases that deliver an equivalently good approximation of the original data curves $x_i(t)$ using h basis functions. Let T be an orthonormal matrix of order h, i.e., $T^T T = T T^T = I$. Then, the orthonormal basis functions defined as

$$\psi_i(t) = T \, \xi_i(t), \quad i = 1, .., h,$$

which is a rigid rotation of ξ_i from a geometrical point of view, are equally good approximates in terms of Equation 7. In practical applications, rotating the harmonics can lead to a simpler and more intuitive interpretation. The most popular rotation technique is the VARIMAX strategy. Details on this method are given in J. O. Ramsay and B. W. Silverman (2005).

In the functional setting, the eigenproblem derived in Equation 5 involves the computation of integrals, which requires computational methods for approximation. All three techniques presented below are based on converting the continuous functional eigenproblem to an approximately equivalent matrix eigenanalysis task:

1. **Discretizing the functions:** Discretize the observed functions $x_i(t)$, $i = 1, ..., n$, to a fine grid of m equally spaced values that span the interval $\mathcal{T}$. This transfers the functional problem to a multivariate one, since this discretizing yields a $n \times m$ data matrix X, where each row corresponds to an observation evaluated at the m grid points. Then, X can be analyzed with common statistical software for multivariate data.

2. **Basis function expansion:** Express the observed curves $x_i(t)$, $i = 1, ..., n$, as linear combinations of known basis functions, e.g., B-splines or Fourier series basis, and basis coefficients, as in Section 2.1.1. Then, the eigendecomposition is applied on the matrix of basis coefficients, which is again a multivariate problem.

3. **Numerical quadrature:** This technique is based on approximating an integral by a weighted sum of function values at discrete points:

$$\int_{\mathcal{T}} x(t) \, dt \approx \sum_{i=1}^{n} w_i \, x(t_i).$$

   Here, n refers to the number of discrete argument values t_i, which themselves are called *quadrature points*, and w_i is referred to as *quadrature weights*. Applying this linear approximation to Equation 5 yields a discrete approximation of the covariance function, which in turn represents a matrix eigenproblem.

More details on these techniques are described in Chapter 6.4 of J. O. Ramsay and B. W. Silverman (2005). The software and packages chosen will determine which method is used in practice. The computational methods used in this analysis are further elaborated in subsequent chapters.

2.2.3 Multivariate Functional Principal Components Analysis

So far, the data considered in the previous section was univariate, i.e., each observation consisted of one single function observed over $\mathcal{T}$. In the case of **Multivariate Functional Principal Components Analysis (MFPCA)**, the data at hand are multivariate curves, i.e., each observation comprises multiple functions. A typical example are multiple related time series per subject, e.g., temperature and precipitation curves for different regions.

One approach to this kind of data could be treating each dimension as a univariate problem and conduct multiple univariate FPCAs. Happ and Greven (2018) show with an example that there may be non-negligible correlation between almost all PC score vectors, thus univariate FPCAs capture joint variation in multiple dimensions only indirectly. Consequently, the interpretation might be more difficult and correlated scores might lead to multicollinearity problems in further analyses.

In order to adequately address the multivariate data structure, i.e., obtain one PC score value for each multi-dimensional observation and mulitvariate PC, MFPCA constructs harmonics that represent a basis of h orthonormal basis functions. In the literature, there are several methods how to set up an MFPCA. The approach of J. O. Ramsay and B. W. Silverman (2005) is based on concatenating the observations of the functions on a fine grid of points or alternatively the coefficients of a suitable basis expansion into a single vector, i.e., one single function. Then, a standard multivariate PCA on these concatenated vectors is carried out, which results in an orthonormal basis when a basis expansion is used. Berrendero, Justel, and Svarc (2011) propose to perform classical PCA for each value of the domain $\mathcal{T}$, i.e., at grid points of $\mathcal{T}$, the data is treated as multivariate and a PCA is applied to find PCs for these multivariate observations. Subsequently, the resulting PCs, which are scalar values at each point, need to be combined to form continuous functional PCs. This is achieved through interpolation, ensuring that the PCs are smooth and continuous functions.

Given that all MFPCAs performed within the scope of this project utilize the R package `MFPCA`, the following algorithm is extracted from the package's developer Happ-Kurz (2020). The approach is based on performing a univariate FPCA on PC scores for every dimension, which themselves are results from univariate FPCAs.

Given independent demeaned samples $x_1(t), ..., x_n(t)$ of a d-dimensional random function $x(t) = (x^{(1)}, ..., x^{(d)})$, which share the same distribution, the MFPCA is performed in four steps. Note that the elements $x^{(j)}$ are defined on domains $\mathcal{T}_j \subset \mathbb{R}^{d_j}$, with potentially different dimensional dimensions $d_j \in \mathbb{N}$.

1. For each dimension $j = 1, ..., d$, derive PC functions $\xi_1^{(j)}, ..., \xi_{h_j}^{(j)}$ and PC scores $\boldsymbol{s}_{i1}^{(j)}, ..., \boldsymbol{s}_{ih_j}^{(j)}$

for each observation $i = 1, ..., n$ performing a univariate FPCA for a suitable chosen number of PCs h_j. Note that the number h_j may vary for each dimension j, and may be determined using scree plots.

2. Combine all PC scores into one matrix $S \in \mathbb{R}^{n \times H}$ with $H = \sum_{j=1}^{d} h_j$ and rows:

$$S_{i,\cdot} = (\boldsymbol{s}_{i1}^{(1)}, ..., \boldsymbol{s}_{ih_1}^{(1)}, ..., \boldsymbol{s}_{i1}^{(d)}, ..., \boldsymbol{s}_{ih_d}^{(d)})$$

3. Estimate the joint covariance matrix $\hat{\Sigma} = \frac{1}{n-1} S^T S$ and derive its eigendecomposition including eigenvectors ψ_g and corresponding eigenvalues λ_g for $g = 1, ..., G$ and some chosen $G \leq H$.

4. Calculate multivariate PC functions $\tilde{\xi}_g$ and scores $\tilde{\boldsymbol{s}}_{ig}$ based on the results from steps 1 and 3:

$$\tilde{\xi}_h^{(j)} = \sum_{l=1}^{h_j} [\psi_h]_l^{(j)} \xi_l^{(j)}, \quad \tilde{\boldsymbol{s}}_{ih} = \sum_{j=1}^{d} \sum_{l=1}^{h_j} [\psi_h]_l^{(j)} \xi_{il}^{(j)} = S_{i,\cdot} \psi_h, \quad h = 1, .., G.$$

In contrast to running univariate FPCAs only, steps 2 and 3 of this algorithm ensure to take covariation between different dimensions into account by using the joint covariance matrix $\hat{\Sigma}$ of all scores for solving the eigenproblem.

As for the univariate case, the original data functions $x_i(t)$ can be approximated by a sum of PC scores and PC functions:

$$x_i(t) \approx \hat{\mu}(t) + \sum_{h=1}^{G} \tilde{\boldsymbol{s}}_{ih} \tilde{\xi}_h,$$

with $\hat{\mu}$ representing the sample mean function defined in Equation 2.

In Section 2.2.2, details on the practical implementation of FPCAs indicated that the core idea is to consider the functional problem as multivariate. This applies to the multivariate functional setting as well, since the proposed algorithm is based on FPCAs. Specifically, this means that the implementation of the MFPCA function in R is based on basis function expansions as derived in Section 2.1.1. In Happ and Greven (2018) some further possibilities are explored how to replace univariate FPCAs with a basis expansion approach in the first step of the algorithm for arbitrary basis functions, and how to impose weights on different dimensions $j = 1, ..., d$ in case of varying domain, range or variation.

2.3 Multivariate Functional Linear Regression

Functional Linear Regression (FLR) extends traditional regression to situations where the predictor or response variables, or both, are functions rather than simple numerical values, making it a powerful tool in fields dealing with complex data structures. An overview over available techniques and recent developments is given by Gertheiss et al. (2023). The

Symbol	Description	Dimension
$\mathbf{y}(t)$	Functional response	$d \times 1$
$\mathbf{x}(s)$	Functional predictor	$p \times 1$
$C_M^{\mathbf{y}}$	Matrix of response coefficients	$M \times n$
$C_L^{\mathbf{x}}$	Matrix of predictor coefficients	$L \times n$
B_{LM}	Coefficient matrix to be estimated	$L \times M$
$\Xi_M^{\mathbf{y}}(t)$	Basis functions for the response	$d \times M$
$\Xi_L^{\mathbf{x}}(s)$	Basis functions for the predictors	$p \times L$
E	Error matrix	$M \times n$
$\mathbf{c}_M^{\mathbf{y}}$	Coefficient vector for response basis expansion	$M \times 1$
$\mathbf{c}_L^{\mathbf{x}}$	Coefficient vector for predictor basis expansion	$L \times 1$
$\boldsymbol{\xi}_r^{\mathbf{x}}(s)$	Basis function for the r-th predictor component	$p \times 1$
$\boldsymbol{\xi}_q^{\mathbf{y}}(t)$	Basis function for the q-th response component	$d \times 1$
$\boldsymbol{\varepsilon}(t)$	Error term	$d \times 1$
Σ	Covariance matrix of distribution of β	$L \times M$
$\Lambda_L^{\mathbf{x}}$	Diagonal matrix of eigenvalues of $\mathbf{x}$	$L \times L$

Table 1: Summary of notations and dimensions.

general model equation for a function-on-function model, that is, both predictor and covariates are univariate functions, is given by (Gertheiss et al., 2023):

$$y(t) = \beta_0(t) + \sum_{j=1}^{p} \int_{\mathcal{S}_j} \beta_j(s,t)\, x_j(s)\, ds + \varepsilon(t), \quad t \in \mathcal{T},$$

where $y(t)$ represents a functional response variable, $\beta_0(t)$ is the intercept function, $\beta_j(s,t)$ is the bivariate coefficient function on domain $\mathcal{S}_j$, $j = 1, ..., p$, $x_1(t), ..., x_p(t)$ are the p covariate functions, and $\varepsilon(t)$ is the error function drawn from a stochastic process with mean zero and covariance function $\Sigma(s,t)$, $s,t \in \mathcal{T}$. The integral of the functional covariates and the bivariate coefficient function represents the influence of the entire trajectory of $x_j(t)$ on the response $y(t)$. The function $\beta_j(s,t)$ captures how the values of $x_j(s)$ at different points s contribute to the value of $y(t)$ at point t. This model can be further extended to include multiple non-functional covariates, as well as additive and random effects (Gertheiss et al., 2023; Volkmann et al., 2023).

Recall that the project at hand deals with multivariate functional data as response variable and may include univariate or multivariate functional covariates as well. Therefore, standard FLR techniques need to be extended to the **Multivariate Functional Linear Regression (mFLR)** model developed by Chiou, Yang, and Y.-T. Chen (2016).

2.3.1 Model Setup

Table 1 summarizes the notation and dimensions of each variable that follows in this section. Let $\mathbf{y}(t) = (y_1(t), ..., y_d(t))^T$, $t \in \mathcal{T}$, be a functional response variable of dimension d on domains $\mathcal{T}_1, ..., \mathcal{T}_d$, and $\mathbf{x}(s) = (x_1(s), ..., x_p(s))^T$, $s \in \mathcal{S}$, multiple functional covariates on

domains $\mathcal{S}_1, ..., \mathcal{S}_p$. Suppose that the variables are standardized using the dimension-wise mean functions and covariance functions defined in Section 2.1.3. Then, the mFLR is defined as:

$$y_k(t) = \beta_0(t) + \sum_{j=1}^{p} \int_{\mathcal{S}_j} \beta_{jk}(s,t) x_j(s)\, ds + \varepsilon_k(t), \quad t \in \mathcal{T}_k,\ k = 1, ..., d, \tag{8}$$

where $\beta_{jk}(s,t)$ is the bivariate regression coefficient function of covariate j and dimension k, and $\varepsilon_k(t)$ is the random error process. In matrix notation, Equation 8 can be expressed as:

$$\mathbf{y}(t) = \left(\int (\mathbf{x}(s))^T \beta(s,t)\, ds \right)^T + \boldsymbol{\varepsilon}(t), \quad t \in \mathcal{T}, \tag{9}$$

with $\beta(s,t) = (\beta_{jk}(s,t))_{j=1,...,p,k=1,...,d}$ and $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), ..., \varepsilon_d(t))^T$. Just as for FPCA and MFPCA, both predictor and covariates can be expressed by eigenfunctions which serve as basis functions, and estimated coefficients:

$$\mathbf{x}(s) = \sum_{r=1}^{\infty} c_r^{\mathbf{x}} \boldsymbol{\xi}_r^{\mathbf{x}}(s) = \Xi^{\mathbf{x}}(s)\, \mathbf{c}^{\mathbf{x}}, \quad s \in \mathcal{S}, \quad \text{and}$$
$$\mathbf{y}(t) = \sum_{q=1}^{\infty} c_q^{\mathbf{y}} \boldsymbol{\xi}_q^{\mathbf{y}}(t) = \Xi^{\mathbf{y}}(t)\, \mathbf{c}^{\mathbf{y}}, \quad t \in \mathcal{T},$$

where $\Xi^{\mathbf{x}}(s) = (\boldsymbol{\xi}_r^{\mathbf{x}}(s))_{r \geq 1}$ and $\Xi^{\mathbf{y}}(t) = (\boldsymbol{\xi}_q^{\mathbf{y}}(t))_{q \geq 1}$ represent the eigenfunctions of $\mathbf{x}$ and $\mathbf{y}$ with $\boldsymbol{\xi}_r^{\mathbf{x}}(s) = (\xi_{1r}^{\mathbf{x}}(s), ..., \xi_{pr}^{\mathbf{x}}(s))^T$ and $\boldsymbol{\xi}_q^{\mathbf{y}}(t) = (\xi_{1q}^{\mathbf{x}}(t), ..., \xi_{dq}^{\mathbf{y}}(t))^T$, and $\mathbf{c}^{\mathbf{x}} = (c_r^{\mathbf{x}})_{r \geq 1}$ and $\mathbf{c}^{\mathbf{y}} = (c_q^{\mathbf{y}})_{q \geq 1}$ are the corresponding coefficients. It should be noted that when an infinite number of basis functions are taken into consideration, the sum of the coefficients and eigenfunctions is no longer an approximation; rather, it is an equivalent expression for the variable in question.

With this representation, a minimizer of

$$\beta(s,t) = \underset{\beta(s,t)}{\arg\min}\, \mathbb{E}\left(\left\| \mathbf{y} - \left(\int (\mathbf{x}(s))^T \beta(s,t)\, ds \right)^T \right\| \right)^2$$

can be defined as:

$$\beta(s,t) = \sum_{r,q=1}^{\infty} \frac{\mathbb{E}(c_r^{\mathbf{x}} \cdot c_q^{\mathbf{y}})}{\lambda_r^{\mathbf{x}}} \boldsymbol{\xi}_r^{\mathbf{x}}(s) \boldsymbol{\xi}_q^{\mathbf{y}}(t)^T. \tag{10}$$

Here, $\lambda_r^{\mathbf{x}}$ represents the eigenvalue corresponding to $\boldsymbol{\xi}_r^{\mathbf{x}}$. Details on the derivation of that parameter estimate can be found in Supplement W1 of Chiou, Yang, and Y.-T. Chen (2016).

The representation of β in Equation 10 can be considered equivalent to a basis function representation. Since the estimated coeffient function is bivariate, a common approach is to use two different basis systems for the expansion, one for each variable. Since one variable corresponds to the domain $\mathcal{T}$ and the other to $\mathcal{S}$, it seems intuitive to choose the basis

system used to represent $\mathbf{y}$ and $\mathbf{x}$. Thus, Equation 10 represents a bivariate basis function expansion for β, which can be equivalently written as:

$$\beta(s,t) = \sum_{r,q=1}^{\infty} b_{rq} \boldsymbol{\xi}_r^{\mathbf{x}}(s) \boldsymbol{\xi}_q^{\mathbf{y}}(t)^T = \Xi^{\mathbf{x}}(s) B \, \Xi^{\mathbf{y}}(t)^T. \tag{11}$$

Here, $B = (b_{rq})_{r,q \geq 1}$ represents the basis coefficient matrix which needs to be estimated in the fitting process.

2.3.2 Model Estimation and Statistical Inference

The parameter function β derived in Equation 11 depends on an infinite amount of basis functions and basis coefficients, rendering it inaccessible for direct calculation. Consequently, the number of basis functions, or the number of PCs, must be constrained. In a manner analogous to the derivation of the FPCA, the appropriate number of PCs can be determined by the sum of the variation accounted for by each PC. Let L be the number of determined PCs for $\mathbf{x}$, and M for $\mathbf{y}$. This reduces Equation 11 to:

$$\beta^{LM}(s,t) \approx \sum_{r=1}^{L} \sum_{q=1}^{M} b_{rq} \boldsymbol{\xi}_r^{\mathbf{x}}(s) \boldsymbol{\xi}_q^{\mathbf{y}}(t)^T = \Xi_L^{\mathbf{x}}(s) B_{LM} \Xi_M^{\mathbf{y}}(t)^T,$$

where B_{LM} is the $L \times M$ basis coefficient matrix. In turn, this allows to reformulate the mFLR derived in Equation 9:

$$\mathbf{y}_M(t) = \left(\int \mathbf{x}_L(s)^T \beta_{LM}(s,t) \, ds \right)^T + \boldsymbol{\varepsilon}_M(t), \quad t \in \mathcal{T}. \tag{12}$$

Note that in this formulation, the error function $\boldsymbol{\varepsilon}_M$ is also represented as basis function expansion using the exact same M basis functions as for the response $\mathbf{y}$, that is,

$$\boldsymbol{\varepsilon}_M(t) = \Xi_M^{\mathbf{y}}(t) \, \boldsymbol{\varepsilon}_M, \tag{13}$$

with basis coefficient vector $\boldsymbol{\varepsilon}_M$. With that and the basis representations of $\mathbf{y}$ and $\mathbf{x}$, Equation 12 can be transformed as follows:

$$\begin{aligned}
& \mathbf{y}_M(t) = \left(\int \mathbf{x}_L(s)^T \beta_{LM}(s,t) \, ds \right)^T + \boldsymbol{\varepsilon}_M(t) \\
\Leftrightarrow \quad & \Xi_M^{\mathbf{y}}(t) \mathbf{c}_M^{\mathbf{y}} = \left(\int (\Xi_L^{\mathbf{x}}(s) \mathbf{c}_L^{\mathbf{x}})^T \, \Xi_L^{\mathbf{x}}(s) B_{LM} \Xi_M^{\mathbf{y}}(t)^T \, ds \right)^T + \Xi_M^{\mathbf{y}}(t) \boldsymbol{\varepsilon}_M \\
\Leftrightarrow \quad & \Xi_M^{\mathbf{y}}(t) \mathbf{c}_M^{\mathbf{y}} = \left(\int (\mathbf{c}_L^{\mathbf{x}})^T (\Xi_L^{\mathbf{x}}(s))^T \Xi_L^{\mathbf{x}}(s) B_{LM} \Xi_M^{\mathbf{y}}(t)^T \, ds \right)^T + \Xi_M^{\mathbf{y}}(t) \boldsymbol{\varepsilon}_M \\
\Leftrightarrow \quad & \Xi_M^{\mathbf{y}}(t) \mathbf{c}_M^{\mathbf{y}} = \left((\mathbf{c}_L^{\mathbf{x}})^T B_{LM} \Xi_M^{\mathbf{y}}(t)^T \int \Xi_L^{\mathbf{x}}(s)^T \Xi_L^{\mathbf{x}}(s) \, ds \right)^T + \Xi_M^{\mathbf{y}}(t) \boldsymbol{\varepsilon}_M \\
\Leftrightarrow \quad & \Xi_M^{\mathbf{y}}(t) \mathbf{c}_M^{\mathbf{y}} = \Xi_M^{\mathbf{y}}(t) B_{LM}^T \mathbf{c}_L^{\mathbf{x}} + \Xi_M^{\mathbf{y}}(t) \boldsymbol{\varepsilon}_M \\
\Leftrightarrow \quad & \mathbf{c}_M^{\mathbf{y}} = B_{LM}^T \mathbf{c}_L^{\mathbf{x}} + \boldsymbol{\varepsilon}_M.
\end{aligned} \tag{14}$$

This transformation implies that the functional regression problem can be reduced to a multivariate regression model using the basis coefficients of functional basis expansions for both the response and the predictors. Consequently, the objective of estimating β is transformed into estimating the $L \times M$ coefficient matrix B_{LM}. This can be performed using standard least square (LS) estimators for multivariate regression.

Therefore, let $Y(t) = (\mathbf{y}_i(t))_{1 \leq i \leq n}$ and $X(s) = (\mathbf{x}_i(s))_{1 \leq i \leq n}$ be n response and corresponding explanatory variables. Let $C_M^{\mathbf{y}}$ be the $M \times n$ matrix of response coefficients, $C_L^{\mathbf{x}}$ the $L \times n$ matrix of predictor coefficients, and E the $M \times n$ matrix of errors, respectively. The LS estimator of Equation 14 minimizes the sum of squared residuals. This can be expressed as:

$$B_{LM} = \underset{B_{LM}}{\arg\min} \, \|C_M^{\mathbf{y}} - B_{LM} C_L^{\mathbf{x}}\|_F^2 \,,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Taking the derivative and setting it equal to zero results in:

$$B_{LM} = C_M^{\mathbf{y}} (C_L^{\mathbf{x}})^T \left(C_L^{\mathbf{x}} (C_L^{\mathbf{x}})^T \right)^{-1} .$$

The coefficient matrices $C_M^{\mathbf{y}}$ and $C_L^{\mathbf{x}}$ and corresponding eigenfunctions $\Xi_M^{\mathbf{y}}$ and $\Xi_L^{\mathbf{x}}$ can be estimated using the spectral decomposition of the covariance functions of Y and X, respectively, as defined in Section 2.1.3. Details on their derivations are given in Supplement W2 of Chiou, Yang, and Y.-T. Chen (2016). Plugging in the estimates yields

$$\hat{B}_{LM} = \hat{C}_M^{\mathbf{y}} (\hat{C}_L^{\mathbf{x}})^T \left(\hat{C}_L^{\mathbf{x}} (\hat{C}_L^{\mathbf{x}})^T \right)^{-1} ,$$

and this results in the following estimate for the bivariate parameter function β:

$$\hat{\beta}(s,t) = \hat{\Xi}_L^{\mathbf{x}}(s) \hat{B}_{LM} (\hat{\Xi}_M^{\mathbf{y}}(t))^T .$$

Here, this matrix representation of $\beta(s,t)$ implies the notation β_{jk} for the j-th $\mathbf{x}$-dimension and the k-th $\mathbf{y}$-dimension.

With that, given a new standardized predictor $\mathbf{x}^*(s) = (x_1^*(s), ..., x_p^*(s))^T$, predictions can be estimated via:

$$\hat{\mathbf{y}}^*(t) = \hat{\Xi}_L^{\mathbf{y}}(t) \hat{B}_{LM} \hat{\mathbf{c}}^{\mathbf{x}^*} ,$$

where $\hat{\mathbf{c}}^{\mathbf{x}^*}$ can be estimated by a weighted LS approach derived in Appendix A.1 of Chiou, Yang, and Y.-T. Chen (2016).

For both $\hat{\beta}$ and $\mathbf{y}^*$, pointwise confidence intervals for any $t \in \mathcal{T}$ and simultaneous confidence bands for all $t \in \mathcal{T}$ can be derived. For any scalar vector $\mathbf{a} = (a_1, ..., a_d)^T$, a $100(1-\alpha)\%$ pointwise confidence interval $[CI_L(t), CI_U(t)]$ for $\mathbb{E}[\mathbf{a}^T \mathbf{y}^*(t)]$ is given by:

$$\mathbf{a}^T \hat{\mathbf{y}}_{LM}^* \pm \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \left[\mathbf{a}^T \hat{\omega}_{LM}(t,t) \, \mathbf{a} \right]^{\frac{1}{2}} , \tag{15}$$

where Φ is the standard normal cumulative distribution function and $\hat{\omega}_{LM}$ is the estimate of the covariance function $\text{cov}(\mathbf{y}_{LM}^*(t)|X, \mathbf{x}^*)$. The simultaneous confidence band $[CB_L(t), CB_U(t)]$ for all $t \in \mathcal{T}$ can be approximated by:

$$\mathbf{a}^T \hat{\mathbf{y}}_{LM}^* \pm \left[\chi_{L,1-\alpha}^2 \mathbf{a}^T \hat{\omega}_{LM}(t,t) \, \mathbf{a} \right]^{\frac{1}{2}} ,$$

where $\chi^2_{L,1-\alpha}$ is the $100(1-\alpha)\%$-th percentile of the chi-squared distribution with L degrees of freedom. Prediction intervals can be determined similarly as Chiou, Yang, and Y.-T. Chen (2016) show in Chapter 3.2.

In order to obtain a confidence band for the coefficient function $\hat{\beta}(s,t)$, Chiou, Yang, and Y.-T. Chen (2016) state that under multiple conditions (derived in Chapter 4 and Appendix A.2) the LS estimate $\hat{\beta}$ is asymptotically normally distributed such that for all $j = 1, ..., p$ and $k = 1, ..., d$,

$$\sqrt{n}\left(\hat{\beta}_{jk}(s,t) - \beta_{jk}(s,t)\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{jk}(s,t)\right),$$

where the covariance matrix Σ is given by $\Sigma_{\cdot k} = B^{\mathbf{x}}(s)(\mathbf{e}_k^T B^{\mathbf{y}}(t)\mathbf{e}_k)$ with

$$B^{\mathbf{x}}(s) = \Xi_L^{\mathbf{x}}(s)^T(\Lambda_L^{\mathbf{x}})^{-1}\Xi_L^{\mathbf{x}}(s) \text{ and } B^{\mathbf{y}}(t) = \Xi_M^{\mathbf{y}}(t)^T\Omega_M\Xi_M^{\mathbf{y}}(t),$$

where $\Lambda_L^{\mathbf{x}} = diag(\lambda_1^{\mathbf{x}}, ..., \lambda_L^{\mathbf{x}})$ and $\Omega_M = cov(\varepsilon_M, \varepsilon_M)$ as derived in Equation 13. With that, a pointwise $100(1-\alpha)\%$ confidence interval for $\beta_{jk}(s,t)$ is given by

$$\hat{\beta}_{jk} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\left[\frac{1}{n}\hat{\Sigma}_{jk}(s,t)\right]^{\frac{1}{2}}.$$

As before for $\mathbf{y}^*$ in Equation 15, these pointwise confidence intervals can be extended to a simultaneous confidence band for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$:

$$\hat{\beta}_{jk}(s,t) \pm \left[\frac{1}{n}\chi^2_{L,1-\alpha}\hat{\Sigma}_{jk}(s,t)\right]^{\frac{1}{2}}.$$

The asymptotic properties of the estimators are presented in Chapter 4 of Chiou, Yang, and Y.-T. Chen (2016). The key results include the consistency and asymptotic normality of the estimators, which provide a basis for valid statistical inference. The rate of convergence and the required regularity conditions are also discussed, with particular emphasis on the importance of selecting an appropriate number of basis functions and ensuring the smoothness of the coefficient functions.

2.3.3 Numerical Details

The previously derived FPCA and MFPCA scores in Section 2.2.2 and Section 2.2.3 can be employed as a means of facilitating the numerical approximation of suitable parameter estimates. In practice, the mFLR minimization problem is transformed into a multivariate regression problem as follows:

1. Perform a FPCA or an MFPCA for the functional target variable and functional covariates. This induces M and L PC scores for every observation i, respectively.

2. Set up the model including the $n \times M$ PC score matrix as multivariate target matrix, and the $n \times L$ PC score matrix as multivariate response. If needed, the covariate matrix can be further extended to multiple non-functional covariates.

3. Fit the resulting multivariate regression model. If the model is used for prediction, the predicted scores need to be transformed with help of the basis functions of the target derived in the first step to yield predicted functions.

With that, a mFLR can be implemented efficiently. In practice, R package `MFPCA` Happ-Kurz (2020) is required for performing MFPCAs or FPCAs, and standard packages are sufficient for linear regression only. The inclusion of additive effects can be implemented using `mgcv` Wood (2001).

2.4 K-means Clustering

In contrast to the preceding sections, which focused on functional data, this section considers standard non-functional data points. According to Wu (2012), **k-means Clustering** is a very simple, robust and highly efficient unsupervised machine learning technique that is employed for detecting patterns within a data set. The algorithm aims to divide the data into k non-overlapping subgroups, so called *clusters*, with similar features. Similarity is assessed based on distance measures like the *Euclidean distance* between the data points and a centroid, i.e., the cluster mean of data points within a cluster.

In mathematical terms, let $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m\}$ be a data set of m data vectors to be clustered and k the chosen number of clusters. The clustering is conducted in four steps:

1. Select k initial centroids randomly, namely $\mathbf{b}_1, ..., \mathbf{b}_k$.

2. Assign every data point in $\mathcal{D}$ to the closest centroid $\mathbf{b}_{j^*}$ by calculating the minimum distance to each centroid $\mathbf{b}_j$:

$$j^*(\mathbf{z}_i) = \underset{j=1,...,k}{\arg\min} \|\mathbf{z}_i - \mathbf{b}_j\|^2,$$

   where $\mathbf{b}_j$ is the centroid of cluster C_j. Each collection of data points assigned to a centroid forms a cluster.

3. Update the centroids based on the points assigned to the respective cluster:

$$\mathbf{b}_j = \sum_{\mathbf{z}_i \in C_j} \frac{\mathbf{z}_i}{|C_j|}.$$

4. Repeat steps 2 and 3 until no data point changes the cluster.

This 4-step procedure, which can be applied to a wide range of data types, is a gradient-descent alternating optimization method minimizing the total within-cluster sum of squares

$$\sum_{j=1}^{k} \sum_{\mathbf{z}_i \in C_j} \|\mathbf{z}_i - \mathbf{b}_j\|^2,$$

that often converges to a local minimum or saddle point.

There is no clear answer to the question how to select an appropriate number of clusters k. A heuristic that is often used in practice is the *elbow method*. Therefore, the within-cluster sum of squares is plotted against multiple possible numbers of k, e.g., $k \in \{1, ..., 10\}$. Then, the presence of a bend in the plot ("elbow") indicates the optimal number of clusters, suggesting a notable drop in the rate of change of the within-cluster sum of squares beyond that point. However, other considerations, including domain knowledge, may suggest a different number of clusters than indicated by the elbow plot and may be more reasonable depending on the application.

In order to assess the similarity between cluster assignments obtained by the proposed k-means algorithm, or equivalently by any clustering algorithm, several measures exist. One popular metric is the **Adjusted Rand Index (ARI)** developed by Hubert and Arabie (1985). It is based on the simple Rand Index (RI) introduced by Rand (1971) which considers all possible pairs of data points and determines how consistently they are clustered together or separately in two partitions. However, this index is highly dependent on the overall number of clusters and the number of elements in each cluster, potentially leading to biased evaluations when comparing clusterings with different numbers of clusters or highly unbalanced clusters (Morey and Agresti, 1984; Fowlkes and Mallows, 1983). The ARI is independent of cluster sizes because it corrects for chance agreement between clusterings by taking into account the expected similarity between random clusterings. Mathematically, let U and V be two different partitions of a data set of n elements. Then the ARI is defined as (Hubert and Arabie, 1985)

$$ARI(U,V) = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

where

- n_{ij} is the number of elements in common in cluster i of partition U and cluster j of partition V,
- a_i is the number of elements assigned to cluster i of partition U, and
- b_j is the number of elements assigned to cluster j of partition V.

The ARI ranges from -1, indicating complete disagreement between clusterings, to 1, perfect agreement, with 0 representing the expected value for random clusterings. An overview over possible other methods for comparing clusters is given by S. Wagner and D. Wagner (2007).

3 Data Description

The goal of this chapter is to dive deeper into the generation process of the data analysed in this project. This includes giving a brief introduction into the boreal biome and its vegetation dynamics, and how they are simulated in the LPJ-GUESS model. The chapter also includes steps for data pre-processing, which convert the data into a form that is suitable for further analyses in the following chapters.

3.1 The Boreal Forest and its Vegetation

The **Boreal Forest**, described in Pfadenhauer and Klötzli (2020), also known as Taiga, is a vast biome covering about 13% of the land surface. It spans across the northern parts of North America, Europe, and Asia, forming a circumpolar belt between approximately $50°N$ and $70°N$. The climatic conditions are characterized by short, cold summers with mean temperatures hardly above $10°C$ and long winters that are extremely cold (down to $-30°C$) and rich in snow. The vegetation is dominated by evergreen conifers, i.e., various species of firs (*Abies*), spruces (*Picea*) and pines (*Pinus*). In addition to conifers, some regions of the boreal forest include deciduous trees like birches (*Betula*) and aspens (*Populus*), particularly in the southern areas where the climate is slightly milder.

The boreal biome is immensely important for regional and global climate (Bonan, 2008). It serves as one of the largest carbon sinks on the planet by storing vast amounts of carbon in its trees, soil, and permafrost, i.e., permanently frozen ground (Pan et al., 2011). This storage contributes to reducing CO_2 in the atmosphere, which in turn helps to mitigate the effects of climate change (Bonan, 2008). Moreover, Swann et al. (2010) state that the boreal biome has major impact on albedo and evapotranspiration. During winter, the snowcovered forest has a high albedo, that is, a high fraction of solar radiation that is reflected by the Earth's surface back into space. In contrast, dark forest canopies that appear due to the coniferous vegetation absorb more sunlight, which can influence local and regional temperature dynamics. In addition, the boreal forest affects the hydrological cycle. Through evapotranspiration, i.e., the combined process of water evaporation from the Earth's surface (such as soil, vegetation, and water bodies) and transpiration from plants, which affects atmospheric moisture and precipitation patterns, the boreal forest may impact weather and climate globally.

3.2 The LPJ-GUESS Model

The **Lund-Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS)** provides a modelling framework to simulate interactions between vegetation, climate and land use at regional to global scales (Smith, Prentice, and Martin T Sykes, 2001; Smith, Wårlind, et al., 2014). It is a dynamic vegetation model (DVM), which is defined as a simulation model used to simulate the growth, distribution and dynamics of vegetation over time in response to various environmental conditions and disturbances (Foley et al., 2000). LPJ-GUESS incorporates a model of plant physiology that is process-based, comprising photosynthesis, respiration, and evapotranspiration. This feature enables the simulation of disturbances, mortality,

PFT	Full Name	Description/Example Species
BNE	Boreal needleleaf evergreen	European spruce (picea abies)
IBS	Pioneering shade-intolerant broadleafed summergreen	Moor birch (betula pubenscens), Common aspen (populus tremula)
otherC	Other conifers	All coniferious PFTs (aggregated)
TeBS	Temperate broadleafed summergreen	Wych elm (ulmus glabra hudson)
Tundra	Tundra	Various grass and shrub PFTs (aggregated)

Table 2: Overview of PFTs occurring in the boreal forest used for simulation.

and establishment, making it particularly suitable for examining disturbance regimes and the regeneration process following disturbances.

Vegetation is a very broad term, as it encompasses the entirety of plant species on an area. As there are an enormous number of different plant species, there is a need for a classification system for appropriate modelling. **Plant Functional Types (PFTs)** define groups of plant species with similar functions and performances in an ecosystem. The PFTs modelled in the framework of LPJ-GUESS are based on indicators such as growth form (tree, shrub or herb) and leaf phenology type (evergreen, summergreen, raingreen). Table 2 provides an overview of the boreal PFTs considered in this study. Since the boreal forest in mainly dominated by needleleafed trees (Pfadenhauer and Klötzli, 2020), two PFTs cover those species. One PFT represents grass and shrubs (*tundra*), while the last two PFTs characterize broadleafed trees: *pioneering broadleaf* is shade-intolerant and early-successional, a often fast-growing tree species that quickly colonize disturbed areas. These trees play a crucial role in ecological succession, stabilizing the soil, providing shade, and creating conditions that allow other, more shade-tolerant species to establish (Kimmins, 2004). *Temperate broadleaf* are deciduous summergreen species that are adapted to the seasonal climate of temperate regions but can also be found in the boreal forest, particularly in transitional zones where boreal and temperate forests meet (Kimmins, 2004).

The simulation model takes gridded climate parameters including temperature, precipitation and downwelling shortwave radiation, as well as atmospheric CO_2 concentrations and soil properties as inputs. After simulating the vegetation over a pre-defined time period, the model outputs vegetation composition in terms of PFTs, which are derived from aboveground carbon or other indicators like leaf area index (LAI) or net primary production (NPP). Details on the mathematical formulations of the underlying vegetation mechanisms including establishment and mortality are given by Smith, Prentice, and Martin T Sykes (2001) and Sitch et al. (2003).

LPJ-GUESS utilizes a hierarchical model structure to accurately depict population dynamics, including disturbances. The area of concern is depicted by grid cells defined by the input climate data. In each grid cell, the model simulates multiple independent patches – 25 in this study – to reflect landscape heterogeneity. This yields 25 independent samples

from each grid cell. Vegetation dynamics within these patches result from growth and competition for light, space and soil resources among woody plant cohorts and a herbaceous understory, influenced by random disturbances. These disturbances occur annually with a chosen yearly probability p_D on patch-level. When such a disturbance happens, all living vegetation carbon in a patch is moved to the litter pool. Note that the simulation model does not distinguish between different types of disturbances such as wildfires and storms; it only plays with the number of disturbances per year while keeping the intensity and size constant. The resulting output of the model, that is, the composition of PFTs in terms of several indicators, is averaged across all patches in a grid cell.

One of the objectives of LPJ-GUESS is to provide a framework for exploring the future impacts of varying climate change scenarios on vegetation behaviour in different regions (Smith, Wårlind, et al., 2014). The input climate data are taken from the **Intersectoral Model Intercomparison Project (ISIMIP)** repository, which has a resolution of $0.5° \times 0.5°$ (Lange and Büchner, 2021). For this project, LPJ-GUESS is run for all climate scenarios available from ISIMIP, namely SSP1-RCP2.6, SSP3-RCP7.0 and SSP5-RCP8.5, and a control scenario. Details on the definition of the climate scenarios are given in the next section.

3.3 Climate Scenarios

According to the Special Report on Emission Scenarios (SRES) of the **Intergovernmental Panel on Climate Change (IPCC)**, climate scenarios are images of the future, or alternative futures, that are not intended to predict but to improve understanding of how systems behave, evolve and interact (Nakicenovic et al., 2000). The emission scenarios developed in this report which cover different narratives of the future, have served as a basic framework for climate change research for many years, helping researchers to gain greater insight into the complex dynamics of the climate and the Earth system. However, SRES suffer from several drawbacks: for example, these scenarios do not consider climate policies that could be implemented through political decisions, including their costs, benefits and risks (Moss et al., 2010). One approach to provide a new basis for long-term and near-term modelling experiments are the **Representative Concentration Pathways (RCPs)** developed by van Vuuren et al. (2011). The RCPs are based on radiative forcing, a measure of the imbalance in the Earth's energy budget caused by changes in greenhouse gases, aerosols, albedo and other factors. There are several RCPs, each corresponding to a different level of radiative forcing by the year 2100, ranging from 2.6 to 8.5 $\mathrm{W/m^2}$ compared to pre-industrial levels. RCPs focus on possible futures of greenhouse gas emissions and their concentrations in the atmosphere deliberately exclude socioeconomic factors. Pathways that take into account key global trends such as population and economic growth, urbanisation and education have been developed by the climate change research community as the **Shared Socioeconomic Pathways (SSP)** (Riahi et al., 2017). The SSPs comprise five different narratives (SSP1-SSP5), focusing on different energy and land use trajectories and associated uncertainties for greenhouse gas and air pollutant emissions.

The climate scenarios investigated in this analysis are combinations of SSPs and RCPs to account for both socioeconomic changes and radiative forcing. They are summarized in

Scenario	SSP and RCP narrative	Description
Control	-	Constant pre-industrial CO_2 emissions (i.e., at level of 1850) including interannual variability (e.g., weather phenomenons like El Niño)
SSP1-RCP2.6	SSP1: Low challenges to mitigation and adaptation: Transition to a sustainable, inclusive development. RCP2.6: Increase of 2.6 W/m^2 of radiative forcing compared to pre-industrial levels	Low warming, best case scenario in line with the Paris agreement
SSP3-RCP7.0	SSP3: High challenges to mitigation and adaptation: Nationalism rises, global focus narrows. RCP7.0: Increase of 7.0 W/m^2 of radiative forcing compared to pre-industrial levels	High warming, medium scenario
SSP5-RCP8.5	SSP5: High challenges to mitigation, low challenges to adaptation: Market-driven progress, resource exploitation. RCP8.5: Increase of 8.5 W/m^2 of radiative forcing compared to pre-industrial levels	Very high warming, worst case scenario (sometimes referred to as business as usual (BAU)). Despite the name this scenario is considered to be unlikely at the moment.

Table 3: Description of the climate scenarios including SSP and RCP narratives.

detail in Table 3.

3.4 Simulation Runs

For each scenario, the simulation spans the years 1850 to 2299, considering different phases. First, from 850 to 1850, a spinup phase is used in order to develop a baseline vegetation in each patch, i.e., the model is spun up recycling the pre-industrial climate of 1800 to 1829. After spinup, that is, in the year 1850, the area of concern is dominated by *needleleaf evergreen* trees in terms of aboveground carbon. For the three climate scenarios, the historical warming is used to simulate until 2015, while for the control scenario no warming is assumed. From 2015 to 2100 is the experimental phase, which includes running the simulation for each climate scenario and a return period of disturbances of 150 years (i.e., $p_D = 0.0067$). After 2100 until the end of the simulation in year 2299, the model is run recycling data from 2095 to 2100 (spindown).

All grid cells within the boreal biome as defined by the World Wildlife Fund (WWF) classi-

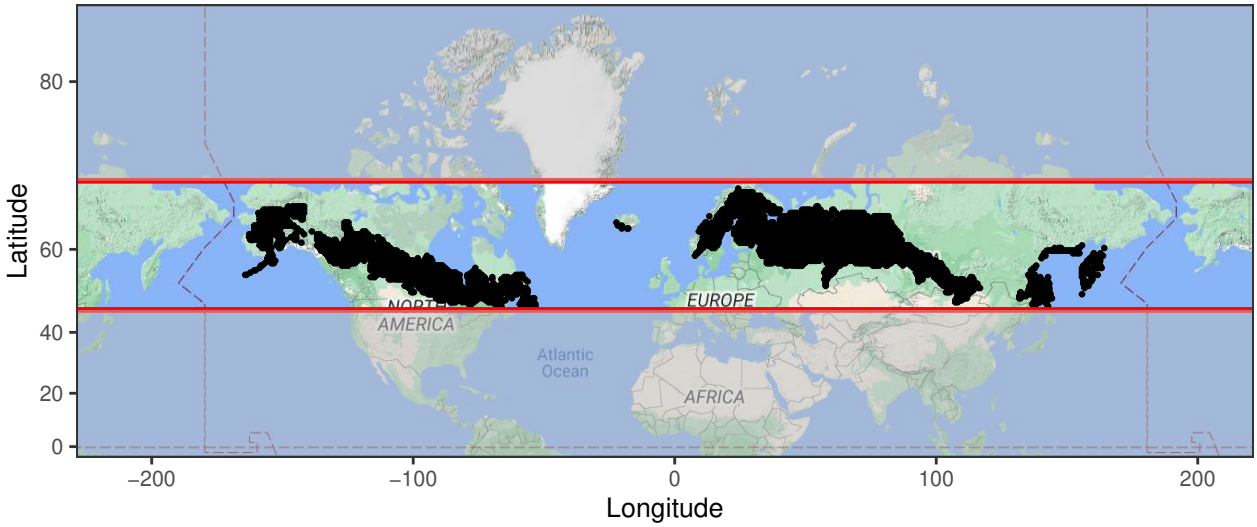


Figure 2: Area of concern. The black dots indicate the grid cells.

fication of terrestrial ecoregions of the world (Olson et al., 2001) are used for simulation. In total, the data comprises 5156 different grid cells, defined by their centroids, each consisting of 25 independent patches. The grid cells in the affected area are visualised in Figure 2. 41% of the grid cells are located in America, while 44% of them are in Asia. The remaining 15% are scattered across Northern Europe.

The output of LPJ-GUESS is a time series from 1850 to 2299 with the disturbance history and aboveground carbon for each grid cell, patch and PFT. A detailed description of the variables is provided in Table 4, and details on selected variables are given in the next section. Note that the final data consists of four separate data sets, one for each scenario.

3.5 Details on Soil, Climate and Ecological Covariates

The soil covariates in Table 4 are taken from the Harmonized World Soil Database. The variable descriptions in the following are taken from Chapter 2.3.3 from its report (FAO and IIASA, 2023). According to their size, mineral particles in a soil are grouped into soil texture classes which can be diagrammatically represented by the soil texture triangle depicted in Figure 3. Each soil comprises sand, silk and clay, and the portion of each of the three components determines properties of the soil. **Sand** grains are particles of size 0.050 to 2 mm and feel gritty when rubbed between fingers. In comparison, **silt** feels like flour and is produced by mechanical weathering of rock. In contrast, **clay** results of chemical weathering and its particles have diameters less than 0.002 mm. For details, see FAO and IIASA (2023). Next to the soil composition, the data set provides three other soil related covariates. **Bulk density** measured in g/cm is defined as the mass of the many particles of the material divided by the total volume they occupy (FAO and IIASA, 2023). The **pH level** measured in soil-water solution is an indicator of acidity and alkalinity of the soil, which determines among others the health of the soil. It ranges from 0 to 14, where 7 is considered to be neutral, with pH values below 7 indicating acidity and above 7 indicating alkalinity. Health is also

	Name	Description	Unit	Range
Location	Year	Simulation year		{1850, 2299}
	Lon	Longitude of grid cell	degrees (°)	$[-164.25, 164.25]$
	Lat	Latitude of grid cell	degrees (°)	$[47.25, 68.75]$
	PID	Patch ID		{0,...,24}
Recovery	PFT	Plant functional type		{BNE, IBS, otherC, TeBS, Tundra}
	dhist	Disturbance history		1 if disturbance occurred in a year, 0 otherwise
	ndist	Number of disturbances in that patch since the start of the simulation		$\mathbb{N}_0$
	age	Patch age, i.e., the time since the last disturbance		$\mathbb{N}_0$
	cmass	Aboveground carbon	kg/m^2	$\mathbb{R}^+$
Ecological	initial_recruitment	Number of new seedlings per PFT right after disturbance		$\mathbb{N}_0$
	recruitment_ten_years	Number of new seedling per PFT in the ten years after disturbance (summed up)		$\mathbb{N}_0$
	time_since_dist	Years since last disturbance		$\mathbb{N}_0$
	previous_state	Vegetation composition before the disturbance	kg/m^2	$\mathbb{R}^+$
	Nuptake_total	Total nitrogen uptake of the grid cell	g/m^2	$\mathbb{R}^+$
	Nuptake	Nitrogen uptake per PFT on grid cell level	mg/g	$\mathbb{R}^+$
Soil	sand_fraction	Sand fraction in soil		$[0, 1]$
	silt_fraction	Silt fraction in soil		$[0, 1]$
	clay_fraction	Clay fraction in soil		$[0, 1]$
	bulk_density_soil	Reference bulk density	g/cm^3	$\mathbb{R}^+$
	ph_soil	pH level measured in a soil-water solution		$[0, 14]$
	soilcarbon	Organic carbon content	g/kg	$\mathbb{R}^+$
Climate	tas_yearlymean	Yearly mean temperature	K	$\mathbb{R}^+$
	tas_yearlymin	Daily minimum temperature per year	K	$\mathbb{R}^+$
	tas_yearlymax	Daily maximum temperature per year	K	$\mathbb{R}^+$
	pr_yearlysum	Yearly precipitation (summed up)	kg/m^2	$\mathbb{R}^+$

Table 4: Variable names of the provided data and additional information.

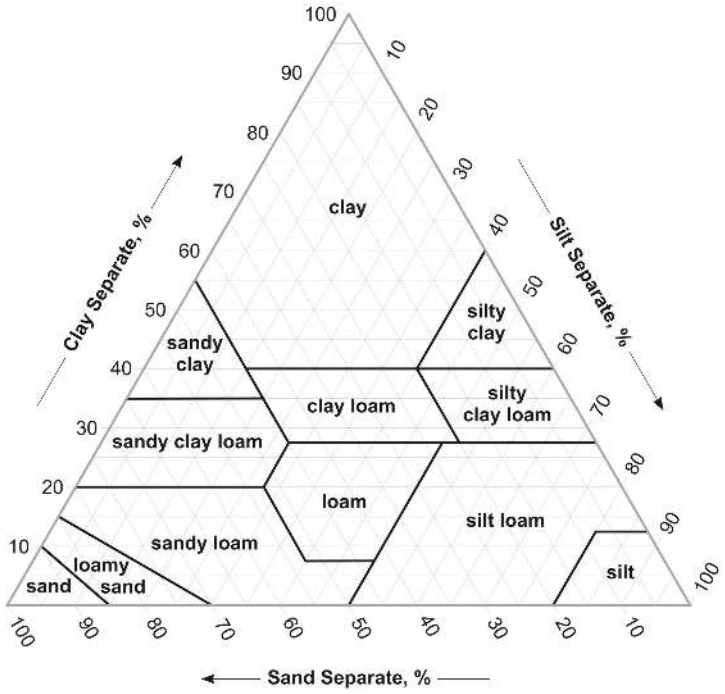


Figure 3: Soil textural triangle. Source: Modified from Thien (1979).

indicated by the **organic carbon content** measured in g/kg. Moderate to high amounts of organic carbon are associated with a good soil structure and fertility (FAO and IIASA, 2023).

The climate variables including **temperature** and **precipitation** are provided for each individual grid cell. The yearly mean temperature corresponds to the annual mean temperature per grid cell, while the annual minimum and maximum temperature represent the average annual minimum and maximum temperature values, respectively. The yearly precipitation is the summed up annual precipitation per grid cell and comprises all kinds of precipitation including rain, snow and hail.

The ecological variables consist of two realisations of random processes. **Initial recruitment**, and therefore **recruitment after ten years**, is the expected number of newly established individuals per PFT in that year. Establishment is realised as a Poisson process and is therefore not the actual number of new seedlings but their expectation. Another indicator for tree growth is given by **nitrogen uptake**. According to Chapin, Matson, and Vitousek (2011), it refers to the amount of nitrogen absorbed from the soil or atmosphere by the plant biomass, e.g., by roots, over a specific time span. High values are indicative of biological productivity and good soil conditions, and may imply growth and adaptation of individual PFTs. Finally, the **vegetation prior to the last disturbance** is given in terms of aboveground carbon.

3.6 Data Pre-Processing

The research objectives derived in Section 1 aim to investigate vegetation recovery and its composition with regard to PFTs after disturbance. LPJ-GUESS provides time series of aboveground carbon only, so to analyse dominance of PFTs in terms of aboveground carbon during recovery the data needs to be adjusted accordingly for each scenario:

1. Select a time period of several years (usually 25) and take patches that experience a disturbance within that period. Only consider patches that have recovered for at least 100 years without being interrupted by another disturbance.

2. The analysis aims at identifying dominant tree species, so the proportion of each PFT is considered instead of absolute values of aboveground carbon (i.e., `cmass`). Therefore, for each disturbed patch, calculate the relative contribution of each of the five PFTs considered in this study to the recovery trajectories.

In the analyses that follow, the time period is chosen to be the beginning of the experimental phase, that is, 2015 to 2040. Thus, the recovery trajectories capture the years 2015 until 2140 which represent a maximum of 126 years in total.

These pre-processing steps ensure that the curves are aligned at the start of the recovery period, regardless of when the actual disturbance occurred in the time frame. As a result, no data registration, i.e., data alignment, or other data sorting algorithms are required. This implies that the data is prepared for analyses like FPCAs or MFPCAs and statistical modelling.

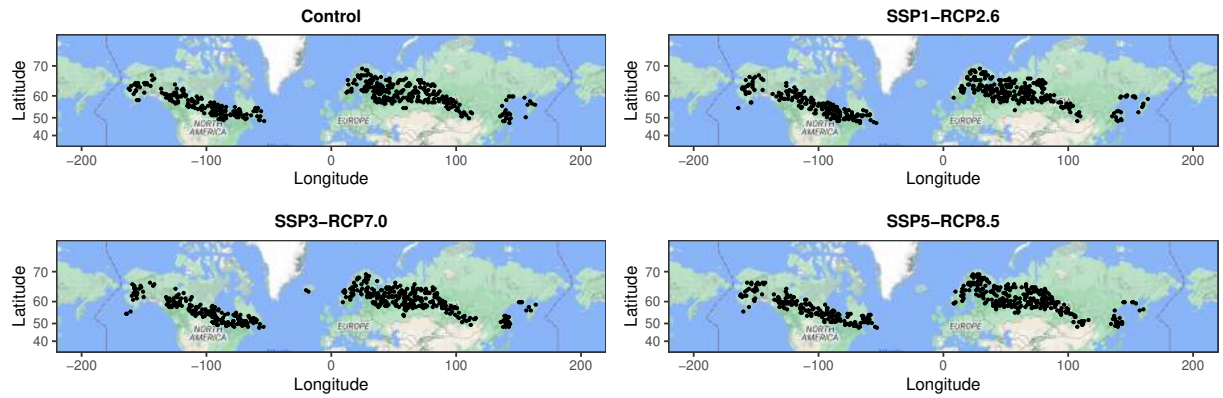


Figure 4: Map of disturbed patches between 2015 and 2040.

4 Exploratory Analysis

Now that a clear understanding of the methodology and data has been established, this chapter delves into the study of recovery curves.

The analysis in this chapter is based on recovery trajectories for patches disturbed between 2015 and 2040. Due to computational reasons, only one patch, i.e., one realisation of each grid cell, is taken into account. This means in particular, that the terms *grid cell* and *patch* are used as synonyms in the following.

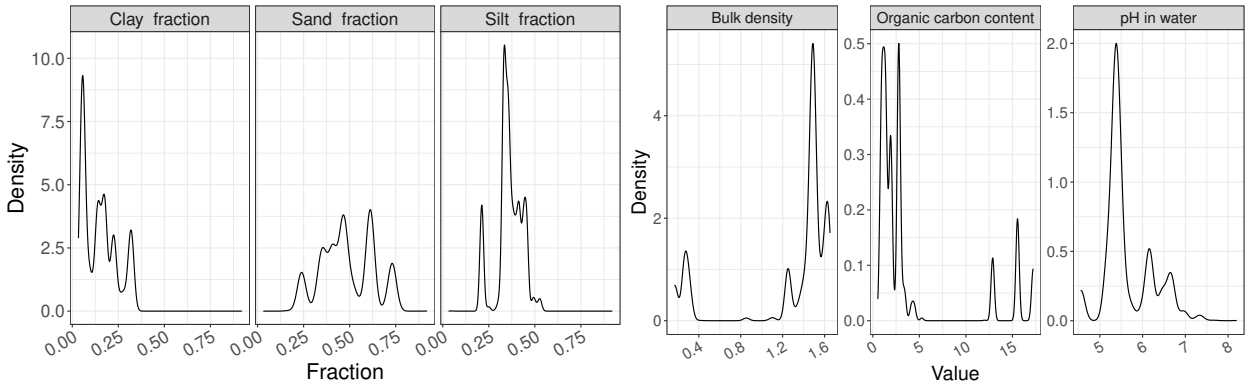
Several R packages implement methods for handling functional data. The most popular one which is also used in this project, is the package `fda` developed by James O. Ramsay, Giles Hooker and Spencer Graves (J. O. Ramsay, Hooker, and Graves, 2009). It provides methods for creating functional data objects, conducting exploratory data analysis like FPCA, and implementing various FDA models like functional regression. The R packge `MFPCA` developed by Clara Happ-Kurz extends the scope of FDA by focusing on multivariate functional principal component analysis (Happ and Greven, 2018; Happ-Kurz, 2020). It comprises tools for the simultaneous decomposition of multiple functional data sets, capturing the major modes of variation across multiple dimensions.

All following code was conducted on a Linux system with an AMD Ryzen 5 5625U and 16 GB of RAM using R 4.4.1.

4.1 Descriptive Statistics

In order to get a first impression of the recovery trajectories for patches disturbed between 2015 and 2040, this section provides a brief description of the pre-processed data set.

For each scenario, the disturbed grid cells are scattered across the boreal biome, as shown in Figure 4. The recovery trajectories comprise 434 disturbed grid cells for the control scenario, 442 for SSP1-RCP2.6, while for SSP3-RCP7.0 and SSP5-RCP8.5 there are 462 and 465 grid



(a) (b)

Figure 5: Soil composition (a) and further soil attributes (b).

cells, respectively. This adds up to 1803 disturbed patches in total. Since four independent runs of the vegetation model LPJ-GUESS are considered, the grid cells disturbed in each scenario may vary or may coincide. In total, 1577 unique grid cells are disturbed, where 226 of them are disturbed in at least two scenarios.

In order to get an intuition of the soil in the boreal biome, Figure 5 shows ridge plots of the soil composition (a) and further soil attributes introduced in Section 3.5 (b). Evaluating the mean values and quantiles at the textural triangle (Figure 3) reveals the soil to be mostly loam, sandy loam and loamy sand. The soil attributes show some variation in the data: bulk density is on average close to 1.3 g/cm^3, while approximate 4.2 g/kg of organic carbon is stored in the soil on average. The mean pH level is at 5.6, which indicates a slightly acid environment.

To gain insight into the climatic conditions at the disturbed grid cells, Figure 6 shows the annual mean, minimum and maximum temperature as well as the summed precipitation averaged over all disturbed grid cells for each scenario between 2015 and 2140. Note that the temperatures are converted to the unit °C for easier interpretation. All three temperature curves reveal the same patterns: the three warming scenarios behave similarly in the years 2015 to 2050 with all three scenarios facing an increase in average temperature (Figure 6a). Thereafter, the mean temperature, as well as the minimum and maximum temperatures, decrease for the SSP1-RCP2.6 scenario, while they continue to increase for the SSP3-RCP7.0 and SSP5-RCP8.5 scenarios until 2100. This pattern is also present when annual precipitation is considered. While the behaviour is similar for the two most extreme scenarios, the SSP1-RCP2.6 scenario experiences a lot of precipitation at the beginning of the study period, which decreases over the time period. Note that for both temperature and precipitation there are no changes in the trends for the control scenario.

As described in Section 3.5, nitrogen uptake is an indicator of vegetation growth. Figure 7 shows the nitrogen uptake per PFT and scenario (a), and the total uptake scenario-wise averaged over all disturbed grid cells (b). The former figure suggests that there is no trend

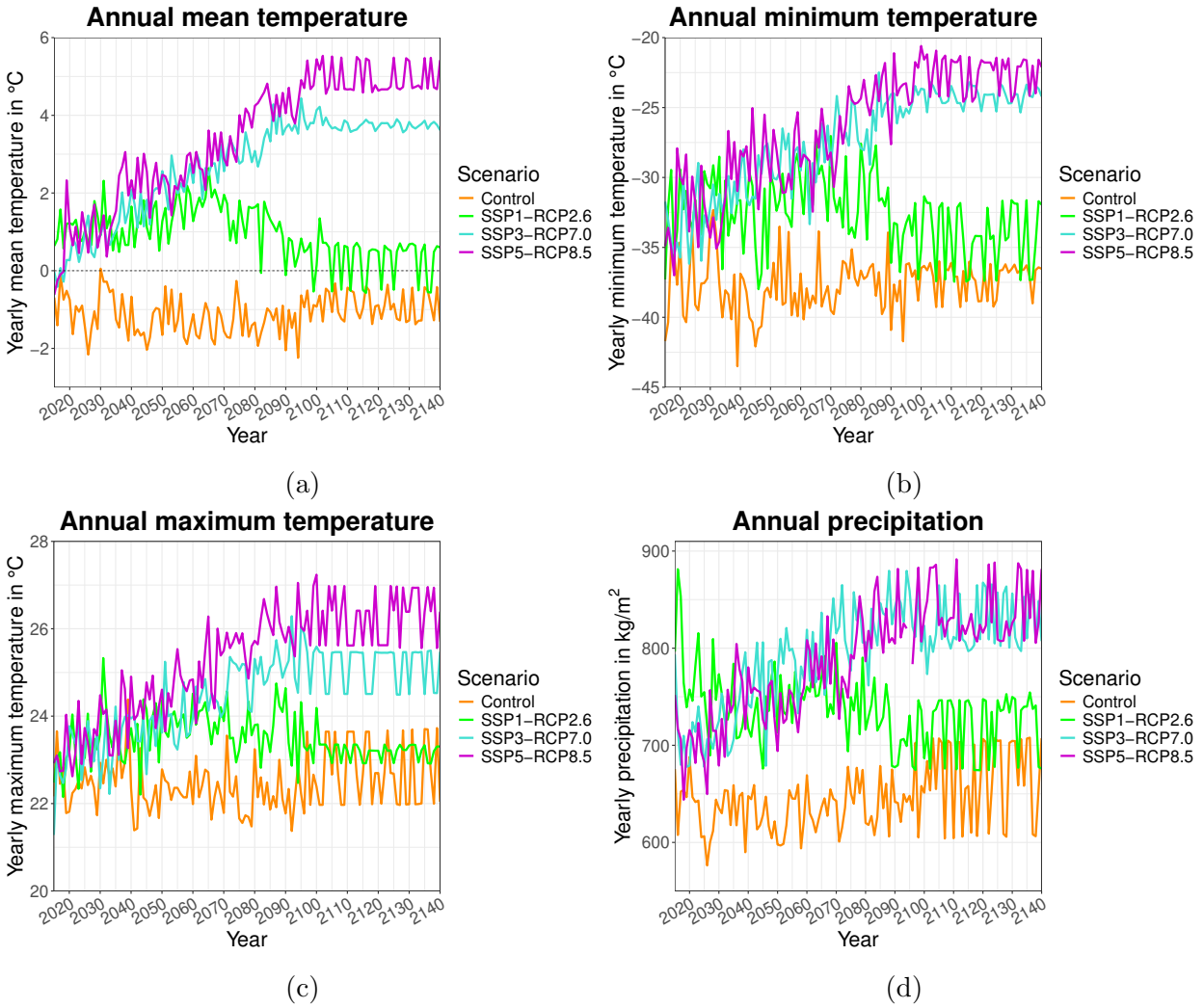
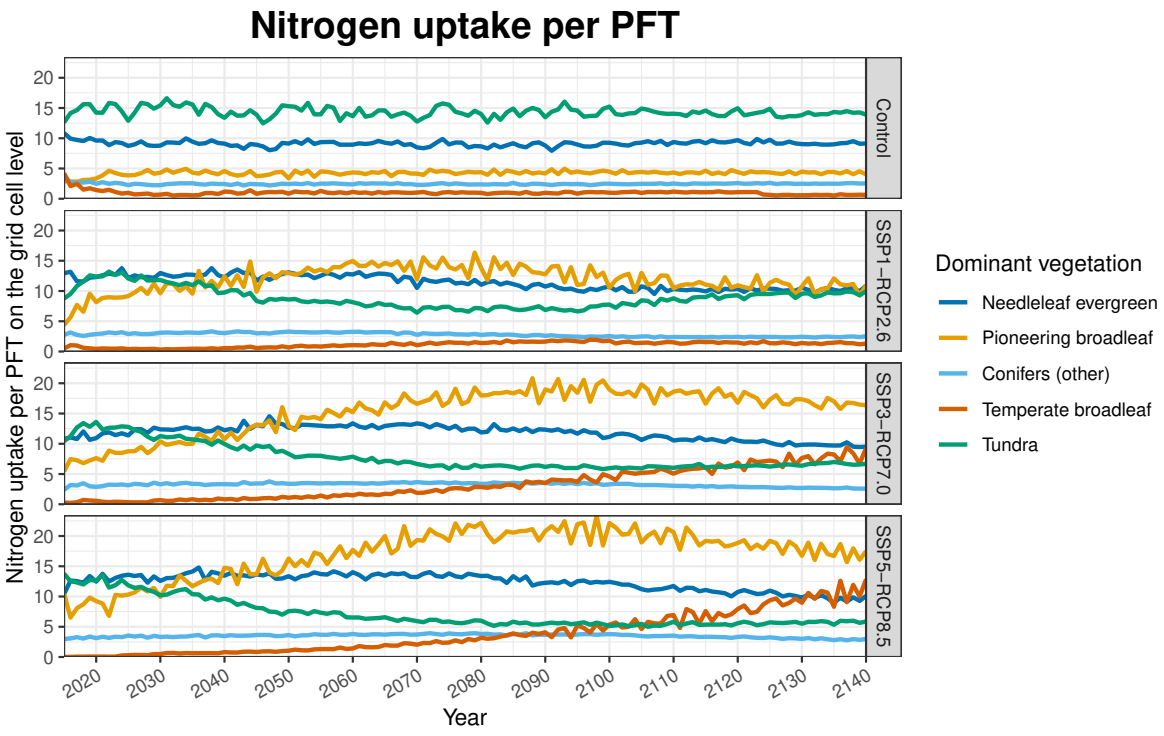
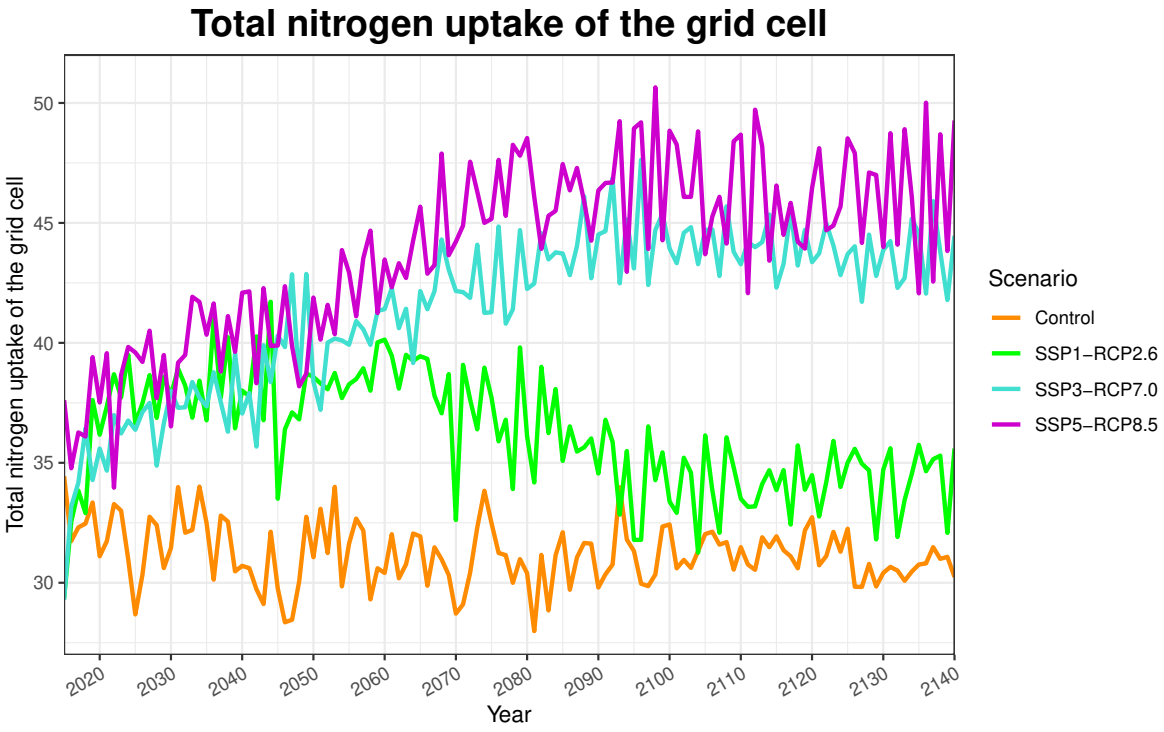


Figure 6: Yearly mean, minimum and maximum temperature and precipitation averaged over all disturbed grid cells for each scenario.



(a)



(b)

Figure 7: Nitrogen uptake per PFT and scenario averaged over all disturbed grid cells (a) and total nitrogen uptake of the grid cell for each scenario averaged over all disturbed grid cells (b).

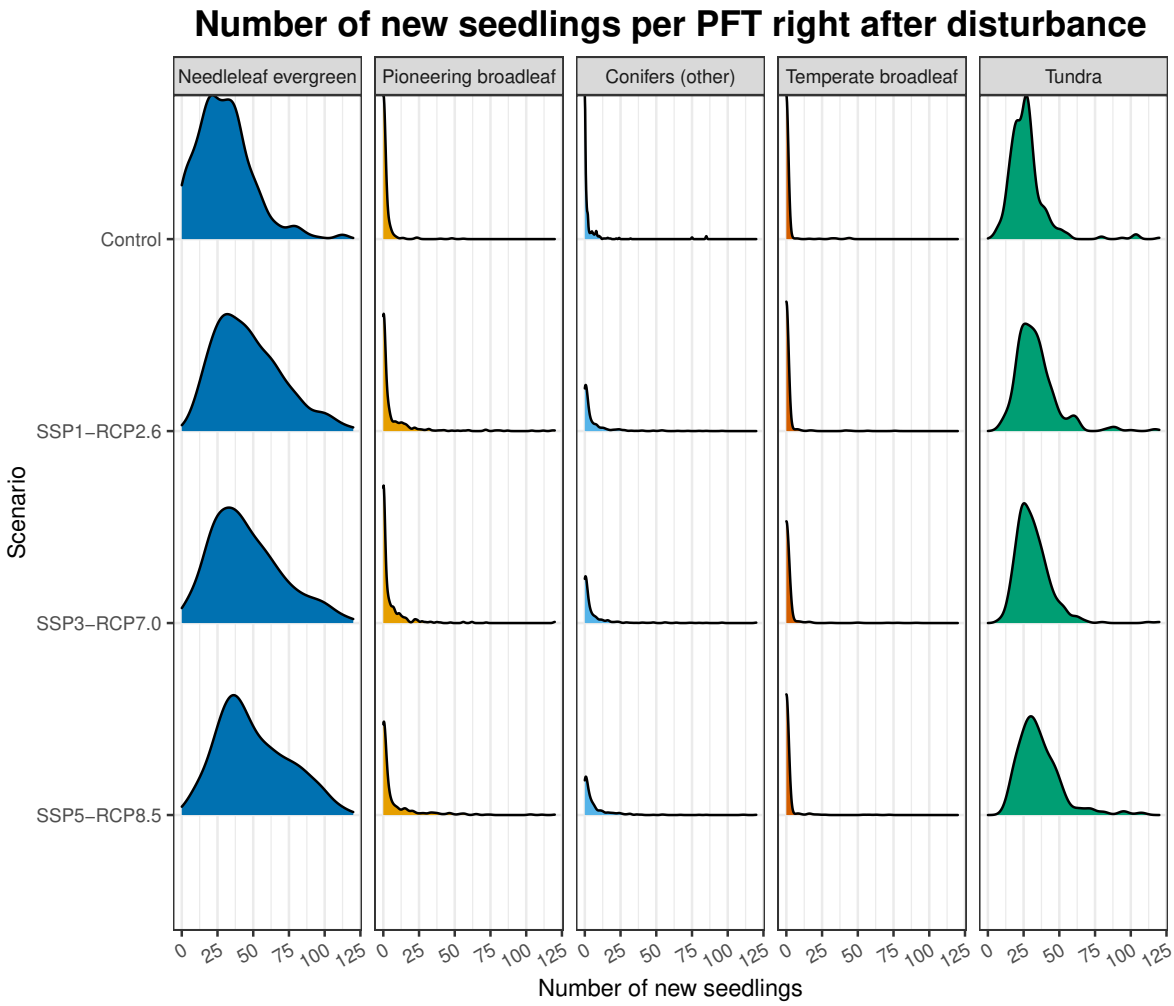


Figure 8: Number of new seedlings immediately after the disturbance.

in uptake in the control scenario, with the highest values for the PFTs *tundra* and *needleleaf evergreen*. In contrast, the warming scenarios show nitrogen dynamics that become more established with increasing radiative forcing. While *tundra* and *needleleaf evergreen* still dominate nitrogen uptake in the early decades of the study period, *pioneering broadleaf* takes the lead in all three scenarios afterwards. In addition, the uptake for *temperate broadleaf* increases in the last decades of the period for the two extreme scenarios SSP3-RCP7.0 and SSP5-RCP8.5. Note that the nitrogen uptake of the PFT *conifers (other)* remains almost constant in all four scenarios. Interestingly, when looking at Figure 7b, the general pattern is very similar to that of the temperature curves in Figure 6. While the control scenario remains at a constant level throughout the study period, nitrogen uptake in the three warming scenarios first increases in the early decades and then decreases before reaching a certain level in the SSP1-RCP2.6 scenario. The values for the two more extreme scenarios continue to increase until 2100 and then remain constant. This increase in total nitrogen uptake is also reflected in Figure 7a, taking into account the increase in total values across all scenarios.

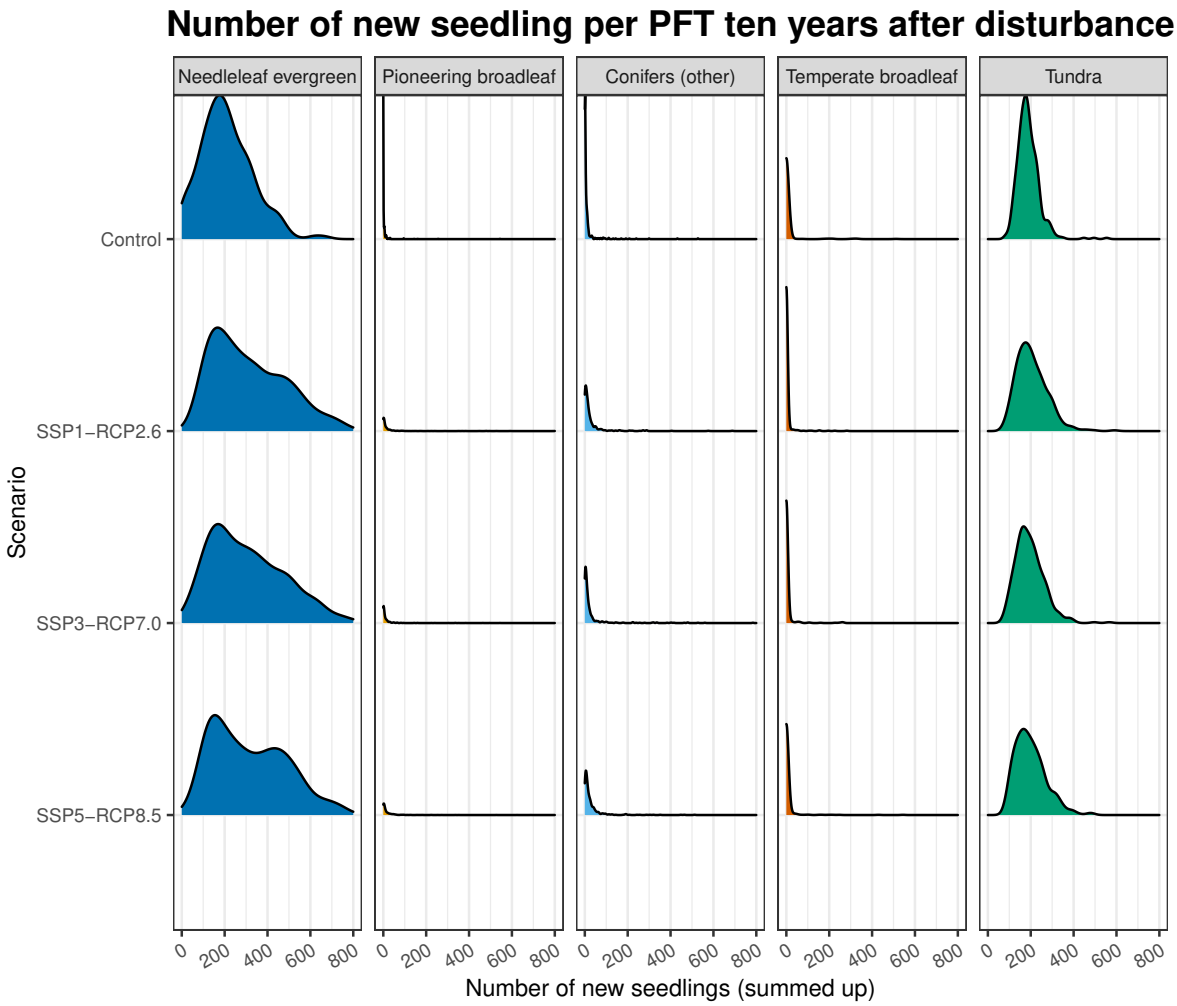


Figure 9: Number of new seedlings in the ten years after the disturbance (summed up).

The three remaining ecological variables are depicted in Figure 8, Figure 9 and Figure 10, respectively. The number of new seedlings per PFT immediately after disturbance, i.e., the initial recruitment realised as Poisson process, visualized as ridge plot in Figure 8 reveals that *needleleaf evergreen* and *tundra* are dominant, which remains true for ten years after the disturbance (Figure 9). While the number of expected seedlings of PFTs *pioneering broadleaf* and *conifers (other)* is of minor importance directly after the disturbance, their numbers of new seedlings almost disappear after ten years. It is noteworthy that there are only minor differences between the three warming scenarios, while the control scenario seems to achieve a lower number of new seedlings in general. Looking at the vegetation composition before the disturbance, transformed into relative proportions shown in Figure 10, highlights the dominance of *needleleaf evergreen*. In all four scenarios, its share is the largest of all the PFTs, but it also shows the most variation in the data. Again, there are no major differences between the scenarios. Note that even in the most extreme scenario SSP5-RCP8.5 the PFT *temperate broadleaf* does not play a major role in the composition.

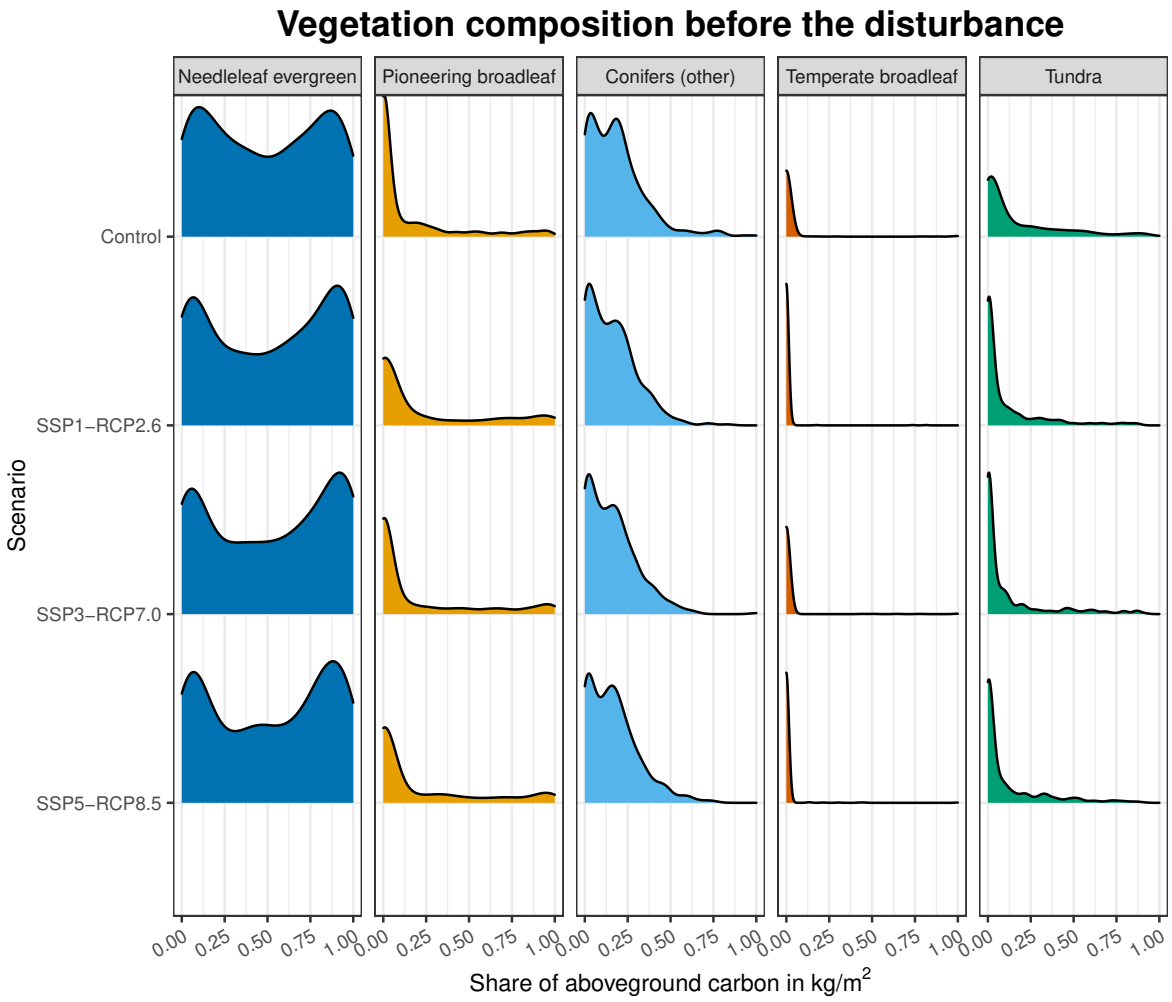


Figure 10: Vegetation composition prior to the disturbance.

A description of the spatial distribution of selected soil, climate and ecological variables is derived in Appendix A.1.

To get a first impression of recovery curves, Figure 11 shows the recovery trajectories for all four scenarios. The bold curve corresponds to the PFT-wise mean. Note that these curves do not result from a basis representation, but the interpolated individual data points. The PFT composition clearly follows similar patterns. In all scenarios, *tundra* dominates in the majority of grid cells in the first years after disturbance. After a short peak, its share decreases and other PFTs take over. Especially the dominant vegetation after 100 years differs between the scenarios. The more extreme the increase in radiative forcing, the more dominant *pioneering broadleaf* becomes, while the importance of *needleleaf evergreen* decreases. PFT *temperate broadleaf* becomes more present in the more extreme scenarios, but is not able to displace needleleafed trees and *pioneering broadleaf*. In addition, Figure 12 shows the mean differences between the warming scenarios and the control scenario, i.e., the difference between the average relative carbon values of each warming scenario and the

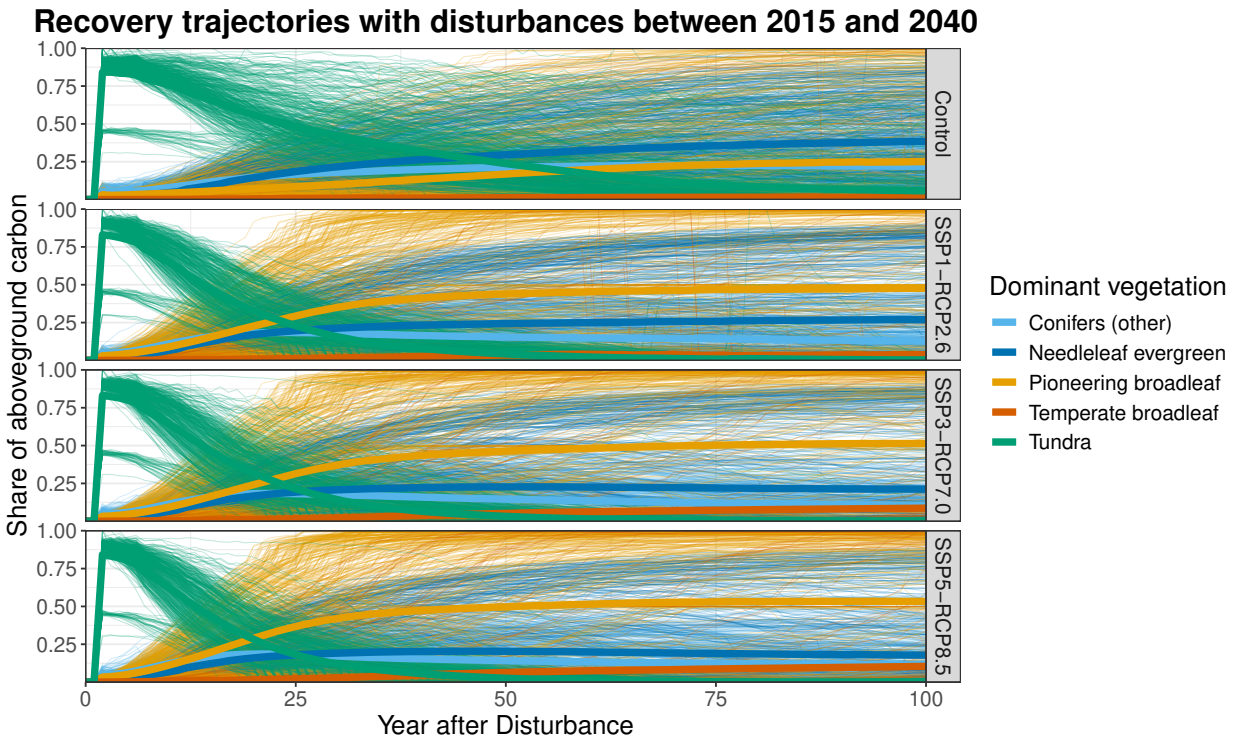


Figure 11: Recovery trajectories for all PFTs and scenarios for patches disturbed between 2015 and 2040.

corresponding values of the control scenario. The figure reveals a rather uniform behaviour with more *pioneering broadleaf* and *temperate broadleaf* than in the control scenario on the one hand, and a smaller proportion of the remaining PFTs on the other one after a few decades of recovery.

4.2 Basis Function Representation

As introduced in Section 2.1.1, the first step in FDA is to find an appropriate basis function expansion for the functional data at hand. Since the recovery trajectories are non-periodic, a B-spline basis with regularization is chosen. Exploring different combinations of parameter settings for the smoothing parameter λ, the degree k and number of basis functions K, as well as which derivative to penalize, yields the following final setup: $\lambda = 1$, $k = 5$ and K is equal to the number of disturbed patches, while the third derivative is penalized. In the absence of a suitable fitting criterion for the data at hand, the overall fit to the data and the corresponding hyperparameters were selected based on visual inspection and domain knowledge.

Figure 13 shows the smoothed fit to the recovery trajectories of PFT *needleleaf evergreen* for all four scenarios and the first 100 years of recovery after a disturbance occurred. The general behavior is similar for all scenarios: after five to ten years after the disturbance, the share of aboveground carbon increases sharply and tends to remain at a high level. A closer look at the scenarios reveals some differences between the control scenario and the three warming

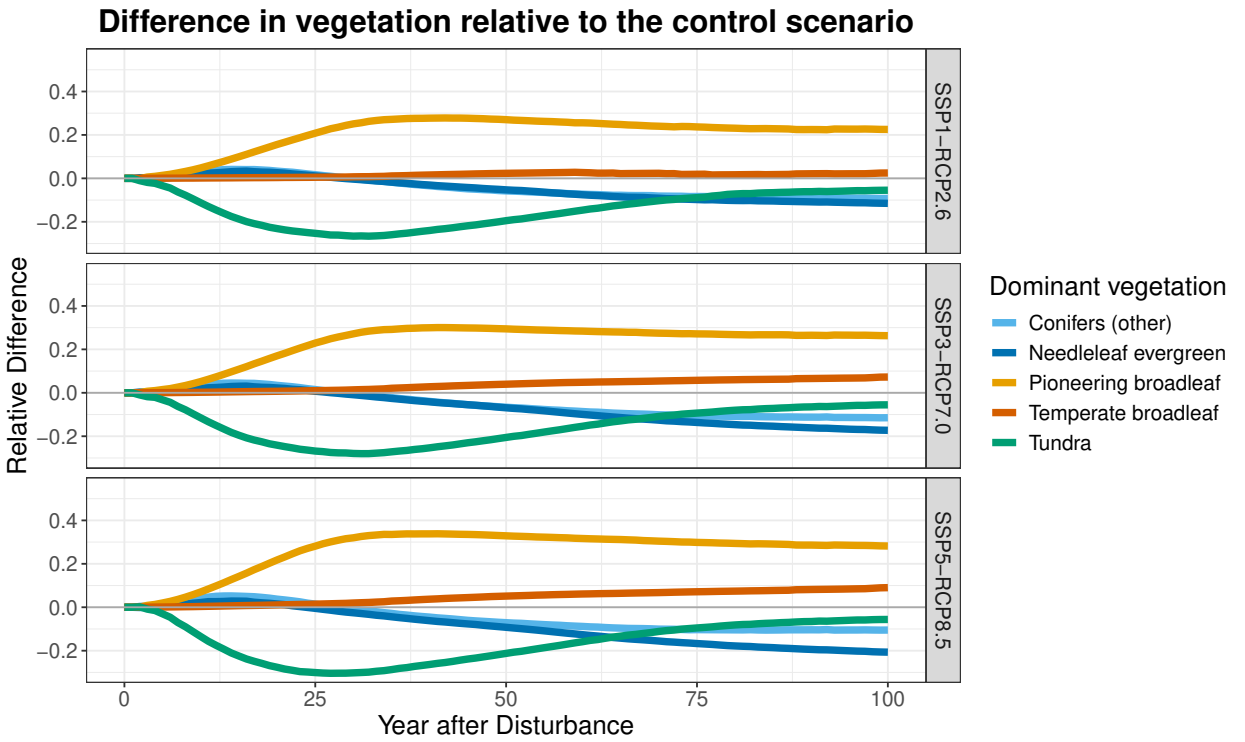


Figure 12: Differences in vegetation composition relative to the control scenario. Values above zero correspond to a higher share of the specific PFT than in the control scenario and vice versa.

scenarios. The functions for the control scenario appear to be more dispersed, while those in the scenarios reach either rather high or rather low levels of *needleleaf evergreen*.

In Figure 14, the smoothed fit for PFT *pioneering broadleaf* is displayed. The share of aboveground carbon of this PFT reaches an almost constant high level after a short period of increase only a few years after the disturbance. There are some substantial differences between the control scenario and the warming scenarios: while in the scenarios, the increase is very sharp for most of the curves, the increase in aboveground carbon is flatter and more dispersed for the control scenario. Note that similar to *needleleaf evergreen*, the warming scenarios show a very similar behaviour in contrast to the control scenario.

For PFT *conifers (others)* shown in Figure 15, the overall behaviour again follows a similar pattern for all scenarios. There is a moderate increase in share of aboveground carbon immediately after the disturbance. In the warming scenarios, most curves peak after about 20 to 40 years of recovery, with some curves increasing over the whole period. This peak is less pronounced in the control scenario. It should be noted that the total share of aboveground carbon is lower than in the two previous PFTs.

Looking at PFT *temperate broadleaf* visualized in Figure 16 reveals some major differences between the scenarios. While there are a substantial number of non-zero curves in the most drastic scenario SSP5-RCP8.5, the majority of grid cells in the control scenario have aboveground carbon shares close to zero. For those grid cells where the vegetation composition includes *temperate broadleaf*, the fraction increases in the first decades after the disturbance and reaches levels close to 100%. The three plots indicating rising radiative forcing clearly show an increase in the number of grid cells covered by *temperate broadleaf* for a more pronounced warming.

Figure 17 shows the smoothed fit to the final PFT *tundra*. Again, the three warming scenarios appear to behave similarly over the recovery period, reaching a high peak shortly after the disturbance and then declining sharply, while the curves in the control scenario show a more dispersed behaviour in the years after the peak. Note that all four simulation runs include grid cells with a substantially lower peak than the rest of the locations. Further investigation did not reveal any spatial patterns in this peak behaviour.

4.3 Univariate FPCA

Now that an appropriate basis function representation is available, the functions are eligible for further analysis. To reduce dimensionality and find patterns within the data, a univariate FPCA is performed on each scenario and PFT separately. This yields 20 different FPCA results. Recall that the four simulation runs representing the four scenarios are somewhat independent runs of LPJ-GUESS. In particular, this means that some patches are disturbed in more than one scenario.

Considering only the first two PCs accounts for more than 90% of the variability for all runs (except one) of the FPCA, as Table 5 indicates. The percentages imply that especially for

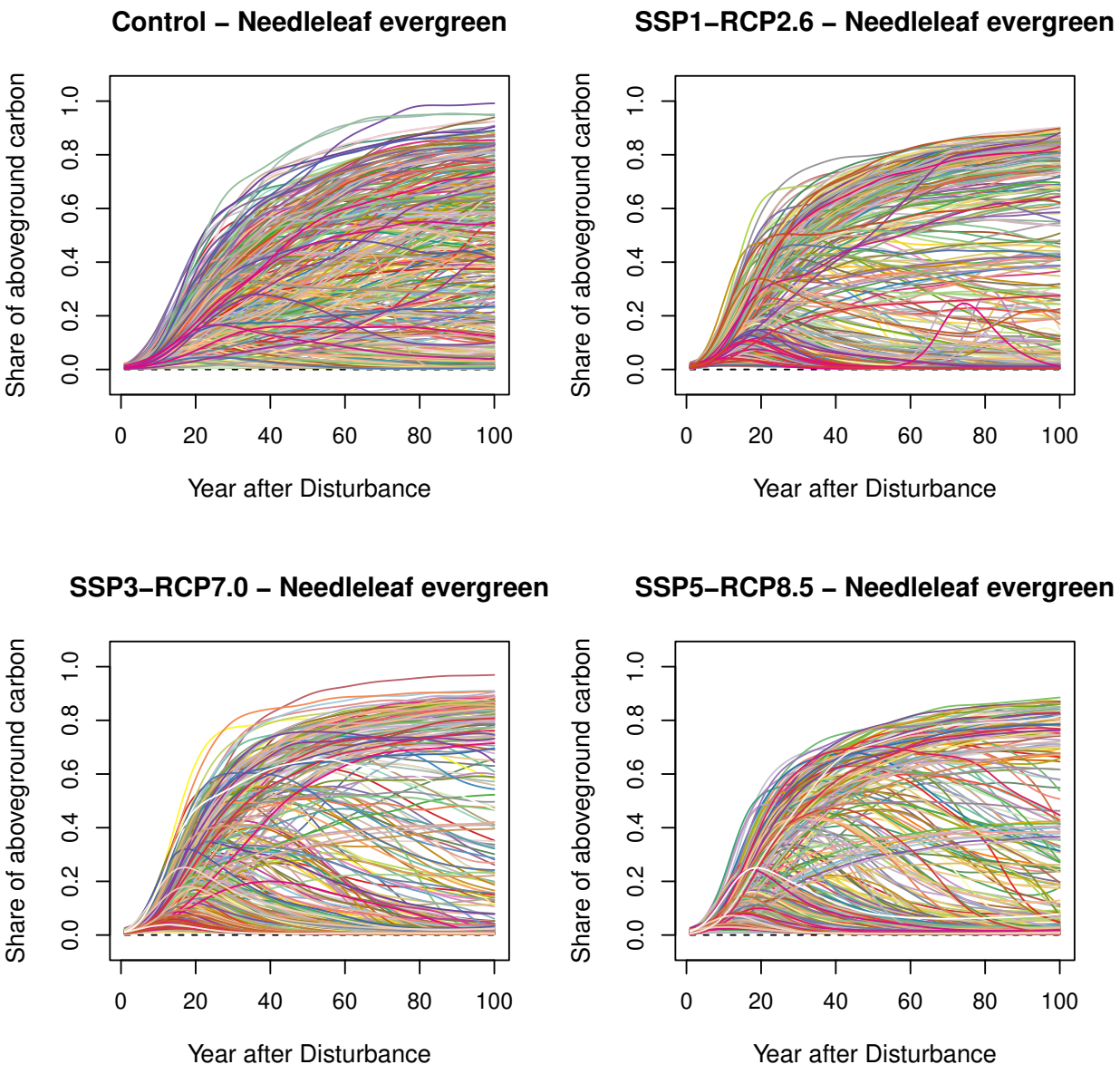


Figure 13: Basis function representation for PFT *needleleaf evergreen*.

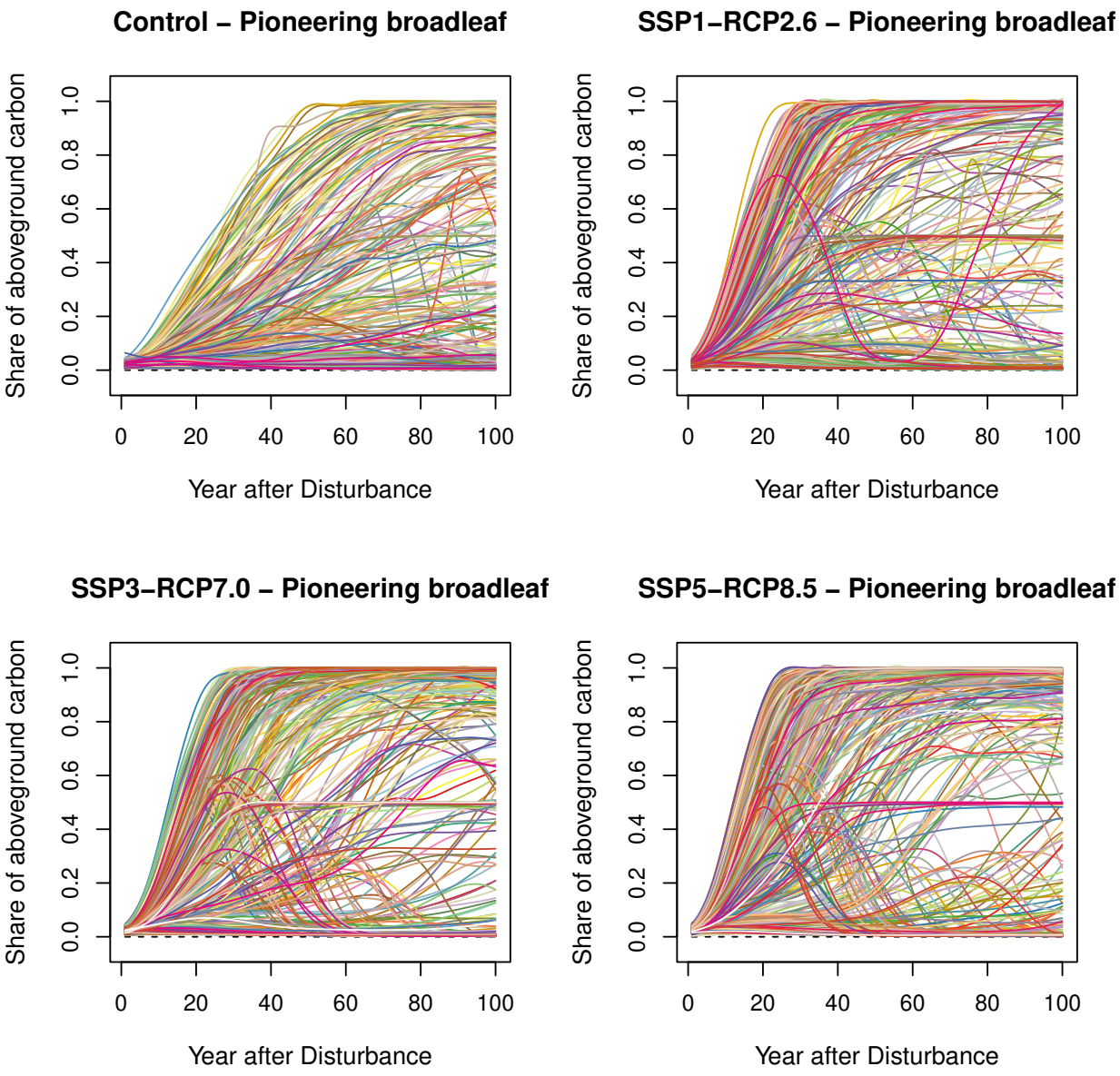


Figure 14: Basis function representation for PFT *pioneering broadleaf*.

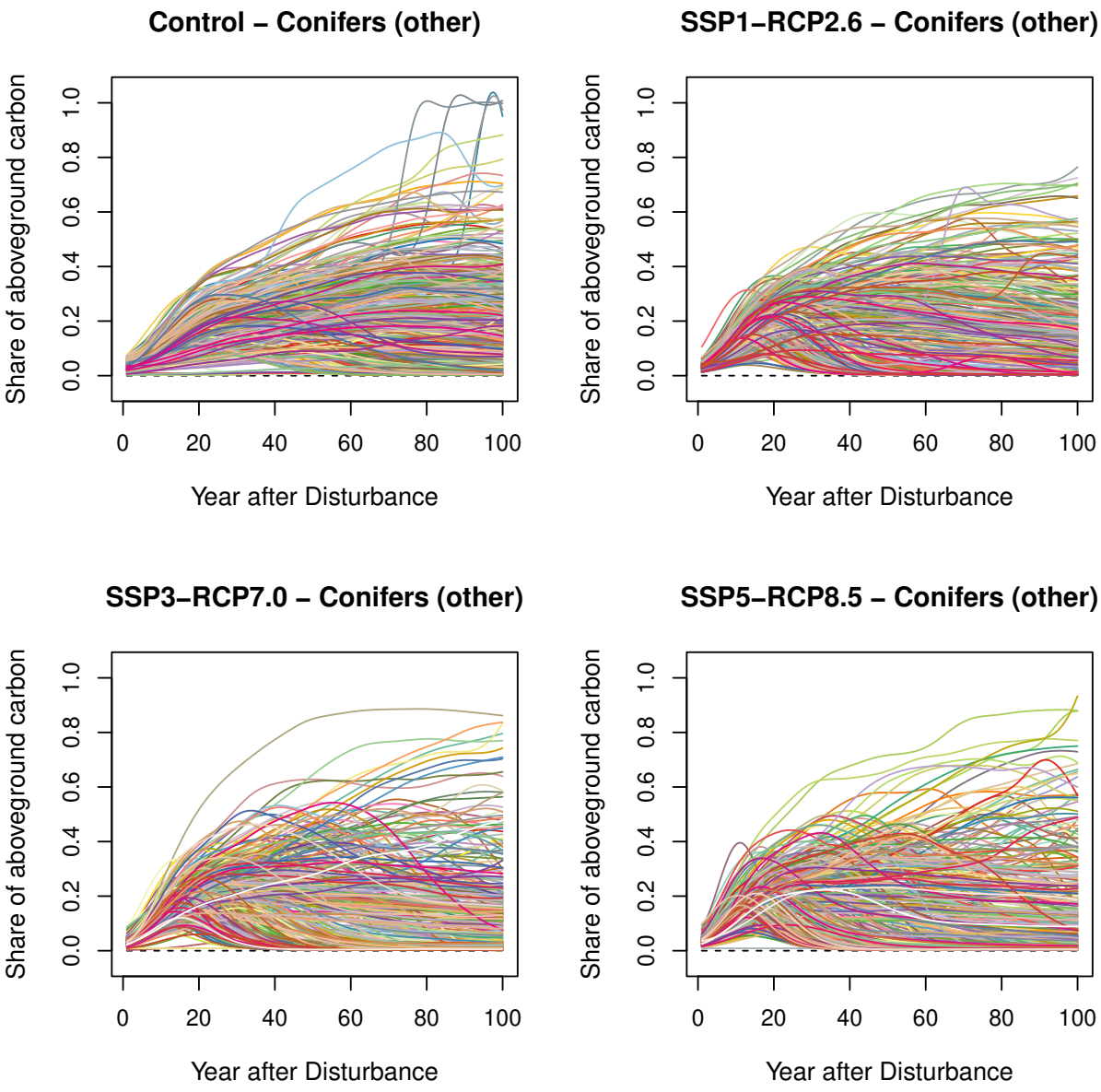


Figure 15: Basis function representation for PFT *conifers (others)*.

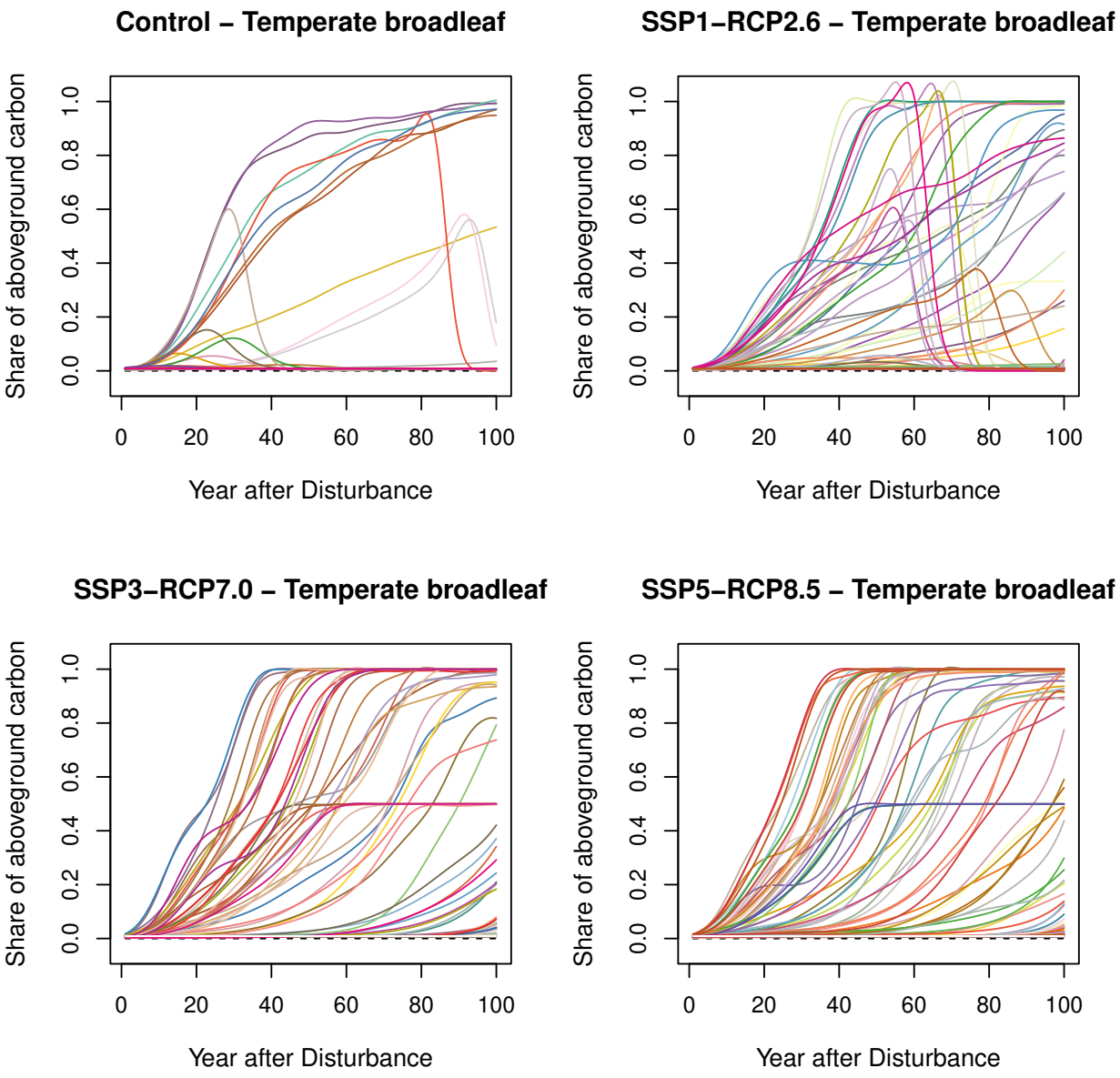


Figure 16: Basis function representation for PFT *temperate broadleaf*.

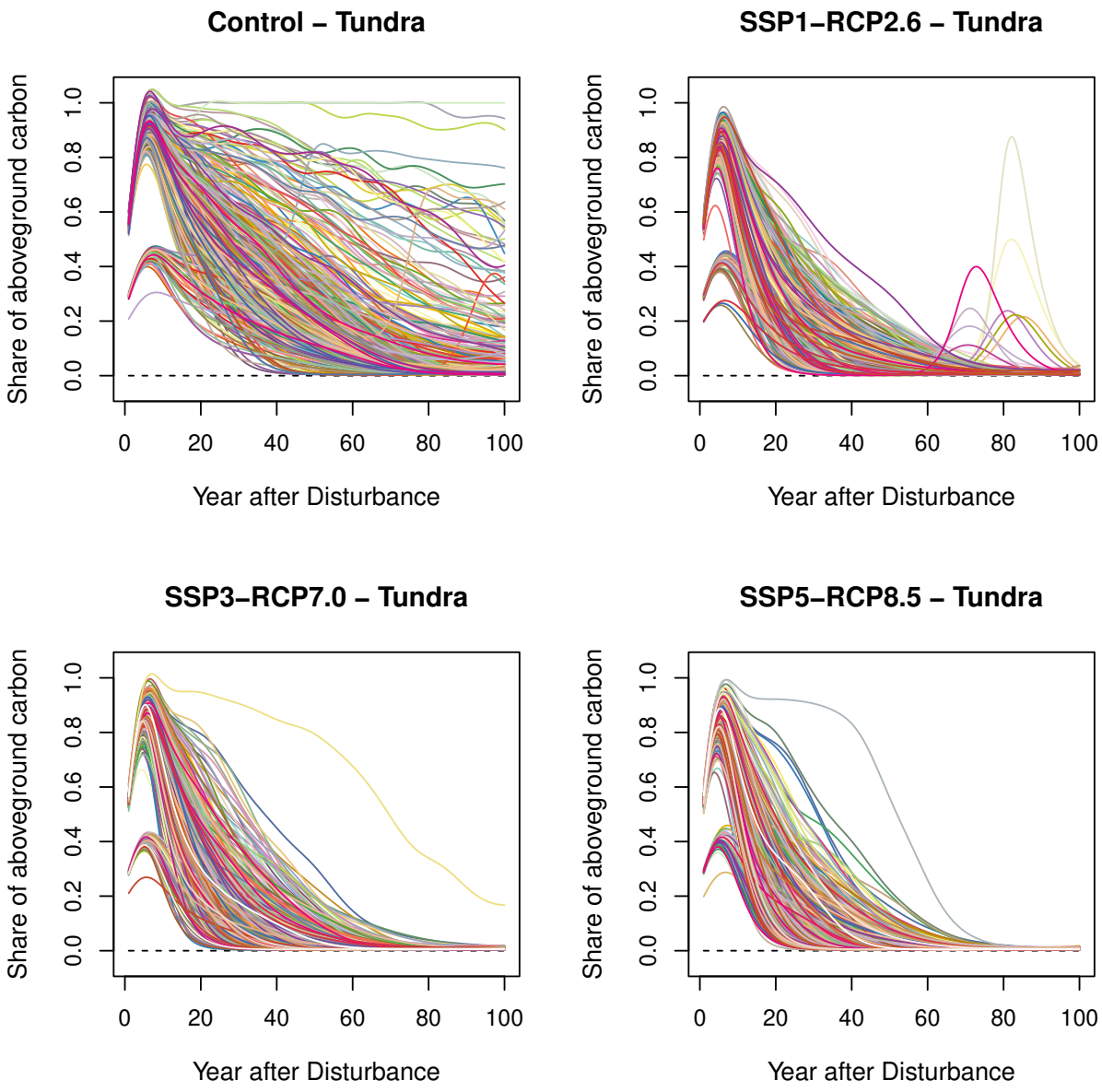


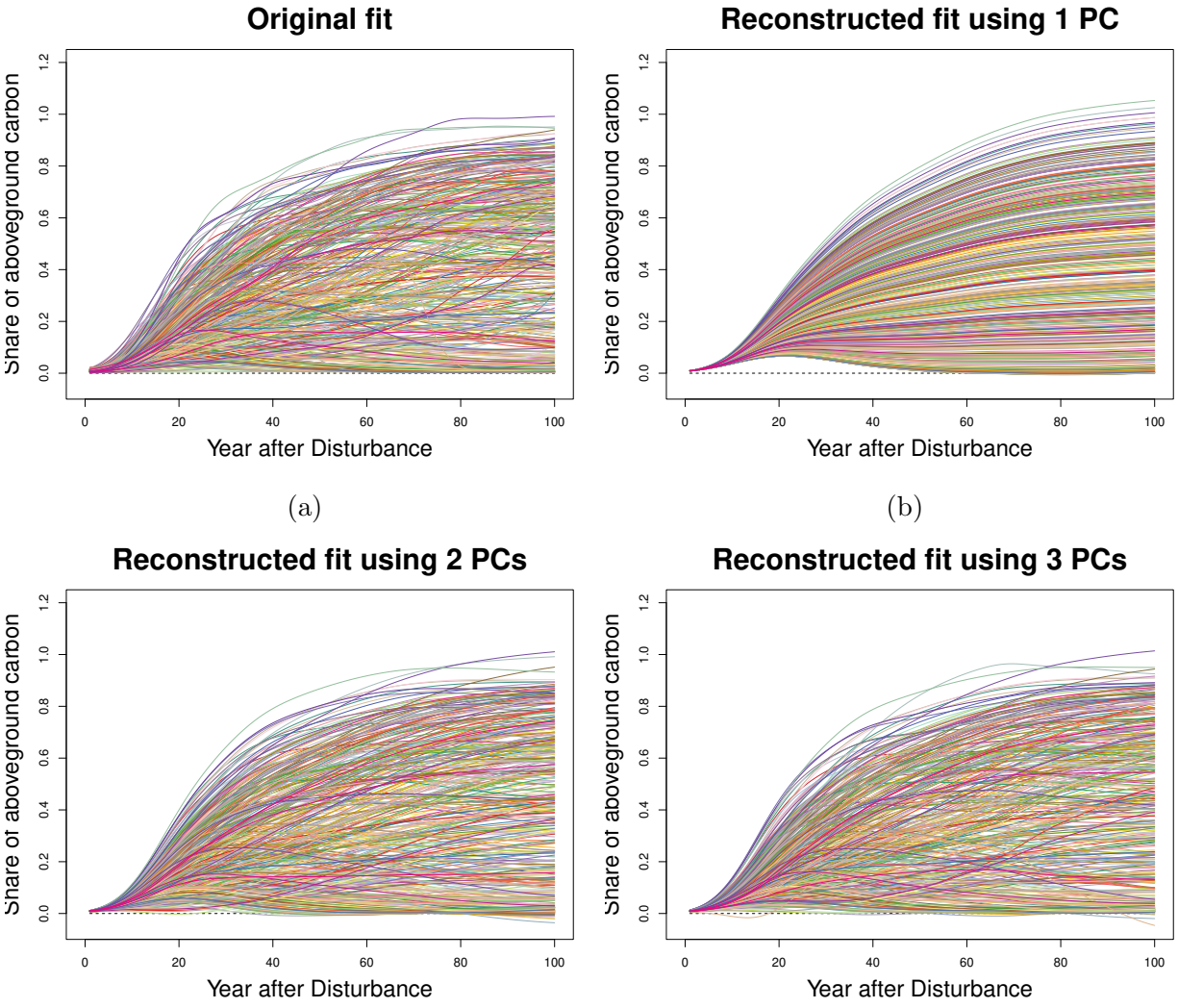
Figure 17: Basis function representation for PFT *tundra*.

		PC 1	PC 2	Sum
Control	**BNE**	94.7	4.7	99.4
	IBS	97.3	2.2	99.5
	otherC	77.8	11.5	89.3
	TeBS	90.4	4.7	95.1
	Tundra	79.7	11.7	91.4
SSP1-RCP2.6	**BNE**	97.4	2.2	99.6
	IBS	96.2	2.6	98.8
	otherC	86.1	6.6	92.3
	TeBS	76.6	14.6	91.2
	Tundra	63.2	28.3	91.5
SSP3-RCP7.0	**BNE**	95.6	3.9	99.5
	IBS	93.7	5.2	98.9
	otherC	81.7	9.7	91.4
	TeBS	86.5	9.4	95.9
	Tundra	70	23.2	93.2
SSP5-RCP8.5	**BNE**	94	5.1	99.1
	IBS	93	5.7	98.7
	otherC	81	10.3	91.3
	TeBS	88.1	6.8	94.9
	Tundra	72.1	22.4	94.9

Table 5: The variability in per cent that is accounted for by each PC for each FPCA that is performed.

the PFT *tundra* it may be beneficial to include the second PC, as the first PC only accounts for 60% to 80%, depending on the climate scenario. Later analyses will reveal the importance of the PFTs *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*. Interestingly, for those PFTs, the first PC accounts for at least 80% of the variability (with one exception), indicating that one component already captures most of the variation in the data. As an example, Figure 18a shows the original fitted curves next to reconstructed curves using one (Figure 18b), two (Figure 18c) and three PCs (Figure 18d), respectively. The reconstructions show that the inclusion of the second PC contributes to a better representation of the behaviour at the end of the study period and accounts for more variation in the data, while the third PC does not add substantial information. Thus, in the following analyses, the first two PCs are considered.

To further explore the relationship between the PFTs for each scenario, Figure 19 shows the correlation between the PC scores for each PFT broken down by scenario. Note that the sign of the PC scores is inherently arbitrary because changing the sign of a PC function simultaneously changes the sign of its corresponding scores. Therefore, when comparing PC scores from different FPCAs, the sign of the correlation can be influenced by the arbitrary sign assignment of the PC in each analysis. As a result, the focus lies on the absolute values in the correlation plots. These correlations give a first hint at the compositional drivers of



(a) (b) (c) (d)

Figure 18: Original fitted curves using a 6-order B-spline basis (a) and reconstructed curves using one PC (b), two PCs (c) and 3 PCs (d) of the univariate FPCA for the control scenario and PFT *needleleaf evergreen*.

vegetation after disturbance in the boreal biome under varying climate. The first PCs of *needleleaf evergreen* and *pioneering broadleaf* are consistently strongly correlated across all scenarios, indicating a robust relationship under different conditions. In the control scenario, there is a strong connection between the second PCs of *needleleaf evergreen* and *conifers (others)*, which vanishes in the three warming scenarios. In contrast, the first PCs of *needleleaf evergreen* and *tundra* show a high correlation in the three warming scenarios, though being close to zero in the control scenario. The same applies to the combinations between the first PC of *conifers (others)* and of *tundra, needleleaf evergreen* and *pioneering broadleaf*. Overall, the warming scenarios show many similarities from which the control scenario deviates, and the relationships tend to be stronger for more extreme climate.

The harmonics resulting from the FPCAs can be examined in more detail. Figure 20 shows the mean shares of aboveground carbon for PFT *needleleaf evergreen* over the recovery period of 100 years including the effects of adding (+) and subtracting (−) two standard deviations of each PC curve. The PCs for all four scenarios are similar and all reflect the same pattern: high values in the first PC indicate much higher aboveground carbon levels than on average after the first 10 years of recovery, and vice versa for low values. The second PC mainly reflects peaks in the first decades after disturbance, where high values indicate a sharper increase in *needleleaf evergreen* than on average, as well as a faster decline. The turning point varies hardly between the scenarios.

The equivalent plot for PFT *pioneering broadleaf* is portrayed in Figure 21. Since the overall functional fit to the data is close to that of *needleleaf evergreen* (compare Figure 13 and Figure 14), the behaviour of the first two PCs is similar as well. High values in the first PC indicate an above average proportion of aboveground carbon after the first few years after disturbance, and vice versa for low values. The second PC mainly reflects the peak of *pioneering broadleaf* after a few decades of recovery, where high values indicate a stronger increase and decrease than the mean. Note that the moment of change between increase and decrease differs between the warming scenarios and the control scenario.

Figure 22 shows similar dynamics for PFT *conifers (others)*. The first PC accounts for variation in the dynamics of recovery starting 10 years after disturbance, where high values indicate an above average share of aboveground carbon, and vice versa for low values. The second PC focuses on the peaking behaviour after a few decades of recovery. In contrast to the PFTs considered before, in the three warming scenarios high values represent a higher peak and a faster decrease than the mean, while low values indicate a small peak in the beginning of the time frame, a small decline and than a huge increase in *conifers (others)* relative to the mean. The behavior of the control scenario is less variable: low values of the second PC represent a higher share of aboveground carbon than the mean until about 90 years post-disturbance, and vice versa for high values.

Recall that Figure 16 revealed a lack of data for PFT *temperate broadleaf* due to the dominance of other PFTs in most of the grid cells, especially in those affected by colder scenarios, i.e., the control scenario and SSP1-RCP2.6. Nevertheless, the harmonics of the corresponding FPCAs in Figure 23 show a consistent behaviour and are therefore less prone to misinterpre-

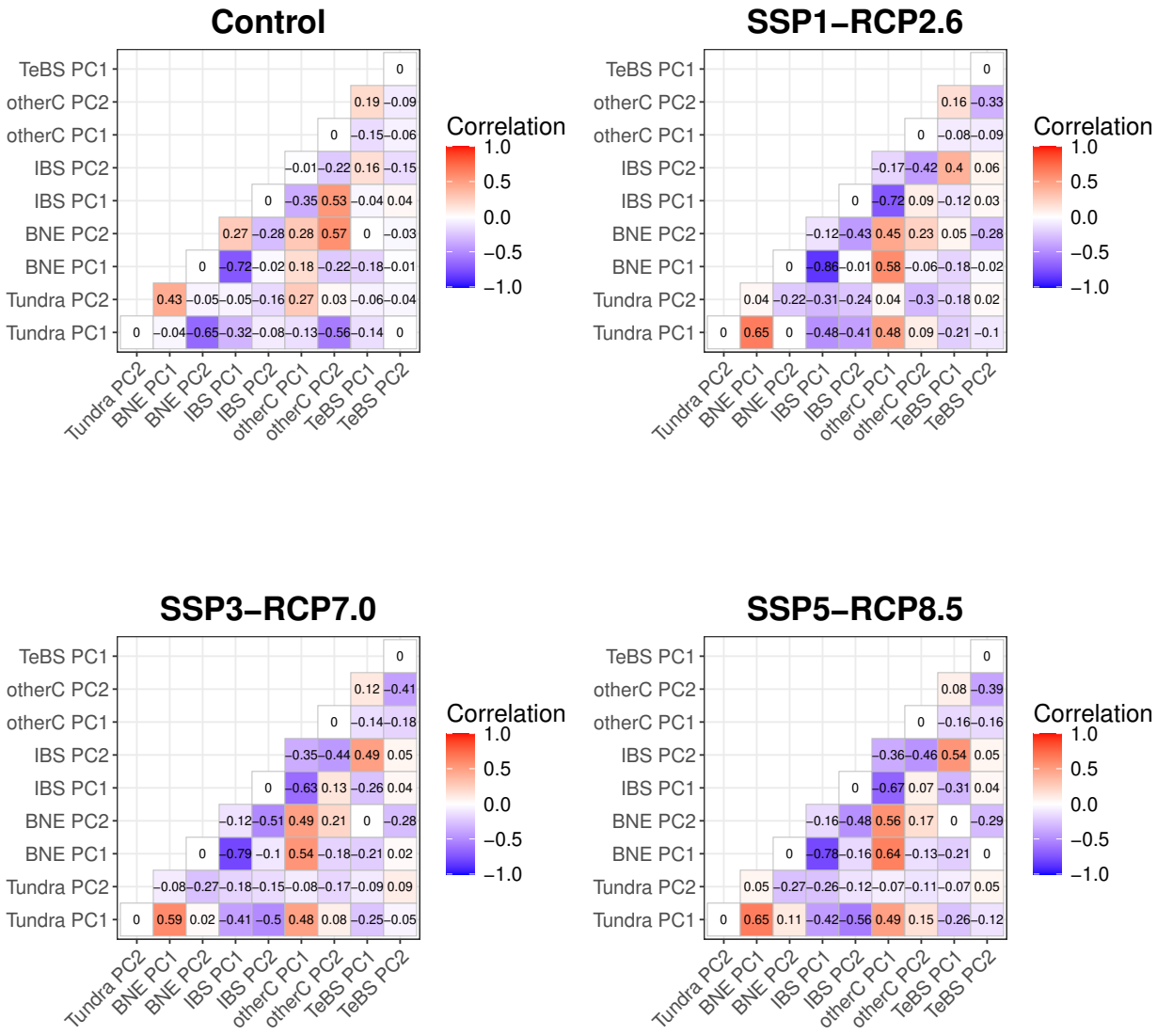


Figure 19: Correlations between the first two PCs derived from separate univariate FPCAs for each scenario and PFT.

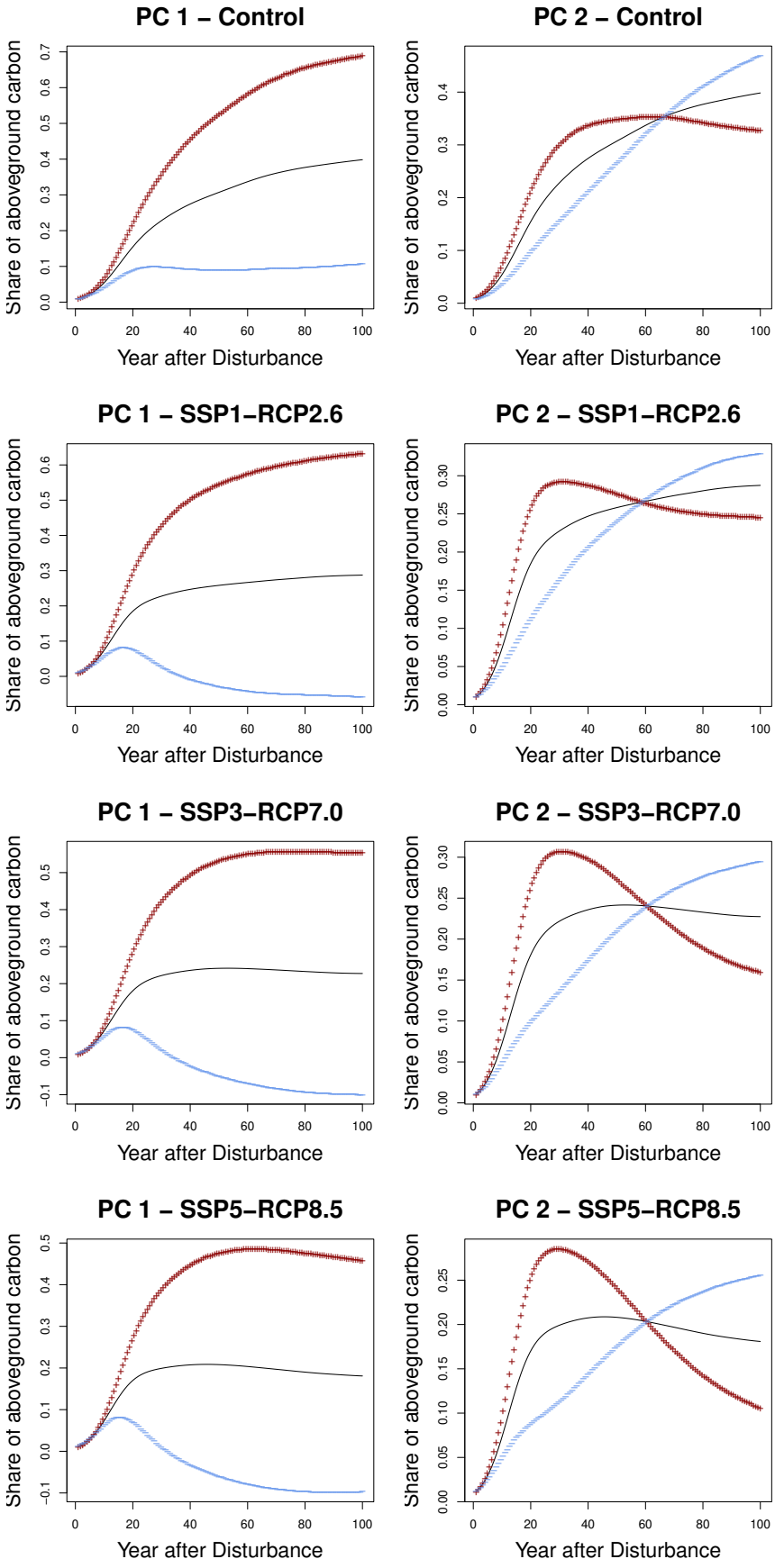


Figure 20: First two PCs for each scenario for PFT *needleleaf evergreen*.

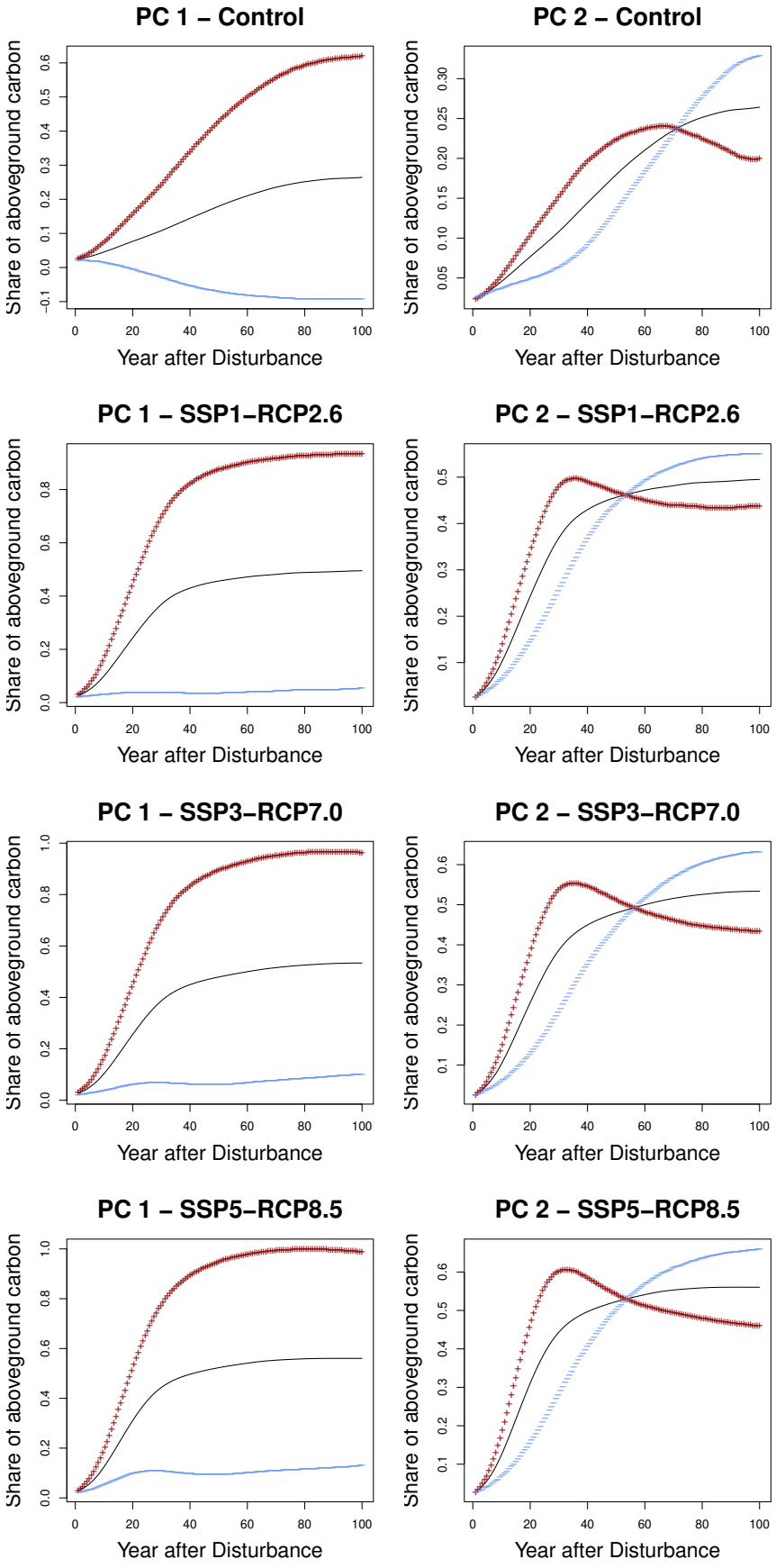


Figure 21: First two PCs for each scenario for PFT *pioneering broadleaf*.

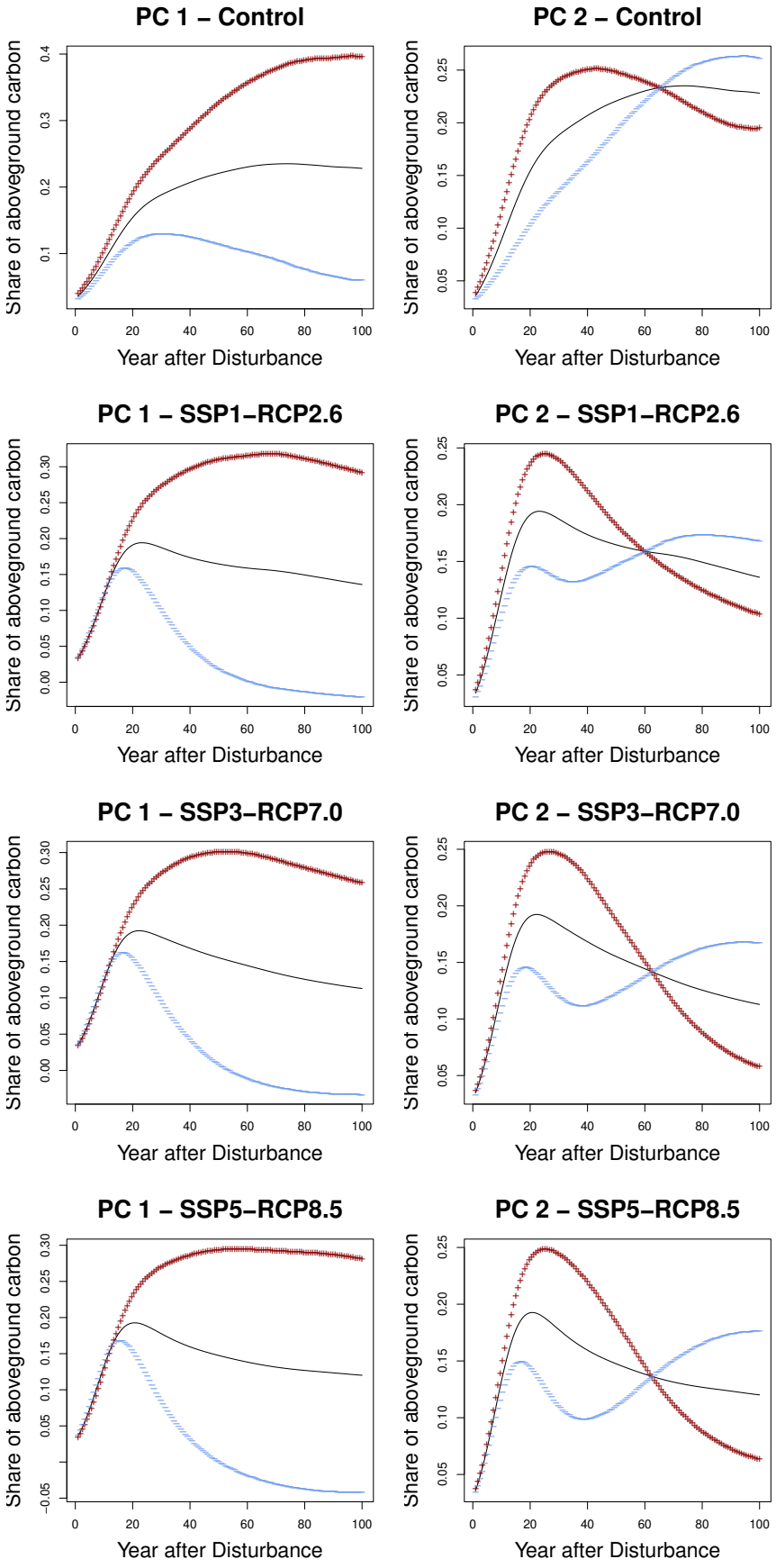


Figure 22: First two PCs for each scenario for PFT *conifers (others)*.

tation. The first PC is very similar for all four scenarios and mainly reflects deviations to the mean function starting after around 10 years of recovery. High values indicate a higher share of aboveground carbon, while low values represent below-average portions. High values in the second PC indicate higher shares than on average for the first 70 years after disturbance in scenario SSP1-RCP2.6, with a peak around 50 years. This behaviour is vice-versa for low values. For the remaining scenarios the dynamics are flipped: high values indicate a lower share than on average with a changing point after about 80 years post-disturbance for scenarios SSP3-RCP7.0 and SSP5-RCP8.5, and about 85 years for the control scenario, and lower shares for the remaining study period, and vice versa for lower values of PC 2.

Finally, Figure 24 shows the first two harmonics for PFT *tundra*. Here major differences between the control scenario and the three warming scenarios become apparent. While high values in the first PC reflect a higher peak in aboveground carbon than the mean, lower values represent a lower peak and a more pronounced decline in the 20-40 years after the disturbance. The decreasing behaviour is very different between the warming scenarios and the control scenario: while in the control scenario this decrease is relatively slow and flat, the three warming scenarios show a rapid decrease in *tundra*. The second PC displays the size of the peak, but focuses on a change in dynamics around 20 years after the disturbance for the warming scenarios and around 55 years for the control scenario. Here, high values indicate a higher peak for both the control and SSP3-RCP8.5 scenario, while for the other two the interpretation is flipped.

To summarize, the harmonics for all PFTs follow similar dynamics with the details shifting in time depending on the scenario. The small amount of variation that the second PC is accounting for (see again Table 5), is visualized by a limited spread in the second PC, reflecting minimal additional information. As already indicated by the correlations in Figure 19, the warming scenarios show similar behaviour, while the control scenario deviates in its dynamics. Note that a VARIMAX rotation was examined, but it was found not to improve the interpretability of the harmonics.

In order to dive deeper into the FPCAs conducted in this chapter and PFTs influencing the vegetation composition, subsection A.2 comprises details on the results obtained by a scenario-wise clustering of PC scores.

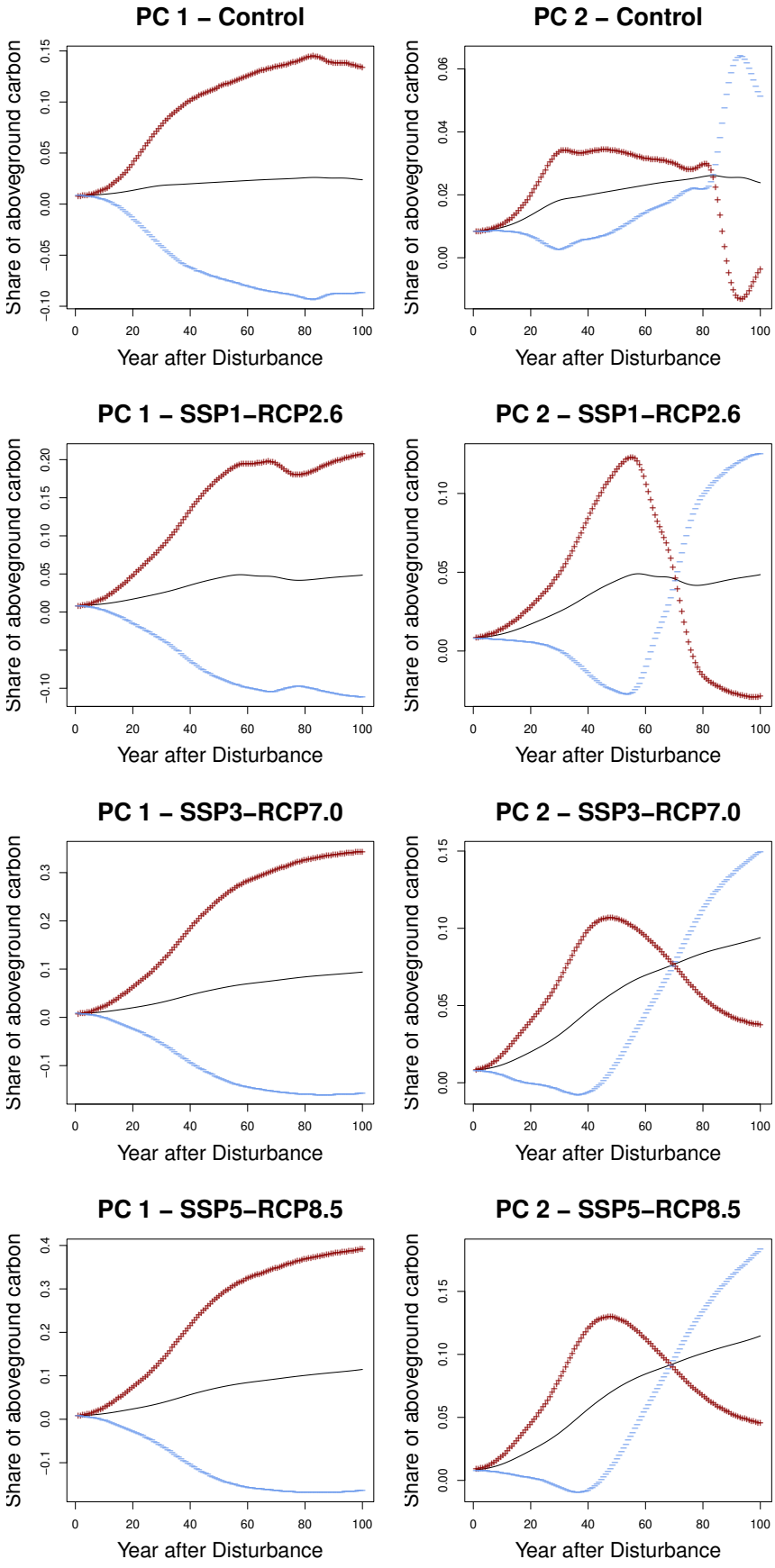


Figure 23: First two PCs for each scenario for PFT *temperate broadleaf*.

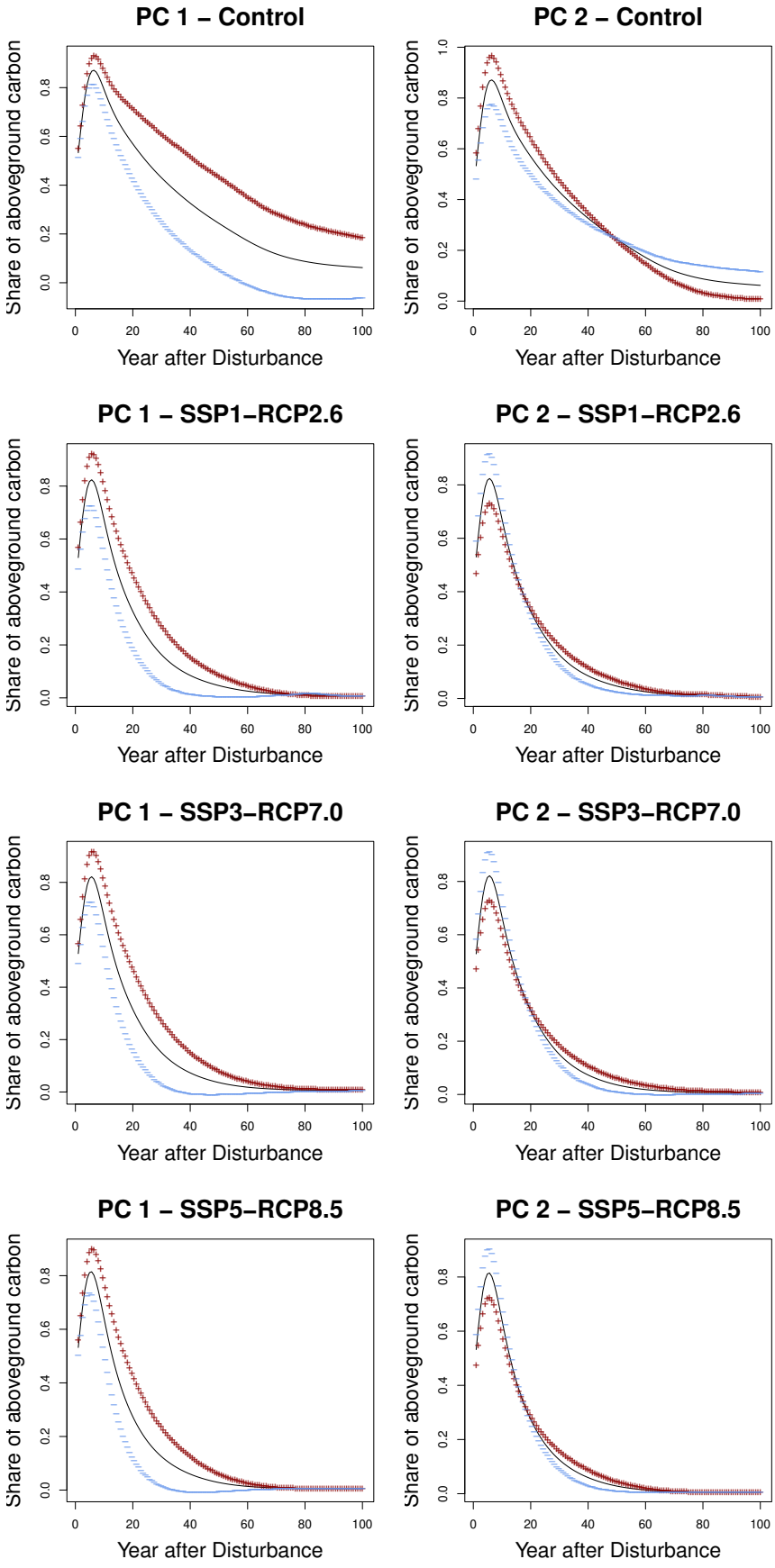


Figure 24: First two PCs for each scenario for PFT *tundra*.

	Variability	Cumulative Sum
PC 1	69.99	69.99
PC 2	13.21	83.20
PC 3	5.93	89.13
PC 4	4.50	93.63
PC 5	2.19	95.82
PC 6	1.64	97.46
PC 7	0.81	98.27
PC 8	0.53	98.80
PC 9	0.34	99.14
PC 10	0.28	99.42

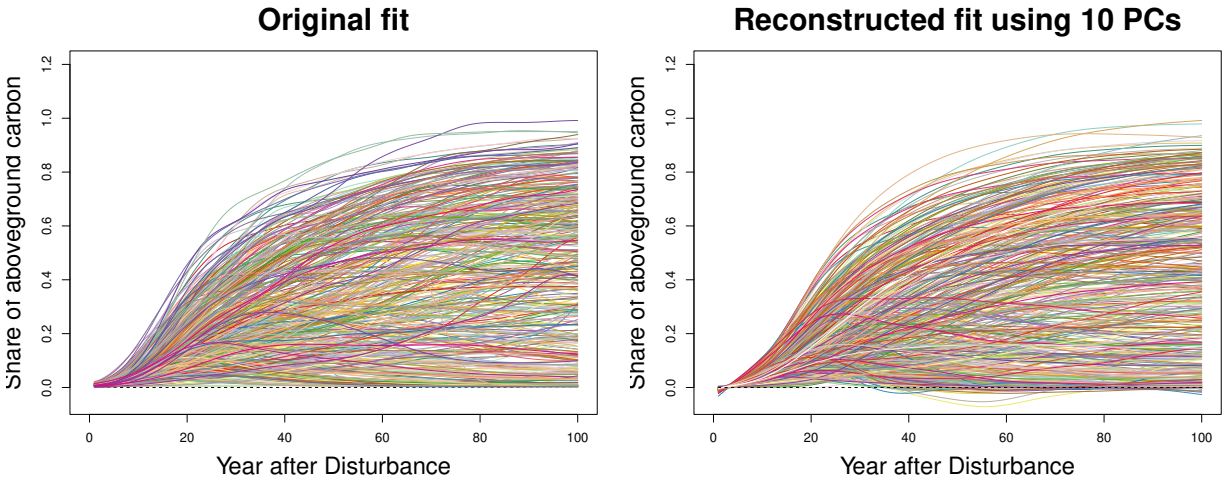
Table 6: The variability in per cent and its cumulative sum that is accounted for by each PC derived by a MFPCA.

4.4 MFPCA

The univariate FPCA approach, when applied to each scenario and PFT separately, is subject to two major limitations. Firstly, as the FPCAs are derived independently, this approach does not account for the covariation between the PFTs and scenarios. Secondly, the overall aim is to model the aboveground carbon proportions of all five PFTs and four scenarios simultaneously in order to draw conclusions about the vegetation composition as a whole. Therefore, in order to consider the multivariate data structure at hand, an MFPCA with ten PCs is conducted for the 1803 disturbed grid cells using the R-package `MFPCA`.

Table 6 shows the variability accounted for by the first ten PCs and the corresponding cumulative sum. As in the univariate setting, the first PC represents the majority of the variability and the subsequent PCs add only small amounts of knowledge. This leads to a drastic reduction in dimensionality if only the first few PCs are used to represent the data. Note that as opposed to the univariate FPCAs, four PCs are required here to achieve more than 90% of variation. Figure 25 shows an example of the reconstruction of the original curves (a) with the first ten PCs (b) for the control scenario and PFT *needleleaf evergreen*. The general behaviour is maintained by the reconstruction, and the dimensionality reduction leads to a removal of noise in the data.

To provide a detailed description of the PCs, the following figures show the first ten PCs for each of the five PFTs. Since an MFPCA is performed on the entire data set, the PCs are comparable between the PFTs and can be interpreted in relation to each other. Figure 26 highlights the dominant patterns and variability captured by the first PC. In black the mean share of aboveground carbon for the respective PFT is portrayed for the whole recovery period of 100 years. The red (+) and blue (−) lines represent the addition and subtraction of two standard deviations to the PC curve, respectively, showing the variability around the mean trend. While high values in the first PC indicate a higher than average proportion of *needleleaf evergreen, conifers (other)* and *tundra*, the second plot shows a lower than average



(a) (b)

Figure 25: Original fitted curves using a 6-order B-spline basis (a) and reconstructed curves using ten PCs (b) of the MFPCA for the control scenario and PFT *needleleaf evergreen*.

proportion of *pioneering broadleaf*, and vice versa for low PC values.

The PFT *temperate broadleaf* is hardly captured by the first PC, but dominates the behaviour of the second PC, as shown in Figure 27. Here, high PC values are indicative of lower proportions of *temperate broadleaf* starting a few years after the disturbance, and vice versa for low values. The effects on the other PFTs are less pronounced. While there is hardly any effect for *tundra*, higher PC values indicate higher proportions of *needleleaf evergreen, pioneering broadleaf* and *conifers (other)* after about 30 years of recovery, and vice versa for low values. Note that the deviations from the mean for these three PFTs are much smaller than for the first PC.

Figure 28 shows the third PC, which again reflects only minor deviations from the mean with respect to the magnitude of the proportion of aboveground carbon. High values of this PC indicate above average proportions of *needleleaf evergreen* and *temperate broadleaf*, while the proportions of *conifers (other)* and *tundra* are below average after a few decades of recovery. The effect on *pioneering broadleaf* captures a shift in behaviour after about 70 years of recovery: while high values initially indicate slightly above-average proportions of aboveground carbon, the proportions after the turning point are slightly below average. All interpretations are reversed for low values.

Similar shifts in dynamics are covered by the fourth PC displayed in Figure 29. Again, this PC reflects minor deviations from the mean proportions. The PCs of *needleleaf evergreen, pioneering broadleaf* and *temperate broadleaf* show a turning point after several decades of recovery. For both broadleafed PFTs, high PC values indicate a slightly below-average share of aboveground carbon up to the turning point, and a somewhat higher proportion thereafter. The behaviour is reversed for *needleleaf evergreen*. For *conifers (other)* high values indi-

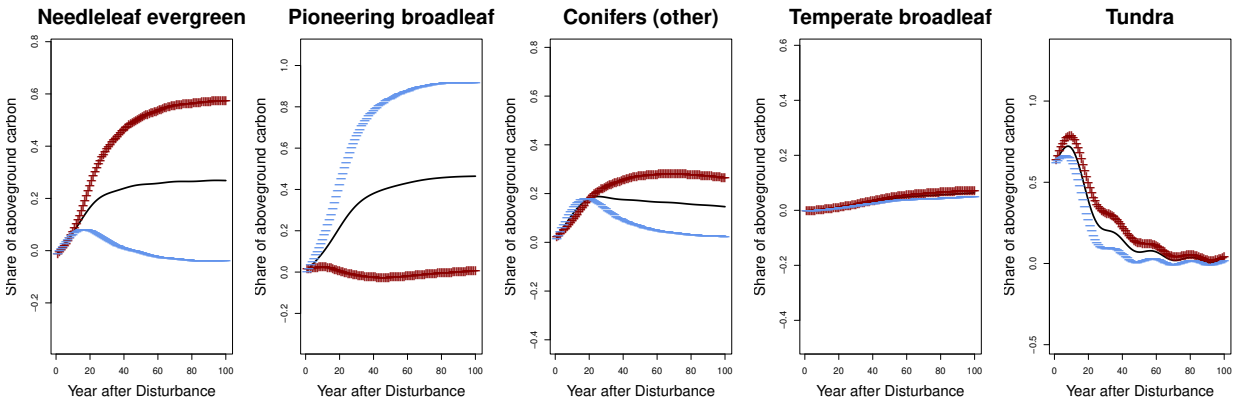


Figure 26: First PC derived by the MFPCA accounting for 69.99% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

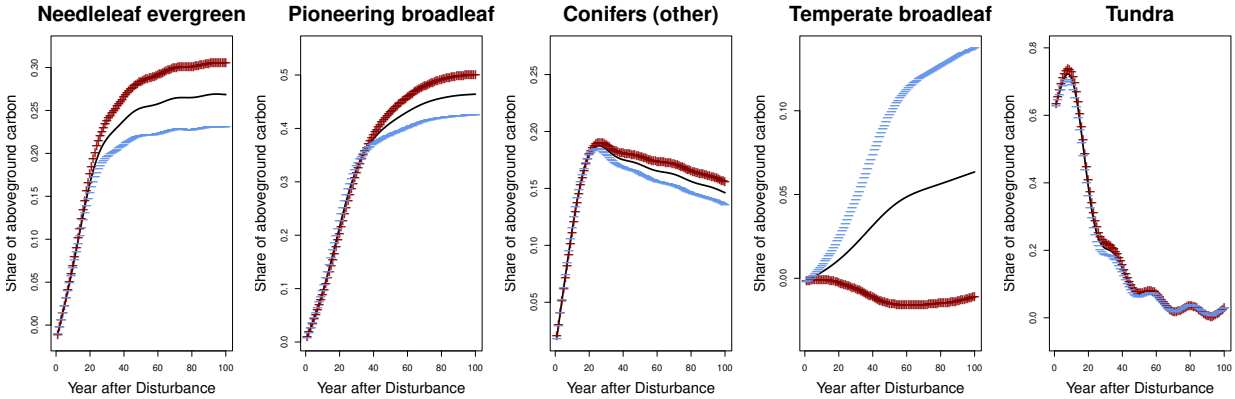


Figure 27: Second PC derived by the MFPCA accounting for 13.21% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

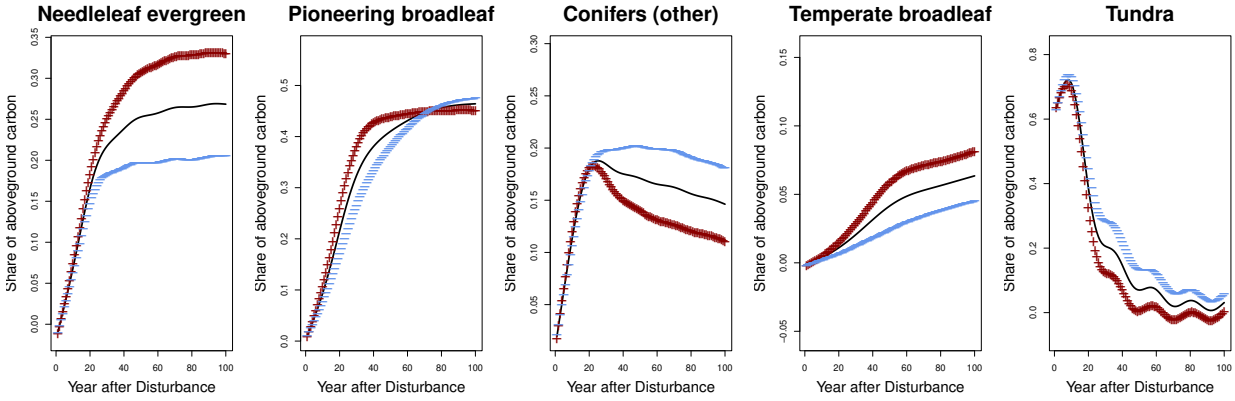


Figure 28: Third PC derived by the MFPCA accounting for 5.93% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

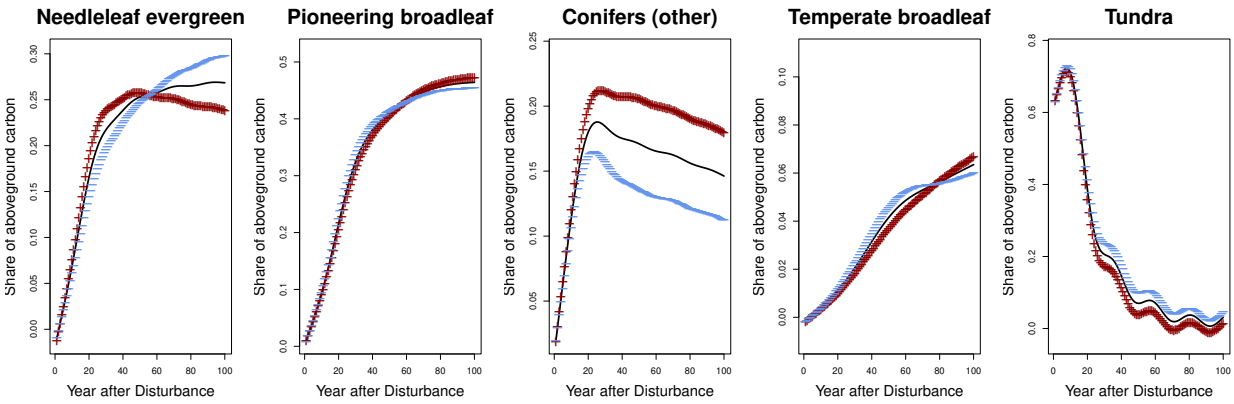


Figure 29: Fourth PC derived by the MFPCA accounting for 4.50% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

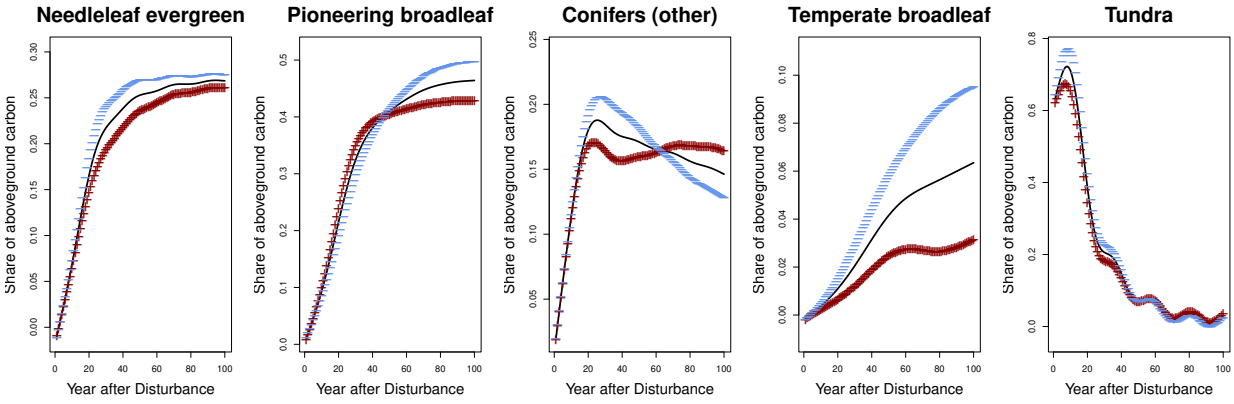


Figure 30: Fifth PC derived by the MFPCA accounting for 2.19% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

cate a higher share of aboveground carbon after a few decades of recovery, while for *tundra*, high values imply slightly lower proportions. All interpretations are vice versa for low values.

The remaining six PCs considered in this MFPCA together account for less than 6% of the variation in the data, which is a non-substantial increase in information. The interpretation is therefore kept short at this point. The fifth (Figure 30), seventh (Figure 32) and ninth (Figure 34) PC focus on variation and dynamics in PFT *temperate broadleaf*, similar to the second PC, while the sixth (Figure 31) and tenth (Figure 35) PC reflect patterns in the shares of *conifers (other)*. The eighth PC visualized in Figure 33 represent some dynamics in *needleleaf evergreen*, but all of these PCs have in common that the deviations to the average proportions of aboveground carbon are very small. Hence, they do not capture major sources of variation but small differences to the mean for selected grid cells. Using only the first few PCs is therefore sufficient for subsequent analyses.

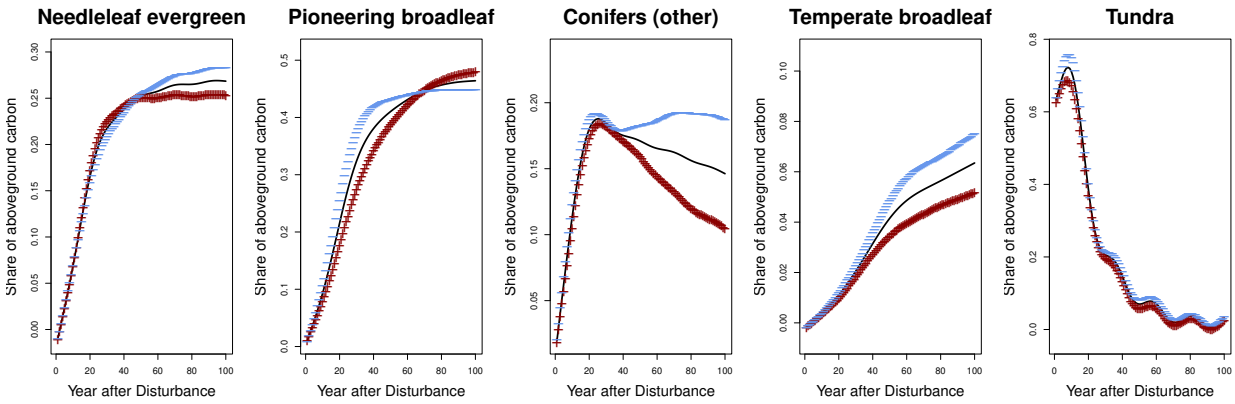


Figure 31: Sixth PC derived by the MFPCA accounting for 1.64% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

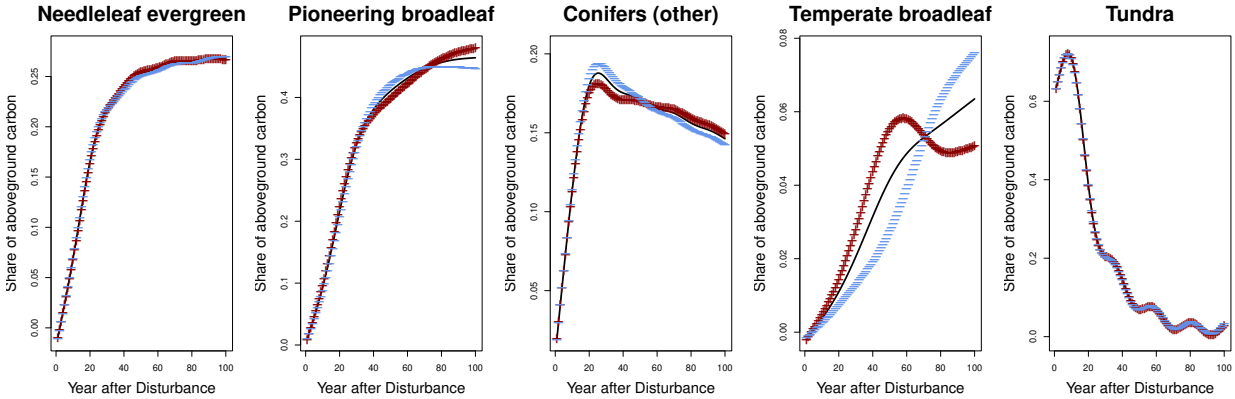


Figure 32: Seventh PC derived by the MFPCA accounting for 0.81% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

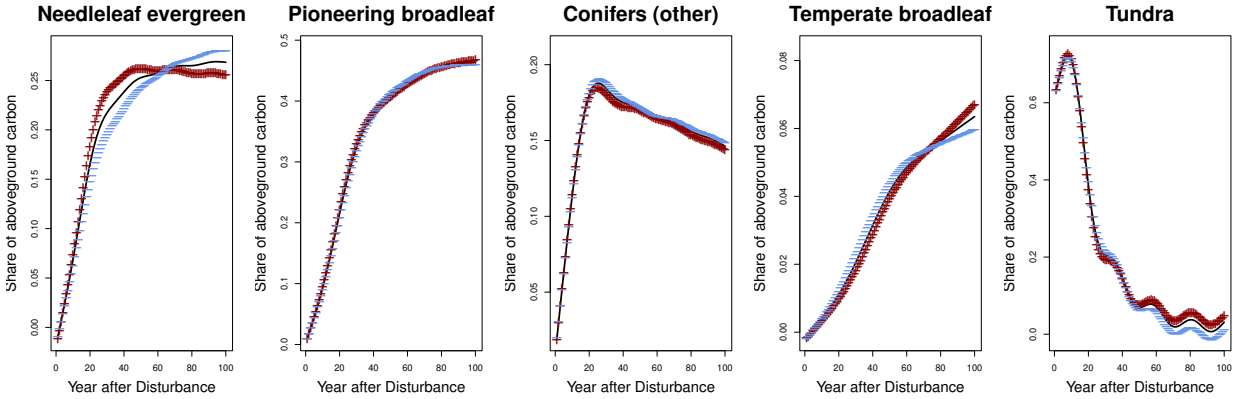


Figure 33: Eighth PC derived by the MFPCA accounting for 0.53% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

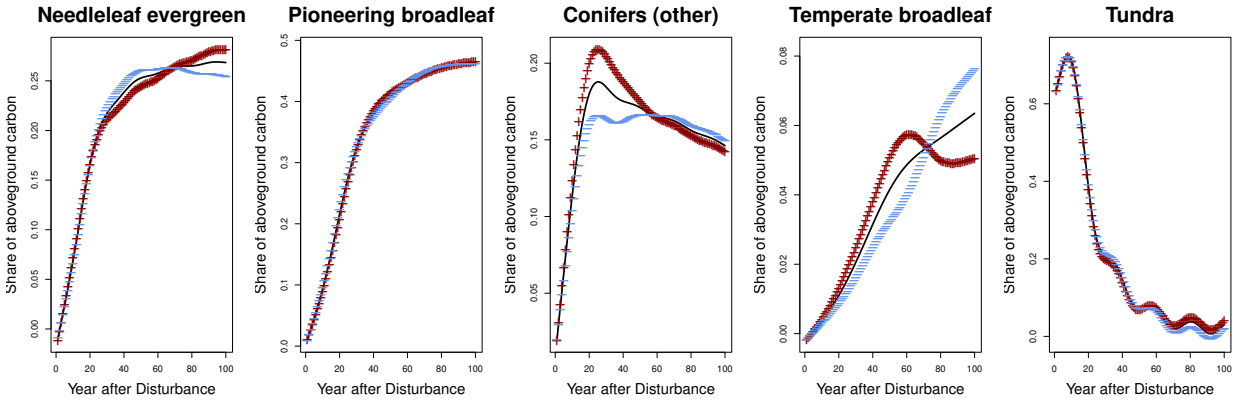


Figure 34: Ninth PC derived by the MFPCA accounting for 0.34% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

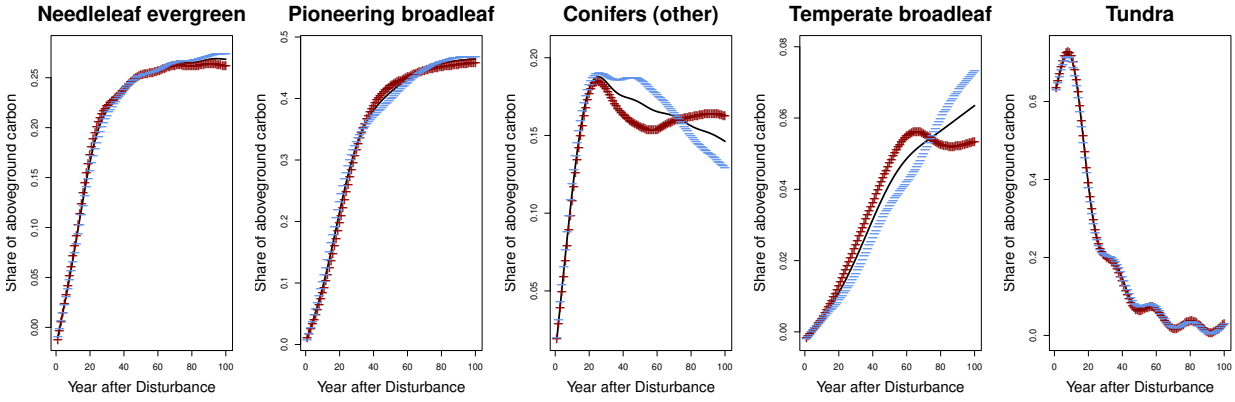


Figure 35: Tenth PC derived by the MFPCA accounting for 0.28% of the variability in the data. The red (+) and blue values (−) indicate the addition and subtraction of two standard deviations to each PC curve.

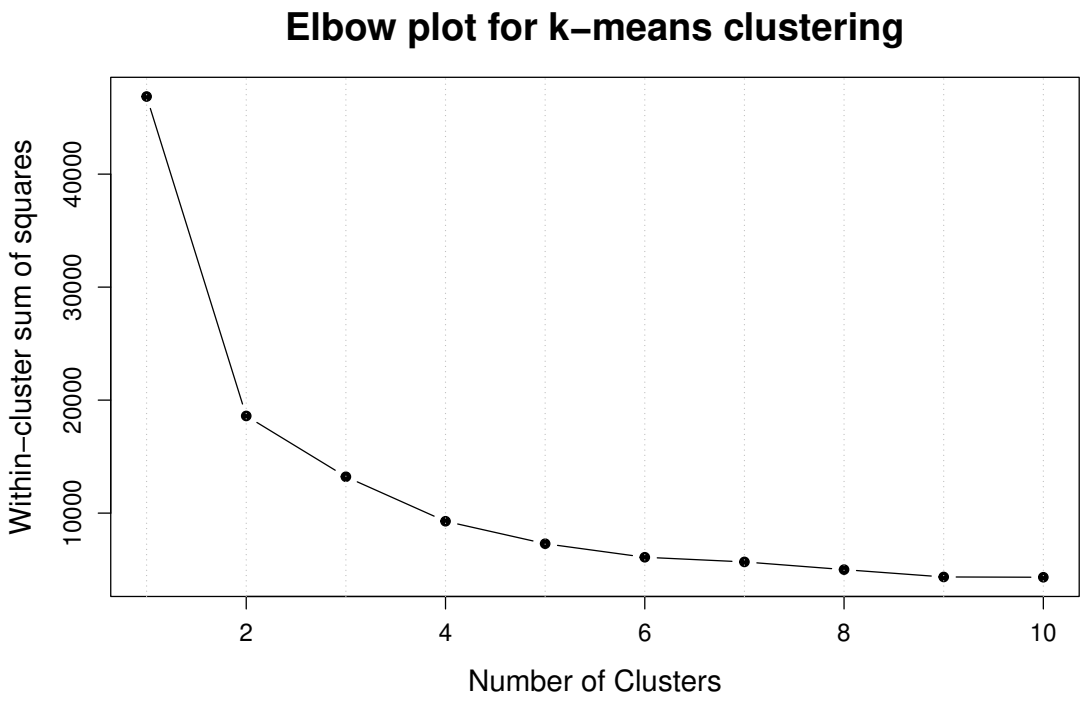


Figure 36: Elbow plot indicating the within cluster sum of squares for different numbers of clusters.

4.5 Clustering of PC scores

Recall that one of the aims of this analysis is to find patterns in the recovery trajectories. Performing a MFPCA yields PC scores for each trajectory, i.e., grid cell, and clustering these scores can provide insights into classifying the recovery behaviour of forests after disturbance. Therefore, the data set of ten available PC scores, that is, the 1803×10 matrix comprising the PC scores for each location, is clustered by a k-means algorithm. In order to determine the appropriate number of clusters, Figure 36 shows an elbow plot indicating the within cluster sum of squares for a variety of numbers of clusters. No clear elbow is visible, leading to the choice of $k = 4$ to be consistent with the derivations in Appendix A.2. The cluster composition is rather unbalanced, as Table 7 suggests. Two clusters, clusters 1 and 4, are clearly dominated by grid cells disturbed in the control scenario, while clusters 2 and

	Control	**SSP1-RCP2.6**	**SSP3-RCP7.0**	**SSP5-RCP8.5**	**Sum**
Cluster 1	100	73	80	58	311
Cluster 2	41	172	193	221	627
Cluster 3	7	20	38	44	109
Cluster 4	286	177	151	142	756
Sum	434	442	462	465	1803

Table 7: Number of curves, i.e., grid cells, in each cluster and each scenario.

3 are mainly driven by the three warming scenarios. To further explore this cluster composition, Figure 37 shows a cluster characterisation in terms of PFT-wise mean proportions of aboveground carbon for all four clusters. Means in this context refer to average proportions across locations in the respective clusters. The figure reveals that the vegetation in the first and second cluster is dominated by *pioneering broadleaf*, with a difference that in cluster 2 there are no other PFTs after a few decades, while in cluster 1 some needleleafed trees appear in minor fractions. Looking back at Table 7 highlights that the regime of dominance of *pioneering broadleaf* differs between the scenarios, as most of the grid cells in cluster 1 are disturbed in the control scenario, while in cluster 2 the control plays a minor role. Here, the more radiative forcing, the more disturbed grid cells are classified in cluster 2. The same applies to cluster 3. Here, after about 25 years of recovery, there is a peak in *pioneering broadleaf*, which is then displaced by *temperate broadleaf*, which dominates from about 35 years of recovery. *Temperate broadleaf* is usually an indicator of a milder climate (Kimmins, 2004), which is also reflected in the distribution of the grid cells (Table 7). Locations disturbed in the control scenario are barely present in cluster 3, but the more radiative forcing, the more grid cells fall into this cluster. Note that cluster 3 is the smallest of all clusters, suggesting that a dominance of *temperate broadleaf* remains rare even with climate change. The largest cluster 4 is dominated by *needleleaf evergreen* after almost 30 years, followed by *conifers (other)* after 40 years. Interestingly, no broadleafed trees manage to grow in these areas throughout the recovery period. Looking at the cluster composition in Table 7 shows that all four scenarios are present in this cluster, but the number of grid cells decreases with increasing radiative forcing and nearly 40% of the grid cells in cluster 4 are disturbed in the control scenario.

In order to gain insight into the spatial distribution of the clusters, Figure 38 illustrates the location of clusters across all four scenarios. It can be observed that cluster 1 is predominantly present in the northern regions of the study area, while clusters 2 and 3 are more prevalent in the southern parts. Cluster 4 does not follow any visible structures.

Recall that the first PC focuses mainly on dynamics in *needleleaf evergreen, pioneering broadleaf* and *conifers (other)*, with high values indicating more *needleleaf evergreen* and *conifers (other)* as well as less *pioneering broadleaf* than on average. The second PC reflects changes in *temperate broadleaf*, with low values indicating an above-average share of aboveground carbon. These patterns can now be applied to the clusters. Figure 39a shows the PC scores of the first two PCs plotted against each other, with colors indicating the cluster. Clearly, cluster 3 comprises grid cells with a low second PC scores to account for more *temperate broadleaf*. The clusters 2, 1 and 4, which are alongside each other, are more or less at the same level in terms of second PC scores, but reach increasing scores of the first PC. Looking at the PFT-wise means in Figure 37 shows that from cluster 2 via cluster 1 to cluster 4, the proportion of *pioneering broadleaf* decreases and is displaced by needleleafed trees. Figure 39b shows the equivalent plot with colors highlighting the scenarios. The clustering clearly finds patterns beyond the scenario in which the grid cells are disturbed, and shows the variety in recovery trajectories.

In summary, clustering MFPCA scores is an appropriate approach for detecting patterns in

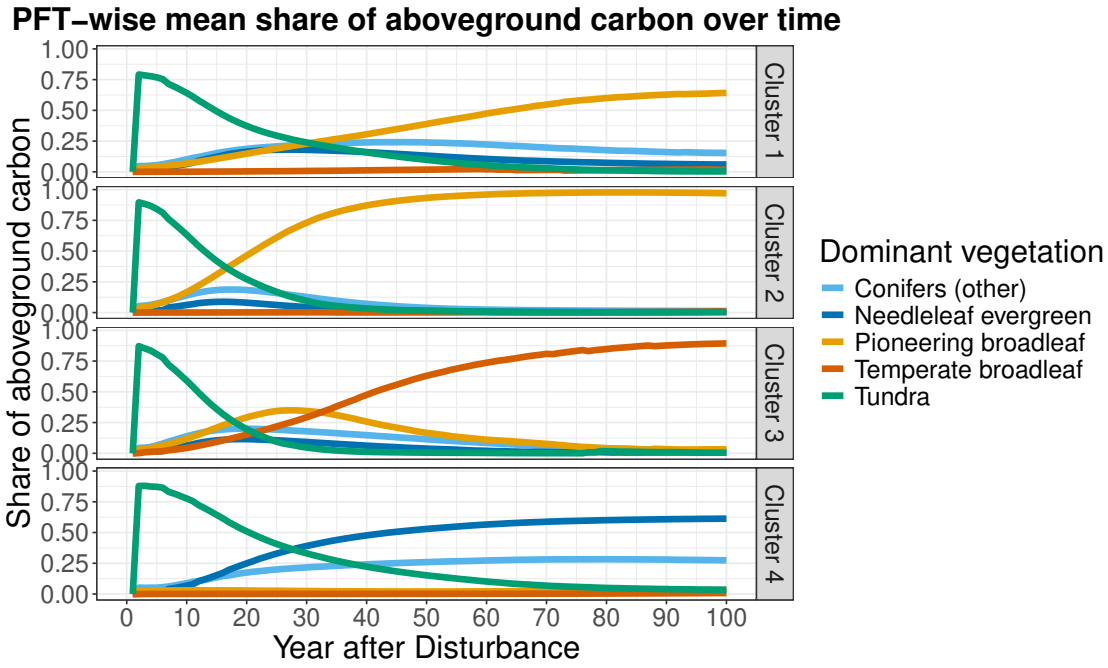


Figure 37: PFT-wise mean shares of aboveground carbon over time. Note that the values are averages over locations and clusters.

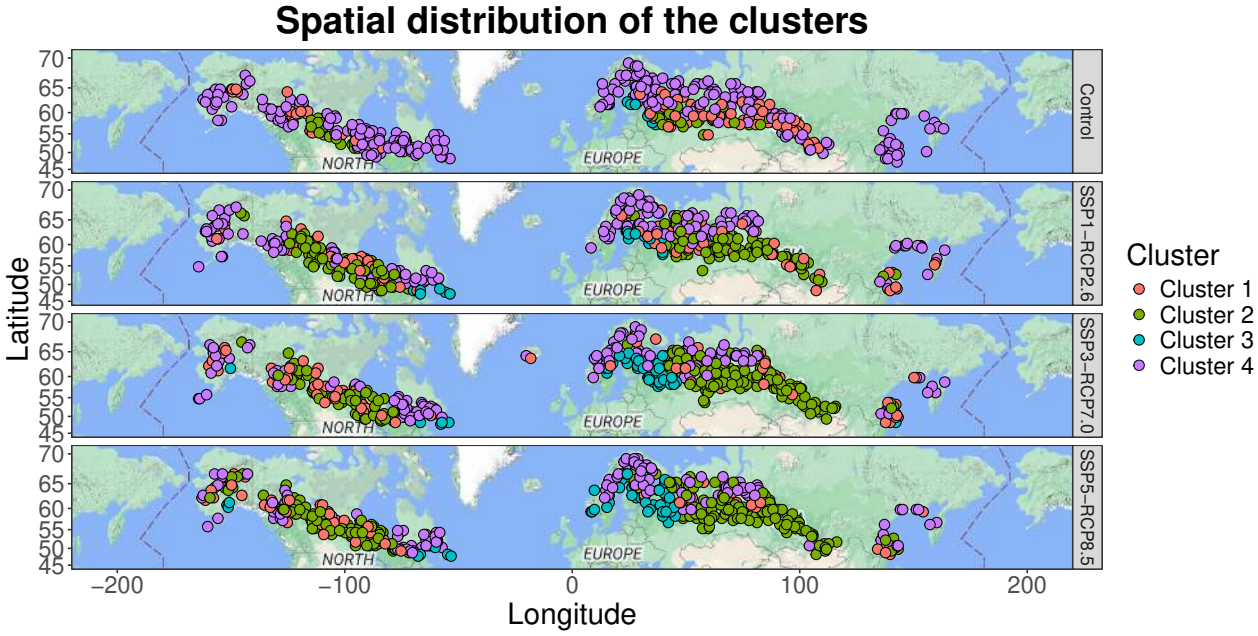
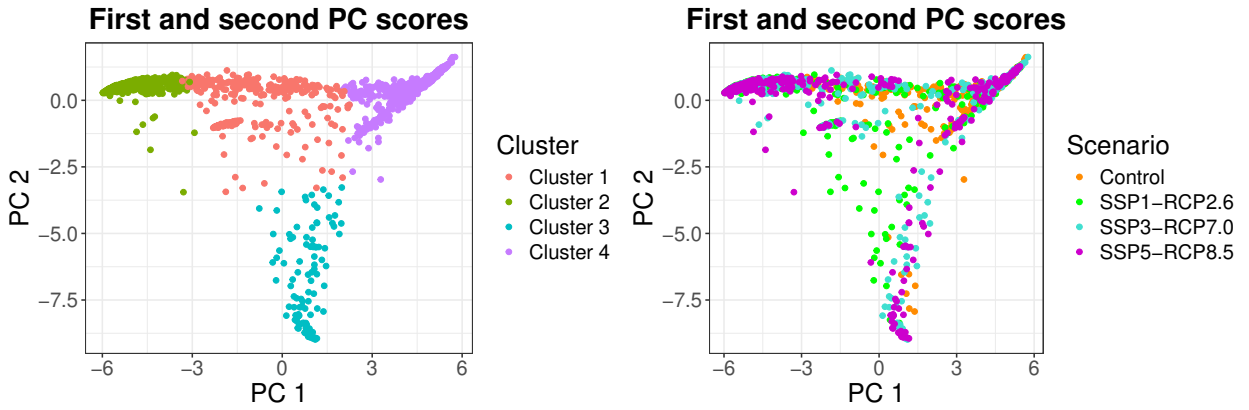


Figure 38: Spatial distribution of the clusters derived by MFPCA.



(a) (b)

Figure 39: First against second PC scores for the performed MFPCA. The left plot shows the resulting clusters using a 4-means clustering algorithm, the right plot visualizes the corresponding scenarios.

recovery trajectories. The clustering reveals three regimes in terms of dominance: dominance of *pioneering broadleaf* (clusters 1 and 2), dominance of *temperate broadleaf* (cluster 3) and dominance of needleleafed trees (cluster 4), which in turn can be separated in terms of scenario focus: clusters 1 and 4 mainly comprise grid cells disturbed in the control scenario, while locations in clusters 2 and 3 are mainly disturbed in the three warming scenarios. Appendix A.3 gives a detailed description of which curve of which scenario and for which PFT belongs to each cluster.

4.5.1 Details on Soil Properties within Clusters

In order to assess whether the clusters, which are derived only on the basis of PFT-wise proportions of aboveground carbon per grid cell, can be further characterised in terms of soil properties (see Section 3.5, Table 4 and Section 4.1), Figure 40a shows the soil composition, i.e., the distribution of clay, sand and silt, for each cluster. The soils of the grid cells in clusters 1 and 2 consist of slightly more clay and less sand than those in the other two clusters. The differences in silt fraction are less pronounced. This implies that silt is less relevant for cluster formation and thus for aboveground carbon proportions.

Figure 40b visualises the bulk density for each cluster. As already seen for the soil composition, the differences between the clusters are rather small. More soils in cluster 4 tend to have a higher bulk density than in the other three clusters, which themselves are more evenly distributed. All four clusters show concentrations around 0.3 and 1.5.

For all properties considered so far, clusters 1 and 2 behave similarly. This applies to the organic carbon content in the soil as well, as depicted in Figure 40c. This phenomenon is not surprising, since Figure 37 revealed that both clusters follow a similar regime in terms of dominant vegetation. There are only minor differences in organic carbon content between the clusters. The variation in small values is more spread in cluster 4 and more evenly

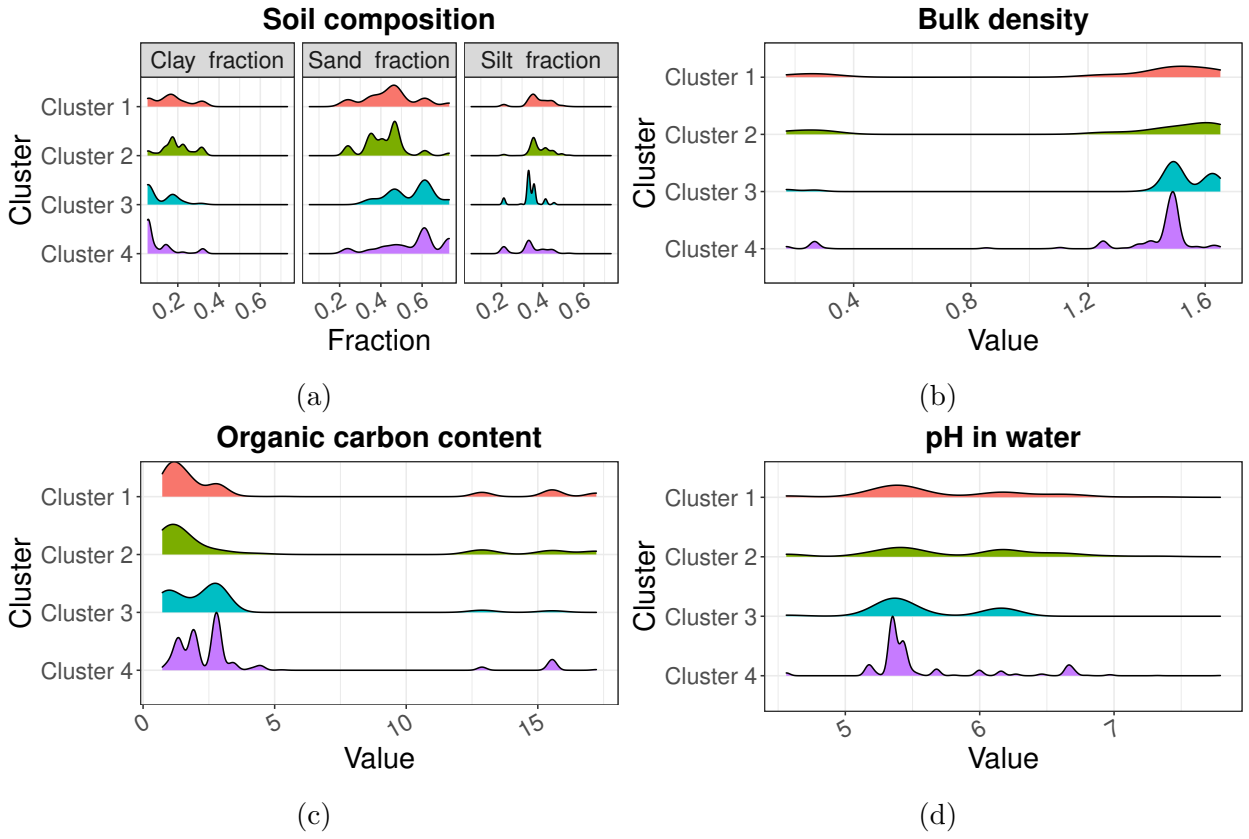


Figure 40: Soil properties for each cluster.

distributed for the remaining three clusters. All clusters have data concentrations around 1 and 15 in common.

Also for pH in water, visualized in Figure 40d, clusters 1 to 3 show only a moderate range of pH values with no extreme variations, whereas the pH value in water in cluster 4 is more variable.

In summary, the soil properties of the clusters do not differ substantially. Cluster 4 shows slightly more variable properties compared to the other three clusters, and clusters 1 and 2 behave very similarly. Both observations are consistent with cluster characteristics, as cluster 4 is the largest cluster with more than 40% of all disturbed grid cells, and clusters 1 and 2 exhibit similar dynamics in dominant vegetation.

4.5.2 Details on Ecological Properties within Clusters

In Section 3.5 ecological properties of disturbed grid cells were introduced. The question now is whether the clusters reveal any features with respect to these properties. Figure 41a shows the total nitrogen uptake averaged over all disturbed grid cells for each cluster. In contrast to the soil properties above, where clusters 1 and 2 behave similarly, here clusters 1 and 4 and clusters 2 and 3 show comparable behaviour. Recall that the first two clusters consist

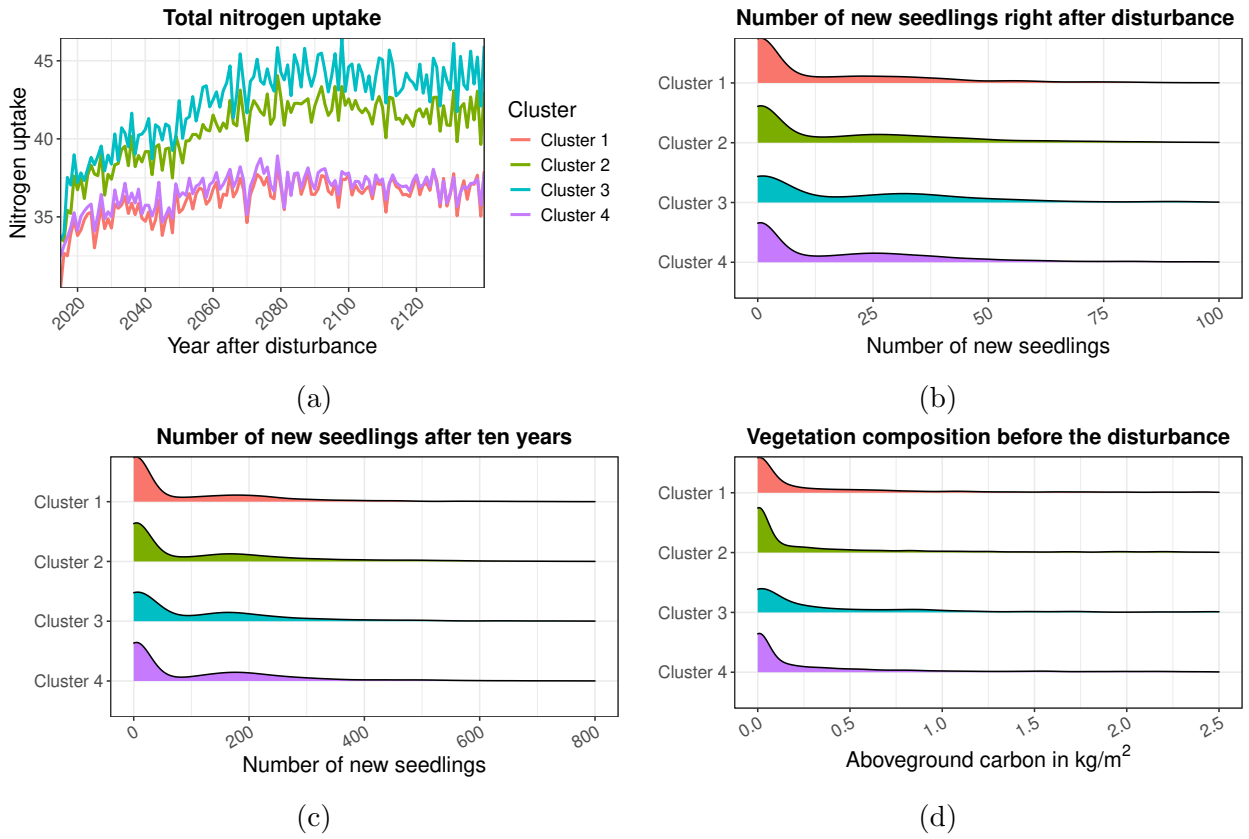


Figure 41: Ecological properties for each cluster.

mainly of grid cells disturbed in the control scenario and the last two clusters are dominated by grid cells disturbed in the three warming scenarios. This means in particular that nitrogen uptake seems to be more sensitive to climate warming, which leads to increased uptake. This pattern was already present in Figure 7b in the descriptive analysis in Section 4.1.

Figure 41b and Figure 41c show the distribution of the number of new seedlings immediately after the disturbance and summed over the first ten years of recovery for each cluster. All four clusters have a peak at zero, but the size of the peak differs, being largest in the first cluster and smallest in cluster 3. For both variables, no major differences between the clusters are evident.

The amount of aboveground carbon prior to the disturbance between 2015 and 2040 visualized in Figure 41d clearly differs between the clusters. All four clusters show a concentration at zero, but while clusters 1 and 4 have a medium high peak, it is highest for cluster 2 and lowest for cluster 3. Interestingly, both clusters 2 and 3 are dominated by broadleafed trees, but this difference in aboveground carbon could indicate a different effect on the distribution of *temperate broadleaf* and *pioneering broadleaf*.

Overall, similar as for soil properties, the differences between the clusters are rather small, indicating no major impact on the patterns within recovery trajectories.

4.5.3 Details on Climatic Properties within Clusters

In addition to soil and ecological variables, functional climate data are also available. Figure 42 shows the mean temperature and precipitation curves for each cluster. Mean values in this context represent the pointwise annual mean values averaged over grid cells in each scenario. The behaviour of the mean annual temperature per cluster, visualised in Figure 42a, is very similar to the mean nitrogen uptake (Figure 41a). The locations covered by clusters 1 and 4, the two clusters focusing on the control scenario, tend to have a lower mean temperature after the first decades of recovery than the two clusters, clusters 2 and 3, dominated by grid cells disturbed in the warming scenarios. The same pattern is visible for the yearly minimum temperature depicted in Figure 42b. Note that here the differences between the clusters are less pronounced than for the yearly mean temperature. For the annual maximum temperature in Figure 42c, clusters 1 and 4 are barely distinguishable and show the lowest maximum temperatures, but the differences between the second and the third cluster increase. Cluster 3, dominated by *temperate broadleaf*, shows the highest annual maximum temperature throughout the whole study period, while cluster 2 remains between cluster 3 and clusters 1 and 4. This suggests that it is the high maximum temperatures, rather than the lower mean and minimum annual temperatures, that are the drivers for higher proportions of *temperate broadleaf*.

The differences between the clusters in terms of annual summed precipitation, shown in Figure 42d, are less pronounced than for temperature. While cluster 2, the cluster dominated by *pioneering broadleaf* and grid cells disturbed in the three warming scenarios, tends to have the most precipitation, cluster 3 has the least precipitation in the early decades of the recovery, which is replaced by cluster 4 in the last decades of the study period. Cluster 1 remains in between throughout the period.

Altogether, while there are only minor differences in precipitation between the clusters, there is a clear distinction between the clusters dominated by the control scenario and those reflecting the grid cells disturbed in the three warming scenarios in terms of minimum, maximum and average annual temperature.

4.5.4 Temporal Consistency of Clusters

The clusters considered so far have all been derived using the full recovery period of 100 years. A recent study of Mandl et al. (2024) suggests that the reorganization phase following a disturbance – essentially the first few years – is highly influential in shaping the eventual forest structure and resilience. To verify that the simulation captures this behaviour, the robustness of the clustering is examined by applying the 4-means algorithm to the PC scores using only certain time points in the trajectory. Recall the definition of PC scores obtained in Equation 6. In this approach, the eigenfunctions are again derived using the entire recovery period, but instead of taking the integral over the entire trajectory for the calculation of PC scores, it is taken over pre-specified single time points. For instance, the first PC score of the first location after 10 years of recovery is obtained by

$$\mathbf{s}_{11} = x_1(10)\xi_1(10).$$

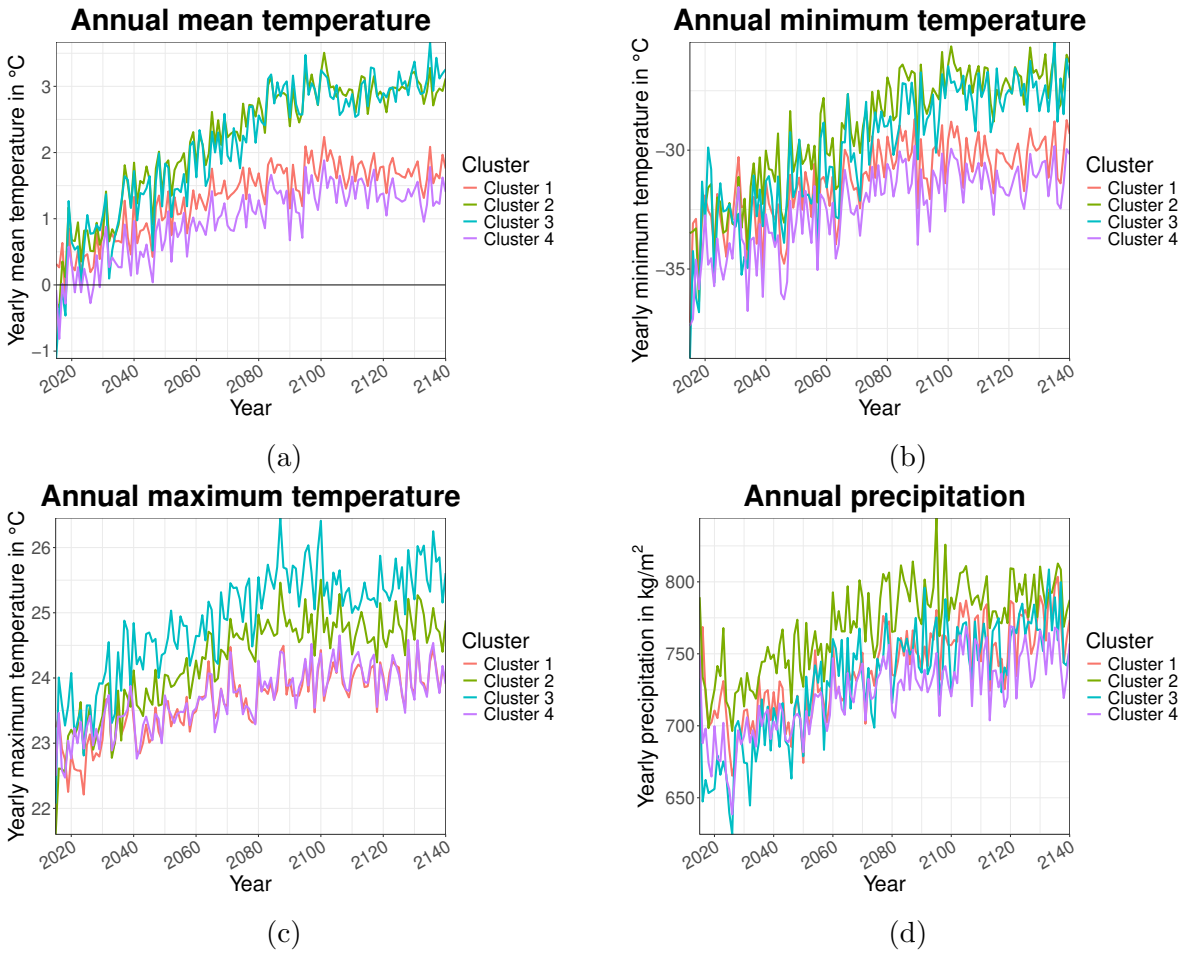
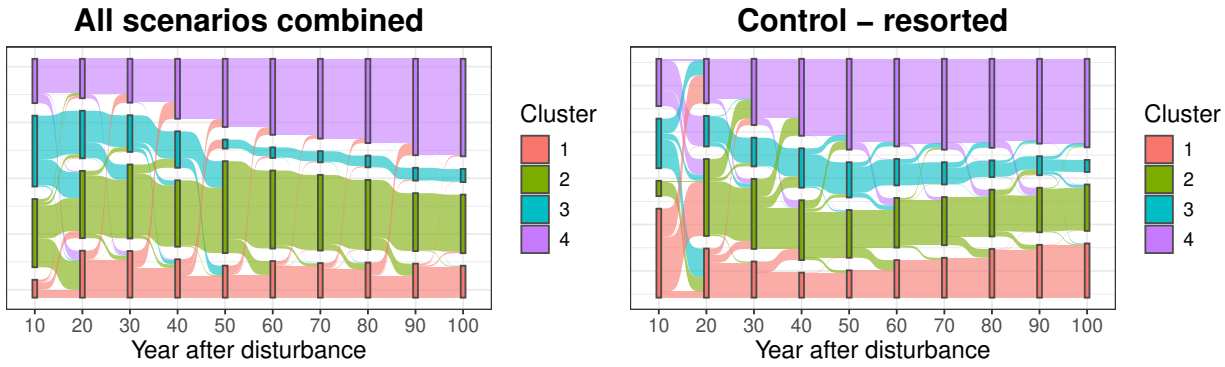


Figure 42: Climatic properties for each cluster.

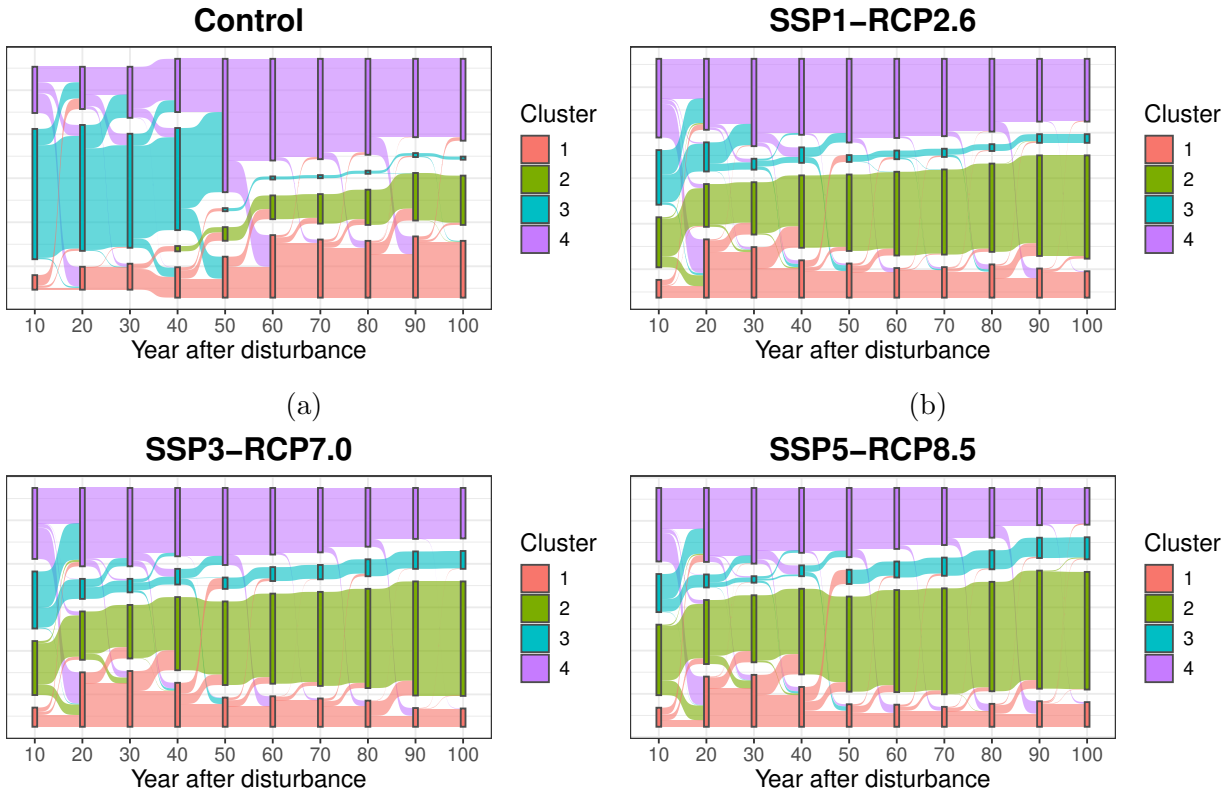
The goal is to derive the PC scores for the time points t, which are evenly spaced between 10 and 100, inclusive. These scores are then subjected to a 4-means algorithm for cluster detection. Following this, the clusters are examined for similarity and consistency. Figure 43a illustrates the cluster development over time for the entire data set as a sankey plot. The clustering appears robust after a few decades of recovery. After 50 years the cluster composition hardly changes, indicating that the vegetation composition is indeed determined by the first years of recovery.

When the clusters are broken down by scenario to examine robustness to different climate conditions, Figure 44 shows the cluster development for all four climate scenarios considered. Note that the scores and clusters are exactly the same as in the previous figure, only separated by scenario. On the one hand, the distribution over time for the three warming scenarios in Figure 44b (SSP1-RCP2.6), Figure 44c (SSP3-RCP7.0) and Figure 44d (SSP5-RCP8.5) is rather similar and follow the same pattern as the development of the entire data set in Figure 43a. One could hypothesise that the clustering appears robust after 20 years, as the clusters change only marginally after each time step considered. However, for the control scenario shown in Figure 44a, the distribution is more variable up to 50 years after the dis-



(a) (b)

Figure 43: Cluster composition over years after disturbance for the entire data set (a) and for the control scenario only (b) with resorted cluster assignments.



(a) (b)

(c) (d)

Figure 44: Cluster composition over years after disturbance for all four scenarios.

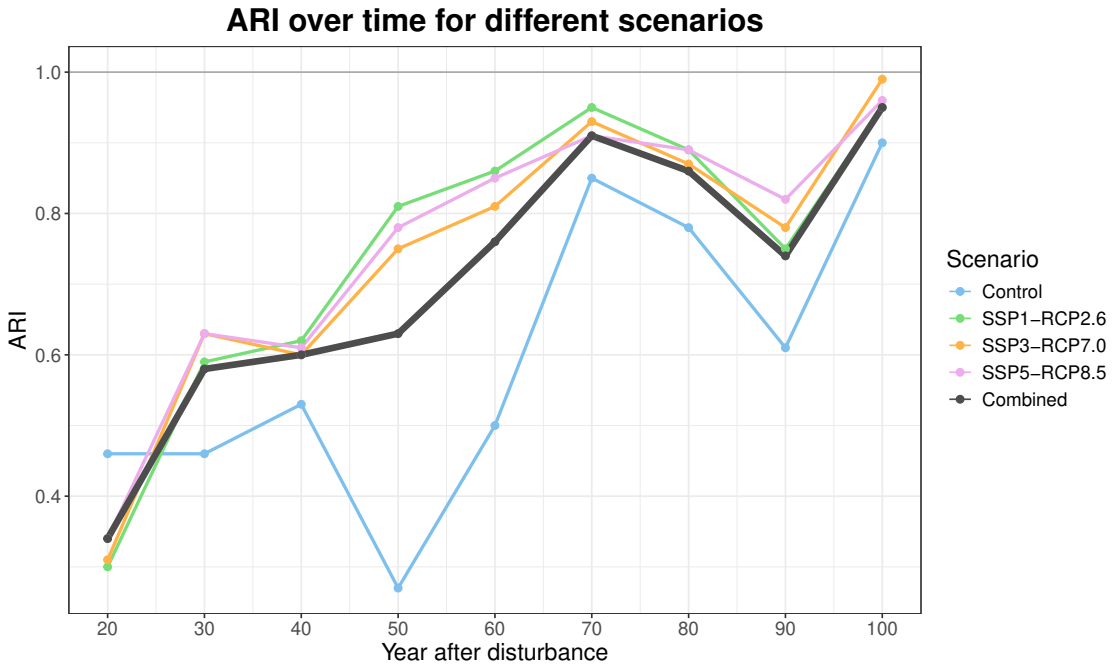


Figure 45: ARIs of cluster composition over years after disturbance for all four scenarios separately and combined.

turbance. This could be due to the order of the clusters. To make the clusters comparable, the cluster assignments are all aligned so that the majority of the cluster at time t remains in the cluster at time $t-10$. For example, if cluster 1 at $t = 80$ contains the majority of the grid cells of cluster 2 at $t = 90$, cluster 1 is renamed cluster 2. This process is done for the entire data set and is not scenario based. Interestingly, the ordering remains appropriate for the three warming scenarios, but operates inconsistently for the control scenario. Figure 43b shows the cluster development for sorting the cluster assignments for the control scenario only. The result is the same as for the warming scenarios, i.e., the cluster composition is already detectable after 20 years of recovery and changes only slightly thereafter. This result – sorting the entire data set does not lead to the same conclusions as for sorting the scenarios separately – indicates the similarity of the three warming scenarios which together comprises the majority of grid cells in the entire data set. The control scenario steps out of line.

For a final mathematical evaluation of the similarity of the clustering over time, ARI values are computed for each time step. Note that the original sorted clusters of the control scenario are taken into account. Figure 45 shows the computed values for all four scenarios and the full data set. For all simulation runs, the ARI values increase almost continuously with time, but show a small decrease between 70 and 90 years of recovery.

To conclude, this temporal clustering approach clearly indicates that the establishment of new vegetation is mainly driven by the first 20 years of recovery after the disturbance when considering the scenarios separately, and by the first 50 years when considering the entire data set.

5 Modelling Approach

Recall that the modelling approach derived in Section 2.3 is based on PC scores for both functional response and covariates. This chapter aims at implementing this model utilizing the MFPCA scores derived in Section 4.4 and the provided covariates introduced in Section 3. First, the covariates need to be pre-processed for matching the data structure. Then, the variable selection process is explained before diving deeper into the general model setup.

5.1 Preparing Functional and Non-Functional Predictors

The available covariates described in Table 4 can be classified into four different groups:

- PFT-dependent and time-varying: `Nuptake`
- PFT-dependent only: all ecological variables except for `Nuptake` and `Nuptake_total`
- time-varying only: all climate variables and `Nuptake_total`
- location dependent only: all soil variables

This classification induces the pre-processing of each covariate. Due to reasons of explainability, redundancy and causality, not all of the provided predictors are included in the final model. Details on variable selection are provided in Section 5.2.

5.1.1 Non-Functional Predictors

The PFT-dependent variables are transformed to PFT-wise covariates to match the general data shape (1803 observations, i.e., disturbed grid cells). For every disturbed grid cell, the soil related variables are already provided, so no further pre-processing is necessary.

5.1.2 Functional Predictors

All time-varying, location dependent variables, e.g., the climate covariates, are functional as the response. This implies that these covariates need to be transformed into multivariate data prior to modelling. In particular, in this approach, all four climate variables `tas_yearlymean, tas_yearlymin, tas_yearlymax`, and `pr_yearlysum` are represented as PC scores in the final model formulation, which result from an MFPCA:

Functional representation
As before, the first step in FDA is to find an appropriate representation for the functional data. Figure 46 shows the functional fit for the annual average temperature for each scenario. Note that the unit of temperature here is K instead of °C as in the descriptive analysis to avoid negative values. As expected, with rising radiative forcing, the annual mean temperature increases. The same applies to the annual minimum temperature displayed in Figure 47 and annual maximum temperature (Figure 48). All three temperature-related covariates underline the similarity between the control and the weakest scenario SSP1-RCP2.6 and between the two more extreme scenarios SSP3-RCP7.0 and SSP5-RCP8.5. Note that there

is no basis representation or smoother involved here. Similarly, Figure 49 visualizes scenario-wise functional fits of the annual sum of precipitation. Interestingly, the figure suggests the hypothesis, that the more radiative forcing, the less variation in precipitation. This observation is consistent with that of other scientists working with ISIMIP climate data.

MFPCA
In general, all functional representations indicate that most of the variation in the data is captured by horizontal lines. This is confirmed by considering an MFPCA for all four climate covariates together with three PCs. Figure 50 shows the first two PCs, which together capture almost all the variability in the data, as the first PC already accounts for 98.56% of the variation. For all climate covariates, the first PC captures a shift of the mean function, with larger values indicating an upward shift and smaller values a downward shift. The second PC accounts for a shift in the trend after varying decades of recovery. Lower values lead to below-average temperature and precipitation values in the first years and above-average values in the last decades of the study period, and vice versa for high values.

Reconstruction
Figure 51 depicts the original functional fit (left) and the reconstructed curves using the first three PCs (right) for all three temperature variables, here for all scenarios together. Decreasing the dimensionality of the data in this way reduces noise and smooths out inter-annual variability. This is particularly useful for statistical modelling in subsequent steps. The same applies to the annual sum of precipitation visualized in Figure 52.

Note that for validity checks, for both functional covariates concerning the nitrogen uptake, i.e., `Nuptake` per PFT and `Nuptake_total`, and all climate covariates a separate FPCA was conducted as well, with similar results concerning the importance of the first PC. As they are not relevant for the model fitted in the subsequent Section 5.3, details of the respective FPCA including the functional representation, the PCs and the reconstructions are not discussed in detail here.

With these transformations, all functional and non-functional covariates are prepared for statistical modelling.

5.2 Variable Selection

Recall that the mFLR model introduced in Section 2.3.2 can be reduced to a multivariate linear regression model with PC scores as multivariate response variables and possibly multivariate predictor variables. This involves two steps of variable selection. First, the number of PCs to include is determined for each functional component. Second, all other covariates are examined to assess whether they should be included or not.

Consider first the functional predictors, i.e., the above mentioned separate FPCAs performed on all climate covariates and both nitrogen uptake variables, as well as the MFPCA on all climate covariates combined. Since each first PC accounts for over 90% of the variance in the data in all performed FPCAs and MFPCAs, only the first PC scores should be included

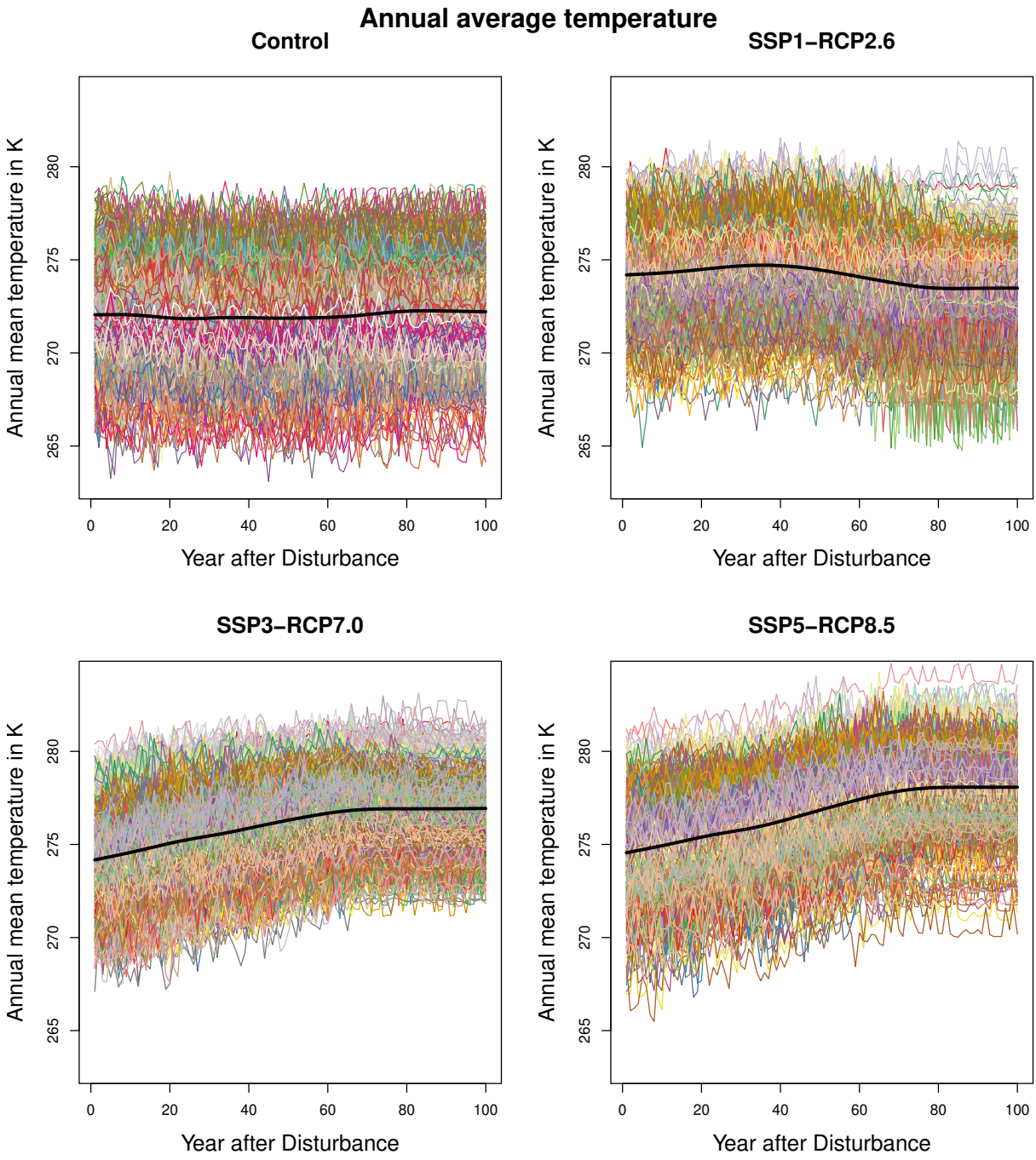


Figure 46: Functional fit for annual average temperature for each scenario.

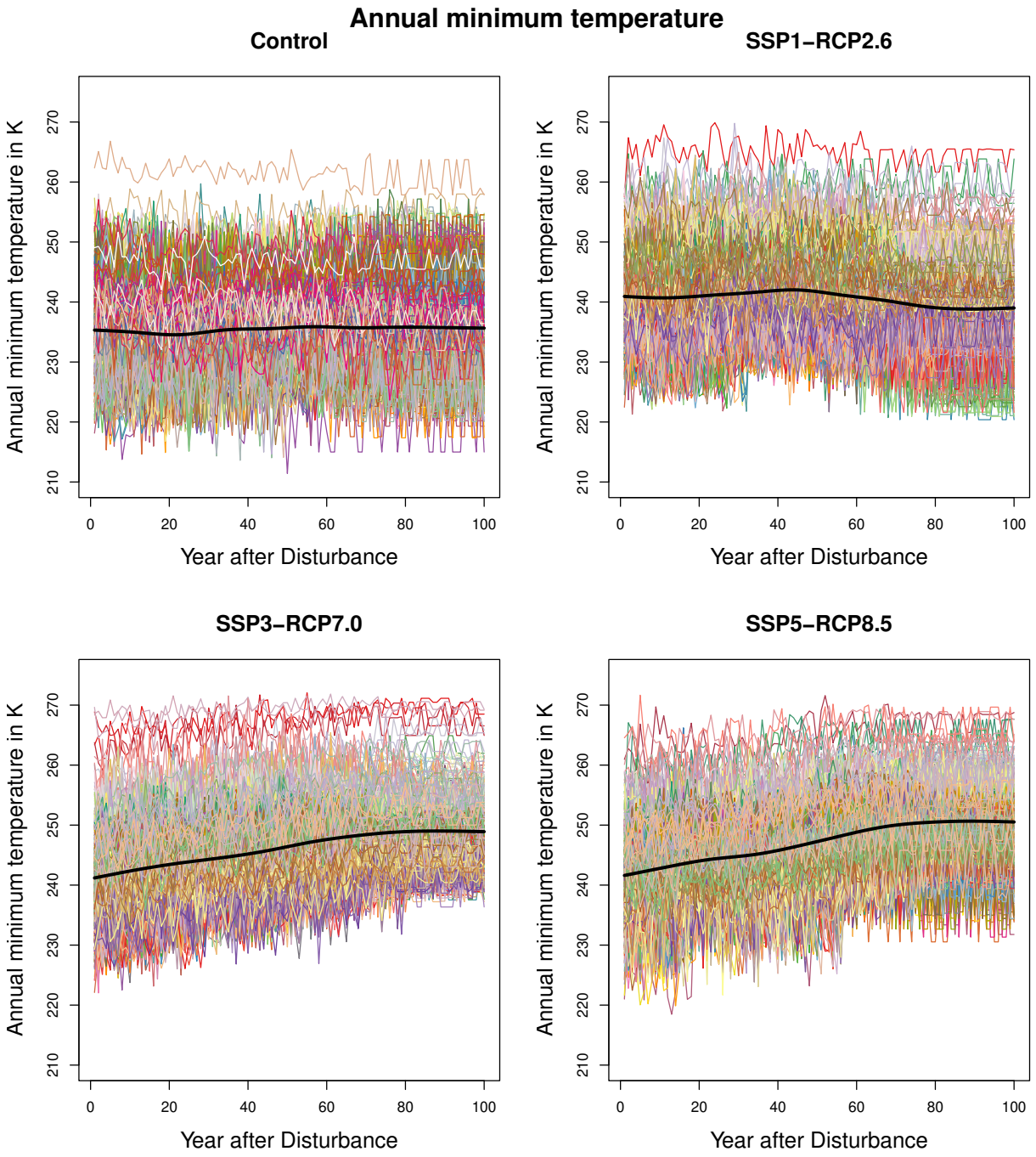


Figure 47: Functional fit for annual minimum temperature for each scenario.

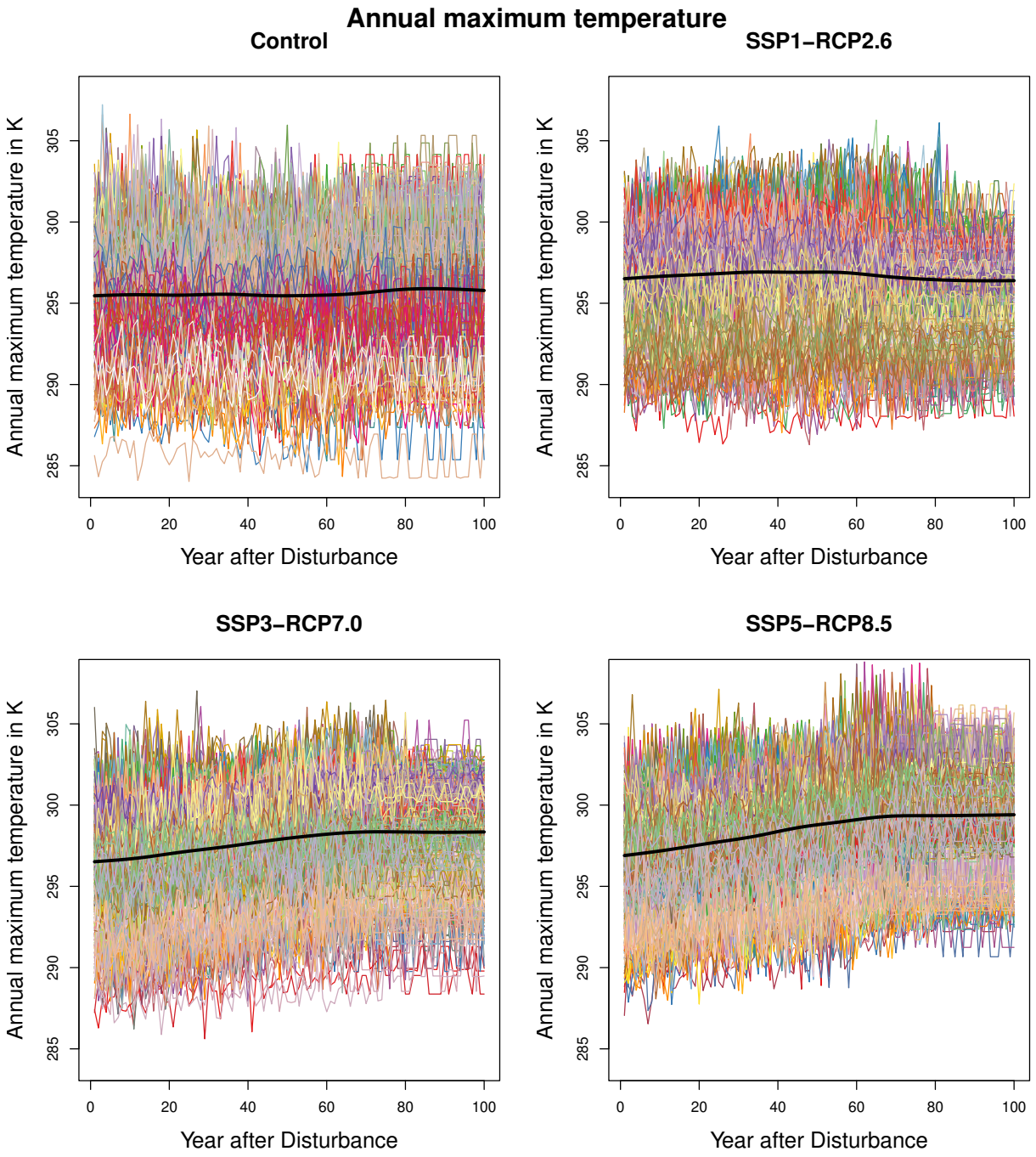


Figure 48: Functional fit for annual maximum temperature for each scenario.

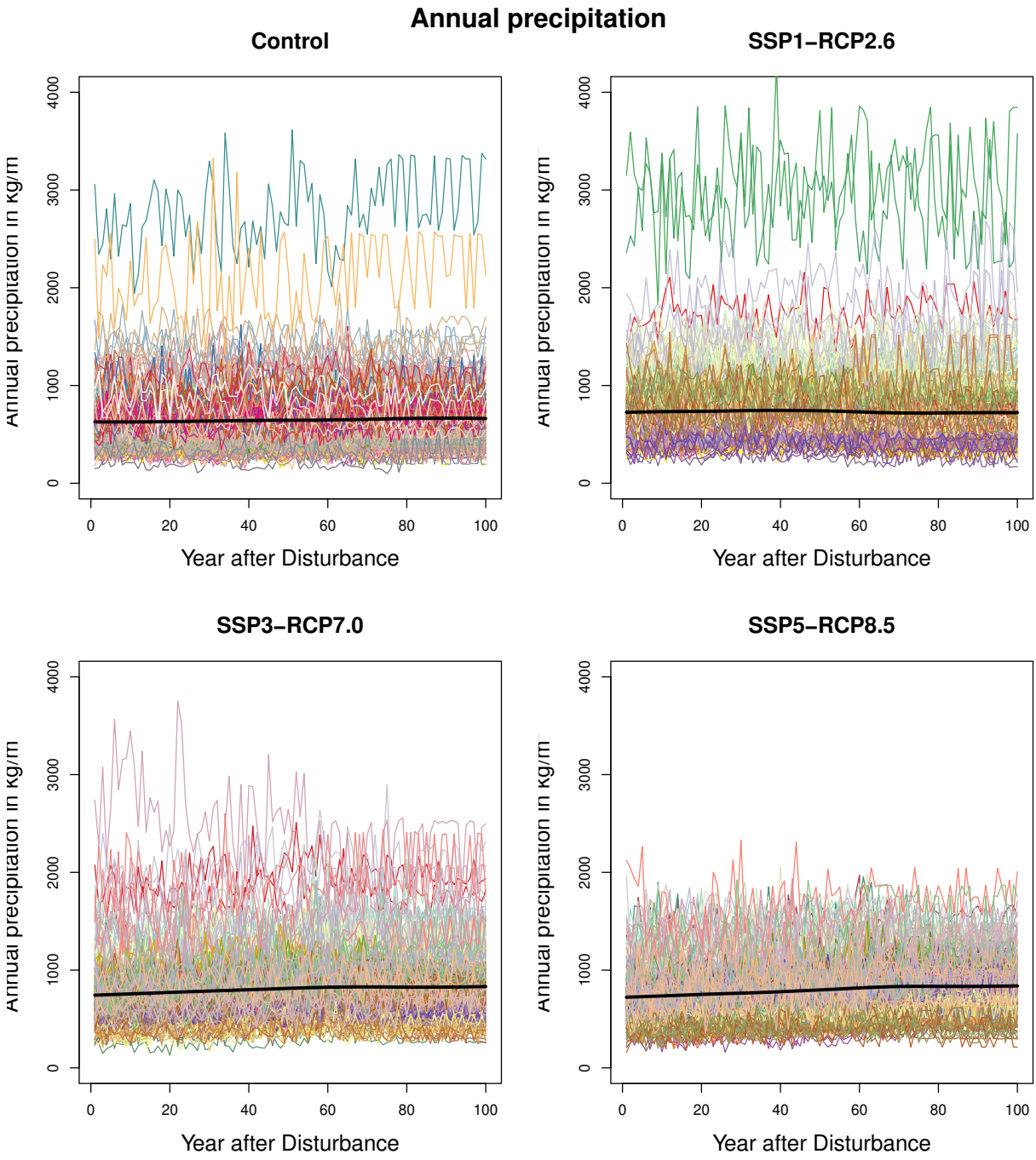
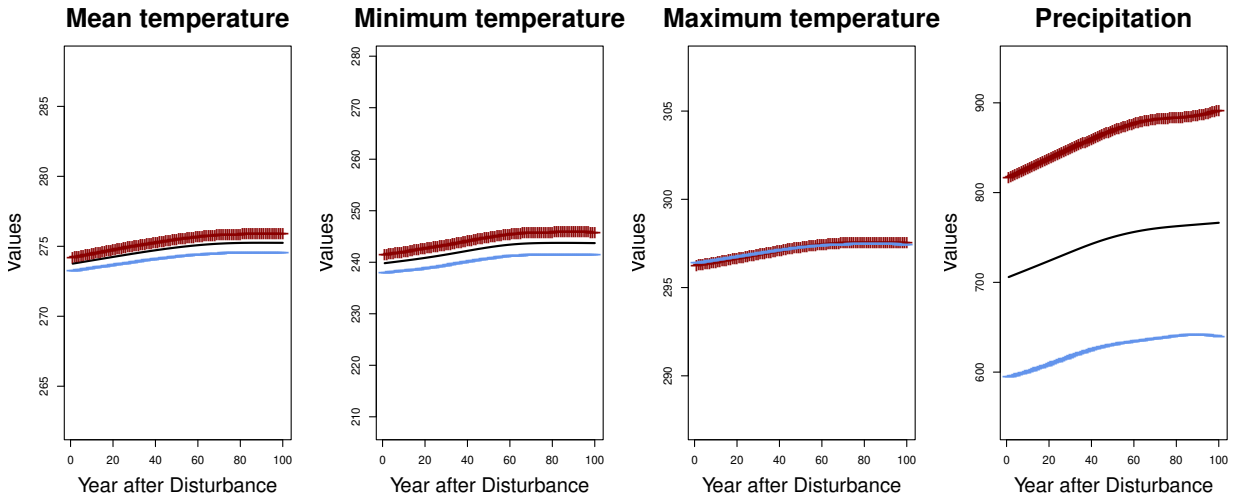
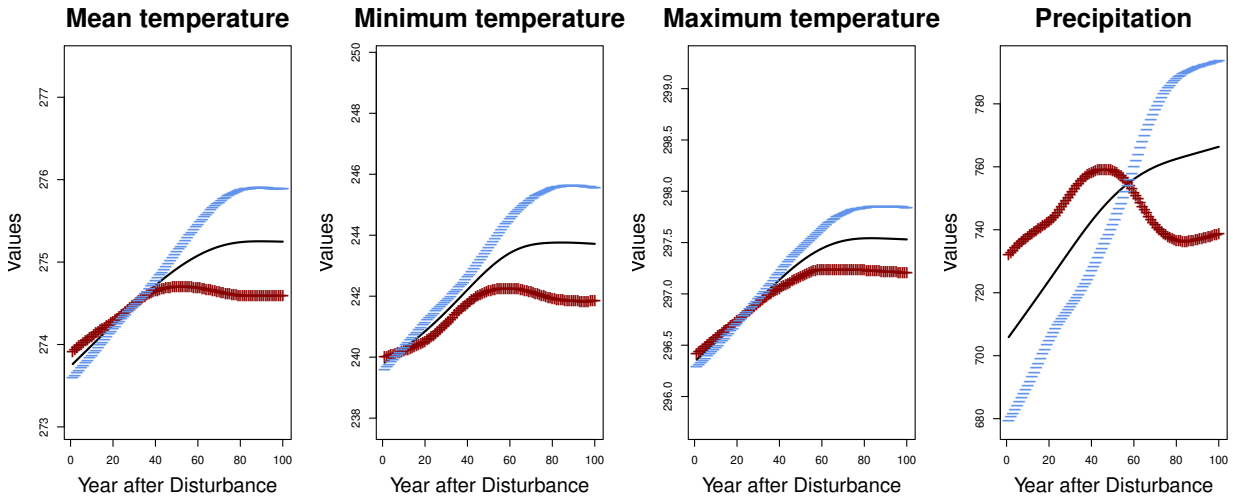


Figure 49: Functional fit for annual precipitation for each scenario.



(a)



(b)

Figure 50: First (a) and second (b) PC derived by the MFPCA performed on all four climate covariates, accounting for 98.56% and 1.18% of the variability, respectively.

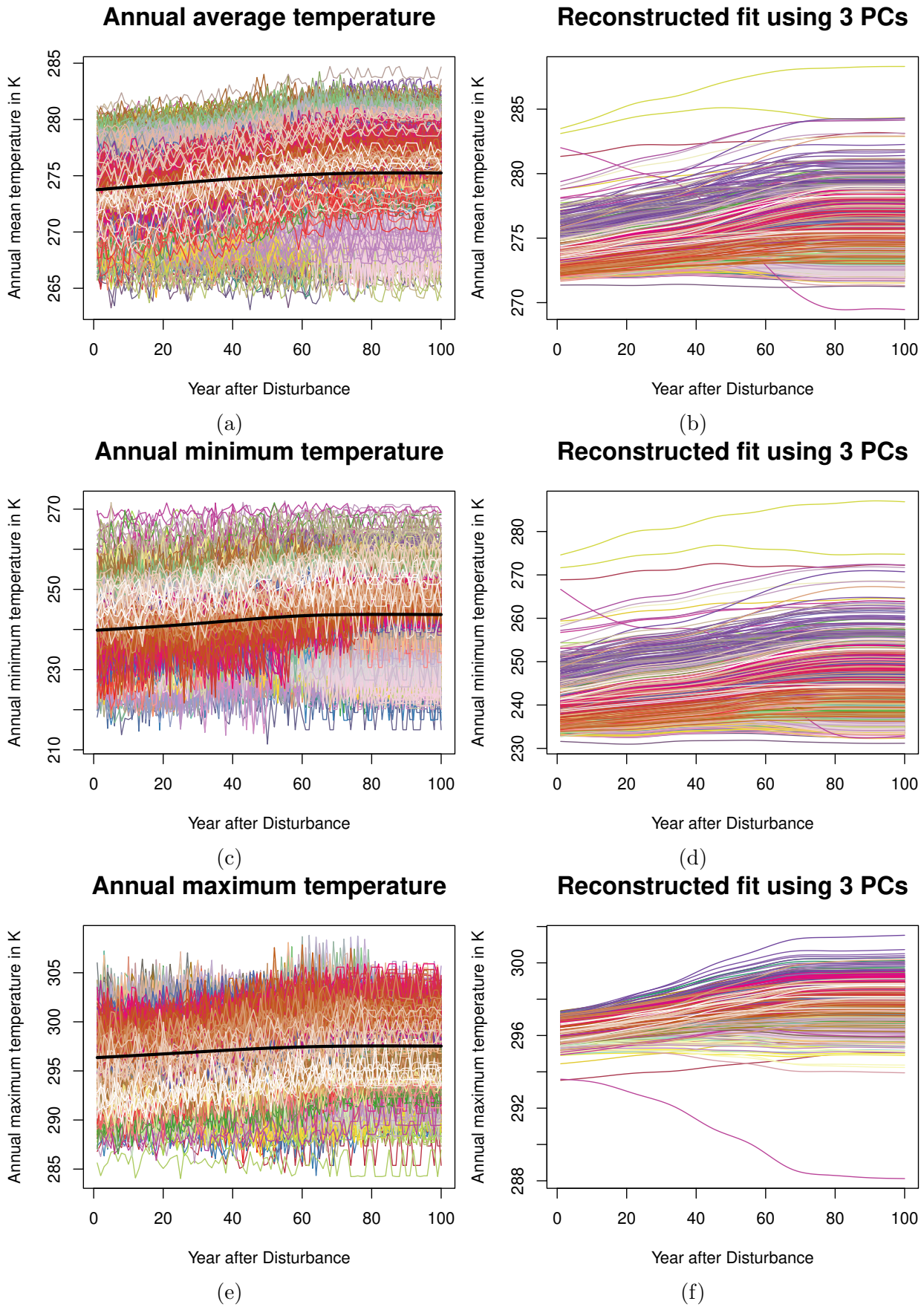
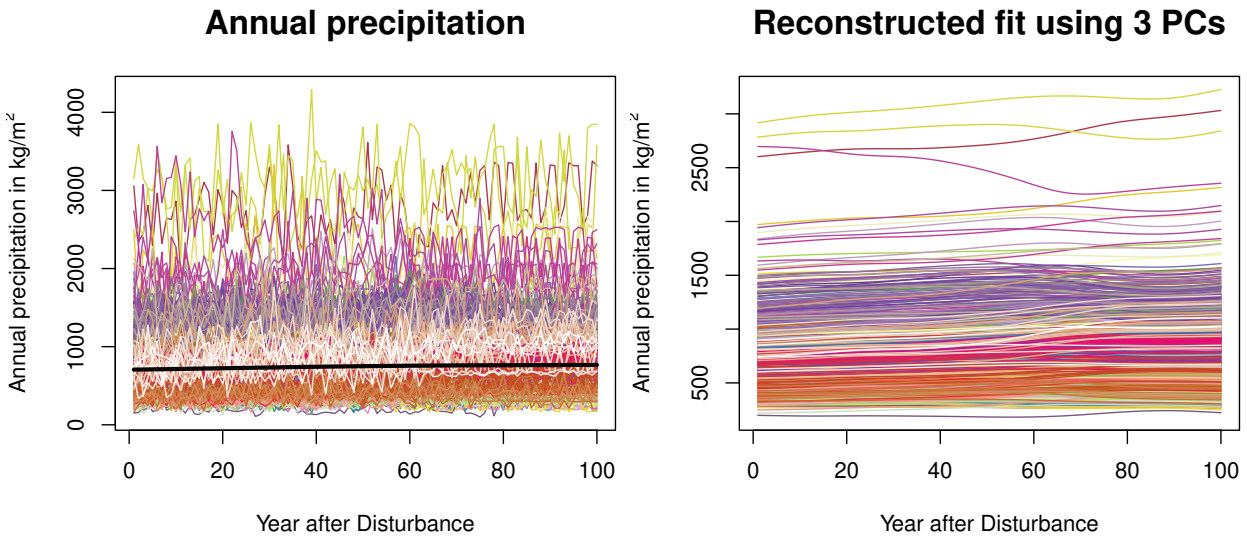


Figure 51: Original functional fit (left) and reconstructed curves using three PCs (right) for all three temperature variables.



(a) (b)

Figure 52: Original functional fit (a) and reconstructed curves using three PCs (b) for the annual sum of precipitation.

in the variable selection process. The situation is more ambiguous when considering the functional response. Table 6 indicates that to capture at least 90% of the variability in the data, the model should include at least the first four PCs. Looking at the third and fourth PCs in Figure 28 and Figure 29 does not yield a substantial gain in information compared to the first two PCs (Figure 26 and Figure 27), which justifies considering only the first two PC scores in the model named `PC1` and `PC2`.

In order to evaluate which of the first PC scores and non-functional covariates should be considered in the model, Figure 53 shows the correlations between each possible predictor (with shortcuts explained in Table 4) and the two response variables `PC1` and `PC2`.

First, consider the linear relationships between the first PC scores of the functional covariates among themselves and between them and the response variables. Obviously, the minimum and maximum annual temperatures are highly correlated with the average annual temperature, and so is the annual precipitation. This implies that the climate variables should be included as MFPCA scores in the model (`PC1_climate`) rather than separately. Almost all first PC scores related to nitrogen uptake are highly correlated with `PC1`. This raises the question of whether aboveground carbon is driven by nitrogen uptake or vice versa. As the causal direction is unclear and nitrogen uptake is considered a response variable rather than an explanatory variable, all nitrogen uptake related variables are excluded from the model.

Furthermore, the matrix illustrates that the three ecological variables `initial_recruitment`, `recruitment_ten_years` and `previous_state` are partially highly correlated for each PFT. This implies that only one of these variables should be included in the model. In this approach, only the covariate `initial_recruitment` is retained, as it was found to best reflect

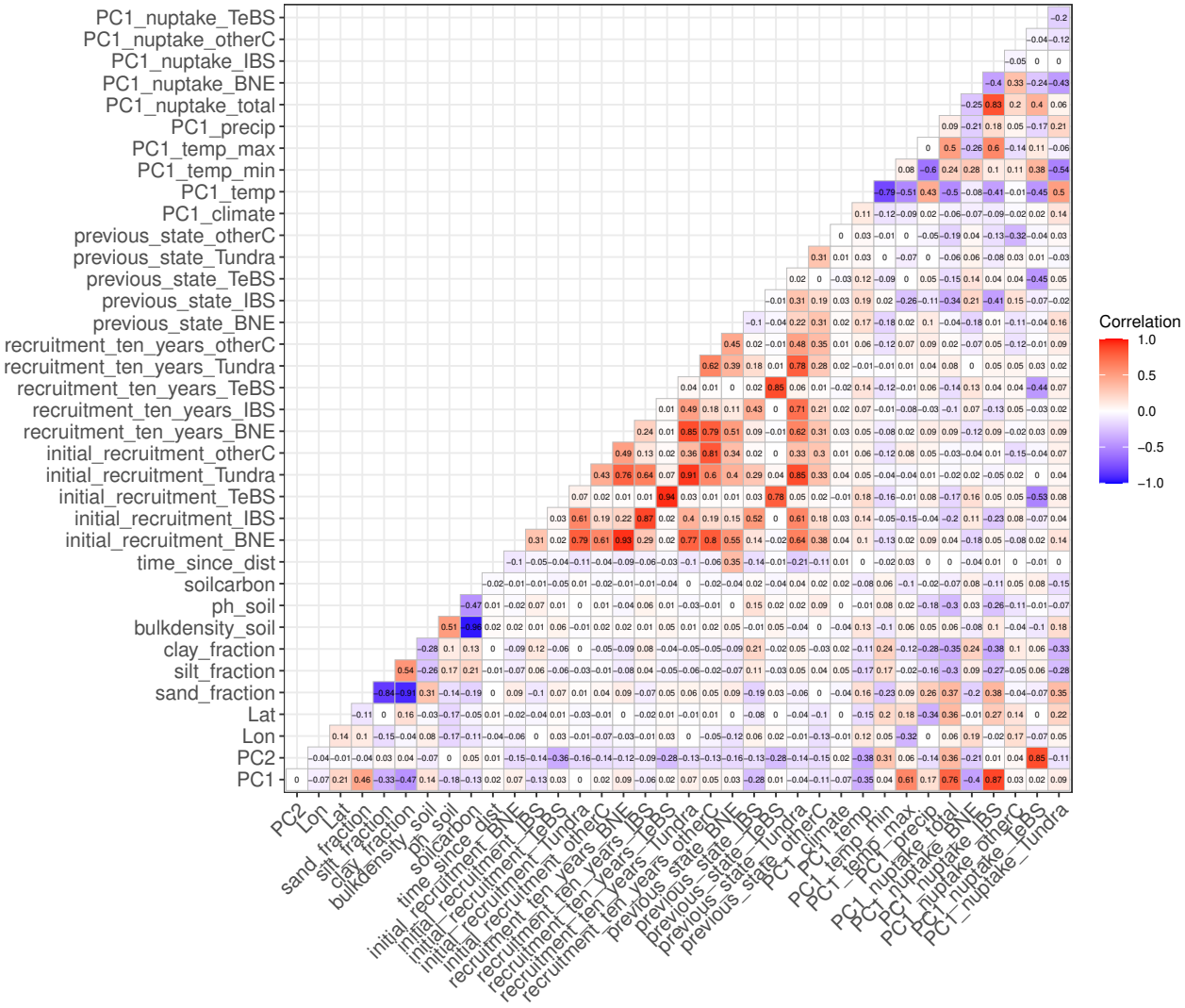


Figure 53: Correlation matrix for all possible covariates and the response variables.

the effect of all three variables together.

Note that the three soil composition variables, i.e., the fractions of sand, silt and clay, add up to 1. Therefore, the variable `clay_fraction` is omitted to avoid multicollinearity issues. Apart from all the exclusions justified above, all other covariates are used for modelling.

5.3 Model Setup and Evaluation

The variable configuration derived above results in the following model formula:

$$
\begin{aligned}
\begin{pmatrix} \text{PC1} \\ \text{PC2} \end{pmatrix} = & \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \end{pmatrix} + \begin{pmatrix} f_1(\text{Lon}, \text{Lat}) \\ f_2(\text{Lon}, \text{Lat}) \end{pmatrix} \\
& + \begin{pmatrix} \beta_{1,1} \\ \beta_{1,2} \end{pmatrix} \cdot \text{sand\_fraction} + \begin{pmatrix} \beta_{2,1} \\ \beta_{2,2} \end{pmatrix} \cdot \text{silt\_fraction} \\
& + \begin{pmatrix} \beta_{3,1} \\ \beta_{3,2} \end{pmatrix} \cdot \text{bulkdensity\_soil} + \begin{pmatrix} \beta_{4,1} \\ \beta_{4,2} \end{pmatrix} \cdot \text{ph\_soil} + \begin{pmatrix} \beta_{5,1} \\ \beta_{5,2} \end{pmatrix} \cdot \text{soilcarbon} \\
& + \begin{pmatrix} \beta_{6,1} \\ \beta_{6,2} \end{pmatrix} \cdot \text{ScenarioSSP1-RCP2.6} + \begin{pmatrix} \beta_{7,1} \\ \beta_{7,2} \end{pmatrix} \cdot \text{ScenarioSSP3-RCP7.0} \\
& + \begin{pmatrix} \beta_{8,1} \\ \beta_{8,2} \end{pmatrix} \cdot \text{ScenarioSSP5-RCP8.5} + \begin{pmatrix} \beta_{9,1} \\ \beta_{9,2} \end{pmatrix} \cdot \text{time\_since\_dist} \\
& + \begin{pmatrix} \beta_{10,1} \\ \beta_{10,2} \end{pmatrix} \cdot \text{initial\_recruitment\_BNE} + \begin{pmatrix} \beta_{11,1} \\ \beta_{11,2} \end{pmatrix} \cdot \text{initial\_recruitment\_IBS} \\
& + \begin{pmatrix} \beta_{12,1} \\ \beta_{12,2} \end{pmatrix} \cdot \text{initial\_recruitment\_otherC} + \begin{pmatrix} \beta_{13,1} \\ \beta_{13,2} \end{pmatrix} \cdot \text{initial\_recruitment\_TeBS} \\
& + \begin{pmatrix} \beta_{14,1} \\ \beta_{14,2} \end{pmatrix} \cdot \text{initial\_recruitment\_Tundra} \\
& + \begin{pmatrix} \beta_{15,1} \\ \beta_{15,2} \end{pmatrix} \cdot \text{PC1\_climate} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}
\end{aligned}
\tag{16}
$$

Here, all functional and non-functional components enter the model linearly, except for the location, which is modelled as a bivariate additive effect of longitude and latitude. Details on that are derived in Section 6. The covariate `Scenario` is included as a categorical variable.

The model is fitted on a 80% training data set using standard packages in R. In order to evaluate the general model fit, Figure 54 shows the model residuals plotted against the fitted values (left) and Q-Q plots (right) for both first (first row) and second (second row) PC scores on the training data set. Usually, the residuals are expected to be randomly distributed around the horizontal axis at zero. However, an examination of the figure reveals a pattern that deviates from this norm, especially for the first PC. The structure of the model is a consequence of the data situation. The model is based on MFPCA scores, which are usually unbounded. However, the MFPCA is performed on proportions of aboveground carbon, which are limited to the interval $[0, 1]$. This, in turn, imposes constraints on the resulting

MFPCA scores, thereby providing an explanation for the structured residuals. Note that these constraints lead to bounds on the MFPCA scores, but the exact bounds are unknown. To address this structure in the first PC scores, a transformation of the predicted MFPCA scores is necessary. Therefore, Figure 55 shows the predicted first PC scores plotted against the true ones (a) on the training data set, which underlines the structure in the predictions. The goal is now to find a function that fits the data's pattern. The overall shape of a possible transformation function resembles to a **Generalized Logistic Function**:

$$\text{glf}(x) = \frac{A}{(1 + e^{-k(x-x_0)})} + C.$$

The parameters A, k, x_0 and C need to be set or derived. Therefore, manually chosen values are used as starting parameters to determine non-linear least-squares estimates of

$$\hat{\texttt{PC1}} \sim \text{glf}(\texttt{PC1}),$$

where $\hat{\texttt{PC1}}$ denotes the estimated first PC scores from the model. The resulting curve and its corresponding estimated parameters is shown in Figure 55b. With that, and the inverse of the generalized logistic function given by

$$\text{glf}^{-1}(y) = x_0 - \frac{1}{k} \log\left(\frac{A}{y - C} - 1\right),$$

the first PC scores are transformed into

$$\texttt{PC1}_{\texttt{trafo}} = \text{glf}^{-1}(\texttt{PC1}).$$

Refitting the previous model with the transformed response yields nearly unstructured predicted versus true first PC scores as visualized in Figure 56. Equivalently, Figure 57 displays the transformed residuals. The structure in question has almost entirely disappeared. In particular, it is still present at the limit ($x = 6$), indicating that the fit of the generalized logistic function is not optimal at that point. The addition of further parameters to the function in order to enhance flexibility did not result in an improved overall fit. Consequently, all subsequent analyses are based on the above-presented approach. Note that the structure in the second PC scores (right plot of Figure 57) is less pronounced, yet demonstrates a tendency to overestimate low PC 2 scores.

In order to evaluate the model's capacity to generalize effectively to previously unseen data, Figure 58 illustrates the comparison between the predicted and actual PC scores for the test data set, comprising 20% of the original data set. As a result of the transformation, the data exhibits no evident structure and is predominantly distributed randomly around the line of equality. This is an indicator for an appropriate model fit.

To provide a final impression of the model's predictive capacity for randomly selected example locations of the test set, Figure 59 depicts the true functional fit and the predicted fits for four distinct example grid cells, one for each scenario. In this instance, the predicted MFPCA scores are back-transformed to functions utilising the predicted first two PC scores and the

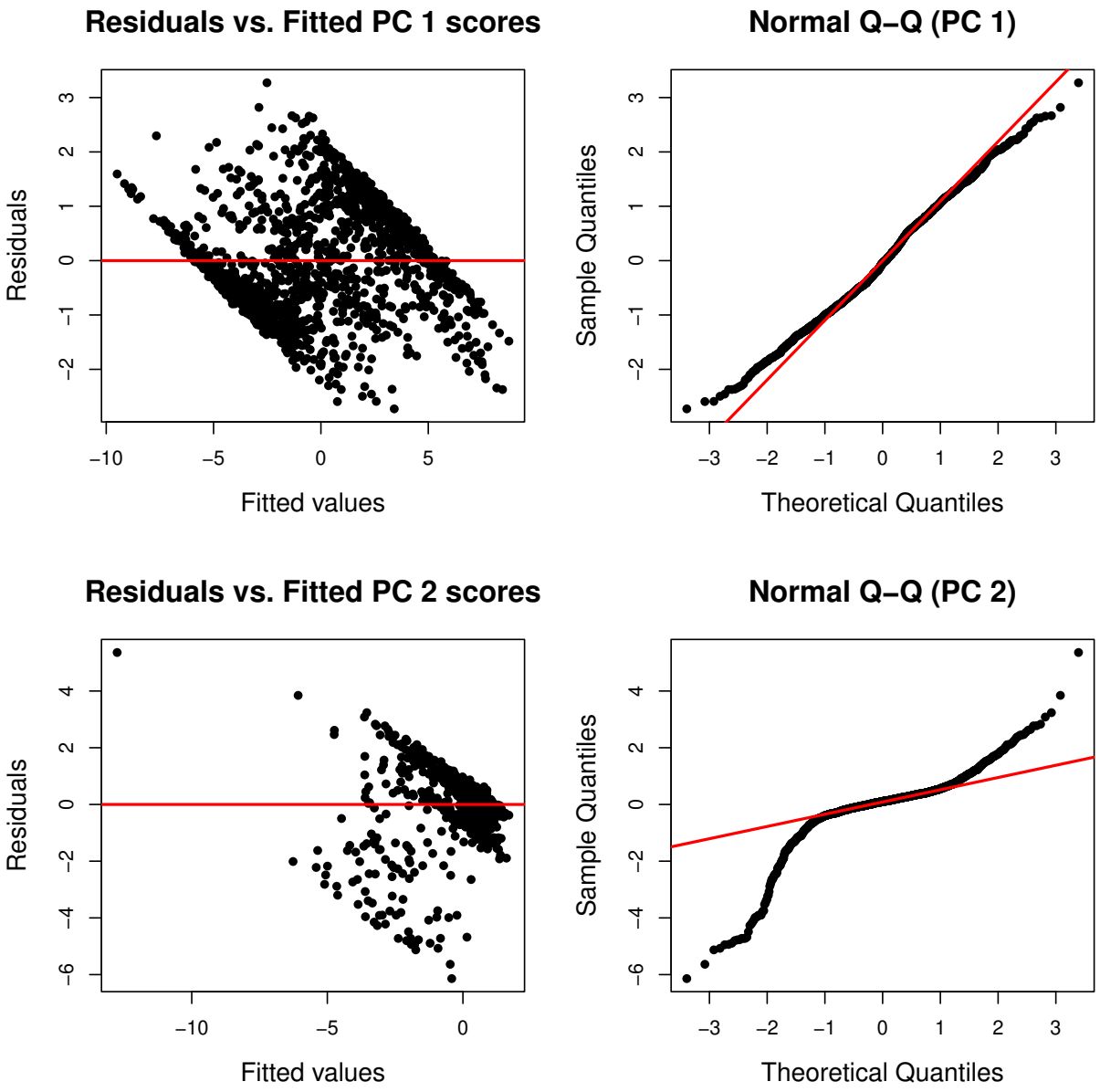
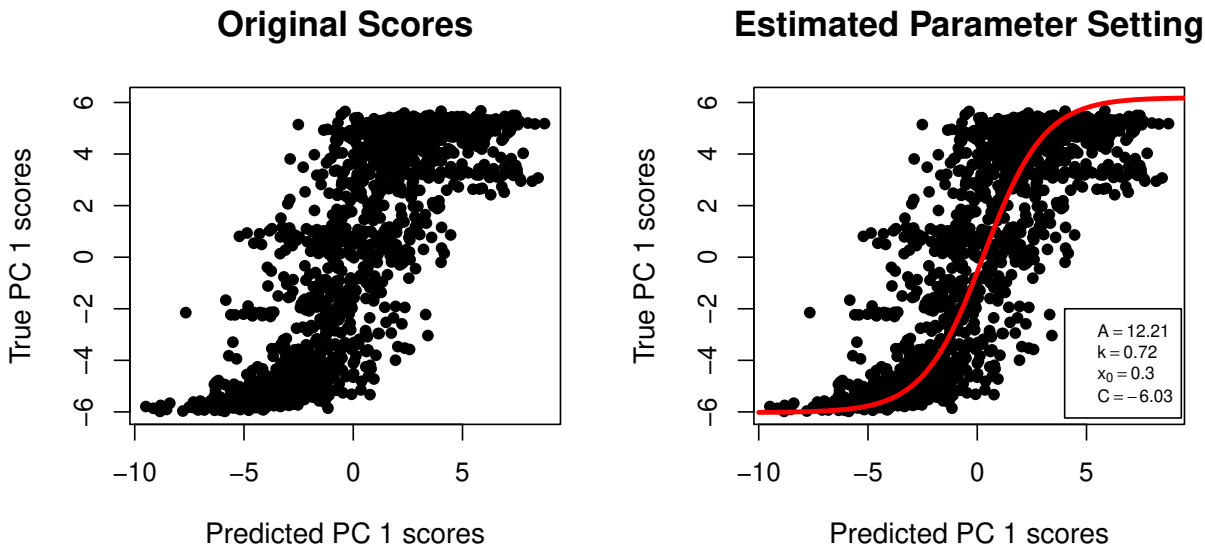


Figure 54: Model residuals and Q-Q plots for both predicted PC scores on the training data set.



(a) (b)

Figure 55: Predicted first PC scores and true ones (a) and fitted generalized logistic function (b).

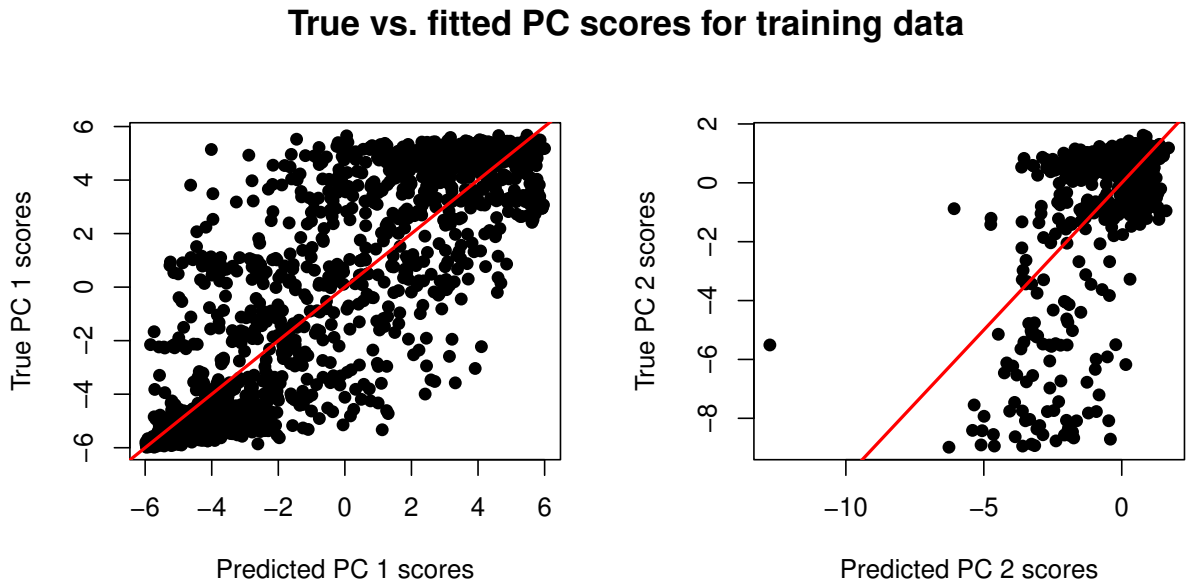


Figure 56: Predicted transformed first PC (a) and second PC scores (b) against the corresponding true PC scores for the training data.

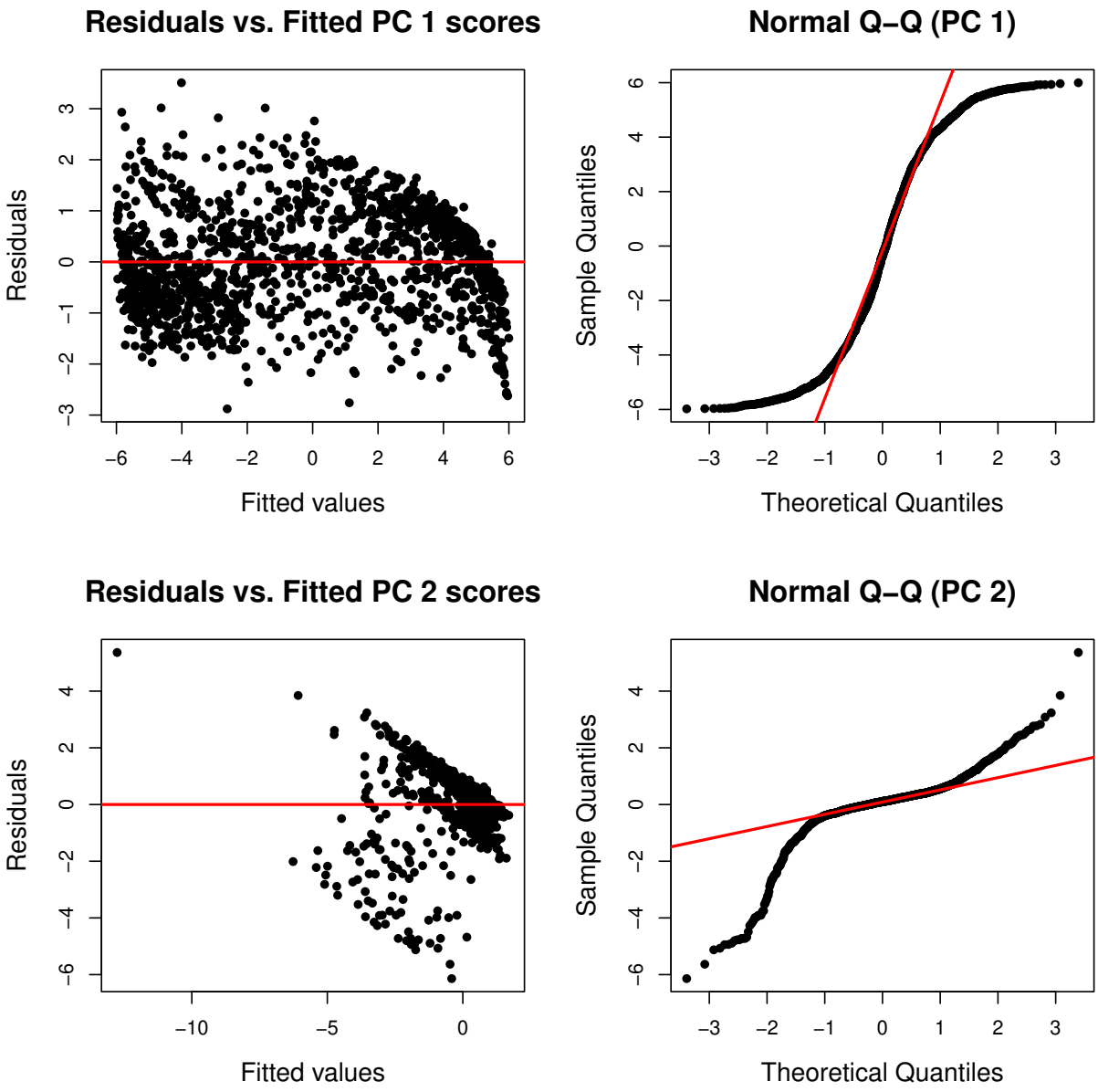


Figure 57: Model residuals and Q-Q plots for both transformed predicted first PC and predicted second PC scores.

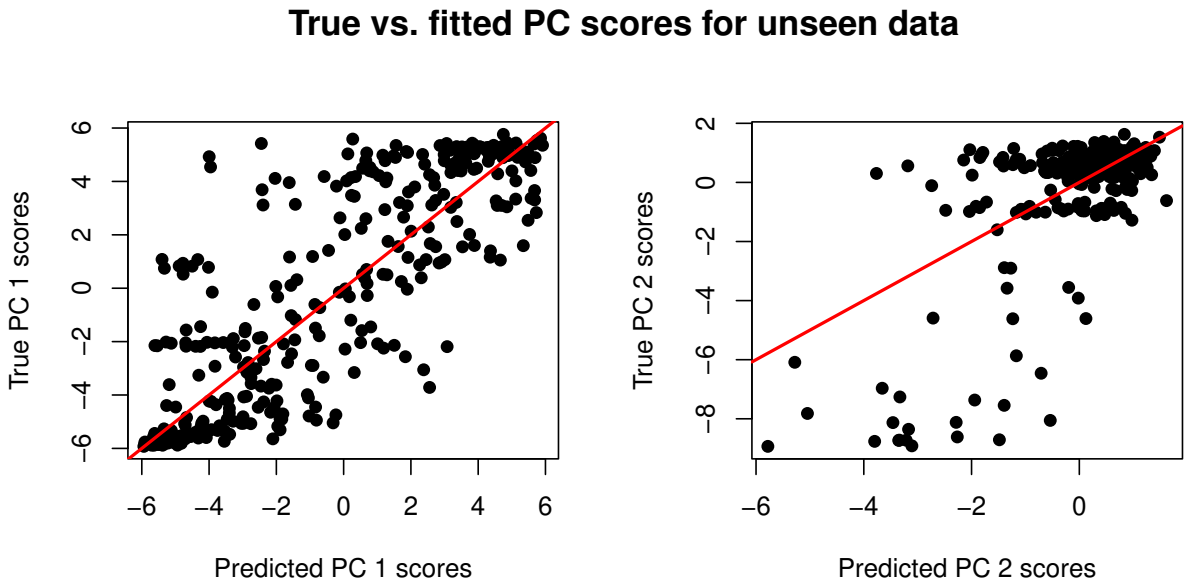


Figure 58: True transformed first PC (a) and second PC scores (b) plotted against predicted ones on the test data set.

derived basis functions. A comparison of the true and the predicted functions demonstrates an appropriate fit, with the dominant PFT being met in all four cases. However, it must be acknowledged that this figure offers only a limited insight into the model's performance on unseen data.

In conclusion, the model performance indicates that the methodology presented in this section is a valid approach to modelling the proportions of aboveground carbon for five PFTs simultaneously. The interpretation of the model is presented in detail in the following chapter.

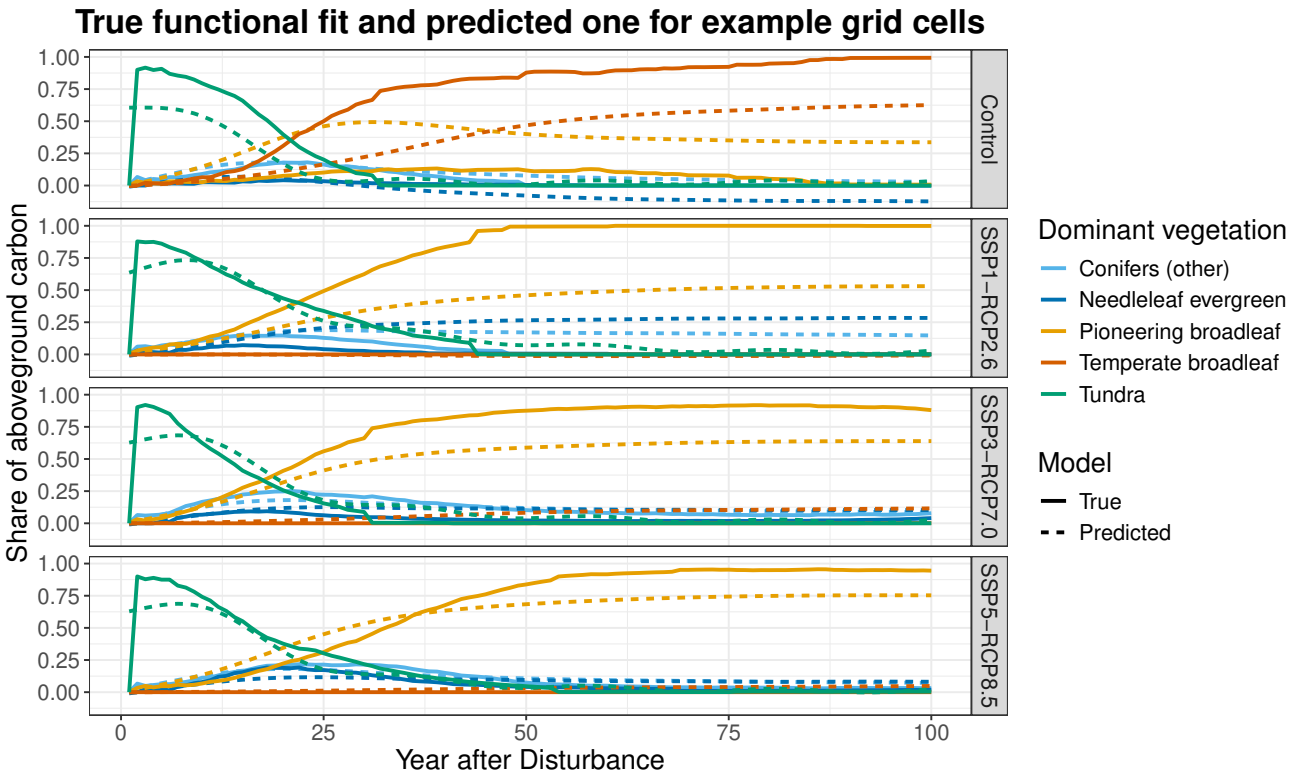


Figure 59: True functional fit and predicted functional fit for four different example grid cells, one for each scenario.

6 Results

In order to understand the estimated model from Section 5, this chapter starts with some guidelines on how to interpret the estimated parameters when modelling PC scores, followed by a detailed description of the results of the model.

6.1 Guidelines on Interpretations

Interpreting the results of the model involves two steps which need to be considered: the transformation of the response variable `PC1` using a generalized logistic function and correctly interpreting the PC scores according to all five PFTs. To address the first concern, assume without loss of generality that one covariate $\mathbf{x}$ is used to fit the model only, i.e.,

$$\begin{aligned} \texttt{PC1} &= \text{glf}(\beta_0 + \beta \cdot \mathbf{x}) \\ &= \frac{A}{1 + e^{-k(\beta_0 + \beta \cdot \mathbf{x} - x_0)}} + C. \end{aligned}$$

Given the non-linear nature of the data transformation in question, interpreting the parameter estimate is a more complex process than it would be for a linear model. By considering the structure of the generalized logistic function, it can be seen that for an increase of $\mathbf{x}$ by one unit, it follows that:

$$\begin{aligned} \beta > 0 &\Rightarrow \beta_0 + \beta \cdot \mathbf{x} \text{ increases} \\ &\Rightarrow 1 + e^{-k(\beta_0 + \beta \cdot \mathbf{x} - x_0)} \text{ decreases} \\ &\Rightarrow \frac{A}{1 + e^{-k(\beta_0 + \beta \cdot \mathbf{x} - x_0)}} + C \text{ increases} \\ &\Rightarrow \hat{\texttt{PC1}} \text{ increases.} \end{aligned}$$

This leads to the conclusion for the model setup in Equation 16 that parameter estimates above 0 generally increase the estimated `PC1` scores, while parameter estimates below 0 decrease them. Note that the same applies to the non-transformed `PC2` scores.

In order to draw conclusions about the vegetation composition that is described by the PCs, the visualizations of the first two PCs come into play. Consider again Figure 26 and Figure 27. The first PC captures vegetation dynamics deviating from the mean for all PFTs but *temperate broadleaf*, which is mostly covered by the second PC. For instance, for *needleleaf evergreen*, high first PC scores in that setting indicate above-average proportions of above-ground carbon and reversed for low values. This means in particular, that for the estimated coefficients for the first PC it holds that:

$$\begin{aligned} \beta > 0 \Rightarrow \hat{\texttt{PC1}} \text{ increases} \Rightarrow &\begin{cases} \oplus & \textit{Needleleaf evergreen, conifers (other)} \text{ and } \textit{tundra} \\ \ominus & \textit{Pioneering broadleaf} \end{cases} \\ \beta < 0 \Rightarrow \hat{\texttt{PC1}} \text{ decreases} \Rightarrow &\begin{cases} \oplus & \textit{Pioneering broadleaf} \\ \ominus & \textit{Needleleaf evergreen, conifers (other)} \text{ and } \textit{tundra} \end{cases} \end{aligned}$$

The interpretation of the second PC follows the same procedure:

$$\beta > 0 \Rightarrow \hat{\texttt{PC2}} \text{ increases} \Rightarrow \begin{cases} \oplus \textit{ Needleleaf evergreen, pioneering broadleaf} \\ \quad \text{and } \textit{conifers (other)} \\ \ominus \textit{ Temperate broadleaf} \end{cases}$$

$$\beta < 0 \Rightarrow \hat{\texttt{PC2}} \text{ decreases} \Rightarrow \begin{cases} \oplus \textit{ Temperate broadleaf} \\ \ominus \textit{ Needleleaf evergreen, pioneering broadleaf} \\ \quad \text{and } \textit{conifers (other)} \end{cases}$$

With this approach, it is now possible to interpret the estimated parameters in more detail in the following section.

6.2 Interpretation of the Parameter Estimates

The estimated coefficients, together with the corresponding t-values and whether or not the coefficients are significant at the 5% level, are summarised in Table 8 for the transformed first PC scores and in Table 9 for the second PC scores. The t-value serves as a measure of importance of a parameter estimate and is derived by the ratio of the coefficient and its standard error:

$$t_{i,k} = \frac{\hat{\beta}_{i,k}}{\text{se}(\hat{\beta}_{i,k})}, \quad i = 0, ..., 15, \; k = 1, 2.$$

A high absolute t-value indicates that the coefficient is significantly different from zero, suggesting that the predictor variable has a significant effect on the response variable. A major challenge of the proposed approach is the quantification of model uncertainty. When considering the model run on one patch, only the uncertainty of the multivariate regression model is taken into account when considering confidence intervals for estimated coefficients, neglecting the MFPCA as an additional source of uncertainty in the model. In order to infer appropriate confidence intervals for the estimated parameters, one approach is to use all 25 available patches for modelling. Since the patches represent 25 realizations of each grid cell, the previously derived model can be refitted to each patch, including deriving MFPCA scores for both functional responses and covariates, and transforming the first response variable by a generalized logistic function. Note that, as expected, the PCs for all patches capture very similar dynamics in vegetation recovery, as do the climate covariate curves. This procedure yields 25 parameter estimates for each linear predictor in the model, which serve as the basis for estimating 95% confidence intervals. A covariate in the model is considered to have a significant effect on the responses if and only if it holds that

$$|t_{j,k}| \geq z_{0.975} = |z_{0.025}| \approx 1.96, \quad j = 0, ..., 15, \; k = 1, 2.$$

where z_α describes the α-quantile of the standard Gaussian distribution. The effects in Table 8 and Table 9 are based on all 25 patches, that is, the parameters are averages over

	Covariate	Parameter	t-Value	Significance
	Intercept	−3.17	−22.72	⊖
Soil	sand_fraction	12.64	98.66	⊕
	silt_fraction	6.31	35.55	⊕
	bulk_density_soil	−1.37	−13.90	⊖
	ph_soil	−0.22	−6.63	⊖
	soilcarbon	−0.16	−18.92	⊖
Ecological	time_since_dist	0.0001	2.92	⊕
	initial_recruitment_BNE	0.002	8.06	⊕
	initial_recruitment_IBS	0.002	9.88	⊕
	initial_recruitment_otherC	−0.007	−9.61	⊖
	initial_recruitment_TeBS	0.003	3.00	⊕
	initial_recruitment_Tundra	−0.0002	−0.60	
Climate & Scenario	PC1_climate	0.00006	7.92	⊕
	ScenarioSSP1-RCP2.6	−1.83	−108.58	⊖
	ScenarioSSP3-RCP7.0	−2.24	−92.61	⊖
	ScenarioSSP5-RCP8.5	−2.59	−92.27	⊖

Table 8: Parameters estimated by the additive model for the transformed first PC scores. Significant negative and positive effects are indicated by ⊖ and ⊕ respectively.

all 25 estimates, and the t-values are calculated by

$$\hat{t}_{i,k} = \frac{\frac{\sqrt{25}}{25}\sum_j \beta_{j,k}}{\sqrt{\frac{1}{24}(\beta_{i,k} - \frac{1}{25}\sum_j \beta_{j,k})}}, \quad i = 0, ..., 15,\ k = 1, 2.$$

In the following, the effects are analyzed in more detail for each covariate group: location, soil properties as well as ecological and climate covariates and scenario.

Effect of the location

As the location is included in the model as a bivariate additive effect, the interpretation of its effect is more complex than for the linearly included variables. As different locations are disturbed in different patches, it is difficult to derive a common estimate for each disturbed location. Therefore, Figure 60 shows an example visualisation of the smooth effects of longitude and latitude for both PC scores, i.e., the predicted values of the bivariate smooth term, for one patch only. For the first PC scores (a), locations in Europe and the northern and western parts of Asia and North America respectively show a positive effect, indicating higher proportions of *needleleaf evergreen, conifers (other)* and *tundra*. In contrast, the centre of North America and the southern part of Asia tend to have negative effects on first PC scores, implying more *pioneering broadleaf* than on average. The effects of location on the second PC (b) are less pronounced. There are some negative effects in the north-western part of Europe and in the western part of North America, resulting in more *temperate broadleaf* than on average in that region. The vegetation in the other parts is only slightly affected by the exact location of the disturbance.

	Covariate	Parameter	t-Value	Significance
	Intercept	0.10	0.60	
Soil	sand_fraction	1.45	11.32	⊕
	silt_fraction	−0.85	−3.88	⊖
	bulk_density_soil	−0.27	−3.38	⊖
	ph_soil	0.08	3.44	⊕
	soilcarbon	−0.03	−4.39	⊖
Ecological	time_since_dist	0.0002	3.73	⊕
	initial_recruitment_BNE	−0.0001	−1.01	
	initial_recruitment_IBS	−0.0001	−3.37	⊖
	initial_recruitment_otherC	−0.002	−5.78	⊖
	initial_recruitment_TeBS	−0.01	−3.37	⊖
	initial_recruitment_Tundra	−0.002	−11.48	⊖
Climate & Scenario	PC1_climate	0.000004	0.76	
	ScenarioSSP1-RCP2.6	−0.18	−9.62	⊖
	ScenarioSSP3-RCP7.0	−0.42	−23.35	⊖
	ScenarioSSP5-RCP8.5	−0.70	−30.93	⊖

Table 9: Parameters estimated by the additive model for PC 2 scores. Significant negative and positive effects are indicated by ⊖ and ⊕ respectively.

Effects of soil properties

Figure 61 shows the t-values of the linear covariates for both response variables, where red indicates positive effects and values in blue negative effects. Significance based on the 97.5%-quantile of the standard Gaussian distribution is indicated by vertical lines at −1.96 and +1.96. The larger the absolute t-value, the more important the variable is in explaining the PC scores.

First consider the effects of the soil covariates on the first PC scores shown in Figure 61a and Table 8. The effects of `silt_fraction` and `sand_fraction` are relative to `clay_fraction`, which is the reference category, and is therefore estimated within the intercept. These components show the two largest positive effects on the first PC scores, indicating that, for example, increasing `sand_fraction` by one unit while keeping `silt_fraction` constant leads to a decrease in clay and an increase in estimated `PC1_trafo` values. The same is applicable when the roles of sand and silt are reversed. In turn, considering the interpretation guidelines derived in Section 6.1, the proportions of *needleleaf evergreen, conifers (other)* and *tundra* increase, while the amount of *pioneering broadleaf* decreases. In contrast, the effects of sand and silt on the second PC values shown in Figure 61b and Table 9 are in the opposite direction. While increasing `sand_fraction` by one unit and maintaining `silt_fraction` leads to higher estimated `PC2` scores relative to `clay_fraction`, the effects are reversed for `silt_fraction` as the corresponding effect estimate is negative. This means that more sand leads to less *temperate broadleaf* than on average, while a higher `silt_fraction` leads to more *temperate broadleaf*, as Section 6.1 indicates.

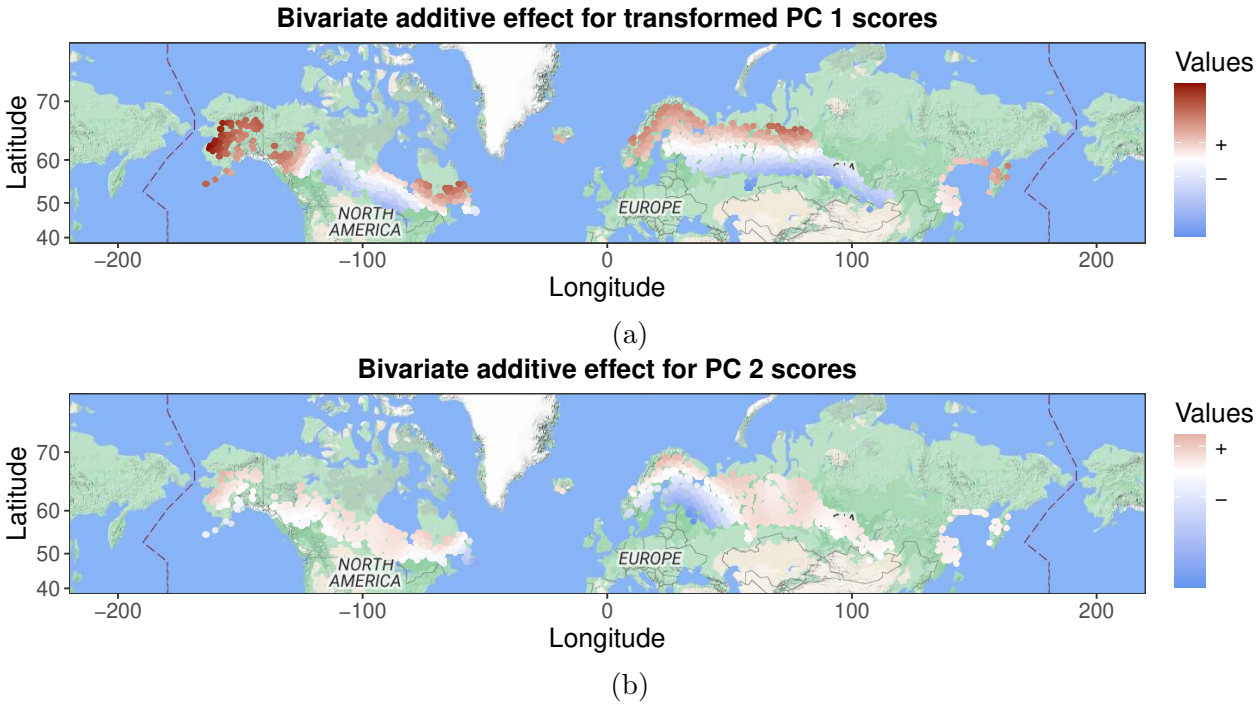
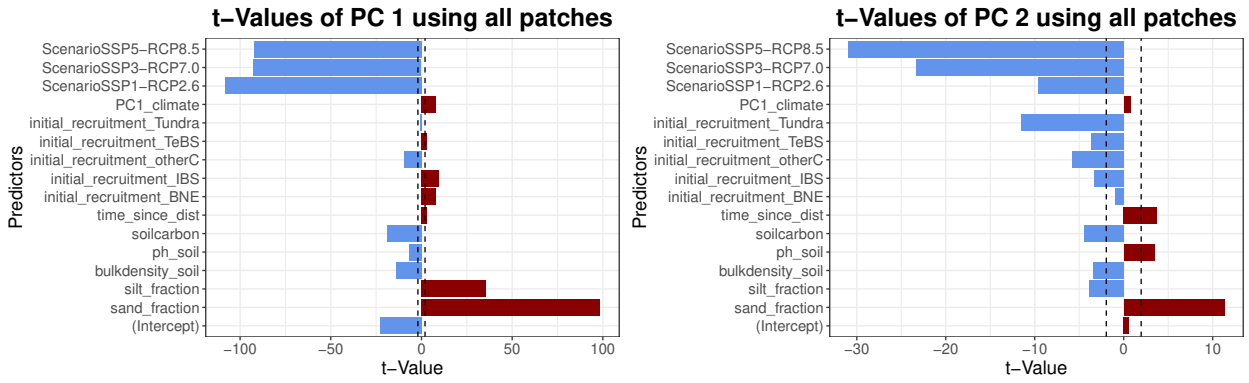


Figure 60: Bivariate additive effect of longitude and latitude for first PC (a) and second PC (b) scores.

In contrast to the soil composition, the three remaining soil covariates, i.e., `ph_soil` and `soilcarbon`, as well as `bulkdensity_soil`, appear to have a small negative effect on the estimated first PC scores, leading to above average proportions of *needleleaf evergreen, conifers (other)* and *tundra* per unit increase. The effects on the second PC scores (Table 9) show a change in sign for the pH value in water, but have similar effect sizes and significance. Unlike the first PC, the effect of `ph_soil` is positive, resulting in less *temperate broadleaf* at higher pH values. The effects of carbon content and bulk density are both negative, indicating more *temperate broadleaf* than on average for a unit increase in these covariates. Note that for both `PC1_trafo` and `PC2`, the effects of all soil properties are significant at 5%.

Effects of ecological covariates

As several ecological covariates were excluded during the variable selection process in subsection 5.2, only two ecological predictors remain in the model, namely `time_since_dist` and the initial number of new seedlings per PFT, modelled as five separate effects. Looking at Figure 61a and Table 8 shows a small positive effect on the first PC scores of the time since the last disturbance for an increase of one year, indicating above average proportions of *needleleaf evergreen, conifers (other)* and *tundra*. This effect is confirmed when looking at Figure 61b and Table 9, which also show a positive effect on the second PC values, resulting also in more *needleleaf evergreen* and *conifers (other)* than on average for an increase by one year. Note that whenever a predictor is shown to have a positive effect on both PC scores, their results are contradictory in one case: the first PC (Figure 26) shows less *pioneering broadleaf* than on average for $\beta > 0$, while the second PC (Figure 27) shows a higher pro-



(a) (b)

Figure 61: t-Values for both transformed first PC scores (a) and second PC scores (b).

portion of aboveground carbon of this PFT. The same applies to negative effects on both PC scores. It is important to relate the magnitude of the PC deviation to the mean curves with the estimated effect size in order to assess which effects are more important. In the case of the years since the last disturbance, the effect of the second PC is clearly predominant and therefore the proportion of *pioneering broadleaf* is above average for longer time periods. Note that the effect of `time_since_dist` is significant for both PC scores, but of minor importance since the corresponding t-values are quite small.

Intuitively, one would expect that a high number of new seedlings immediately after the disturbance would lead to higher levels of the respective PFT in the following years. This effect is not captured by all the PFTs considered in the model as Figure 61a indicates. While the positive effect of the number of new seedlings of *needleleaf evergreen* indicates more *needleleaf evergreen, conifers (other)* and *tundra* per unit increase in new seedlings, the effects for *pioneering broadleaf* and *conifers (other)* are somewhat counterintuitive. The former has a significant positive effect, leading to higher estimated `PC1_trafo` values and therefore below average *pioneering broadleaf* during the recovery period, but higher proportions of needleleafed trees. The effect is reversed for *conifers (other)*, which has a negative effect and is therefore an indicator of less needleleaf and more *pioneering broadleaf*. The first PC hardly affects *temperate broadleaf* and *tundra*, which is reflected in small positive and negative t-values, respectively. Note that all but the effect for *tundra* are significant at 5%. For the second PC scores, all estimated effects of the number of new seedlings are negative (see Table 9 and Figure 61b), with being largest for *tundra*. This means in particular that for all PFTs, an increase of the total number of new seedlings after disturbance by one seedling leads to more *temperate broadleaf* during recovery and less of the remaining tree species than on average. Here, all effects but the one of *needleleaf evergreen* are significant. Interestingly, according to the t-values, the effect of *tundra* is the most important one on the growth of *temperate broadleaf*.

The effect of the number of new seedlings immediately after the disturbance shows that tree growth is not simply given by the recruitment immediately after the disturbance, but that

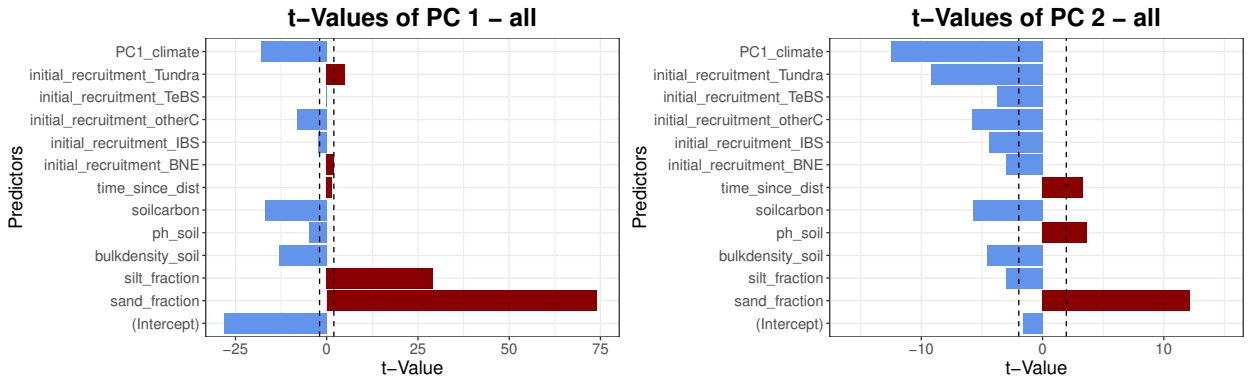
other mechanisms such as dispersal and displacement are important.

Effects of climate covariates and the scenario
All climate covariates, that is, the annual mean, minimum and maximum temperature and summed precipitation, are functional and are therefore included as PC scores derived by an MFPCA in the model (see Equation 16). To adequately interpret the estimated parameters for `PC1_climate`, the visualisation of the first PC in Figure 50a is required. On the one hand, Figure 61a shows a significant positive effect on the first PC scores. This means that an increase in climate PC scores, which themselves indicate above-average annual mean, minimum and maximum temperatures and precipitation throughout the study period (see Figure 50a), leads to lower estimated `PC1_trafo` values and therefore more *pioneering broadleaf* than on average and fewer needleleafed tree species. On the other hand, Figure 61b shows a small, non-significant positive effect on the second PC scores, indicating less *temperate broadleaf* than on average for higher temperatures and more precipitation. This result is counterintuitive, but as the effect on the second PC scores is not significant, the result should not be overinterpreted.

Finally, the effect of the scenario entered into the model as a categorical variable, with the control scenario as the reference category, is examined. The parameter estimates in Table 8 show that all three warming scenarios have a significant negative effect on the first PC scores, indicating more *pioneering broadleaf* and less needleleafed trees with increasing radiative forcing. Looking at the importance in terms of t-values in Figure 61a, the effects of each scenario are among the most important, with SSP1-RCP2.6 having the largest t-value in absolute terms of all three warming scenarios. Similarly, all scenarios have a negative effect on `PC2`, as indicated by Table 9 and Figure 61b. Here the magnitude of the effect estimate and its importance increase with radiative forcing. Consequently, the proportion of *temperate broadleaf* increases with the intensity of the scenario.

One hypothesis as to why the effect of the climate covariates is so small and even non-significant for the second response variable involves the effect of the scenario in the model. The climate induced in the model by the ISIMIP data used as input to LPJ-GUESS incorporates (among other things) changing temperature and precipitation dynamics. By including both effects for climate covariates and the scenario, these dynamics are superimposed in the model and therefore their parameter estimates interfere with each other. Nevertheless, it makes sense to include `PC1_climate` as separate covariate to disentangle its separate effect. This is further examined in Section 6.3.

To conclude, especially the sand and silt fraction as well as the scenario show to have a huge impact on the first PC scores, leading to higher proportions of *needleleaf evergreen, conifers (other)* and *tundra* after disturbances as well as more *pioneering broadleaf* than on average, respectively. On the second PC scores, the effect of the scenario together with the number of new seedlings immediately after disturbance of PFT *tundra* have a strong negative effect, leading to more *temperate broadleaf* during recovery.



(a) (b)

Figure 62: t-Values for both first PC scores (a) and second PC scores (b) for the model fitted without the covariate `Scenario` on all climate scenarios combined.

6.3 Scenario-Based Models

A limitation of the proposed model is that it is not suitable for truly disentangling scenario-wise effects of the covariates. The modelling approach includes a categorical effect for the scenario, but to obtain parameter estimates for each scenario, interactions with each predictor of interest would be required. Since the model in Equation 16 includes the effects of twelve linear covariates (excluding the scenario), including these interactions would lead to the estimation of $12 \cdot 4 + 3 = 51$ parameter coefficients, and interpretation would be more challenging. Therefore, an alternative approach is proposed, which involves refitting the previous model to each scenario separately, i.e., excluding the effect of `Scenario` and fitting the model to scenario-wise data only. This results in four separate models for the four climate scenarios, which is again run on all 25 patches for an adequate uncertainty assessment. Here, the MFPCA for the climate curves is re-fitted to scenario-wise recovery trajectories. Note that the data sets per scenario still contain about 350 observations each, indicating a valid modelling setup. In order to compare the effects of the individual models with the overall model, Figure 62 shows the t-values of the model fitted to all scenarios together, but excluding the effect of `Scenario`. Most of the effects change only slightly, with one exception: the effect of the climate PC scores `PC1_climate` is now significantly negative for both response variables, indicating higher proportions of broadleafed trees during recovery for increased temperature values and more precipitation. This result is in support of the hypothesis in the previous section that the climate covariates are somewhat captured by the effects of each scenario.

Figure 63 shows the t-values for the scenario-wise models for the first PC scores. Analysing the differences yields the following results for each covariate group:

Soil covariates
All soil composition covariates and soil properties show effects in the same direction for all four scenarios, but vary in importance and significance at 5%. The soil composition, i.e., the covariates `sand_fraction` and `silt_fraction`, show large significant positive effects in all

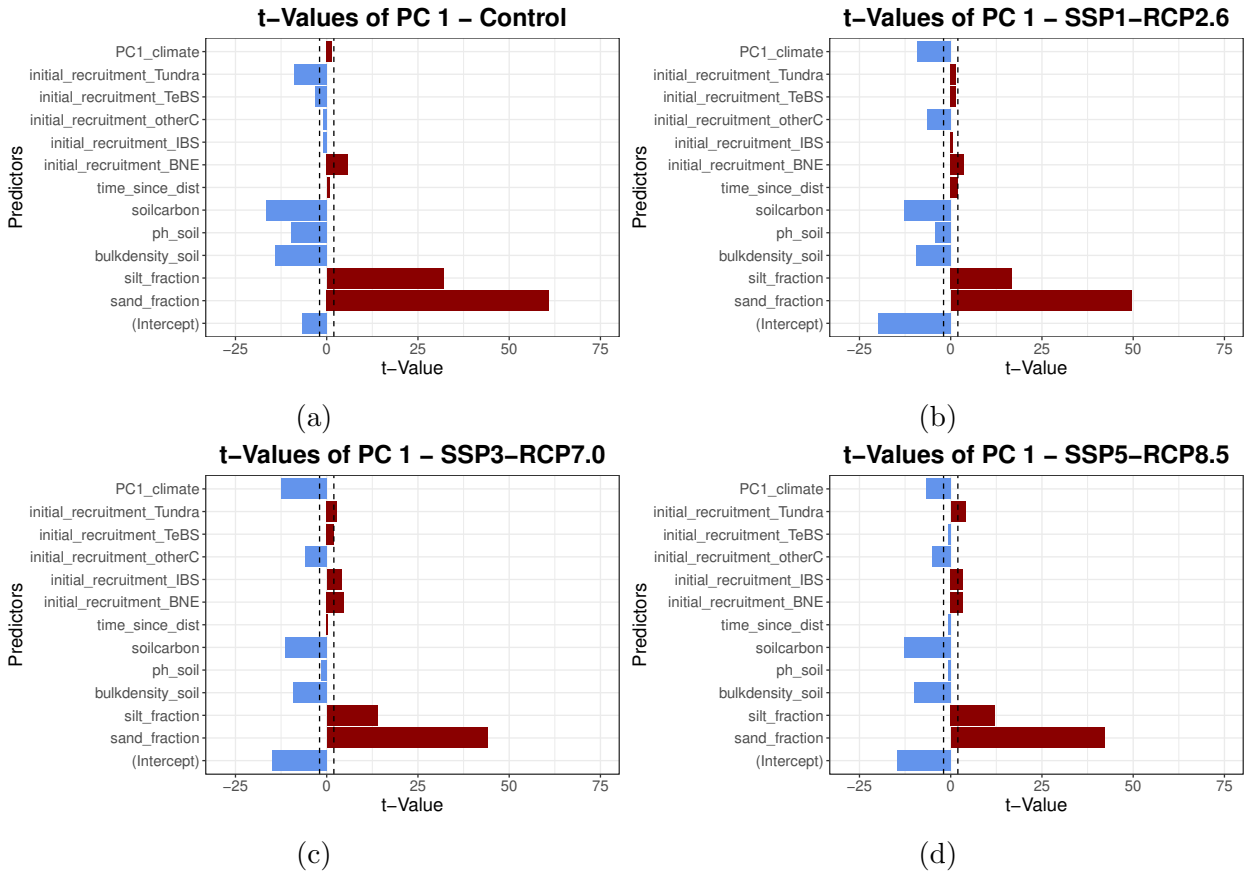


Figure 63: t-Values for transformed first PC scores and each scenario.

four scenarios, with the largest effect in the control scenario. As this scenario is mainly dominated by needleleafed trees, it could be hypothesised that soil composition is more important for these tree species than for PFT *pioneering broadleaf*. Similarly, the effects of `bulkdensity_soil` and `soilcarbon` are most important in the control scenario and show smaller but still significant effects in the three warming scenarios. For the last soil property, `ph_soil`, there is a dynamic within the scenarios: while its effect is moderately negative and significant in the control scenario, its importance decreases in the SSP1-RCP2.6 scenario and is no longer significant but still negative in the two most extreme scenarios.

Ecological covariates

The effect of years since last disturbance appears to be positive for the whole data set (Figure 62a) and for all four individual models except for the SSP5-RCP8.5 scenario. In general, this covariate is of minor importance for all setups, as it is not significant in any of the models considered in this section. In contrast, the effects of the number of new seedlings vary greatly with the different climate conditions. In the overall model, only the effects of *tundra* (positive) and *pioneering broadleaf* and *conifers (other)* (both negative) have significant effects on the first PC scores. While the effect of PFT *needleleaf evergreen* is barely insignificant in the overall model, it has a moderately positive significant effect in all scenarios, indicating more needleleafed trees during recovery. For *pioneering broadleaf* and *tundra*

there is a dynamic in the importance of their effects: while both have a negative effect in the control scenario, their effects turn positive in the SSP1-RCP2.6 scenario and even become positively significant in the two most extreme scenarios. This suggests that as the climate warms, higher levels of *pioneering broadleaf* and *tundra* at the beginning of the recovery period lead to more coniferous tree species in general. In contrast, the effect of *conifers (other)* remains negative and significant for all five models considered in this section. Finally, the effect of the number of new seedlings of *temperate broadleaf* alternates between the scenarios, being negative in the control and SSP5-RCP8.5 scenarios and positive in the remaining two scenarios. The effect is only significant for the control and the SSP3-RCP7.0 scenario.

Climate PC scores
In the overall model (Figure 62a) the effect of `PC1_climate` is significantly negative, and this is supported by the three warming scenarios illustrated in Figure 63. Only the control scenario shows a small non-significant positive effect. This indicates that in a warmer climate, as reflected in the three warming scenarios, the importance of temperature and precipitation changes is greater than in the pre-industrial climate and has a positive influence on the occurrence of broadleafed trees.

Figure 64 shows the equivalent visualizations of scenario-wise t-values for second PC scores. Again, the differences between the scenarios are broken down into groups of covariates:

Soil covariates
Similar to the first PC scores, all soil properties show the same direction of effect for all scenario-based models and the overall model in Figure 62b, here with the exception of `ph_soil`. While the sand fraction shows a large positive significant effect in all five models considered, the effect size and significance of silt differs between scenarios. Although it has a negative effect in all models, it is only significant for the SSP1-RCP2.6 scenario and the overall model. However, its importance is much lower than that of its counterpart, sand. Indeed, the SSP2-RCP2.6 scenario appears to have significant effects for all soil properties, as does the overall model, including the covariates `bulkdensity_soil`, `ph_soil` and `soilcarbon`, whereas none of these effects are significant in the control scenario. PH in water is positively significant in scenario SSP5-RCP8.5, while in SSP3-RCP7.0 the covariates `soilcarbon` and `bulkdensity_soil` are significantly negative. Overall, there is no clear pattern in parameter magnitude as a consequence of climate warming.

Ecological covariates
The effect of the number of years since the last disturbance happens to be positively significant in the overall model in Figure 62b as well as for both the control scenario and the SSP5-RCP8.5 scenario. It is consistently positive for all models fitted in this section, indicating less *temperate broadleaf* for a greater distance from the previous disturbance. In the model fitted to all scenarios together, the effect of the number of new seedlings immediately after the disturbance is significant and negative for all five PFTs, which is not the case in the scenario-wise models. While in the control, SSP3-RCP7.0 and SSP5-RCP8.5 scenarios the effects of the PFTs *conifers (other), temperate broadleaf* and *tundra* are significantly negative and decrease in importance with increasing radiative forcing, the effects of the

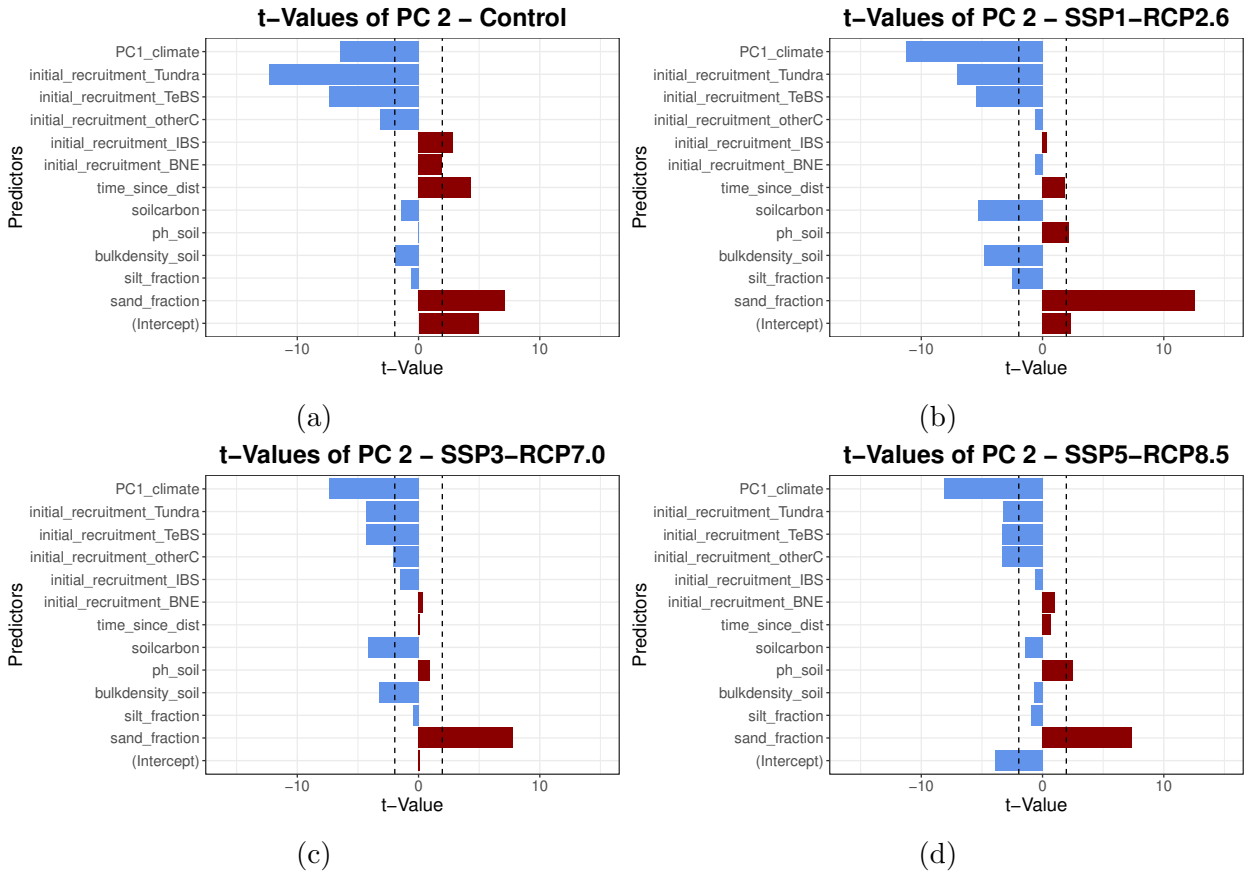


Figure 64: t-Values for PC 2 scores and each scenario.

remaining two PFTs are less stable and change in direction and significance. The effect of *needleleaf evergreen* is not significant in any scenario and the sign of the effect varies between scenarios. Similarly, the effect of *pioneering broadleaf* is significantly positive in the control scenario and remains positive in all other scenarios, but is no longer significant. In general, the effects of *tundra* and *temperate broadleaf* seem to have the most important impact on the occurrence of *temperate broadleaf* during recovery.

Climate PC scores

Finally, in the overall model in Figure 62b, the effect of `PC1_climate` is significantly negative and represents the most important effect in absolute terms. This holds for all scenarios except the control, suggesting that other covariates are more important for *temperate broadleaf* occurrence in this case. Overall, the effect is largest in SSP1-RCP2.6.

Altogether, the differences between the scenarios are highly dependent on the predictor in question. For some covariates, such as `initial_recruitment_Tundra`, a clear dynamic develops with increasing radiative forcing. The changes in other covariates are less straightforward, e.g., in `time_since_dist`. Overall, this scenario breakdown shows the differences in how climate warming affects the way the predictors interact with the recovery trajectories.

7 Discussion

The proposed approach highlights the applicability of functional data analysis (FDA) methods to simulated vegetation data under varying climatic conditions. This thesis has established a modelling framework for multidimensional recovery trajectories based on the concept of multivariate functional principal component analysis (MFPCA) as well as the combination of functional and additive regression techniques. In the following, the main findings of this thesis are summarized and an outlook on potential future research directions and analysis approaches is provided.

Summary
Section 2 began with theoretical background on the FDA building blocks used to establish the modelling framework. This included details on basis expansions, univariate and multivariate functional principal component analysis, as well as multivariate functional linear regression models (mFLR). For the purposes of this thesis, this model was further extended to include additive components. After details on the simulation process with LPJ-GUESS and on the climate scenarios considered in this project in Section 3, the input data – recovery trajectories indicating the proportion of aboveground carbon of five different PFTs at various locations in the boreal forest after disturbances between the years 2015 and 2040 and under four different climate projections – were first transformed into functional form using basis function expansions with B-splines in Section 4. The MFPCA approach makes use of these representations by performing a multivariate PCA on the corresponding basis coefficients. In that way, the functional PCA problem turns into a multivariate one, which is also the principal that underlies the functional regression methodology derived in Section 5. Instead of actually modelling functions, the approach proposed in this project takes the principal component (PC) scores derived by MFPCAs for both functional response variable and functional climate predictors as multivariate variables in the model, and includes additional multiple linear predictors to cover soil properties and ecological covariates, and an additive effect for the location. To account for the bounded structure of the PC scores resulting from the restriction of aboveground carbon proportions to the interval $[0, 1]$, the first PC scores were transformed using a generalised logistic function. This ensured an appropriate model fit.

The objectives of this thesis were firstly to find patterns in recovery trajectories and secondly to identify drivers for the observed vegetation composition after disturbance. The first goal was addressed by clustering the derived PC scores using a 4-means algorithm in Section 4.5. This revealed that vegetation recovery followed three distinct dominance patterns. terns: PFTs *pioneering broadleaf* and *temperate broadleaf*, as well as needleleafed trees. Looking at the scenarios in which the respective grid cells were disturbed, it was found that the milder climate favours the dominance of broadleaf tree species, while the recovery of the grid cells disturbed in the control scenario mainly follows the third regime. The regimes also show regional patterns, with the conifer regime concentrated in locations closer to the pole than the other two. A closer look at the other variables provided (soil properties, ecological and climatic variables) did not reveal any major differences between the dominance regimes induced by the clusters. In contrast, clustering of recovery trajectories at specific time points in the study period showed the robustness of clustering after a few years post-disturbance,

suggesting that the early years of recovery laid the foundation for tree establishment throughout the entire recovery.

In Section 6, the multivariate functional additive model identified the most important factors influencing vegetation composition during recovery from disturbance. An analysis of the amount of variance captured by the PCs indicated that the first two PC scores were sufficient to adequately describe recovery behaviour. The functional climatic covariates, i.e., the mean, minimum and maximum annual temperate, as well as annual summed up precipitation curves, were also included as one single PC score resulting from an MFPCA. The model showed the importance of soil composition for the establishment of new trees. In particular, higher amounts of sand relative to clay in the soil seemed to facilitate the growth of all tree species considered in LPJ-GUESS. The number of new seedlings immediately after the disturbance mostly showed significant effects on the recovery, especially the number of *tundra* seedlings for the growth of *temperate broadleaf*. Here some of the effects were rather counterintuitive, e.g., the negative effect of the number of *conifers (other)* on the establishment of needleleaf, which leaves room for further analysis. The most controversial effects were those of the explanatory variables `scenario` and the first climate PC values. Since these two covariates were interdependent, as the climate data serve as input to the scenario simulation in LPJ-GUESS, the effects were highly confounded, which was confirmed by the scenario-based models. In general, the scenarios had a large positive effect on the establishment of broadleafed trees, that is, as the climate warms, the area in which broadleaf trees can grow moves northwards. A potential drawback of the model setup was that it is not possible to disentangle the effects of temperature and precipitation, as a single MFPCA is fitted to the combined data. However, models with separate FPCAs for mean annual temperature and total annual precipitation showed that it is not possible to separate the two effects and obtain reasonable parameter estimates due to the interplay of the two variables. Especially in the northern parts of the globe, temperature is a strong indicator of precipitation and vice versa. Overall, the proposed modelling strategy adequately handled the functional nature of the data and contributed to a better understanding of vegetation dynamics.

Limitations

The proposed framework is shown to adequately model the data, but leaves room for future research directions. Firstly, the MFPCA approach does not take into account that the recovery trajectories that serve as input for the MFPCA are proportions of five PFTs. On the one hand, this means that the data are compositional, i.e., the restriction that all five proportions have to add up to one, but on the other hand it also means that the data are restricted to the interval $[0, 1]$.

For the former problem, there exists an interesting approach called Riemannian functional principal component analysis (RFPCA) developed by Dai and Müller (2018). The idea extends the methodology of MFPCA to handle functional data that lie on nonlinear spaces, such as Riemannian manifolds and spheres. Traditional MFPCA is primarily used for data in Euclidean space, but this approach is not directly applicable when data points are on a curved space due to the inherent nonlinearity. By taking the square root of the proportions,

the values can be considered to lie on the positive part of a sphere. This in turn implies that only one part of the sphere is suitable for the data, which causes problems whenever the combination of proportions hits the limits of that part. Nevertheless, RFPCA can serve as a starting point for further developments in the handling of compositional data when performing FPCAs and MFPCAs.

The latter problem, i.e., the restriction to $[0, 1]$, leads to limits on the PC scores when processed in a FPCA or MFPCA, but the current mFLR approach is not well suited for handling bounded PC scores. A major challenge is that the exact bounds of these scores are often unknown, making it difficult to apply standard mFLR techniques. This limitation has been tackled to some extent by transforming the PC scores with a generalized logistic function, but other approaches, including non-parametric transformations, could be explored. In general, there is a lack of methods for FLRs and mFLRs that can deal with bounded input data, so the development of new strategies for this case remains an open research question.

Another limitation of the modelling approach is the quantification of uncertainty in the estimated parameters. As the model is fitted to PC scores, there are two main sources of uncertainty in the model: the general uncertainty of a multivariate model, and that of performing an FPCA or MFPCA prior to modelling. In Section 2.2.3, the idea of Chiou, Yang, and Y.-T. Chen (2016), which is based on the asymptotic theory of the distribution of parameter estimates, is briefly outlined, but its concrete implementation was omitted in this project as the chosen approach included the 25 patches. In their paper, the authors propose theoretical results on the confidence bands of the mFLR, which can be further investigated for the presented modelling framework. The multi-patch repetition approach used in this project is highly tailored to the data provided for this analysis, and in many climate projections this multi-patch approach is a valid strategy for deriving confidence intervals. However, in the absence of multiple patches, a bootstrap approach, in which the data are resampled to estimate the variability in the parameter coefficients, may be a reasonable approach for an adequate uncertainty quantification.

Outlook

In addition to the ideas that follow as a consequence of the limitations mentioned, the data set at hand offers more possibilities for model selection. For example, in the proposed approach, the 25 patches were only used to ensure an appropriate uncertainty assessment. However, the 25 patches could be considered as 25 independent samples, which can be modelled as a random effect in e.g., multivariate functional additive mixed models, as suggested by Volkmann et al. (2023). The mixed models framework could also be valuable to further disentangle the effect of the climate scenario by including a random effect for the scenario. Furthermore, an alternative approach would be to view the data from a different perspective, e.g., as a time series rather than functional data. This allows the use of a variety of additional methods, including ARIMA models. For details on the application of this approach to climate data see Pruscha (2012).

In conclusion, this project provides a valid framework for dealing with functional vegetation data, but can also serve as a starting point for further improvement of the proposed method-

ology. This work contributes to a better understanding of the effects of climate change on the vegetation of the boreal forest and shows how changes will alter the character of this biome if the targets set in the Paris Agreement are not met, i.e., if the climate scenario SSP1-RCP2.6 is too weak a projection into the future. By quantifying these potential shifts, this work underscores the critical role of scientific insight in guiding future action. Statistics can serve as an important basis for the interplay between science, research, policy and society to show the changes that will come if we fail to stop the process we started. It is up to all of us to take action.

Acknowledgements

I would like to thank my supervisor Prof. Dr. Thomas Nagler for his great support and open-mindedness towards climate issues. It was really great to see that he got so involved in climate change and forestry. Together with Jana Gauss, he got me very passionate about functional data and the collaboration was a lot of fun.

I also received a lot of support and valuable insights into the work of climate scientists from the institute in Freising under the direction of Prof. Dr. Anja Rammig. Lucia Layritz in particular inspired me a lot with her expertise and enthusiasm for forest recovery. I am very grateful for this great collaboration.

Last but not least, I would like to thank my fellow library buddies for the great support we have given each other, as well as my sister and my boyfriend for digging through all these pages to give me valuable feedback.

References

Allen, Craig D et al. (2010). "A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests". In: *Forest ecology and management* 259(4), pp. 660–684.

Baltzer, Jennifer L et al. (2021). "Increasing fire and the decline of fire adapted black spruce in the boreal forest". In: *Proceedings of the National Academy of Sciences* 118(45), e2024872118.

Berrendero, J.R., Ana Justel, and Marcela Svarc (Sept. 2011). "Principal components for multivariate functional data". In: *Computational Statistics Data Analysis* 55, pp. 2619–2634. DOI: 10.1016/j.csda.2011.03.011.

Bonan, Gordon (July 2008). "Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests". In: *Science (New York, N.Y.)* 320, pp. 1444–9. DOI: 10.1126/science.1155121.

Chapin, F. Stuart, Pamela A. Matson, and Peter M. Vitousek (2011). *Principles of Terrestrial Ecosystem Ecology*. 2nd. Springer: New York. ISBN: 978-1-4419-9503-2. DOI: 10.1007/978-1-4419-9504-9.

Chiou, Jeng-Min, Ya-Fang Yang, and Yu-Ting Chen (2016). "Multivariate functional linear regression and prediction". In: *Journal of Multivariate Analysis* 146, pp. 301–312.

Coop, Jonathan D (2023). "Postfire futures in southwestern forests: Climate and landscape influences on trajectories of recovery and conversion". In: *Ecological Applications* 33(1), e2725.

Dai, Xiongtao and Hans-Georg Müller (2018). "Principal component analysis for functional data on Riemannian manifolds and spheres". In.

FAO and IIASA (2023). *Harmonized World Soil Database version 2.0*. Rome and Laxenburg. URL: https://doi.org/10.4060/cc3823en.

Foley, Jonathan A et al. (2000). "Incorporating dynamic vegetation cover within global climate models". In: *Ecological Applications* 10(6), pp. 1620–1632.

Fowlkes, Edward B and Colin L Mallows (1983). "A method for comparing two hierarchical clusterings". In: *Journal of the American statistical association* 78(383), pp. 553–569.

Gauthier, Sylvie et al. (2015). "Boreal forest health and global change". In: *Science* 349(6250), pp. 819–822. DOI: 10.1126/science.aaa9092.

Gertheiss, Jan et al. (2023). "Functional Data Analysis: An Introduction and Recent Developments". In: *arXiv preprint arXiv:2312.05523*.

Happ, Clara and Sonja Greven (2018). "Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains". In: *Journal of the American Statistical Association* 113(522), pp. 649–659. DOI: 10.1080/01621459.2016.1273115.

Happ-Kurz, Clara (2020). "Object-Oriented Software for Functional Data". In: *Journal of Statistical Software* 93(5), pp. 1–38. DOI: 10.18637/jss.v093.i05.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2(1), pp. 193–218.

Ilisson, Triin and Han YH Chen (2009). "The direct regeneration hypothesis in northern forests". In: *Journal of Vegetation Science* 20(4), pp. 735–744.

Kauermann, G., H. Küchenhoff, and C. Heumann (2021). *Statistical Foundations, Reasoning and Inference: For Science and Data Science*. Springer Series in Statistics. Springer In-

ternational Publishing. ISBN: 9783030698270. URL: https://books.google.de/books?id=Lt1FEAAAQBAJ.

Kimmins, J. P. (2004). *Forest Ecology: A Foundation for Sustainable Forest Management and Environmental Ethics in Forestry*. Prentice Hall.

Lange, Stefan and Matthias Büchner (2021). "ISIMIP3b bias-adjusted atmospheric climate input data (v1.1)". In.

Ma, Shibin et al. (2022). "Identification of forest disturbance and estimation of forest age in subtropical mountainous areas based on Landsat time series data". In: *Earth Science Informatics* 15(1), pp. 321–334.

Mack, Michelle C et al. (2021). "Carbon loss from boreal forest wildfires offset by increased dominance of deciduous trees". In: *Science* 372(6539), pp. 280–283.

Malhi, Yadvinder, DD Baldocchi, and PG Jarvis (1999). "The carbon balance of tropical, temperate and boreal forests". In: *Plant, Cell & Environment* 22(6), pp. 715–740.

Mandl, Lisa et al. (2024). "Unmixing-based forest recovery indicators for predicting long-term recovery success". In: *Remote Sensing of Environment* 308, p. 114194. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2024.114194.

McDowell, Nate G et al. (2020). "Pervasive shifts in forest dynamics in a changing world". In: *Science* 368(6494), eaaz9463.

Morey, Leslie C and Alan Agresti (1984). "The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement". In: *Educational and Psychological Measurement* 44(1), pp. 33–37.

Moss, Richard H. et al. (2010). "The next generation of scenarios for climate change research and assessment". In: *Nature* 463, pp. 747–756. DOI: https://doi.org/10.1038/nature08823.

Nakicenovic, Nebojsa et al. (2000). *IPCC Special Report: Emissions scenarios: Summary for policymakers*. Intergovernmental Panel on Climate Change: Genf. ISBN: 92-9169-113-5.

Olson, David M et al. (2001). "Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity". In: *BioScience* 51(11), pp. 933–938.

Ovenden, Thomas S et al. (2021). "Life after recovery: Increased resolution of forest resilience assessment sheds new light on post-drought compensatory growth and recovery dynamics". In: *Journal of Ecology* 109(9), pp. 3157–3170.

Pan, Yude et al. (2011). "A large and persistent carbon sink in the world's forests". In: *Science* 333(6045), pp. 988–993. DOI: 10.1126/science.1201609.

Pfadenhauer, Jörg and Frank Klötzli (Jan. 2020). "Global Vegetation, Fundamentals, Ecology and Distribution". In: DOI: 10.1007/978-3-030-49860-3.

Pruscha, Helmut (2012). *Statistical analysis of climate series: analyzing, plotting, modeling, and predicting with R*. Springer Science & Business Media.

Pugh, Thomas AM et al. (2019). "Important role of forest disturbances in the global biomass turnover and carbon sinks". In: *Nature geoscience* 12(9), pp. 730–735.

Radovan Hladky Josef Lastovicka, Lukas Holman and Premysl Stych (2020). "Evaluation of the influence of disturbances on forest vegetation using Landsat time series; a case study of the Low Tatras National Park". In: *European Journal of Remote Sensing* 53(1), pp. 40–66. DOI: 10.1080/22797254.2020.1713704. URL: https://doi.org/10.1080/22797254.2020.1713704.

Ramsay, J. O., Giles Hooker, and Spencer Graves (2009). *Functional Data Analysis with R and MATLAB*. 1st. Springer Publishing Company, Incorporated. ISBN: 0387981845.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer. ISBN: 9780387400808. URL: http://www.worldcat.org/isbn/9780387400808.

Rand, William M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66(336), pp. 846–850. DOI: 10.1080/01621459.1971.10482356.

Riahi, Keywan et al. (2017). "The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview". In: *Global Environmental Change* 42, pp. 153–168. ISSN: 0959-3780. DOI: https://doi.org/10.1016/j.gloenvcha.2016.05.009.

Senf, Cornelius and Rupert Seidl (2022). "Post-disturbance canopy recovery and the resilience of Europe's forests". In: *Global Ecology and Biogeography* 31(1), pp. 25–36.

Serra-Burriel, Feliu, Pedro Delicado, and Fernando M. Cucchietti (2021). "Wildfires Vegetation Recovery through Satellite Remote Sensing and Functional Data Analysis". In: *Mathematics* 9(11). ISSN: 2227-7390. DOI: 10.3390/math9111305. URL: https://www.mdpi.com/2227-7390/9/11/1305.

Silverman, B.W. and J.O. Ramsay (2002). *Applied Functional Data Analysis: Methods and Case Studies*. English. Other: Due for publication in June 2002. Springer, New York, NY: United States.

Sitch, Stephen et al. (2003). "Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model". In: *Global change biology* 9(2), pp. 161–185.

Smith, Benjamin, I Colin Prentice, and Martin T Sykes (2001). "Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space". In: *Global ecology and biogeography*, pp. 621–637.

Smith, Benjamin, D Wårlind, et al. (2014). "Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model". In: *Biogeosciences* 11(7), pp. 2027–2054. DOI: 10.5194/bg-11-2027-2014.

Smith-Tripp, Sarah M et al. (2024). "Landsat assessment of variable spectral recovery linked to post-fire forest structure in dry sub-boreal forests". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 208, pp. 121–135.

Strang, Gilbert (2023). *Introduction to Linear Algebra*. 6th. Wellesley-Cambridge Press.

Swann, Abigail L et al. (2010). "Changes in Arctic vegetation amplify high-latitude warming through the greenhouse effect". In: *Proceedings of the National Academy of Sciences* 107(4), pp. 1295–1300.

Thien, Steve J. (1979). "A flow diagram for teaching texture-by-feel analysis". In: URL: https://api.semanticscholar.org/CorpusID:222342434.

United Nations (2015). *Paris Agreement*. Article 1(a). URL: https://unfccc.int/sites/default/files/english_paris_agreement.pdf.

van Vuuren, Detlef P. et al. (2011). "The representative concentration pathways: an overview". In: *Climatic Change* 109(5). DOI: https://doi.org/10.1007/s10584-011-0148-z.

Volkmann, Alexander et al. (2023). "Multivariate functional additive mixed models". In: *Statistical Modelling* 23(4), pp. 303–326.

Wagner, Silke and Dorothea Wagner (2007). *Comparing clusterings: an overview*.

Wood, Simon N (2001). "mgcv: GAMs and generalized ridge regression for R". In: *R news* 1(2), pp. 20–25.

Wu, Junjie (2012). *Advances in K-means Clustering: A Data Mining Thinking*. Springer Theses. Springer: Berlin, Heidelberg.

Young, G. Alastair (Apr. 2014). "Inference for Functional Data with Applications by Lajos Horváth and Piotr Kokoszka". In: *International Statistical Review* 82(1), pp. 155–156. URL: https://ideas.repec.org/a/bla/istatr/v82y2014i1p155-156.html.

Zhu, Fangyan et al. (2020). "Characterizing the effects of climate change on short-term post-disturbance forest recovery in southern China from Landsat time-series observations (1988–2016)". In: *Frontiers of Earth Science* 14, pp. 816–827.

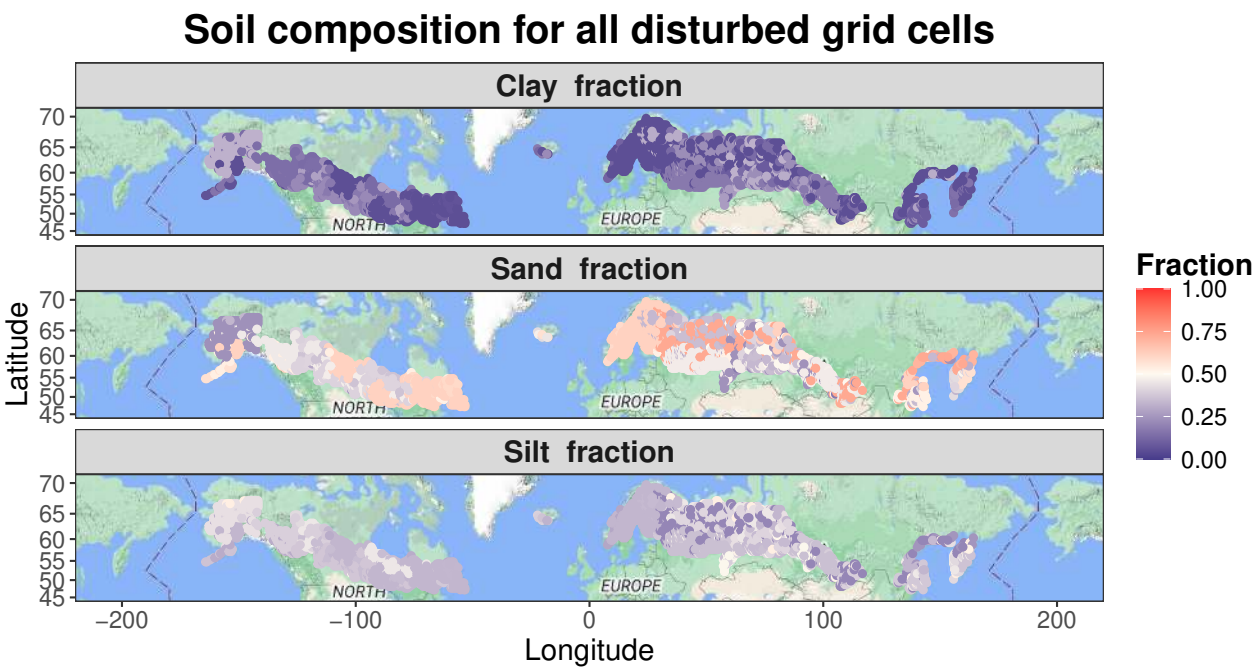


Figure 65: Spatial distribution of the soil composition for all disturbed grid cells between the years 2015 and 2040.

A Appendix

The appendix contains a collection of additional material, including descriptive maps of the non-functional variables provided, a clustering of the scenario- and PFT-wise PC scores derived from multiple univariate FPCAs in Section 4.3, and more details on the clusters based on the PC scores derived from the MFPCA described in Section 4.5. Note that all analyses in this section are based on data from one patch only.

A.1 Details on spatial descriptive statistics

To get an idea of which soil component dominates in which area, Figure 65 shows the proportions of clay, sand and silt for the whole study area. Most of Europe, Asia and the eastern part of North America are dominated by sand, with the highest proportions in northern Asia. In the western part of North America, the proportion of silt is higher than in the rest of the area. Clay plays a minor role in almost all regions. The distribution of bulk density shown in Figure 66a is fairly similar across the area, with most grid cells showing values around 1.5 and only a few locations with values below 0.8. These locations are somewhat scattered across the study area with no clear regional pattern. In contrast, the pH values in water (Figure 66b) show major regional differences, with the northern parts of Europe and Asia and the central part of North America having higher values than the rest of the study area. According to Figure 66c, organic carbon content is strongly related to bulk density, as the majority of locations have similar low organic carbon content values, and some locations scattered across the boreal forest – as with bulk density – have very high content values. This suggests that there is a relationship between high values of bulk density and low values

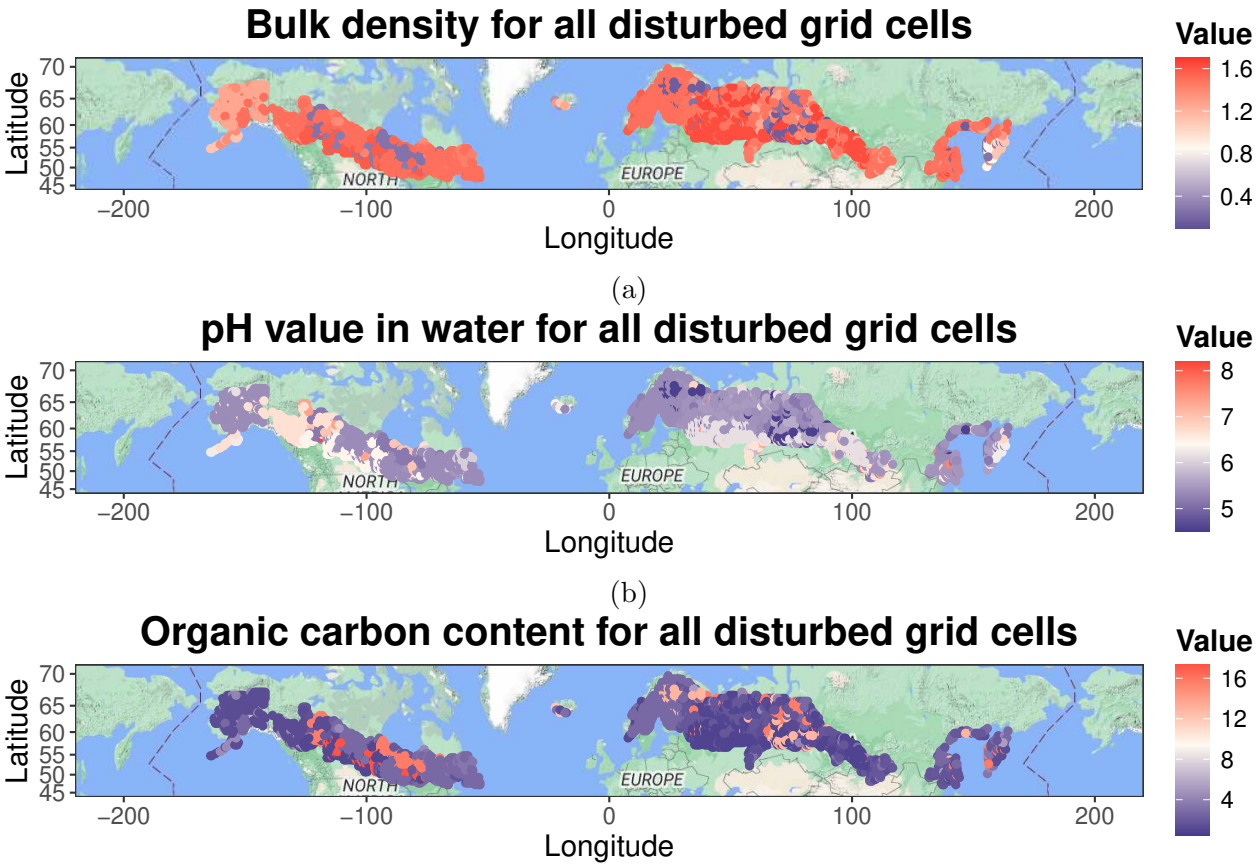
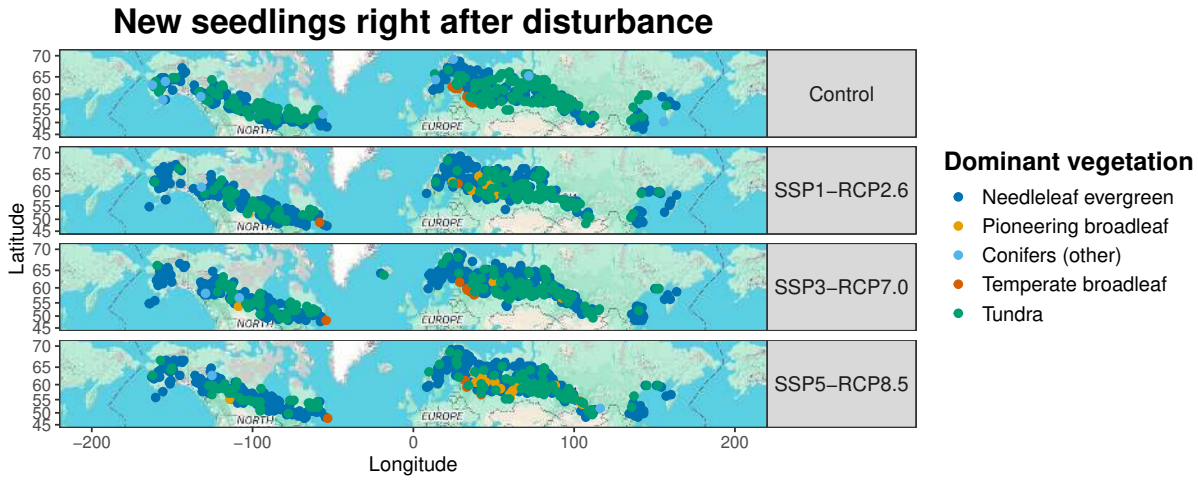


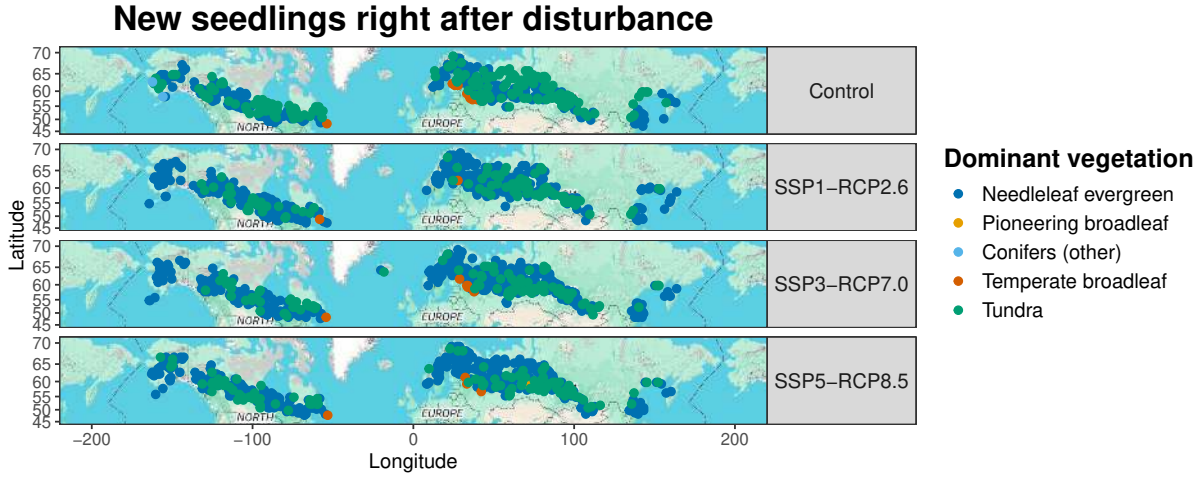
Figure 66: Spatial distribution of the soil properties bulk density (a), pH value in water (b) and organic carbon content (c) for all disturbed grid cells between the years 2015 and 2040.

of organic carbon content, and the other way around.

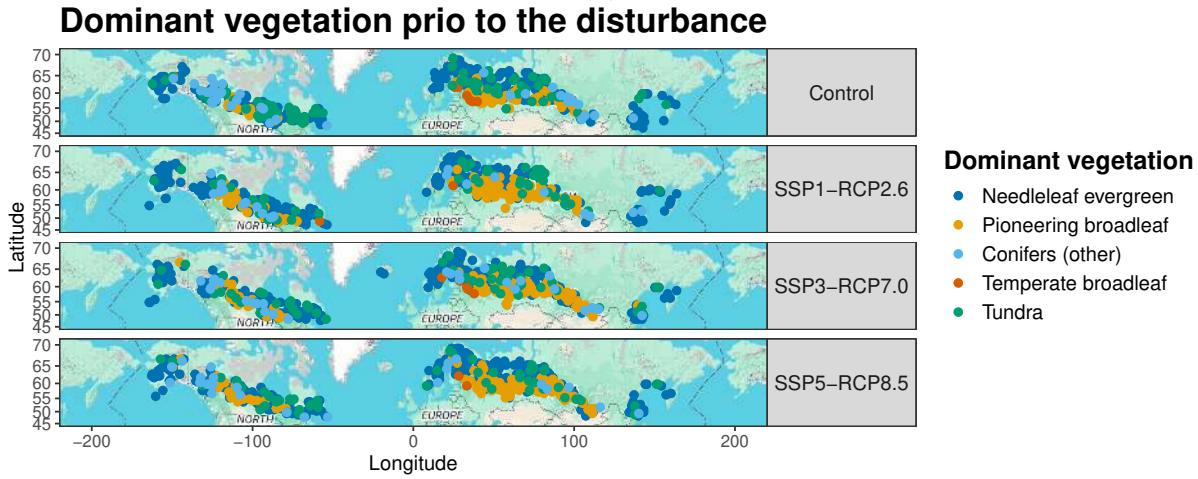
To examine ecological variables, Figure 67a shows the dominant tree species for each location and scenario, i.e., the PFT with the highest number of new seedlings immediately after the disturbance. Most of the grid cells are dominated by *tundra*, while coniferous and deciduous trees are less prominent. The latter tend to become more prevalent as the climate warms. In contrast, the distribution after ten years of recovery, that is, the maximum of the sum of recruitment in the first ten years after the disturbance visualised in Figure 67b, shows only minor differences between the scenarios and is mostly dominated by *tundra*. The hypothesis in Section 1 and Section 4.5.4 that the first years of recovery already determine the vegetation composition over most of the trajectory is supported by the fact that the dominance is less distributed across different PFTs and the similarity to the distribution immediately after the disturbance. Figure 67c shows the dominant vegetation types at each location in terms of the amount of aboveground carbon prior to the considered disturbance. In all scenarios there is a trend towards more *tundra* and needleleafed trees in the northern parts of the study area and more deciduous tree species in the southern parts. Particularly in North America, *pioneering broadleaf* becomes more widespread with increasing radial forcing.



(a)



(b)



(c)

Figure 67: Spatial distribution of the number of new seedlings directly after disturbance (a), of the number of new seedlings summed up over the first ten years of recovery (b) and of the dominant vegetation right before the disturbance (c). Dominance is indicated by the highest number of seedlings or the highest amount of aboveground carbon.

	Control	SSP1-RCP2.6	SSP3-RCP7.0	SSP5-RCP8.5
Cluster 1	105	39	151	67
Cluster 2	153	57	193	260
Cluster 3	119	171	39	95
Cluster 4	57	175	79	43

Table 10: Number of curves, i.e., grid cells, in each cluster and each scenario.

A.2 Details on Clusters derived by univariate FPCAs

In order to find similar behavior and patterns in the functional data describing recovery trajectories, the PC scores derived from univariate FPCAs for each scenario and each PFT separately presented in Section 4.3 are clustered with a k-means algorithm. The data used to cluster consists of ten PC scores – the first two component scores for each of the five PFTs. That is, clustering is performed for each scenario separately, combining the five PFTs. To determine an appropriate number k of clusters, Figure 68 shows an elbow plot for all four scenarios considered in this study. In the control scenario, there can be assumed that there is an elbow at $k = 2$ and $k = 4$. Since there are no clear indicators for k in the warming scenarios, $k = 4$ is chosen for all subsequent analyses. The resulting four clusters are rather unbalanced as Table 10 indicates. While for the control, there are three large clusters, both SSP1-RCP2.6 and SSP3-RCP7.0 are dominated by two large clusters. SSP5-RCP8.5 comprises one larger cluster and three smaller ones. Note that since only one patch is considered, each curve corresponds to a grid cell in the data set.

To get a deeper insight into how the PFTs influence the clustering process, the first two PC scores are plotted against each other for each PFT. Figure 69 shows the PC scores for PFT *needleleaf evergreen*, with colours indicating the corresponding cluster of the grid cells. In general, there is an increase in the first PC scores with a warmer climate. The four clusters are distinguishable within each cluster, although the more pronounced the increase in radial forcing, the less overlap between the clusters.

Figure 70 shows the equivalent plot for *pioneering broadleaf*. Again, the grouping structure is clearly present in the data. The more extreme the scenario, the smaller the first PC scores and the more distinguishable the clusters. In addition, with radial forcing the variation in the second PC increases.

Looking at *conifers (others)* portrayed in Figure 71 reveals no clear dissociation of the clusters, however a slight grouping remains. Again, the more extreme the scenario, the higher the PC 1 values.

There is a clear lack of data for *temperate broadleaf* shown in Figure 72. This species tends to be more common in milder climates, which explains the increase in data points with climate warming. For the SSP5-RCP8.5 scenario, some grouping structure can be assumed. However, the results for this PFT should not be over-interpreted.

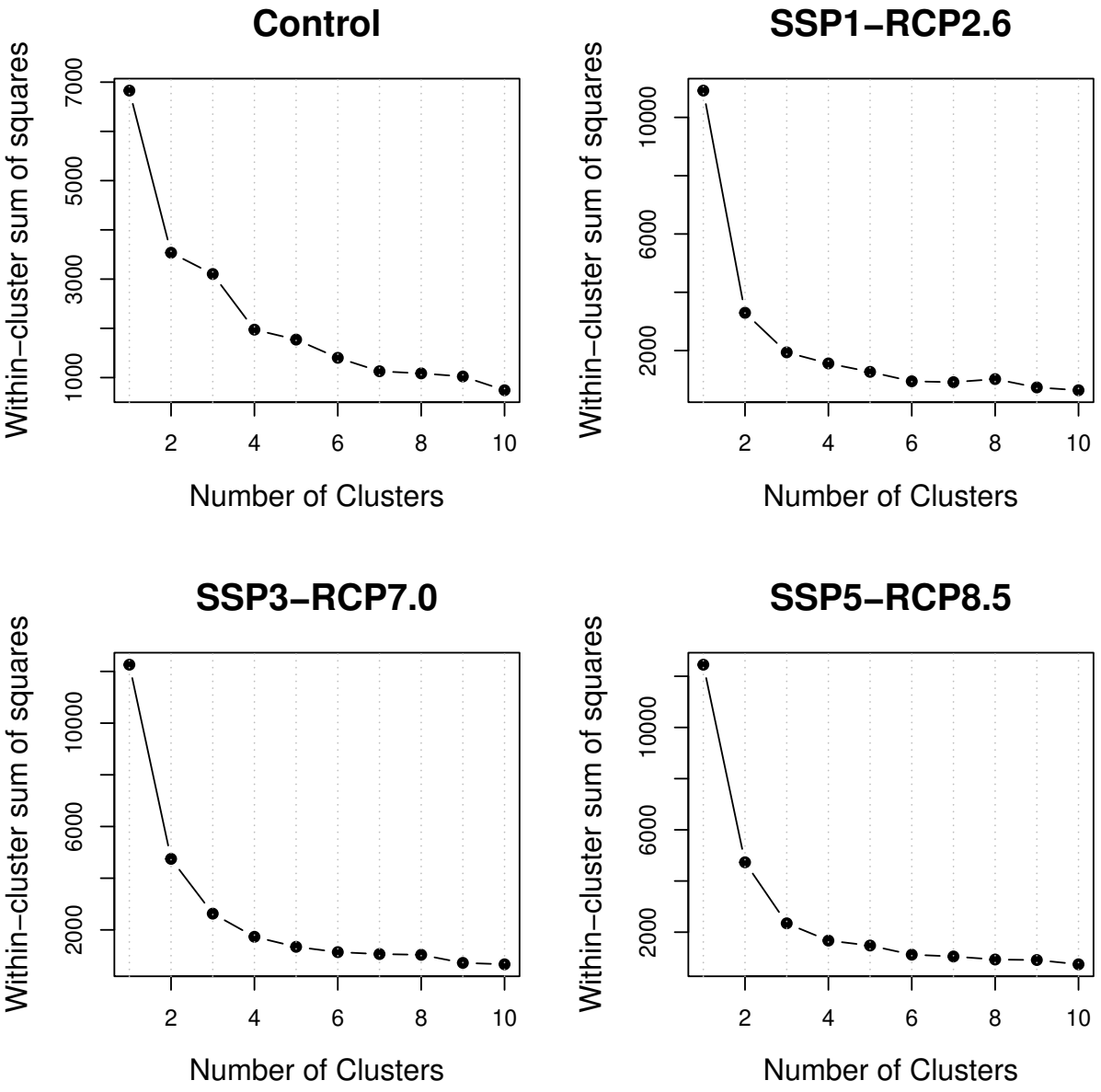


Figure 68: Elbow plots for each scenario indicating the within cluster sum of squares for different numbers of clusters.

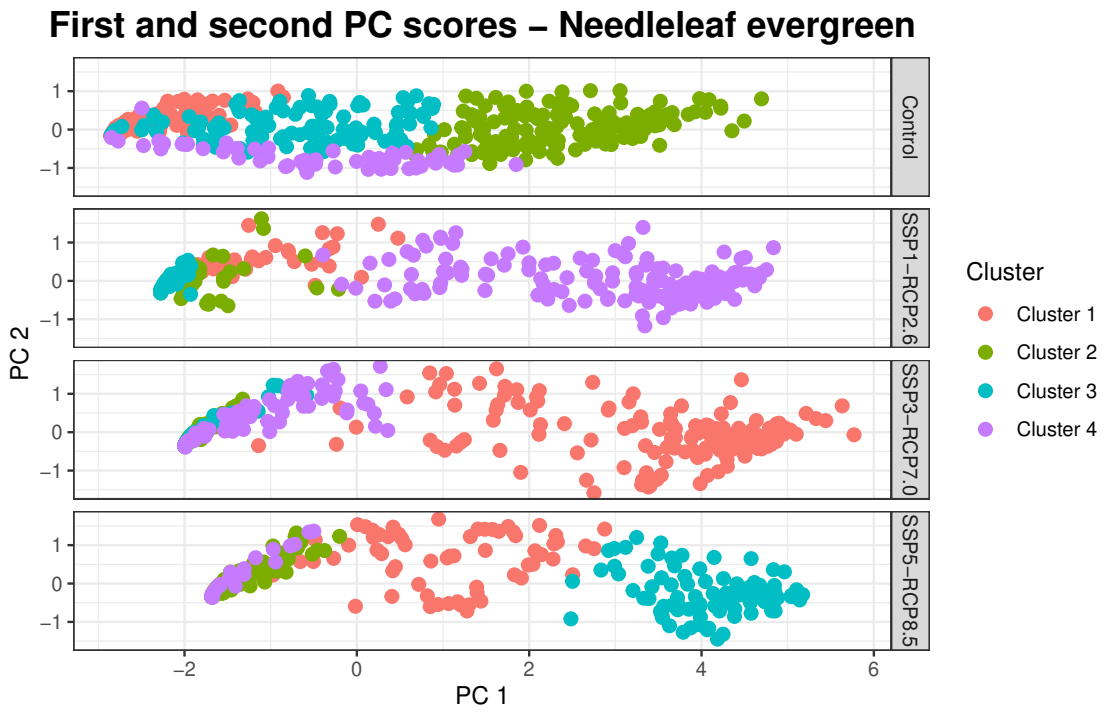


Figure 69: First PC scores plotted against second PC scores for PFT *needleleaf evergreen* and all four scenarios as well as one patch.

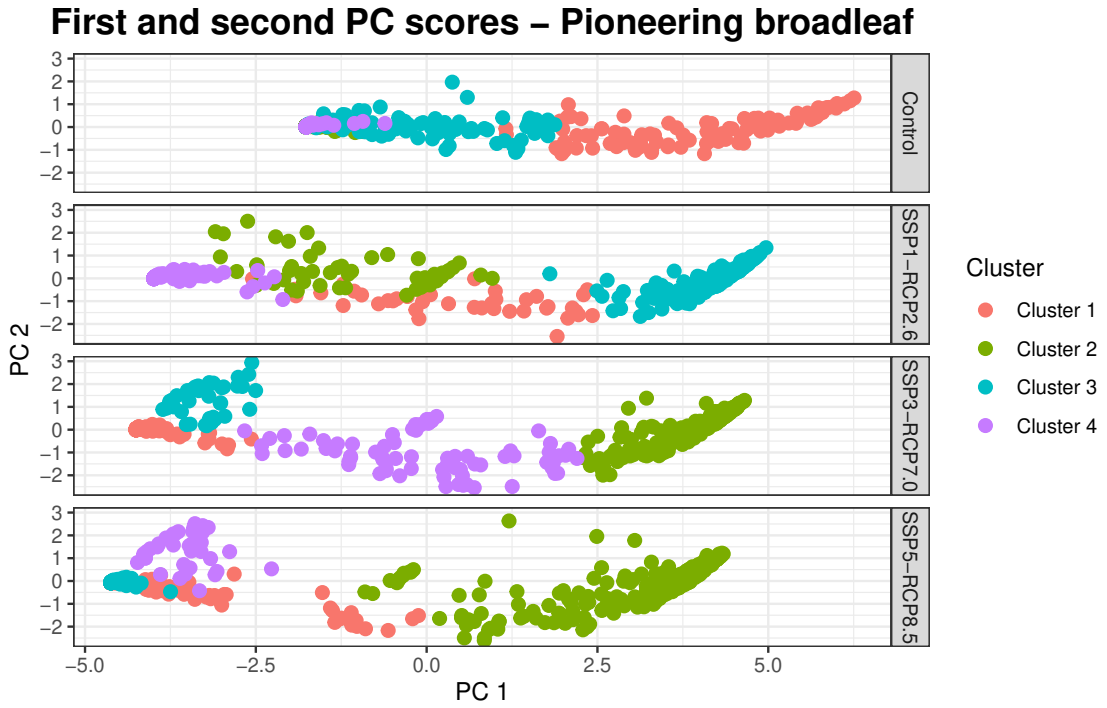


Figure 70: First PC scores plotted against second PC scores for PFT *pioneering broadleaf* and all four scenarios as well as one patch.

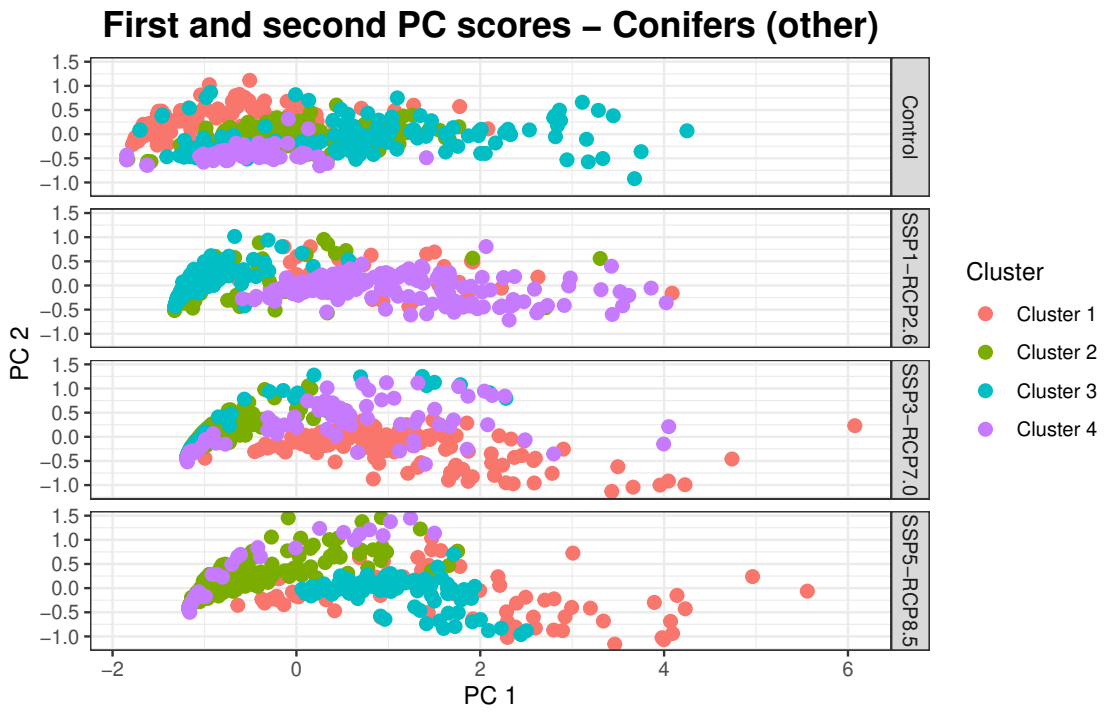


Figure 71: First PC scores plotted against second PC scores for PFT *conifers (others)* and all four scenarios as well as one patch.

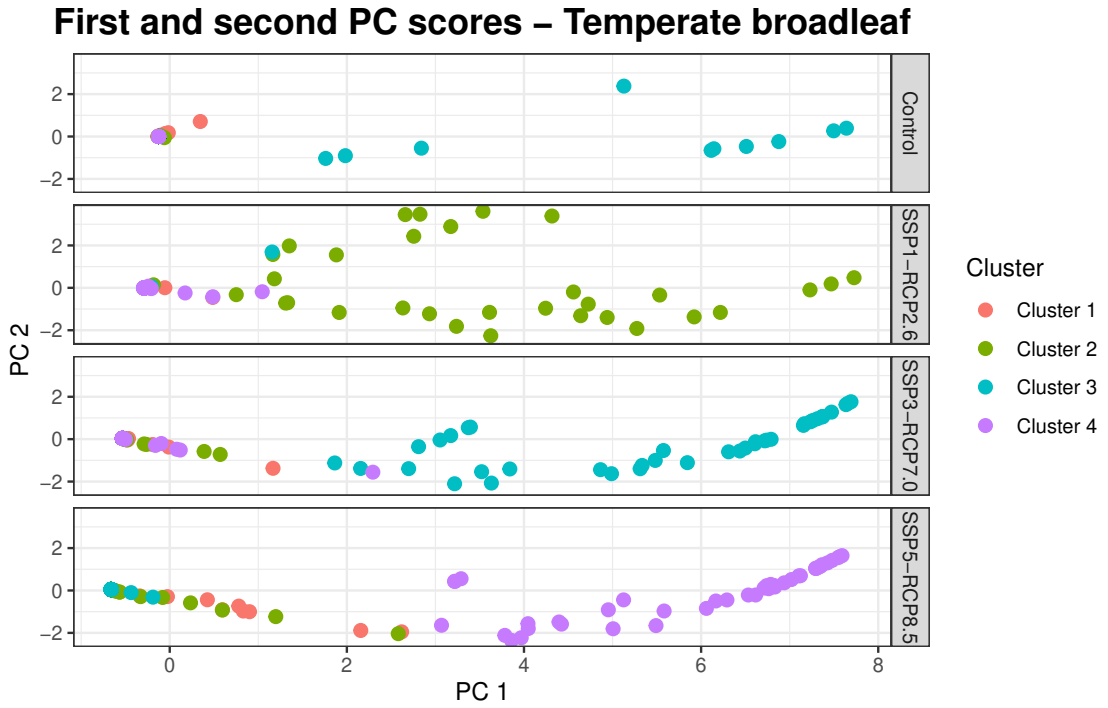


Figure 72: First PC scores plotted against second PC scores for PFT *temperate broadleaf* and all four scenarios as well as one patch.

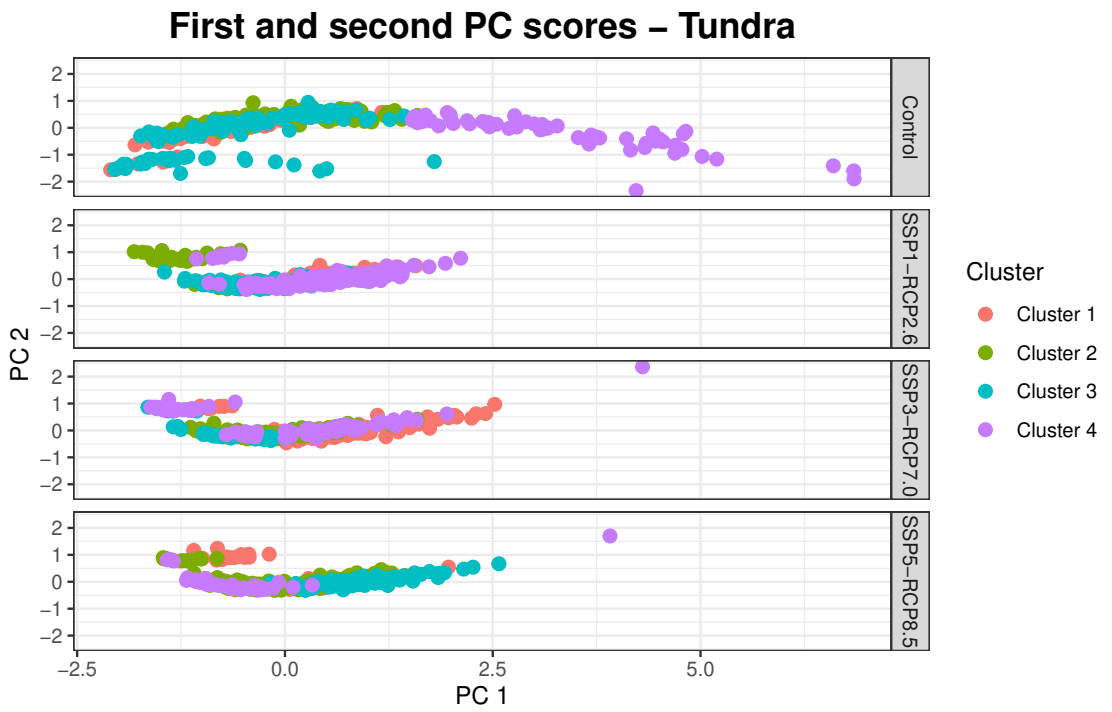


Figure 73: First PC scores plotted against second PC scores for PFT *tundra* and all four scenarios as well as one patch.

For PFT *tundra*, there is a visual clustering possible as Figure 73 indicates. The clustering is slightly detectable but the visual clusters are not met entirely.

Overall, plotting the first two PC scores indicates that the clustering algorithm is mainly driven by PFTs *needleleaf evergreen* and *pioneering broadleaf*, since for these two PFTs the clusters define clear patterns in the data structure.

A.2.1 Clustering for the control scenario

In order to gain more insights into which curves belong to which cluster, Figure 74 shows the clustered curves for the control scenario for each cluster and PFT *needleleaf evergreen*. The color indicates the cluster, while the dark curves represent the within-cluster mean functions. Cluster 2 represents grid cells with a rather high share of *needleleaf evergreen*, while cluster 1 reflects grid cells with low shares. The third and forth cluster cover the functions in between and vary in terms of sharpness of the growth behavior after disturbance.

Figure 75 portrays the equivalent plot for PFT *pioneering broadleaf*. Cluster 1 is driven by grid cells with a very high share of aboveground carbon. Clusters 2 and 4 cover all grid cells with rather low shares, the mean functions hardly vary from zero. Cluster 3 represents all curves in between.

While cluster 3 was driven by medium shares of *needleleaf evergreen* and *pioneering broadleaf*

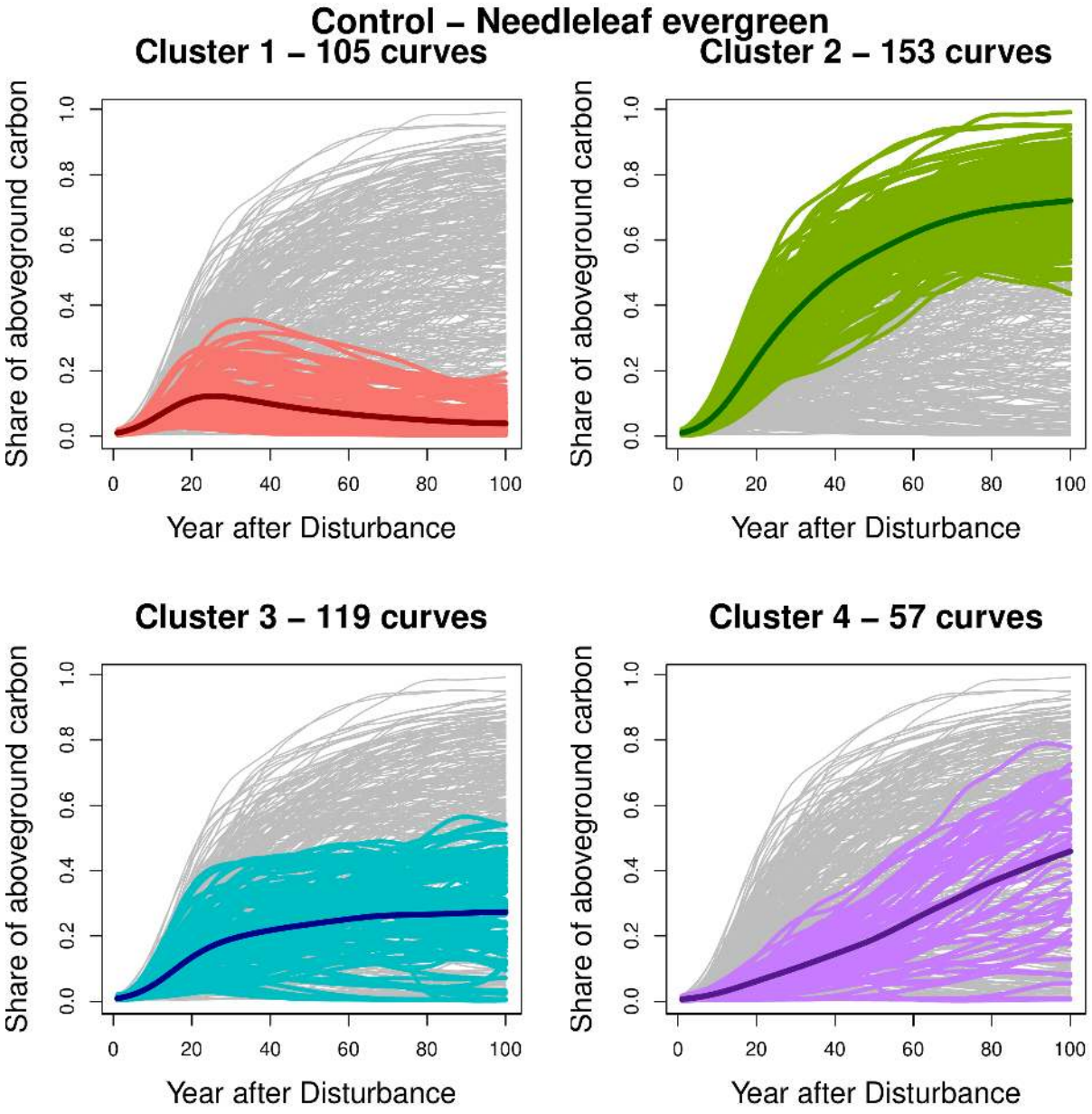


Figure 74: Clustered curves for the control scenario and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

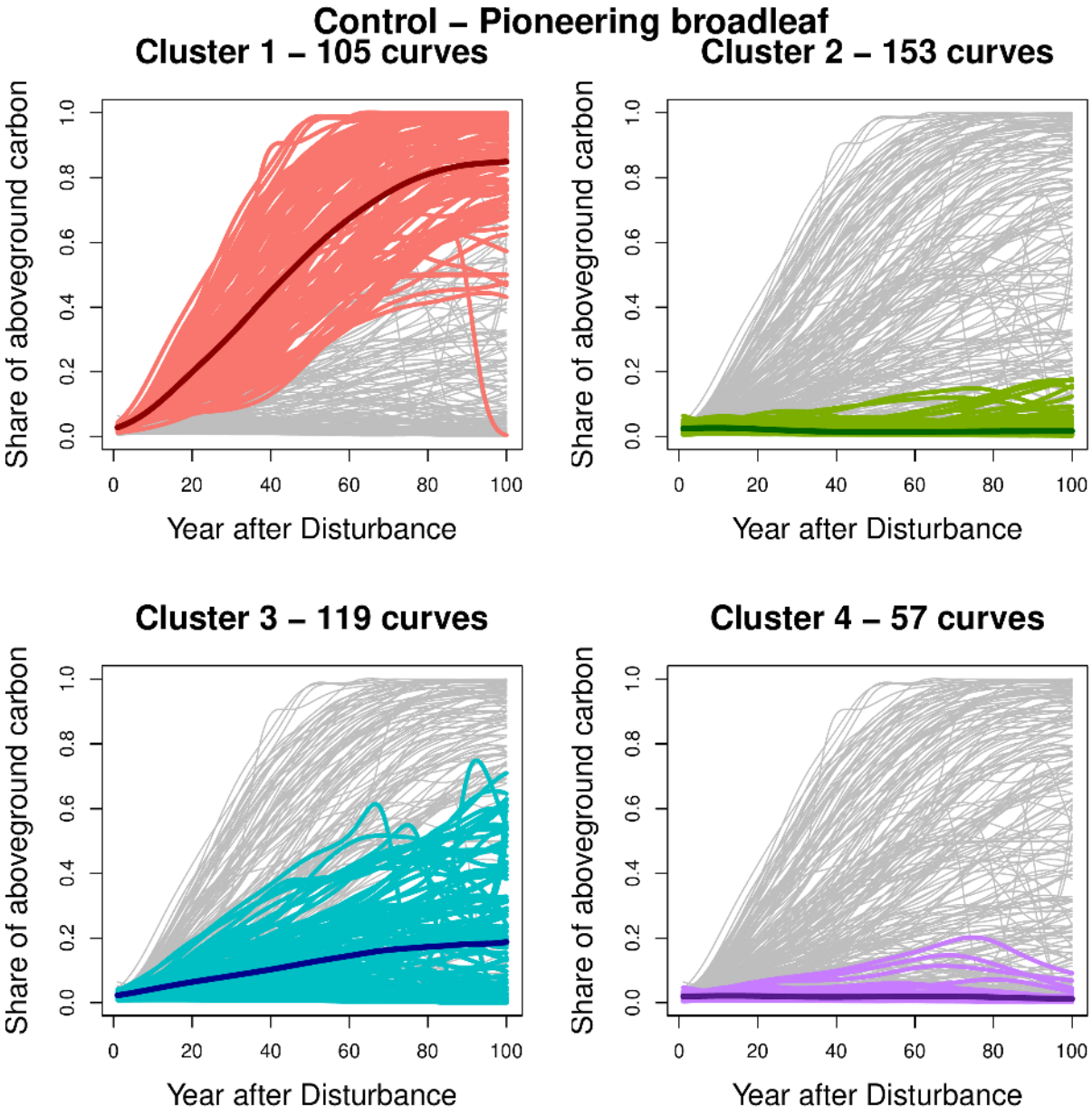


Figure 75: Clustered curves for the control scenario and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

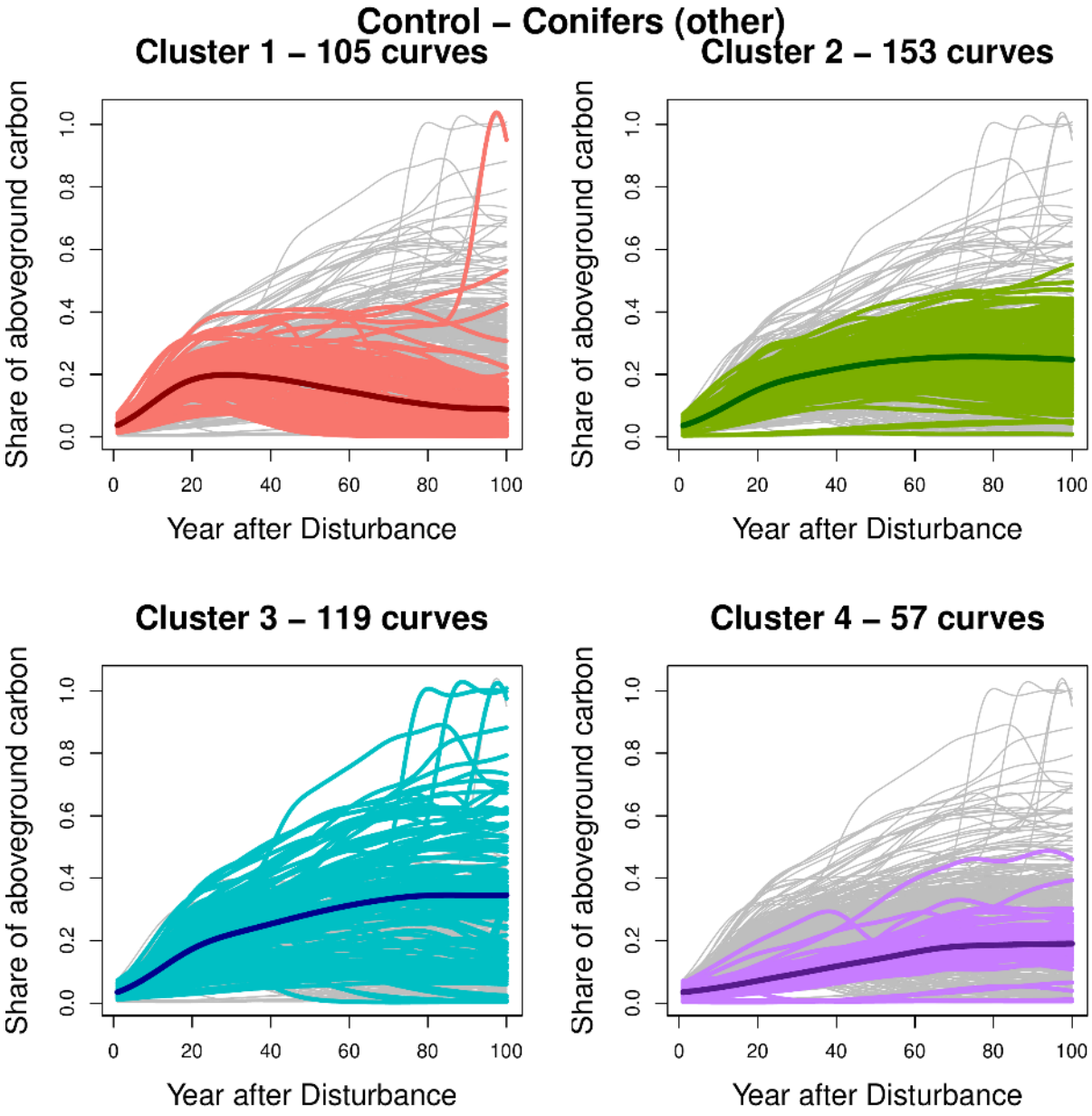


Figure 76: Clustered curves for the control scenario and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

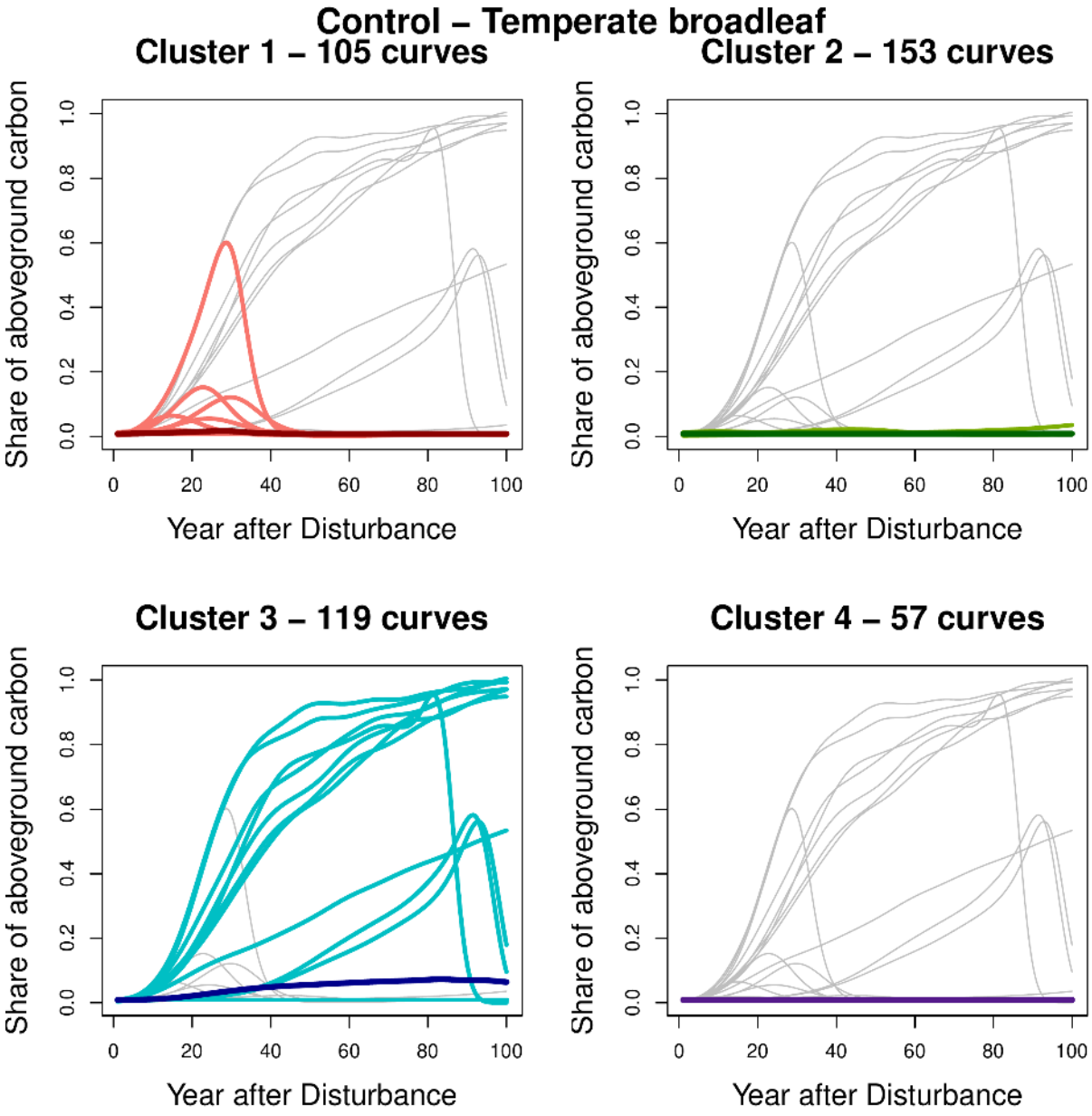


Figure 77: Clustered curves for the control scenario and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

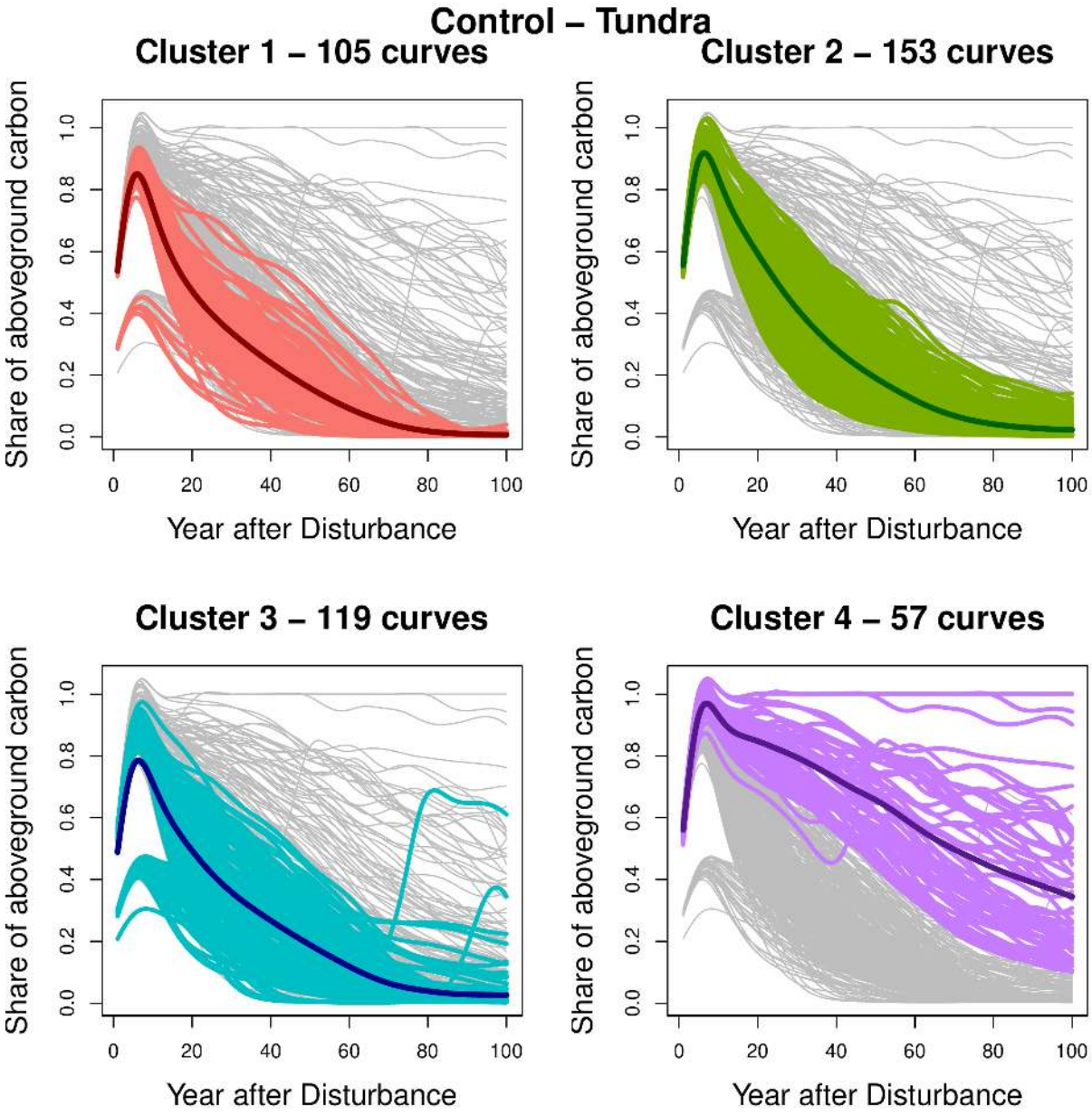


Figure 78: Clustered curves for the control scenario and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

so far, Figure 76 reveals high shares of *conifers (others)*. The remaining three clusters represent medium shares of aboveground carbon, differing in size and timing of the peaks if existent.

For *temperate broadleaf* depicted in Figure 77, no valid interpretation is possible due to lack of data. One could possibly argue that cluster 3 represents those curves with a rather high share, but this result should not be overestimated.

The clustered curves for PFT *tundra* are shown in Figure 78. Here, cluster 4 represents all the grid cells with a high share of *tundra* throughout the whole recovery period while cluster 2 represents those curves with a rather high peak. The differences between the remaining two clusters are less pronounced.

In total, the clusters for the control scenario are highly driven by the share of *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*, while *tundra* and *temperate broadleaf* seem less crucial for finding grouping structures in the PC scores. Cluster 1 is dominated by high shares of *pioneering broadleaf*, while cluster 2 covers grid cells with high shares of *needleleaf evergreen*. Cluster 3 combines high shares of *temperate broadleaf* and high shares of *conifers (others)*, while cluster 4 comprises mostly curves with high shares of *tundra*.

A.2.2 Clustering for scenario SSP1-RCP2.6

In order to assess whether the same structures in the clusters are present in the clustering of PC scores for scenario SSP1-RCP2.6, Figure 79 shows the clustered curves for PFT *needleleaf evergreen*. Here, the grouping structure is even more apparent than in the equivalent figure for the control scenario: the dominating cluster 4 represents curves with a high share of *needleleaf evergreen*, while clusters 2 and 3 represent grid cells with a low share of aboveground carbon. Cluster 1 covers grid cells with a medium share, but is very small in size in comparison to the other clusters in general.

Cluster 3 is dominated by *pioneering broadleaf* as indicated in Figure 80. Cluster 4, one of the two largest clusters, represents grid cells with nearly no *pioneering broadleaf*, while cluster 1 and 2 cover those with a medium share of aboveground carbon. Note that those two clusters show large differences in the general dynamics of the curves.

Figure 81 reveals that clusters 2 and 3 represent low to medium shares of *conifers (others)*, differing in terms of size and timing of the peak in share of aboveground carbon. Moreover, the dominant cluster 4 and the small cluster 1 cover grid cells with high shares of *conifers (others)*.

The lack of data for PFT *temperate broadleaf* remains an issue for scenario SSP1-RCP2.6 (Figure 82). Cluster 2 comprises most of the non-zero shares in aboveground carbon, but again, this result should not be over-interpreted.

The clusters of the final PFT *tundra* shown in Figure 83 are barely distinguishable. Cluster

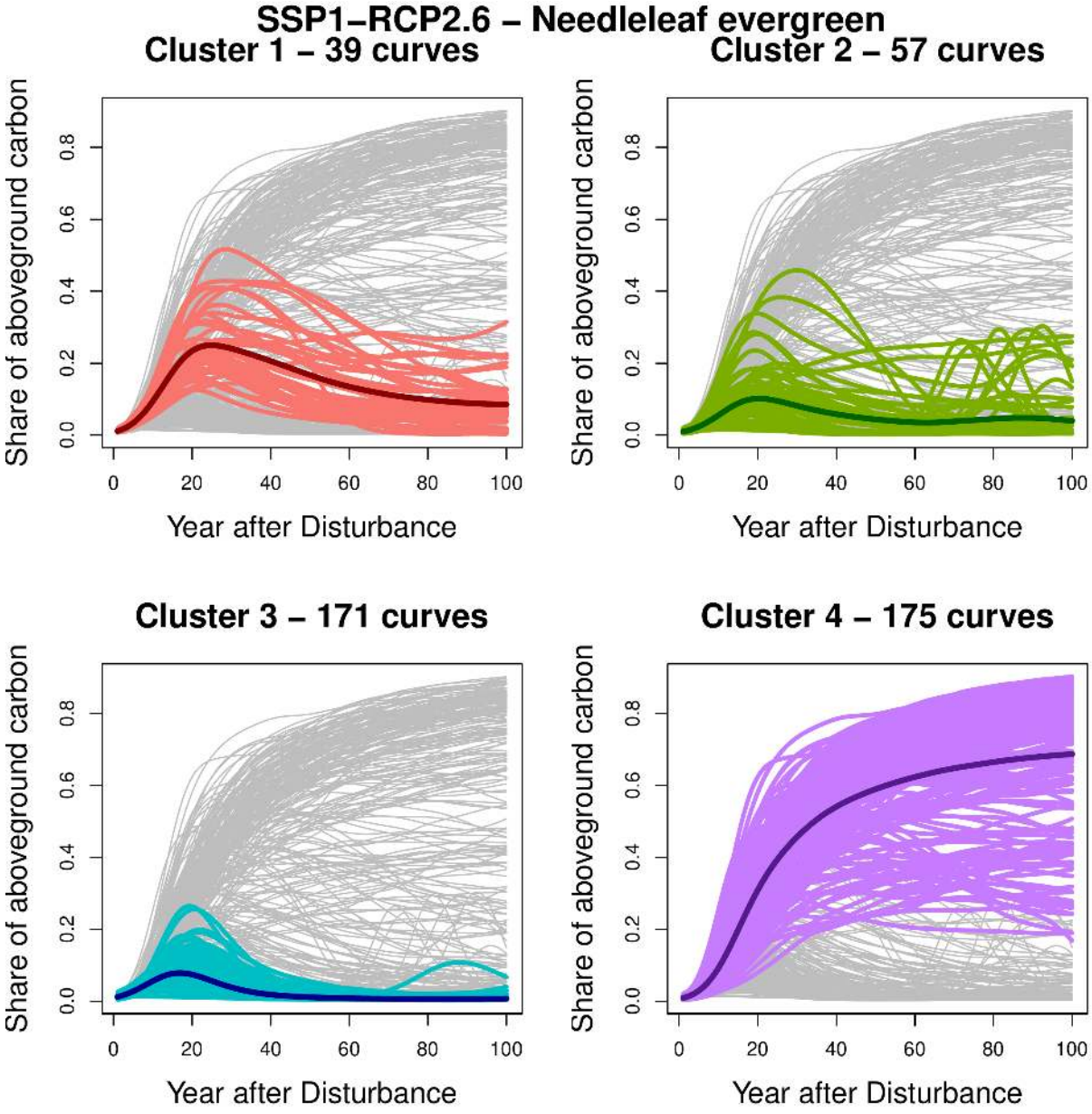


Figure 79: Clustered curves for scenario SSP1-RSCP2.6 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

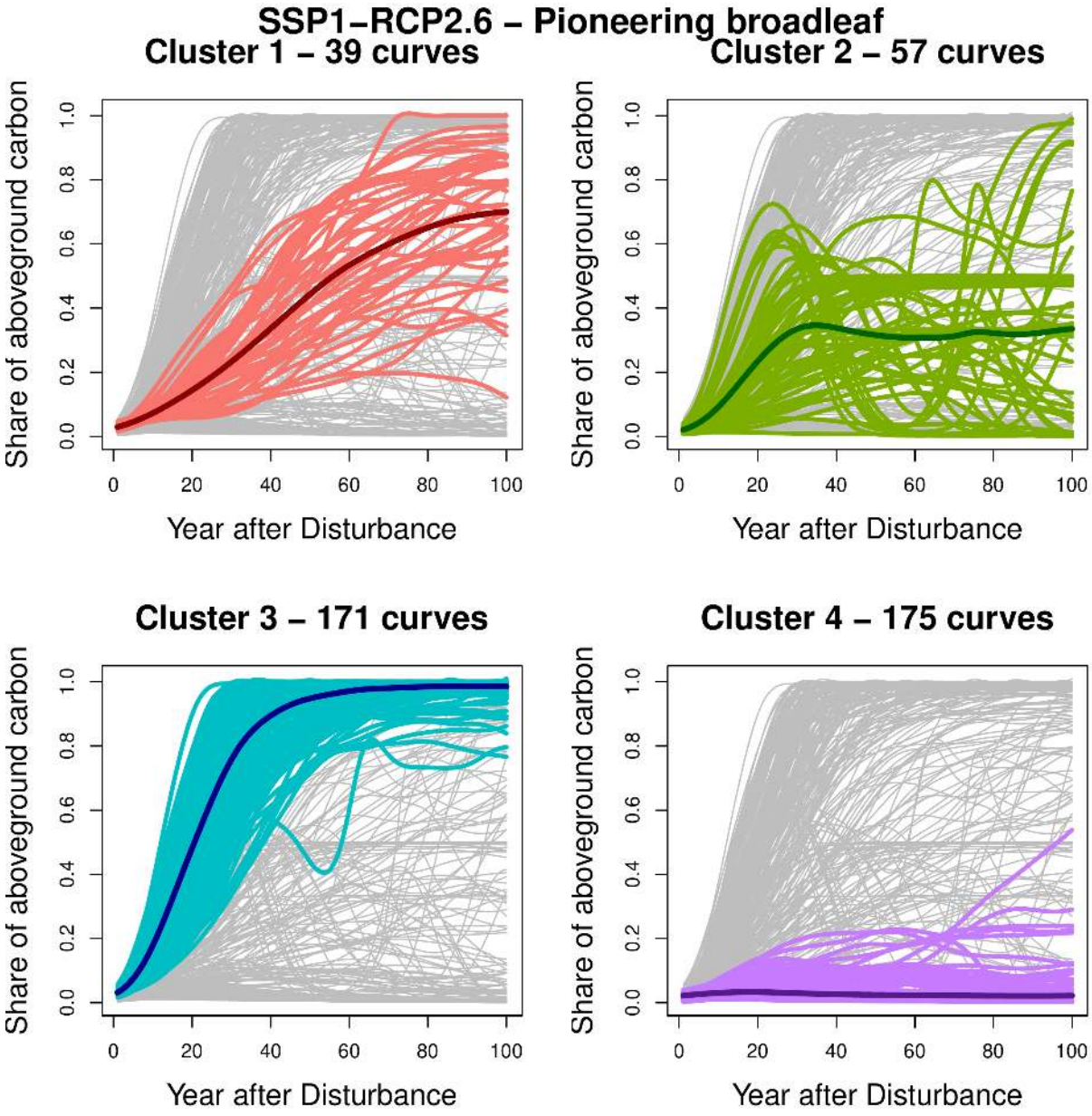


Figure 80: Clustered curves for scenario SSP1-RSCP2.6 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

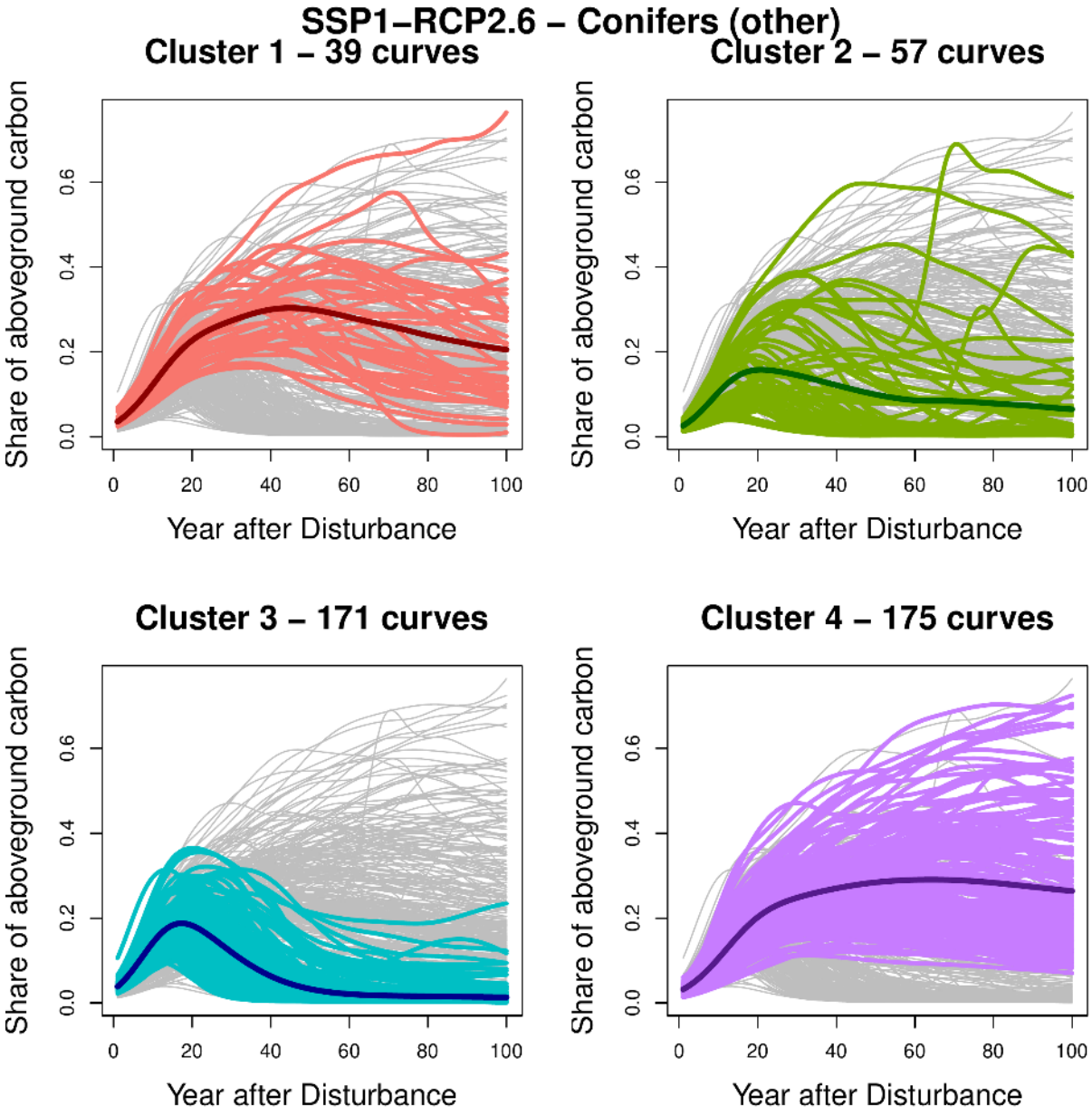


Figure 81: Clustered curves for scenario SSP1-RSCP2.6 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

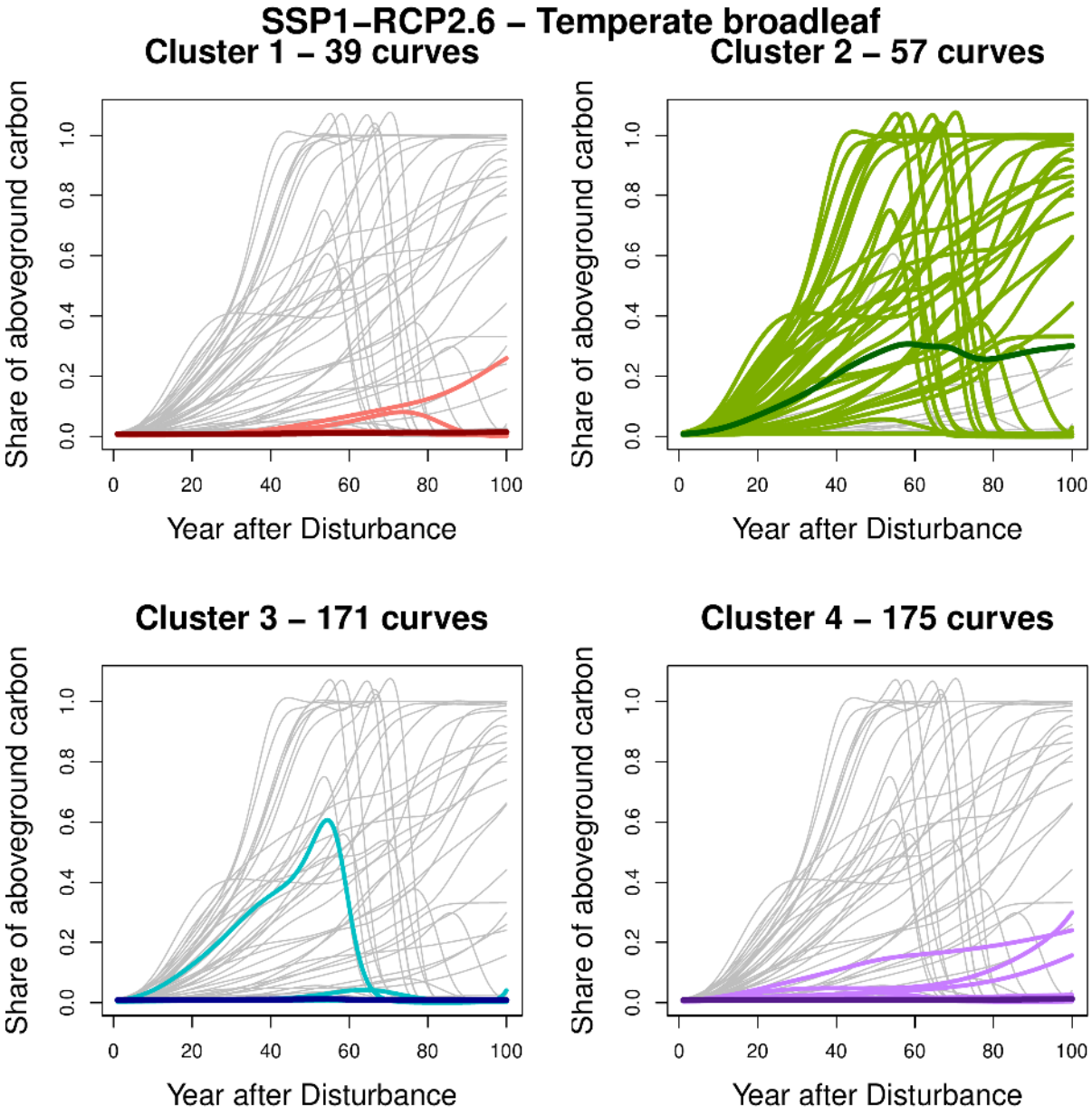


Figure 82: Clustered curves for scenario SSP1-RSCP2.6 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

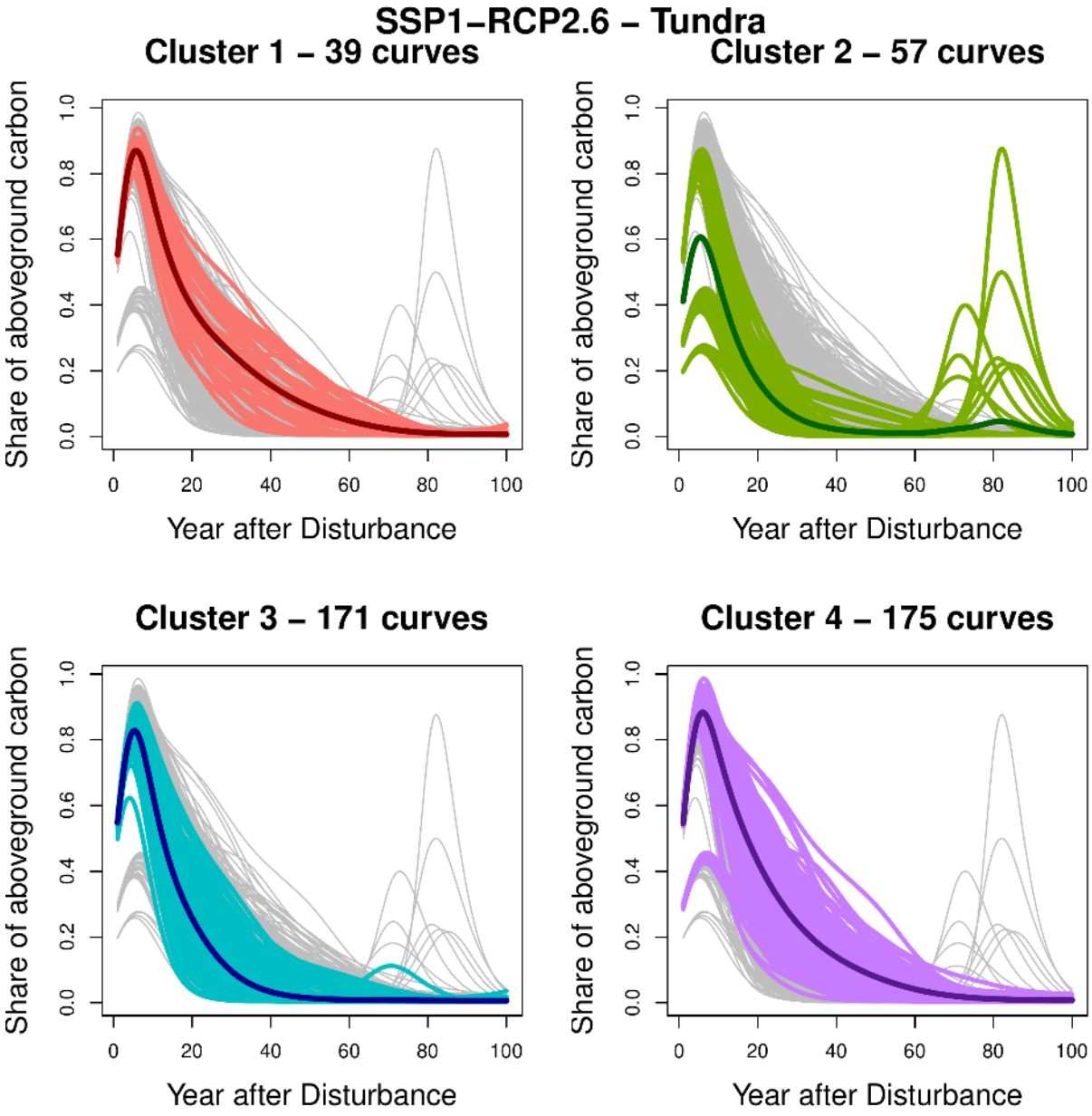


Figure 83: Clustered curves for scenario SSP1-RSCP2.6 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

2 comprises grid cells with second peaks in *tundra* in the last decades of the study period, whereas the remaining clusters do not show any characteristics.

As a result, the clustering mechanisms for scenario SSP1-RCP2.6 are again mainly defined by PFTs *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*. Cluster 1 is dominated by medium shares of *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*. In cluster 2, *temperate broadleaf* seems to be of importance, while cluster 3 reflects high shares of *pioneering broadleaf*. Cluster 4 is characterised by high shares of *needleleaf evergreen* and *conifers (others)*.

A.2.3 Clustering for scenario SSP3-RCP7.0

In the following the clusters for scenario SSP3-RCP7.0 representing a mediate climate warming are under examination. Clustering PC scores of PFT *needleleaf evergreen*, depicted in Figure 84, results in similar patterns as seen in the scenario before, with clusters 2 and 3 representing low shares, cluster 1 high shares and cluster 4 medium shares of aboveground carbon. Note that the two dominating clusters, namely clusters 1 and 2, are mainly driven by the extreme cases, either very high shares of aboveground carbon, or very low shares.

Cluster 2 is characterized by very high shares of *pioneering broadleaf* shortly after disturbances, as visualized in Figure 85. The second largest cluster, cluster 1, which is dominated by *needleleaf evergreen*, has nearly mean zero for shares of *pioneering broadleaf*. The two smaller clusters 3 and 4 represent medium shares of aboveground carbon, differing highly in size and timing of the peak if existent at all.

For PFT *conifers (others)* shown in Figure 86 the behaviour resembles to that of the two previous scenarios: two clusters (2 and 3) reflect minor shares of aboveground carbon, while clusters 1 and 4 represent mediate to high shares of *conifers (others)*. Note that clusters 1 and 2 are the largest clusters, suggesting that another PFT is responsible for cutting cluster 4 of cluster 1 and cluster 3 of cluster 2.

Figure 87 portrays the clustered curves for PFT *temperate broadleaf*. Again, one cluster, here cluster 3, comprises nearly all non-zero curves. Note that despite the overall increase of non-zero curves, the results should be interpreted carefully as for the preceding scenarios.

For PFT *tundra* depicted in Figure 88, no substantial differences between the clusters are detectable. The smallest cluster 3 tends to reflect grid cells with a fast decrease in share of aboveground carbon. The remaining three clusters are hardly distinguishable.

To summarize the results for scenario SSP3-RCP7.0, similar to the scenarios before, the clustering algorithm is mainly influenced by the PFTs *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*. Cluster 1 is driven by high shares of *needleleaf evergreen* and *conifers (others)*, while cluster 2 is dominated by *pioneering broadleaf*. Cluster 3 comprises nearly all non-zero curves of *temperate broadleaf*, while cluster 4 represents medium to high shares of *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*.

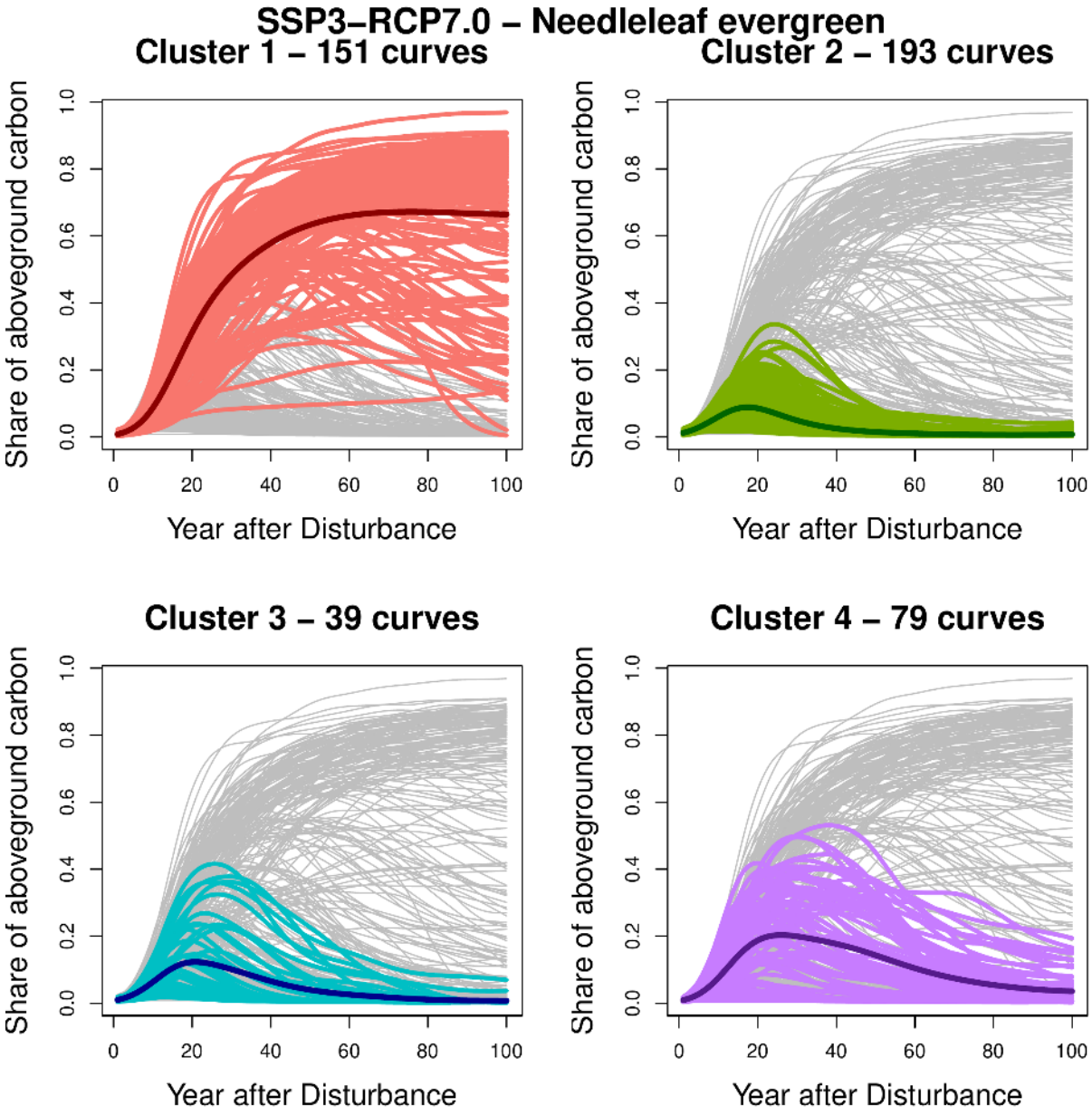


Figure 84: Clustered curves for scenario SSP3-RSCP7.0 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

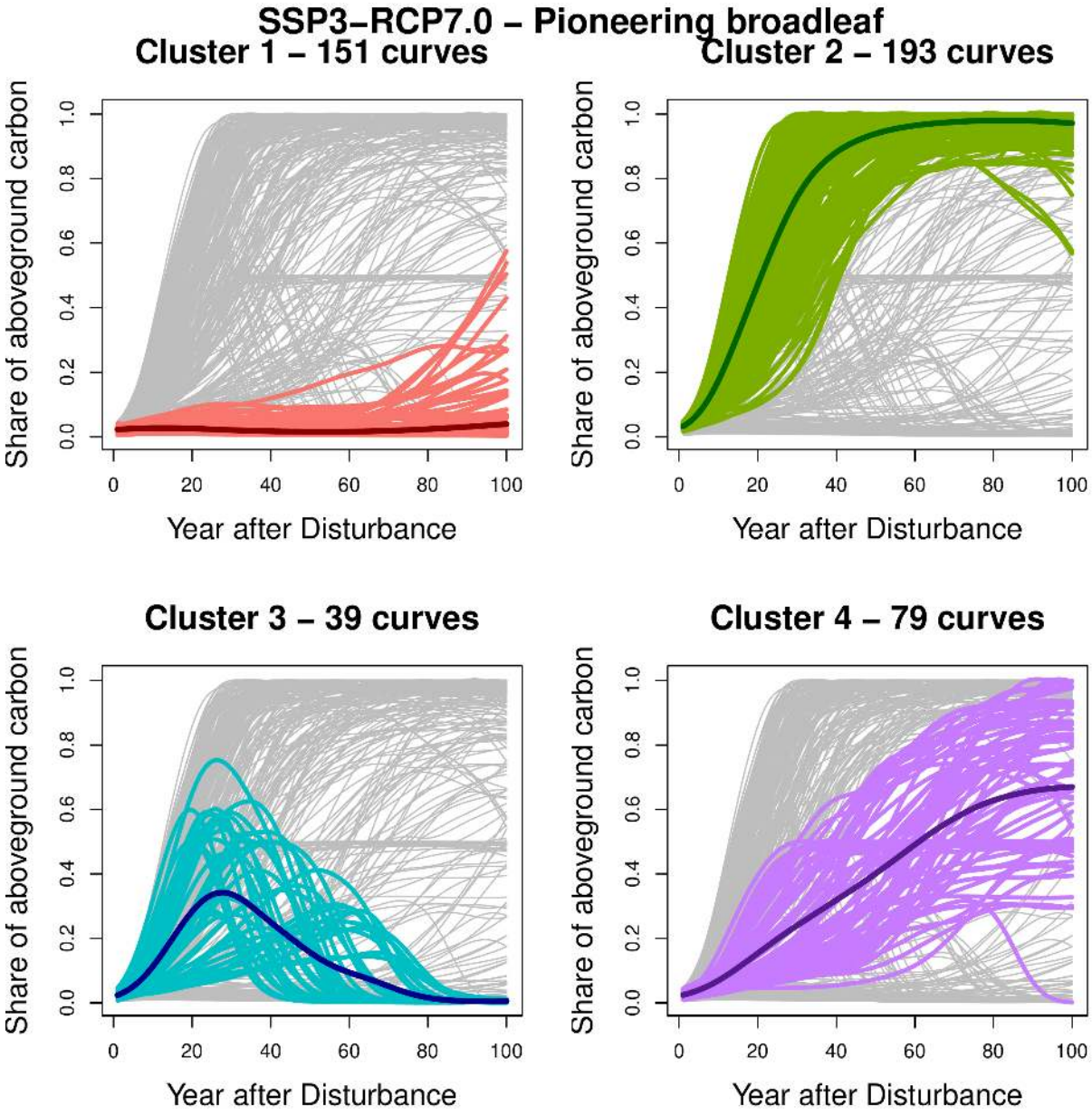


Figure 85: Clustered curves for scenario SSP3-RSCP7.0 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

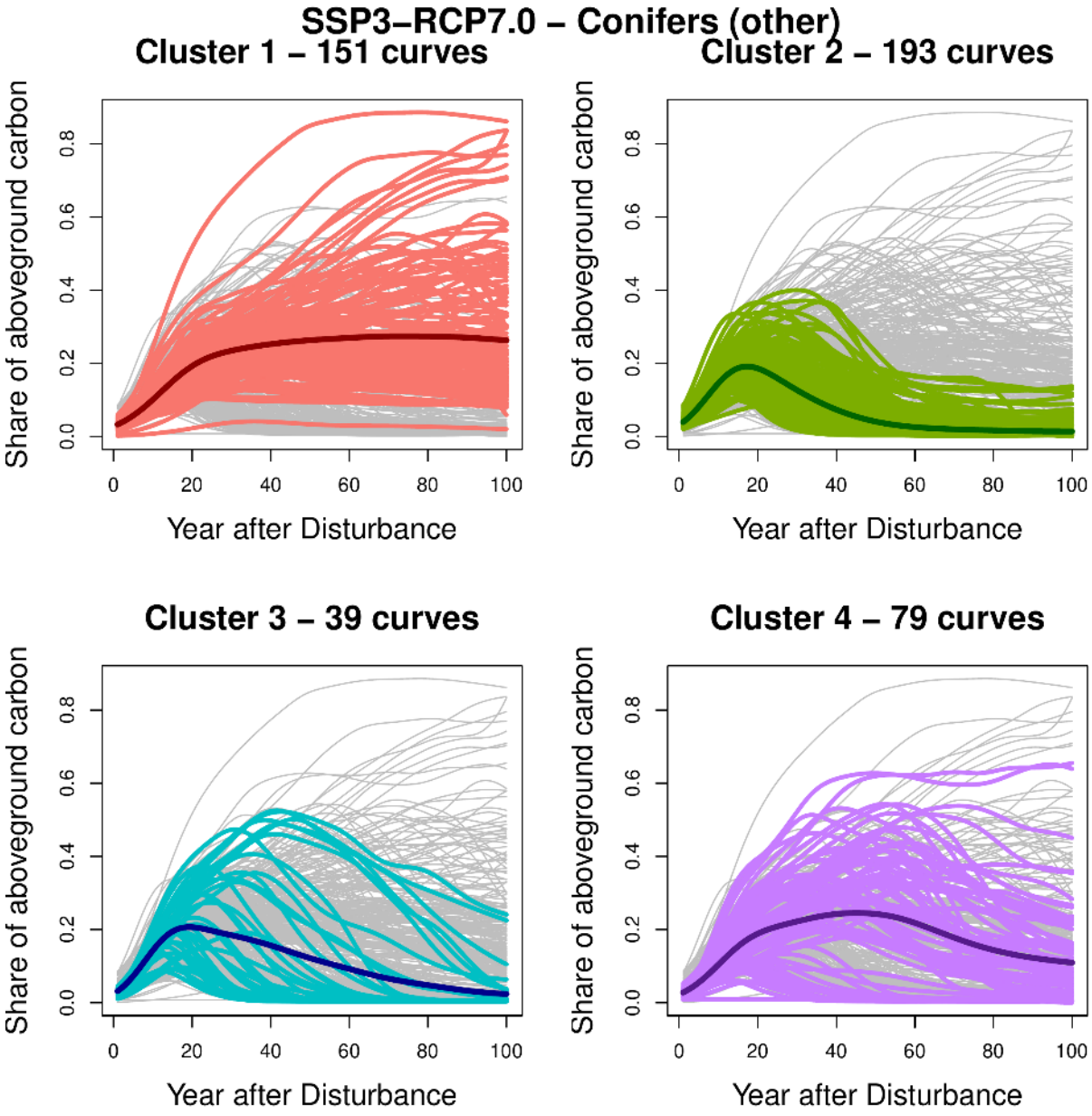


Figure 86: Clustered curves for scenario SSP3-RSCP7.0 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

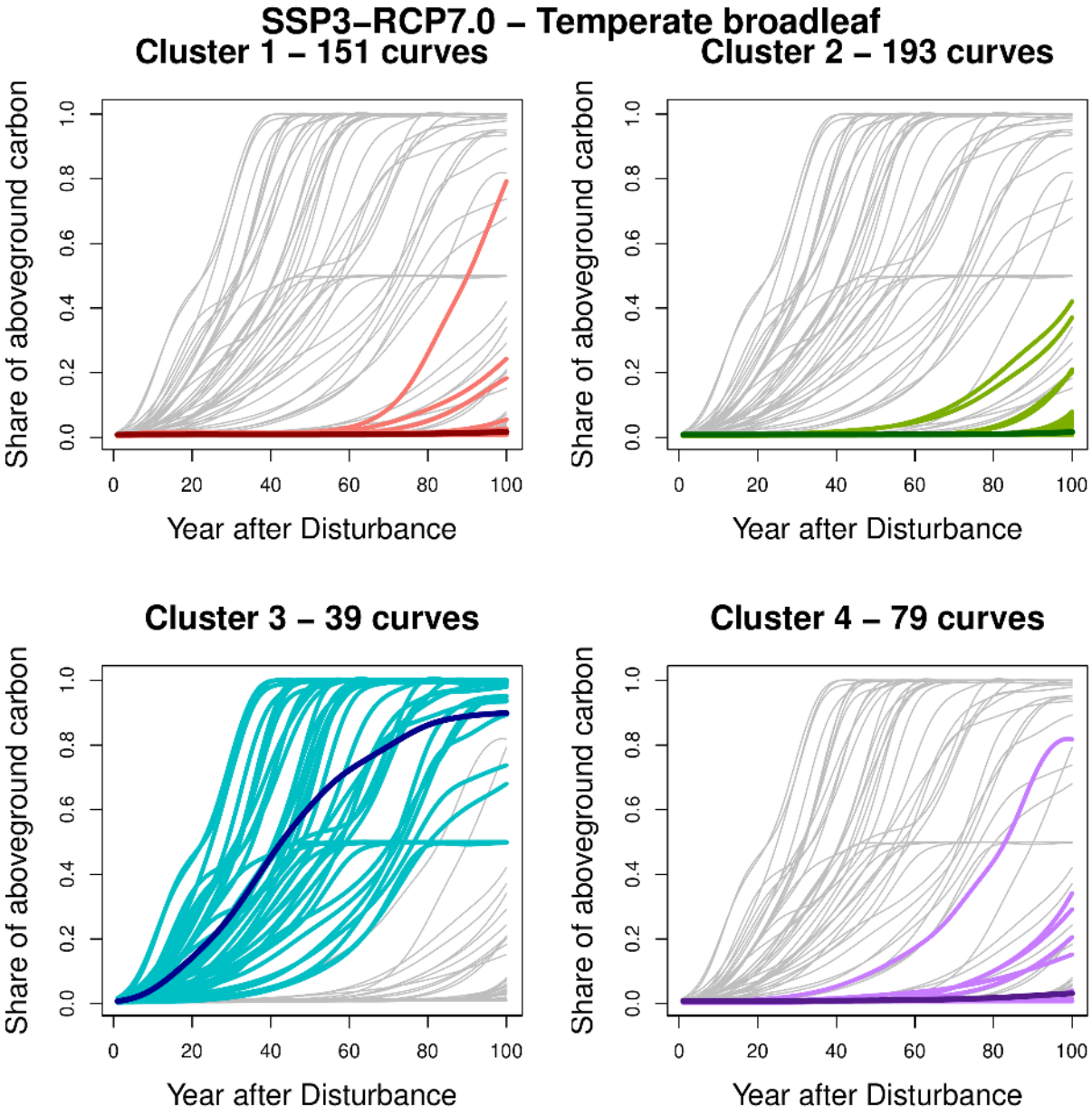


Figure 87: Clustered curves for scenario SSP3-RSCP7.0 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

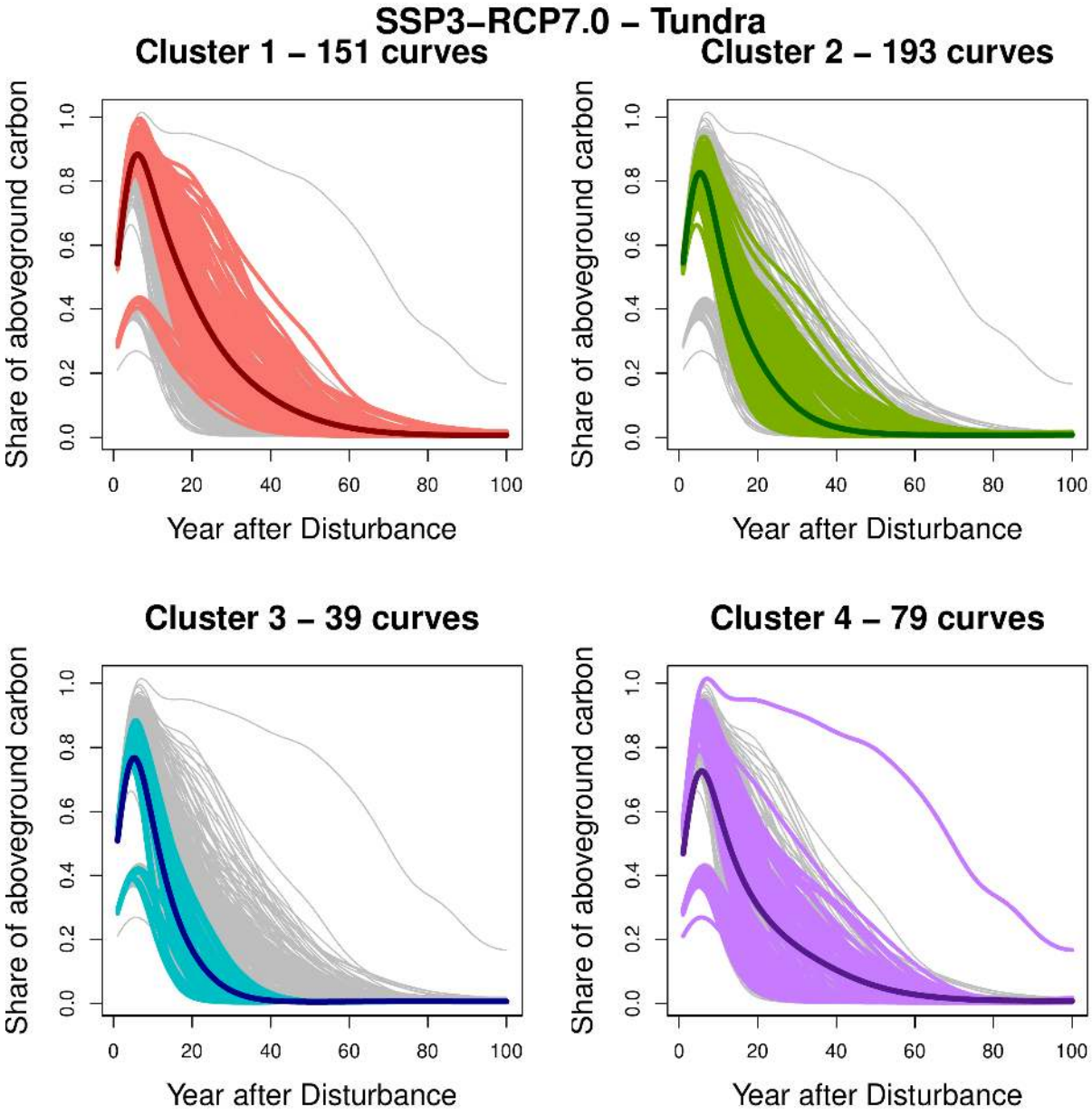


Figure 88: Clustered curves for scenario SSP3-RSCP7.0 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

A.2.4 Clustering for scenario SSP5-RCP8.5

To conclude, this final chapter on univariate clustering dives deeper into the clusters resulting from univariate FPCAs for the most extreme scenario SSP5-RCP8.5. For PFT *needleleaf evergreen* depicted in Figure 89, clusters 2 and 4 reflect grid cells with a low share of aboveground carbon of PFT *needleleaf evergreen*. Cluster 1 represents medium shares, that is, curves with a peak in the beginning of the recovery period and those with a constantly increasing share. Cluster 3 covers grid cells with a strong increase in aboveground carbon in the first decades after disturbance.

While cluster 3 is dominated by *needleleaf evergreen*, the largest cluster 2 is driven by high shares of PFT *pioneering broadleaf* as depicted in Figure 90. Cluster 3 comprises grid cells with nearly no *pioneering broadleaf*, while clusters 1 and 2 reflect curves with medium shares of aboveground carbon.

For *conifers (others)* shown in Figure 91, the same structure as in previous scenarios becomes apparent: two clusters, namely clusters 2 and 4 reflect similar curves with low shares of aboveground carbon, while cluster 1 and 3 cover grid cells with medium to high shares of *conifers (others)*.

Similar as in the scenarios before, one cluster, here cluster 4, comprises nearly all non-zero curves of *temperate broadleaf* (Figure 92). For this scenario, the lack of data is less pronounced since milder climate yields more *temperate broadleaf* as previous analyses revealed. As a consequence, the results for SSP5-RCP8.5 allow for a valid interpretation in contrast to the previous sections.

Finally, Figure 93 shows the clustered curves for PFT *tundra*. The patterns in all clusters are rather similar, with cluster 4 representing the sharpest decreases in aboveground carbon. Note that only clusters 1 and 2 represent grid cells with lower peaks.

Overall, the clusters derived for scenario SSP5-RCP8.5 are again mostly influenced by *needleleaf evergreen, pioneering broadleaf* and *conifers (others)*. Cluster 1 reflects mediate to high shares of *needleleaf evergreen, pioneering broadleaf* and *conifers (others)*, while cluster 2 is dominated by *pioneering broadleaf*. Cluster 3 is characterized by a high share of *needleleaf evergreen* together with high shares of *conifers (others)*. In contrast to the scenarios before, PFT *temperate broadleaf* gains in importance and mainly dominates cluster 4.

To conclude, this clustering approach gives first insights into grouping structures within each scenario, but does not consider correlations between the scenarios. This is covered by the MFPCA conducted in Section 4.4. The yielded clusters show some similarities between the scenarios and enforce the assumption that the vegetation composition after disturbances are mainly influenced by three PFTs: *needleleaf evergreen*, *pioneering broadleaf* and *conifers (others)*.

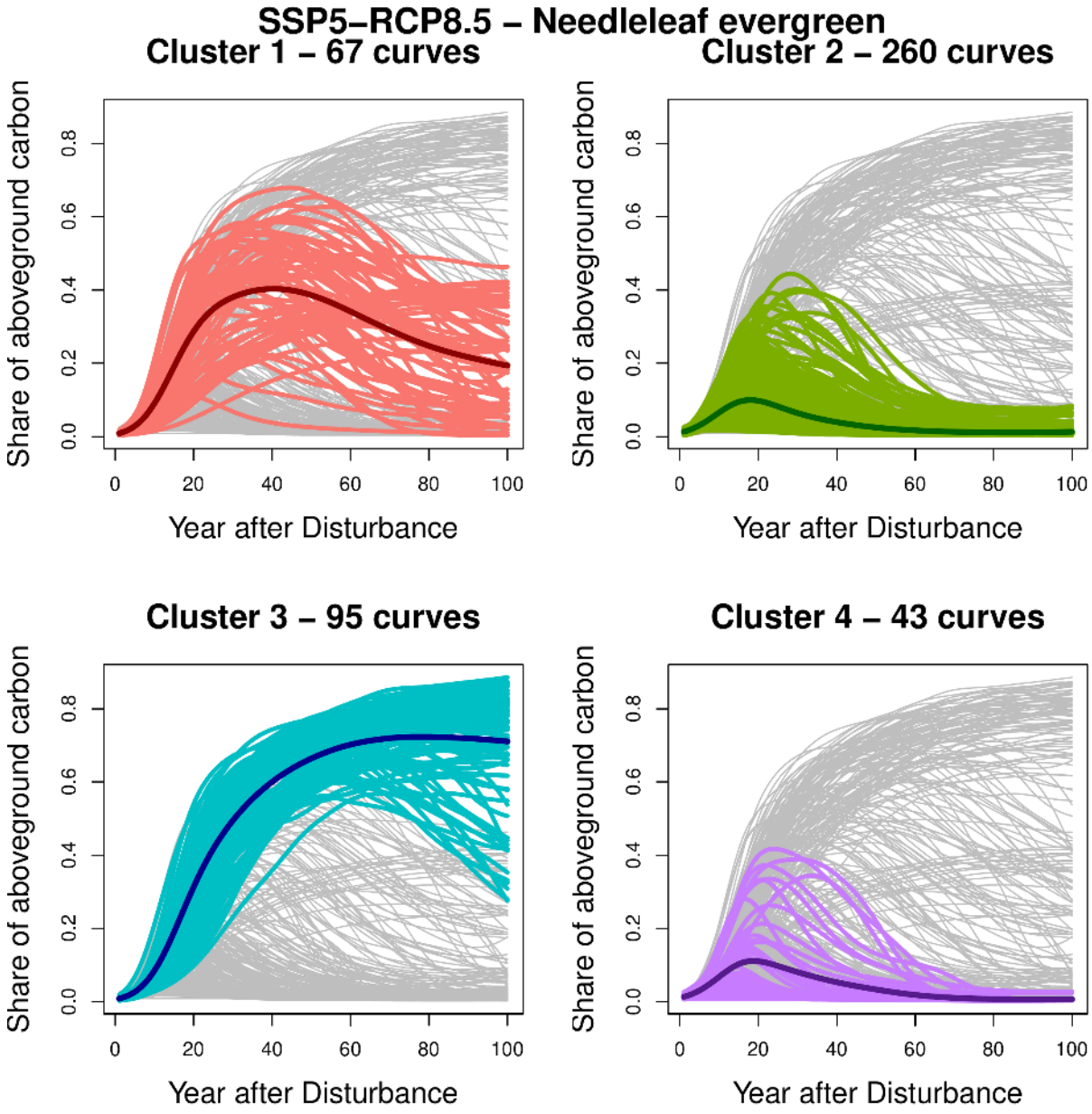


Figure 89: Clustered curves for scenario SSP5-RSCP8.5 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

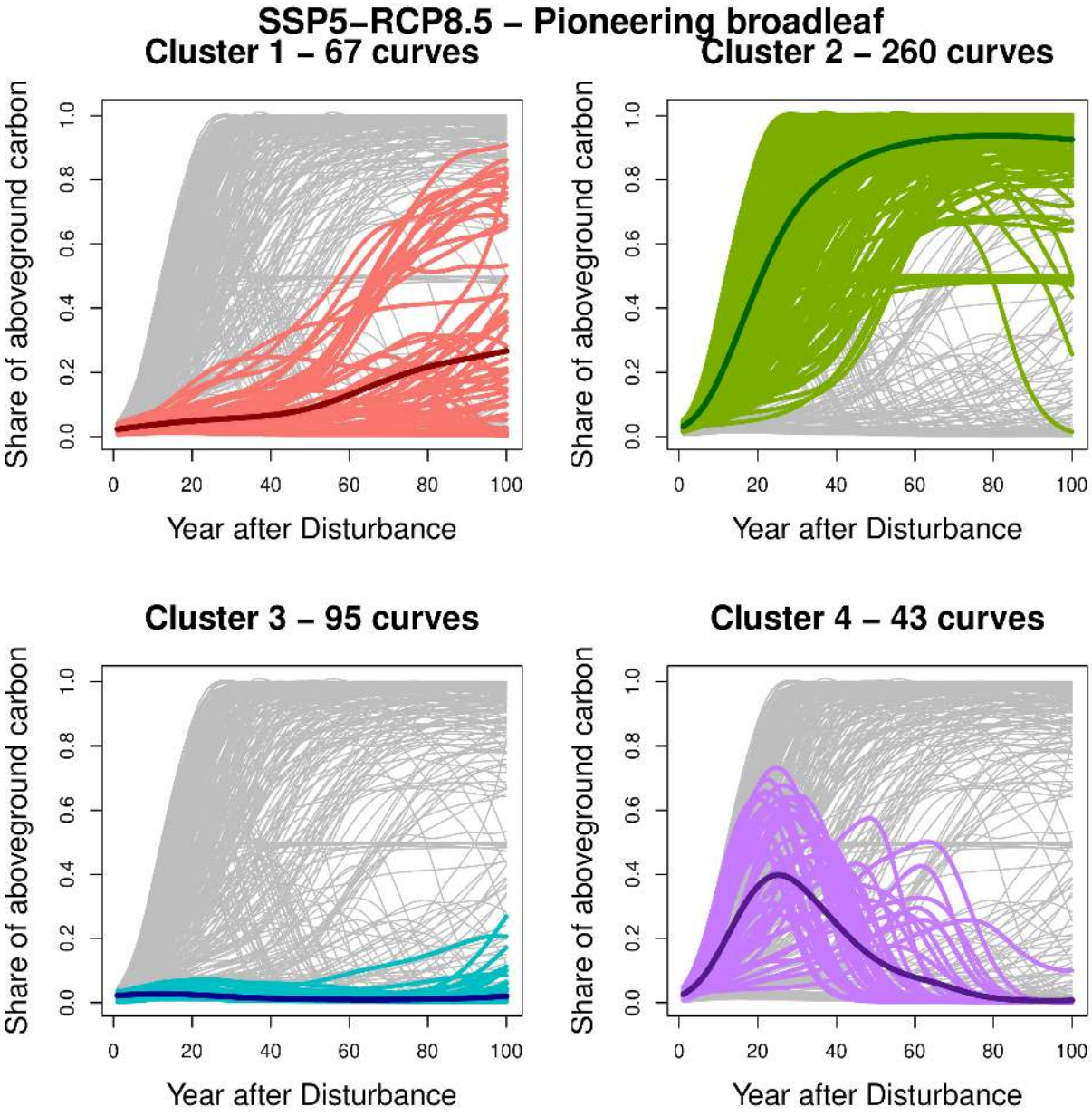


Figure 90: Clustered curves for scenario SSP5-RSCP8.5 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

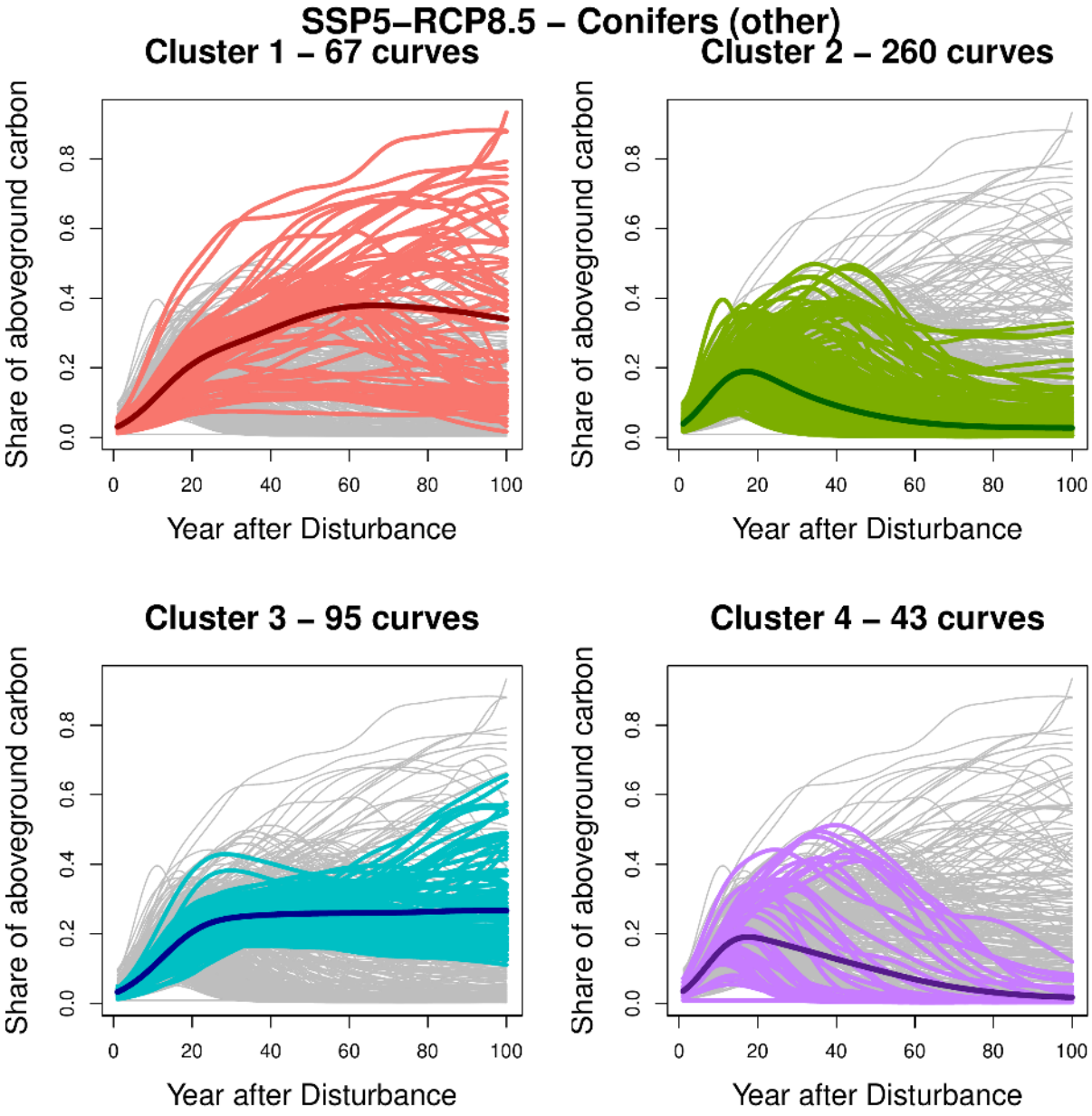


Figure 91: Clustered curves for scenario SSP5-RSCP8.5 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

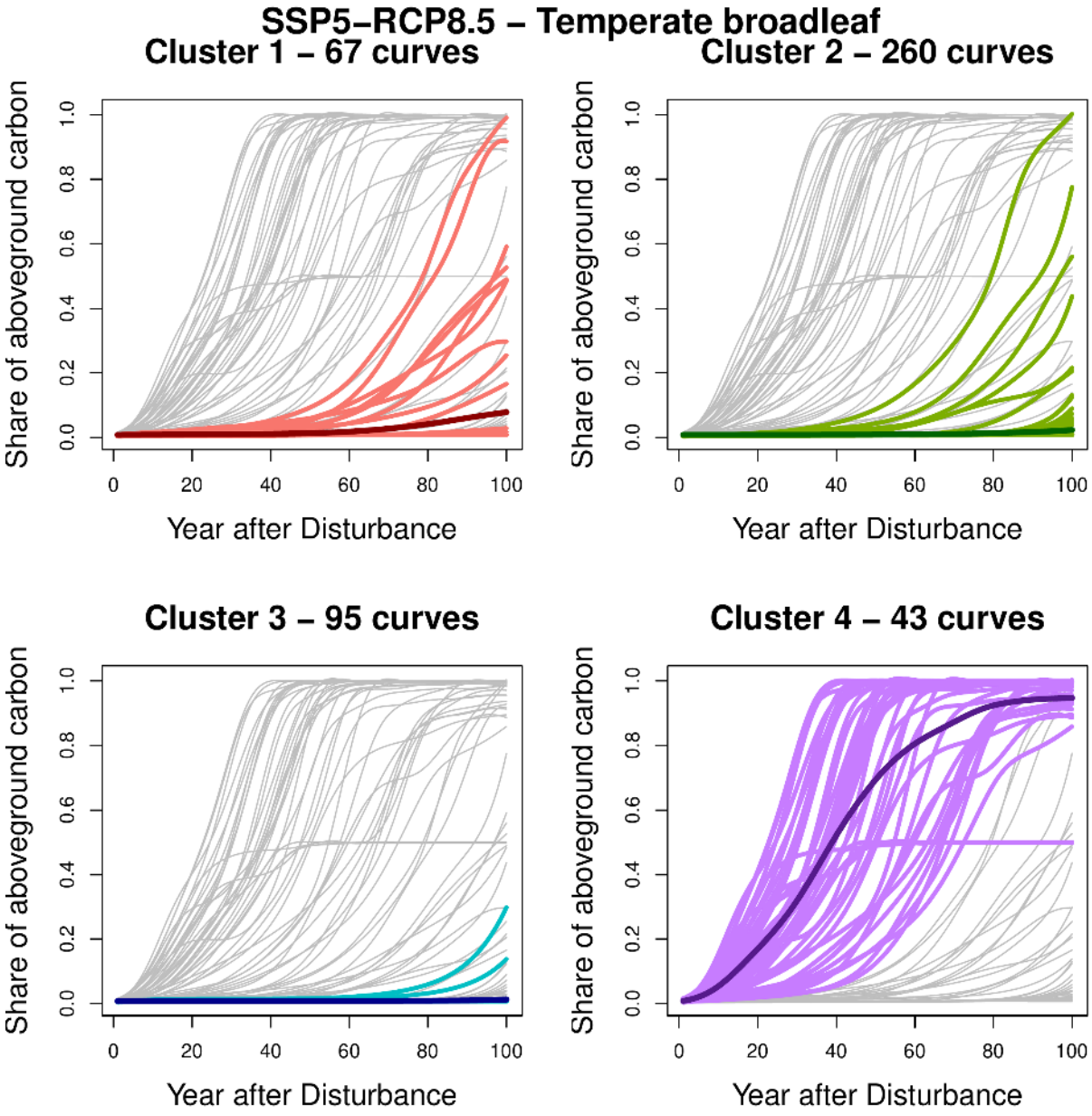


Figure 92: Clustered curves for scenario SSP5-RSCP8.5 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

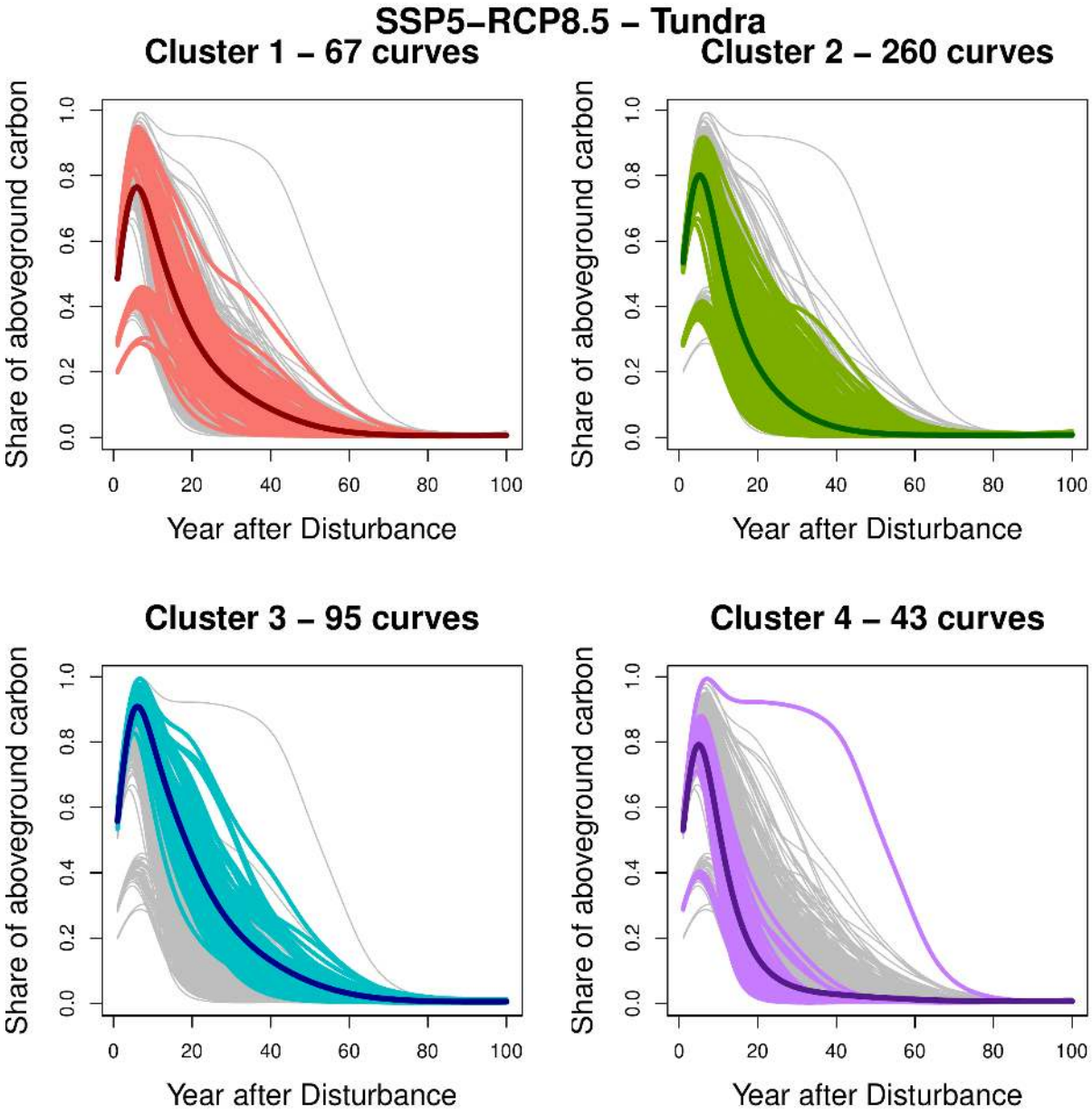


Figure 93: Clustered curves for scenario SSP5-RSCP8.5 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

A.3 Details on Clusters derived by MFPCA

In Section 4.5, the PC scores derived from an MFPCA based on the entire data set of 1803 disturbed grid cells, five tree species and 100 years of recovery were clustered for pattern recognition. Similar to Appendix A.2, this section examines which curve belongs to which cluster. Unlike the clusters derived by univariate FPCAs, the clusters are comparable across PFTs and scenarios as the MFPCA is based on the entire recovery trajectory data base. Recall that the detailed description of the cluster assignment in **??** revealed a rather unbalanced clustering, with two large clusters (2 and 4) dominated by *pioneering broadleaf* and coniferous species respectively, a medium-sized cluster (1) also dominated by needleleafed species, and a small cluster (3) dominated by *temperate broadleaf*, according to the cluster-specific mean proportions of aboveground carbon shown in Figure 37.

A.3.1 Clustering for the control scenario

Figure 94 shows the clustering of curves belonging to the control scenario and PFT *needleleaf evergreen*. While clusters 2 and 4 show nearly no occurrence of that PFT, cluster 4 is mainly driven by high shares of *needleleaf evergreen*. The medium sized cluster 1 shows moderate proportions of aboveground carbon.

Looking at PFT *pioneering broadleaf* in Figure 95 shows two clusters dominated by this species: cluster 1 and 2, which is in line with the result obtained in Figure 37. However, the dynamics of the increase in aboveground carbon are different. While cluster 2 includes locations with a strong increase in the first decades after the disturbance, the increase in cluster 1 is more flat. Clusters 3 and 4 show only low occurrences of *pioneering broadleaf*.

The PFT-specific mean proportions depicted in Figure 37 show medium proportions of *conifers (other)* in clusters 1 and 4, and this is supported by the cluster-wise curves in Figure 96. The behaviour of the curves in clusters 1 and 2 is very similar, but this scenario and PFT combination represents only a small fraction of the curves in the respective clusters.

For *temperate broadleaf* there is again the difficulty of lack of data for the control scenario. Figure 97 shows the clustered curves for this species, and as expected, cluster 3 contains most of the curves with non-zero aboveground carbon proportions. The mean cluster-specific shares in the other three clusters are close to zero, indicating no substantial occurrence of *temperate broadleaf*.

Cluster 4, the largest cluster dominated by *needleleaf evergreen* and *conifers (other)* show the highest variation in shares of *tundra*, as Figure 98 visualizes. It comprises locations with a lower peak in the first years after disturbance, as does cluster 1, which covers medium sized proportions of aboveground carbon. Clusters 1 and 2 show slightly sharper decreases in *tundra* after the high peak in the beginning of the recovery period.

In summary, the control scenario is mainly represented by clusters 1 and 4, since it plays only a minor role in the two remaining scenarios in terms of number of grid cells. Clusters

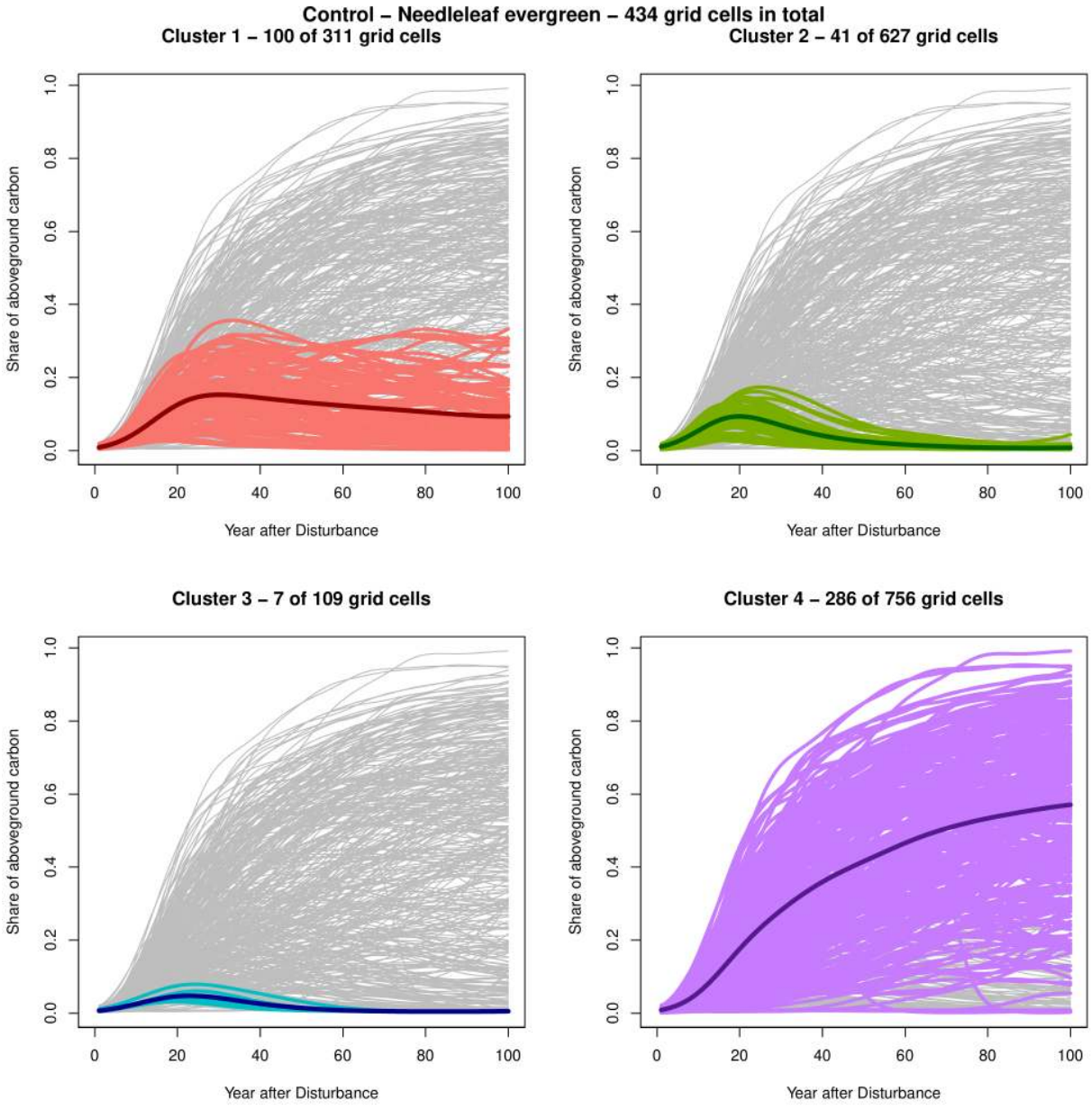


Figure 94: Clustered curves for the control scenario and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

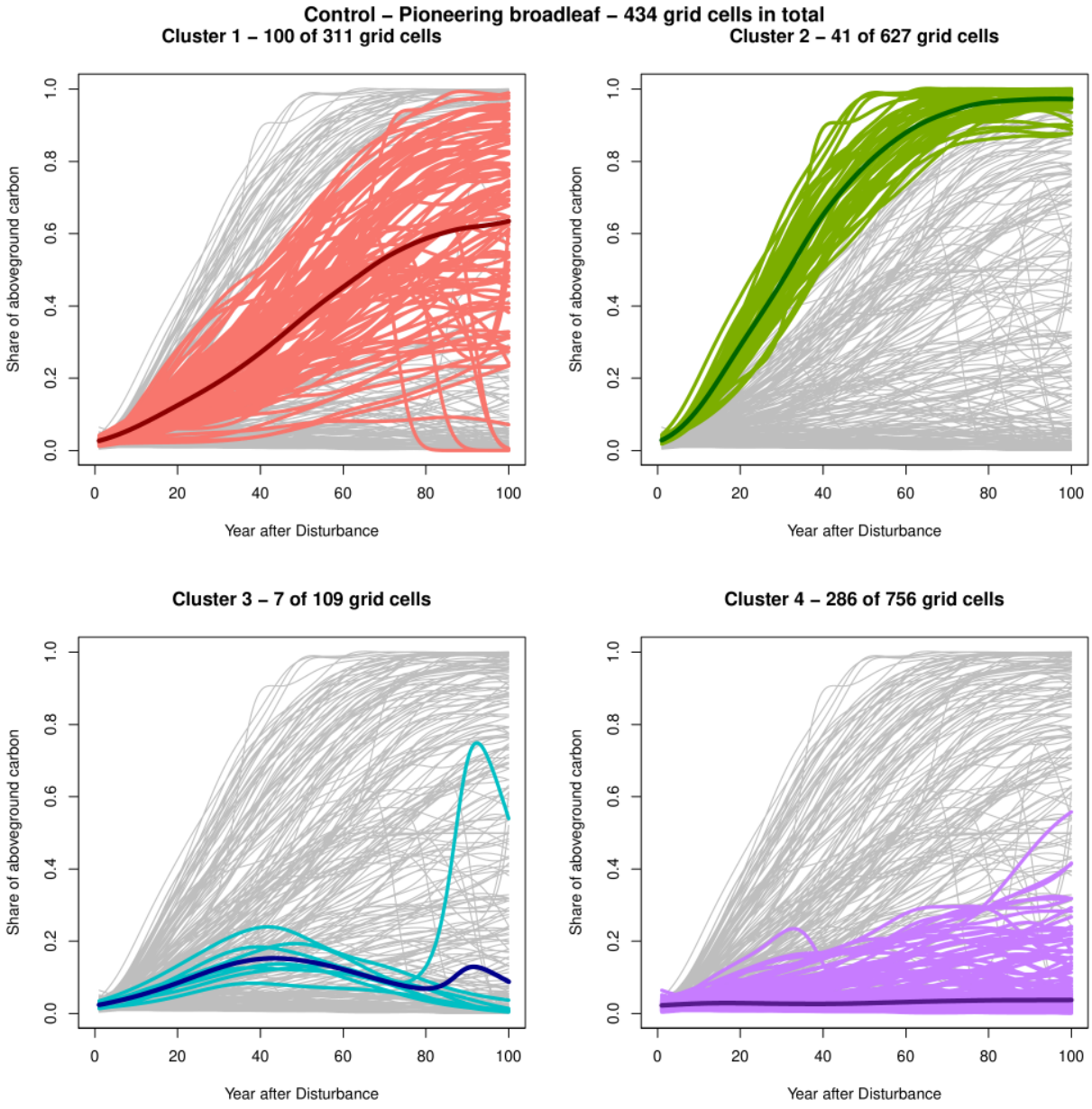


Figure 95: Clustered curves for the control scenario and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

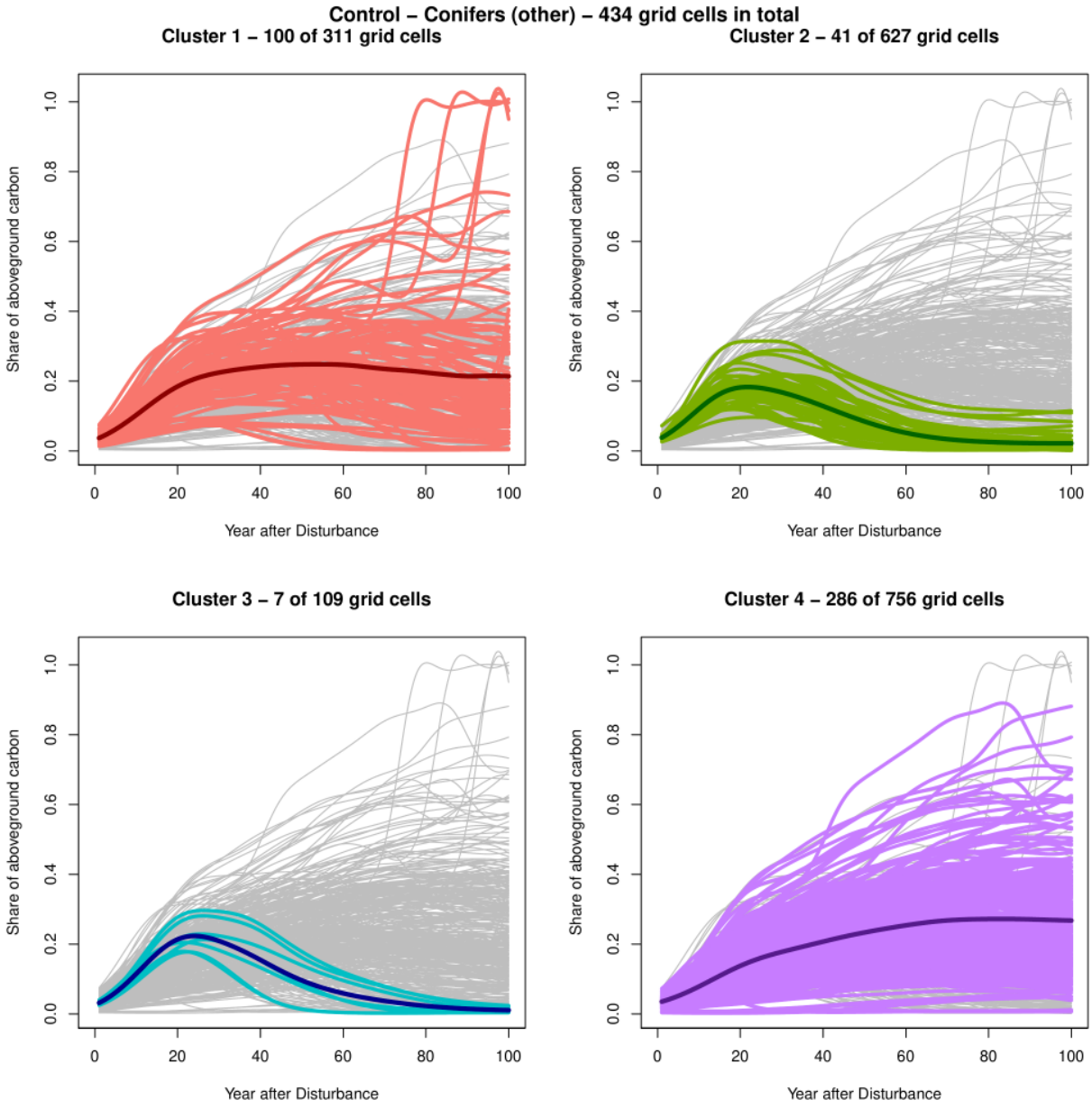


Figure 96: Clustered curves for the control scenario and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

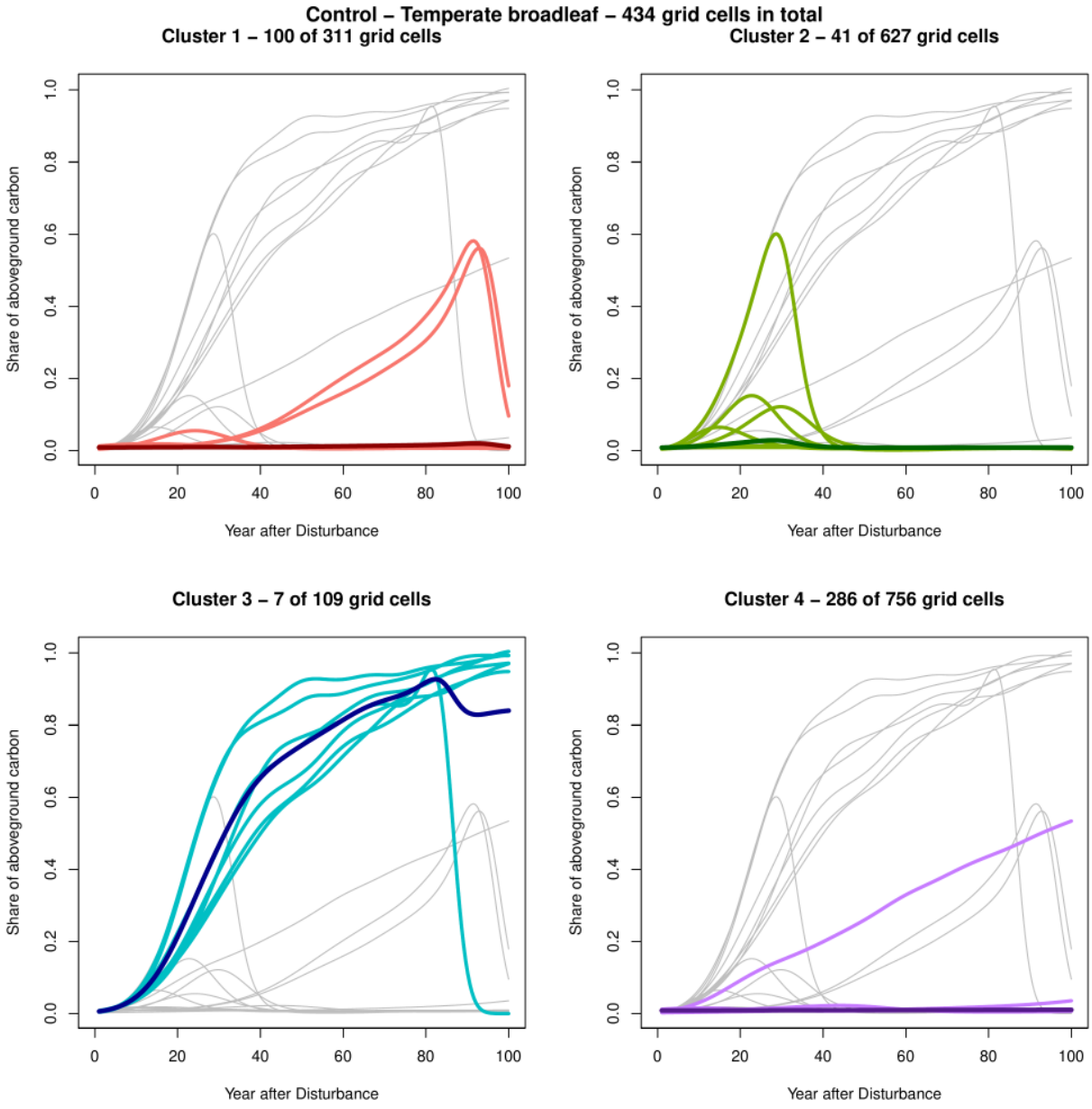


Figure 97: Clustered curves for the control scenario and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

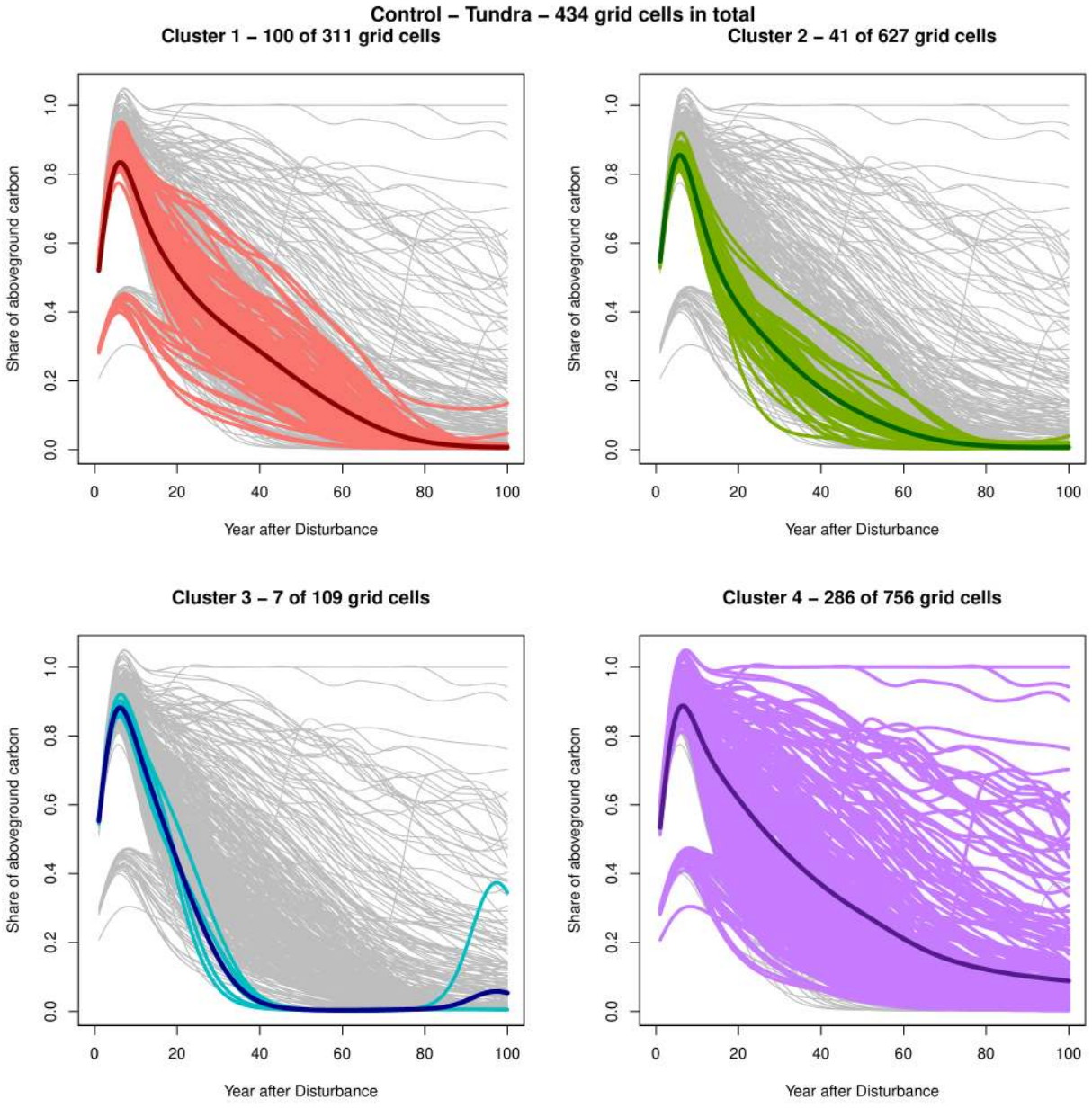


Figure 98: Clustered curves for the control scenario and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

1 and 2 are dominated by high shares of *pioneering broadleaf*, while cluster 3 covers most of the non-zero curves of *temperate broadleaf*. Cluster 4 is mainly driven by high shares of *needleleaf evergreen, conifers (other)* and *tundra*.

A.3.2 Clustering for scenario SSP1-RCP2.6

The visualisation of the clustered curves for the SSP1-RCP2.6 scenario indicating low warming and PFT *needleleaf evergreen* is shown in Figure 99. Here, cluster 4 is mainly dominated by high proportions of aboveground carbon, while *needleleaf evergreen* plays a minor role in the other three clusters. In cluster 2, the mean proportion peaks only slightly and is close to zero throughout the entire recovery period. Clusters 1 and 3 represent medium sized shares of *needleleaf evergreen*.

Figure 37 revealed that clusters 1 and 2 are mainly dominated by *pioneering broadleaf*, and this is supported by the clustered curves shown in Figure 100. Interestingly, the behaviour of these curves is very different between these clusters: while in cluster 2 the increase in aboveground carbon is sharp and reaches its maximum after only a few decades of recovery, the increase in cluster 1 is slower and does not reach maximum values. This result is in line with that of Section 4.5, since cluster 2 is exclusively dominated by *pioneering broadleaf*, while in cluster 1 some other species are still present in the last decades of recovery. Cluster 3, the cluster dominated by *temperate broadleaf*, shows medium sized curves, while in cluster 4 *pioneering broadleaf* is barely present.

Similar to the control scenario, high proportions of *conifers (other)* are mainly covered by clusters 1 and 4, as shown in Figure 101. Clusters 2 and 3 show curves with a small peak after about 20 years on average and decreasing proportions thereafter.

Figure 102 confirms that cluster 3 contains most of the curves with non-zero proportions of *temperate broadleaf*. Although cluster 1 contains some curves with high peaks in aboveground carbon, the mean function of this cluster is close to zero, as are those of the remaining two clusters 2 and 4.

For *tundra* in Figure 103, only cluster 2 contains no curves with a small peak immediately after the disturbance. Clusters 1 and 4 are almost indistinguishable in terms of curve behaviour, and cluster 3, the smallest cluster, consists mainly of grid cells with a sharp decrease in *tundra* after the peak.

The overall result is the same as for the control scenario, which is to be expected since the entire data set is used to perform the MFPCA underlying the clustering. Clusters 1 and 2 are dominated by *pioneering broadleaf*, with cluster 1 having moderate proportions of aboveground carbon of *needleleaf evergreen* and *conifers (other)*. Cluster 3 includes almost all non-zero curves of *temperate broadleaf*, while cluster 4 represents curves with high proportions of *tundra* and needleleafed species.

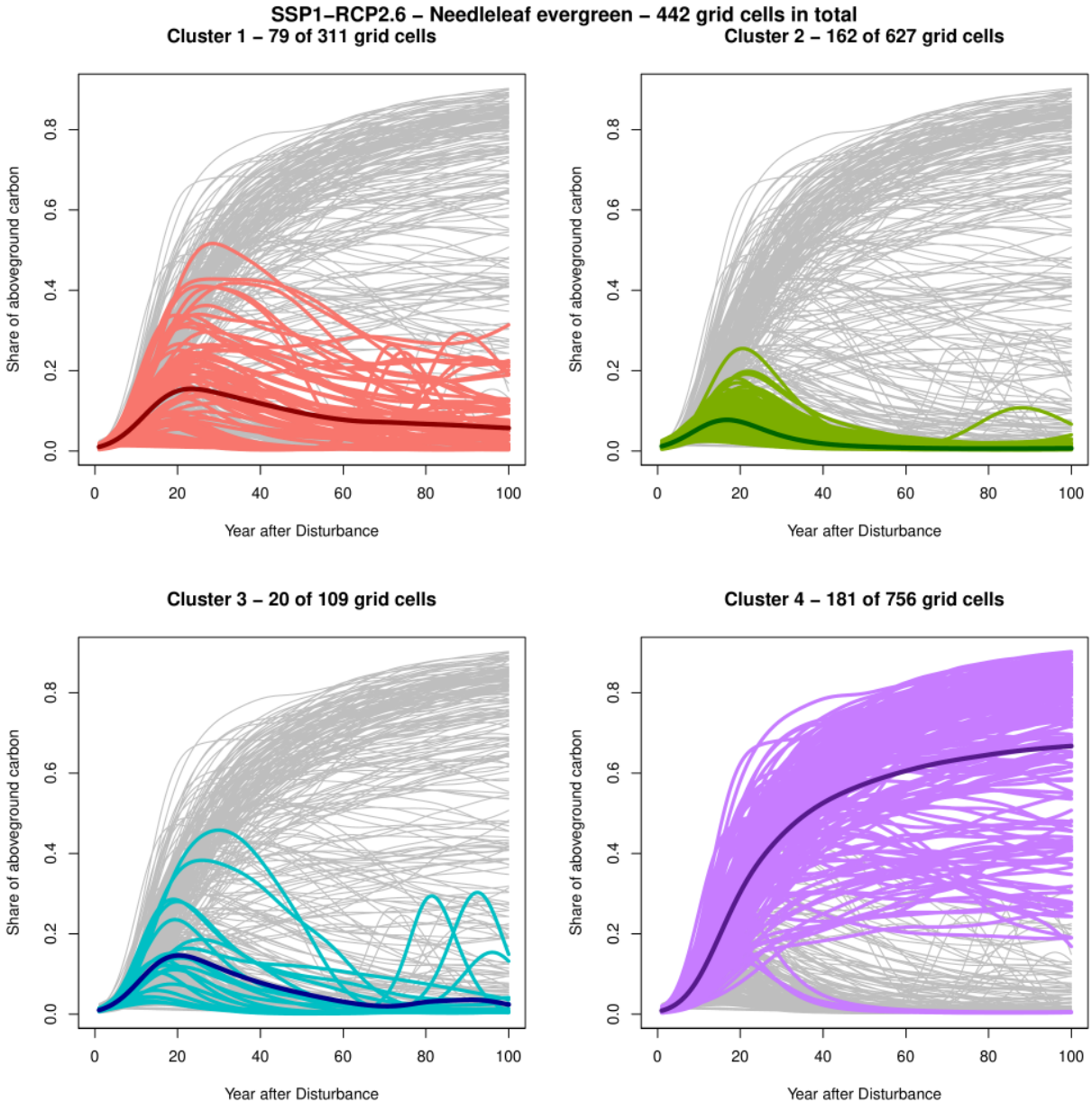


Figure 99: Clustered curves for scenario SSP1-RCP2.6 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

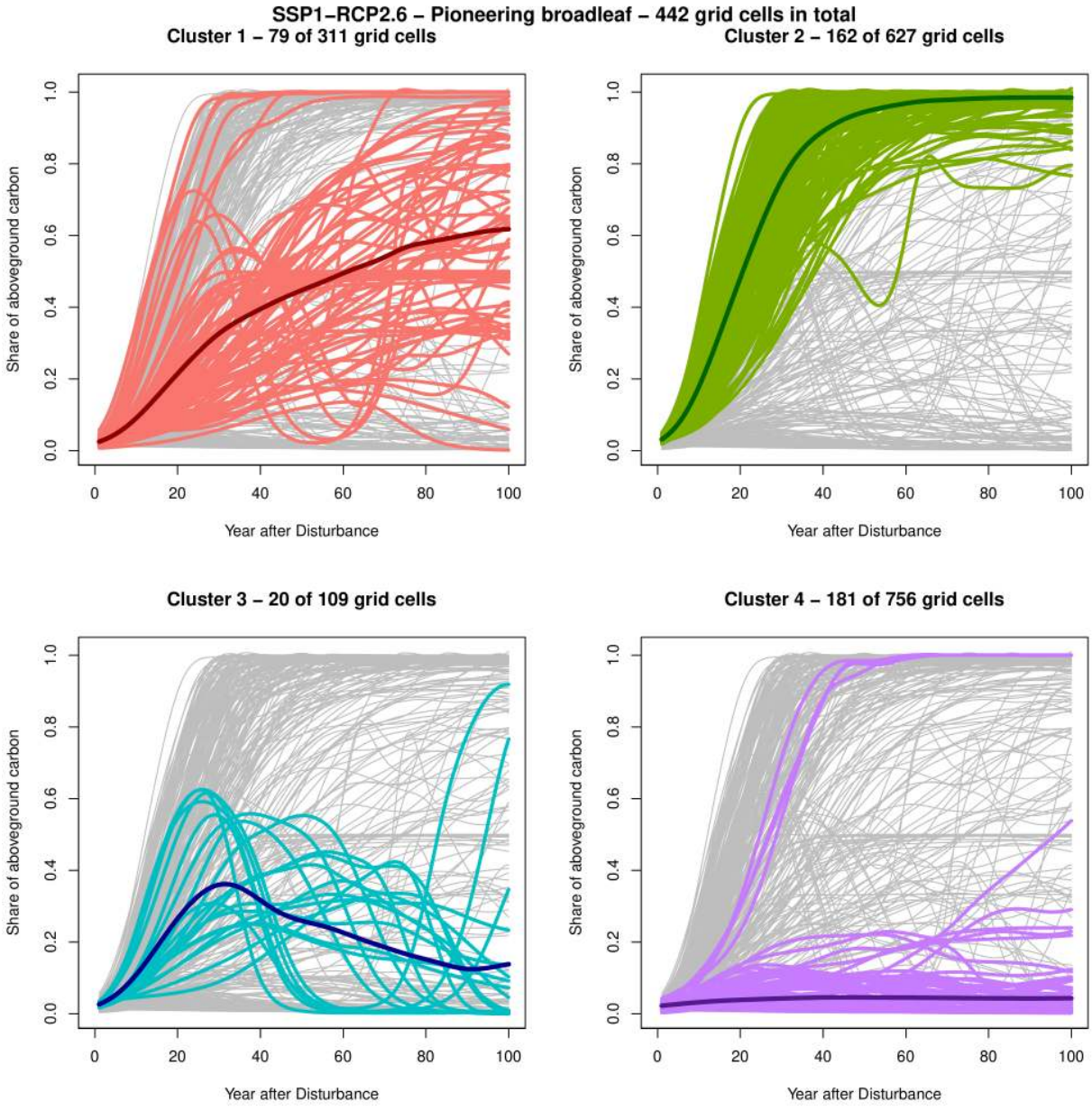


Figure 100: Clustered curves for scenario SSP1-RCP2.6 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

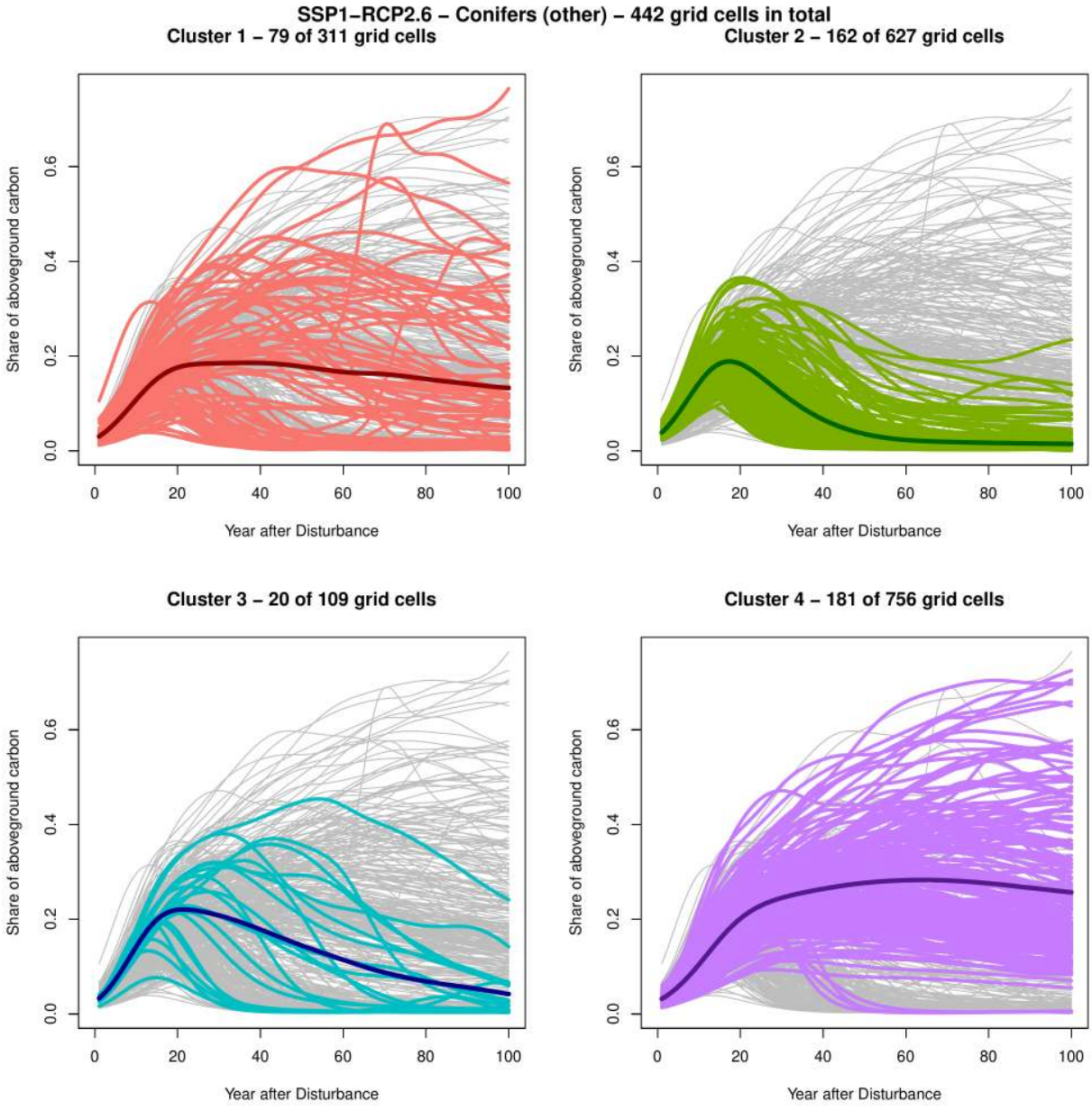


Figure 101: Clustered curves for scenario SSP1-RCP2.6 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

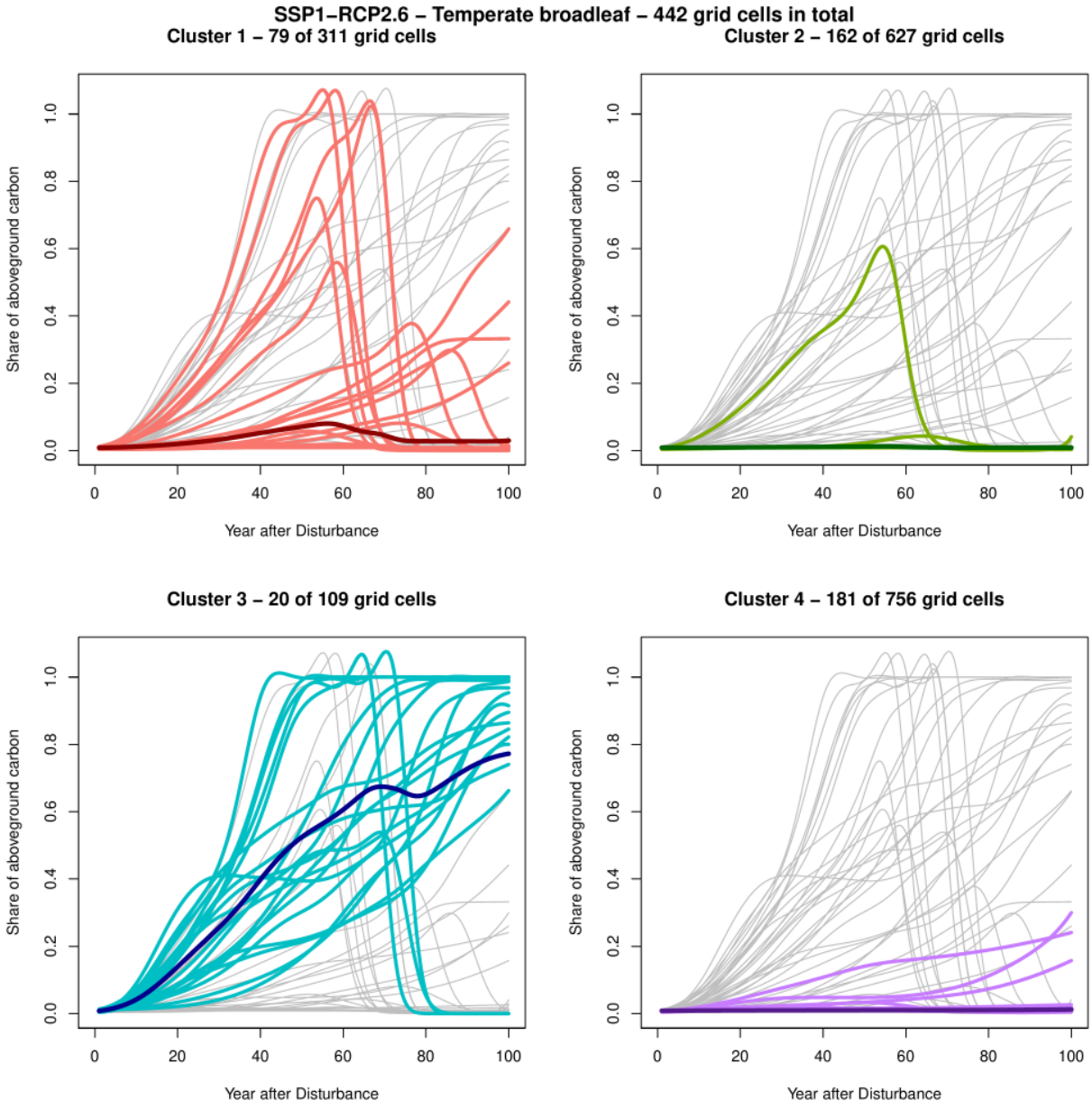


Figure 102: Clustered curves for scenario SSP1-RCP2.6 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

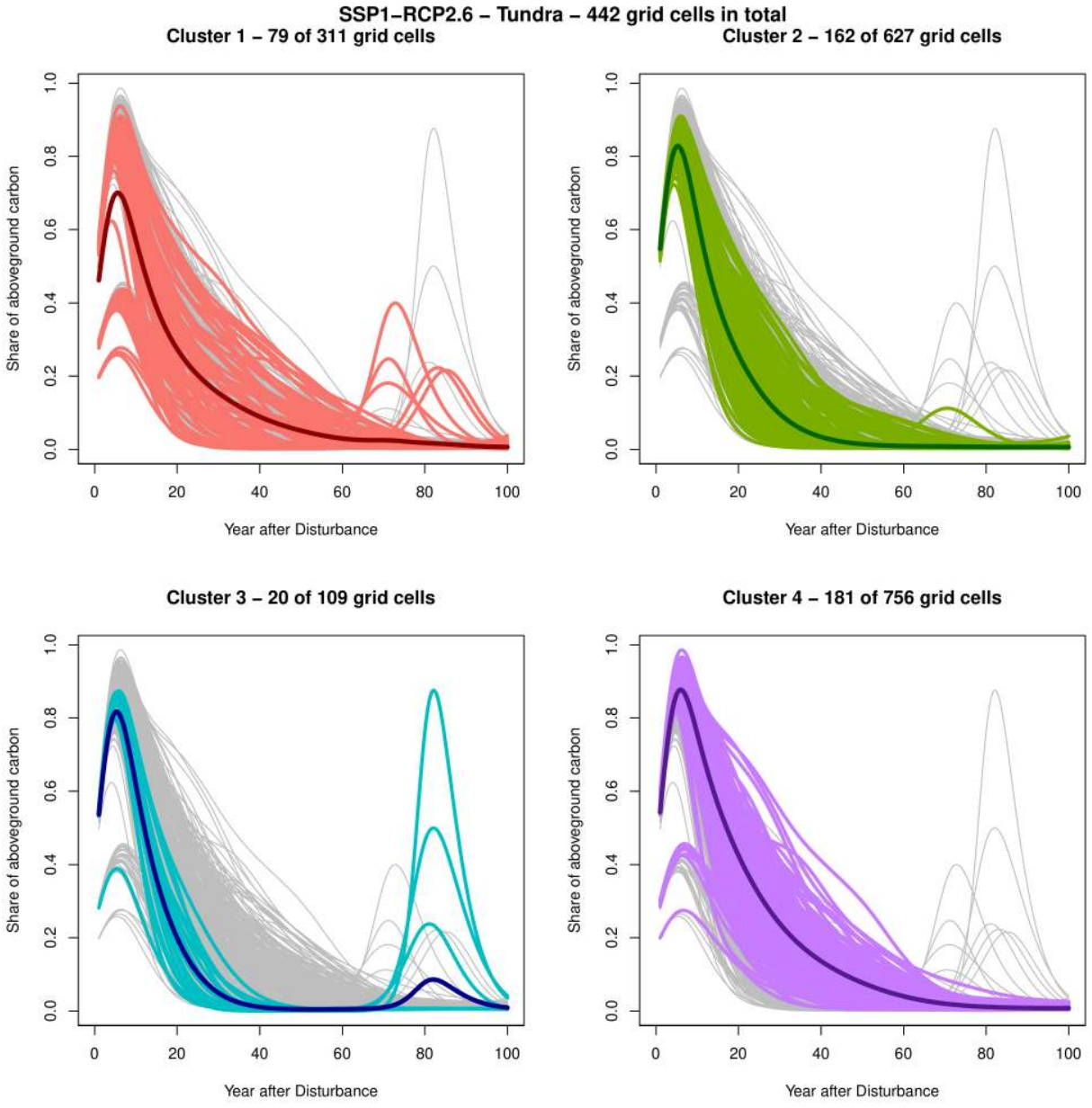


Figure 103: Clustered curves for scenario SSP1-RCP2.6 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

A.3.3 Clustering for scenario SSP3-RCP7.0

As for the two scenarios considered previously, cluster 4 of the SSP3-RCP7.0 scenario is mainly dominated by *needleleaf evergreen*, as visualised by Figure 104. Clusters 2 and 3, dominated by *pioneering broadleaf* and *temperate broadleaf*, represent small proportions of aboveground carbon with a small peak after about 20 years of recovery. Cluster 1 includes all moderate increases in *needleleaf evergreen*.

Again, clusters 1 and 2 are dominated by *pioneering broadleaf* (Figure 105), with cluster 2 reaching values close to one in a sharp increase after the disturbance. Cluster 4 shows almost no occurrence of *pioneering broadleaf* on average, while cluster 3 comprises grid cells with a small peak in aboveground carbon after a few decades of recovery and a decline thereafter. This is due to displacement by *temperate broadleaf*.

For the PFT *conifers (other)* shown in Figure 106, clusters 1 and 4 show medium to high proportions of aboveground carbon, while clusters 2 and 3 represent grid cells with a small peak in the early decades of recovery and a subsequent decline. Again, the proportion in cluster 1 is higher on average than in cluster 2, as cluster 2 is dominated exclusively by *pioneering broadleaf* in the last decades of recovery.

As the SSP3-RCP7.0 scenario reflects high climate warming, the occurrence of *temperate broadleaf* increases. This is visualised in Figure 107, where again cluster 3 contains most of the curves with high proportions of *temperate broadleaf*. Note that the number of grid cells within cluster 3 increases with increasing radial forcing, while the number of elements decreases, especially in cluster 1. This highlights the trend towards more broadleafed species for more extreme climate warming.

Figure 108 shows the clustered curves for PFT *tundra*. For the first time, all four clusters contain grid cells with a small initial peak of aboveground carbon. Similarly, the third cluster represents grid cells with a rapid decline in *tundra*, while cluster 1 shows the lowest average peak. Clusters 2 and 4, the two largest clusters, differ in the speed of decline after the peak.

Overall, the patterns and curve behaviour are very similar to the two scenarios considered earlier. Clusters 1 and 2 are again dominated by *pioneering broadleaf*, although the latter is generally more variable in terms of tree species. Cluster 3 gains in importance due to climate warming and is dominated by *temperate broadleaf*. In contrast, cluster 4, dominated by *needleleaf evergreen* and *conifers (other)*, begins to play a less prominent role as a milder climate favours more broadleafed species.

A.3.4 Clustering for scenario SSP5-RCP8.5

This final section of the appendix concludes with further details of the clustered curves for the most extreme climate scenario, SSP5-RCP8.5. Figure 109 again shows the dominance of *needleleaf evergreen* in the fourth cluster, while clusters 2 and 3 represent locations with a small peak after about 20 years and a decline thereafter. Cluster 1 reflects grid cells with a

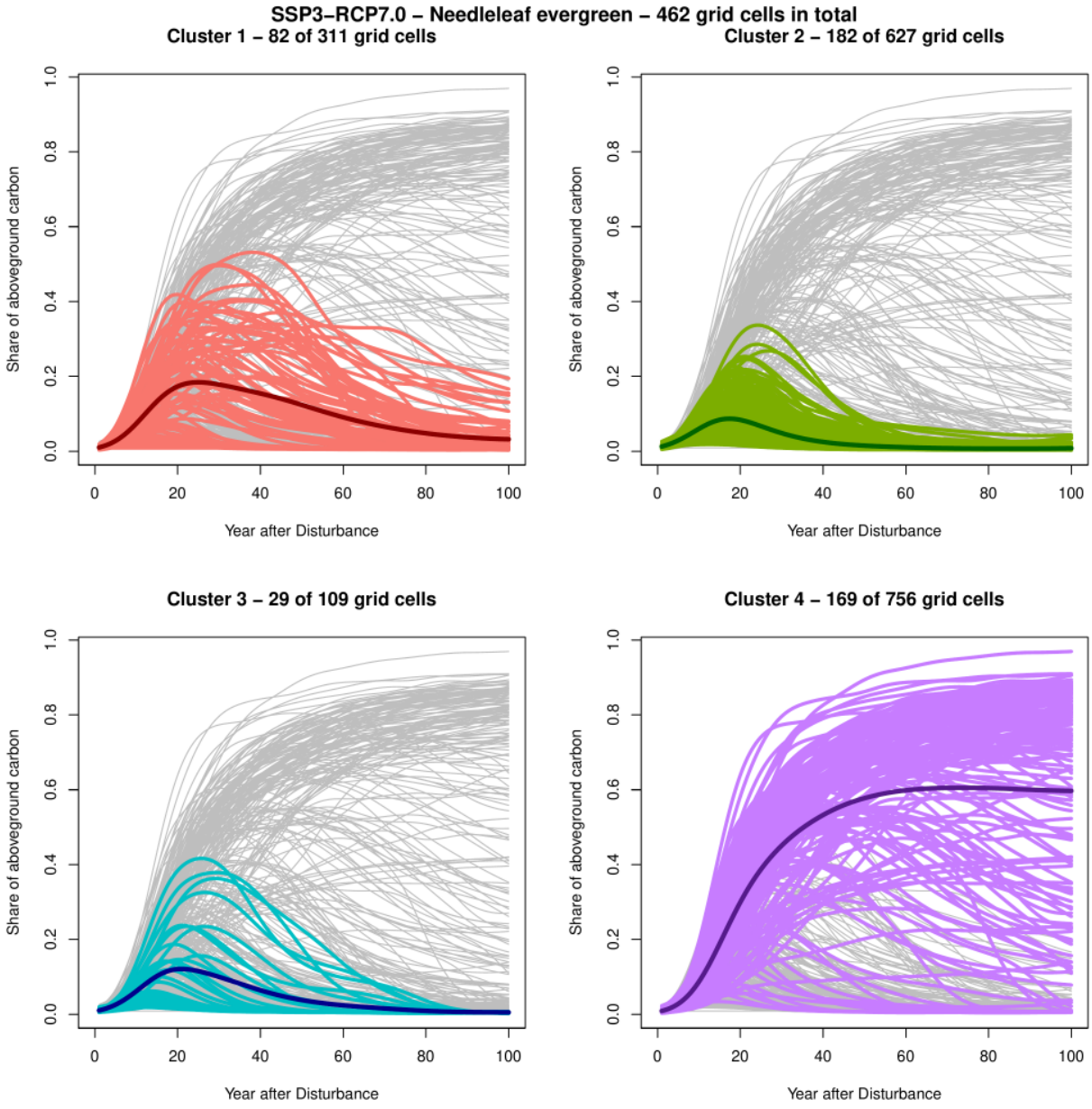


Figure 104: Clustered curves for scenario SSP3-RCP7.0 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

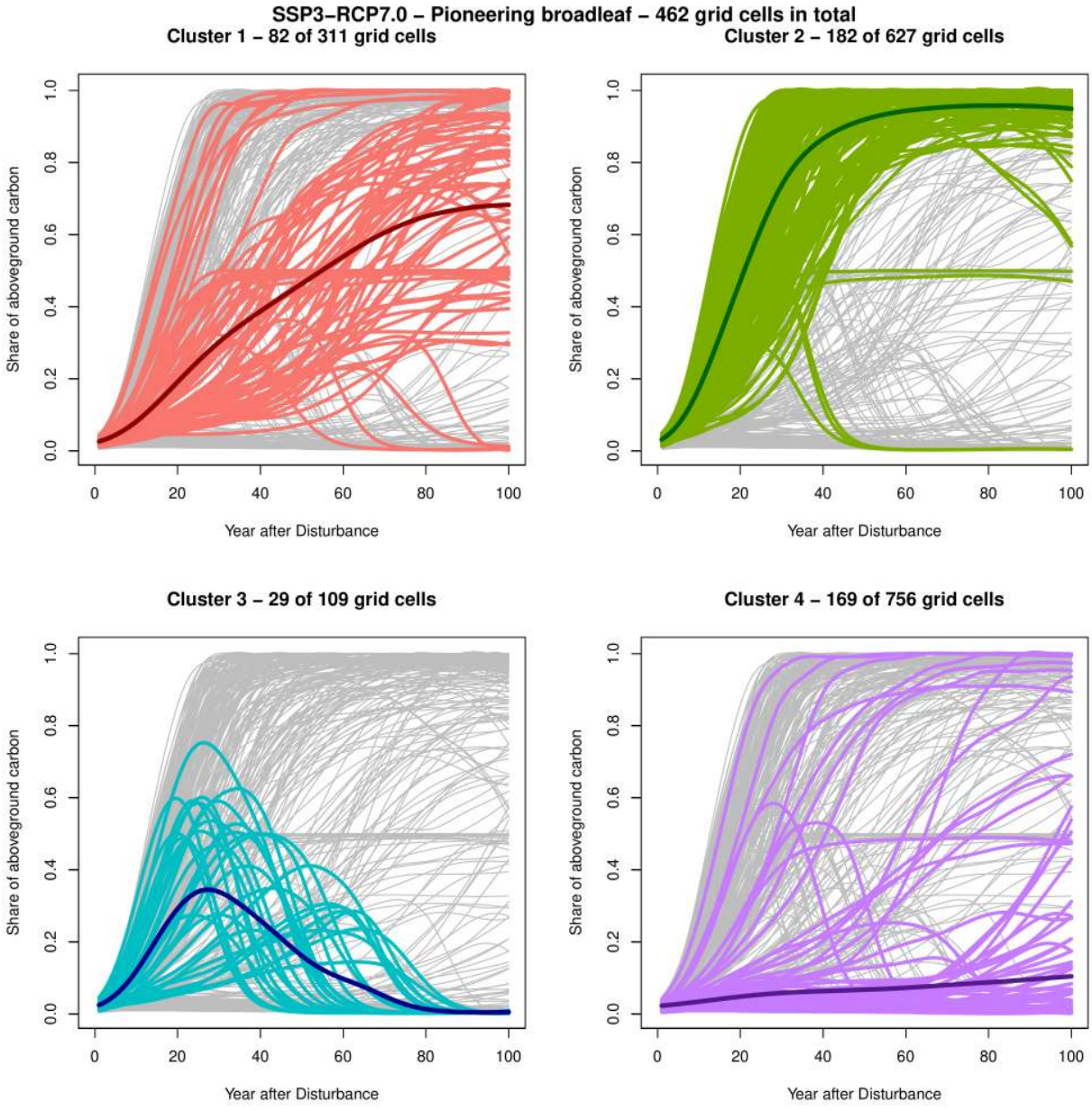


Figure 105: Clustered curves for scenario SSP3-RCP7.0 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

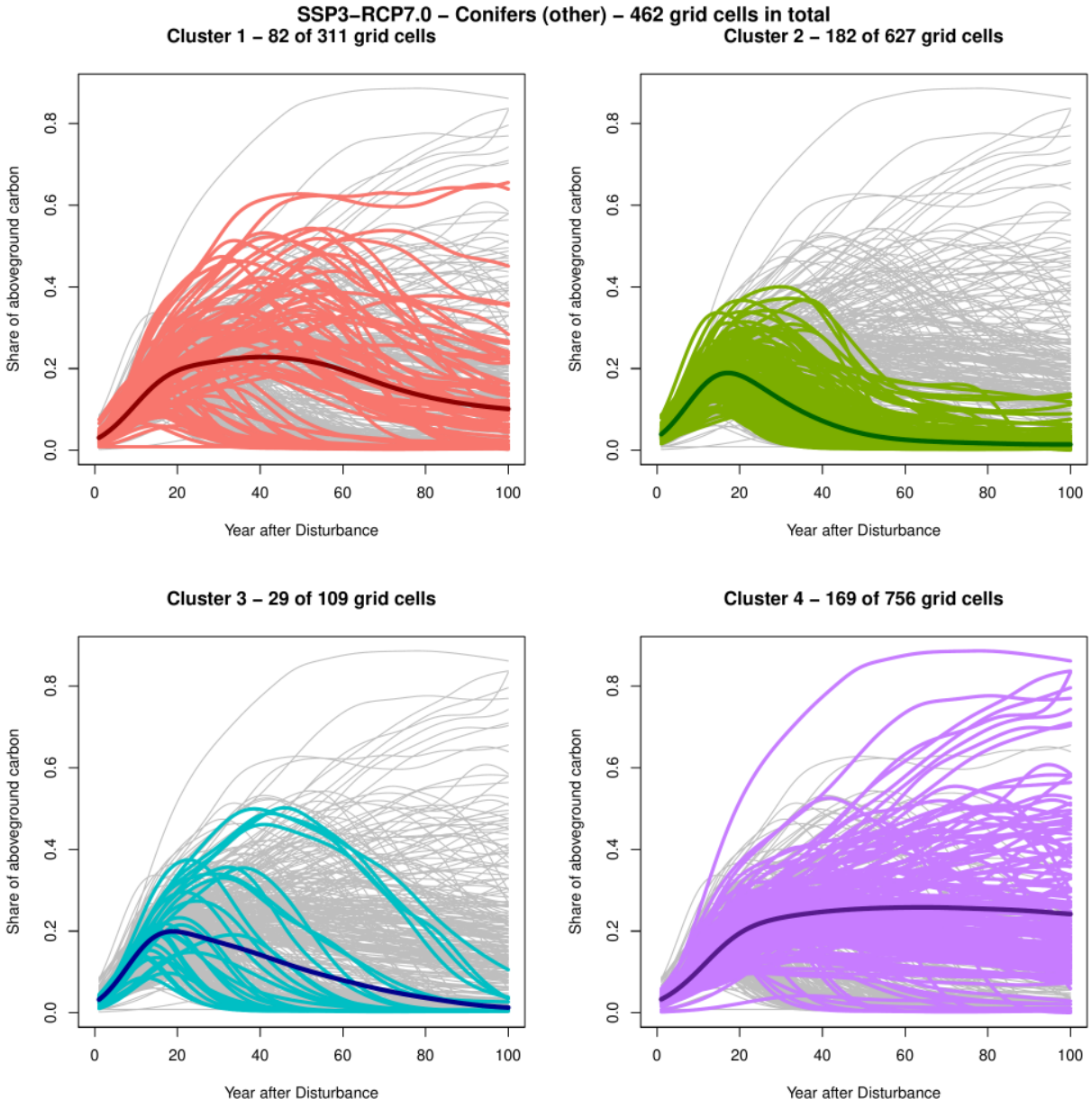


Figure 106: Clustered curves for scenario SSP3-RCP7.0 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

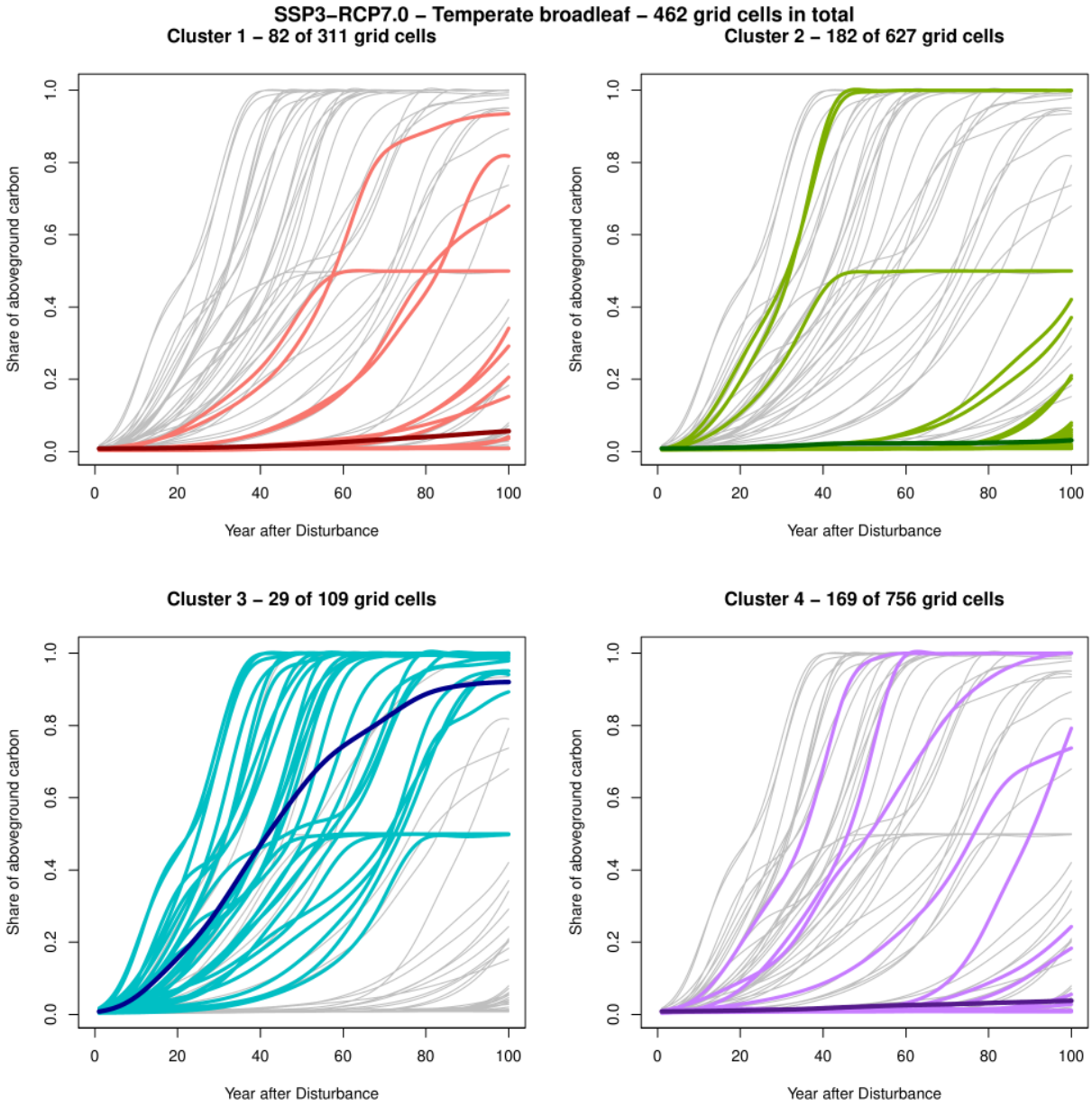


Figure 107: Clustered curves for scenario SSP3-RCP7.0 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

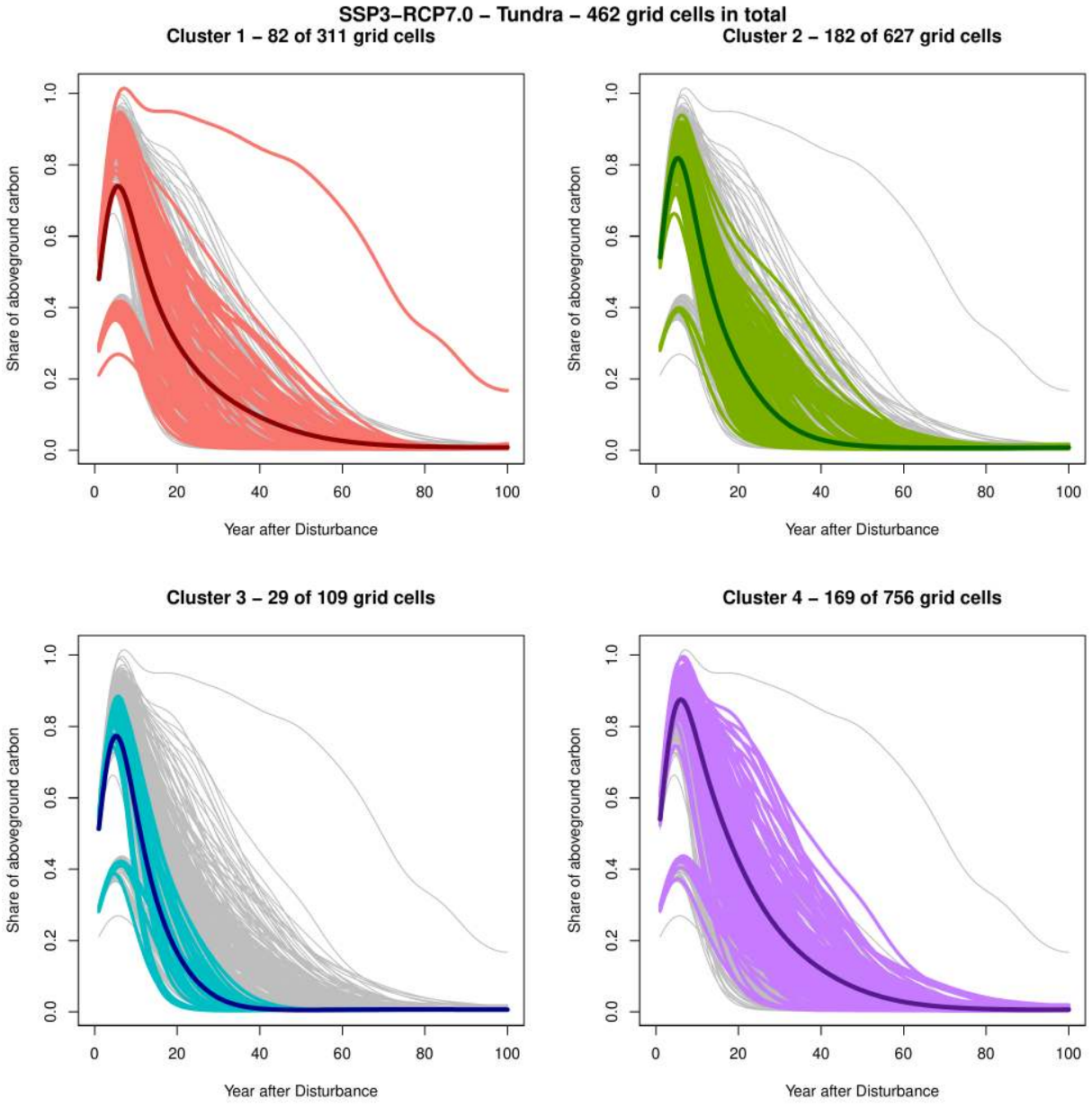


Figure 108: Clustered curves for scenario SSP3-RCP7.0 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

moderate amount of aboveground carbon. On the one hand, while cluster 4 comprised about 66% of the grid cells in the control scenario, it now covers only about 31% of the disturbed grid cells. On the other hand, the proportion of cluster 3 increased strongly from 1.6% to now 9.4%. This underlines the decreasing importance of *needleleaf evergreen* for increasing radial forcing.

The behaviour of the clusters with respect to *pioneering broadleaf* shown in Figure 110 is very similar to the previous scenario. Clusters 1 and 2 show large proportions of *pioneering broadleaf*, with cluster 2 being almost completely dominated by this PFT. Cluster 3 covers grid cells with a moderate peak in the third decade of recovery, while cluster 4 shows the lowest amount of aboveground carbon, being dominated by needleleafed tree species.

Conifers (other) is one of the dominant vegetation types in clusters 1 and 4, the two clusters concentrated on grid cells disturbed in the control scenario, as visualized in Figure 111. Clusters 2 and 4 show both a small peak in the second decade after disturbance and a rapid and moderate decline, respectively.

As the SSP5-RCP8.5 scenario indicates very high warming, the PFT *temperate broadleaf* has a higher proportion of above-ground carbon. Figure 112 shows that again cluster 3 represents the grid cells dominated by *temperate broadleaf*, while the mean curves in clusters 2 and 4 are barely above zero. Cluster 1 also contains some curves with high proportions, but on average is still far from the amounts of *temperate broadleaf* reached by cluster 3.

Finally, Figure 113 shows the clustered curves for the SSP5-RCP8.5 scenario and PFT *tundra*. All four clusters contain curves with a low peak at the beginning of the study period, but in general the clusters are difficult to distinguish. This confirms the result obtained for the previous scenarios, that the aboveground carbon of *tundra* is less influential in finding patterns in the recovery trajectories.

In summary, clusters 1 and 2 show the dominance of *pioneering broadleaf* in all scenarios, with some other tree species also occurring in cluster 1. Cluster 3 represents the prevalence of *temperate broadleaf*. Looking at the cluster composition, we can see that the importance of this cluster increases with increasing radial forcing. In contrast, cluster 4, which focuses on high proportions of coniferous trees, includes many grid cells that are disturbed in the control scenario and declines in importance for more extreme scenarios. The results in this section underline the results obtained in Section 4.5.

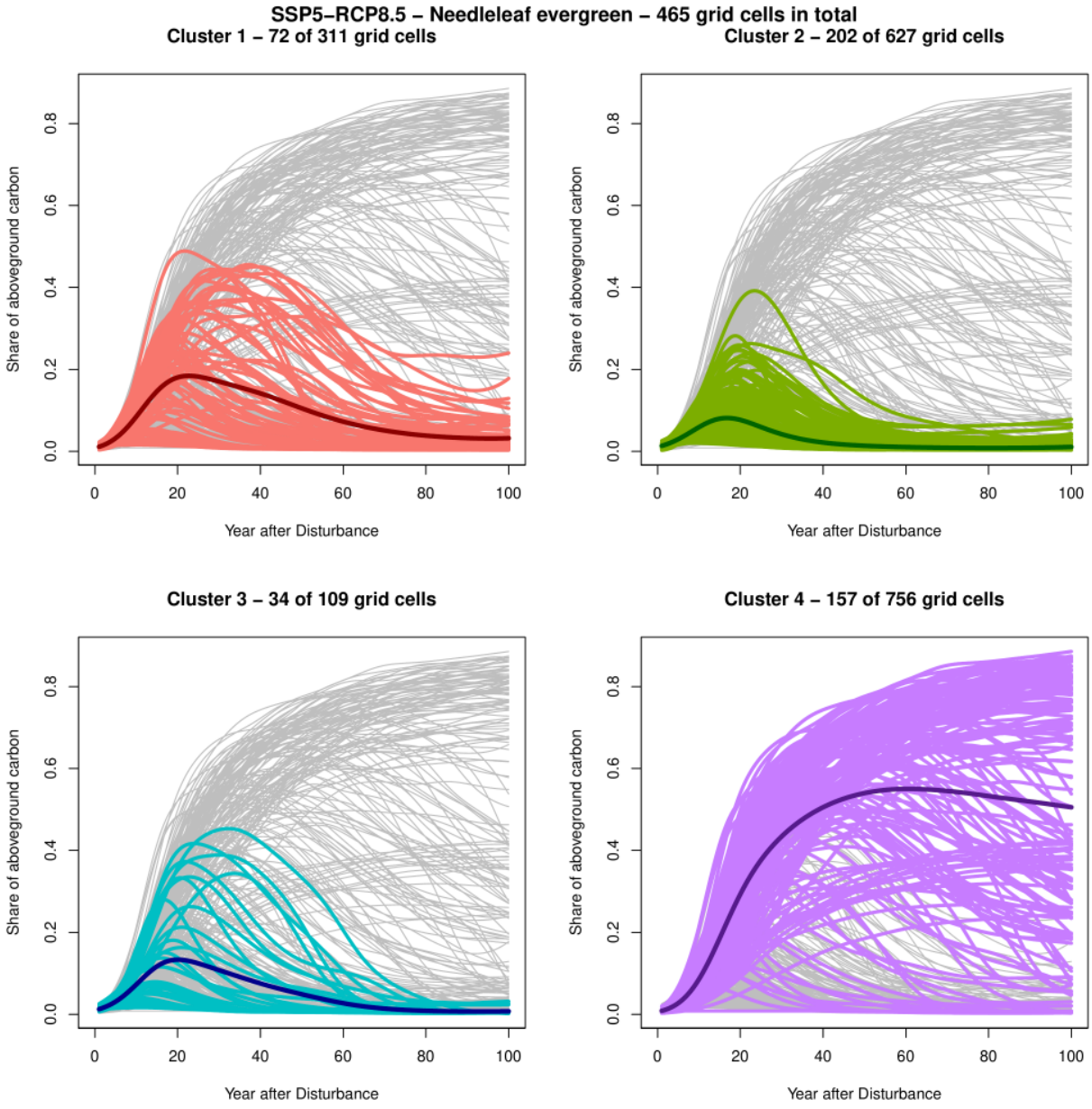


Figure 109: Clustered curves for scenario SSP5-RCP8.5 and PFT *needleleaf evergreen*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

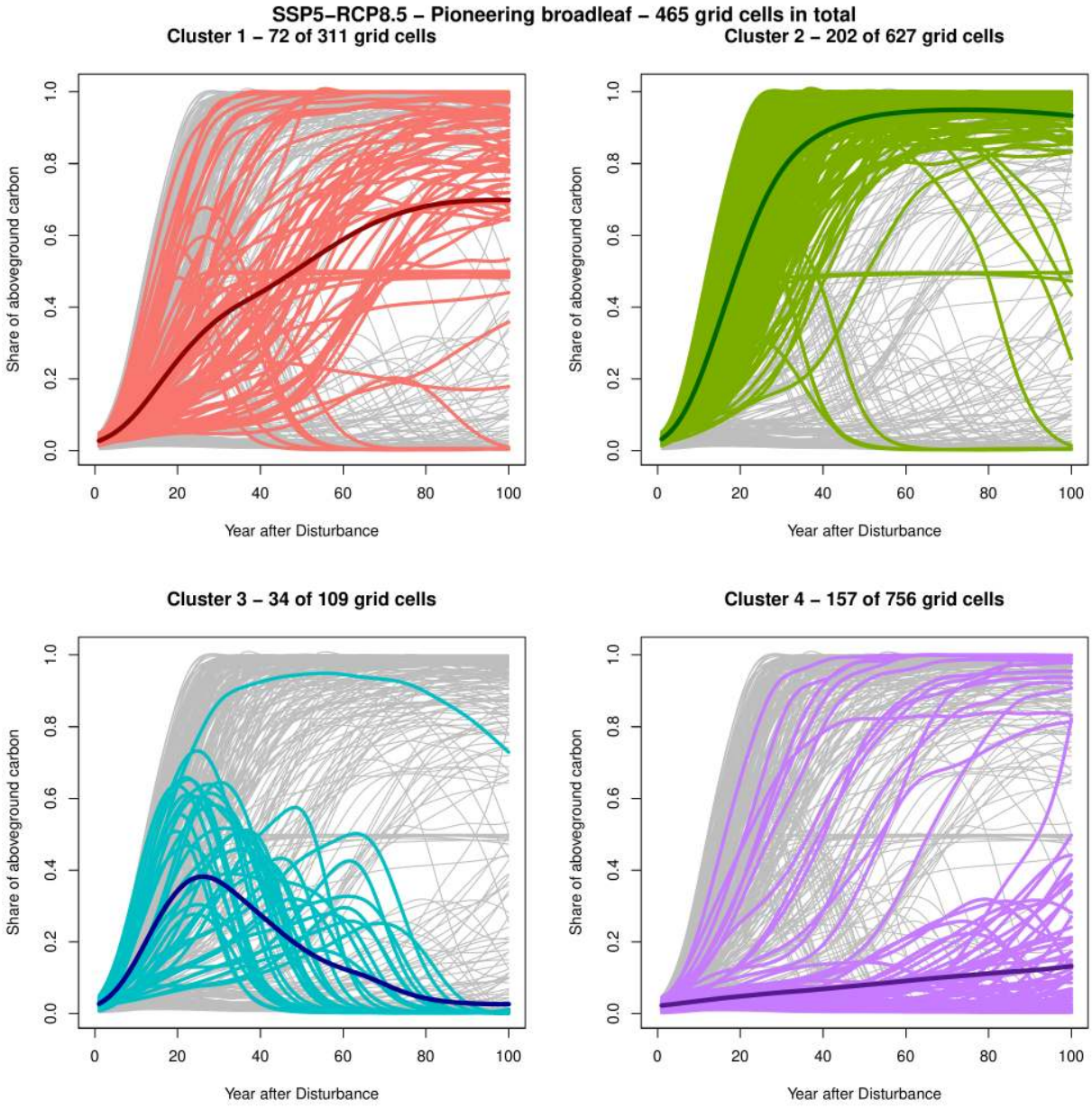


Figure 110: Clustered curves for scenario SSP5-RCP8.5 and PFT *pioneering broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

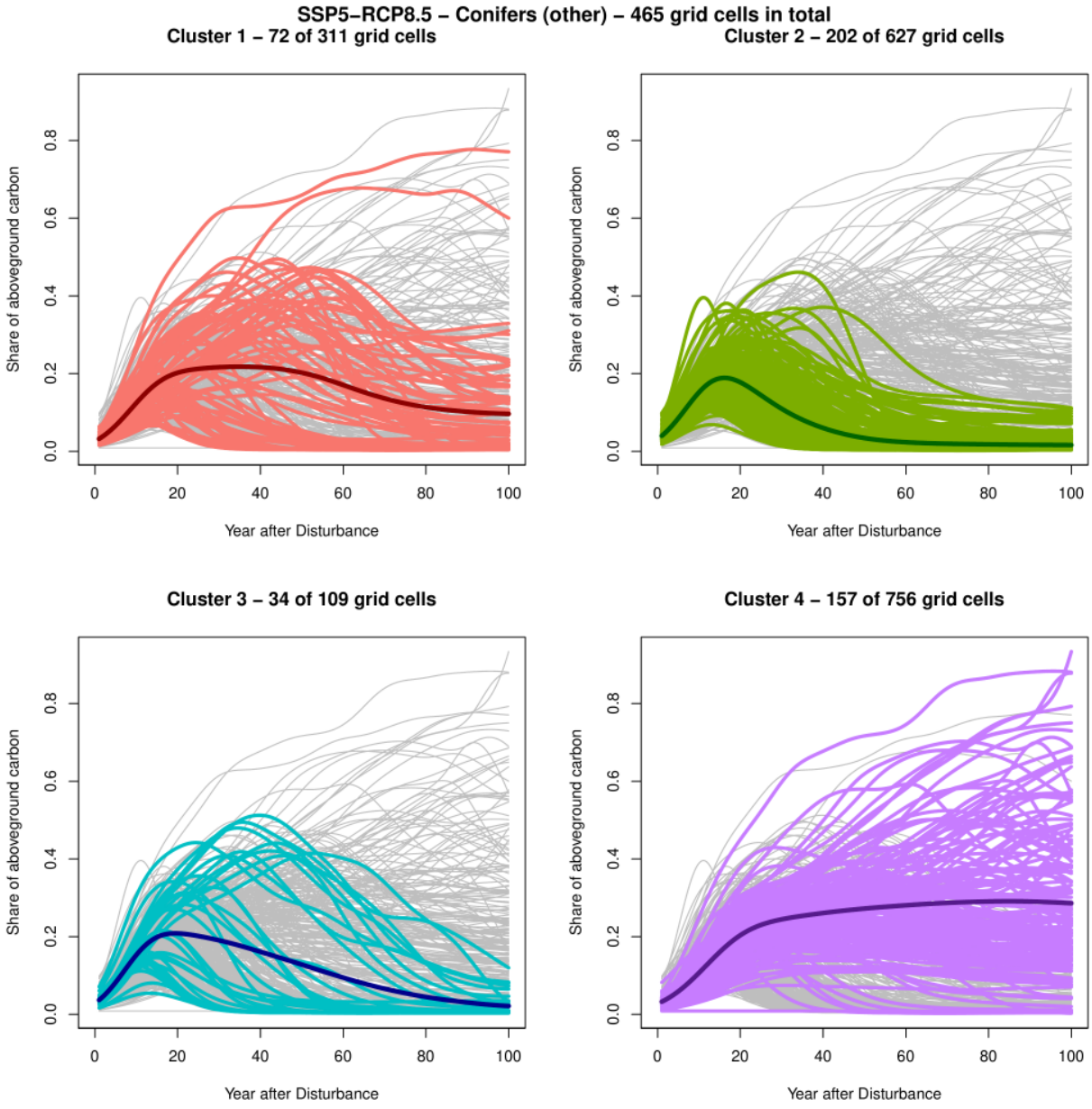


Figure 111: Clustered curves for scenario SSP5-RCP8.5 and PFT *conifers (others)*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

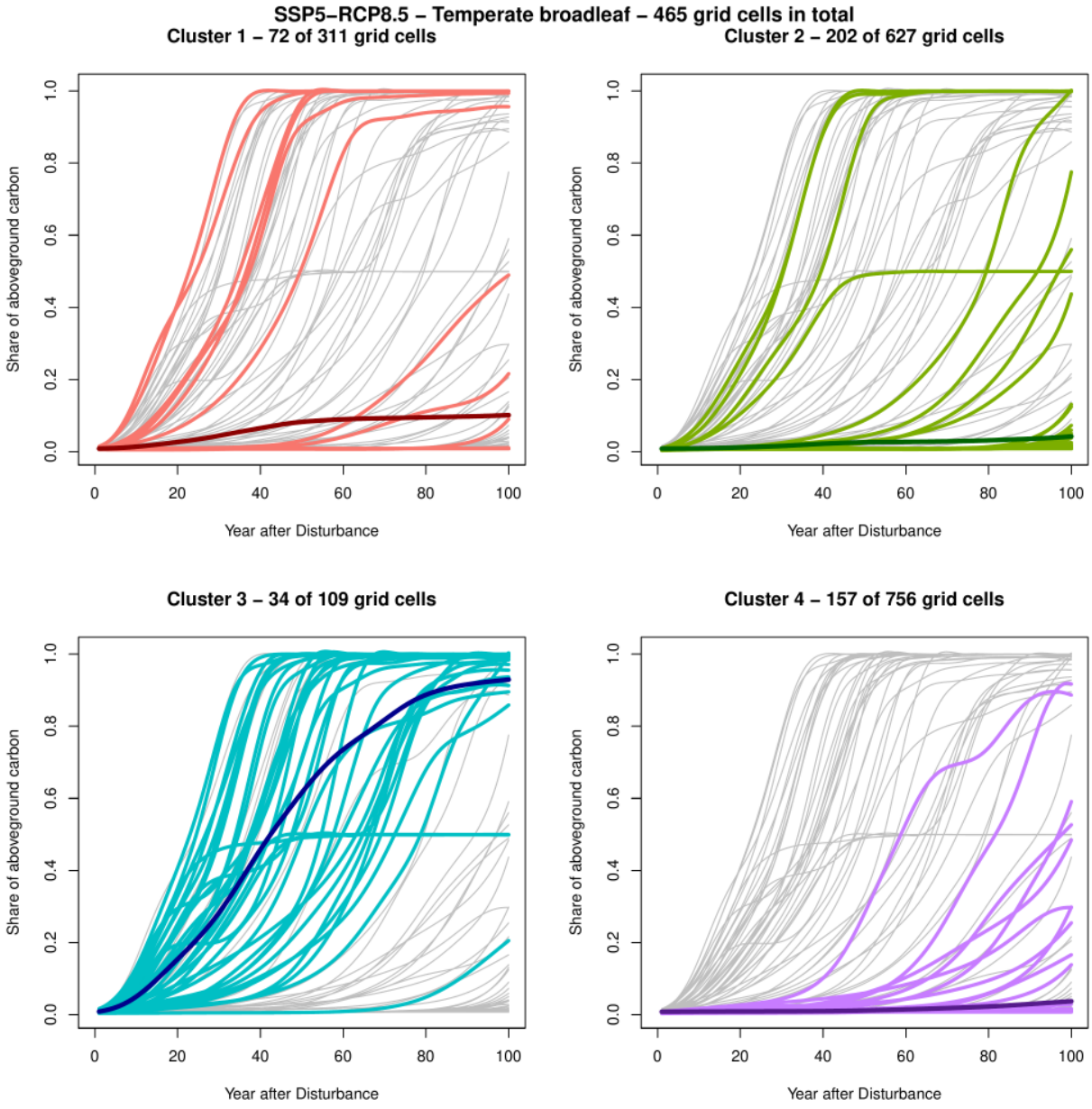


Figure 112: Clustered curves for scenario SSP5-RCP8.5 and PFT *temperate broadleaf*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

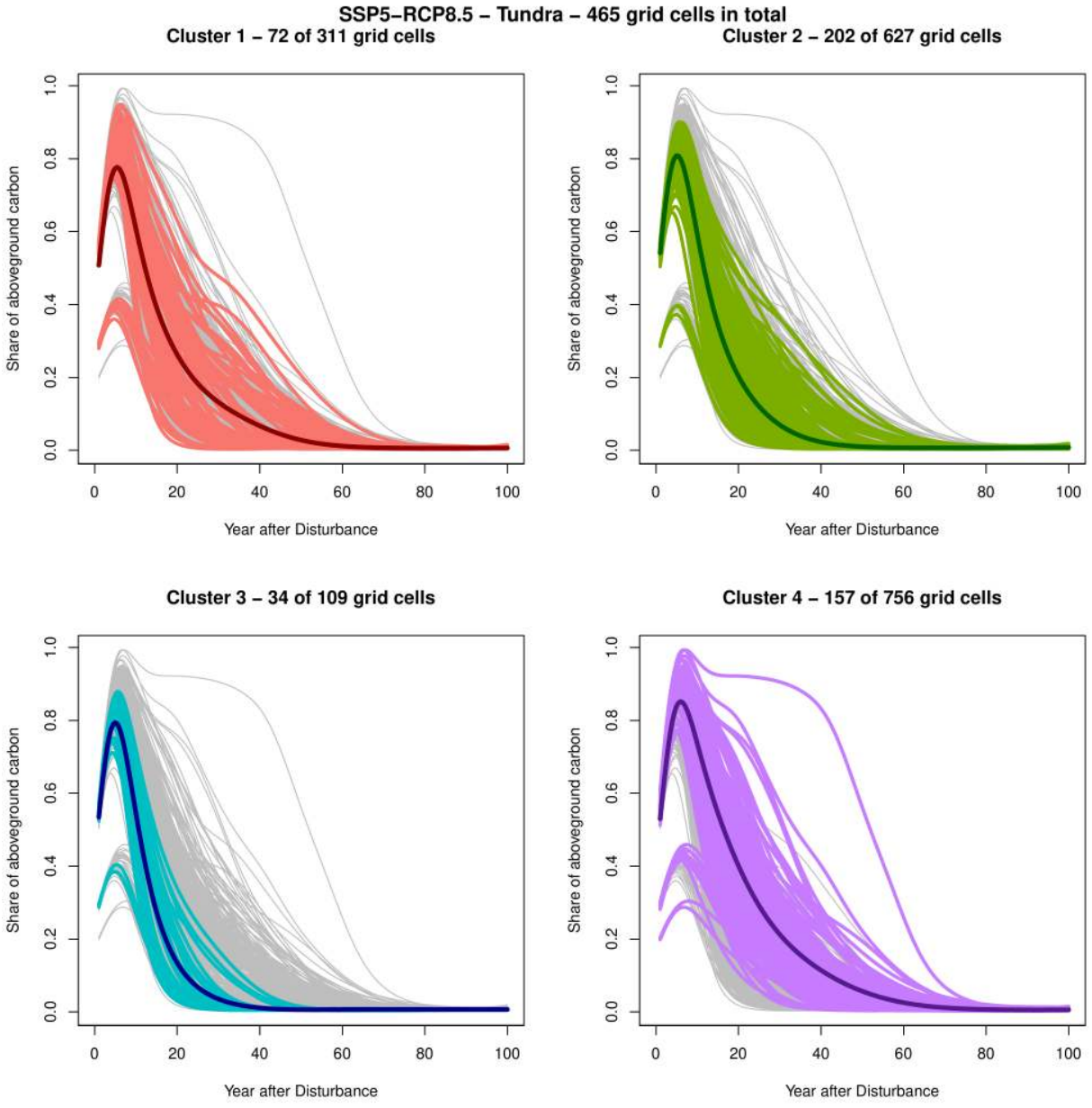


Figure 113: Clustered curves for scenario SSP5-RCP8.5 and PFT *tundra*. The colored curves indicate the belonging to the respective cluster. The dark curves represent the cluster-specific mean functions.

B Electronic Appendix

All codes and figures can be found on GitHub[1]. The repository is built as follows:

- Folder 00_Database consists of a file to build the database from the raw output data of the LPJ-GUESS dynamic vegetation model. It brings the data into a suitable form for further analysis. This code has been provided by project partner Lucia Layritz.
- The 01_Description folder contains files for descriptive analysis, including maps, plots for ecological, soil and climate variables, and recovery trajectories.
- The first approach to performing FPCA can be found in 02_FPCA. The univariate scenario and PFT-wise FPCAs are conducted in the file `FPCA_univ.R`, while the derived PC scores are clustered in `FPCA_clustering.R`. The final file, `FPCA_ex.R`, plots a sample curve of functional data. All calculations are based on the R package `fda` (version 6.1.8) developed by J. O. Ramsay, Hooker, and Graves (2009).
- To address the multivariate structure of the recovery trajectories for five different PFTs, file `MFPCA_all.R` in 03_MFPCA/MFPCA performs an MFPCA using the R package `MFPCA` (version 1.3-10) developed by Happ-Kurz (2020). Therefore, in the file `MFPCA_calculation.R`, some functions from the original `MFPCA` package are modified to fit the requirements of the data at hand. The folder 03_MFPCA/Clustering contains files for clustering the PC scores and additional investigations of cluster-specifying properties with respect to soil, ecological and climatic variables. The file `MFPCA_temporal_clustering_single.R` can be used to derive results on the temporal consistency of the cluster assignment.
- Folder 04_Model consists of the main part of this project, the modelling approach. Because of the multivariate structure of the data, a mFLR is fitted to the previously derived MFPCA scores. As the climate covariates are also in functional form, an MFPCA is performed to temperature and precipitation curves in the file `MFPCA_climate.R`. The resulting scores are used for modelling in both `mFLR_Patch1.R` and `mFLR_AllPatches.R`, the former fitting a generalised additive model (GAM) using the R package `mgcv` (version 1.9-1) (Wood, 2001) to data from only one patch and the latter to all 25 available patches to derive appropriate confidence bands. In the `mFLR_scenario.R` file, scenario-wise GAMs are fitted to derive impacts for each climate scenario. All required data and models are stored in the respective subfolders.
- All plots that are within this report are stored as pdf files in 05_Plots.

The data for the scripts is not publicly available and therefore not part of this repository, which may cause some of the scripts to not compile correctly.

[1] https://github.com/TheresaMeier/MA_FDA_veg

Declaration of Authorship

I hereby declare that this master's thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the master's thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the master's thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future master's thesis submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, September 24, 2024

Theresa Meier