

ISSN 1860-0387

DISSERTATION 1|2018

DISSERTATION 1|2018  
Helmholtz Zentrum für Umweltforschung – UFZ  
Department of Ecological Modelling

Sebastian Lehmann

## Linking pattern and processes in natural forests: innovative modelling methods across scales

Helmholtz Zentrum für  
Umweltforschung – UFZ  
Permoserstraße 15  
04318 Leipzig  
[www.ufz.de](http://www.ufz.de)

NICHT ZUM VERKAUF BESTIMMT.

Sebastian Lehmann | Linking pattern and processes in natural forests: innovative modelling ...

1|2018

 HELMHOLTZ  
CENTRE FOR  
ENVIRONMENTAL  
RESEARCH – UFZ



**Linking pattern and  
processes in natural forests:  
innovative modelling methods across scales**

**Sebastian Lehmann**

**A cumulative thesis submitted for the degree of  
Doctor of Natural Sciences (Dr.rer.nat.)**

**University of Osnabrück  
Institute of Environmental Sciences  
Department of Mathematics/Computer Science**

**Leipzig, July 2017**

# Linking pattern and processes in natural forests: innovative modelling methods across scales

## PhD dissertation

Sebastian Lehmann ([sebastian.lehmann@ufz.de](mailto:sebastian.lehmann@ufz.de))  
Helmholtz Centre for Environmental Research - UFZ  
Department of Ecological Modelling  
Leipzig, Germany

University of Osnabrück  
Department of Mathematics/Computer Science  
July 2017

## Supervisor

Prof. Dr. Andreas Huth (UFZ Leipzig, University of Osnabrück)  
Dr. Thorsten Wiegand (UFZ Leipzig)

## Publications

**Lehmann, S.**, Huth, A. (2016). “Fast calibration of a dynamic vegetation model with minimum observation data”. *Ecological Modelling*, 301, 98-105.

May, F., Wiegand, T., **Lehmann, S.**, Huth, A. (2016). “Do abundance distributions and species aggregation correctly predict macroecological biodiversity patterns in tropical forests?”. *Global Ecology and Biogeography*, 25(5), 575-585.

Brinck, K., Fischer, R., Groeneveld, J., **Lehmann, S.**, De Paula, M. D., Pütz, S., ..., Huth, A. (2017). “High resolution analysis of tropical forest fragmentation and its impact on the global carbon cycle”. *Nature Communications*, 8, 14855.

# Abstract

---

The dynamics of ecosystems are often analysed by using the concept of processes. It is essential to explore relations across scales and understand the linkage of pattern and processes (e.g. using ecosystem modelling), because the scale at which an ecological process acts may differ from the scale at which the process generates patterns. In this thesis natural forest are the focus of investigation as they provide important services like conservation of biodiversity and carbon storage. In the first chapter we give an introduction on the different parts of a scientific inference framework in which this thesis is embedded. We emphasize the importance of innovative modelling methods and present in the following chapters three modelling studies in the context of forests.

In the first study we present an inverse modelling framework based on stochastic search methods. This framework allows the fast calibration and uncertainty assessment of vegetation models. The framework is tested using a dynamic forest model in a virtual, single species and multiple species experiment. The developed methods were able to provide reliable parameter and uncertainty estimates using 10 times less computational resources than comparable methods based on Bayesian techniques.

In the second study we developed a neutral model of tropical rain forest including size structure and neighbourhood competition. The model was able to explain the main dynamics of six spatial and non-spatial biodiversity patterns using a single parameter set. This includes the species abundance distribution, the species-area relationship and the individual tree size distribution. A global sensitivity analysis revealed a highly correlative structure between mortality, dispersal and migration, which points to possible model enhancements.

In the third study we focused on the quantification of global deforestation, as this is the main driver for biodiversity loss and species extinction. We present results regarding the structure and state of forest fragments on a global scale and in three different forest biomes using an innovative cluster detection algorithm, unique data compression methods and high-resolution (30 m) forest cover maps. The analysis showed that the forests worldwide contain 409 million forest fragments and 36% of forested area lies within 100 m of forest edges. Additionally, the fragment size and shape index frequency distributions have a similar shape across the forest biomes, which can be described through power-laws.

This thesis demonstrates that complex questions regarding the understanding of natural forests can be answered using innovative modelling methods in conjunction with recent advances in available observation data. The developed methods are formulated in such a general way that they can serve as a basis to test future hypotheses about ecological communities and vegetation dynamics.



# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The need of innovative modelling methods . . . . .	3
1.1.1	Inverse modelling . . . . .	4
1.1.2	Modelling biodiversity and neutral theory . . . . .	5
1.1.3	The state of global deforestation . . . . .	6
1.2	Main objectives of the thesis . . . . .	7
1.2.1	Chapter 2 . . . . .	8
1.2.2	Chapter 3 . . . . .	8
1.2.3	Chapter 4 . . . . .	8
<b>2</b>	<b>Fast calibration of a dynamic vegetation model</b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	Material and methods . . . . .	10
2.2.1	Nonlinear regression model . . . . .	10
2.2.2	Parameter estimation . . . . .	11
2.2.3	General calibration procedure . . . . .	12
2.2.4	Modelling the dynamics of tropical forest . . . . .	13
2.2.5	Data generation . . . . .	14
2.3	Results . . . . .	16
2.3.1	Virtual data test case . . . . .	16
2.3.2	Real case - single species group . . . . .	18
2.3.3	Real data case - multiple species group . . . . .	21
2.4	Discussion . . . . .	22
<b>3</b>	<b>Size-structured neutral theory</b>	<b>25</b>
3.1	Introduction . . . . .	25
3.2	Material and methods . . . . .	27
3.2.1	Study site . . . . .	27
3.2.2	The size-structured spatially explicit neutral model . . . . .	27
3.2.3	Linking model and data . . . . .	30
3.3	Results . . . . .	34
3.3.1	Prediction of ecological patterns . . . . .	34
3.3.2	Global parameter screening . . . . .	37
3.4	Discussion . . . . .	40

<b>4</b>	<b>The global state of forest fragmentation</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Material and methods . . . . .	44
4.2.1	Database: Hansen’s forest cover map . . . . .	44
4.2.2	The advanced cluster analysis algorithm . . . . .	44
4.2.3	Compression of labelled forest fragment maps . . . . .	45
4.2.4	Forest fragment characteristics . . . . .	45
4.2.5	Structural fragment analysis for different ecoregions . . . . .	47
4.2.6	Mapping forest conditions in relation to fragmentation . . . . .	48
4.3	Results . . . . .	49
4.3.1	Global analysis of forest fragment structure for the year 2000 . . .	49
4.3.2	Global analysis of forest fragment condition for the year 2000 . .	53
4.4	Discussion . . . . .	56
4.5	Conclusion . . . . .	57
<b>5</b>	<b>Discussion and perspectives</b>	<b>59</b>
5.1	Model calibration and uncertainty assessment . . . . .	59
5.2	Understanding the mechanisms that shape biodiversity . . . . .	61
5.3	Analysis of high-resolution forest cover maps . . . . .	61
5.4	Outlook . . . . .	63
<b>A</b>	<b>Appendix for Chapter 2</b>	<b>65</b>
A.1	Uncertainty assessment . . . . .	65
A.2	Identifiability . . . . .	66
A.3	Model selection . . . . .	66
<b>B</b>	<b>Appendix for Chapter 3</b>	<b>69</b>
B.1	Parameter screening . . . . .	69
B.1.1	The morris method for global parameter screening . . . . .	69
B.1.2	Individual screening results . . . . .	69
B.2	Calibration . . . . .	71
<b>C</b>	<b>Appendix for Chapter 4</b>	<b>73</b>
C.1	Data structures of the cluster analysis algorithm . . . . .	73
C.2	Compression method used on labelled images . . . . .	73
C.3	Summary of results for a tree cover threshold of 25% . . . . .	74
C.4	Edge detection method . . . . .	75
	<b>List of Figures</b>	<b>79</b>
	<b>List of Tables</b>	<b>82</b>
	<b>Bibliography</b>	<b>83</b>

# Chapter 1

## Introduction

---

Linking pattern and processes is a fundamental method for scientific inference and also used in ecology to understand which mechanisms shape our ecological environment (Ricklefs and Miller, 1990; Grimm et al., 1996; Hilborn and Mangel, 1997). This approach can be better understood when viewed as a multi-step framework with separated but connected tasks. In the following, we describe different steps of this framework briefly in the context of ecology.

### Observation and data acquisition

Depending on the scale of the observation different techniques are used, where varying difficulties can emerge. On a local scale in classical field ecology, observation is usually carried out with the help of measurement teams. For example, the CTFS-ForestGEO is a worldwide network of over 60 forest inventories with a focus on tropical regions based on sample plots, where a plot is typically between 25 ha and 50 ha in size. In each plot, all free-standing trees with a trunk-diameter (at breast height) of at least 1 cm are identified to its species, tagged and measured every 5 years (Anderson-Teixeira et al., 2015). The CTFS-ForestGEO network thus monitors the growth and survival of about 6 million trees and 10 000 species.

As field observations are the basis of ecology, the number of variables requiring observation increases (McRoberts and Tomppo, 2007) and thereby the complexity, cost and time which is necessary to conduct the inventories. With recent technological improvements in earth observation systems large amounts of remote sensing data are publicly available. For example: NOAA/NASA pathfinder 8-km land data set (Smith et al., 1997), US Geological Survey (Clark et al., 2007), ALOS Palsar (Rosenqvist et al., 2007), NASA reverb (Cechini et al., 2011) and ESA Sentinel (Veefkind et al., 2012). The availability of such data reduces costs and increases precision on temporal and spatial scales. Recent advances in remote sensing make it even possible to estimate various soil properties using hyper spectral data (Ge et al., 2011). Although processing and analysing high resolution remote sensing data requires large amounts of computational resources, sometimes only available to larger institutes with high performance clusters, recent research accounts for the facilities by combining data from field plots with remote sensing data to draw inference (Rödig et al., 2017).



## Pattern extraction

The extraction of a pattern is some form of pre-processing and-or aggregation of the raw observational data in a form that is understandable and interpretable (Wiegand et al., 2003; Grimm et al., 1996, 2005). The perfect pattern represents a kind of repetitive structure across multiple observations. Repetition implies that some form of prediction is possible (MacArthur, 1972). Especially power-laws attract scientific interest, as their scale invariance<sup>1</sup> exposes an universality which certain mechanisms can generate (Mitzenmacher, 2004). Key examples for patterns in the field of ecology are for example, the hollow-curved relative species abundance distribution (RAD) (Fisher et al., 1943; Preston, 1962; Hubbell, 2001) on a specific trophic<sup>2</sup> level and the species-area relationship (SAR) which can be described through a power-law on intermediate scales (Arrhenius, 1921; Preston, 1962) or a tri-phase curve over continental scales (Rosenzweig, 1995; Rosindell and Cornell, 2007). The search for a pattern is supported by ecological hypotheses of relationships one expects under the current theoretical understanding of the ecological system. A pattern thus exposes itself some kind of “law” at minimum as a “data trend”.

## Ecosystem modelling

An ecological model is an abstract representation of an ecological system through the description of interacting processes in a preferable mechanistic or process-based way. Its spatial scale can vary from the individual level of a plant cell to a global scale. A model is usually formulated quantitatively in terms of a mathematical formalism including equations, state variables and their relationship through a set of rules. Models which are analytically tractable have the desirable property of mathematical elegance and high explanatory power. Most ecological models are only analytically tractable to a certain extend due to their process complexity (Wissel, 1992). As complexity increases, modelling approaches typically shift from analytical to simulation based (Acevedo, 2012).

A simulation model includes numerical techniques to solve problems for which analytical solutions are impractical or impossible. A simulation model can be deterministic, where the modelling outcome is fully determined by its parameters and initial state. Or stochastic, where the outcome possesses inherent randomness (Tangirala, 2014). For a simulation model the formal description is transferred into a computer language. This poses challenges and opportunities. It can be difficult to find suitable algorithms in terms of numerical stability and coding complexity to solve the problem at hand (e.g. partial differential equation). This also includes questions for the practicability of a numerical algorithm in terms of computational resources (Süli and Mayers, 2003). Often the best compromise has to be found between the accurateness of a numerical solution and the computational resources needed for this solution (Press et al., 1982). This compromise also depends on the scientific question, its temporal and spatial scales and includes the

---

<sup>1</sup>Scale invariance is related to the concept of self-similarity, where features of objects do not change with scale.

<sup>2</sup>The trophic level of an organism is its position within the food chain.

careful identification of modelling parts where major computing time is spend and which parts need an accurate numerical representation to minimize the bias on the modelling output.

### Experiments

Experimentation in the context of ecology denotes the task of varying system processes in such a degree that one is able to accept or decline a priori stated hypotheses about the systems behaviour and mechanistic relationships between processes. For a field ecologist an experiment is carried out under well-defined conditions (e.g. experimental drought experiments in forests (Davidson et al., 2004; Nepstad et al., 2007)). Important field experiments include the evidence of faunistic equilibrium (Simberloff and Wilson, 1969); the Park Grass experiment (Silvertown et al., 2006) as a resource for questions of biodiversity and natural selection; grassland biodiversity experiments as the JENA Experiment (Weigelt et al., 2010) or the diversity/productivity experiments by Tilman and Downing (1994); Tilman et al. (2001) and the functional trait experiments by Kraft and Ackerly (2010). The execution of field experiments is costly in human and technical resources. Either because it is difficult to observe a system in isolation without external influences, or depending on the temporal and spatial scale of the process under view, it may take considerable efforts to observe an ecological system as a whole.

There is also the possibility to perform experiments based on simulation models. This includes the extensive analysis of the models behaviour in terms of parameter sensitivity, process interaction and comparison of the model output with existing field observations. Additionally, it is possible to investigate the dynamics of experiments on much larger time frames (hundreds to thousand of years) or test the systems response under scenarios which are either impossible or impractical in reality (e.g. IPCC scenarios of climate change, Pachauri et al. (2015)). Experiments based on computer simulations can thus either reveal how ecological processes interact, which processes are represented too simplistic in the model or by improving our understanding of processes which can not be observed directly. Computer simulations can therefore complement field experiments for increasing insights about ecological systems.

## 1.1 The need of innovative modelling methods

As available computer resources roughly doubles each year (Moore et al., 1998), there is an increasing effort in solving complex modelling problems using advanced methods based on computer programs. Examples include the combination of remote sensing and local observations to derive biomass maps (Saatchi et al., 2007; Avitabile et al., 2016) as part of “pattern extraction” or to apply vegetation models on spatial scales above plot size (e.g. global vegetation models, Sitch et al. (2003)) as part of “ecosystem modelling” or the structural analysis of ecosystem models under different scenarios as part of “experiments” (e.g. evaluation of vegetation models under climate scenarios, Cramer et al. (2001)).

With increasing amount of available computer resources answers are usually not retrieved faster, but one seeks answers to more complex questions. It is therefore essential to carefully design methods which use computing resources effectively or improve upon existing methods by relaxing assumptions and dependencies. The basis of this thesis are thus a synthesising collection of innovative modelling methods or the establishment and adaptation of methods from fields other than ecology. Each single chapter of this thesis covers in detail a part of the inference framework, where the developed methods are used to answer specific questions in the context of forest ecosystems. In the following we will present the background to the topics that were analysed in this thesis (Section 1.1.1 - Section 1.1.3).

### 1.1.1 Inverse modelling

Inverse modelling can be described as the task of finding the casual factors that produced a certain set of observations. Confronting a model with data is done through several components. This includes: model calibration, identifiability, uncertainty and sensitivity analysis (Tarantola, 2005). Model calibration aims to estimate unknown model parameters in a way that the model output matches reference or observation data (Brown, 1994). It is therefore an indirect modelling approach to quantify processes through parameters. This serves different purposes in ecology modelling. At first, it is an important but not necessary model validation and evaluation strategy to test the models plausibility. If the calibrated model is able to predict the observation data under a specified error tolerance level, one assumes that the internal model processes are a reasonable approximation of the underlying processes that generated the observation data. This approach can further be used as a hypothesis testing framework, where conflicting theories (models) are discriminated by their ability to reproduce the observations (Wiegand et al., 2003; Grimm et al., 2005). Additionally, an existing and validated model can be calibrated against specific local environmental conditions to serve as a starting point for the actual modelling experiments.

The challenge of calibration depends on the model complexity. For models, where a direct mathematical formulation is available, calibration is usually done through a statistical fit using standard software like R, SPSS, Matlab or Mathematica. This is vastly different for simulation based computer models, where standard software does not provide a useful and easy to use framework. Until recently, computer models in ecology were often manually calibrated. That means important parameters are changed by hand and tested whether the model is able to better explain the observations than before the parameter adjustment. This is a highly time consuming and difficult task, as modern simulation models include often nonlinear and interactive processes. The change of the dynamic of one process has direct consequences on the dynamics of other processes. Manual calibration is therefore tiresome and only an option in the absence of other available methods.

Recently, Bayesian techniques entered the field of inverse modelling in ecology (El-lison, 2004; McCarthy, 2007). In Bayesian statistics parameter values are viewed as probability distributions instead of the frequentist view of specific but uncertain true

parameter values. With assumptions about the models structure and observation errors a (approximative) likelihood function is constructed. The likelihood of a model using a specific parameter(-set) is equal to the probability of the observation given the parameters (Beaumont et al., 2002; Tarantola, 2005). In combination with a priori assumptions about the distribution of parameters one calculates, through the use of Bayes' theorem, the posterior model distribution. The posterior is proportional to the likelihood multiplied by the prior (Lee, 2012).

As the calculation of the necessary normalization factor for the posterior distribution is difficult in practise, the posterior is approximated through sampling with Markov-Chain-Monte-Carlo methods (MCMC). The desired posterior is then the equilibrium distribution of the Markov chain (Andrieu et al., 2003). A standard MCMC method is the Metropolis-Hastings algorithm, where a random walk is generated using a proposal density and a method to reject some of the proposed moves (Hastings, 1970). This concept has been successfully applied for the automatic model calibration and uncertainty assessment for forest models (Van Oijen et al., 2005b; Hartig et al., 2014). Depending on the dimension of the parameter space and process correlations through model complexity, MCMC methods need large amounts of computational resources ( $10^5 - 10^6$  model evaluations). That means a modern forest model like FORMIND (Köhler and Huth, 1998; Fischer et al., 2016) needs months for full calibration. Possible parallelization is also limited by the sequential construction characteristic of a Markov chain. Because of this practical limitations of the MCMC methods, there is a need for methods which allow fast calibration of dynamic vegetation models by reducing the amount of model evaluations. This framework should be applicable to different models and multiple types of observation data.

We show that the task of calibration can be accomplished efficiently by stochastic search methods. Parameter and predictive uncertainty can then be assessed through the asymptotic normality of the likelihood using Taylor expansion around the parameter estimate (Hogg and Craig, 1995).

### 1.1.2 Modelling biodiversity and neutral theory

Understanding the mechanisms and processes that determine the patterns of biodiversity is a fundamental question in ecology. A potential approach to explain species diversity within species rich communities is provided by neutral theory, where dispersal and demographic stochasticity are the main mechanisms which shape the structure of a community (Hubbell, 2001). Neutrality means that at a given trophic level species are equivalent in birth, death, dispersal and speciation rates. Differences between species are thus irrelevant to their success. Neutral theory does not claim, that the assumption are per se true, but to analyse the consequences about these assumptions (Rosindell et al., 2011).

Initial studies using spatially implicit neutral models focused on the explanation of the species abundance distribution (SAD) (McGill et al., 2007; Volkov et al., 2003, 2005; Alonso and McKane, 2004; He, 2005; Etienne and Alonso, 2005, 2007; Etienne et al., 2007b; Rosindell et al., 2010). The SAD is a key pattern of diversity, where the number of species with a certain number of individuals is counted and aggregated into classes (his-

torically to the power of two, Preston (1962)). The lasting debate about which and how mechanisms should be included into neutral models to explain the SAD demonstrated, that this pattern alone cannot be used to discriminate between different biological assumptions. Later spatially explicit models were able to reproduce beta-diversity patterns (Condit et al., 2002) and species-area relationships (SAR) (Rosindell and Cornell, 2007, 2009; O'Dwyer and Green, 2010). The theoretical framework was recently extended by the incorporation of size structure in an analytically tractable manner (O'Dwyer et al., 2009) but without including interactions through neighbourhood competition.

In addition to the analytical analysis of neutral models, spatially explicit simulation models were used to understand which mechanisms explain the maintenance of species diversity. To test for some ecological hypotheses, Alonso and Sole (2000) used an individual based niche model of diverse tropical forest. The model included size structure, as demographic process can vary with size (Peters, 1986), and neighbourhood competition. Chave et al. (2002) used a neutral model against a niche-based trade-off model to test the effect of density dependence on spatial patterns and species richness. Both approaches worked on a lattice and were not tested extensively against field observations.

The capability of a model to simultaneously explain several ecological patterns at once (with a single parameter set) helps discriminating contrasting models and theories. It is also an important indicator of possible misspecification of mechanisms, as a single pattern only provides limited information about the underlying ecological processes (Wiegand et al., 2003; Grimm et al., 2005). Recently, this has been tested for a neutral model, where trees are placed continuously in a local plot and mortality increases through neighbourhood competition (May et al., 2015). This approach was able to explain several spatial and non-spatial patterns of diversity but was not able to explain the species-area relationship and the distance decay of similarity simultaneously (May et al., 2015, 2016).

Within the framework of neutral theory, two aspects denote important extensions of the current setting. To reflect the complex nonlinear interactions present in forests, the inclusion of size structure should be linked through competition with other processes. This rises the question what new patterns can be explained and if such an approach improves upon previous limits of simultaneous pattern prediction. The possibility to answer this question depends also on the availability of advanced calibration methods previously discussed. To disentangle process interactions and nonlinearities, it is important to understand the global sensitivity of model parameters and processes on each specific pattern. A comprehensive task which has not been done so far for neutral models. The availability of methods for global sensitivity analysis (using a limited amount of model evaluations, Morris (1991)) makes such a systematic analysis feasible.

### 1.1.3 The state of global deforestation

Global forests cover about 30% of total land surface (Bonan, 2008; Hansen et al., 2010). Estimates of global forest area from remote sensing data range from 3.5 billion ha (Hansen et al., 2010) to 4 billion ha (FAO, 2015). Forests provide important ecosystem services such as carbon storage, regulation of the hydraulic cycle and forest products

such as timber (Bonan, 2008). Global forests are classified in different biomes<sup>3</sup> depending on their environmental conditions (Olson et al., 2001b). The tropical forest biome plays an important role in forest research, as it represents the largest terrestrial pool of global carbon budget (Bonan, 2008; Baccini et al., 2012) and accounts for 50% of the carbon stored in the global vegetation (Houghton et al., 2009; Pan et al., 2011). Moreover tropical forests have a special role in the conservation of biodiversity. They are home of the majority of species of plants and animals on earth (Wilson et al., 1988; Raven, 1988). For example, it has been estimated that there exist more than 53 000 tree species in the tropics (Slik et al., 2015).

Deforestation and fragmentation of forests by human activities is the main driver for biodiversity loss and species extinction (Pereira et al., 2010; Rands et al., 2010). Fragmentation leads to altered microclimate conditions in forest edges (Ewers and Banks-Leite, 2013), such as higher light incidence, reduced humidity and higher temperatures. This results in increased tree mortality and subsequently lower above-ground carbon stocks in edges (Laurance et al., 2002; Harper et al., 2005; Broadbent et al., 2008; Laurance et al., 2011). The alteration of forest structure has not only an impact on global carbon budgets, it also affects habitat quality as a result of changing forest dynamics within forest edges (Groeneveld et al., 2009; Pütz et al., 2014). Thus, quantifying the state of forest fragmentation is an important task for monitoring the progress of global deforestation and change in habitat conditions.

Previous studies aimed at the analysis of forest fragmentation were conducted on low-resolution remote sensing data (Wade et al., 2003; Riitters et al., 2000). Studies using high-resolution forest cover maps evaluated only local structural quantities of forest fragments using moving window methods (Riitters et al., 2016), nearest neighbourhood methods (Haddad et al., 2015) or were conducted on a regional scale (Pütz et al., 2014).

Improving upon these studies requires a deeper insight into the global distribution of forest fragment quantities. This can be achieved using cluster analysis and classification methods, where a forest fragment is seen as a single independent cluster. The availability of global high resolution (30 m) forest cover maps (Hansen et al., 2013; Sexton et al., 2013) makes this a difficult task in terms of available computing resources. To make such an analysis feasible, existing algorithms for cluster analysis like the Hoshen-Kopelman algorithm (Hoshen and Kopelman, 1976b) have to be extended to work on such large data sets. Additionally, classification maps, which visualize state variables of deforestation, require the handling of large intermediate processing results. A task that can be accomplished by data compression methods.

## 1.2 Main objectives of the thesis

The objective of this thesis is to develop three independent modular frameworks. Each presented study focuses on a specific part of the scientific inference framework presented

---

<sup>3</sup>A biome is a community of organisms that share common characteristics depending on the environment they exists in.

in Chapter 1. The individual frameworks are then used to answer specific ecological questions. We describe the main objectives of each chapter in the following briefly.

### **1.2.1 Chapter 2**

The objective of this chapter is to develop a framework for fast calibration and uncertainty assessment of forest simulation models. This task can be seen as a component of the “experiments” part of the scientific inference framework (Chapter 1). We give an overview over the steps needed for the successful calibration of vegetation models and answer three specific questions. (1) Are methods of stochastic optimization able to correctly identify parameters values for a dynamic model of tropical rain forest? (2) What is the minimum amount of observation data needed to make reasonable parameter estimates? (3) Are these estimates made from observations at one point in time able to make predictions about the temporal behaviour of the vegetation model?

### **1.2.2 Chapter 3**

The objective of this chapter is to incorporate the aspect of individual size into a neutral model of tropical forest. This task can be seen as a component of the “ecosystem modelling” part of the scientific inference framework (Chapter 1). This is achieved through the incorporation of neighbourhood competition in a way, that size structure has an effect on all ecological processes. We compare the model predictions to a set of spatial and non-spatial field observations from Barro Colorado Island (Panama) and analyse the sensitivity of processes on different biodiversity patterns. In the course of this we answer the following questions: (1) How can size-structure be incorporated into a spatially explicit neutral model tropical rain forest while linking the effect of size growth to other community processes? (2) What are emerging new patterns and is the model able to explain various patterns of diversity in a case study? (3) How does the model parameters and processes affect these patterns? (4) What are the consequences on neutral theory? (5) Is the model a useful extension to the current framework of neutral theory?

### **1.2.3 Chapter 4**

The objective of this chapter is to analyse forest fragmentation on a global scale using high resolution forest cover maps. This task can be seen as a component of the “pattern extraction” part of the scientific inference framework (Chapter 1). The aim is to characterize the structure and condition of each forest fragment, irrespective of its size. Within this study we answer the following questions: (1) How is it possible to construct an efficient cluster analysis method using standard computational resources? (2) How can the structure of forest fragmentation be characterized on a global scale and for three forest biomes? Can we observe differences between the biomes? (3) How can the condition of forest fragments be characterized and is it possible to visualize this in a global state map?

# Chapter 2

## Fast calibration of a dynamic vegetation model with minimum observation data

---

This chapter has been published as Lehmann, S., Huth, A. (2016). “Fast calibration of a dynamic vegetation model with minimum observation data”. *Ecological Modelling*, 301, 98-105. DOI: <https://doi.org/10.1016/j.ecolmodel.2015.01.013>

### Abstract

The estimation and uncertainty analysis of parameters for dynamic vegetation models is a complex process. If one is mainly interested in parameter estimation, this can be done with simple global stochastic search methods, while uncertainty analysis is carried out with traditional first-order analysis, which significantly reduces the number of needed model evaluations. Within a nonlinear regression framework, where the misfit between model and observations is expressed as a sum of weighted squares, we model the dynamics of tropical forest with a size-structured Sinko-Streifer model and demonstrate the general calibration procedure on a virtual data set. A second case study on real data for a single species shows that surprisingly total stem number, basal area and aboveground biomass are the minimum observations needed for successful calibration. A third case study on real data for a three species group shows the prediction of successional states while only using the former reduced set of observations for calibration. The methodology is well suited for time consuming models, where only limited amount of forest site observations are available.

### 2.1 Introduction

The calibration of dynamic vegetation models is a complex and time consuming process. Calibration refers to the procedure of adjusting the models parameters in such a way that the models response matches experimental observations. In the past this often has been done through manual calibration, making this a difficult task, especially if these models are complex and parameters may affect many processes at once. Recent research



has applied Monte Carlo techniques under a Bayesian context to estimate parameter properties (Van Oijen et al., 2005a, 2013; Hartig et al., 2013), but this needs a large amount of model evaluations to make reliable estimates of the underlying posterior distributions, because ecological models often have strong parameter correlations. If one is primarily interested in parameter estimates and uncertainty assessment is less important, methods of global stochastic optimization in a frequentist context offer an alternative solution. The here presented methods are widely used in other research fields, particularly in the field of hydrological models (Duan et al., 1993; Tolson and Shoemaker, 2007a; Gallagher and Doherty, 2007) although their application is new in the field of ecology.

In this study we give an overview over the steps needed for successful calibration of vegetation models and answer three specific questions. (1) Are methods of stochastic optimization able to correctly identify parameters values for a model of tropical rain forest. (2) What is the minimum amount of observation data needed to make reasonable estimates. (3) Are the parameter estimates made from observations at one point in time, able to make predictions about the temporal behaviour of the vegetation model in the past for a multi-species version of the model (succession).

## 2.2 Material and methods

In this section we give, among others, a more general introduction into methods, assumptions, possibilities and expected results for ecological model calibration within a statistical framework. This should help readers with a basic understanding to broaden their view on the underlying concepts.

### 2.2.1 Nonlinear regression model

We assume that the observed data vector  $y$  with dimension  $|y| = n$  can be written in the following parametric form

$$y = m(\theta) + \epsilon \quad (2.1)$$

where  $m(\theta)$  is the output of a (deterministic) model  $m$  with the to-be-estimated parameter vector  $\theta \in \Theta$  with dimension  $|\theta| = k$  and  $\epsilon \sim N(0, S)$  is an additive multivariate normal distributed error term including systematic and measurement error. The distribution of  $y$  is the result of convoluting the error with the point mass  $\delta_{m(\theta)}$ , therefore,  $y \sim N(m(\theta), S)$ . We further assume independent observations, which means that  $S = \sigma^2 S_W$ , where  $S_W = \text{diag}(1/w_1, 1/w_2, \dots, 1/w_n)$  is a diagonal matrix representing relative weights.

### 2.2.2 Parameter estimation

The unknown parameter vector  $\theta$  gets estimated through maximum-likelihood estimation (MLE). The likelihood of  $y$  can be written as

$$L(\theta|y) = \frac{1}{\sqrt{(2\pi)^n |S|}} \exp\left(-\frac{1}{2} u^T S^{-1} u\right) \quad (2.2)$$

where  $u = y - m(\theta)$ . Taking the logarithm of  $L(\theta|y)$  and leaving out constants that do not affect the location of the maximum results in the objective function

$$Q(\theta) = u^T W u \quad (2.3)$$

$$= \sum_{i=1}^n w_i (y_i - m(x)_i)^2 \quad (2.4)$$

where  $W = S_W^{-1} = \text{diag}(w_1, w_2, \dots, w_n)$  and  $m(x)_i$  is the  $i$ -th component of the model output  $m(x)$ . Hence maximizing  $L(\theta|y)$  is equivalent to minimizing  $Q(\theta)$  and the parameter estimate  $\hat{\theta}$  gets now determined through

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta) \quad (2.5)$$

The first question that arises is, how to assign values to the weights  $w_i$ , if the variances in the observation data are unknown beforehand? Generally little is known about the distribution of measuring error in ecological settings, albeit the use of the normal distribution is often justified from either the central limit theorem or as an approximative distribution. The situation gets more difficult, if one is trying to calibrate a model against different types of data, like stem-diameter-distribution, basal area or biomass. Also subjective requirements of the modeller may play an important role, like robustness (Huber, 2011) of the estimates or the minimization of a certain goodness of fit measure.

If nothing is known about the variances a method known as IRLS (iteratively recursive least squares) is applicable (Green, 1984). As an iterative procedure one starts with constant weights which get successively refined after each calibration step until the estimated parameters eventually converge. The refinement is based on the residuals where, in its simplest form, the squared residuals  $u^T u$  of one estimation step serve as inverse weights for the next step. It can be shown that this results in the highest likelihood for the likelihood function (Equation 2.2).

Following such an approach multiple full calibration steps are needed, which are time consuming and impractical for the here considered case of complex models.

A simpler approach is to assume an existing functional relationship like  $w_i = |y_i|^{-\xi}$  with  $\xi \geq 0$  between the observations and the weights. Three values for  $\xi$  are frequently adopted. For  $\xi = 0$  we have the case of constant variance, which is often assumed or acceptable to use after a suitable transformation (e.g. Box and Cox (1982)). For Poisson distributed field data with large rate parameter  $\lambda$  ( $> 10$ ),  $\xi = 1$  serves as a good approximation. In Van Oijen et al. (2005a, 2013)  $\xi = 2$  is used, which implies that the standard deviation of the errors are proportional to the observations.

The second question is how to find the minimum of  $Q$ ? Traditional nonlinear optimization algorithms like gradient descent or quasi newton methods are only local search methods which additionally rely on the ability of computing first and/or second-order derivatives. The first property enforces one to make multiple optimization trials for multimodal problems and the second one is impractical due to the need for additional costly model evaluations. Hence we use a set of randomized search methods which do not exhibit the mentioned problems and do not take much effort to be implemented in an appropriate programming language.

- Adaptive Simulated Annealing (ASA): A parameter-free variant of the well known simulated annealing algorithm (Kirkpatrick, 1984). In contrast to the original annealing algorithm, a candidate solution gets sampled from a Cauchy distribution and the cooling schedule is adapted to a fixed number of allowable function evaluations (Ingber, 1993, 1996).
- Dynamically dimensioned search (DDS): A simple algorithm originally developed for the calibration of watershed models, where a random candidate solution is drawn around the current best solution with the search radius  $r$  being the only free parameter. Through the successive reduction of the number of simultaneously perturbed parameters, DDS is also efficient for high dimensional problems (Tolson and Shoemaker, 2007a).
- Adaptive Differential Evolution (JADE): An improved variant of the population based differential evolution algorithm (Storn and Price, 1997). Except the population size NP all other parameters are adjusted dynamically to the problem (Zhang and Sanderson, 2009).

### 2.2.3 General calibration procedure

Under the assumption of an existing simulation model the general calibration procedure is as follows:

First individual parameter ranges define the parameter space  $\Theta$ , including expert knowledge regarding the ecological meaningfulness. An error model is specified according to available observation data or suitable assumptions about the error distribution. A random candidate solution  $\theta \in \Theta$  is sampled from the parameter space, its objective function value  $Q(\theta)$  is evaluated and successively improved through one of the stochastic search methods (Section 2.2.2) until a maximum number of simulations is reached.

For further analyses the determined maximum likelihood estimate  $\hat{\theta}$  can be used to access uncertainty and identifiability properties. Either by estimating confidence intervals for  $\hat{\theta}$  and the prediction  $m(\hat{\theta})$ , which requires a numerical approximation of the Jacobian at  $\hat{\theta}$ , or by estimating profile likelihoods using multiple calibration trials. See Section A.1 for a detailed description and Section A.2 for the proper handling of identifiability issues.

For selecting a model from a set of candidate models we use the Akaike information criterion (AIC) (Section A.3).

### 2.2.4 Modelling the dynamics of tropical forest

For modelling the dynamics of tropical forest we use a size-structured Sinko-Streifer model (Muller-Landau et al., 2006; Kohyama, 1991; Condit, 1998; Moorcroft et al., 2001). Let  $n_i(t, x)$  ( $\text{cm}^{-1} \text{m}^{-2}$ ) be the number of trees per area of species  $i$  at time  $t$  with a diameter at breast height (DBH) of  $x$  cm. The time evolution of species  $i$  is then described through the following partial differential equation (PDE):

$$\frac{\partial n_i(t, x)}{\partial t} = -\frac{\partial g_i(t, x)n_i(t, x)}{\partial x} - m_i(t, x)n_i(t, x) \quad (2.6)$$

With the boundary conditions  $n_i(0, x) = 0 \text{ cm}^{-1} \text{m}^{-2}$  and  $n_i(t, x_0)g_i(t, x_0) = r_i(t)$ , which implies that we start with fallow land and assume a seedling input rate at the smallest stem diameter  $x_0$  from outside. Where  $g_i(t, x)$  ( $\text{cm a}^{-1}$ ) is the stem diameter grow function,  $m_i(t, x)$  ( $\text{a}^{-1}$ ) is the mortality function and  $r_i(t)$  ( $\text{m}^{-2} \text{a}^{-1}$ ) is the recruitment function of species  $i$ .

If the limits  $g_i(x) := \lim_{t \rightarrow \infty} g_i(t, x)$ ,  $m_i(x) := \lim_{t \rightarrow \infty} m_i(t, x)$  and  $r_i := \lim_{t \rightarrow \infty} r_i(t)$  exist, then the equilibrium solution  $n_i(x)$  can be written as

$$n_i(x) = \frac{r_i}{g_i(x_0)} \exp \left( - \int_{x_0}^x \frac{m_i(y)}{g_i(y)} dy \right) \quad (2.7)$$

where the variable of integration  $y$  (cm) represents a stem diameter interval. The number of trees  $N_i(x_a, x_b)$  ( $\text{m}^{-2}$ ) of species  $i$  in the diameter class  $[x_a, x_b]$  is

$$N_i(x_a, x_b) = \int_{x_a}^{x_b} n_i(y) dy \quad (2.8)$$

and thus the total number of trees  $N_i$  ( $\text{m}^{-2}$ ) is  $N_i := N_i(x_0, \infty)$ . The probability density in equilibrium  $p_i(x)$  can then be written as  $p_i(x) = n_i(x)/N_i$ .

Important statistics about a species include the cumulative basal area  $\text{BA}_i(t, x)$  ( $\text{m}^2 \text{ha}^{-1}$ ) of trees larger than  $x$  which is

$$\text{BA}_i(t, x) = \int_x^\infty y^2 n_i(t, y) dy \quad (2.9)$$

From this it follows that the basal area of trees larger than  $x$  in the whole stand is  $\text{BA}(t, x) = \sum_i \text{BA}_i(t, x)$ . Additionally the aboveground biomass of species  $\text{BM}_i(t)$  ( $\text{t ha}^{-1}$ ) can be approximated through a power law

$$\text{BM}_i(t) = c \cdot \int_{x_0}^\infty y^d n_i(t, y) dy \quad (2.10)$$

where  $c$  and  $d$  are site specific constant. Again the biomass of the whole stand is  $\text{BM}(t) = \sum_i \text{BM}_i(t)$ .

For some combinations of grow, mortality and recruitment functions analytical expressions for the equilibrium solution can be derived. For example, if grow, mortality and recruitment are constant the probability distribution  $p_i(x)$  reduces to an ordinary

exponential distribution with rate parameter  $m_i/g_i$ . For additional examples see Muller-Landau et al. (2006). In this study we also use constant mortality and constant recruitment but a grow function of the following form

$$g_i(t, x) = (1 - \alpha_i(t, x))g_i \quad (2.11)$$

$$\alpha_i(t, x) = \begin{cases} 1 & \text{if } \beta_i^{-1}BA(t, x) > 1 \\ \beta_i^{-1}BA(t, x) & \text{else} \end{cases} \quad (2.12)$$

In this formulation the constant grow  $g_i$  of a species gets linearly reduced through the time dependend factor  $\alpha_i(t, x)$ . The amount of reduction is directly proportional to the basal area above a certain diameter in the whole stand as a factor of light competition through the species specific constant  $\beta_i$  ( $\text{m}^2 \text{ha}^{-1}$ ). Hence the equilibrium distribution can be written as

$$n_i(x) = \frac{r_i}{g_i^\infty(x_0)} \exp\left(-\int_{x_0}^{\infty} \frac{m_i}{g_i^\infty(x)}\right) \quad (2.13)$$

$$p_i(x) = \frac{n_i(x)}{N_i} = \frac{m_i}{r_i} n_i(x) \quad (2.14)$$

where  $g_i^\infty(x)$  denotes the mean diameter grow in equilibrium. As this model does not have a closed analytical solution we use a numerical approximation using a finite difference upwind scheme (Courant et al., 1952).

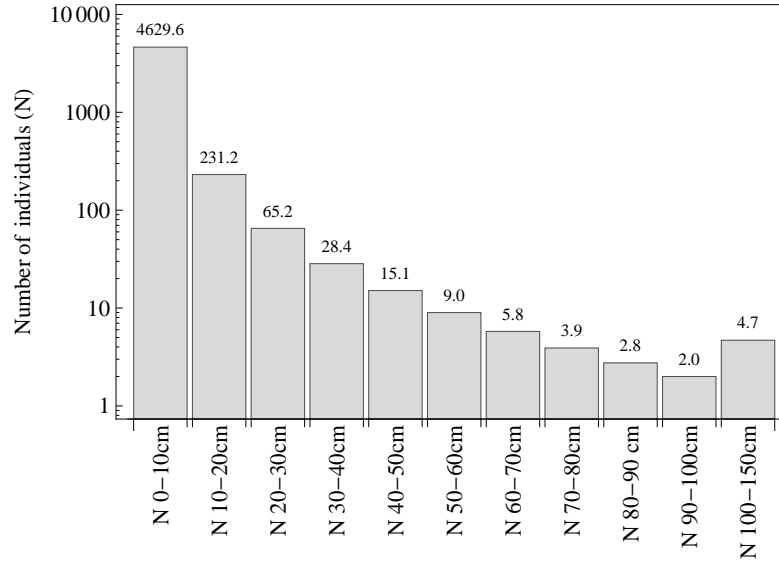
## 2.2.5 Data generation

The dynamic forest model is calibrated against three different data sets.

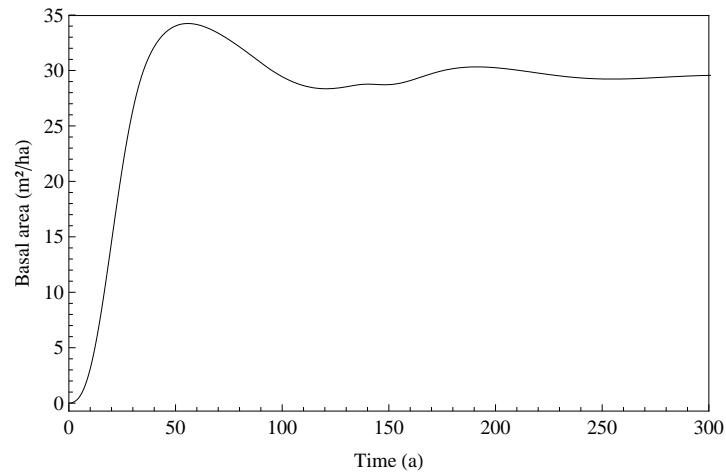
### Virtual data

A virtual 1 ha inventory using a single species is generated from the forest model in its equilibrium state (Section 2.2.4) with the following parameters:  $g = 1 \text{ cm a}^{-1}$ ,  $\beta = 30 \text{ m}^2 \text{ha}^{-1}$ ,  $m = 0.02 \text{ a}^{-1}$ ,  $r = 100 \text{ m}^{-2} \text{a}^{-1}$  and  $x_0 = 1 \text{ cm}$ . The model output is summarised into 11 DBH classes each 10 cm wide, except the last class which is 50 cm wide, consisting of the number of individual trees in the corresponding stem diameter class i.e.  $d = \{N(0, 10), N(10, 20), \dots, N(100, 150)\}$ . See Fig. 2.1 and Fig. 2.2 for a graphical representation.

A structural analysis of the model equations results in non-unique identifiability. It can be seen from Equation 2.13 and Equation 2.14 that the shape of the distribution only depends on  $m/g^\infty(x)$  and its scale on  $r/m$ . Thus, as there is always a representation for  $g^\infty(x)$  in the form  $g^\infty(x) = \lambda g'^\infty(x)$ , the parameters  $g, \beta, m$  and  $r$  are structurally unidentifiable. Therefore, we fix the specific mortality rate  $m = 0.02 \text{ a}^{-1}$  to its true value, as it is an easy to measure parameter for a forest ecologist and are left with 4 parameters for calibration.



**Figure 2.1:** Stem-diameter-distribution in equilibrium from the dynamic forest model using the virtual data set.



**Figure 2.2:** Basal area over time from the dynamic forest model using the virtual data set.

### Real data - single species group

The study area is represented by a 50 ha tropical forest inventory from Barro Colorado Island (BCI), Panama (N52°10', W79°51') with observation data from the years 1985-2010 (6 censuses). This plot mainly consists of old growth lowland moist forest that hosts more than 300 tree and shrub species with more than 1 cm in diameter at breast height. Observation data includes the diameter stem distribution over all species in classes each 10cm wide, averaged over all censuses. Classes are merged with their neighbouring classes until a single class does not have less than 5 trees. Additionally the average basal area  $BA=33.89 \text{ m}^2 \text{ ha}^{-1}$  and the aboveground biomass  $BM=310.82 \text{ t ha}^{-1}$  over all species is included. The biomass is approximated through a power-function (Chave et al., 2005) where  $BM(d) = 10 \cdot d^{2.6} \text{ (t ha}^{-1}\text{)}$  is the biomass of a single tree with a DBH of  $d \text{ (m)}$ . This results in the 11 classes (N(0-10], N(10-20], N(20-30], N(30-40], N(40-50], N(50-60], N(60-70], N(70-90], N(90-350], BA, BM).

As the same model structure as for the virtual data is used, we fix the mortality to  $m = 0.02 \text{ a}^{-1}$ , which is well known value for tropical forests (Phillips and Gentry, 1994).

### Real data - three species group

The same tropical forest inventory from BCI with the same rules for data generation is used, but species get partitioned into three plant functional types (PFT) according to their maximum height and diameter growth. This results in the following data classes:

- PFT0: 9 classes, (N(0-10], N(10-20], N(20-30], N(30-40], N(40-50], N(50-70], N(70-350],  $BA=13.50 \text{ m}^2 \text{ ha}^{-1}$ ,  $BM=146.84 \text{ t ha}^{-1}$ )
- PFT1: 8 classes, (N(0-10], N(10-20], N(20-30], N(30-40], N(40-50], N(50-260],  $BA=14.22 \text{ m}^2 \text{ ha}^{-1}$ ,  $BM=93.10 \text{ t ha}^{-1}$ )
- PFT2: 7 classes, (N(0-10], N(10-20], N(20-30], N(30-40], N(40-330],  $BA=6.17 \text{ m}^2 \text{ ha}^{-1}$ ,  $BM=70.89 \text{ t ha}^{-1}$ )

We fix the individual mortality rates to mean estimates from the field data:  $m_0 = 0.017 \text{ a}^{-1}$ ,  $m_1 = 0.023 \text{ a}^{-1}$  and  $m_2 = 0.056 \text{ a}^{-1}$  for PFT0, PFT1 and PFT2 respectively.

## 2.3 Results

### 2.3.1 Virtual data test case

We test the ability of the three proposed random search methods to find parameters estimates in a vegetation model test case with known error structure. For that purpose we consider the virtual 1 ha inventory from Section 2.2.5 with 4 to-be-estimated parameters.

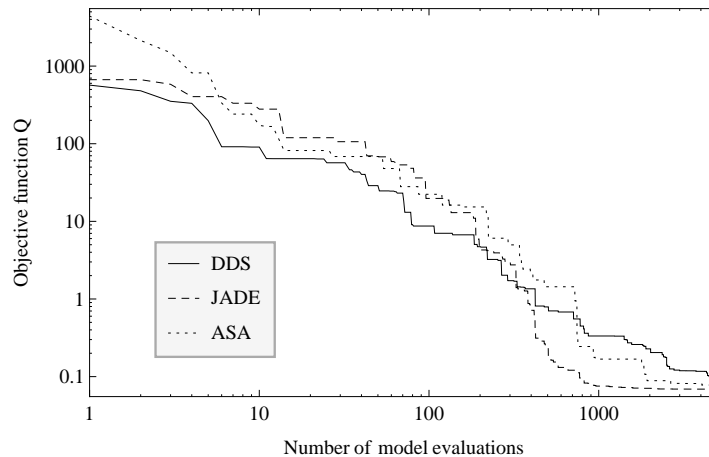
We generate the virtual data  $y$  through a single realization of  $y = d + \epsilon$  where  $\epsilon \sim N(0, 0.1^2 \cdot \text{diag}(d_1, d_2, \dots, d_{11}))$  is a multivariate normal distributed error with Poisson-like error structure. We fit the model by starting every search with a random estimate

within the given parameter ranges, use a budget of 5000 model evaluations and give the median estimate (in reference to the objective function) of five runs for each of the different algorithms. For the DDS algorithm with use a search radius of  $r = 0.2$  (Tolson and Shoemaker, 2007a) and for the JADE algorithm a population size of  $NP = 20$ .

All three algorithms give similar estimates (Table 2.1) for the three parameters  $g$ ,  $\beta$  and  $r$  within the given budget, which are in relation to the error structure, at a reasonable distance to the true values. From a convergence point of view all algorithms show similar behaviour at the beginning of the optimization process. The population of the JADE algorithm needs more time to converge but, at the end, surpasses the other algorithms and gives the largest likelihood value (Fig. 2.3 and Table 2.1). The constant radius of the DDS algorithm prevents it from improving the parameter estimate further, although Tolson and Shoemaker (2007a) suggest to use a second search with a smaller radius in such a situation. Although the ASA algorithm performs quite well for this example, its performance will decrease with increasing number of parameters as all components of the candidate solution get perturbed at once.

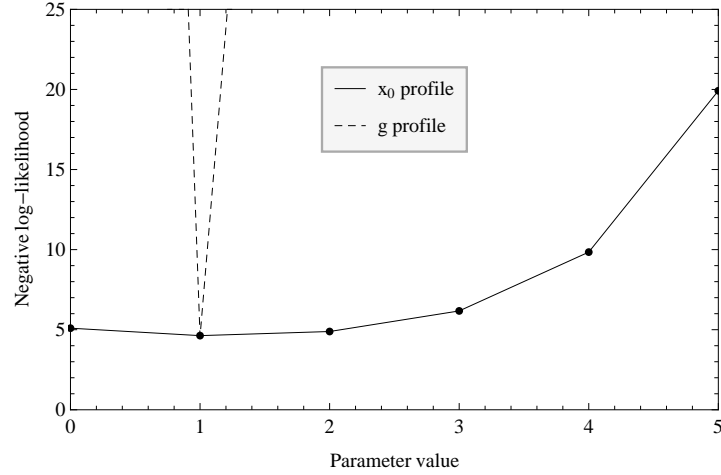
Contrary to the estimation of the other parameters the estimation of  $x_0$  deviates largely from the true value. The reason for this is a high correlation of  $x_0$  with  $g$  and  $\beta$  (Table 2.2). This uncertainty is also represented as a flat valley from  $[0, 5]$  in the corresponding profile plot (Fig. 2.4).

Our analysis shows that the proposed random search methods are all able to find reasonable parameter estimates and a high level of data aggregation does not prevent this. Secondly the presented methods in Section A.1 are successfully able to quantify parameter uncertainties and identifiability issues even in a small sample scenario.



**Figure 2.3:** Convergence graph of the different optimization algorithms for calibrating the virtual data shown on a log-log scale.





**Figure 2.4:** Negative log-likelihood profile for the parameters  $x_0$  and  $g$ . The valley between  $[0, 5]$  for  $x_0$  indicates poor identifiability.

**Table 2.1:** Median parameter estimates and log-likelihood for five independent optimization runs for the different algorithms.

	Range	True	DDS	JADE	ASA
$x_0$ (cm)	0 – 5	1	0.26	0.35	1.43
$g$ (cm a <sup>-1</sup> )	0 – 5	1	0.96	0.99	1.01
$\beta$ (m <sup>2</sup> ha <sup>-1</sup> )	0 – 100	30	29.00	29.30	29.89
$r$ (m <sup>-2</sup> ha <sup>-1</sup> )	0 – 200	100	99.87	99.86	99.87
$\log L(\theta y)$			-11.36	-4.64	-8.14

### 2.3.2 Real case - single species group

In this section we examine the ability of the presented forest model from Section 2.2.4 in predicting a real forest structure with different aggregation stages. For this we use inventory data from a forest on Barro Colorado Island (BCI) representing a single species group (Section 2.2.5). For the calibration process we always use the JADE algorithm, with a population size of  $NP = 20$  and a budget of 5000 model evaluations, as it has shown slight advantages in convergence characteristics (Section 2.3.1).

#### Weight assignment

In our error model formulation in Section 2.2.2 we introduced different ways to specify data specific weights in the formulation of the objective function  $Q$ . We test different assumption about the error model and their impact on the goodness of fit, represented by the bias corrected Aitken information criterion (AICc) and the mean absolute percentage error (MAPE). Additionally we include the mean coefficient of variation  $\overline{CV}$  as a measure of uncertainty and the mean correlation coefficient  $\overline{CC}$  as a measure of identifiability over all parameters. We test four different models: weights inversely proportional to the

**Table 2.2:** Approximative coefficients of variation (CV), 95% confidence intervals (CI) and pairwise correlation values estimated using the true values for the model and error structure.

	CV (%)	95% CI	$x_0$	$g$	$\beta$	$r$
$x_0$	26.0	[0.49, 1.51]	1			
$g$	4.4	[0.91, 1.09]	-0.88	1		
$\beta$	2.0	[28.82, 31.18]	0.87	-0.56	1	
$r$	0.1	[99.72, 100.28]	-0.03	-0.03	0.06	1

variance of the data (calculated from the 6 repeated measurements), constant weights ( $\xi = 0$ ), weights inversely proportional to the squared data ( $\xi = 2$ ) (Van Oijen et al., 2005a, 2013) and an IRLS model with two iterations (doubling the number of model evaluations).

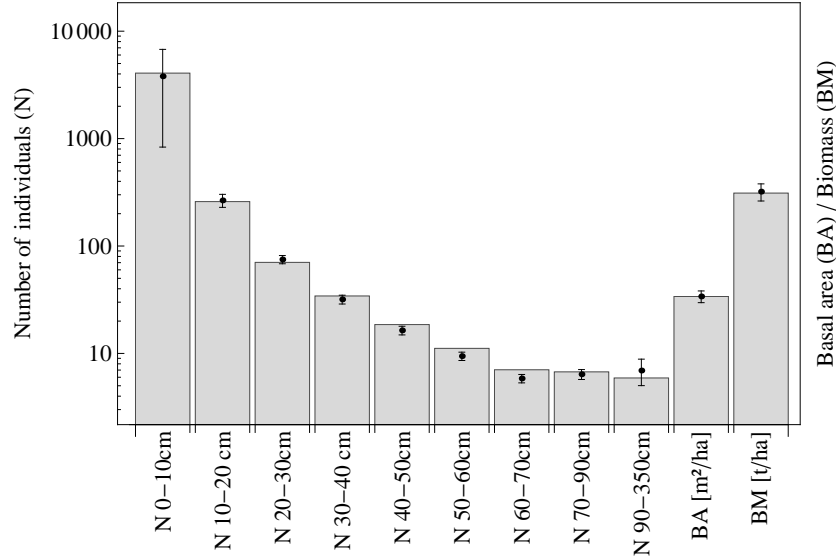
The underlying assumptions about the error model affect results regarding uncertainty ( $\overline{CV}$ ) and identifiability ( $\overline{CC}$ ) (Table 2.3). Regarding the relative error (MAPE) all models give similar results which is a consequence of good agreement between the model and the observations, irrespective a specific weighting approach (Fig. 2.5). Looking at AICc values misleads in this case, as one would prefer the IRLS model, which assumes that nothing is known about error variances. The distribution of variances from multiple measurements and the estimated ones from the IRLS model differ ( $p = 0.22$  in a two-sided K-S test), but are in agreement with variances proportional to the squared observations ( $p = 0.83$ ). Hence in absence of known error variances the usage of the latter model is appropriate.

**Table 2.3:** Effect of different assumptions about the error model on goodness of fit represented by the mean absolute percentage error (MAPE) and a bias corrected Aitken information criterion (AICc). Estimates on parameter uncertainty are represented by the mean coefficient of variation ( $\overline{CV}$ ) and estimates on parameter identifiability by the mean correlation coefficient ( $\overline{CC}$ ) over all parameters.

Weights	MAPE (%)	AICc	$\overline{CV}$ (%)	$\overline{CC}$
$\propto \text{Var}(\text{data})^{-1}$	8.39	81.92	13.21	0.36
$\propto \text{data}^{-2}$	7.23	78.95	8.23	0.40
constant	8.13	59.84	2.71	0.80
IRLS	9.26	37.70	5.20	0.90

### Model selection

For the model selection we test three cases using weights inversely proportional to the data variance. The first model includes all parameters, namely  $x_0$ ,  $g$ ,  $\beta$  and  $r$ . In the second case we fix  $x_0 = 0$  cm and in the third one we additionally disable diameter grow reduction through setting  $\beta = 0$  m<sup>2</sup> ha<sup>-1</sup>. For model comparison we use the Akaike-weights calculated through the AICc (see Section A.3). It can be seen from Table 2.4 that



**Figure 2.5:** Prediction of the forest model against field data on a log-scale in a single species experiment using weights inversely proportional to the data variance. Filled bars indicate mean observations for diameter steam distribution, basal area and aboveground biomass. Error bars indicate the approximative 95% confidence interval for the marked prediction.

the model with  $x_0$  fixed has the highest weight among the three models. The increase in parameter dimensionality of the full model does not justify the slightly better fit. In contrast the model without grow reduction is not able to provide an acceptable fit quality.

**Table 2.4:** Effect on the AICc and corresponding weights when model complexity is reduced. As an additional measure of goodness of fit the mean absolute percentage error (MAPE) is included.

Model	k	MAPE (%)	AICc	Weight
full	4	8.39	81.92	0.14
$x_0$ fixed	3	8.94	78.30	0.86
$x_0, \beta$ fixed	2	47.80	114.07	$\sim 0$

## Data selection

We try to estimate the minimum amount of observation data needed to give reliable parameter estimates for the model of tropical rain forest. For this we subsequently reduce the amount of data, against which the model will be calibrated. We consider four specific cases:

- Set 1. Full data set (11 classes)  
 $(N(0-10], N(10-20], N(20-30], N(30-40], N(40-50], N(50-60], N(60-70], N(70-90], N(90-350]), BA, BM)$
- Set 2. DBH data  $(N(0-10], N(10-30], N(30-350])$  (3 classes)
- Set 3. Total stem count  $N$ , basal area, biomass (3 classes)
- Set 4. Total stem count  $N$ , biomass (2 classes)
- Set 5. Biomass, basal area (2 classes)

We compare the fits by evaluating the MAPE against the full data set as a measure of goodness of fit, the mean coefficients of variation ( $\overline{CV}$ ), the mean of the absolute pairwise correlation values ( $\overline{CC}$ ) and the mean bias of the parameter estimates in comparison to the estimates using the full data set. We use the calibration method from Section 2.3.2, where  $x_0$  is fixed, weights inversely proportional to the data variance and check the uniqueness of the solution by doing multiple full optimization trials.

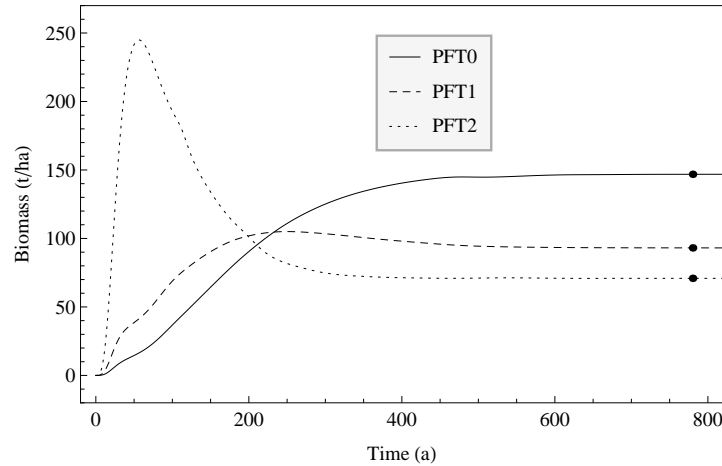
The results can be seen in Table 2.5. Using Set 2 or 3 increases parameter uncertainty by 44% and 74% respectively, but Set 2 worsens the goodness of fit including a noticeable bias in the parameter estimates. Hence, using only total stem count, basal area and biomass is enough to give reliable estimates, but using less than that results in partial non-identifiability of model parameters.

**Table 2.5:** Effect of observation data reduction to the models predictions represented by the mean absolute percentage error (MAPE), the mean of the coefficients of variation ( $\overline{CV}$ ), the mean of the absolute pairwise correlation values ( $\overline{CC}$ ) and the mean bias of the parameter estimates in comparison to the estimates using the full data set.

Set	MAPE (%)	$\overline{CV}$ (%)	$\overline{CC}$	Bias (%)
1.	8.94	13.78	0.29	-
2.	16.61	19.86	0.37	20.9
3.	7.39	23.99	0.35	5.9
4.	$g$ not identifiable			
5.	$r$ not identifiable			

### 2.3.3 Real data case - multiple species group

In this section we examine the ability of our forest model from Section 2.2.4 in predicting a real forest structure using multiple species using inventory data from BCI representing a multiple species group (Section 2.2.5). Results from the single species experiments are adopted, hence we use the reduced model with 3 parameters per species, weights inversely proportional to the data variance and only calibrate against total stem count, basal area and biomass per species resulting in a reduction from 24 classes to 9 classes.



**Figure 2.6:** Biomass over time for the three species model, calibrated against stem count, basal area and biomass. Dots represent field data measurements.

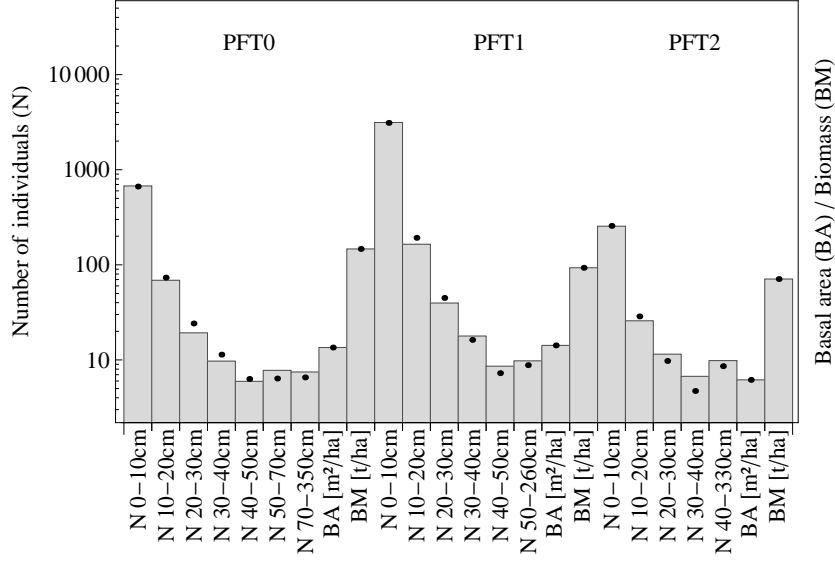
For the increase of parameter dimensionality we adjust the maximum number of model evaluations to 50000 and use a population size of  $NP=50$ .

Using three species groups clear successional states can be observed in the simulation (Fig. 2.6). Although we only used 9 classes for calibration, a comparison of the prediction with 24 classes results in a MAPE of 9.2% showing good agreement (Fig. 2.7).

## 2.4 Discussion

The here proposed calibration framework using methods of stochastic optimization is formulated in such a general way that it has a wide area of applicability. It can be applied to any ecological (simulation) model with defined input parameters and model output which can be related to field observations. That does not mean that all parameters have to or can be inversely calibrated nor that every model output has to have a corresponding field observation. We have hereby shown the successful application of this framework for a simple dynamic model of tropical rain forest. Nonetheless, the ecological model complexity or general model structure does not matter, as long as the mismatch between model output is represented as a valid (e.g. enough to identify the parameters) objective function.

There exist numerous classes and variants of other optimization algorithms. What is “best” depends on specific requirements and is additionally problem dependent, but promising more complex techniques include algorithms like Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen and Ostermeier, 2001) and response surface assisted techniques (Regis and Shoemaker, 2005). Using only a quadratic approximation in a maximum likelihood framework for uncertainty assessment, gives a speed advantage of a factor of 10 to 100 compared to MCMC techniques, which explore the complete posterior distribution. Furthermore the MLE corresponds to the maximum a posteriori (MAP) estimate in a Bayesian framework with uniform prior and in conjunction



**Figure 2.7:** Prediction of the equilibrium model on a log-scale in a three species experiment denoted by their plant functional type (PFT), using weights inversely proportional to the data variance and calibrated using total stem count, basal area and biomass per species. Filled bars represent mean observations and dots model predictions.

with the estimated covariance matrix can be used as an initial estimate for the posterior distribution, if more precise uncertainty analyses are needed through the use of a MCMC.

We have emphasised the importance of correct specification of the error model, as this has a strong impact on parameter estimates, uncertainty and identifiability measurements. Thereby iteratively reweighted least squares (IRLS) is a useful technique, when nothing is known about error variances. We suggest that, for the here considered case, the variance of multiple measurements can be approximated through a functional relationship, where the variances are proportional to the squared mean observations, neglecting the need for multiple calibration trials as in the IRLS case. The remaining scale parameter  $s$  can be estimated from the residuals (Equation A.2 in Section A.1) and must not be specified a priori (Van Oijen et al., 2005a, 2013), as this distorts uncertainty estimates additionally.

Our most important finding is that, for the here considered Sinko-Streifer model, the minimum amount of observation data needed to give reliable parameter estimates are total stem count  $N$ , basal area  $BA$  and aboveground biomass  $BM$  at the cost of higher parameter uncertainty. It is surprising and a highlight of this study that only this minimum amount observation data is needed. Using a lower number of observations results in partial unidentifiable parameters. As every of the observations contributes a different type of information, a further reduction in the number of observations is possible by fixing additional parameters to direct measurements, as in the inevitable case of the mortality parameter  $m$ . For example, through the relationship  $N = r/m$ , one could use

an estimated recruitment number  $r$  and use only basal area and biomass for estimating the remaining parameters mean growth  $g$  and growth reduction  $\beta$ . Another way of circumventing identifiability issues is the usage of penalization terms in the objective function Equation 2.4 like Tikhonov regularization, which again has a direct interpretation in a Bayesian context. It should be investigated in future studies to which extend this special findings are applicable to other vegetation models as well.

The parameter estimates made from observations at one point in time, are able to make predictions about the temporal behaviour of the vegetation model in the past for a multi species version of the model, where we used 9 observations to make 24 model predictions, as long as there exist estimates for the specific mortality rates. The reliability of the successional states and their transient behaviour should be evaluated in future studies.

Therefore, the presented methods can also be carried out at forest sites where a limited number of observations in time and space are available. After a careful structural examination of the model and the observation data a combination of direct and inverse modelling will lead to successful calibration results.

# Chapter 3

## Size-structured neutral theory explains multiple diversity patterns in a tropical forest

---

### Abstract

Neutral theory has been used to explain the maintenance of species diversity within communities. While previous attempts mainly focused on the prediction of species-abundance distributions and species-area relationships, we incorporate size-structured demography as a community assembly mechanism. We show that the linkage of reproduction, dispersal, mortality, size growth and local competition successfully predicts a wide variety of macroecological patterns. It is the first neutral model that is able to match the species-abundance distribution, species-area relationship, distance-decay of similarity, pair-correlation function and individual size distribution simultaneously, when compared to the Barro Colorado Island (BCI) tree community. However, the model was able to match the number of species in a specific size class only partly. A global parameter screening revealed a highly sensitive and correlative structure between mortality, species richness and dispersal. We conclude that dispersal through a single dispersal kernel has to be enhanced to account for the complex spatial composition of species in tropical forests. Our study demonstrates the importance of confronting models simultaneously with multiple patterns. Embedding the model into a framework of fast model calibration and sensitivity analysis additionally advances the possibilities of hypothesis testing and process understanding.

### 3.1 Introduction

Understanding the mechanisms and processes that determine the patterns of biodiversity is a fundamental question in ecology. One approach to explain the maintenance of species diversity within species rich communities is provided by neutral theory, where dispersal and demographic stochasticity are the main drivers that shape the community structure (Hubbell, 2001). The theory initiated a controversial debate in community ecology. The objective of neutral theory is not to assume that the assumptions are



per se true, but rather about the consequences of making these assumptions (Rosindell et al., 2011). First results using spatial implicit models focused only on the mechanistic (or statistical) explanation of the species abundance distribution (SAD) (McGill et al., 2007). This includes proper analytical results (Volkov et al., 2003; Alonso and McKane, 2004; Etienne and Alonso, 2005, 2007), the analysis of different speciation modes other than point-mutation (Etienne et al., 2007b; Rosindell et al., 2010) and investigations whether the exclusion of the regional species pool and additional density dependence can lead to similar results (He, 2005; Volkov et al., 2005). Results demonstrates that the species abundance alone cannot be used without additional information, to discriminate between different biological assumptions. Later spatially explicit models were able to reproduce realistic beta-diversity patterns (Condit et al., 2002) and species-area relationships (SAR) (Rosindell and Cornell, 2007, 2009; O'Dwyer and Green, 2010). Recently the theoretical framework was extended by the incorporation of size structure (O'Dwyer et al., 2009). Although the individual size growth was modelled independent from the other processes (excluding neighbourhood competition), this helped resolving the non-random sampling bias (Rosindell et al., 2012) introduced through the arbitrary determination of a reproductive size threshold.

It has been shown that the capability of simultaneous predicting several ecological patterns helps discriminating contrasting models and theories (May et al., 2015; Xiao et al., 2016), because a single pattern only provides limited information (Wiegand et al., 2003; Grimm et al., 2005). This approach has seemingly found its limits by failing to predict the species-area relationship and the distance decay of community similarity simultaneously in tropical forests under neutral assumptions (May et al., 2015, 2016). As this might be an indicator of non-neutral effects in macro-ecological patterns, it is crucial to understand the interaction of ecological processes and how they drive specific biodiversity patterns. As demography can vary with size (Peters, 1986), theory should capture the variation in demographic rates with size. The inclusion of size structure in biodiversity models dates back to the niche-based DivGame simulator of rainforest (Alonso and Sole, 2000), but it is an extension for simulation models based on neutral theory.

This study takes on the two aspects of individual size and simultaneous prediction of patterns by incorporating size structure in a individual based spatially explicit neutral model of tropical forest. This is done in a way that size structure has an effect on all the other ecological processes. We additionally compare the model predictions to a set of spatial and non-spatial field observation from Barro Colorado Island (Panama) and analyse the sensitivity of processes on different biodiversity patterns. In the course of this we answer three specific questions: How can size structure be incorporated into a spatially explicit neutral model tropical rain forest while linking the effect of size growth to other model processes? (2) What are emerging new patterns and is the model able to predict various patterns of diversity using a single parameter set? (3) If so, how do model parameters/processes affect these patterns? (4) What are the consequences on neutral theory? (5) Is the model a useful extension of the current framework of neutral theory?

## 3.2 Material and methods

### 3.2.1 Study site

The tropical forest on Barro Colorado Island (BCI), Panama (9.15° N, 79.85° W) is a 15 km<sup>2</sup> island located in Lake Gatun, an artificial water body created by the construction of the Panama Canal (Condit et al., 2001). It is covered with semi-deciduous tropical lowland rainforest and estimations on the minimum forest age range from 300 to 1500 years (Bohlman and O'Brien, 2006; Meyer et al., 2013; Lobo and Dalling, 2014). The climate is characterized by average daily maximum and minimum temperatures of 30.8 °C and 23.4 °C and an annual precipitation sum around 2600 mm with a dry season from January to April (Condit et al., 2001). A 50-ha rainforest observation plot is located on the central plateau of the island, with terrain altitudes varying between 120 and 160 m above sea level (Lobo and Dalling, 2014). Since its establishment in the early 1980s, each tree in the 1000 m x 500 m area with a minimum stem diameter at breast height (DBH) of 1 cm has been measured during censuses in five years intervals (Condit, 1998; Hubbell et al., 1999, 2005). Estimates of mean canopy height are  $24.6 \pm 8.2$  m and mean above ground biomass (AGB) are  $281 \pm 20$  t/ha (Chave et al., 2003).

### 3.2.2 The size-structured spatially explicit neutral model

Our stochastic simulation model is a direct extension of the classical spatially implicit neutral model from Hubbell (2001). We maintain the distinction between a non-spatial metacommunity (regional species pool) and the local community. Our model differs in important aspects from previous approaches. Individuals in the local community are placed in continuous space, thus allowing a direct comparison with spatially explicit forest census data using point-pattern analysis. Temporal dynamics are typically constrained by a zero-sum assumption to account for competition of space, where a dead individual is instantly replaced with a new individual, hence keeping the total population size constant. In our approach the population size is allowed to vary through the separation of birth/immigration and mortality. The model explicitly models the size growth of each individual tree, therefore each tree has an individual impact on the competition dynamics in its neighbourhood. The modelling of tree size additionally allows the prediction of new patterns.

In the model each tree  $i$  is characterized by a set of properties namely its position in space  $(x_i, y_i)$ , its DBH  $d_i$  and its species identity  $S_i$ . The model internals can be best described through a set of interacting building blocks: metacommunity, competition, recruitment, mortality and size growth, which are described in the following.

#### Metacommunity

We use a new concept for constructing a non-spatial metacommunity. Previous approaches used either a dynamic metacommunity, which develops over longer time periods than the local community (Hubbell, 2001), or were based on a sequential construc-

tion scheme (Etienne and Alonso, 2005). The former needs the explicit computation of metacommunity dynamics including a “burn-in”-phase to stabilize the species distribution, whereas the latter approach yields only a fixed abundance distribution for the metacommunity, needing multiple simulation trials for accurate predictions in the local community (May et al., 2015).

In our approach species are sampled from the metacommunity according to their expected species rank abundance. The distribution of the metacommunity is based on the Fisher log-series, as it is the limiting distribution of a metacommunity for high diverse systems (Etienne and Alonso, 2007). This results in two free parameters: the number of species  $S$  and the ratio of birth-to-death  $p$  (instead of  $J_m$ , the fixed size of the metacommunity and  $\theta$  the “fundamental” biodiversity number). For the Fisher log-series the two concepts are linked through the formula  $\theta = -S/\log(1-p)$ . The expected species rank abundance (SRA) distribution is analytically difficult to determine. Therefore, we use a specific construction algorithm as a numerical approximation.

The construction algorithm is defined through the following steps, where an example is shown in Fig. 3.1: 1) Creation of a single metacommunity with  $S$  species, where the number of species which have  $n$  individuals is distributed according to a log-series distribution with parameter  $p$ . 2) The  $S$  species are sorted in decreasing order with respect to the number of their individuals. Each species is assigned to a rank, where the rank results directly from its sorted position (the species with the highest number of individuals has the first rank). This step results in a single SRA distribution. 3) Several random SRA distributions are created through step 1) and 2) and the expected SRA distribution is calculated over this set of random distributions. The construction is terminated when the estimated mean is stable, where the stability is determined through the maximum relative deviation between two estimates.

The expected SRA distribution is used to sample species from the metacommunity, which immigrate into the local community. The higher the expected rank of a certain species, the higher its probability of being sampled. Therefore, the identification number of a specific species is fixed for the whole simulation process.

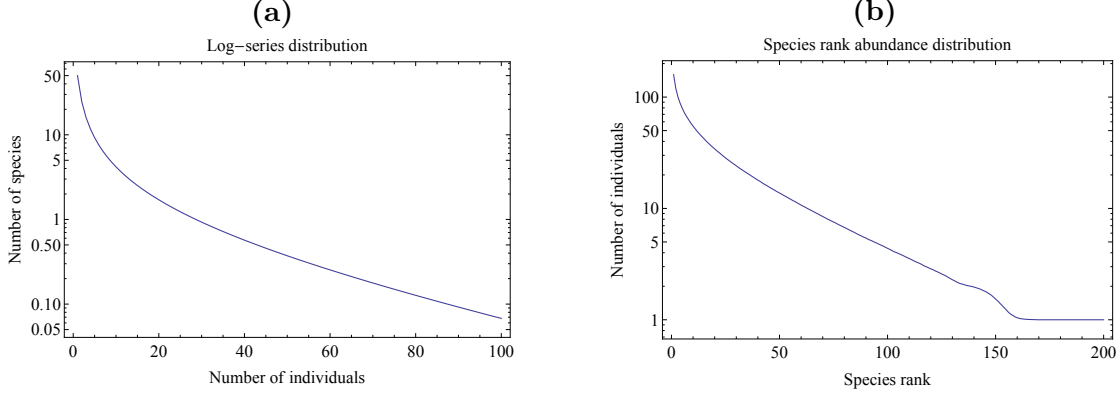
## Competition

Inter-and intraspecific competition is an important aspect in our model and has an effect on individual size growth and the probability of dying. We model the general competition index  $C_{i,j}(\alpha, \sigma)$  of a tree  $j$  on a focal tree  $i$  ( $i, j = 1, \dots, n$ ) through a two-dimensional isotropic gaussian function (see also Brown et al. (2011)):

$$w_{i,j}(\sigma) = \exp\left(\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{-2\sigma^2}\right) \quad (3.1)$$

$$C_{i,j}(\alpha, \sigma) = \alpha \cdot w_{i,j}(\sigma) \quad , \quad (3.2)$$

where  $\alpha$  is the amplitude and  $\sigma$  (m) the standard deviation (control of width) of the function  $w_{i,j}$ . The function  $w_{i,j}$  weights the competition exerted by tree  $j$  on the focal tree  $i$ . As amplitude we use the basal area  $B_j$  (m<sup>2</sup>) of tree  $j$ , hence  $\alpha_j := B_j = \pi/4 \cdot d_j^2$ .



**Figure 3.1:** Based on the log-series distribution of species (a) the species rank abundance (SRA) distribution (b) is constructed as described in Section 3.2.2. The SRA distribution is used to sample species from the metacommunity, which immigrate into the local community. For this example we used  $S = 200$  species and a birth-to-death ratio of  $p = 0.98$ .

The modelling of neighbourhood competition through weighted power-scaling on tree diameter is a widely used concept (Uriarte et al., 2004, 2010; Canham et al., 2004), where the weighting and scaling can be formulated in different ways (e.g. assuming sensitivity on the size of the target tree  $i$  or asymmetric competition for modelling shading). The total net competition effect  $C_i(\sigma)$  ( $\text{m}^2$ ) on tree  $i$  is then the sum of the competition indices of all neighbouring trees:

$$C_i(\sigma) = \sum_{j=1, i \neq j}^n C_{i,j}(B_j, \sigma) \quad (3.3)$$

The effect of competition is thus a crowding effect that influences size growth and mortality, where we do not make difference between conspecific and heterospecific effects. For both processes we used different neighbourhood ranges ( $\sigma_g$  for competition effects on growth,  $\sigma_m$  for competition effects on mortality). In a practical implementation only trees are considered that are within a distance  $R$ , where the weight  $w_{i,j}$  is larger than a predefined constant.  $R$  typically ranges up to 15 m (Uriarte et al., 2004).

### Mortality and individual size growth

We model the decrease of survival probability of a tree  $i$  with an exponential decay depending on the amount of experienced competition. This results in a mortality probability of

$$m_i = 1 - (1 - m_b)e^{-\alpha_m C_i(\sigma_m)} \quad , \quad (3.4)$$

where  $m_b$  is the annual base mortality,  $\alpha_m$  ( $\text{m}^{-2}$ ) the reduction factor and  $\sigma_m$  (m) the standard deviation of the competition kernel (see Section 3.2.2). Likewise, we model the decrease in potential size growth  $g$  ( $\text{cm a}^{-1}$ ). This results in an individual size growth

$g_i$  ( $\text{cm a}^{-1}$ ) of

$$g_i = g \cdot e^{-\alpha_g C_i(\sigma_g)} \quad , \quad (3.5)$$

where  $\alpha_g$  ( $\text{m}^{-2}$ ) is the reduction factor and  $\sigma_g$  (m) the standard deviation of the competition kernel. Competition is thus parameterised by four parameters:  $\alpha_m, \sigma_m$  for mortality and  $\alpha_g, \sigma_g$  for size growth.

### Recruitment

Recruits can be offspring of mother trees or immigrants from the metacommunity. Each year a constant number of  $n$  ( $\text{a}^{-1}$ ) recruits enter the local community. With immigration probability  $m$  an individual recruit is a sample from the metacommunity and placed randomly within the simulated plot area. With probability  $1 - m$  a random mother tree is selected, from which a seed is dispersed according to a radially symmetric log-normal dispersal kernel (Greene et al., 2004) with parameters  $(d_m, d_s)$ , where  $d_m$  (m) is the mean dispersal distance and  $d_s$  (m) is the dispersals standard deviation. The local community is assumed to be a torus (periodic boundary conditions) for seedlings which disperse outside the plot bounds.

### 3.2.3 Linking model and data

A way to test the ability of the models simplified processes to capture the main dynamics of a tropical forest is to calibrate its parameters against a set of field observations. This is not just an exercise in model fitting but a complex and time consuming task of structural fitting the internal model dynamics. It is essential that a set of several observable patterns is selected which provide enough information to make all parameters identifiable with as least uncertainty as possible. If the model is able to successfully reproduce the observed patterns, we assume that the model processes are a (possible) valid representation of the main processes which shape the structure of community assembly for the considered study site.

### Macroecological patterns

We selected three non-spatial and three spatial macroecological patterns against which the model is calibrated. Each pattern provides structural information regarding specific processes (with partial informational overlapping) and represents contrasting aspects of forest structure.

As patterns for non-spatial structure we selected the species abundance distribution (SAD), the diameter stem size distribution (SSD) and the species-individual distribution (SID). The species abundance distribution (SAD) (Hubbell, 2001) is a direct measure of species diversity, where the number of species with a specific number of individuals is counted and aggregated through grouping in classes to the power of 2 (Preston octave curve, Preston (1948)). The diameter stem size distribution (SSD) describes the distribution of individual tree size. We use classes with a width of 10 cm for trees with

a DBH  $< 100$  cm and one class for trees which are larger. The species-individual distribution (SID) combines informations from the previous two patterns. It uses the same size classes as the diameter stem size distribution but represents the mean number of individuals per species in each size class.

As patterns for spatial structure we selected the pair-correlation function  $g(r)$ , the distance decay of community similarity  $F(r)$  and the species-area relationship SAR. The pair-correlation function  $g(r)$  describes how the neighbourhood density of trees, irrespective of its species, varies as a function of distance from a reference tree (Wiegand and Moloney, 2013). Values of  $g(r) = 1$  indicates complete spatial randomness (CSR),  $g(r) < 1$  dispersion and  $g(r) > 1$  aggregation. The distance decay of similarity  $F(r)$ , as a measure of beta-diversity, estimates the probability that two randomly selected trees, which are a distance of  $r$  (m) apart, belong to the same species (Condit et al., 2001). It quantifies species mixing versus intraspecific aggregation. For independent species patterns  $F(r) = 1 - D = \sum_{i=1}^S p_i^2$ , where  $D$  is the Simpson diversity index and  $p_i$  is the proportion of species  $i$  on all species  $S$  (Condit et al., 2001; Wiegand and Moloney, 2013). The species area relationship (SAR) quantifies the number of different species within a certain area, and it has been empirically shown that this relationship follows systematic mathematical relationships (Preston, 1962). While both  $F(r)$  and the SAR consider the position and species identity of each individual, the SAR is more affected by rare species and  $F(r)$  is dominated by the most abundant species (Morlon et al., 2008).

We estimated the six patterns  $S_{\text{obs}}^i(t)$  ( $i, t = 1, \dots, 6$ ) from six BCI censuses (1985-2010 in 5 year steps) using all living trees with a diameter at breast height that is larger than 10 cm as a reproductive size threshold, because neutral theory applies only to adult trees (Hubbell, 2001). The pair-correlation function and the distance decay of similarity were estimated for all distances 1 – 50 m. This covers the scale at which trees compete for light and resources (Stoll and Newbery, 2005) and where seed dispersal takes place (Muller-Landau et al., 2008).

The observed patterns  $S_{\text{obs}}^i = 1/n \sum_{t=1}^n S_{\text{obs}}^i(t)$  were estimated as the mean from all censuses and the simulated patterns as the mean of subsequent observation within a simulation of 2000 years, starting at 1000 years with a 100 year observation interval. This ensures that the model is in a steady state and shows lower uncertainty for the predicted pattern mean.

### Quantification of model deviation

For a specific parameter set  $x$  the mismatch between an observed pattern  $S_{\text{obs}}$ , which is also called “summary statistic” (Beaumont et al., 2010), and a simulated pattern  $S_{\text{sim}}$  is quantified using the mean relative error MRE (Tofallis, 2015):

$$\text{MRE}(S_{\text{obs}}, S_{\text{sim}}(x)) = \frac{1}{n} \sum_{j=1}^n \left| \frac{S_{\text{sim}_j}(x) - S_{\text{obs}_j}}{S_{\text{obs}_j}} \right|, \quad (3.6)$$

where  $S_{\text{obs}_j}$  is the  $j$ -th element of the observed pattern,  $S_{\text{sim}_j}(x)$  is the  $j$ -th element of the simulated pattern for the parameter set  $x$  and  $n = \dim(S_{\text{obs}})$ . The mean relative

error  $\text{MRE}(x)$  for a parameter set  $x$  and  $m$  patterns  $S_{\text{obs}}^i, S_{\text{sim}}^i(x)$  with  $i = 1, 2, \dots, m$  is then our objective function  $Q(x)$ :

$$Q(x) := \text{MRE}(x) = \sum_{i=1}^m w_i \cdot \text{MRE}(S_{\text{obs}}^i, S_{\text{sim}}^i(x)) \quad \text{with} \quad \sum_{i=1}^m w_i = 1 \quad , \quad (3.7)$$

where the  $w_i$  are specific weights given to each of the  $n$  different patterns. In this study we use the trivial case of equal weights:  $w_i = 1/n$  for  $i = 1, 2, \dots, m$ , but other weightings including knowledge about measurement uncertainty or prediction preferences are possible.

### Model calibration

For the model calibration we rely on algorithms that only need a small amount of simulation runs and are applicable at a distributed computing system (e.g. high performance cluster (HPC)). This renders most MCMC techniques obsolete, as their possible parallelisation is limited and they usually need about  $10^6$  simulation trials. Instead we use stochastic search methods which have proven to be efficient to calibrate dynamic forest models (Lehmann and Huth, 2015). We decided to use the JADE method (Zhang and Sanderson, 2009). It is a population based approach that is especially suited for parallel processing. We use a maximum of  $10^5$  simulation trials on 50 parallel threads, which needs about 2 days for completion. There are 12 free parameters, which are summarised in Table 3.1. To see what is the best possible prediction the model is capable of, we did not fix any parameters by estimates from field data or previous studies and only chose meaningful parameter ranges based upon this data. The calibration method yields a set of best parameters  $\hat{x}$  found and a list of parameters (points in parameter space) visited. Additional to the best set of parameter values, we form a list  $L$ , by using all sets of parameter values whose model deviation (see Section 3.2.3) is less than 5% to the deviation of the best set found. If we assume that the model using  $\hat{x}$  is a reasonable approximation to the processes which generated the field observations, then  $L$  provides parameter sets which also give reasonable approximations. This enables us to further quantify the models structure. We then estimate the correlation matrix of  $L$  and calculate the marginal distribution for every parameter (Hartig et al., 2011).

To simplify the analysis, we apply a classification scheme to the correlation matrix and the marginal distributions. For each parameter let  $c_n$  be the number of parameters (excluding the parameter itself) which have an absolute correlation coefficient  $> 0.5$  and CV the coefficient of variation of its marginal distribution. The correlation of the parameter is then assumed to be “high” if  $c_n > 3$ , “moderate” if  $c_n = 3$  and “low” if  $c_n < 3$ . The uncertainty of the parameter estimate is assumed to be “high” if  $\text{CV} > 10\%$ , “moderate” if  $5\% \leq \text{CV} \leq 10\%$  and “low” if  $\text{CV} < 5\%$ .

### Global sensitivity analysis

To quantify the global sensitivity of the model parameters and patterns we use the elementary effects method of Morris (1991) and its recent developments (Campolongo

**Table 3.1:** Overview of the 12 free models parameters, their units and plausible input ranges for the chosen study site.

Name	Description	Unit	Range
$S$	Number of species in the regional pool	-	200-1000
$n$	Number of local recruits	$\text{a}^{-1}$	450-550
$\theta$	Fundamental biodiversity number	-	40-60
$m$	Immigration probability	%	1-20
$d_m$	Mean dispersal distance	m	1-50
$d_s$	Std.-dev. of dispersal distance	m	1-50
$g$	Potential stem diameter growth	$\text{cm a}^{-1}$	0.1-2
$c_g$	$\sigma$ of growth competition kernel	m	0.1-5
$c_m$	$\sigma$ of survival competition kernel	m	0.1-5
$\alpha_g$	Reduction factor on growth	$\text{m}^{-2}$	0-2000
$\alpha_m$	Reduction factor on survival	$\text{m}^{-2}$	0-2000
$m_b$	Base mortality	$\% \text{ a}^{-1}$	0.1-10

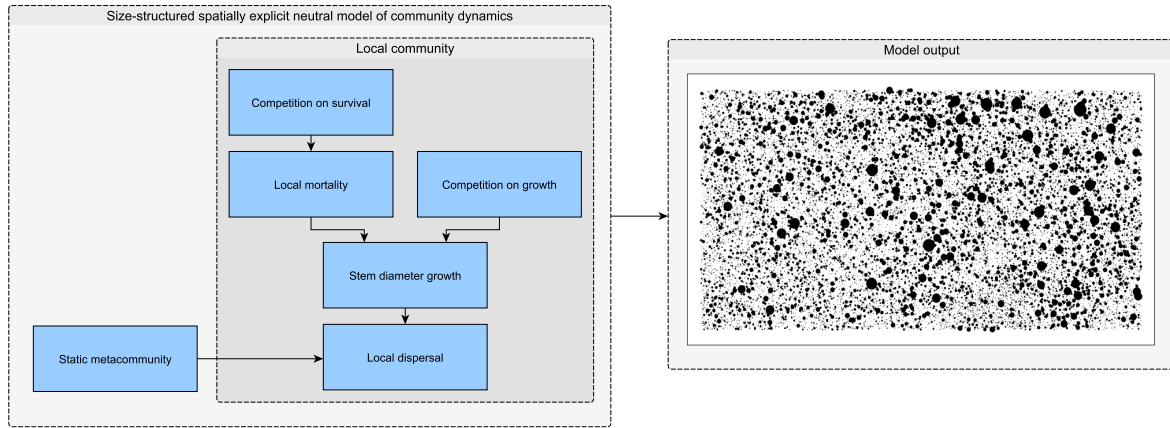
et al., 2007) as a qualitative screening framework. The method estimates which parameters (input factors) have insignificant, linear/additive or nonlinear/interactive effects. For each input factor two sensitivity measures are calculated: the absolute mean  $\mu^*$  of elementary effects, as a measure of overall influence and its standard deviation  $\sigma$ , as a measure of higher order effects. The approach is explained in detail in Section B.1.1.

To analyse the results of the screening process, one inspects the  $\mu^*, \sigma$ -plot for each of the six patterns. To simplify the potential conclusions, we apply additional aggregations on the plots. First, we calculate the row-wise mean of the  $\mu^*, \sigma$ -matrices to get an overall  $\mu^*, \sigma$ -plot of parameter sensitivity over all patterns. This is further aggregated by grouping parameters into four distinct process classes: migration ( $S, \theta, m, n$ ), dispersal ( $d_m, d_s$ ), size growth ( $g, c_g, \alpha_g$ ) and mortality ( $m_b, c_m, \alpha_m$ ). This results in an  $\mu^*, \sigma$ -plot of process sensitivity over all patterns. At last, we calculate the column-wise mean of the  $\mu^*, \sigma$ -matrices to get an overall  $\mu^*, \sigma$ -plot of pattern sensitivity over all parameters.

### Model schedule

Each simulation starts with an empty local community and a pre-calculated expected rank-abundance distribution for the metacommunity. For a predefined number of 2000 years a timeloop is executed, where for each year the following schedule takes place: 1) the total net competition effect for every individual is calculated, 2) probability of dying for every individual is calculated and evaluated, 3) individuals which survived, growth in diameter size, 4) local recruitment takes place. A flow-chart of this scheduling can be seen in Fig. 3.2.





**Figure 3.2:** Flow-chart of model processes with a snapshot from above of a simulated model output. The snapshot shows the local community with an area of  $1000 \text{ m} \cdot 500 \text{ m} = 50 \text{ ha}$ . Circles denote single trees, where the circles diameter equals the stem diameter of a tree resized by a factor of 5.

### 3.3 Results

#### 3.3.1 Prediction of ecological patterns

The model calibration is performed as explained in Section 3.2.3. An overview of the resulting parameter estimates is shown in Table 3.2. Some of the estimates can be compared to previous estimates from the literature or field data. The estimated number of local recruits  $n = 516 \text{ a}^{-1}$  is close to the mean number of recruits (with a DBH  $> 10 \text{ cm}$ ) of  $497.76 \text{ a}^{-1}$  from the study site over all six censuses. The base mortality of  $1.2 \text{ \%/a}$  in conjunction with competition results in an effective mortality of  $2.05 \text{ \%/a}$  which is in the range of the observed mortality of  $2.34 \text{ \%/a}$  (mean over five censuses). The estimated potential stem diameter growth of  $0.7 \text{ cm/a}$  is close to the mean observed potential stem diameter growth of  $0.76 \text{ cm/a}$ , where potential diameter growth was estimated as the mean stem size growth for trees with a stem diameter from  $50 \text{ cm}$  to  $100 \text{ cm}$ . The mean dispersal distance of  $23.8 \text{ m}$  is lower than values from the literature based on seed trap data. For example Condit et al. (2002) estimated a mean dispersal distance of  $39 \text{ m}$  for 65 tree species in BCI and Muller-Landau (2001) estimated a mean of  $39.5 \text{ m}$  for 81 tree species. The fundamental biodiversity number of  $56.9$  is higher and the immigration probability of  $3.7\%$  is lower than previous estimates on the same study site, e.g.  $47.2$  and  $10\%$  (Volkov et al., 2003),  $50$  and  $10\%$  (Hubbell, 2001),  $48.5$  and  $7.9\%$  (McGill, 2003). Through geometrical properties of the plot, it is possible to functional link the immigration rate to the dispersal distance through  $m \approx Pd_m/(\pi A)$  (Chisholm and Lichstein, 2009), where  $P = 3000 \text{ m}$  is the plots perimeter and  $A = 500\,000 \text{ m}^2$  its area. This results in an estimate of  $m = 4.5\%$  which is in agreement with our estimate.

We use all parameter sets that are within a  $5\%$  bound of the best parameter set found, as a measure of interactive effects and uncertainty of parameter estimates (see

**Table 3.2:** Summary of calibration results for the 12 free parameters. The last two columns show the uncertainty and correlation classification based on the scheme from Section 3.2.3.

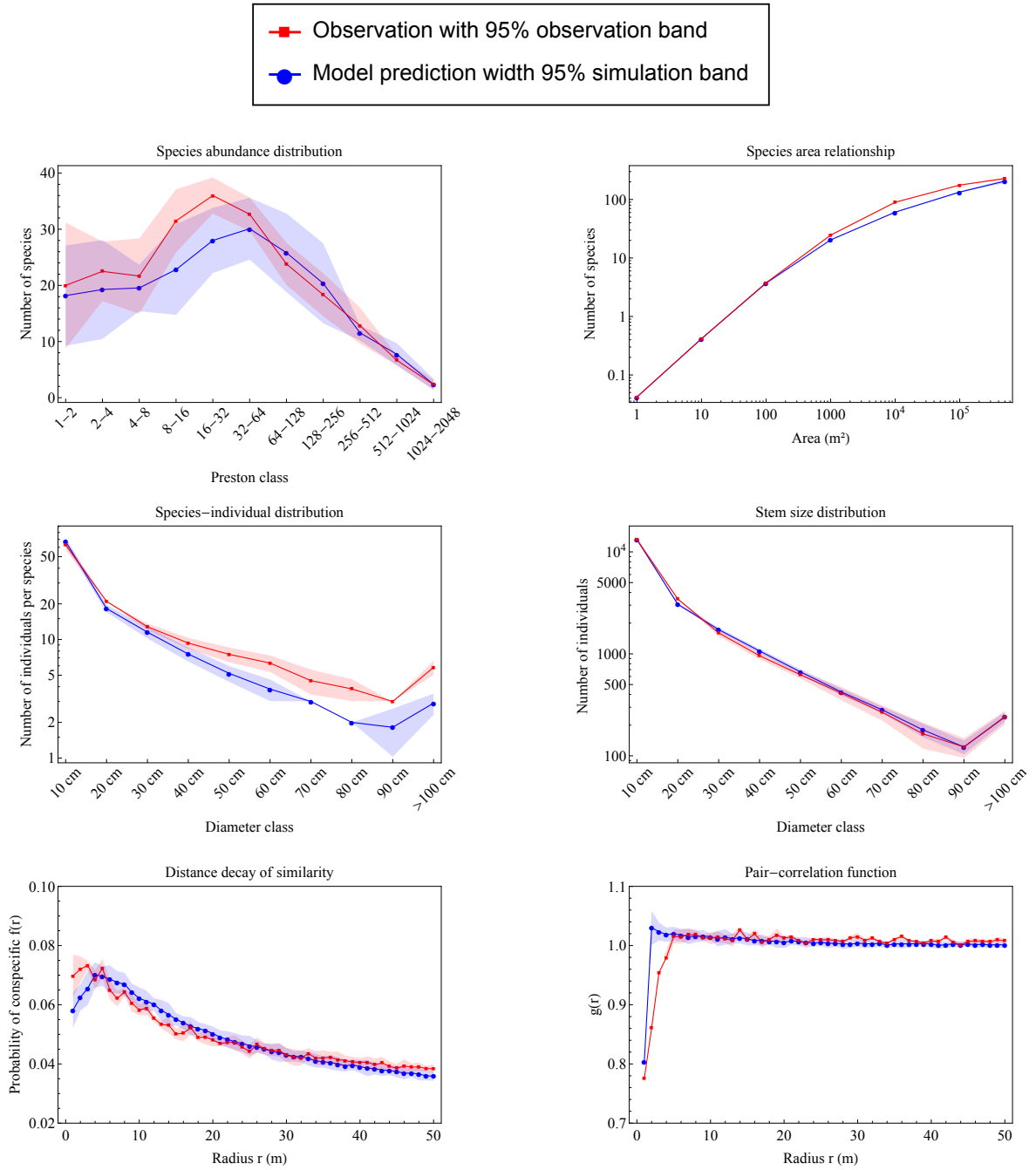
Name	Description	Unit	Estimate	Uncertainty	Correlation
$S$	Number of species in the regional pool	-	653	moderate	low
$n$	Number of local recruits	1/a	516	low	low
$\theta$	Fundamental biodiversity number	-	56.9	low	high
$m$	Immigration probability	%	3.7	high	high
$d_m$	Mean dispersal distance	m	23.8	low	moderate
$d_s$	Std.-dev. of dispersal distance	m	21.2	high	moderate
$g$	Potential stem diameter growth	cm/a	0.7	low	high
$c_g$	$\sigma$ of growth competition kernel	m	1.3	low	low
$c_m$	$\sigma$ of survival competition kernel	m	0.4	moderate	moderate
$\alpha_g$	Reduction factor on growth	$m^{-2}$	617	high	moderate
$\alpha_m$	Reduction factor on survival	$m^{-2}$	344	high	moderate
$m_b$	Base mortality	%/a	1.2	low	moderate

Section 3.2.3). The full correlation matrix is shown in the appendix (see Fig. B.2). The parameters with the highest correlation (absolute linear correlation  $> 0.5$ ) are the fundamental biodiversity number  $\theta$  (correlates with  $S$ ,  $m$ ,  $d_m$  and  $d_s$ ), immigration probability  $m$  (correlates with  $S$ ,  $\theta$ ,  $d_m$  and  $d_s$ ) and potential diameter growth  $g$  (correlates with  $c_g$ ,  $c_m$ ,  $\alpha_m$ ,  $\alpha_g$  and  $m_b$ ). The complete marginal parameter distributions are shown in the Appendix (Fig. B.3). The parameter estimates with the highest uncertainty are  $m$  (CV = 13%),  $d_s$  (CV = 16%),  $\alpha_m$  (CV = 11%) and  $\alpha_g$  (CV = 25%).

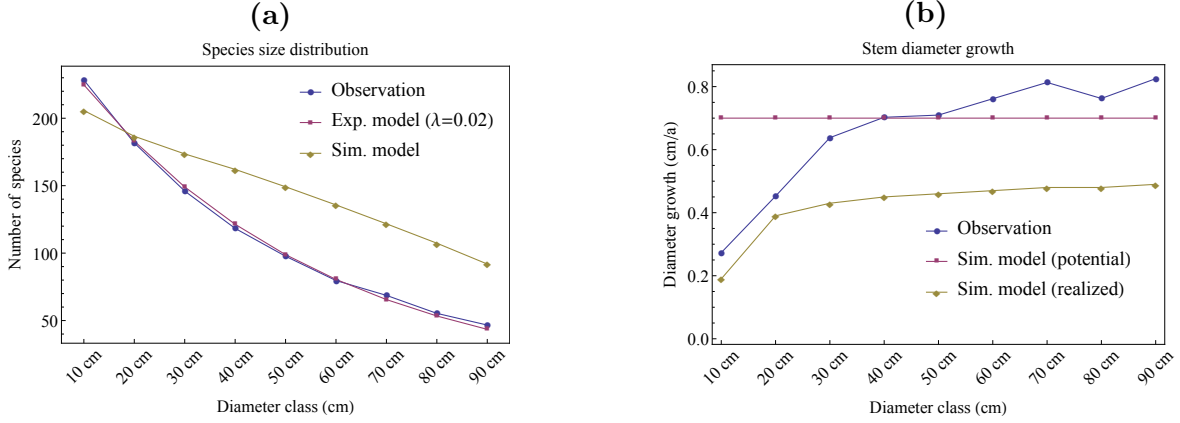
The individual errors of the model predictions against each pattern are shown in Table 3.3 and result in a mean relative error of 10.9%. Albeit its simplicity, the model is able to reproduce five out of six, and the general properties for each observed pattern (see Fig. 3.3). This includes: the hollow-curved SAD, the fat-tailed distribution of individual size (which can not be approximated by an exponential distribution for this study site), the decreasing number of individuals per species with increasing size, the power-law behaviour of the SAR on a local scale, the observed dispersion effect for small radii around focal trees, represented by the pair-correlation function and the decay of

**Table 3.3:** Relative error of the best found parameter set against each individual pattern.

Pattern	Description	Relative error (%)
SAD	Species abundance distribution	12.4
SSD	Diameter stem size distribution	5.6
SID	Species individual size distribution	29.0
SAR	Species area relationship	12.1
F	Distance decay of similarity	4.8
G	Pair-correlation function	1.2
Mean		10.9



**Figure 3.3:** Visualization of the model predictions against the mean field observations for the six observed patterns: species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G).



**Figure 3.4:** Shown is the a) Species size distribution including the observation data, a fitted theoretical exponential distribution and the result from the simulation model b) Stem diameter growth of observation data, the potential stem diameter growth and the realized stem size growth of the simulation model.

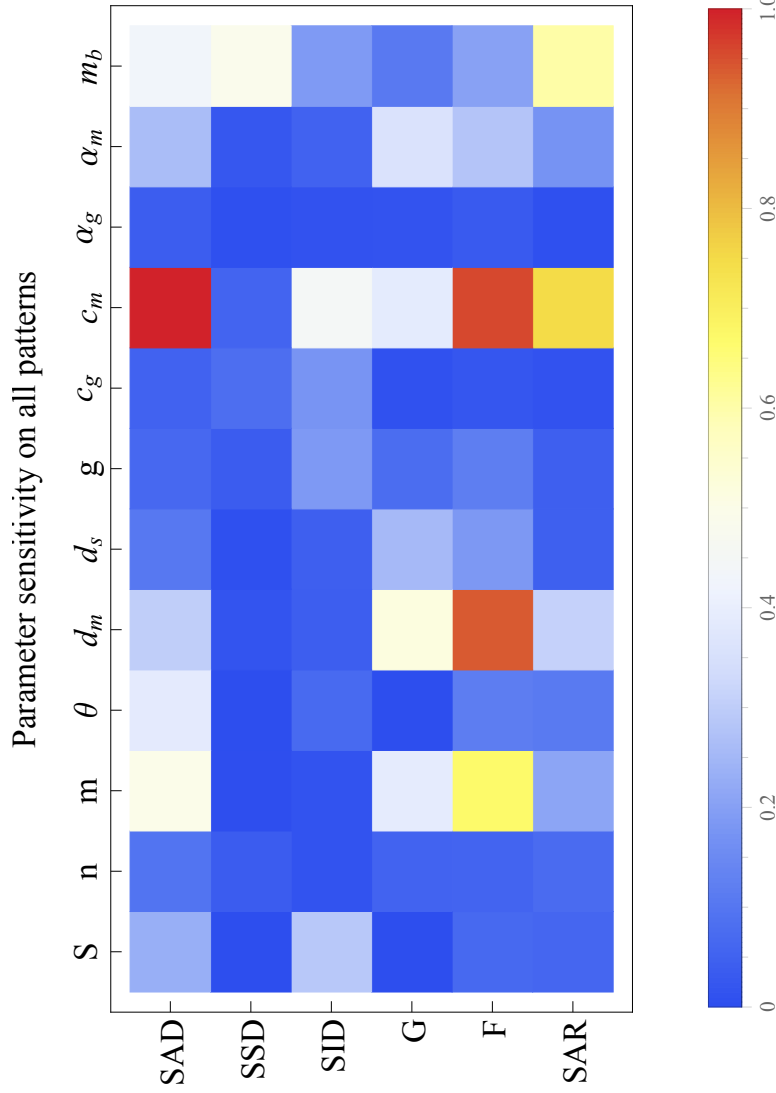
similarity with increasing distance.

The Species-individual distribution has the highest relative error of 29.0% over all patterns. The number of individuals per species is always lower for trees with a DBH  $> 20$  cm. Inspecting results from diameter growth shown in Fig. 3.4 reveals that the realized stem diameter growth is lower and that the number of species within a certain size class is higher than observed. Just increasing the potential size growth  $g$  has almost no effect, as the effective size growth is limited by the reduction factor.

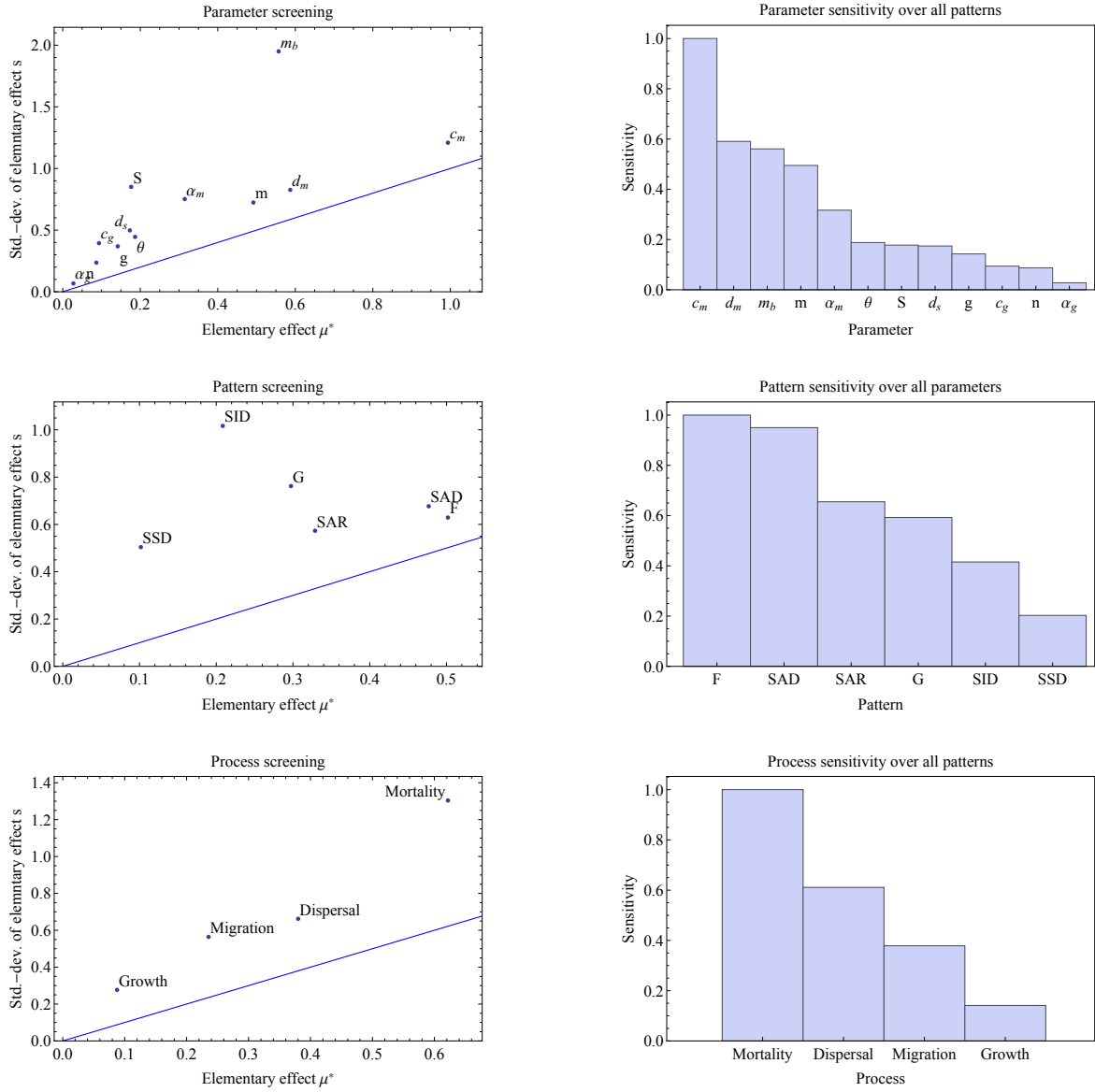
### 3.3.2 Global parameter screening

The qualitative model screening is carried out as explained in Section 3.2.3. The heat map of  $\mu^*$ -values shown in Fig. 3.5 is a first indicator that  $c_m$ ,  $d_m$ ,  $m_b$  and  $m$  are the most sensitive parameters, while the SAD and F are the most sensitive patterns. The aggregated screening results shown in Fig. 3.6 allows a closer examination. All parameters, processes and patterns show nonlinear behaviour and/or interaction effects indicated by the boundary line  $\mu^*/\sigma$  (based on a classification scheme proposed by Sanchez et al. (2014)). The overall most sensitive parameters and patterns conform with our estimates from the heat map, while parameters and patterns related to size growth are the least sensitive. Individual  $\mu^*, \sigma$ -plots on each pattern are shown in the Appendix Fig. B.1. Aside mortality related parameters the two most additional sensitive parameters on each individual pattern are: the fundamental biodiversity number  $\theta$  and the immigration probability  $m$  on the SAD, the number of species in the metacommunity  $S$  and the potential diameter growth  $g$  on the SID, the potential diameter growth  $g$  and  $\sigma$  of the growth competition kernel on the SSD, the dispersal mean  $d_m$  and the immigration probability  $m$  on all spatial patterns G, F and SAR. Grouping the parameters into process classes as discussed in Section 3.2.3 shows that the ranking of processes is:

mortality, dispersal, migration and size growth, from most to least sensitive.



**Figure 3.5:** Shown is a heat map of scaled absolute elementary effects  $\mu^*$  of each parameter on each individual pattern. The patterns are the species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G).



**Figure 3.6:** Global model screening results on all parameters, patterns and processes. Points above the boundary line show nonlinear behaviour and/or interaction effects. The patterns shown are the species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G).

### 3.4 Discussion

In this study we extended the classical neutral theory by incorporating individual tree size. In contrast to previous studies (O'Dwyer et al., 2009) tree size is in our model directly connected to mortality through intra&inter-specific crowding competition, based on actual tree diameter as assumed in statistical neighbourhood models (Uriarte et al., 2004). Individual tree size is therefore affected and effects the processes in its local neighbourhood. The model also extends beyond previous simplifications of simulation models based on neutral theory (Chave et al., 2002). As the model does not rely on constant community size (zero-sum), the number of individuals fluctuate in time. This can be interpreted as non-constant resource availability (Etienne et al., 2007a). Additionally, we presented a new construction scheme for the metacommunity by sampling from a mean rank abundance curve. This avoids the inevitable uncertainty regarding model predictions by not sampling from a specific realisation.

Recent advances in the calibration of dynamic vegetation models (Lehmann and Huth, 2015) enabled us to evaluate the ability of the model to reproduce observational patterns within a short time-frame using high performance computing clusters. This is especially useful in the context of future hypothesis testing on community models, as this technique allows to discriminate between conflicting theories rapidly.

The introduction of tree size allows the prediction of new biodiversity patterns. As examples we used in this study the individual tree size distribution, the species size distribution and as a structural conjunction: the species-individual distribution. We found that the model is able to predict the main dynamics of all six spatial and non-spatial patterns simultaneously (using a single parameter set) and fits five of the six patterns within a narrow error bound. That means that the presented model, despite its simplicity, incorporates enough mechanism to correctly predict several fundamental diversity pattern. It is therefore a possible realistic representation of the dominant ecological processes which shape the structure of diverse tropical forest communities.

From our perspective a highlight of this study is the combination of model calibration and global model behaviour screening to identify uncertainties, correlations, sensitivities, nonlinear and interactive effects among parameters and processes. We estimated that the fundamental biodiversity number  $\theta$ , the immigration probability  $m$  and the potential stem diameter growth  $g$  are highly correlative parameters, while the estimate of  $m$  is also highly uncertain. The parameters  $\theta$  and  $m$  correlate the most with the number of species in the metacommunity  $S$  and both dispersal kernel related parameters. As the dispersal is a sensitive process on all spatial patterns there exists an indirect link to the prevalent species composition. The same effect can be observed on the potential stem diameter growth  $g$ , which correlates with all mortality related parameters. That means, even though size growth is estimated to be the least sensitive process in the model its indirect connection to mortality (the main driver in shaping community structure) reveals its importance on the overall dynamics.

We detected that the observed species size distribution can be perfectly described through an exponential distribution (see Fig. 3.4a) for trees with a DBH  $\leq 100$  cm at the study site. This distribution can be interpreted as the equilibrium distribution of a

one dimensional transport equation, where individual species recruit into the smallest diameter class, grow in size and die at a constant rate. However, the model was not able to correctly predict this distribution and the directly linked size distribution of individuals per species in a simultaneous framework. We found through additional analysis that the local mortality and the effective stem size growth rate are too slow. Nonetheless, this results in a good prediction of the stem size distribution. Additionally, we estimated a lower immigration rate, higher diversity and lower mean dispersal distance than previous studies. Such an effect is also seen in May et al. (2015) for the parameters that represented the best compromise in fitting all patterns simultaneously. Increasing the immigration rate (while decreasing diversity) or mean dispersal distance would distort the spatial structure and directly affects competition indices and thus size growth and tree mortality. We argue here that the representation of dispersal through a single local dispersal kernel and random placement through immigration might be too simplistic. This illustrates the importance of confronting the model with several patterns at once, as each carries different aspects of the underlying ecological processes and provides the basis for subsequent identification of misspecified mechanisms (Wiegand et al., 2003; Grimm et al., 2005).

The results let us conclude that the presented model is an important extension of existing models in the context of neutrality. It is the first neutral model to date that is able to predict various spatial, non-spatial and size-related biodiversity pattern simultaneously. We presented a complete framework including the combination of model calibration and global screening, which can be used to test future hypotheses about community structure in forests. In the future more complex dispersal and competition modes can be tested, as the model has shown some limitations regarding the interaction of growth, dispersal, mortality and species composition. As this points to the possible involvement of non-neutral processes, future work could rely on the framework of Uriarte et al. (2010) and Fortunel et al. (2016).





# Chapter 4

## The global state of forest fragmentation: linking high-resolution maps and innovative cluster analysis

---

This chapter is currently under review as Lehmann, S., Fischer, R. Huth, A. (2017). “The global state of forest fragmentation: linking high-resolution maps and innovative cluster analysis”. *Landscape Ecology*

### Abstract

Cluster analysis of forest cover is a fundamental method for quantifying the state of forest fragmentation. High demands on computing resources are needed for current high resolution forest cover maps. We demonstrate how important features of forest fragmentation can be captured using an innovative method for cluster analysis in combination with data compression. We present for the first time results regarding the structure and size of all forest fragments using a 30 m map, globally and in three biomes: tropical, temperate and boreal forests. Our results show that the forests worldwide contain 409 million fragments and that 36% of global forest area lies within 100 metres of forest edges. By providing a world map showing the global state of forest fragmentation, we discover that the forest area in poor condition is twice as high in temperate and boreal forests as in tropical forests.

### 4.1 Introduction

Global forests cover 30% of total land surface (Bonan, 2008; Hansen et al., 2010). Forests provide important ecosystem services such as carbon storage, regulation of the hydrologic cycle and forest products (Bonan, 2008). Deforestation and fragmentation of forests by human activities occurs worldwide and is the main driver for biodiversity loss and species extinction (Pereira et al., 2010; Rands et al., 2010). Fragmentation leads to altered microclimate conditions in forest edges (Ewers and Banks-Leite, 2013), such as higher light incidence, reduced humidity and higher temperatures. This results in increased tree mortality and subsequently lower above-ground carbon stocks in edges (Laurance et al.,

2002; Harper et al., 2005; Broadbent et al., 2008; Laurance et al., 2011). The alteration of forest structure has an impact not only on global carbon budgets, it also affects habitat quality as a result of changing forest dynamics within forest edges (Groeneveld et al., 2009; Pütz et al., 2014). Thus, quantification of forest fragmentation is an important task for monitoring the progress of global deforestation and change in habitat conditions.

Previous studies aimed at forest fragmentation were either evaluating local structural quantities using moving windows (Wade et al., 2003; Riitters et al., 2000, 2016), nearest neighbour approaches (Haddad et al., 2015) or concentrated on a regional scale (Pütz et al., 2014) only. None of these studies provide a deeper insight into the global distribution of forest fragment quantities.

In this study we analyse forest fragmentation at a global scale based on maps of forest cover (Hansen et al., 2013). In the course of this, we answer three specific questions. (1) Is it possible to construct an efficient method using standard computational resources to calculate properties of each forest fragment on a global scale? (2) How can forest fragmentation be characterized at global scale and for three forest biomes? Can we observe differences between the biomes? (3) How can the condition of forest fragments be characterized and can this be visualized in a global map? On one hand the answers to these questions provide a theoretical insight into the inherent structure and an outlook about underlying ecological processes of forest fragmentation. On the other hand a global view on deforestation provides information on future decisions regarding conservation.

## 4.2 Material and methods

### 4.2.1 Database: Hansen’s forest cover map

To demonstrate our developed methods we analysed the high resolution forest cover map from Hansen with tree cover data from 2000 (Hansen et al., 2013). The map has a resolution of 1 arc second per cell (pixel) ( $\approx 30$  m at the equator) and is based on Landsat data. The map extends from  $80^{\circ}\text{N}$ ,  $180^{\circ}\text{W}$  to  $180^{\circ}\text{E}$ ,  $60^{\circ}\text{S}$  with a total of  $1\,440\,000 \cdot 560\,000$  pixels ( $\approx 10^8$  pixels, 51.3 GB compressed GeoTIFF size). The map’s surface area of  $47\,174.12 \times 10^6$  ha is  $3718.22 \times 10^6$  ha less than the WGS-84 ellipsoid surface area of  $50\,892.34 \times 10^6$  ha due to missing data around the poles. Following Hansen et al. (2013), we also used a bi-level forest/non-forest classification threshold value of 50% forest cover for the main analysis in order to match the published results for the total forested area. Other values reported in the literature are 0% (Riitters et al., 2016) and 30% (Sexton et al., 2013).

### 4.2.2 The advanced cluster analysis algorithm

The classical Hoshen-Kopelman algorithm (Hoshen and Kopelman, 1976a), as a special case of the union-find algorithm and its recent developments (Bailey and Johnston, 2007), serve as a basis for our new developed cluster analysis algorithm.

The bi-level forest/non-forest grid is considered a graph, in which each forest pixel

is a single vertex. The cluster analysis algorithm performs a top-down raster scan on the map grid and assigns labels (represented as integers) to occupied (forest) cells. As multiple labels can refer to the same cluster (e.g. U-shaped clusters), a set of labels which refer to the same cluster is termed an “equivalence class”, while a representative element of this class is termed “root”.

Each time an occupied cell is encountered a check is done to determine, whether this cell has any occupied neighbouring cells (here: 4-connected, cardinal directed cells N/S/E/W), which have already been scanned. If so, a *union* operation is performed to specify that this neighbouring cell belongs to the same equivalence class. Then a *find* operation is performed to find the representative member of the equivalence class with which to label the current cell. If the current cell has no neighbouring cells occupied by forest it is assigned a new, previously unused label. For analyses that go beyond the numerical characterisation of forest fragments (see Section 4.3.2) the grid has to be scanned a second time, where only *find* operations are used to re-label cells with their final assignment of the representative element. For details on the used data structures see Section C.1. The cluster analysis algorithm and its subsequent modules are implemented as a C++ program and is available on request.

### 4.2.3 Compression of labelled forest fragment maps

In case of a full two-pass cluster analysis a map with temporary labels is created in a first pass and a map with final labels in the second pass. For modern high-resolution maps the labelled images take up large amounts of space. For example, the 30 m label map from Section 4.2.1, using 64-bit labels for the worst case checkerboard scenario, would consume about 6000GB ( $1.440.000 \cdot 560.000 \cdot 8/1024^4$ ) of space - too much to be held in random access memory (RAM). Even if we have a hard drive that can hold the map, reading it completely would take more than 8 hours, at an average sequential reading rate of  $200\text{MB s}^{-1}$ . So, we are limited, both by storage size and access speed. To overcome these limitations an online compression mechanism has been used. It compresses at least enough to compensate the additional processing time needed for decompression in comparison to using no compression. The here developed method is specially adapted for correlations in labelled maps and is possible to reduce the size of the labelled map from Section 4.2.1 from 6TB down to 8.8GB, a reduction factor of  $\sim 680$ . Details on the algorithm are presented in Section C.2.

### 4.2.4 Forest fragment characteristics

We used an equirectangular projection of the forest cover map using the WGS84 ellipsoid (EPSG:4326 projection), which enables us to use simple methods to calculate length and area of each map element (cell or pixel). To calculate distances, we can not use the famous Vincenty formula (Vincenty, 1975), because this method measures distances on a geodesic between two points on an ellipsoid and not along a circle of latitude (rows in equirectangular projection) or along a meridian (columns in equirectangular projection). Nonetheless, exact formulas for the calculation of pixel length and area along a circle

of latitude do exist (see Section 4.2.4 and Section 4.2.4). As pixel edge lengths and pixel areas stay the same along a circle of latitude in equirectangular projection, we precalculate pixel lengths and pixel areas for every row of the map to enable faster processing.

### Forest edge detection and edge length calculation

The von Neumann neighbourhood (N/W/E/S) serves as a basis for the analysis of neighbouring cells. To be recognised as an edge of a forest fragment, each of the four sides of a forest pixel is tested, to determine whether at least  $p_e$  pixels are non-forest pixels in the specific cardinal direction. See Fig. C.1 in the supplementaries for a graphical explanation. If a pixel is classified as an edge of a forest fragment, its length is calculated depending on whether it is an edge along the meridian at a certain latitude (vertical) or along a circle of latitude at a certain longitude (horizontal). The selection of  $p_e$  effects results regarding edge length and edge affected area. The obvious choice of  $p_e = 1$  pixel at the equator is not useful for high resolution maps, as the length of a pixel on such a map corresponds to 30 m near the equator. We choose  $p_e$  to be the number of pixels which equals a distance of  $\approx 60$  m. This results in  $p_e = 2$  vertical and horizontal pixels at the equator and up to  $p_e = 12$  horizontal pixels near the poles. The question, how considering diagonal pixels in the edge detection algorithm changes result, is not part of this study.

To calculate the horizontal distance  $d_h$  (m) at latitude  $\phi$  (rad) for a longitudinal difference of  $\Delta\lambda$  (rad) we use an exact formula (Osborne, 2013):

$$d_h(\phi, \Delta\lambda) = \frac{a}{\sqrt{1.0 - e_2 \cdot \sin^2(\phi)}} \cdot \cos(\phi) \cdot \Delta\lambda \quad , \quad (4.1)$$

where  $a$  (m) is the length of the semi-major axis of the WGS84 ellipsoid and  $e_2$  (m) is its squared eccentricity. To calculate the vertical distance  $d_v$  (m) on the meridian of a latitude pair  $\phi_1$  and  $\phi_2$  we use a polynomial approximation (see e.g. Osborne (2013)):

$$\begin{aligned} v(\phi) &= a_0 \sin(\phi) - a_2 \sin(2\phi) + a_4 \sin(4\phi) \\ &\quad - a_6 \sin(6\phi) + a_8 \sin(8\phi) \\ d_v(\phi_1, \phi_2) &= v(\phi_2) - v(\phi_1) \end{aligned} \quad (4.2)$$

The coefficients are  $a_0 = 6367449.156$ ,  $a_2 = 16038.509$ ,  $a_4 = 16.833$ ,  $a_6 = 0.022$  and  $a_8 = 0.00003$  in metre.

### Fragment area calculation

Although approximate methods like the trapezoid formula can be used to calculate pixel area, we use the exact formula for improved accuracy. Because a priori tabulated values are utilised during the analysis, there is no additional negative impact on the running time of the program. Calculation of the surface area  $A(\phi, \Delta\lambda)$  ( $m^2$ ) of a latitude-

longitude rectangle from latitude  $\phi$  to zero using the exact formula (Mularie, 2000) yields:

$$\begin{aligned} t &= \sqrt{1.0 - b^2/a^2} \\ z_m &= 1.0 - t \cdot \sin(\phi) \\ z_p &= 1.0 + t \cdot \sin(\phi) \\ A(\phi, \Delta\lambda) &= 2\Delta\lambda \cdot b^2 \frac{\tanh^{-1}(t \cdot \sin(\phi))}{t} + \frac{\sin(\phi)}{z_p \cdot z_m} \end{aligned} \quad (4.3)$$

where  $a$  (m) is the length of the semi-major axis of the WGS84 ellipsoid and  $b$  (m) the length of its semi-minor axis. The area of a rectangle from latitude  $\phi_1$  to  $\phi_2$  is then simply the difference  $A(\phi_1, \phi_2, \Delta\lambda) := A(\phi_1, \Delta\lambda) - A(\phi_2, \Delta\lambda)$ .

### Estimation of edge-affected area

To estimate the edge-affected area of a fragment we use the Didham&Ewers core area model (Didham and Ewers, 2012). The edge-affected area depends on the shape of the fragment: the higher the perimeter-to-core ratio, the higher the fraction of edge-affected area. The size index (Patton, 1975)

$$S_I = \frac{P}{2\sqrt{\pi A_T}} \quad (4.4)$$

where  $P$  (m) is the fragment's border length (perimeter) and  $A_T$  (m<sup>2</sup>) its total area, is a scale independent measure to quantify the deviation of a fragment's shape from circularity. It ranges from 1.0 for a circle to  $+\infty$  for increasing shape complexity. The edge-affected area  $A_E$  (m<sup>2</sup>) is estimated using the size index  $S_I$  depending on the edge effect depth  $d_E$  (m<sup>2</sup>):

$$A_E \approx d_E P - S_I^2 \pi d_E^2 \quad (4.5)$$

This formula has its maximum at  $Pd_E = 2A_T$ . When  $Pd_E > 2A_T$  we set  $A_E = A_T$ .

The spatial range for edge effects varies from 10 m to more than 1000 m (de Paula et al., 2016). Throughout this study we use an edge effect depth  $d_E$  of 100 m as a rather conservative (Laurance, 2000; Laurance et al., 2011), but not important choice, because we do not evaluate additional habitat impacts like carbon loss through edge effects.

### 4.2.5 Structural fragment analysis for different ecoregions

In addition to analysing the forest cover map, we intersected the forest fragment map with a map of 14 terrestrial ecoregions (Olson et al., 2001a) rasterised at 90 arc seconds. The eight forest biomes are merged into three zones: tropical (tropical and subtropical moist broadleaf forests, tropical and subtropical dry broadleaf forests, tropical and subtropical coniferous forests), temperate (temperate broadleaf and mixed forests, temperate coniferous forests) and boreal (boreal forests/taiga). This results in four maps,

where each analysis can be done in a single pass over the forest cover map (see Section 4.2.2). This takes  $\sim 01:30\text{h}$  on a 3.20Ghz CPU using at most 150MB of system memory. For this part of the study we calculated for every forest fragment the size, edge length, edge area and the shape index. Based on this we derived four distributions, where the independent variable in each case is represented by 10 fragment size classes  $[0 - 1), [10^0 - 10^1), \dots, [10^8 - 10^9)$  in ha:

1. number of fragments (fragment size distribution)
2. forest area (fragment area distribution)
3. forest edge area (fragment edge area distribution)
4. mean shape index (fragment shape index distribution)

For the calculation of edge-affected area we assume an edge depth of  $d_E = 100\text{m}$ . (see Section 4.2.4). To test how the fragment distribution and the shape index distribution can be described through a functional relationship, we fit a power law  $p(x) \sim x^{-\alpha}$  via maximum likelihood method (Virkar et al., 2014) to all bins except the first and the last. This selection minimized the test statistic of a two-sample Kolmogorov–Smirnov test.

#### 4.2.6 Mapping forest conditions in relation to fragmentation

For a graphical representation of forest fragment condition we created a 4-colour map showing the relation of forest fragment core area to the total fragment area, as an index of local forest condition. This value  $\xi$ , ranging from 0 to 1, expresses the relative amount of edge-affected area from each forest fragment. Higher values represent better fragment conditions, as the fraction of unaffected core area is higher. For better visual differentiation, we grouped these values into 4 distinct classes:

1. “poor”:  $\xi \leq 25\%$  (core area/total fragment area)
2. “moderate”:  $\xi \leq 50\%$
3. “good”:  $\xi \leq 75\%$
4. “very good”: otherwise

The calculation of the classification map needs two passes over the forest cover map (see Section 4.2.2) and uses the developed compression scheme (Section 4.2.3) for efficient procession of intermediate results. For presentation purposes the classification map has been resampled using a factor of 1:250 ( $\approx 6\text{ km}$  pixel length at the equator, Fig. 4.4).

**Table 4.1:** Summary of results for the forest fragment analysis using a threshold of 50% for forest/non-forest classification and  $d_E = 100$  m edge effect depth. Note that the global results are not the summation of the individual biomes, but the result of a separate analysis including non-forest biomes.

Feature	Unit	Boreal	Temperate	Tropical	Global
Num. of Fragments	$10^6$	122.97	73.78	83.76	409.22
Land surface area	$10^6$ ha	1512.10	1691.34	2349.38	47174.12
Forested area	$10^6$ ha	710.49	679.33	1452.38	3300.24
Edge length	$10^6$ km	60.02	51.65	51.58	221.20
Edge area	$10^6$ ha	317.37	304.00	294.85	1177.26
Edge area / total forested area	%	44.7	44.7	20.3	35.7

## 4.3 Results

### 4.3.1 Global analysis of forest fragment structure for the year 2000

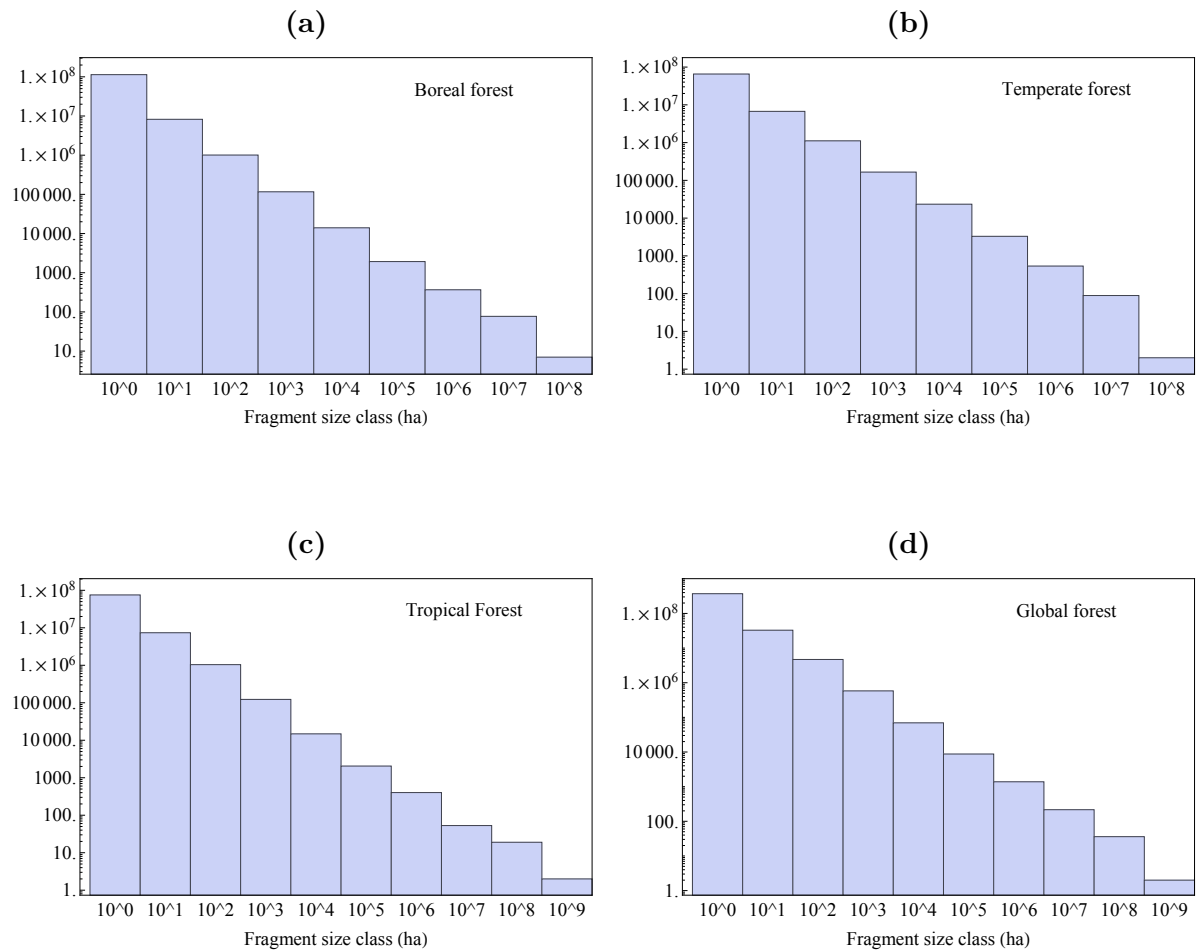
For the forest fragment analysis we analysed forest cover at global scale and separately for the three forest biomes (temperate, boreal and tropical). For every fragment we calculated fragment size, fragment area, fragment edge area and the shape index (see Section 4.2.5).

A summary of the analysis results is shown in Table 4.1. Total forested area is estimated to be  $3300.24 \times 10^6$  ha which matches the result from other studies (Hansen et al., 2010). The largest proportion of forested area lies within the tropical biome (62% of its surface area), while its fragmentation condition is the best among the other biomes, indicated by the relation of edge area to total forested area. 35.7% of the global forest area lies within 100 m of forest edges (20.3% for tropical, 44.7% for temperate and 44.7% for boreal forests). An additional analysis using a different threshold for vegetation cover is shown in the Appendix (Table C.1).

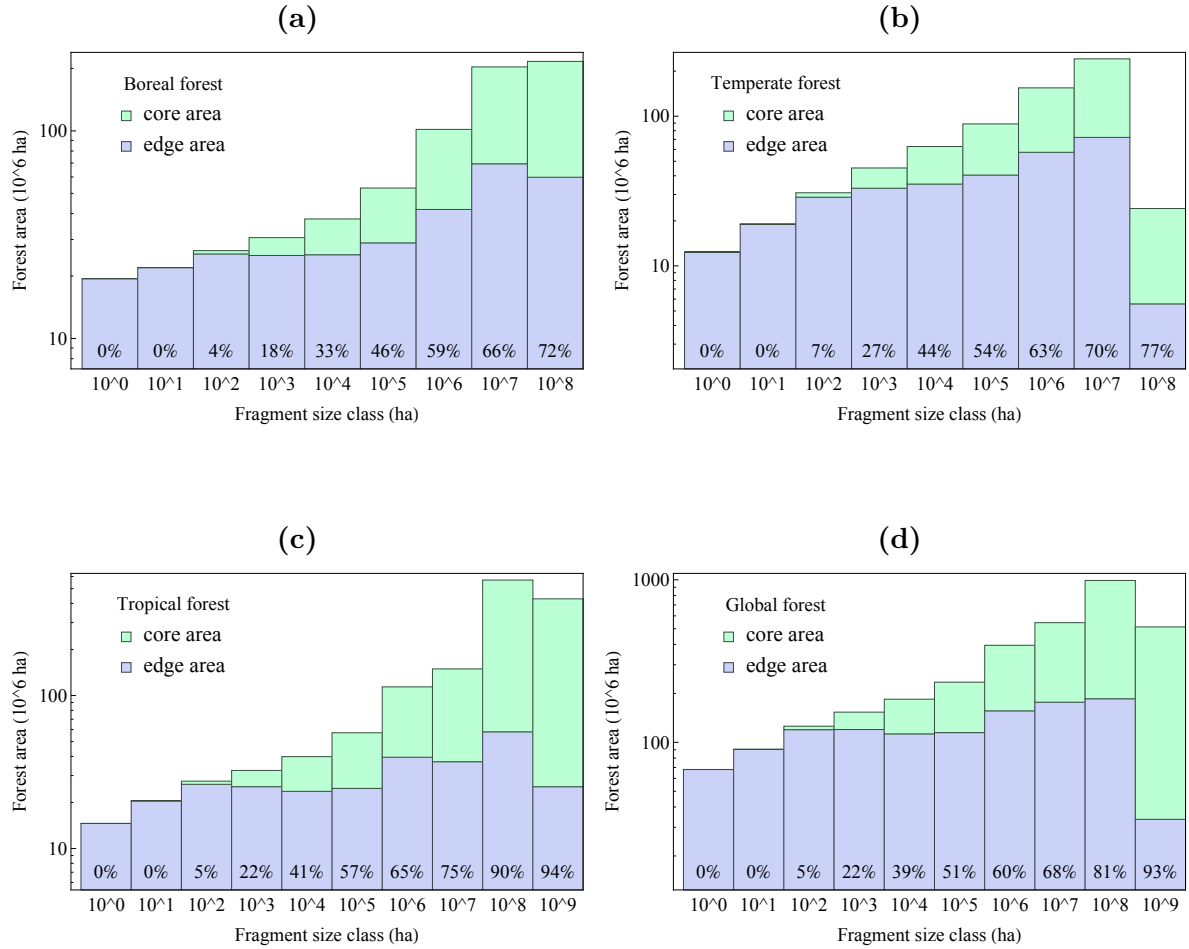
**Table 4.2:** Summary of power-law fits to the a) fragment size distribution and (Fig. 4.1) and b) shape index distribution (see also Fig. 4.3) of different forest biomes. Shown is the estimated scaling exponent  $\alpha$ , its approximative 95% confidence interval and p-value using a two-sample Kolmogorov-Smirnov test. All values indicate strong evidence for the assumption of a power-law distribution.

(a)				(b)			
Biome	$\alpha$	95% CI	p-value	Biome	$-\alpha$	95% CI	p-value
Global	1.86	[1.86, 1.86]	1	Global	1.35	[1.34, 1.35]	1
Tropical	1.87	[1.87, 1.87]	1	Tropical	1.30	[1.29, 1.30]	1
Temperate	1.80	[1.80, 1.80]	1	Temperate	1.33	[1.32, 1.33]	1
Boreal	1.92	[1.92, 1.92]	1	Boreal	1.39	[1.39, 1.40]	1

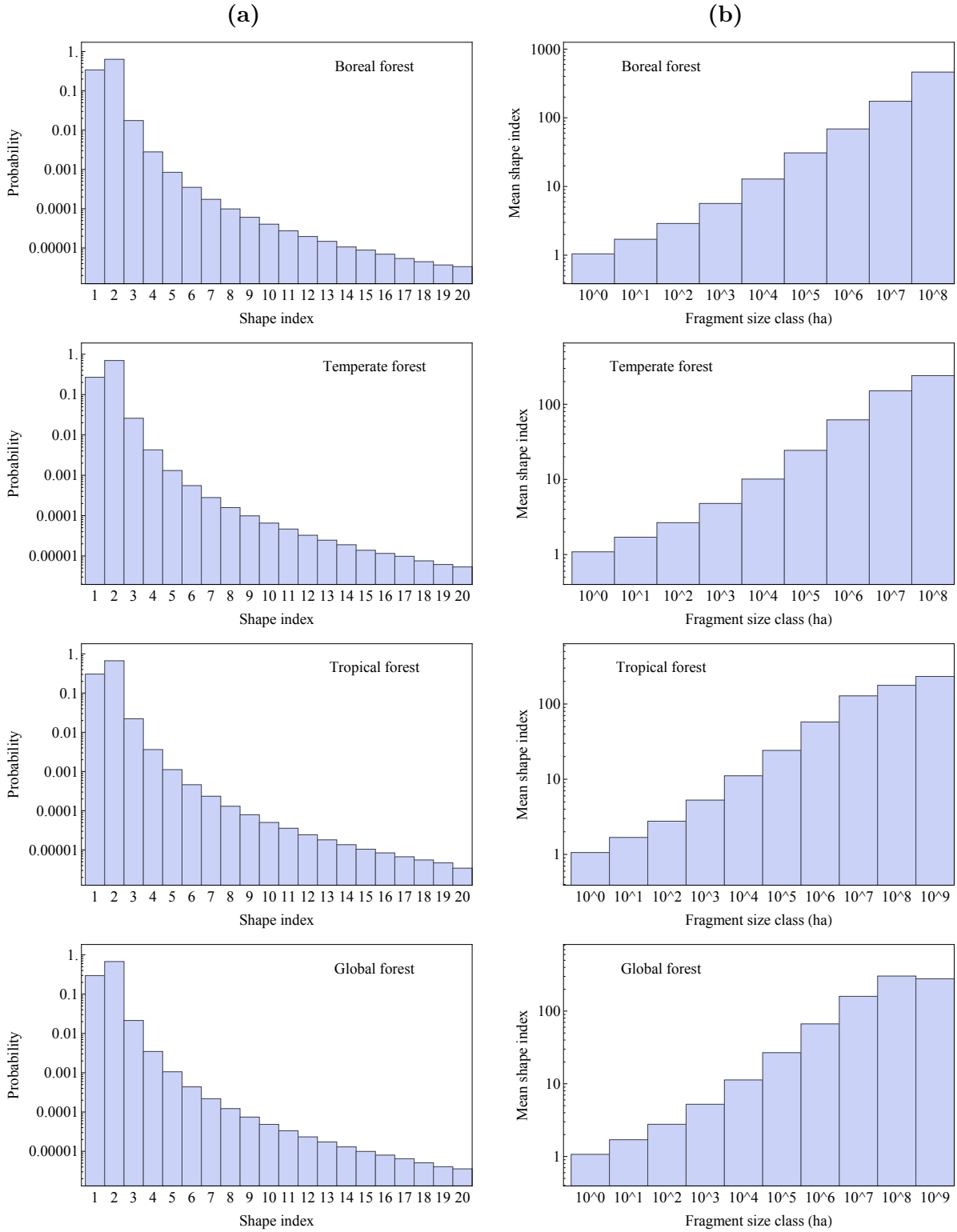




**Figure 4.1:** Number of forest fragments for different size classes on a log-log scale, classified by forest biome. (a) boreal, (b) temperate, (c) tropical, (d) global



**Figure 4.2:** Core and edge area for different fragment size classes on a log-log scale classified by forest biome where (a) boreal, (b) temperate, (c) tropical, (d) global. Percentages show the portion of core area within the fragment size class.



**Figure 4.3:** Analysis of forest fragments (a) Probability distribution of the shape index (values  $> 20$  are omitted) and (b) mean shape index in relation to fragment size classes (shape index distribution) on a log-scale, for each biome and global forests.

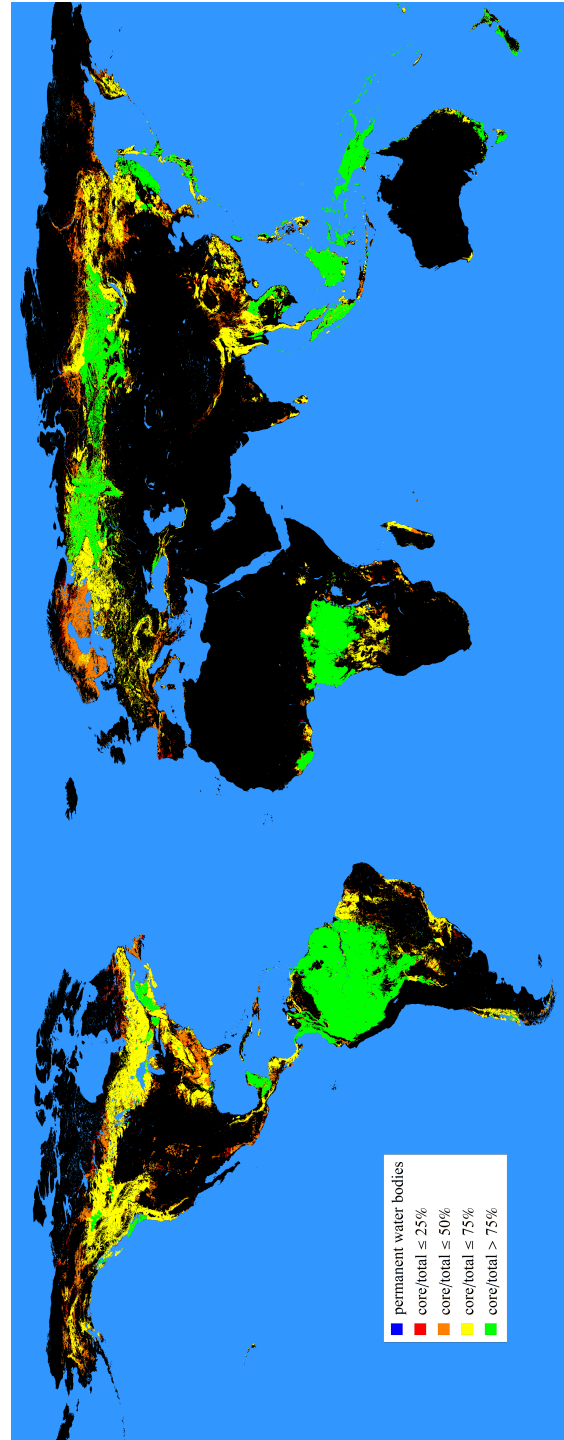
Fig. 4.1 displays the derived fragment size distribution for the boreal, temperate and tropical biome. The number of fragments shows a similar steep decline with increasing fragment size. A fitted power law is in good agreement with the data, while the scaling exponent  $\alpha$  lies in all biomes between 1.8 and 1.9 (see Table 4.2a). Fig. 4.2 shows how forest area is distributed over different fragment size classes for the boreal, temperate and tropical biome. Total forested area is increasing with fragment size class for each biome, except for the given largest fragment size class in temperate and tropical forests. With increasing fragment size an increasing portion of its area is estimated to be core area, while for small forest fragments almost the whole area corresponds to edge area. The largest fragment size class ( $10^8 - 10^9$  ha) reaches a core area of 72% , 77% and 94% in the boreal, temperate and tropical forests. A better insight into the structure of forest fragmentation is given by the distribution of the shape index in Fig. 4.3. More than 97% of the global forest fragments have a shape-index  $< 2$  and thereby a more circular shape. The circular shape is dominant for small fragments. Fragments larger in size also increase in shape complexity. Nearly 30% of the global forest fragments have a shape index  $< 1$ . This is plausible because an edge of a forest fragment has to have a minimum distance to another edge to contribute to the edge-length according to our method (Section 4.2.4). This indicates a large amount of small non-isolated fragments across the landscape. A fitted power law is in good agreement with the mean shape index distributions, while the scaling exponent  $\alpha$  is in all biomes around  $-1.3$  (see Table 4.2b).

### 4.3.2 Global analysis of forest fragment condition for the year 2000

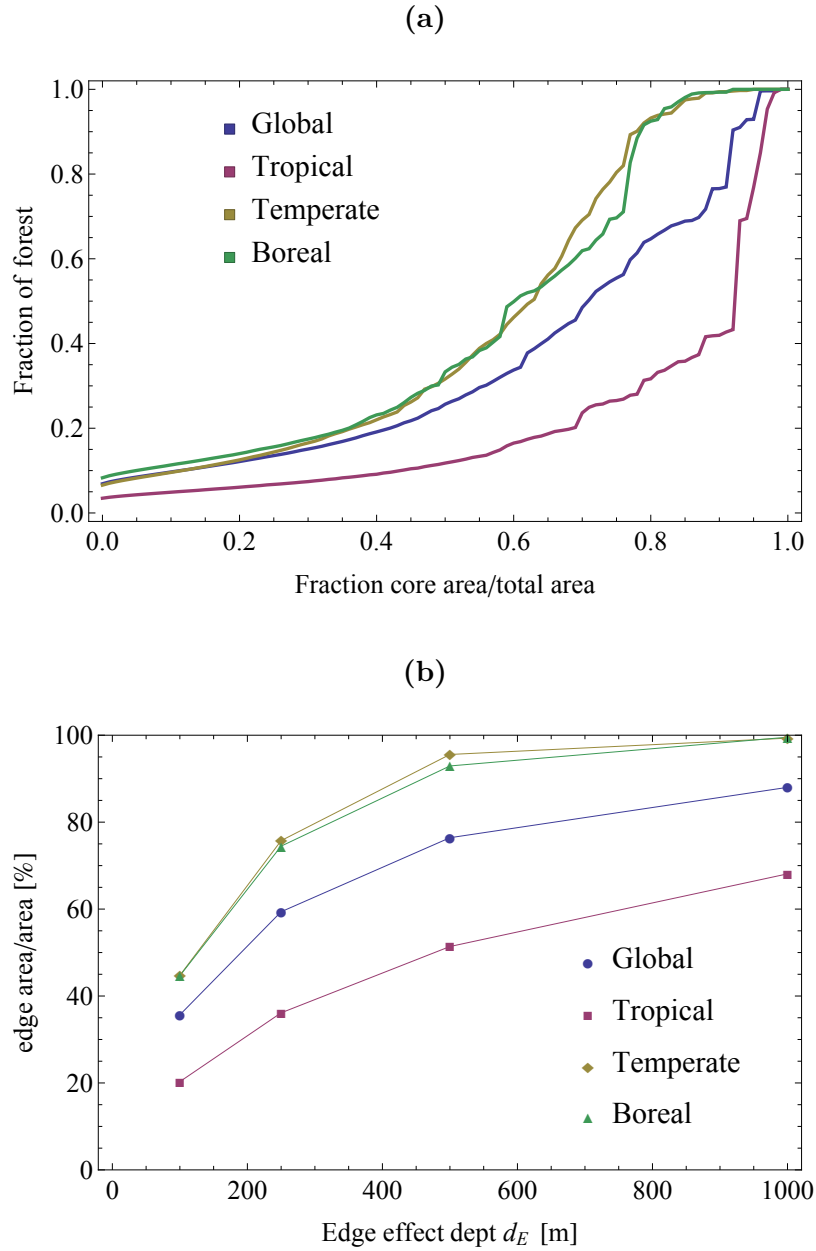
The map (construction details in Section 4.2.6) shown in Fig. 4.4 gives a simplified overview of the condition of forest fragments across the world, while a more detailed summary is shown in Table 4.3. Within the tropical biome most of the fragments are in very good conditions with a small amount of edge-affected area (73.5% of total forested area). Only 6.7% of the forested area in the tropics is in poor condition. For temperate forests, an area subject to strong anthropogenic influences, 69.1% of total forested area is in good or very good condition (Table 4.3). However, the fraction of forest area in poor condition is with 14.2% more than twice as high as in the tropics. This is also the case for the boreal biome, where 15.4% of total forested area is in poor condition and 69.4% is in good or very good condition, with larger unaffected areas in Russia's boreal forest.

For a detailed view of forest condition we derived a cumulative distribution (Fig. 4.5a), where the fraction of core area to total area of each forest fragment is binned and its area accumulated in relation to total forested area. For example, the probability that the area of a tropical forest fragment has a core area  $\leq 80\%$  (20% of its area lies within a 100 m edge) is about 25%, while the probability that the area of a boreal forest fragment has a core area  $\leq 80\%$  is about 95%. The fragment state distributions for temperate and boreal forests are rather similar and are always above the distribution for tropical forests. This illustrates the results from Table 4.3 that forest fragments in the tropical biome are in a general better state than in the temperate or boreal biome.

The proportion of forest area that lies within different distances to the forest edges



**Figure 4.4:** Global map showing a 4-colour classification of forest fragments depending on the relation of core to total fragment area. The shown map is a resampled version of the original 30-m map (factor of 1:250).



**Figure 4.5:** Condition of forest fragments globally and classified by forest biome. (a) Fraction of core area to total area of a forest fragment in relation to total forested area, as an index of fragment state (edge effect depth of  $d_E = 100$  m), (b) Proportion of forest area that lies within different distances to the forest edges.

**Table 4.3:** Proportion of forest fragment area in a certain condition expressed as fragment core area versus total fragment area assuming an edge effect depth of  $d_E = 100$  m.

Biome	$\xi$ (fragment core area / total area)			
	poor $\leq 25\%$	moderate $\leq 50\%$	good $\leq 75\%$	very good $> 75\%$
Global	13.4%	11.5%	30.0%	45.0%
Tropical	6.7%	4.9%	14.9%	73.5%
Temperate	14.2%	16.8%	48.7%	20.4%
Boreal	15.4%	15.1%	38.8%	30.6%

(Fig. 4.5b) emphasises these findings. For the temperate and boreal biome nearly 100% of forest area lies within 1000 m of forest edges, while this holds only for 68% of tropical forest area.

## 4.4 Discussion

Here we have presented a global fragment analysis based on high-resolution forest cover maps. This analysis was possible by using a new methodological framework including an advanced cluster analysis algorithm combined with a unique data compression. The practicability of the method was demonstrated through the analysis of a global 30 m forest cover map (Hansen et al., 2013). Thereby it is possible to conduct global forest fragment analysis for edge detection, connectivity, thresholding and intersection with general maps (e.g. elevation or fire data) in a short time frame on a desktop computer. Unlike previous global studies on the structure of forest fragments (Wade et al., 2003; Riitters et al., 2000, 2016; Haddad et al., 2015) the method presented here is able to explicitly capture the individual properties of each single fragment (area, circumference, edge-area etc.), independent of the fragment size.

The calculation of edge-affected area is an important aspect of the fragment analysis, which includes the detection and evaluation of forest edges. As forest fragmentation can occur on small spatial scales ( $< 100$  m) our study benefits from the high (30 m) resolution of the tree-cover map. Future studies could include more sophisticated methods for fragment and edge detection to avoid cases where either almost isolated forest areas are counted as single fragments or small clusters of non-forest area in undisturbed forest areas contribute to the estimated forest edge length.

One of our main findings, that 35.7% of the global forest area lies within 100 m of forest edges, highlights the importance of current observation and conservation efforts, because the increased fraction of 44.7% in temperate and 44.7% in boreal forests in comparison with 20.3% for tropical forests is an indicator for anthropogenic impacts on the structure of tropical forests. For better comparability, we additionally estimated that 76% of global forest area lies within 500 m of forest edge, a value higher than the estimates from previous studies using different methods (60% within 700 m (Haddad et al., 2015) and 62% within 700 m (Riitters et al., 2016)).

While the forest fragments core and edge area is heterogeneously distributed across the different forest biomes the fragment size distribution and the shape index distribution have a similar shape, where each can be described through power-laws. This could be an indicator for a dominant underlying ecological process, which shapes the structure of forest fragmentation.

## 4.5 Conclusion

The here presented approach to quantify global forest fragmentation is an important extension of existing methods. Through the combination of an advanced cluster-analysis algorithm with a unique data compression scheme it is now practicable to estimate properties of each single forest fragment on a global scale using high-resolution maps. The method has a broad range of applications not covered in this study. It can be used to quantify fragmentation of other ecosystems or additional analysis by combining it with ecological informations from other available maps (e.g. fire or elevation data).

Land-use causes fragmented landscapes all around the world. To maintain ecosystem services and improve decisions regarding biodiversity conservation, a global monitoring of forest fragmentation is important. We show that the state of forest fragmentation is very heterogeneous across different forest biomes with a continuous need for conservation and restoration. For this reason our developed high-resolution map of the global state of forest fragmentation is a helpful tool for future decisions. In conclusion, our analysis highlights the importance of forest fragmentation in forest ecology and conservation biology.





# Chapter 5

## Discussion and perspectives

---

In this thesis we have shown how innovative modelling methods help to improve upon the existing knowledge of specific aspects of natural forests. At the core of each chapter a modelling framework was developed. These frameworks, implemented in separate C++ programs (totalling over 30 000 lines of code), were already used in other studies and can be applied to future ecological questions due to their general formulation. In the following, we will address what has been archived, what are possible extension to the frameworks, and what are future research directions.

### 5.1 Model calibration and uncertainty assessment

We have shown in Chapter 2 that simple methods of stochastic optimization are able to properly identify parameters for a size structured model of tropical rain forest. These methods are thus a practical alternative for computational more demanding methods based on Bayesian calibration. Stochastic search methods do not rely on the ability to calculate first and/or second order derivatives and are not prone to slow convergence or trapping into local minima, if the objective function is non-smooth or otherwise pathological (Hoos and Stützle, 2004). We emphasised the importance of using a set of diagnostic tools for inspecting parameter uncertainty, general identifiability and appropriate model complexity. Our most important finding was that the minimum amount of observation data needed to calibrate the forest model are total stem number, basal area and biomass. That means, these three properties of forest structure carry enough information to make a plausible estimate of the underlying stem size distribution. The framework can thus be applied at forest sites where only a limited amount of observations is available. As the developed framework is independent of the model under investigation, it has already been applied in other studies (May et al., 2015; Rödiger et al., 2017).

Theoretically, the number of model parameters that can be calibrated using this framework is unlimited. For the dynamic vegetation model from Chapter 2, three parameters per species (nine in total) were calibrated, May et al. (2015) calibrated six parameters and for the neutral model from Chapter 3, 12 parameters were calibrated. With increasing number of parameters one usually increases the number of realized model evaluations to make accurate estimates. We used 10 000 times the number of parameters which have to be calibrated as a rule of thumb, although one of the used calibration methods (Dy-

namically Dimensioned Search, Tolson and Shoemaker (2007b)) was already applied to a watershed problem with 30 parameters using only 2000 model evaluations. This demonstrates that the calibration methods are applicable to high dimensional calibration problems using a small amount of model evaluations, for example, a dynamic vegetation model like FORMIND Köhler and Huth (1998); Fischer et al. (2016), which uses about 50 parameters.

If the available data to calibrate the model does not contain enough information to distinguish all processes, multiple or even an unlimited number of solutions to the calibration problem can exist. That means in this case (some) parameter estimates are highly uncertain. Reducing this uncertainty further is a question of data availability on specific scales. The usage of remote sensing data in combination with local field measurements offers the possibility to infer parameters of previously imprecisely described processes.

We therefore advocate a future extension of the framework through a procedure which automatically identifies unidentifiable parameters and what patterns contain information for which model processes. As this can be a time consuming task one possibility is to use computational inexpensive meta modelling techniques. In a meta modelling framework the simulation model is only executed at specific design positions (e.g. based on Latin Hypercube sampling, McKay et al. (2000)) in the parameter space. The model response between the design points is then interpolated. This approach would provide the following improvements to the existing framework: the usage of advanced stochastic search methods for calibration based on radial basis functions and kriging methods (Regis and Shoemaker, 2005; Gu et al., 2012) and fast execution of MCMC methods using Gaussian processes (Conti et al., 2009; Fricker et al., 2013). The estimation of parameters and processes, which are identifiable through available observations, is then a task of automatic and fast evaluation of the meta model. Although experiments based on these methods show promising results, their effectiveness depends on the complexity (non-pathological, smoothness) of the underlying model response.

In addition to reducing uncertainty of parameter estimators, quantifying uncertainty in processes and parameters is also an important aspect of future extensions. As we have only used linear approximation and error propagation at the estimated parameter optimum, this played only a secondary role in our study. The quantification of uncertainty should include the evaluation of structural model uncertainty, input parameter uncertainty and measurement uncertainty of observations. A task which is hard to solve satisfactorily. Recent developments in hydrology using sequential data assimilation and particle filtering show promising results regarding the evaluation of different aspects of uncertainty (e.g. Moradkhani et al. (2005) or Liu and Gupta (2007) for a review of methods). These methods work in the context of state-space models, which can generate estimates of unobserved variables. A state-space model typically consists of two equations: a measurement equation which links observed to unobserved state variables and a transition equation which describes the dynamics of the state variables (Luo et al., 2013).

## 5.2 Understanding the mechanisms that shape biodiversity

We have shown in Chapter 3 that a neutral model, where size-growth and mortality are directly connected through competition for space, is able to explain various spatial, non-spatial and size-related diversity patterns in a tropical forest. This has been achieved by a careful extension of the existing theory. In addition to the incorporation of plant size growth, the metacommunity is represented through a static distribution to avoid variations in the regional species distribution. And a variable number of individuals in the local community reflects varying resource availability.

We evaluated the predictive power of the neutral model using the inverse modelling framework from Chapter 2 and extended the framework through a qualitative global sensitivity analysis based on the elementary effects methods (Morris, 1991). This allowed the additional evaluation of uncertainties, correlations, sensitivities, nonlinear and interactive effects among parameters and processes. Albeit its simplicity, the model was able to explain the main dynamics of six biodiversity patterns including the species abundance distribution, the species-area relationship and the tree size distribution. However, the model was not able to explain the number of species for different tree size classes precisely. We found that there possibly exist missing or misspecified mechanisms. This results in a local species richness which is slightly too diverse and an effective tree size growth which is too low. This emphasizes the importance of confronting a model with several patterns at once, as each pattern carries varying aspects of the underlying ecological processes (Wiegand et al., 2003; Grimm et al., 2005).

The combination of fast model calibration and global sensitivity analysis can be used to test future hypotheses about ecological communities and vegetation dynamics. This includes advanced dispersal mechanisms, as the estimated dispersal parameters were lower than expected. The difference of some patterns within our model points to possible non-neutral processes. Future work could therefore rely on the studies of Uriarte et al. (2010) and Fortunel et al. (2016) to test the impact of species-specific neighbourhood processes on size growth and mortality. Trait differences expressed as fitness differences between species can be incorporated through species-specific competition coefficients. The relation of niche differences (degree on which intraspecific competition exceeds interspecific competition) in relation to fitness differences between species emerges new patterns through which conditions for coexistence can be analysed (Kraft et al., 2015).

## 5.3 Analysis of high-resolution forest cover maps

We have presented in Chapter 4 a forest fragment analysis on a global scale using high-resolution forest cover maps. This was possible using a new methodological framework, where an advanced cluster analysis algorithm is combined with a specialized data compression method. The feasibility of the framework was demonstrated through the analysis of a global forest cover map with a resolution of 30 m (Hansen et al., 2013).

In contrast to previous studies (Wade et al., 2003; Riitters et al., 2000, 2016; Haddad

et al., 2015) the developed method is able to explicitly capture individual properties of each single forest fragment, e.g. area, circumference or edge-area. One of our main findings is that 35.7% of global forest area lies within 100 m of forest edges. On a biome level, tropical forests show the best condition with 20.3%. This value is higher in temperate (44.7%) and boreal (44.7%) forest. That could be an indicator for anthropogenic impacts on the structure of forests in these areas. Additionally, we found that the forest fragments core and edge area show different distributions across the three forest biomes, while the fragment size and shape index frequency distributions have a similar shape across the biomes. These two distributions can be described through power-laws, which could be an indicator for a dominant underlying ecological process.

The present framework for forest fragment analysis has a broad range of future applications. It was already applied for the estimation of carbon loss in tropical forests through deforestation (Brinck et al., 2017). This was archived through the additional usage of three pan-tropical biomass maps (Saatchi et al., 2007; Baccini et al., 2012; Avitabile et al., 2016), where we calculated the biomass of each single forest fragment and estimated the expected carbon-loss through edge effects.

To analyse how the structure and state of forest fragmentation changes with environmental factors, elevation, fire or climate data maps could be used for further classification. The framework would be also applicable to time-series of global forest cover maps to analyse how forest fragmentation changes with time. This task needs the accurate estimation of forest gain and losses depending on the spatial origin within a fragment (edge/core). The temporal analysis of forest fragmentation could be based on newly available global PALSAR forest/non-forest maps with a resolution of 25 m (Shimada et al., 2014).

The framework could be extended in several ways to better understand the main ecological drivers that affect forest deforestation. This includes estimates of additional metrics to characterize forest fragment structure, which improves the understanding of the effects of fragmentation. Current metrics make estimates on an area, edge and form level. Patch metrics like the proximity index using neighbourhood analysis can be used to characterize the connectivity of the landscape (Rutledge, 2003). The proximity index is inspired by island biogeography and quantifies the spatial context of a focal patch in relation to its neighbours. The index has a large value when the focal patch is surrounded by larger and/or closer patches and decreases as patches become smaller and/or more sparse (Gustafson and Parker, 1994). Because the calculation of the proximity index depends on neighbourhood distances, it is highly demanding in computing time for high-resolution maps.

In addition to the estimation of the scaling exponent of the fragment size distribution in Chapter 4, other spatial scaling characteristics could be estimated. This enables the linkage with percolation theory, which makes predictions about several critical exponents at a phase transition (Stauffer and Aharony, 1994). One of such properties is the fractal dimension, which is a measure of geometrical structure and describes how the detail of a pattern changes with scale (Mandelbrot, 1967).

## 5.4 Outlook

The outlook of this thesis is the combination of the here presented methods into a large scale modelling and analysis framework for natural forests. The basis of such a project is the application of an extended neutral model from Chapter 3 on a regional or continental scale (e.g. the Amazon or Congo basin). In the extended model tree size growth and local species composition are not only depending on competition effects but also on several other environmental conditions like light availability, soil properties, terrain and climate conditions. The model dynamics should replicate observational dynamics on spatial and also temporal scales. As data availability from field plots on such a scale is typically sparse, multiple data sources have to be combined. From the available data, including remote-sensing data, several patterns and aggregated statistics have to be extracted. These patterns should describe fundamental ecological processes across scales including spatial and non-spatial patterns of diversity. In addition to the usage of directly extracted patterns from remote sensing using the framework from Chapter 4, we advocate the implementation of an optimized 3D model for the simulation of radiative transfer (e.g. DART, Gastellu-Etchegorry et al. (2004)). With such a model it could be possible to estimate structural forest attributes along environmental gradients from spectral remote sensing data. Calibration, validation and the process analysis of such a complex model can then be performed with the already discussed extensions of the calibration framework from Chapter 2.

An important application of this modelling framework is the quantification of the impact of climate change and deforestation on species diversity on a regional or global scale.



# A Appendix for Chapter 2

---

## A.1 Uncertainty assessment

A popular and fast way way to access parameter and prediction uncertainty is through the asymptotic normality  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$  for large samples of the maximum likelihood estimator, where the covariance matrix  $I^{-1}$  is the inverse of the information matrix. We approximate the covariance matrix through a first-order taylor expansion around  $\hat{\theta}$  using finite differences:

$$I^{-1} \approx \sigma^2(J^T W J)^{-1} =: X \quad (\text{A.1})$$

where  $J$  is the Jacobian of  $m(\hat{\theta})$ . An unbiased estimator for  $\sigma^2$  is the residual variance

$$\hat{\sigma}^2 = \frac{1}{n-k} Q(\hat{\theta}) \quad (\text{A.2})$$

The corresponding parameter correlation matrix  $C$  can be calculated through  $C = DXD$ , where  $D = (I_n \circ X)^{-1}$  with  $I_n$  the identity matrix of dimension  $n$  and  $\circ$  the Hadamard product.

An approximative confidence interval at the significance level  $\alpha$  can then be constructed through

$$\hat{\theta}_i \pm t_{n-p, 1-\frac{\alpha}{2}} \sqrt{X_{i,i}}, \quad (\text{A.3})$$

where  $t_{n-p, 1-\frac{\alpha}{2}}$  is the inverse CDF of a Student  $t$  distribution with  $n-p$  degrees of freedom at level  $1-\alpha/2$ . Confidence intervals for the prediction  $m(\hat{x})$  can be constructed through linear error propagation

$$Y := JXJ^T \quad (\text{A.4})$$

$$m(\hat{\theta})_i \pm t_{n-p, 1-\frac{\alpha}{2}} \sqrt{Y_{i,i}} \quad (\text{A.5})$$

If the sample size  $n$  is small or the confidence region is non-quadratic, a profile based approach is better suited (Meyer and Hill, 1992), but can be expensive, if the simulation evaluations are costly. For profiling the parameter vector  $\theta = \theta_1 \cup \theta_2$  gets partitioned into the parameter set  $\theta_1$  of interest and the parameter set  $\theta_2$  of nuisance. The profile likelihood  $L_p$  of  $\theta_1$  is then  $L_p(\theta_1) = L(\theta_1, \hat{\theta}_2)$ , where  $\hat{\theta}_2$  is the maximum likelihood estimate of  $\theta_2$  for given  $\theta_1$ .



## A.2 Identifiability

Proper identifiability is an important yet often disregarded problem in parameter estimation. But it is a property a model must satisfy for correct inference. In this context (unique) identifiability is denoted by a single global minimum of the objective function  $Q$ .

Identifiability consists of two main parts which can be demonstrated at the following simple population model:

$$\dot{n}(t) = \alpha n(t) - \beta n(t) - \gamma \quad \text{with } \gamma \ll (\alpha - \beta) \quad (\text{A.6})$$

The two parameters  $\alpha$  and  $\beta$  are structurally unidentifiable regardless of the quality of observation data and is thus a model inherent property. On the other hand the practical identifiability of  $\gamma$  depends on the accurateness of observation data.

There are different ways of inspecting identifiability:

- Run multiple minimization tasks from different starting points and check whether  $Q(\theta_1) = Q(\theta_2)$  for  $\theta_1 \neq \theta_2$
- Detect structural unidentifiability from model equations (see Section 2.3.1 for an example)
- Approximative singularity of the parameter covariance matrix  $X$  and detection of strong pairwise linear dependence through high off-diagonal absolute elements in the correlation matrix  $C$

More systematic approaches include the usage of profile likelihoods (Raue et al., 2009) and the detection of the most sensitive subset of identifiable parameters (Brun et al., 2001).

## A.3 Model selection

Selecting a model among a set of candidate models is an important task in science. Usually the selection criterion is based on the ability of explaining a fixed set of observations (goodness of fit). To not overemphasize complex models, as this may lead to overfitting, a trade off between the goodness of fit and the number of free parameters should be made. This is for example achieved by the Akaike information criterion (Akaike, 1974), which is defined as

$$\text{AIC}_i = 2k_i - 2 \log L_i(\hat{\theta}|y) \quad (\text{A.7})$$

where  $k_i$  is the number of parameters and  $L_i(\hat{\theta}|y)$  is the (maximized) likelihood of the  $i$ -th model for  $p$  candidate models. For small samples (rule of thumb  $n/k < 40$ ) a bias-corrected version AICc of the AIC should be used (Burnham and Anderson, 2002):

$$\text{AICc}_i = \text{AIC}_i + \frac{2k_i(k_i + 1)}{n - k_i - 1} \quad (\text{A.8})$$

The AIC value is only interpretable in comparison with different models, where the preferred model is the one with the lowest AIC value. A discrimination among models can be attained by using Akaike-weights  $w_i$ , where

$$\Delta_i = \text{AIC}_i - \text{AIC}_{\min} \quad \text{and} \quad w_i = \frac{\exp(-0.5\Delta_i)}{\sum_{k=1}^p \exp(-0.5\Delta_k)} \quad (\text{A.9})$$

The weights can be used to make multimodel inference by building a weighted average model and interpreted as the probability of model  $i$  being the best model for the given data set.



## B Appendix for Chapter 3

---

### B.1 Parameter screening

#### B.1.1 The morris method for global parameter screening

The morris method for global parameter screening consists of randomised local one-at-a-time (OTA) experiments, where each model input parameter  $x_i$  is assumed to vary over  $p$  levels or  $k$  parameters. The region of experimentation  $\Omega$  is thus a  $k$ -dimensional  $p$ -level grid. Let  $x_i^+ = (x_1, \dots, x_i + \Delta, \dots, x_k)$ , then the elementary effect  $\phi_{i,j}$  of the parameter  $i$  on pattern  $j$  is calculated as:

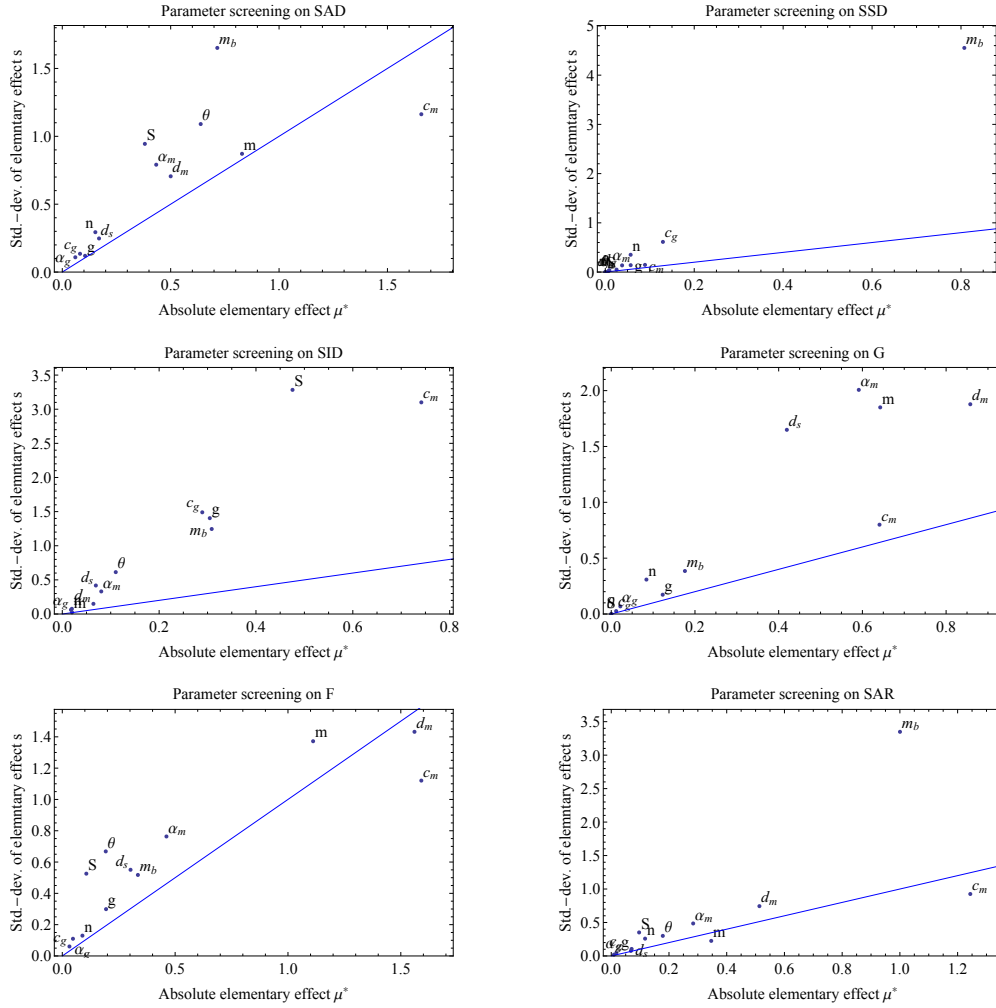
$$\phi_{i,j} = \frac{\text{MRE}(S_{\text{obs}}^j, S_{\text{sim}}^j(x_i^+)) - \text{MRE}(S_{\text{obs}}^i, S_{\text{sim}}^i(x))}{\Delta} \quad (\text{B.1})$$

where  $\Delta$  is a value in  $\{1/(p-1), \dots, 1 - 1/(p-1)\}$ . A convenient choice for the parameters is an even number for  $p$  and  $\Delta = p/(2(p-1))$ . The random sampling of different  $x$  creates a finite distribution  $\phi_{i,j} \sim F_{i,j}$  of elementary effects for each parameter on each output pattern, where the sensitivity measures  $\mu_{i,j}$  and  $\sigma_{i,j}$  are the mean and standard deviation of the distribution  $F_{i,j}$ .

Morris (1991) suggests sampling  $r$  elementary effects from each  $F_{i,j}$  via an efficient design that construct  $r$  trajectories of  $k+1$  points. Each trajectory provides  $k$  elementary effects per parameter and thus the total number of experiments is  $r(k+1)$ . For our analysis we choose  $p = 10$  levels and  $r = 50$  trajectories, which results in 650 model evaluations.

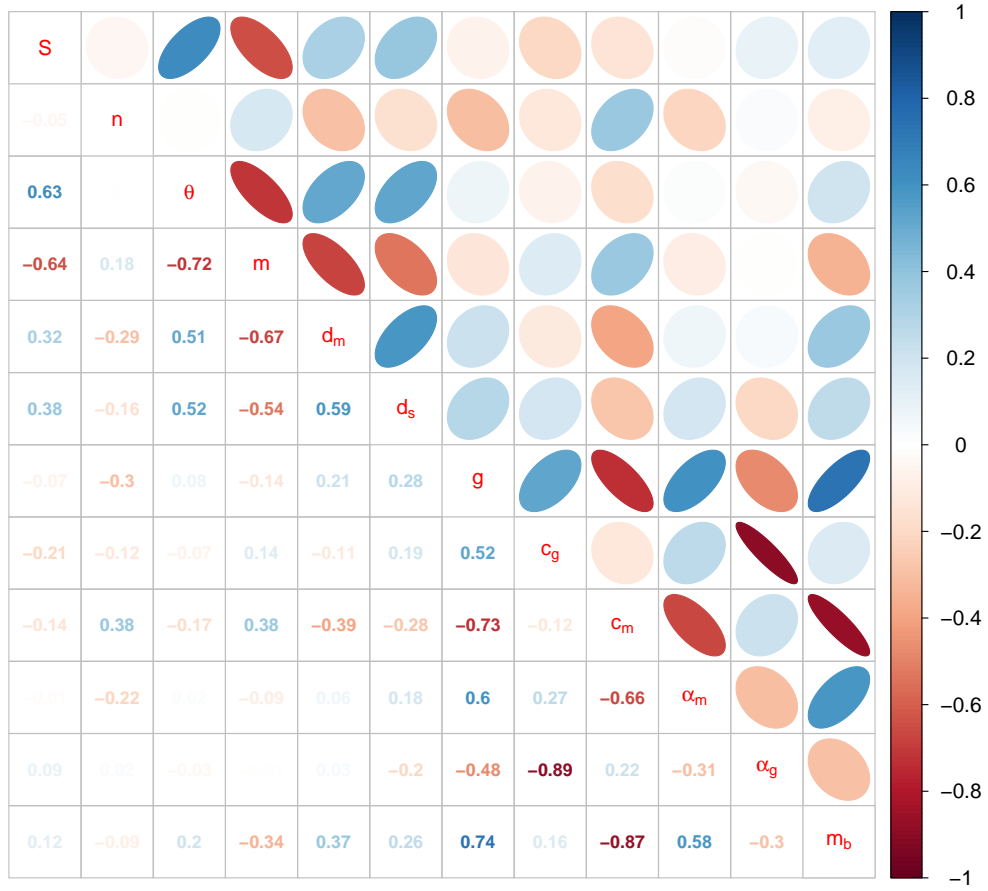
We use recent developments of the method (Campolongo et al., 2007). This includes an improved sampling strategy by maximizing the spread between the trajectories and thus maximizing the coverage of the parameter space. Additionally, we use the absolute mean  $\mu_{i,j}^*$  of  $F_{i,j}$  instead of  $\mu_{i,j}$  to estimate the overall influence of a factor. The measure  $\mu^*$  is on its own satisfactory to provide a reliable ranking in comparison with computational more demanding variance based methods, although we loose the information on the sign of the effect.

#### B.1.2 Individual screening results



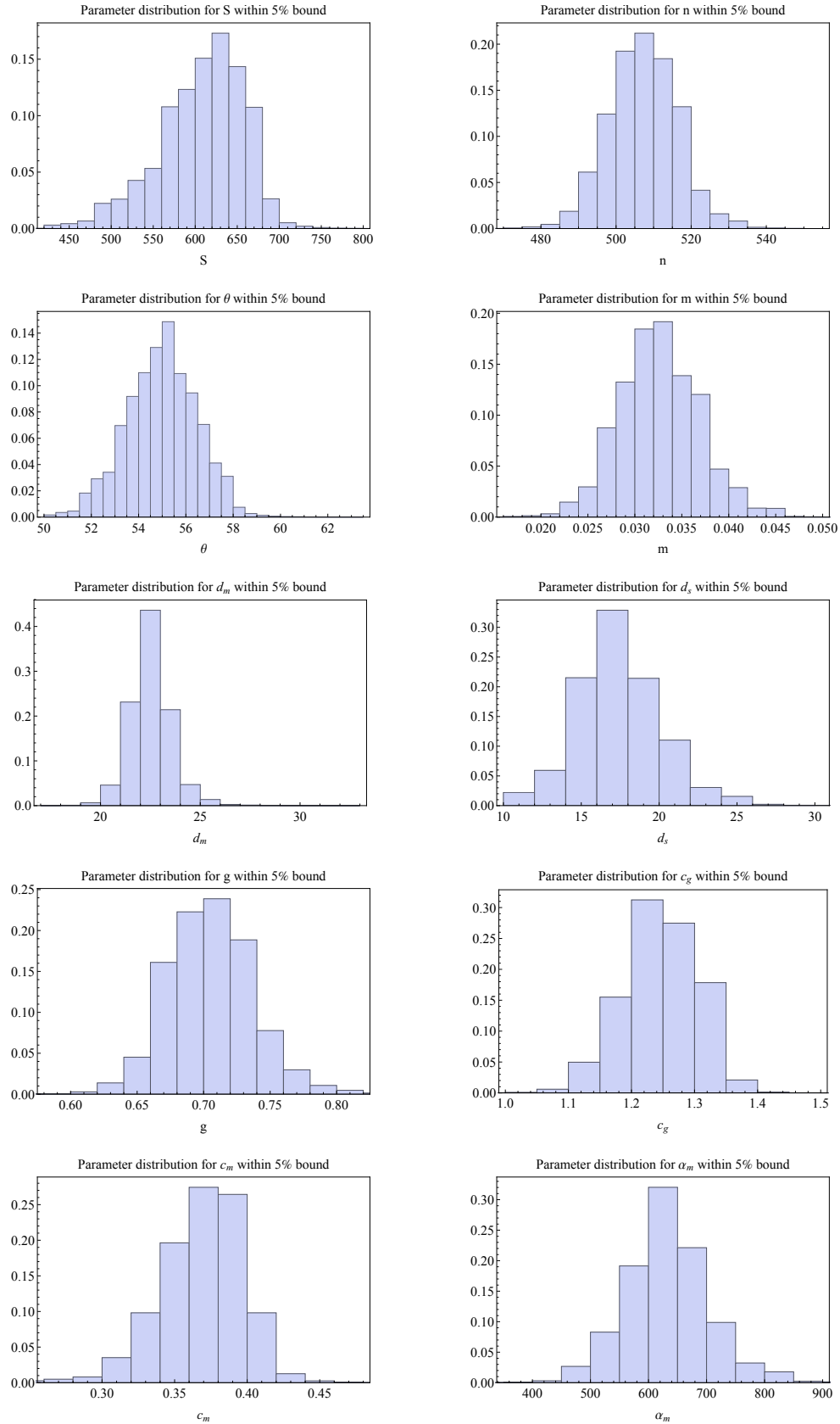
**Figure B.1:** Results for global parameter screening on each individual pattern. Points above the boundary line show nonlinear behaviour and/or parameter interaction effects.

## B.2 Calibration



**Figure B.2:** Correlation matrix for parameter sets where the objective function is within 5% of the best parameters found.

## B Appendix for Chapter 3



**Figure B.3:** Parameter distribution for parameter sets where the objective function is within 5% of the best parameters found.

# C Appendix for Chapter 4

---

## C.1 Data structures of the cluster analysis algorithm

The use of intelligent data structure is crucial for the efficiency of the cluster analysis algorithm. For this a label array 'csize' representing a set of linked lists and a cluster-data array 'cdata' is used. The index into the 'csize' and 'cdata' arrays indicates the label. A positive value in the 'csize' array indicates cluster size in pixels and negative values redirection to previous nodes in the path to the root. The 'cdata' array is used to accumulate any additional structural information in relation to a certain label. Therefore, after the complete first-pass, every index, whose value of the 'csize' array is positive, represents a cluster with additional information about its structure in the 'cdata' array for the same index.

Throughout the analysis the two arrays occupy more and more memory as labels are added to them. If only a single pass analysis is needed and system memory is limited, one can flush all roots to a temporary file which are not linked by nodes from the actual row of the forest map. This allows the algorithm to use only a fixed amount of memory for arbitrary raster map sizes.

## C.2 Compression method used on labelled images

As a prior analysis showed, a labelled forest map consists of runs (continuous sequence of the same value) of zeros (no forest) and runs of a specific label, where labels tend to be similar (low absolute difference between adjacent labels) along rows. Additionally, it should be possible to reach a certain row rapidly, without the need to decompress the previous rows, which implies that the algorithm only considers cells on a single row independent of other rows. Therefore we focus on a method that is simple to compute using run length (RLE) and variable length encoding (VLE), because more general compression methods like ZIP, using a variant of the LZ77 algorithm (Ziv and Lempel, 1977), are not able to capture these special data correlations.

We use gamma codes for encoding integers (Elias, 1975), where a number  $x \geq 1$  is coded as  $N = \lfloor \log_2(x) \rfloor$  zero bits, followed by a one bit and followed by the remaining  $N$  bits of  $x$ . This ensures the prefix property of the codes. Thus, the binary representation of  $x$  uses  $2\lfloor \log_2(x) \rfloor + 1$  bits. The usage of gamma codes implies a certain probability distribution that should roughly match the source distribution.

A single row is decomposed into a set of runs. A run is either a zero runs or a labelled run. A zero run is coded with a gamma code of '1', followed by the gamma code of the



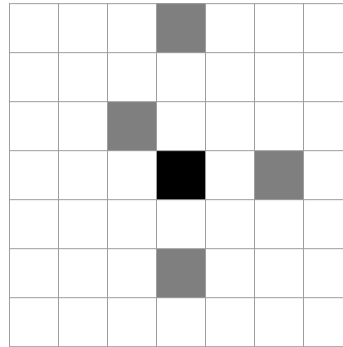
run length. A labelled run is coded with the gamma code  $x = |d| + 2$ , followed by a single bit, indicating whether  $d$  is positive or negative and followed by the gamma code of the run length. The variable  $d$  is the difference between the label of the actual run and the label of the previous labelled run, or zero if no labelled run has been seen so far.

### C.3 Summary of results for a tree cover threshold of 25%

**Table C.1:** Summary of results for the forest fragment analysis using a threshold of 25% for forest/non-forest classification and  $d_E = 100$  m edge effect depth. Note that global results are the result of a separate analysis including non-forest biomes and not the summation of the individual biomes.

Feature	Unit	Boreal	Temperate	Tropical	Global
Num. of Fragments	$10^6$	104.32	84.86	87.21	450.06
Land surface area	$10^6$ ha	1512.10	1691.34	2349.38	47174.12
Forested area	$10^6$ ha	962.48	767.10	1595.16	4165.22
Edge length	$10^6$ km	59.33	53.92	53.67	247.02
Edge area	$10^6$ ha	356.16	314.66	313.88	1363.51
Edge area / area	%	37.0	41.0	19.7	32.7

## C.4 Edge detection method



**Figure C.1:** Edge detection method within the cluster analysis algorithm for a forest fragment (shown in black) near the equator, neighboured by four forest pixels (shown in grey). The forest fragment has two edges considered as forest edge: the top edge because the distance to the next northern forest pixel is two pixels and the left edge, because there are no additional forest pixels to the west.



# List of Figures

---

2.1	Stem-diameter-distribution in equilibrium from the dynamic forest model using the virtual data set. . . . .	15
2.2	Basal area over time from the dynamic forest model using the virtual data set. . . . .	15
2.3	Convergence graph of the different optimization algorithms for calibrating the virtual data shown on a log-log scale. . . . .	17
2.4	Negative log-likelihood profile for the parameters $x_0$ and $g$ . The valley between $[0, 5]$ for $x_0$ indicates poor identifiability. . . . .	18
2.5	Prediction of the forest model against field data on a log-scale in a single species experiment using weights inversely proportional to the data variance. Filled bars indicate mean observations for diameter steam distribution, basal area and aboveground biomass. Error bars indicate the approximative 95% confidence interval for the marked prediction. . . . .	20
2.6	Biomass over time for the three species model, calibrated against stem count, basal area and biomass. Dots represent field data measurements. . . . .	22
2.7	Prediction of the equilibrium model on a log-scale in a three species experiment denoted by their plant functional type (PFT), using weights inversely proportional to the data variance and calibrated using total stem count, basal area and biomass per species. Filled bars represent mean observations and dots model predictions. . . . .	23
3.1	Based on the log-series distribution of species (a) the species rank abundance (SRA) distribution (b) is constructed as described in Section 3.2.2. The SRA distribution is used to sample species from the metacommunity, which immigrate into the local community. For this example we used $S = 200$ species and a birth-to-death ratio of $p = 0.98$ . . . . .	29
3.2	Flow-chart of model processes with a snapshot from above of a simulated model output. The snapshot shows the local community with an area of $1000 \text{ m} \cdot 500 \text{ m} = 50 \text{ ha}$ . Circles denote single trees, where the circles diameter equals the stem diameter of a tree resized by a factor of 5. . . . .	34
3.3	Visualization of the model predictions against the mean field observations for the six observed patterns: species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G). . . . .	36

3.4	Shown is the a) Species size distribution including the observation data, a fitted theoretical exponential distribution and the result from the simulation model b) Stem diameter growth of observation data, the potential stem diameter growth and the realized stem size growth of the simulation model. . . . .	37
3.5	Shown is a heat map of scaled absolute elementary effects $\mu^*$ of each parameter on each individual pattern. The patterns are the species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G). . . . .	38
3.6	Global model screening results on all parameters, patterns and processes. Points above the boundary line show nonlinear behaviour and/or interaction effects. The patterns shown are the species abundance distribution (SAD), species-area relationship (SAR), species-individual distribution (SID), stem size distribution (SSD), distance decay of similarity (F) and the pair-correlation function (G). . . . .	39
4.1	Number of forest fragments for different size classes on a log-log scale, classified by forest biome. (a) boreal, (b) temperate, (c) tropical, (d) global	50
4.2	Core and edge area for different fragment size classes on a log-log scale classified by forest biome where (a) boreal, (b) temperate, (c) tropical, (d) global. Percentages show the portion of core area within the fragment size class. . . . .	51
4.3	Analysis of forest fragments (a) Probability distribution of the shape index (values $> 20$ are omitted) and (b) mean shape index in relation to fragment size classes (shape index distribution) on a log-scale, for each biome and global forests. . . . .	52
4.4	Global map showing a 4-colour classification of forest fragments depending on the relation of core to total fragment area. The shown map is a resampled version of the original 30-m map (factor of 1:250). . . . .	54
4.5	Condition of forest fragments globally and classified by forest biome. (a) Fraction of core area to total area of a forest fragment in relation to total forested area, as an index of fragment state (edge effect depth of $d_E = 100$ m), (b) Proportion of forest area that lies within different distances to the forest edges. . . . .	55
B.1	Results for global parameter screening on each individual pattern. Points above the boundary line show nonlinear behaviour and/or parameter interaction effects. . . . .	70
B.2	Correlation matrix for parameter sets where the objective function is within 5% of the best parameters found. . . . .	71
B.3	Parameter distribution for parameter sets where the objective function is within 5% of the best parameters found. . . . .	72

C.1	Edge detection method within the cluster analysis algorithm for a forest fragment (shown in black) near the equator, neighboured by four forest pixels (shown in grey). The forest fragment has two edges considered as forest edge: the top edge because the distance to the next northern forest pixel is two pixels and the left edge, because there are no additional forest pixels to the west. . . . .	75
-----	--	----



# List of Tables

---

2.1	Median parameter estimates and log-likelihood for five independent optimization runs for the different algorithms. . . . .	18
2.2	Approximative coefficients of variation (CV), 95% confidence intervals (CI) and pairwise correlation values estimated using the true values for the model and error structure. . . . .	19
2.3	Effect of different assumptions about the error model on goodness of fit represented by the mean absolute percentage error (MAPE) and a bias corrected Aitken information criterion (AICc). Estimates on parameter uncertainty are represented by the mean coefficient of variation ( $\overline{CV}$ ) and estimates on parameter identifiability by the mean correlation coefficient ( $\overline{CC}$ ) over all parameters. . . . .	19
2.4	Effect on the AICc and corresponding weights when model complexity is reduced. As an additional measure of goodness of fit the mean absolute percentage error (MAPE) is included. . . . .	20
2.5	Effect of observation data reduction to the models predictions represented by the mean absolute percentage error (MAPE), the mean of the coefficients of variation ( $\overline{CV}$ ), the mean of the absolute pairwise correlation values ( $\overline{CC}$ ) and the mean bias of the parameter estimates in comparison to the estimates using the full data set. . . . .	21
3.1	Overview of the 12 free models parameters, their units and plausible input ranges for the chosen study site. . . . .	33
3.2	Summary of calibration results for the 12 free parameters. The last two columns show the uncertainty and correlation classification based on the scheme from Section 3.2.3. . . . .	35
3.3	Relative error of the best found parameter set against each individual pattern. . . . .	35
4.1	Summary of results for the forest fragment analysis using a threshold of 50% for forest/non-forest classification and $d_E = 100$ m edge effect depth. Note that the global results are not the summation of the individual biomes, but the result of a separate analysis including non-forest biomes. . . . .	49



4.2	Summary of power-law fits to the a) fragment size distribution and (Fig. 4.1) and b) shape index distribution (see also Fig. 4.3) of different forest biomes. Shown is the estimated scaling exponent $\alpha$ , its approximative 95% confidence interval and p-value using a two-sample Kolmogorov-Smirnov test. All values indicate strong evidence for the assumption of a power-law distribution. . . . .	49
4.3	Proportion of forest fragment area in a certain condition expressed as fragment core area versus total fragment area assuming an edge effect depth of $d_E = 100$ m. . . . .	56
C.1	Summary of results for the forest fragment analysis using a threshold of 25% for forest/non-forest classification and $d_E = 100$ m edge effect depth. Note that global results are the result of a separate analysis including non-forest biomes and not the summation of the individual biomes. . . .	74

# Bibliography

---

- Acevedo, M. F. *Simulation of ecological and environmental models*. CRC Press, 2012.
- Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- Alonso, D. and McKane, A. J. Sampling hubbell’s neutral theory of biodiversity. *Ecology Letters*, 7(10):901–910, 2004.
- Alonso, D. and Sole, R. V. The divgame simulator: a stochastic cellular automata model of rainforest dynamics. *Ecological Modelling*, 133(1):131–141, 2000.
- Anderson-Teixeira, K. J., Davies, S. J., Bennett, A. C., Gonzalez-Akre, E. B., Muller-Landau, H. C., Joseph Wright, S., Abu Salim, K., Almeyda Zambrano, A. M., Alonso, A., Baltzer, J. L., et al. Ctf-forestgeo: a worldwide network monitoring forests in an era of global change. *Global Change Biology*, 21(2):528–549, 2015.
- Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
- Arrhenius, O. Species and area. *Journal of Ecology*, 9(1):95–99, 1921.
- Avitabile, V., Herold, M., Heuvelink, G., Lewis, S. L., Phillips, O. L., Asner, G. P., Armston, J., Ashton, P. S., Banin, L., Bayol, N., et al. An integrated pan-tropical biomass map using multiple reference datasets. *Global change biology*, 22(4):1406–1420, 2016.
- Baccini, A., Goetz, S., Walker, W., Laporte, N., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P., Dubayah, R., Friedl, M., et al. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2(3):182–185, 2012.
- Bailey, D. G. and Johnston, C. T. Single pass connected components analysis. In *Proceedings of image and vision computing New Zealand*, pages 282–287, 2007.
- Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- Beaumont, M. A. et al. Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, 41(379-406):1, 2010.

## Bibliography

- Bohlman, S. and O'Brien, S. Allometry, adult stature and regeneration requirement of 65 tree species on barro colorado island, panama. *Journal of Tropical Ecology*, 22(02): 123–136, 2006.
- Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *science*, 320(5882):1444–1449, 2008.
- Box, G. and Cox, D. An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77(377):209–210, 1982.
- Brinck, K., Fischer, R., Groeneveld, J., Lehmann, S., De Paula, M. D., Pütz, S., Sexton, J. O., Song, D., and Huth, A. High resolution analysis of tropical forest fragmentation and its impact on the global carbon cycle. *Nature Communications*, 8:14855, 2017.
- Broadbent, E. N., Asner, G. P., Keller, M., Knapp, D. E., Oliveira, P. J., and Silva, J. N. Forest fragmentation and edge effects from deforestation and selective logging in the brazilian amazon. *Biological conservation*, 141(7):1745–1757, 2008.
- Brown, C., Law, R., Illian, J. B., and Burslem, D. F. Linking ecological processes with spatial and non-spatial patterns in plant communities. *Journal of ecology*, 99(6): 1402–1414, 2011.
- Brown, P. J. Measurement, regression, and calibration. *Journal of Chemometrics*, 8(5): 371–372, 1994.
- Brun, R., Reichert, P., and Künsch, H. R. Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015–1030, 2001.
- Burnham, K. P. and Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.)*. Springer, 2002.
- Campolongo, F., Cariboni, J., and Saltelli, A. An effective screening design for sensitivity analysis of large models. *Environmental modelling & software*, 22(10):1509–1518, 2007.
- Canham, C. D., LePage, P. T., and Coates, K. D. A neighborhood analysis of canopy tree competition: effects of shading versus crowding. *Canadian Journal of Forest Research*, 34(4):778–787, 2004.
- Cechini, M., Mitchell, A., and Pilone, D. Nasa reverb: standards-driven earth science data and service discovery. In *AGU Fall Meeting Abstracts*, volume 1, page 1406, 2011.
- Chave, J., Muller-Landau, H. C., and Levin, S. A. Comparing classical community models: theoretical consequences for patterns of diversity. *The American Naturalist*, 159(1):1–23, 2002.
- Chave, J., Condit, R., Lao, S., Caspersen, J. P., Foster, R. B., and Hubbell, S. P. Spatial and temporal variation of biomass in a tropical forest: results from a large census plot in panama. *Journal of Ecology*, 91(2):240–252, 2003.

- Chave, J., Andalo, C., Brown, S., Cairns, M., Chambers, J., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., et al. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, 145(1):87–99, 2005.
- Chisholm, R. A. and Lichstein, J. W. Linking dispersal, immigration and scale in the neutral theory of biodiversity. *Ecology Letters*, 12(12):1385–1393, 2009.
- Clark, R. N., Swayze, G. A., Wise, R., Livo, K. E., Hoefen, T., Kokaly, R. F., and Sutley, S. J. Usgs digital spectral library splib06a. *US Geological Survey, Digital Data Series*, 231, 2007.
- Condit, R. *Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a comparison with other plots*. Springer Science & Business Media, 1998.
- Condit, R., Robinson, W. D., Ibáñez, R., Aguilar, S., Sanjur, A., Martínez, R., Stallard, R. F., García, T., Angehr, G. R., Petit, L., et al. The status of the panama canal watershed and its biodiversity at the beginning of the 21st century long-term ecological studies reveal a diverse flora and fauna near the panama canal, harbored within a corridor of forest stretching from the caribbean to the pacific, but deforestation, land degradation, erosion, and overhunting remain threats. *BioScience*, 51(5):389–398, 2001.
- Condit, R., Pitman, N., Leigh, E. G., Chave, J., Terborgh, J., Foster, R. B., Núñez, P., Aguilar, S., Valencia, R., Villa, G., et al. Beta-diversity in tropical forest trees. *Science*, 295(5555):666–669, 2002.
- Conti, S., Gosling, J. P., Oakley, J. E., and O’hagan, A. Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3):663–676, 2009.
- Courant, R., Isaacson, E., and Rees, M. On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics*, 5(3):243–255, 1952.
- Cramer, W., Bondeau, A., Woodward, F. I., Prentice, I. C., Betts, R. A., Brovkin, V., Cox, P. M., Fisher, V., Foley, J. A., Friend, A. D., et al. Global response of terrestrial ecosystem structure and function to co2 and climate change: results from six dynamic global vegetation models. *Global change biology*, 7(4):357–373, 2001.
- Davidson, E. A., Ishida, F. Y., and Nepstad, D. C. Effects of an experimental drought on soil emissions of carbon dioxide, methane, nitrous oxide, and nitric oxide in a moist tropical forest. *Global Change Biology*, 10(5):718–730, 2004.
- de Paula, M. D., Groeneveld, J., and Huth, A. The extent of edge effects in fragmented landscapes: Insights from satellite measurements of tree cover. *Ecological Indicators*, 69:196–204, 2016.

- Didham, R. K. and Ewers, R. M. Predicting the impacts of edge effects in fragmented habitats: Laurance and yensen’s core area model revisited. *Biological Conservation*, 155:104–110, 2012.
- Duan, Q., Gupta, V. K., and Sorooshian, S. Shuffled complex evolution approach for effective and efficient global minimization. *Journal of optimization theory and applications*, 76(3):501–521, 1993.
- Elias, P. Universal codeword sets and representations of the integers. *IEEE transactions on information theory*, 21(2):194–203, 1975.
- Ellison, A. M. Bayesian inference in ecology. *Ecology letters*, 7(6):509–520, 2004.
- Etienne, R. S. and Alonso, D. A dispersal-limited sampling theory for species and alleles. *Ecology letters*, 8(11):1147–1156, 2005.
- Etienne, R. S. and Alonso, D. Neutral community theory: how stochasticity and dispersal-limitation can explain species coexistence. *Journal of Statistical Physics*, 128(1-2):485–510, 2007.
- Etienne, R. S., Alonso, D., and McKane, A. J. The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, 248(3):522–536, 2007a.
- Etienne, R. S., Apol, M. E. F., Olf, H., and Weissing, F. J. Modes of speciation and the neutral theory of biodiversity. *Oikos*, 116(2):241–258, 2007b.
- Ewers, R. M. and Banks-Leite, C. Fragmentation impairs the microclimate buffering effect of tropical forests. *PLOS one*, 8(3):e58093, 2013.
- FAO. *Global Forest Resources Assessment 2015*. UN Food and Agriculture Organization, Rome, 2015.
- Fischer, R., Bohn, F., de Paula, M. D., Dislich, C., Groeneveld, J., Gutiérrez, A. G., Kazmierczak, M., Knapp, N., Lehmann, S., Paulick, S., et al. Lessons learned from applying a forest gap model to understand ecosystem and carbon dynamics of complex tropical forests. *Ecological Modelling*, 326:124–133, 2016.
- Fisher, R. A., Corbet, A. S., and Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42–58, 1943.
- Fortunel, C., Valencia, R., Wright, S. J., Garwood, N. C., and Kraft, N. J. Functional trait differences influence neighbourhood interactions in a hyperdiverse amazonian forest. *Ecology Letters*, 19(9):1062–1070, 2016.
- Fricker, T. E., Oakley, J. E., and Urban, N. M. Multivariate gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56, 2013.

- Gallagher, M. and Doherty, J. Parameter estimation and uncertainty analysis for a watershed model. *Environmental Modelling & Software*, 22(7):1000–1020, 2007.
- Gastellu-Etchegorry, J., Martin, E., and Gascon, F. Dart: a 3d model for simulating satellite images and studying surface radiation budget. *International journal of remote sensing*, 25(1):73–96, 2004.
- Ge, Y., Thomasson, J. A., and Sui, R. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, 5(3):229–238, 2011.
- Green, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1984.
- Greene, D. F., Canham, C. D., Coates, K. D., and LePage, P. T. An evaluation of alternative dispersal functions for trees. *Journal of Ecology*, 92(5):758–766, 2004.
- Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmański, J., and Wissel, C. Pattern-oriented modelling in population ecology. *Science of the Total Environment*, 183(1-2): 151–166, 1996.
- Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., and DeAngelis, D. L. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science*, 310(5750):987–991, 2005.
- Groeneveld, J., Alves, L., Bernacci, L., Catharino, E., Knogge, C., Metzger, J., Pütz, S., and Huth, A. The impact of fragmentation and density regulation on forest succession in the atlantic rain forest. *Ecological Modelling*, 220(19):2450–2459, 2009.
- Gu, J., Li, G., and Dong, Z. Hybrid and adaptive meta-model-based global optimization. *Engineering Optimization*, 44(1):87–104, 2012.
- Gustafson, E. J. and Parker, G. R. Using an index of habitat patch proximity for landscape design. *Landscape and Urban Planning*, 29(2-3):117–130, 1994.
- Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., Lovejoy, T. E., Sexton, J. O., Austin, M. P., Collins, C. D., et al. Habitat fragmentation and its lasting impact on earth’s ecosystems. *Science Advances*, 1(2):e1500052, 2015.
- Hansen, M. C., Stehman, S. V., and Potapov, P. V. Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences*, 107(19):8650–8655, 2010.
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman, S., Goetz, S., Loveland, T., et al. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013.

- Hansen, N. and Ostermeier, A. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.
- Harper, K. A., Macdonald, S. E., Burton, P. J., Chen, J., Broszofski, K. D., Saunders, S. C., Euskirchen, E. S., Roberts, D., Jaiteh, M. S., and Esseen, P.-A. Edge influence on forest structure and composition in fragmented landscapes. *Conservation Biology*, 19(3):768–782, 2005.
- Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. Statistical inference for stochastic simulation models—theory and application. *Ecology letters*, 14(8):816–827, 2011.
- Hartig, F., Dislich, C., Wiegand, T., and Huth, A. Technical note: Approximate bayesian parameterization of a complex tropical forest model. *Biogeosciences Discussions*, 10(8):13097–13128, 2013.
- Hartig, F., Dislich, C., Wiegand, T., and Huth, A. Technical note: Approximate bayesian parameterization of a process-based tropical forest model. *Biogeosciences*, 11(4):1261–1272, 2014.
- Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- He, F. Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Functional Ecology*, 19(1):187–193, 2005.
- Hilborn, R. and Mangel, M. *The ecological detective: confronting models with data*, volume 28. Princeton University Press, 1997.
- Hogg, R. V. and Craig, A. T. *Introduction to mathematical statistics*. (5<sup>th</sup> edition). Prentice Hall, Upper Saddle River, New Jersey, 1995.
- Hoos, H. H. and Stützle, T. *Stochastic local search: Foundations and applications*. Elsevier, 2004.
- Hoshen, J. and Kopelman, R. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Physical Review B*, 14(8):3438, 1976a.
- Hoshen, J. and Kopelman, R. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Physical Review B*, 14(8):3438, 1976b.
- Houghton, R., Hall, F., and Goetz, S. J. Importance of biomass in the global carbon cycle. *Journal of Geophysical Research: Biogeosciences*, 114(G2), 2009.
- Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography (MPB-32)*, volume 32. Princeton University Press, 2001.

- Hubbell, S. P., Foster, R. B., O'Brien, S. T., Harms, K., Condit, R., Wechsler, B., Wright, S. J., and De Lao, S. L. Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283(5401):554–557, 1999.
- Hubbell, S. P., Condit, R., and Foster, R. Barro colorado forest census plot data. *URL* <http://ctfs.si.edu/datasets/bci>, 2005.
- Huber, P. J. *Robust statistics*. Springer, 2011.
- Ingber, L. Simulated annealing: Practice versus theory. *Mathematical and computer modelling*, 18(11):29–57, 1993.
- Ingber, L. Adaptive simulated annealing (asa): Lessons learned. In *Control and cybernetics*. Citeseer, 1996.
- Kirkpatrick, S. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, 34(5-6):975–986, 1984.
- Köhler, P. and Huth, A. The effects of tree species grouping in tropical rainforest modelling: Simulations with the individual-based model formind. *Ecological Modelling*, 109(3):301–321, 1998.
- Kohyama, T. Simulating stationary size distribution of trees in rain forests. *Annals of Botany*, 68(2):173–180, 1991.
- Kraft, N. J. and Ackerly, D. D. Functional trait and phylogenetic tests of community assembly across spatial scales in an amazonian forest. *Ecological Monographs*, 80(3): 401–422, 2010.
- Kraft, N. J., Godoy, O., and Levine, J. M. Plant functional traits and the multidimensional nature of species coexistence. *Proceedings of the National Academy of Sciences*, 112(3):797–802, 2015.
- Laurance, W. F. Do edge effects occur over large spatial scales? *Trends in Ecology & Evolution*, 15(4):134–135, 2000.
- Laurance, W. F., Lovejoy, T. E., Vasconcelos, H. L., Bruna, E. M., Didham, R. K., Stouffer, P. C., Gascon, C., Bierregaard, R. O., Laurance, S. G., and Sampaio, E. Ecosystem decay of amazonian forest fragments: a 22-year investigation. *Conservation Biology*, 16(3):605–618, 2002.
- Laurance, W. F., Camargo, J. L., Luizão, R. C., Laurance, S. G., Pimm, S. L., Bruna, E. M., Stouffer, P. C., Williamson, G. B., Benítez-Malvido, J., Vasconcelos, H. L., et al. The fate of amazonian forest fragments: a 32-year investigation. *Biological Conservation*, 144(1):56–67, 2011.
- Lee, P. M. *Bayesian statistics: an introduction*. John Wiley & Sons, 2012.



## Bibliography

- Lehmann, S. and Huth, A. Fast calibration of a dynamic vegetation model with minimum observation data. *Ecological Modelling*, 301:98–105, 2015.
- Liu, Y. and Gupta, H. V. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43(7), 2007.
- Lobo, E. and Dalling, J. W. Spatial scale and sampling resolution affect measures of gap disturbance in a lowland tropical forest: implications for understanding forest regeneration and carbon storage. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1778):20133218, 2014.
- Luo, Y., Nie, J., and Young, E. R. Model uncertainty, state uncertainty, and state-space models. In *State-Space Models*, pages 91–112. Springer, 2013.
- MacArthur, R. H. *Geographical ecology: patterns in the distribution of species*. Princeton University Press, 1972.
- Mandelbrot, B. B. How long is the coast of britain. *Science*, 156(3775):636–638, 1967.
- May, F., Huth, A., and Wiegand, T. Moving beyond abundance distributions: neutral theory and spatial patterns in a tropical forest. In *Proceedings of the Royal Society B: Biological Sciences*, volume 282, page 20141657. The Royal Society, 2015.
- May, F., Wiegand, T., Lehmann, S., and Huth, A. Do abundance distributions and species aggregation correctly predict macroecological biodiversity patterns in tropical forests? *Global Ecology and Biogeography*, 2016.
- McCarthy, M. A. *Bayesian methods for ecology*. Cambridge University Press, 2007.
- McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature*, 422(6934):881–885, 2003.
- McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, 10(10):995–1015, 2007.
- McKay, M. D., Beckman, R. J., and Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.
- McRoberts, R. E. and Tomppo, E. O. Remote sensing support for national forest inventories. *Remote Sensing of Environment*, 110(4):412–419, 2007.
- Meyer, K. and Hill, W. Approximation of sampling variances and confidence intervals for maximum likelihood estimates of variance components. *Journal of Animal Breeding and Genetics*, 109(1-6):264–280, 1992.

- Meyer, V., Saatchi, S., Chave, J., Dalling, J., Bohlman, S., Fricker, G., Robinson, C., Neumann, M., and Hubbell, S. Detecting tropical forest biomass dynamics from repeated airborne lidar measurements. *Biogeosciences*, 10(8):5421–5438, 2013.
- Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.
- Moorcroft, P., Hurtt, G., and Pacala, S. W. A method for scaling vegetation dynamics: the ecosystem demography model (ed). *Ecological monographs*, 71(4):557–586, 2001.
- Moore, G. E. et al. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.
- Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water resources research*, 41(5), 2005.
- Morlon, H., Chuyong, G., Condit, R., Hubbell, S., Kenfack, D., Thomas, D., Valencia, R., and Green, J. L. A general framework for the distance–decay of similarity in ecological communities. *Ecology letters*, 11(9):904–917, 2008.
- Morris, M. D. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.
- Mularie, W. Department of defense world geodetic system 1984, its definition and relationships with local geodetic systems. *National Geospatial-Intelligence Agency, Tech. Rep.*, 152, 2000.
- Muller-Landau, H. *Seed Dispersal in a Tropical Forest: Empirical Patterns, Their Origins, and Their Consequences for Community Dynamics*. Princeton University, 2001.
- Muller-Landau, H. C., Condit, R. S., Harms, K. E., Marks, C. O., Thomas, S. C., Bunyavejchewin, S., Chuyong, G., Co, L., Davies, S., Foster, R., et al. Comparing tropical forest tree size distributions with the predictions of metabolic ecology and equilibrium models. *Ecology Letters*, 9(5):589–602, 2006.
- Muller-Landau, H. C., Wright, S. J., Calderón, O., Condit, R., and Hubbell, S. P. Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology*, 96(4):653–667, 2008.
- Nepstad, D. C., Tohver, I. M., Ray, D., Moutinho, P., and Cardinot, G. Mortality of large trees and lianas following experimental drought in an amazon forest. *Ecology*, 88(9):2259–2269, 2007.
- O’Dwyer, J., Lake, J., Ostling, A., Savage, V., and Green, J. An integrative framework for stochastic, size-structured community assembly. *Proceedings of the National Academy of Sciences*, 106(15):6170–6175, 2009.

## Bibliography

- O'Dwyer, J. P. and Green, J. L. Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecology letters*, 13(1):87–95, 2010.
- Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al. 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. *BioScience*, 51(11):933–938, 2001a.
- Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al. 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. *BioScience*, 51(11):933–938, 2001b.
- Osborne, P. *The mercator projections*. Zenodo, 2013.
- Pachauri, R. K., Meyer, L., Plattner, G.-K., Stocker, T., et al. *IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, 2015.
- Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., et al. A large and persistent carbon sink in the world's forests. *Science*, 333(6045):988–993, 2011.
- Patton, D. R. A diversity index for quantifying habitat "edge". *Wildlife Society Bulletin (1973-2006)*, 3(4):171–173, 1975.
- Pereira, H. M., Leadley, P. W., Proença, V., Alkemade, R., Scharlemann, J. P., Fernandez-Manjarrés, J. F., Araújo, M. B., Balvanera, P., Biggs, R., Cheung, W. W., et al. Scenarios for global biodiversity in the 21st century. *Science*, 330(6010):1496–1501, 2010.
- Peters, R. H. *The ecological implications of body size*, volume 2. Cambridge University Press, 1986.
- Phillips, O. L. and Gentry, A. H. Increasing turnover through time in tropical forests. *Science*, 263(5149):954–958, 1994.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical recipes in C*, volume 2. Cambridge University Press, 1982.
- Preston, F. W. The commonness, and rarity, of species. *Ecology*, 29(3):254–283, 1948.
- Preston, F. W. The canonical distribution of commonness and rarity: Part i. *Ecology*, 43(2):185–215, 1962.

- Pütz, S., Groeneveld, J., Henle, K., Knogge, C., Martensen, A. C., Metz, M., Metzger, J. P., Ribeiro, M. C., de Paula, M. D., and Huth, A. Long-term carbon loss in fragmented neotropical forests. *Nature communications*, 5, 2014.
- Rands, M. R., Adams, W. M., Bennun, L., Butchart, S. H., Clements, A., Coomes, D., Entwistle, A., Hodge, I., Kapos, V., Scharlemann, J. P., et al. Biodiversity conservation: challenges beyond 2010. *Science*, 329(5997):1298–1303, 2010.
- Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.
- Raven, P. H. Our diminishing tropical forests. *Biodiversity*, pages 119–122, 1988.
- Regis, R. G. and Shoemaker, C. A. Constrained global optimization of expensive black box functions using radial basis functions. *Journal of Global Optimization*, 31(1): 153–171, 2005.
- Ricklefs, R. and Miller, G. *Ecology*. W.H. Freeman, 1990.
- Riitters, K., Wickham, J., O’Neill, R., Jones, B., Smith, E., et al. Global-scale patterns of forest fragmentation. *Conservation Ecology*, 4(2):3, 2000.
- Riitters, K., Wickham, J., Costanza, J. K., and Vogt, P. A global evaluation of forest interior area dynamics using tree cover data from 2000 to 2012. *Landscape Ecology*, 31(1):137–148, 2016.
- Rödig, E., Huth, A., Bohn, F., Rebmann, C., and Cuntz, M. Estimating the carbon fluxes of forests with an individual-based forest model. *Forest Ecosystems*, 4(1):4, 2017.
- Rosenqvist, A., Shimada, M., Ito, N., and Watanabe, M. Alos palsar: A pathfinder mission for global-scale monitoring of the environment. *IEEE Transactions on Geoscience and Remote Sensing*, 45(11):3307–3316, 2007.
- Rosenzweig, M. L. *Species diversity in space and time*. Cambridge University Press, 1995.
- Rosindell, J. and Cornell, S. J. Species–area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology letters*, 10(7):586–595, 2007.
- Rosindell, J. and Cornell, S. J. Species–area curves, neutral models, and long-distance dispersal. *Ecology*, 90(7):1743–1750, 2009.
- Rosindell, J., Cornell, S. J., Hubbell, S. P., and Etienne, R. S. Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, 13(6):716–727, 2010.

- Rosindell, J., Hubbell, S. P., and Etienne, R. S. The unified neutral theory of biodiversity and biogeography at age ten. *Trends in ecology & evolution*, 26(7):340–348, 2011.
- Rosindell, J., Jansen, P. A., and Etienne, R. S. Age structure in neutral theory resolves inconsistencies related to reproductive-size threshold. *Journal of plant ecology*, 5(1): 64–71, 2012.
- Rutledge, D. T. *Landscape indices as measures of the effects of fragmentation: can pattern reflect process?* Department of Conservation, Wellington, 2003.
- Saatchi, S. S., Houghton, R., Dos Santos Alvala, R., Soares, J., and Yu, Y. Distribution of aboveground live biomass in the amazon basin. *Global Change Biology*, 13(4): 816–837, 2007.
- Sanchez, D. G., Lacarri re, B., Musy, M., and Bourges, B. Application of sensitivity analysis in building energy simulations: Combining first-and second-order elementary effects methods. *Energy and Buildings*, 68:741–750, 2014.
- Sexton, J. O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins, K. M., Channan, S., DiMiceli, C., et al. Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of modis vegetation continuous fields with lidar-based estimates of error. *International Journal of Digital Earth*, 6(5):427–448, 2013.
- Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., and Lucas, R. New global forest/non-forest maps from alos palsar data (2007–2010). *Remote Sensing of environment*, 155:13–31, 2014.
- Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., and Biss, P. M. The park grass experiment 1856–2006: its contribution to ecology. *Journal of Ecology*, 94 (4):801–814, 2006.
- Simberloff, D. S. and Wilson, E. O. Experimental zoogeography of islands: the colonization of empty islands. *Ecology*, 50(2):278–296, 1969.
- Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J., Levis, S., Lucht, W., Sykes, M. T., et al. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the lpj dynamic global vegetation model. *Global Change Biology*, 9(2):161–185, 2003.
- Slik, J. F., Arroyo-Rodr guez, V., Aiba, S.-I., Alvarez-Loayza, P., Alves, L. F., Ashton, P., Balvanera, P., Bastian, M. L., Bellingham, P. J., Van Den Berg, E., et al. An estimate of the number of tropical tree species. *Proceedings of the National Academy of Sciences*, 112(24):7472–7477, 2015.
- Smith, P. M., Kalluri, S. N., Prince, S. D., and DeFries, R. The noaa/nasa pathfinder avhrr 8-km land data set. *Photogrammetric Engineering and remote sensing*, 63(1): 27–32, 1997.

- Stauffer, D. and Aharony, A. *Introduction to percolation theory*. CRC press, 1994.
- Stoll, P. and Newbery, D. M. Evidence of species-specific neighborhood effects in the dipterocarpaceae of a bornean rain forest. *Ecology*, 86(11):3048–3062, 2005.
- Storn, R. and Price, K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.
- Süli, E. and Mayers, D. F. *An introduction to numerical analysis*. Cambridge University press, 2003.
- Tangirala, A. K. *Principles of system identification: theory and practice*. CRC Press, 2014.
- Tarantola, A. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.
- Tilman, D. and Downing, J. A. Biodiversity and stability in grasslands. *Nature*, 367(6461):363–365, 1994.
- Tilman, D., Reich, P. B., Knops, J., Wedin, D., Mielke, T., and Lehman, C. Diversity and productivity in a long-term grassland experiment. *Science*, 294(5543):843–845, 2001.
- Tofallis, C. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362, 2015.
- Tolson, B. A. and Shoemaker, C. A. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43(1), 2007a.
- Tolson, B. A. and Shoemaker, C. A. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43(1), 2007b.
- Uriarte, M., Canham, C. D., Thompson, J., and Zimmerman, J. K. A neighborhood analysis of tree growth and survival in a hurricane-driven tropical forest. *Ecological Monographs*, 74(4):591–614, 2004.
- Uriarte, M., Swenson, N. G., Chazdon, R. L., Comita, L. S., John Kress, W., Erickson, D., Forero-Montaña, J., Zimmerman, J. K., and Thompson, J. Trait similarity, shared ancestry and the structure of neighbourhood interactions in a subtropical wet forest: implications for community assembly. *Ecology letters*, 13(12):1503–1514, 2010.
- Van Oijen, M., Rougier, J., and Smith, R. Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiology*, 25(7):915–927, 2005a.

## Bibliography

- Van Oijen, M., Rougier, J., and Smith, R. Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiology*, 25(7):915–927, 2005b.
- Van Oijen, M., Reyer, C., Bohn, F., Cameron, D., Deckmyn, G., Flechsig, M., Härkönen, S., Hartig, F., Huth, A., Kiviste, A., et al. Bayesian calibration, comparison and averaging of six forest models, using data from scots pine stands across europe. *Forest Ecology and Management*, 289:255–268, 2013.
- Veefkind, J., Aben, I., McMullan, K., Förster, H., De Vries, J., Otter, G., Claas, J., Eskes, H., De Haan, J., Kleipool, Q., et al. Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120:70–83, 2012.
- Vincenty, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975.
- Virkar, Y., Clauset, A., et al. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, 8(1):89–119, 2014.
- Volkov, I., Banavar, J. R., Hubbell, S. P., and Maritan, A. Neutral theory and relative species abundance in ecology. *Nature*, 424(6952):1035–1037, 2003.
- Volkov, I., Banavar, J. R., He, F., Hubbell, S. P., and Maritan, A. Density dependence explains tree species abundance and diversity in tropical forests. *Nature*, 438(7068):658–661, 2005.
- Wade, T. G., Riitters, K. H., Wickham, J. D., and Jones, K. B. Distribution and causes of global forest fragmentation. *Conservation Ecology*, 7(2), 2003.
- Weigelt, A., Marquard, E., Temperton, V. M., Roscher, C., Scherber, C., Mwangi, P. N., Felten, S., Buchmann, N., Schmid, B., Schulze, E.-D., et al. The jena experiment: six years of data from a grassland biodiversity experiment. *Ecology*, 91(3):930–931, 2010.
- Wiegand, T. and Moloney, K. A. *Handbook of spatial point-pattern analysis in ecology*. CRC Press, 2013.
- Wiegand, T., Jeltsch, F., Hanski, I., and Grimm, V. Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *Oikos*, 100(2):209–222, 2003.
- Wilson, E. O. et al. The current state of biological diversity. *Biodiversity*, 521(1):3–18, 1988.
- Wissel, C. Aims and limits of ecological modelling exemplified by island theory. *Ecological Modelling*, 63(1-4):1–12, 1992.

- Xiao, X., O'Dwyer, J. P., and White, E. P. Comparing process-based and constraint-based approaches for modeling macroecological patterns. *Ecology*, 97(5):1228–1238, 2016.
- Zhang, J. and Sanderson, A. C. Jade: Adaptive differential evolution with optional xternal archive. *IEEE Transactions on Evolutionary Computation*, 13(5), 2009.
- Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 1977.



## Acknowledgement

An dieser Stelle möchte ich mich herzlich bei allen Personen bedanken, die zum Gelingen dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt meinen Betreuern Prof. Dr. Andreas Huth und Dr. Thorsten Wiegand für die uneingeschränkte Unterstützung, ihr Vertrauen, den eingeräumten Freiheiten und den inspirierenden Diskussionen. Besonders bei Prof. Dr. Andreas Huth möchte ich mich für die ideenreiche Betreuung bedanken, die seit meiner Zeit als Praktikant am UFZ bestand hat. Er konnte mich von Beginn an für die ökologische Modellierung begeistern und hat es in den letzten 8 Jahren geschafft meine Stärken zu fördern und mir bei Schwierigkeiten mit Unterstützung und Hilfe beiseite zu stehen.

Für die hilfreichen Diskussionen geht mein Dank an Felix May (neutrale Theorie), Juliane Mai und Matthias Cuntz (Optimierung und Unsicherheitsanalyse) und meinen Bürokollegen Frank Masurowski (GIS Systeme) und Juliane Horn (ökologische Modellierung).

Herzlich möchte ich mich auch bei meiner Familie, Christina Richter und meiner Verlobten Sandra Wachsmuth bedanken, die stets großes Interesse an meiner Promotion hatten und mich mit Verständnis allzeit unterstützten.

# Declaration of authorship

## Erklärung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise unentgeltlich geholfen.

**Kapitel 2:** Koautoren des veröffentlichten Artikels: Andreas Huth.

**Kapitel 3:** Koautoren des Manuskripts: Thorsten Wiegand, Andreas Huth.

**Kapitel 4:** Koautoren des Manuskripts: Rico Fischer, Andreas Huth.

Weitere Personen waren an der inhaltlichen und materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

---

Ort, Datum

---

Unterschrift