EXPLORING ROBUSTNESS AND UNCERTAINTIES OF PROJECTIONS WITH FOREST ECOSYSTEM MODELS

DISSERTATION ZUR ERLANGUNG DES
DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von
Johannes Oberpriller

aus
Landshut

im Jahre
2022

EXPLORING ROBUSTNESS AND UNCERTAINTIES OF PROJECTIONS WITH FOREST ECOSYSTEM MODELS

DISSERTATION ZUR ERLANGUNG DES
DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von
Johannes Oberpriller

aus
Landshut

im Jahre
2022

Das Promotionsgesuch wurde eingereicht am:
16. März 2022

Die Arbeit wurde angeleitet von:
Prof. Dr. Florian Hartig

Unterschrift:

_____

Johannes Oberpriller

Declaration of Manuscripts

This thesis is composed of manuscripts, which are largely identical with manuscript accepted or submitted for publication:

Chapter 3:

Johannes Oberpriller Christine Herschlein,Peter Anthoni, Almut Arneth, Andreas Krause, Anja Rammig, Mats Lindeskog, Stefan Olin, Florian Hartig; "*Climate and parameter sensitivity and induced uncertainties in carbon stock projections for European forests (using LPJ-GUESS 4.0)*"; Submitted to *Geoscientific Model Development*

Author Contributions: JO and FH conceived and designed the study and wrote a first draft. JO implemented the case studies, ran the experiments, and analyzed the results. All authors contributed to discussing and interpreting the results, and to the preparation of the manuscript.

Chapter 4:

Johannes Oberpriller, Volodymyr Trotsiuk, Lukas Heiland, Lisa Hülsmann, Florian Hartig; "*A site-specific Bayesian calibration of a physiological forest model shows intraspecific functional variation in tree species across Europe*";Submitted to *Ecography*

Author Contributions: JO and FH conceived and designed the study and wrote a first draft. JO implemented the case studies,ran the experiments, and analyzed the results. LHe and JO prepared the data. All authors contributed to discussing and interpreting the results, and to the preparation of the manuscript.

Chapter 5:

Johannes Oberpriller, Melina de Souza Leite, Maximilian Pichler; "*Fixed or random? On the reliability of mixed-effects models for a small number of levels in grouping variables*"; Provisionally accepted in *Ecology and Evolution*

Author Contributions: MP, JO and MSL designed the study. MP and JO ran the simulations, analyzed the results and wrote a first draft. All authors contributed equally to revising the manuscript and interpreting and discussing results.

Chapter 6:

Johannes Oberpriller, David R. Cameron, Michael C. Dietze, Florian Hartig; *"Towards robust statistical inference for complex computer models*"; Published in *Ecology Letters, 24.6, pp. 1251–1261, 2021.*

Author Contributions: FH and JO conceived and designed the study. JO implemented the case studies, ran the experiments, and analyzed the results. All authors contributed to discussing and interpreting the results, and to the preparation of the manuscript.

Manuscripts **not included** in this thesis:

Johannes Oberpriller, Almut Arneth, Christine Herschlein, Anja Rammig,Florian Hartig; "*Modellunsicherheiten in Klimafolgeprojektionen*"; Published in *Fränkisch Geographische Gesellschaft, Band 67, pp.70-81, 2021*

Juliano Sarmento Cabral, Alma Mendoza-Ponce, André Pinto da Silva,**Johannes Oberpriller**, Anne Mimet, Julia Kieslinger, Thomas Berger, Jana Blechschmidt,Maximilian Brönner, Alice Classen, Stefan Fallert, Florian Hartig, Christian Hof, Markus Hoffmann, Thomas Knoke, Andreas Krause, Anne Lewerentz, Perdita Pohle, Uta Raeder,Anja Rammig, Sarah Redlich, Sven Rubanschi, Christian Stetter, Wolfgang Weisser, Daniel Vedder, Peter H. Verburg, Damaris Zurell; *"The road to integrate climate change effects on land-use change in regional biodiversity models*"; Submitted to *People and Nature*

David R. Cameron, Florian Hartig,**Johannes Oberpriller**, Björn Reineking, Marcel van Oijen, Michael C. Dietze; *"Identifying why the Bayesian calibration of process-based models with unbalanced quantities of calibration data can be challenging: The significance of model sturctural deficiencies and data biases"*; In preperation for *Methods in Ecology and Evolution*

CONTENTS

SUMMARY

Forests act as important CO_2 sinks and might help to reduce the impacts of global and climate change. We can explore such scenarios with forest ecosystem models as their mechanistic structure in principle allows forecasting into never-observed conditions. However, to make realistic projections, we have to adjust the model to fit the observed data. To do so and to assess uncertainties of projections, researchers use methods like a sensitivity and uncertainty analysis but also Bayesian calibration. However, the naive application and and the associated assumptions of these methods do often not reflect the empirical knowledge about forest ecosystems. To adress these issues, this doctoral thesis analyzes the robustusness of and applies these numerical methods to state-of-the-art forest ecosystem models. We asked the following questions: The first one was: What are the main contributors of uncertainty in forest ecosystem models? Can we use uncertainty analysis and calibration of forest ecosystem models to analyze ecological patterns on environmental gradients? The next question deals with the remaining variance: To what extent can random effects be used to represent ecological variation and how much data points are required to estimate these variations precisely? And the last question investigates if the findings above are robust when we have structural model error: What are the consequences and solutions of calibrating of and projecting with models with structural errors?

The **first chapter** introduces the role of forest ecosystem models and their associated uncertainities for projecting forest dynamics under climate change and emerging challenges. In the **second chapter**, we explain key concepts and methods which are essential to understand our research results. In particular these are types of forest ecosystem models, their associated uncertainties (due to initial conditions, model inputs, model structure and parameters), sensitivity and uncertainty analysis and Bayesian calibration. In the **third chapter**, we analyze sensitivities and uncertainties of carbon projections across European forests under climate change with a dynamic vegetation model (LPJ-GUESS 4.0) addressing the effect of both model parameters and environmental drivers. We find that carbon projections are most sensitive to photosynthesis-related parameters, while environmental drivers induce most uncertainty. Moreover, environmental drivers modify the uncertainties of other parameters. This study shows that environmental drivers are strong contributors and modifiers of uncertainties in other ecosystem processes. In the **fourth chapter**, we analyze the consequences and possibilities to represent intraspecific variation in the calibration of a forest ecosystem model. To do so, we calibrate the 3-PG model against biomass derived from inventory data across Germany and Sweden with a hierarchical Bayesian calibration scheme. We find evidence for intraspecific variation that can be partly explained by environmental conditions. This study shows the potential of using forest ecosystem models to infer not measurable ecological information. In the **fifth chapter**, we analyze if with a low number of levels it is better to model a grouping variable as a random or as a fixed-effect. We find with varying intercepts and slopes in the data-generating process, using a random slope and intercept model, and, in case of a singular fit switching to a fixed-effects model, avoids overconfidence in the results. This study shows how to make ecological inference with mixed-effects models more robust for a small number of levels. In the **sixth chapter**, we explain why model error causes bias and underestimated uncertainties, especially when calibrated against unbalanced data, and propose a framework for robust inference with complex computer simulations. As possible solutions we discuss data rebalancing and adding bias corrections during or after the calibration procedure. We illustrate the methods in a case study, using a dynamic vegetation model. From this, we conclude that developing better methods for robust inference of complex computer simulations is essential for generating reliable predictions. The **last chapter** discusses the relevance and significance of our studies for forecasting and inference with forest ecosystem models and outlines further research questions.

1 INTRODUCTION

1.1 GLOBAL AND CLIMATE CHANGE DEMAND PROJECTIONS WITH FOREST ECOSYSTEM MODELS

Global and climate change pose challenges in all areas of human life, e.g. agriculture (e.g. Howden et al., 2007) or ecosystem conservation (e.g. Scholze et al., 2006). Rising temperatures and a higher frequency of extreme weather events (e.g. severe droughts, heat waves and floods (see IPCC, 2014) are at least in part a consequence of increased anthropogenic CO_2 emissions since the industrial revolution (e.g. Solomon et al., 2009). By now it is well accepted that a drastic reduction in net CO_2 emissions is required to meet the commitments of the Paris Agreement (e.g. Liu and Raftery, 2021).

Forests cover one third of the terrestrial area and act as important CO_2 sinks (Luyssaert et al., 2008), but they also provide other ecosystem services such as nutrient cycling, water- and air purification and wildlife habitat maintenance (e.g. Brockerhoff et al., 2017). In the face of progressing climate change, however, their potential for carbon storage and their ability to maintain their immense value for the natural system are uncertain (e.g. Krause et al., 2019). To explore potential future scenarios for their role in mitigating climate change and maintaining forest ecosystem services, robust predictions and a quantification of the associated uncertainties are required (see Gustafson, 2013).

Forest ecosystem models are in principle able to project forest dynamics under climate change (Kearney et al., 2010; Rastetter, 2017). The rationale behind this is that, having sufficiently described, implemented and parameterised the underlying processes of the climate-driven forest ecosystem, the mechanistic formulation extrapolates sufficiently well into never observed conditions (e.g. Radchuk et al., 2019). Thus, forest ecosystem models have a wide range of applications such as policy reports but also in basic research (Cramer et al., 2001; Snell et al., 2014; Fisher et al., 2018; IPCC, 2014).

For decision making, however, it is often not enough to know the most likely future scenario, but to evaluate different management strategies decision-makers need to understand the uncertainties of these scenarios (O'Hagan, 2012). If uncertainties are not quantified correctly potential irreversible effects, e.g. species extinctions, might not be considered and thus their impacts underestimated (Uusitalo et al., 2015). To provide decision-makers and researchers a realistic picture of forests under climate change, we have to evaluate the nature and extent of the uncertainties in forest ecosystem model predictions (e.g. Burgman, 2005).

Uncertainty in model predictions comes from various parts, which are identical to the most important model components (see Latif, 2011): model inputs, initial conditions, model structure and model parameters. Fundamentally, uncertainty can be split into three different categories (see also Walker, 2013): Risk, uncertainty and deep uncertainty. By risk we mean a situation in which we know the probabilities of the possible future scenarios. By uncertainty we refer to a situation in which possible future scenarios are known, but not their probability. By deep or complete uncertainty we mean situations in which not even all possible future scenarios are known (Knight, 1921). The uncertainties in projections of forest ecosystems under climate are a mixture of these different fundamental types of uncertainty (Yousefpour and Hanewinkel, 2016). While parametric and initial uncertainty are risks, model structural uncertainty could be seen as risk, uncertainty or deep uncertainty. Climate inputs are uncertainties (e.g. there are no probabilities for the anthropogenic CO_2 emissions) or deep uncertainties (e.g. due to model structural uncer-

tainty in the climate models, see Lawrence et al., 2020). Overall, all fundamental categories as well as all parts of the model contribute to the predictive uncertainty of forest ecosystem models.

1.2 NUMERICAL METHODS FOR QUANTIFYING FOREST ECOSYSTEM MODELS UNCERTAINTIES

Recent computational advances allowed the ecosystem modeling community to evaluate the model more often with different parameterizations or climatic conditions and to use these multiple model evaluations to analyze model behavior and quantify uncertainities (e.g. Clark, 2020). Thus, methods like a sensitivity (SA, e.g. Saltelli, 2002) or uncertainty analysis (UA, e.g. Saltelli et al., 2019) but also Bayesian calibration (e.g. Hartig et al., 2012) are now the basis for improving and analyzing forest ecosystem models and their uncertainties.

Sensitivity and uncertainty analysis aim at identifying model components that strongly influence the predictions of the model (SA) and contribute a significant amount of uncertainty (UA) (Caswell, 2019). Reasons for the strong influence on the predictions or uncertainties can be, among others, that the model reacts very sensitively to the model component or that the model component itself is very uncertain. Although this is true for all model components, in most SAs/UAs initial conditions and structural uncertainty are ignored, and model inputs are usually not propagated probabilistically, but as a scenario analysis (for a comprehensive review see Wu and Li, 2006). Thus, most SA/UAs concentrate on parametric uncertainty.

To reduce the parametric uncertainty in the parameters identified via a SA/UA, model calibration compares and adjusts them to data (see Hartig et al., 2012). Recently, the field has moved from informal methods to methods based in probability theory like Bayesian calibration (e.g. Clark, 2005). Bayesian methods allow to incorporate prior knowledge, which then gets updated by a model-data comparison (e.g. McElreath, 2016). Bayesian calibration also offers a clear pathway for calibrating ecosystem models, propagating uncertainties to model outputs and quantifying evidence in results (e.g. Dietze, 2017a).

However, the naive application and and the associated assumptions of these methods do often not reflect the empirical knowledge about forest ecosystems. For example, it is by now well accepted that the importance of different processes for forest growth and function varies with climatic conditions (e.g. Bonan, 2008). Thus, comprehensive analyses of forest ecosystems should allow for changes in parameter values and their uncertainty along environmental gradients. Moreover, there is often additional biological variation in ecological parameters that is not accounted for in the calibration (Clark, 2005). For example intraspecific variation is either not represented at all in forest ecosystem models and species act as homogeneous units, or individuals of a species are completely independent of each other. It is unclear (but proposed, e.g Laubmeier et al., 2020) whether one can correctly estimate these variations in parameters representing biological traits with random effects and if so how many data points are required to estimate the variations precisely or when it is beneficial to fall back to a homogeneous unit (analogous in a regression model: no modeling of the grouping variable) or completely independent individuals assumption (analogous in a regression model: fixed effect for a grouping variable). This question is especially interesting as it is also unclear for simple regression models (see Harrison et al., 2017). Finally, we know that forest ecosystem models have a certain amount of structural deficiencies (e.g. Van Oijen, 2017). When we calibrate models with structural error to data, however, almost all calibration algorithms do not account for structural error (e.g. Kennedy and O'Hagan, 2001). Thus, it is not clear how much we can rely on their results and uncertainty estimates and how we should correct for structural model error.

1.3 RESEARCH QUESTIONS

The aim of this thesis is to advance inference and calibration of forest ecosystem models. The particular focus of this thesis lies on the computational challenges and inferential properties of numerical methods applied to forest ecosystem models and their predictive uncertainties. Our main questions are:

Using numerical experiments to analyze sensitivities, uncertainties and patterns along environmental gradients: What are the main contributors of uncertainty in forest ecosystem models? Can we use uncertainty analysis and calibration of forest ecosystem models to analyze ecological patterns on environmental gradients? The next question deals with the remaining variance:

Random effects to model unexplained variance in ecological experiments: To what extent can random effects be used to represent ecological variation and how much data points are required to estimate these variations precisely? And the last question investigates if the results of inferred results of forest ecosystem models are robust against structural model error:

Consequences and solutions of projections with wrong models: What are the consequences and solutions of calibrating and projecting with models with structural errors?

The following chapter outlines the concepts and methods to approach these questions. Chapters 3-6 consist of the research papers published or prepared during this PhD. Chapter 7 concludes this dissertation with a joint discussion of the results and lessons learned.

2 CONCEPTS AND METHODOLOGIES

The aim of this chapter is to explain key concepts and methods which are essential to understand our research results. The first section 2.1 deals with the concepts and ideas of forest ecosystem models. The associated uncertainties of a forest ecosystem model are explained in section 2.2. The next section 2.3 explains how numerical experiments help to understand and improve forest ecosystem models, how uncertainties can be attributed to different parts and how uncertainties of model projections can be reduced.

2.1 FOREST ECOSYSTEM MODEL

Forests under climate change are simulated with different types of models: empirical models, gap-type models, and process-based models (for a review of different model types see Reyer, 2015). While empirical models due to their purely statistical approach have problems extrapolating to new environments, forest gap models (see Bugmann, 2001) can capture long-term forest dynamics, but oversimplify climate-growth relationships (e.g. Schenk, 1996).

Process-based models mechanistically describe biochemical and physiological processes in forests and thus are most suitable to simulate forest response under climate change (Makela et al., 2000; Radchuk et al., 2019). Depending on their scope of application, process-based models can be divided into two categories. Stand-scale process-based models simulate climatic impact on forest stands with a focus on physiological characteristics and stand structure (see Fontes et al., 2011). To do so, they need detailed stand-level input data often including inventory data to set up the simulations and a species-specific parametrization. In contrast, biogeochemical or dynamic global vegetation models often concentrate on plant functional types instead of species, start from bare ground, spin up vegetation pools and simulate forest dynamics on a continental or global scale (e.g. the LPJ-GUESS model, Smith et al., 2001). Their focus is more on simulating cycling of carbon, nutrient or water pools instead of stand scale structure (Brilli et al., 2017). As the natural spatial scale of these processes is continental or even global, biogeochemical or dynamic global vegetation models are typically applied on these scales. Both process-based model categories, however, share the goal of mechanistically describing the processes in forests.

Although the concrete implementation of processes depends on the specific model (and even on the version), most process-based models have a similar structure (Xia et al., 2013). They usually describe forest development via submodules of different process groups: establishment, light absorption, productivity, water/nutrient cycling, biomass allocation, soil, mortality and management (Fig. 2.1). Based on these processes the forest pools and structures are updated at specific time steps. The updating intervals depend on the model and can also differ between modules. In the establishment submodule new trees are established based on bioclimatic limits of the investigated species and enough available resources (space, light, water and radiation). In the light absorption module light absorption is simulated based on light extinction, often considering the horizontal canopy structure. The productivity submodule calculates gross primary production and net primary production based on the quantum efficiency of photosynthesis. Quantum efficiency usually varies between species and usually depends on the environmental conditions (sometimes including soil). In the nutrient and water cycling submodule evapotranspiration based on tree transpiration and soil evaporation is calculated. In some models additionally the cycling of nutrients, e.g. phosphorus or nitrogen, is simulated. In the biomass allocation submodule different parts of the trees allocate the available biomass with the distribution across parts depending on environmental and soil conditions. The mortality submodule calculates mortality often based on density dependence and other factors such as pests, diseases and droughts. The

management module implements management on the investigated sites.

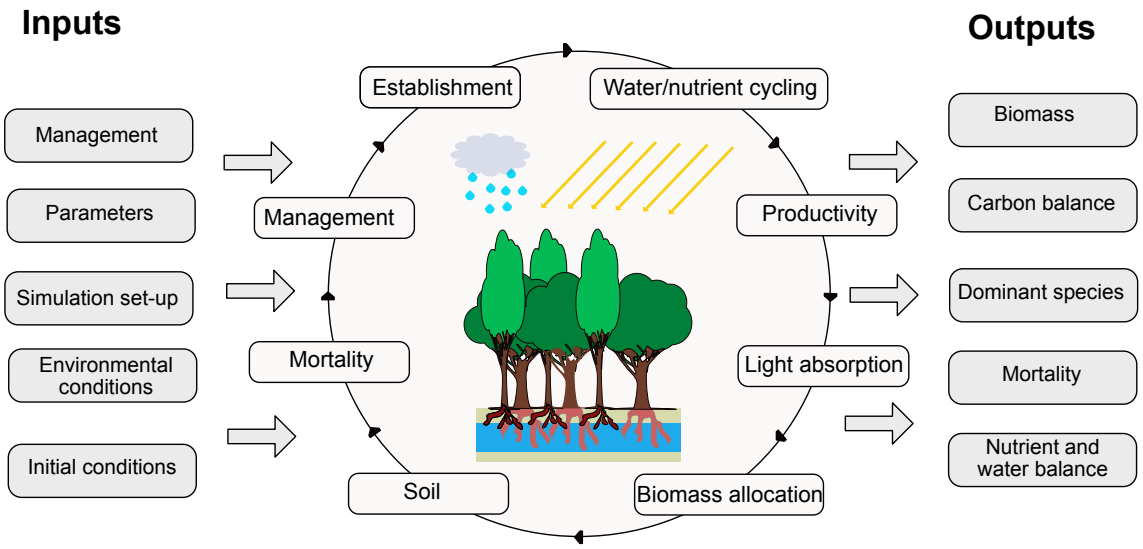


Fig. 2.1: Structure of a process-based forest ecosystem model, which shows common elements of process-based models. Individual models can deviate in their inputs, outputs and represented processes. Figure is inspired by (Trotsiuk et al., 2020b).

2.2 UNCERTAINTIES

In the context of this thesis, we are often interested in the uncertainty of model predictions. Model predictions are a consequence of assumptions (e.g. model structure) and numerical values (e.g. inputs, parameters) chosen by the modeler to represent reality as accurately as possible. Despite all the care, it is usually impossible to construct a perfect model. For example measurements of physical parameters always have a certain error (e.g. Rabinovich, 2006) and processes of complex environmental systems necessarily are simplifications that lead to errors in the predictions (Evans et al., 2013). In principle, any modeling component that affects the predictions can be uncertain and thus contribute to predictive uncertainty. Thus, the most important model components are also the biggest uncertainty contributors. These are: Initial conditions, model inputs, parameters and model structure (Latif, 2011).

2.2.1 *Initial conditions*

Initial conditions are the initial state of the model, i.e. the state of the ecosystem at the beginning of the simulation (e.g. Collins, 2002). Initial conditions uncertainty is especially important for short-term predictions, because for such simulations the internal dynamics do not play a strong role and predictions are more strongly dependent on the initial conditions. In ecosystem modeling studies, initial conditions might be an influential source of uncertainty, when the model is initialized with very uncertain data, e.g. forest distributions derived by remote-sensing data. For studies starting from bare ground or initializing the models with inventory data as most ecosystem modeling studies do, initial conditions uncertainty is less important.

2.2.2 *Model inputs*

For ecosystem models, model inputs (drivers, also called boundary conditions or forcings) are climate conditions (Collins and Allen, 2002). We can split the uncertainty of the climatic conditions into two components: the influence of humans and the intrinsic uncertainty of the climate

(e.g. Schurer et al., 2013). Climatic uncertainty becomes more important the further into the future we want to predict (e.g. Wang et al. 2011). One of the reasons is that climate conditions becomes more uncertain with increasing temporal distance from present time (e.g. Tebaldi and Knutti, 2007).

2.2.3 *Model parameters and model structure*

The last major complex of model uncertainties is the uncertainty of the structure of the implemented processes and their parameters (e.g. Kennedy and O'Hagan, 2001; Keenan et al., 2011a). The model structure generally refers to all structural decisions in the model, in a narrower sense usually the structure of the implemented processes, i.e. the (differential) equations that describe the temporal development of the system. Parameters are variable quantities in the model structure. We discuss these two sources of uncertainty together because parametric and structural uncertainty are connected to each other (Keenan et al., 2011a). For example, when the model structure is changed, parameters need to be updated to retain similar model predictions. Thus, when analyzing predictive uncertainty, we have to deal with model structural uncertainty and parametric uncertainty simultaneously.

Parametric uncertainty is the uncertainty about the correct values of model parameters (see Hartig et al., 2012). One reason why parameters are uncertain is that parameters often do not correspond to directly measurable physical quantities, but even if they do, there are additional sources of uncertainty in parameters of forest ecosystem models. Biological parameters are often not deterministic, but rather have a stochastic component (Kremer, 1983) or depend in a complex manner on other processes or the environment. For example when describing the average response of a species, variations around the average response arise because of intraspecific variation. For all these reasons ecosystem model parameters are uncertain.

Model structural uncertainty is the uncertainty about the correct mathematical (differential) equations and assumptions of the model (e.g. Wieder et al., 2015). For example the correct process and implementation of photosynthesis in trees is unclear and a field of ongoing debate (see Porcar-Castell and Palmroth, 2012). Moreover, empirical studies often concentrate on one aspect of the system, e.g. growth temperature relationships, and not on the entire ecosystem. It is, however, unclear how the empirically derived results of small parts of the entire ecosystem can be put together to correctly describe the entire forest dynamics. Overall, there are many reasons why forest ecosystem models obey structural uncertainty.

2.2.3.1 Hierarchical structure as part of structural uncertainty

Biological data is often hierarchical (Clark, 2005), e.g. the same plant species in different environments. When analyzing such data sets, we have to account for the hierarchy to reflect that observations are not identically and independently distributed (i.i.d.) (e.g. Arnqvist, 2020), which is a basic assumption of most calibration algorithms. Thus, we have to directly describe the hierarchy in the model or calibration algorithm. This does not only apply to forest ecosystem models, but is also important for simple regression models. When using a fixed effect for a grouping variable in a multi-level regression, we assume the groups are independent of each other (e.g. McLean et al., 1991). Conversely, we can use a random-effect and thereby express that groups are realizations from an underlying distribution. Thus, hierarchical modeling is an important issue for all types of modeling.

One characteristic of hierarchical modeling is that there is an information exchange between the different levels of a grouping variable (Dixon, 2016). For example when modeling the same species in different environments, environments with a lot of data are informative for environments with almost no data (McLean et al., 1991). This can be useful if it reflects the biology of the

investigated system, however, when it does not, can lead to a bias in the estimates or posterior distributions. Thus, we have to be careful what calibration assumptions we make.

2.3 NUMERICAL METHODS TO UNDERSTAND AND IMPROVE FOREST ECOSYSTEM MODELS AND QUANTIFY UNCERTAINITIES

To understand, develop and improve forest ecosystem models, we have to perform numerical experiments to judge predictions. The most common experiments are a sensitivity analysis, an uncertainty analysis and a model calibration. In the following subsections we will introduce those methods, their underlying rationalities and challenges.

2.3.1 *Sensitivity analysis*

Although there is no strict consensus about an unambiguously definition of a sensitivity analysis, the common definitions agree that it relates variations in model outputs to variations in model inputs and thereby is agnostic about different magnitudes of uncertainty in different model inputs (Jørgensen and Bendoricchio, 2001). Our definition throughout this thesis is that a sensitivity analysis calculates the change in model output per unit or percentual change of the respective input (see Fig. 2.2).

A distinction is made between local and global sensitivity analysis. The former looks at the sensitivities in a small area around the reference parameterisation and as a common practice changes only one parameter at a time (for an example see Hill and Tiedeman, 2007). This means that a local sensitivity analysis requires less model evaluation and can also be implemented for models with long runtimes such as climate models (Pianosi et al., 2016). A global sensitivity analysis calculates the sensitivities in the space of all plausible parameter values and inputs and thus also requires more model evaluations (Iooss and Lemaître 2015). Testing the entire space of plausible parameters has the advantage that we do not need a reference parameterization, i.e. the definition of a "best" parametrization which might be difficult for forest ecosystem models as it assumes the existence of unique parameters fitting all environmental conditions (Pianosi et al., 2016). Overcoming this problem makes a global sensitivity analysis attractive for forest ecosystem models and other models that evaluate sufficiently fast.

Following our definition the goal of a sensitivity analysis is to identify which parameters have strong influence on the model predictions. However, there are also other benefits, e.g. testing the robustness of model outputs to the different model inputs or model structure (assumptions about some processes) (for an example see Paton et al. 2013) or verifying if the model behavior is consistent with modelers' expectations (for an example see Devenish et al., 2012). Overall, a sensitivity analysis offers insights into the model.

2.3.2 *Uncertainty analysis*

Just like for the definition of a sensitivity analysis, for the definition of an uncertainty analysis there is still no clear agreement. Here, we define uncertainty analysis as attributing uncertainty in the model outputs to uncertainty in the model inputs. Following our definitions, uncertainty and sensitivity analysis are technically very similar to each other (see Fig. 2.2).

The benefits of an uncertainty analysis are different from the benefits from a sensitivity analysis. The primary use is identifying and quantifying the uncertainty contributions of different model inputs. This might help to further reduce uncertainty by collecting additional information on these factors (factor priorisation) (Saltelli et al., 2008) and also it might allow to reduce the computational load in a calibration by fixing some parameters, which do not contribute much uncertainty (Saltelli, 2002). Due to these and many more (Saltelli et al., 2019) important benefits, an uncertainty analysis is prescribed in guidelines for impact assessment.

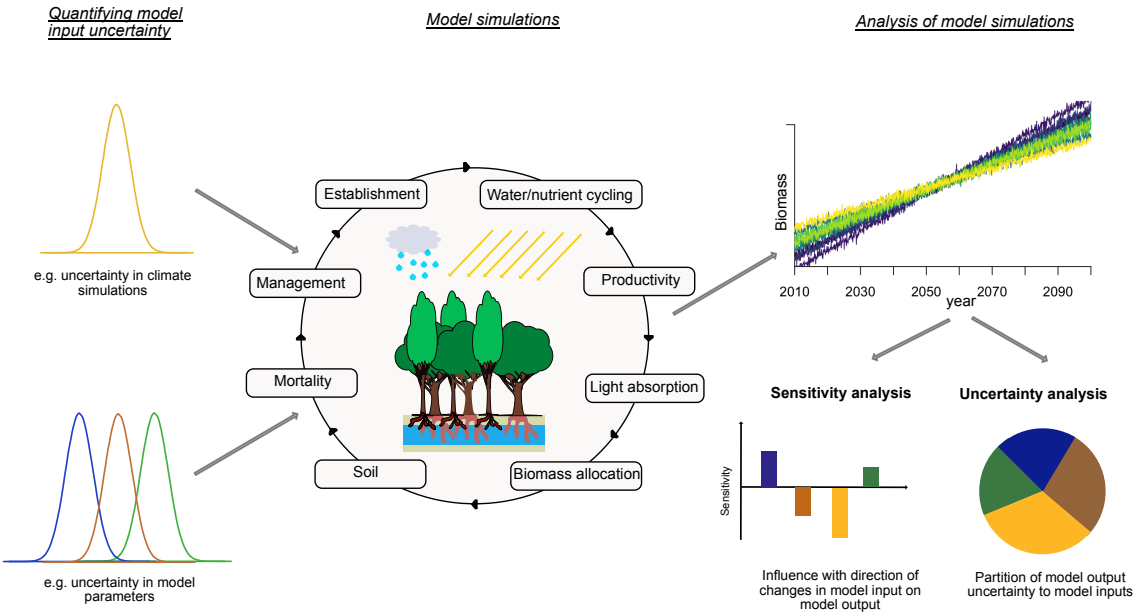


Fig. 2.2: Schematic illustration of the concept of a sensitivity and uncertainty analysis and their relationship to each other.

2.3.3 *Model fitting and likelihood*

To make realistic predictions with forest ecosystem models, we have to compare predictions with data and adjust the model parameters to fit the data (e.g. Van Oijen, 2017). Despite many informal ways to do so, there are mainly two formal methodologies that are used in most practical applications: Bayesian calibration and maximum likelihood estimation. Both methods rely on the concept of a likelihood.

The likelihood as a statistical measure of the goodness of fit (e.g. Gelman et al., 2020) connects the model and data and is proportional to the probability of observing the data given the model and parameters. To do so we have to specify an error, measurement or observational model. The error model describes the statistics of the difference between model and data. This model accounts for all parts of the error that lead to random scattering according to the mathematical model, i.e. it also can account for simplifications leading to random scattering (e.g. Watson et al., 2013).

2.3.3.1 Maximum likelihood estimator

The maximum likelihood estimator (MLE), as the name says, is the set of parameter values that maximizes the likelihood function, usually through numerical optimization schemes. The reason why it is applied in many studies is that it is the best linear unbiased estimator. The inferred parameter values in a frequentist view can be interpreted as the most likely parameters. To judge if these parameters are significantly different from a reference parametrization, additional numerical operations are required as the likelihood is only proportional to the probability of the data given the model and parameters. Moreover, with additional numerics, we can estimate uncertainties in these parameters, usually expressed as the confidence intervals.

2.3.3.2 Bayesian Calibration

The aim of Bayesian calibration (also called model inversion, or inverse modeling; e.g. Hartig et al., 2012) is to infer the plausible ranges for the model parameters from information about the observed outputs and thereby quantify the parametric uncertainty (McElreath, 2016). Bayesian

calibration allows us to update our prior information in the light of evidence into posterior information by using Bayes theorem. To do so, we treat unknown parameters as random variables describing the uncertainty about the "correct" parametrization under our imperfect knowledge (McElreath 2016). Model calibration can also be interpreted as a filter: the uncertainty determines the possible range of a certain parameter. By matching the model predictions generated from this parameter with observed data, we can determine the likely parameter range and discard parameter values that do not generate outputs that match the data.

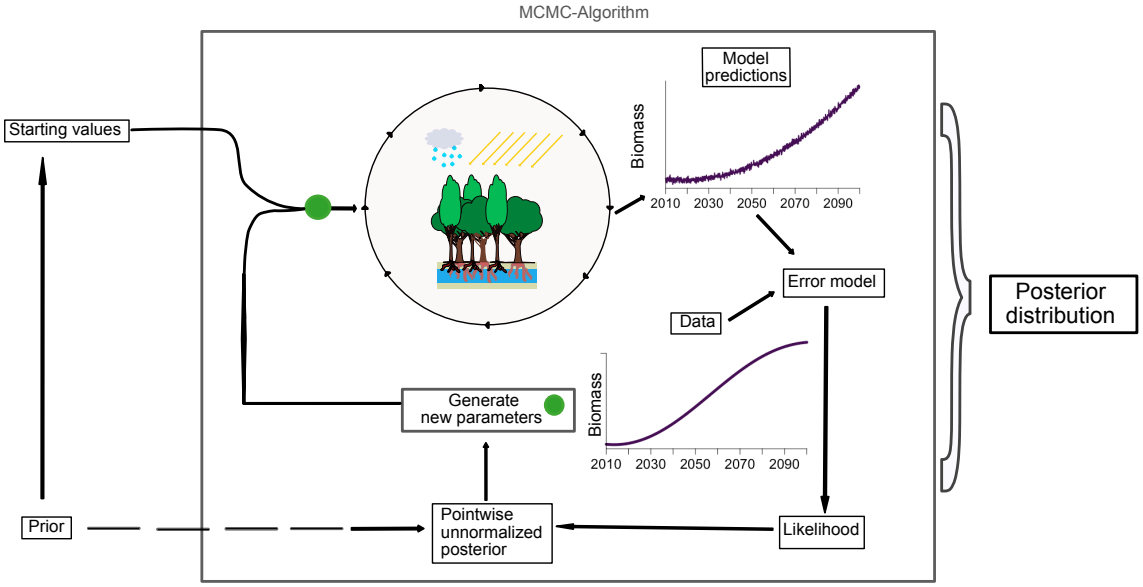


Fig. 2.3: Workflow of a Bayesian Calibration with a Markov Chain Monte Carlo sampler. We start with the prior and generate starting values, then we run the model, compare model predictions with the data through an error model, which results in the likelihood. Given this likelihood, we calculate pointwise the unnormalized pointwise posterior using Bayes theorem and the prior. We then generate new parameters. We repeat this until the parameter chains have converged, which represent the posterior distribution.

2.3.3.2.1 PRIOR

Prior information is our information about the parameters without looking at the data and can stem from former calibrations, other measurements or experiments, expert knowledge or physical constraints. In principle, the prior should capture the important model characteristics. Thus, when feeding prior samples into the model there should be some combinations near the data. The absence of such simulations in these prior predictive checks indicates that priors might be invalid (e.g. Gelman et al., 2020).

2.3.3.2.2 MCMC ALGORITHMS AND POSTERIOR DISTRIBUTION

Markov chain monte carlo methods (MCMC) are methods for sampling from probability distributions and are the working horse of Bayesian inference with system models (Pullen and Morris, 2014). The reason that in Bayesian inference we want to sample from a probability distribution is that we can describe the posterior by the Bayes theorem, but calculating it is intractable for many models (e.g. Andrieu et al., 2003). MCMC algorithms (see Fig. 2.3) can sample from the posterior distribution and thus approximate the distribution with many samples. This is achieved by constructing new parameters given some "rules" from the old parameters. The resulting stationary distribution of this so constructed Markov chain equals the posterior (e.g. Green et al., 2015).

After convergence of the MCMC algorithms (e.g. measured via Gelman-Rubin diagnostics; Gelman and Rubin, 1992), we obtained the posterior distribution. Following Speagle (2020), there are three main benefits of the posterior. First, the posterior distribution allows us to first make educated guesses about the calibrated parameters and their correlations. Second, it provides constraints for parameters and predictions. Third, it allows us to generate predictions with associated uncertainties. Thus, having the posterior distribution allows us to basically calculate all model parameter related quantities and their uncertainties.

2.3.3.2.3 COUPLING FOREST ECOSYSTEM MODELS WITH BAYESIAN CALIBRATION

Most forest ecosystem models are written in low level languages like C++ (e.g. LPJ-GUESS, see Smith et al., 2001) or FORTRAN (e.g. BASFOR), which do not support Bayesian calibration algorithms. Thus, when we want to calibrate a forest ecosystem model, we have to either change parameters by hand, use existing model wrappers to languages like R or write a wrapper by ourselves.

During my PhD I improved a LPJ-GUESS wrapper that was specifically written for the 3.0 version of LPJ-GUESS (Bagnara et al., 2019), so that it is now suitable for all LPJ-GUESS versions. These improvements included a scanning of the instruction file to determine necessary drivers and parameters. Based on these files, it generates templates, which can then be manipulated to run LPJ-GUESS with another parametrization. Additionally, it automatically extracts the default parametrization and default inputs for a better comparison of old and new parameters. Due to these improvements, numerical experiments for model developments and evaluations of all LPJ-GUESS model versions are now possible.

3

CLIMATE AND PARAMETER SENSITIVITY AND INDUCED UNCERTAINTIES IN CARBON STOCK PROJECTIONS FOR EUROPEAN FORESTS (USING LPJ-GUESS 4.0)

Status: 2nd Revision *Geoscientific Model Development*

Authors: Johannes Oberpriller, Christine Herschlein, Peter Anthoni, Almut Arneth, Andreas Krause, Anja Rammig, Mats Lindeskog, Stefan Olin, Florian Hartig

Author Contributions: JO and FH conceived and designed the study and wrote a first draft. JO implemented the case studies, ran the experiments, and analyzed the results. All authors contributed to discussing and interpreting the results, and to the preparation of the manuscript.

ABSTRACT Understanding uncertainties and sensitivities of projected ecosystem dynamics under environmental change is of immense value for research and climate change policy. Here, we analyze sensitivities (change in model outputs per unit change in inputs) and uncertainties (changes in model outputs scaled to uncertainty in inputs) of vegetation dynamics under climate change, projected by a state-of-the-art dynamic vegetation model (LPJ-GUESS v4.0) across European forests (the species Picea abies, Fagus sylvatica and Pinus sylvestris), considering uncertainties of both model parameters and environmental drivers. We find that projected forest carbon fluxes are most sensitive to photosynthesis-, water- and mortality-related parameters, while predictive uncertainties are dominantly induced by environmental drivers and parameters related to water and mortality. The importance of environmental drivers for predictive uncertainty increases with increasing temperature. Moreover, most of the interactions of model inputs (environmental drivers and parameters) are between environmental drivers themselves or between parameters and environmental drivers. In conclusion, our study highlights the importance of environmental drivers not only as contributors to predictive uncertainty in their own right, but also as modifiers of sensitivities and thus uncertainties in other ecosystem processes. Reducing uncertainty in mortality related processes and accounting for environmental influence on processes should therefore be a focus in further model development.

3.1 INTRODUCTION

Terrestrial ecosystem models have emerged in the last three decades as a central tool for decision making and basic research on vegetation ecosystems (Cramer et al., 2001; Fisher et al., 2018; IPCC, 2014; Smith et al., 2001; Snell et al., 2014). Projections from different vegetation models, however, often disagree on important details, for example regarding the observable past (Bastos et al., 2020) or the future carbon uptake of forest ecosystems (Huntzinger et al., 2017; Krause et al., 2019). Among the possible reasons for such differences is the uncertainty in climate scenarios (Saraiva et al., 2019), model structural uncertainty (Bugmann et al., 2019; Oberpriller et al., 2021b; Prestele et al., 2016), initial condition uncertainty (Dietze, 2017a) as well as uncertainty about the model parametrization (Grimm, 2005), which in turn make models' projections themselves uncertain (Dietze, 2017b). It is widely appreciated that understanding which exact factors drive these uncertainties is of immense value for directing research (Tomlin, 2013), but also to interpret and understand projections (Dietze et al., 2018). For example, the IPCC started in its Fifth Assessment Report to systematically analyze uncertainties and attribute them to model inputs (IPCC, 2014) similar to other predictive sciences (e.g. nuclear reactor safety Chauliac et al., 2011; energy assessment for buildings Tian et al., 2018; or policy analysis Maxim and Sluijs, 2011).

The two main tools to propagate uncertainties in model inputs (drivers, parameters, and model structure) to model outputs are sensitivity analysis (SA) and uncertainty analysis (UA) (Cariboni et al., 2007; Caswell, 2019; Saltelli, 2002; Saltelli et al., 2008). The key difference between these two methods is that an UA considers the magnitude of uncertainty in the model inputs (e.g. parameters, typically determined via expert elicitations and previous studies (Matott et al., 2009), while a SA is agnostic about the magnitudes of uncertainty in different inputs, and simply calculates the change in the output per unit or percentual change of the respective input (Jørgensen and Bendoricchio, 2001). This difference aside, both methods share the goal of identifying inputs with a high influence on model outputs, with the underlying idea that better constraining these will increase robustness and reliability of model projections (Balaman, 2019).

Although the benefits for understanding model behavior and predictive uncertainties are obvious, relatively few SAs and UAs have been applied to complex ecosystem models and especially the widely used dynamic global vegetation models (DGVMs) that project terrestrial ecosystem responses to climate change or land management (see, e.g., Courbaud et al., 2015; Cui et al., 2019; Huber et al., 2018; Reyer et al., 2016; Tian et al., 2014; Wang et al., 2013). A reason for this is arguably the complex structure of most DGVMs (Fer et al., 2018), which makes SAs and UAs computationally demanding and difficult to interpret, especially when performing state-of-the-art global SAs and UAs that compute sensitivities and uncertainties across the entire parameter space (Saltelli et al., 2008) rather than just locally around a reference parameter set (see, e.g., Hamby, 1994). Moreover, several studies highlight that sensitivities and uncertainties of DGVMs also exist with respect to environmental drivers (Barman et al., 2014; Wu et al., 2017; Wu et al., 2018), especially solar radiation (Barman et al., 2014; Wu et al., 2018), temperature (Barman et al., 2014) and precipitation (Wu et al., 2017), and it is reasonable to expect that there can be interactions between parameter and environmental sensitivities, meaning that certain parameters are more sensitive in some environments than in others. It therefore seems important to investigate parametric sensitivities in conjunction with their environmental sensitivities in one combined analysis.

In this study, we concentrate on a well-established and widely applied DGVM, the Lund-Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) (Gerten et al., 2004; Sitch et al., 2003; Smith, 2001). Three previous SAs or UAs for the LPJ family identified the intrinsic quantum efficiency of CO_2 uptake (alpha_C3) and the photosynthesis scaling parameter (from leaf to canopy) (alpha_a) as the main contributors of sensitivity for net primary production (NPP) (about 50-60% of the overall sensitivity, Zaehle et al., 2005; Pappas et al., 2013) or foliage projective cover (Jiang et al., 2012). Additionally, these previous studies show that LPJ-GUESS projections of

NPP and vegetation carbon pools showed high sensitivity to tree structure-related (sapwood to heartwood turnover rate, longevity of trees, Pappas et al., 2013; Wramneby et al., 2008; Zaehle et al., 2005), establishment-related (maximum sapling establishment rate, minimum forest floor photosynthetically active radiation for tree establishment, Jiang et al., 2012; Wramneby et al., 2008; Zaehle et al., 2005), mortality-related (threshold for growth suppression mortality, (Pappas et al., 2013) and water-related parameters (minimum canopy conductance not associated with photosynthesis, maximum daily transpiration Pappas et al., 2013; Zaehle et al., 2005). Regarding uncertainties, strong impacts on LPJ-GUESS projections of NPP and vegetation carbon pools (FPC for Jiang et al., 2012) were found for photosynthesis related parameters (Jiang et al., 2012; Zaehle et al., 2005), but also for water-related (minimum canopy conductance not associated with photosynthesis, Zaehle et al., 2005) as well as structure-related parameters (tree leaf to sapwood area ratio, crown area to height function Jiang et al., 2012), whereas soil hydrology parameters were not identified as very sensitive in earlier studies.

Since the publication of these studies, however, the structure of the LPJ-GUESS model changed substantially. The most important changes are the inclusion of the nitrogen cycle (Smith et al., 2014) and new management modules (Lindeskog et al., 2021). Since these changes, no study has systematically examined how model sensitivities and uncertainties were affected by the new model structure. Moreover, previous SAs and UAs ignored management parameters, which, however, are expected to have large impacts on carbon pools and fluxes (Lindeskog et al., 2021).

A further limitation of most previous studies for LPJ-GUESS and other models (e.g. Mäkelä et al., 2020) is that they either analyzed sensitivities and uncertainties to parameter changes, or to changes in the environmental drivers, but not both. As discussed earlier, however, there are good reasons to expect that the sensitivity of parameters will change if environmental drivers change. Given that previous sensitivity analyses used different choices for these boundary conditions (different sensitivities for the climate scenarios and sites in Jiang et al. (2012); for different elevations in Pappas et al. (2013); different sites in Wramneby et al. (2008)), this not only limits the comparability between studies, but also questions the generality of the results for all climatic conditions. Only Jiang et al. (2012) combined parameter and driver sensitivities, but used for the latter only a number of fixed climate scenarios instead of a range of possible values, which prohibits a systematic joint analysis. Moreover, it would be interesting to compare the relative importance of drivers and parameters for the predictive uncertainty of model simulations and how these change between environmental zones (here we use the classification of Metzger et al. (2005)) and thus on an environmental gradient. When sensitivities or uncertainties of parameters belonging to a specific process increase on an environmental gradient, this indicates that the process itself becomes more important on the gradient (Saltelli, 2002). By comparing such changes to existing ecological hypotheses, we can test if model sensitivities and thus process descriptions are in line with ecological expectations.

To answer these questions, we analyzed sensitivities and uncertainties in LPJ-GUESS for 200 randomly distributed sites across Europe (see Supporting Information S1, Fig. S1.1). We address the issue of interactions between environmental and parametric sensitivities by simultaneously investigating uncertainty in environmental drivers (precipitation, temperature, solar radiation, CO_2, nitrogen deposition) with parametric uncertainty in the most important processes (photosynthesis, establishment, nitrogen, water cycle, mortality, disturbance/management, and growth) for dynamic climate change from 2001-2100 and steady climate from 2100-2200. We simulated the most abundant tree species in Europe (Fagus sylvatica, Pinus sylvestris and Picea abies) individually and in mixed stands, as these species are suffering from climate change (e.g. Buras et al., 2018; Walentowski et al., 2017) and could benefit from mixed stands (e.g. Pretzsch et al., 2015). To test climate change impacts, we randomly sampled climate projections within the boundaries of RCP2.6 and RCP8.5. Thereby, our key objectives were to understand the sensitivities and uncertainties of LPJ-GUESS due to environmental drivers and parameters. We were especially interested in 1) overall sensitivities and uncertainties across European forests, 2) uncertainties per

environmental zone and 3) uncertainties on a temperature gradient. Moreover, we investigated, 4) if and how environmental conditions change the uncertainties of environmental processes.

3.2 METHODS AND MATERIAL

3.2.1 *The LPJ-GUESS vegetation model*

LPJ-GUESS is a process-based ecosystem model that simulates vegetation growth, vegetation dynamics and biogeography as well as biogeochemical (e.g. nitrogen and carbon) and water cycles (Lindeskog et al., 2013; Olin et al., 2015; Smith et al., 2014). Ecosystem dynamic processes in the model include establishment, growth, mortality, and competition for light, space and soil resources. To simulate these processes, the model combines time steps on different scales from daily (e.g. phenological and photosynthesis processes) to yearly (e.g. allocation of net primary production to tree carbon components) basis. LPJ-GUESS includes forest gap dynamics succession of cohorts (each represented by an average individual) of different plant functional types (PFTs) or species. Each PFT/species has a unique parameter set.

In this study, we use a model version that was slightly modified from Lindeskog et al. (2021), which is based on the LPJ-GUESS 4.0 version, with a re-parameterization for spruce (Picea abies), pine (Pinus sylvestris) and beech (Fagus sylvatica) (see Supporting Information S1, Fig S1.2 for Pin. syl. and Pic. abi.). To account for the stochastic components of establishment, mortality and patch destroying disturbances, LPJ-GUESS simulates several replicate patches (25 for the simulation with the reference parametrization and 1 for each simulation in the SA and UA) representing "snapshots" of the grid-cell. In this model version, fire is based on the BLAZE model (Rabin et al., 2017). Thereby annually burned area is generated based on fire weather and fuel continuity and distributed to monthly intervals based on climatology (Giglio et al., 2010). Tree mortality is then estimated by computing firelines based on weather and converted into height-dependent survival probabilities (see Haverd et al., 2014) depending on empirical biome specific parameters.

A first set of key parameters from our expert elicitation (see below) for **establishment** are the bioclimatic limits (i.e. minimum growing degree days (gdd5min_est), minimum 20-year coldest month (tcmin_est), maximum 20-year coldest month (tcmax_est) and minimum forest photoactive radiation at forest floor (parff_min)), which build the environmental envelope for establishment. Given the bioclimatic limits are fulfilled, at regular intervals new PFTs are established (here: 1 year) given enough space, light, soil water and photoactive radiation at forest floor is available for establishment (Smith, 2001). Moreover, each of our three investigated species has a maximum establishment rate (est_max) (Smith, 2001).

Structure of trees in the model is mainly linked to the simulated growth of trees, which is triggered by allocating all net primary production (NPP) besides a reproduction debt of 10% (reprfrac) to tree components thereby satisfying mechanical (e.g. allometric eq. for the relationship between height and diameter with allometric parameters (k_allom2, k_allom3) (e.g. Huang et al., 1992), the relationship between tree leaf to sapwood area (k_latosa) (e.g. Robichaud and Methven, 1992), the relationship between crown area and height (k_rp) (packing constraint Zeide, 1993), the maximum crown area (crownarea_max) and leaf longevity (leaflong)) and functional balance as well as demographic constraints (Sitch et al., 2003). Each living tissue is assigned a turnover rate transferring sapwood into heartwood (turnover_sap) and leaves (turnover_leaf) and fine roots (turnover_root) to litter. Investment into above and belowground growth is influenced by the resource stress as individuals are competing for light, space, nitrogen and water. Competition for light is determined by the photosynthetic response and light extinction in the canopy. Competition for space (self-thinning) is represented in the model via allometric equations between crown area and stem diameter (Sitch et al., 2003). Competition for nitrogen and water is determined by tree individual demand for nitrogen and water and soil availability of

nitrogen and water and the PFT-specific root profile. Competition between species will favor certain life-history strategies in particular situations, for example shade-tolerant (e.g. Fagus sylvatica and Picea abies) or intermediate-shade tolerant (e.g. Pinus sylvestris) growth responses, and dynamically changing root-to-shoot ratios.

Tree mortality (natural or via harvest) in the model responds to growth efficiency (ratio of annual NPP to leaf area) being too low over a 5-year period, e.g. due to light competition, maximum longevity of a PFT or changes in environmental conditions (e.g. tolerance to drought (drought_tolerance) changes water uptake) exceeding the species suitable range. Light competition is modeled using the foliage projective cover (FPC), defined as the area of ground by foliage directly above it, using Beer's Law (Smith et al., 2011). The resulting shading mortality is distributed proportional to species' FPC growth in the respective year due to their biomass increase. Mortality is modeled inversely proportional to the growth efficiency (with a given species-specific threshold (greff_min), e.g. Waring, 1983). Moreover, negative NPP of a species kills all individuals of the respective cohort. Background mortality probability increases with tree age, reaching one at the maximum longevity (longevity). Mortality has also a stochastic component. Natural disturbances are implemented in the model as process-based wildfires (with a given fire resistance for each species (fireresist)) and as patch-destroying disturbances (e.g. windthrow and landslides) with the same yearly occurrence probability for all patches (inverse of distinterval). Additional mortality arises from forest management activities, determined by thinning intensity (percentage of all trees cut, thinning_intensity) and cutting intervals (cut_interval), which can be set for each species individually. For a more detailed description of the management module and the additional management parameters see Lindeskog et al. (2021).

Nitrogen input is implemented in the model through nitrogen deposition (prescribed) and biological nitrogen fixation. The latter is simulated empirically as a linear function with intercept (nfix_a) and slope (nfix_b) of the five-year averaged actual evapotranspiration (Cleveland et al., 1999). The resulting amount of nitrogen accumulates in the ecosystem equally over the year and directly adds to the available mineral soil nitrogen pool. When nitrogen is in living tissue, a fraction (nrelocfrac) is re-translocated before leaf- and root shedding.

Photosynthesis is modeled as a function of absorbed photosynthetically active radiation, temperature optimum temperature range for photosynthesis determined by pstemp_low and pstemp_high, Larcher, 1983, intercellular CO_2 (i.e. non-water stressed ratio of intercellular to ambient CO_2 (lambda_max)), and canopy conductance thereby considering a species-specific respiration coefficient (respcoeff) (Smith, 2001) and nitrogen availability. The photosynthesis scheme is a modified version of the Farquhar photosynthesis model, but instead of prescribed values for the Rubisco capacity it is optimized for maximum net CO_2 assimilation at the canopy level (Smith et al., 2014).

Water availability for plants is based on precipitation and snowmelt in the two-layer soil hydrology submodule (for details see Hickler et al., 2004; Smith et al., 2001). Vegetation transpiration and evaporation (with a maximum evapotranspiration rate (emax)) from bare ground and leaves reduce water availability as well as runoff from saturated soil (Sitch et al., 2003). Water vapor exchange by the vegetation canopy is calculated on a daily basis within the photosynthesis scheme (e.g. minimum canopy conductance not associated with photosynthesis (gmin)). The water supply and transpirative demand are calculated on a daily basis and converted into a drought-stress coefficient. Given this coefficient, the investment in roots at the costs of leaves is calculated.

3.2.2 *Simulation setup*

We selected 200 study sites (see Supporting Information S1, Fig. S1.1) spatially and environmentally stratified over Europe by applying random stratified sampling (using the R package splitstackshape Mahto, 2019) with longitudinal and latitudinal coordinates as well as mean precipitation, solar radiation and temperature as categories based on IPSL-CM5 Earth System Model CMIP5 (Dufresne et al., 2013) climate data. We chose 200 sites as a compromise between the high computational demand of running LPJ-GUESS multiple times for all sites and a good spatial as well as environmental coverage of Europe. For these sites, we performed simulations for each of the three most common species in Europe (Fagus sylvatica, Pinus sylvestris and Picea abies) as monospecific stands and additionally all three species together as mixed stands.

The simulation period was from 1861 to 2199. To start the simulations with equilibrium C pools and fluxes, we spun up LPJ-GUESS vegetation and soil carbon and nitrogen pools to preindustrial equilibrium by recycling the 1861 to 1900 climate, the 1861 CO_2 concentration (Meinshausen et al., 2011) and nitrogen deposition. For the transient and future simulation runs, we used the bias-corrected monthly IPSL-CM5 Earth System Model CMIP5 (Dufresne et al., 2013). From this data set, we extracted temperature, precipitation, number of wet days per month, and incoming solar radiation from 1861 to 2099 for RCP4.5 as base scenario and RCP2.6/RCP8.5 as lower/upper boundaries for the climate ranges (see below). In addition to these data, monthly nitrogen deposition was extracted from Lamarque et al. (2013) and soil texture data from Batjes (2005). All these driving data had a spatial resolution of 0.5°x 0.5°. We recycled detrended data from 2090-2099 for all environmental drivers except CO_2 and nitrogen deposition and used these as potential stable climates for the 2100-2199 period.

3.2.3 *Selection of parameters and drivers and their ranges*

The a priori selection of the most influential parameters that can be specified in the parameter file and their ranges was based on our expert knowledge (following the SHELF expert elicitation protocol, (see Gosling, 2018)) and a literature review. The resulting eleven (= 33%) parameters common for all species and 22 (= 20%) species-specific parameters (see Table 3.1) were grouped to the specific processes they contribute most to (Table 3.1, Grouping).

From the environmental drivers of the model, we selected incoming solar radiation, temperature, precipitation, atmospheric CO_2 and nitrogen deposition for our analysis. To obtain uncertainties for temperature, precipitation and solar radiation, we calculated the mean deviations of RCP8.5/RCP2.6 to our base scenario RCP4.5 plus/minus one standard deviation as maximal/minimal per site. As the CO_2 data is global and not site-specific, we calculated ranges from the global data set (RCP2.6 as minimum, RCP8.5 as maximum) averaged over time and plus/minus a standard deviation. For nitrogen deposition, we used RCP6.0 as maximum and RCP2.6 as minimum with the same procedure as for the other drivers.

Table 3.1: The model inputs investigated in the sensitivity analysis can be group in a) common parameters b) species-specific parameters and c) drivers. The ranges for the parameters have been determined from experts and literature, default parameter values that changed from Hickler et al. (2012) due to the reparameterization are explained in Supporting Information S1, Table S1.1 .* denotes an averaging over sites.

a) Common Paramaters

Grouping	Parameter	Explanation	Unit	Default Value	Min. Value	Max. Value	Literature sources
Mortality / Management	distinterval	average return time for generic patch-destroying disturbances	year	920	200	1000	-
Nitrogen	nfix_a	First term in N fixation eqn	-	0,102	0,102	0,367	-
Nitrogen	nfix_b	Second term in N fixation eqn	-	0,524	-0,754	0,524	Cleveland et. al, 1999
Nitrogen	nreleasefrac	Fraction of N retranslocated prior to leaf and root shedding	-	0,5	0,1	0,8	-
Photosynthesis/Light	lambda_max	Non-water-stressed ratio of intercellular to ambient CO2 pp	-	0,8	0,6	0,8	Pappas et al. 2013
Structure/Phenology	reprfrac	Fraction of NPP allocated to reproduction	-	0,1	0,05	0,3	-
Structure/Phenology	turnover_root	Rate of fine root turnover	1/year	0,7	0,65	0,75	-
Structure/Phenology	crownarea_max	maximum crown area	mm^2	40	20	60	-
Structure/Phenology	k_allom2	height =kallom2* -diameter ^(kallom3)	-	60	30	80	-
Structure/Phenology	k_rp	crown area = kallom1- ^height^(k_rp)	-	1,6	1,3	1,6	-
Water	emax	Maximum evapotranspiration rate	mm/day	5	2	6	Köstner 2000

b) Species-specific Paramaters

Group	Parameter	Explanation	Unit	Default Value	Min. Value	Max. Value	Default Value	Min. Value	Max. Value	Default Value	Min. Value	Max. Value	Literature sources
-	-	-	-	*Pinus sylvestris*			*Picea abies*			*Fagus sylvatica*			
Establishment	parff_min	Min forest floor PAR for grass growth/tree estab	J/m^2/day	2500000	1500000	3500000	1000000	750000	1600000	1000000	750000	1600000	-
Establishment	gddSmin_est	Min GDD on 5 deg C base for establishment	°C day	500	250	700	350	300	700	1300	1050	1450	-
Establishment	tcmin_est	Min 20-year coldest month mean temp for establishment	°C	-29	-100	-15	-29	-100	-15	-6,5	-8	-5	-
Establishment	tcmax_est	Max 20-year coldest month mean temp for establishment	°C	5,5	-1,0	6	3	-2	6	7	5	8	Schiltbaskli et al 2017
Establishment	est_max	Max sapling establishment rate	1/m^2/year	0,2	0,1	0,25	0,1	0,05	0,2	0,2	0,05	0,25	-
Establishment	alphar	Shape parameter for recruitment-pvr growth rate relationship	-	10	4	15	4	2	5	2	0,8	5	-
Mortality / Management	longevity	Expected longevity under lifetime non-stressed conditions (yr)	year	500	300	900	300	200	1000	400	250	650	-
Mortality / Management	fireresist	fire resistance	-	0,4	0,05	0,7	0,1	0,05	0,8	0,1	0,05	0,8	-
Mortality / Management	cutinterval	Time until trees are cut	year	90	40	140	90	60	120	105	80	140	-
Mortality / Management	greff_min	Threshold for growth suppression mortality	kgC/m^2/yr	0,21	0,07	0,26	0,135	0,03	0,19	0,02	0,001	0,13	Pappas et al. 2013
Mortality / Management	drought_tolerance	Implements drought-limited establishment plus water uptake, from 0: total to 1: not at all drought-limited	-	0,25	0,1	0,4	0,48	0,2	0,65	0,39	0,2	0,49	-
Mortality / Management	thinning_intensity	percentage of treshold crowncoverage that is kept after thinning	-	0,9	0,45	1	0,9	0,5	1	0,9	0,55	1	-
Photosynthesis/Light	respcoeff	Respiration coefficient	-	1	0,8	2,2	1	0,8	2,2	1	0,5	1,5	-
Photosynthesis/Light	pstemp_low	Approx lower range of temp optimum for photosynthesis	°C	10	6,75	15	10	6,75	14	15	8	20	Renwick et al. 2019/Pallandt-Verheijen et al. 2015
Photosynthesis/Light	pstemp_high	Approx higher range of temp optimum for photosynthesis (deg C)	°C	25	16	30	25	16	30	25	20	30	Zhang et al 2014
Structure/Phenology	cton_leaf_min	minimum leaf C:N ratio	-	31,90	27,32	38,37	38,37	31,9	43,16	24,06	22,7	27,19	-
Structure/Phenology	sla	Specific leaf area	m^2/kgC	8,56	7,812	9,3	11,52	8,7	15,1	43,08	28,33	48,23	Mencuccini, M., Bonosi, L., 2001.; Pallandt-Verheijen et al. 2015; Xiao et al. 2006
Structure/Phenology	turnover_sap	Rate of sapwood turnover	fraction/year	0,085	0,05	0,1	0,065	0,04	0,09	0,085	0,05	0,1	-
Structure/Phenology	k_latosa	Tree leaf to sapwood xs area ratio	-	3000	1800	5200	4000	2500	7000	5000	2500	8000	-
Structure/Phenology	k_allom2	height =kallom2* -diameter ^(kallom3)	-	30	15	60	Values from common Parameters			Values from common Parameters			Zhang et al 2014
Structure/Phenology	k_chilb	Coefficient in equation for budburst chilling time requirement	-	100	80	800	100	80	800	600	250	800	-
Water	gmin	minimum canopy conductance not assoc with photosynthesis	mm/s	0,3	0,22	0,38	0,3	0,22	0,38	0,5	0,42	0,58	Pappas et al 2013;

c) Drivers

Environmental Drivers	insol	Mean deviations solar radiation from standard scenario RCP 4.5 per site	W/m^2	RCP 4.5	-63,9*	65,2*	-	-	-	-	-	-	-
Environmental Drivers	temp	Mean deviations temperature from standard scenario RCP 4.5 per site	°C	RCP 4.5	-5,40*	5,82*	-	-	-	-	-	-	-
Environmental Drivers	prec	Mean deviations precipitation from standard scenario RCP 4.5 per site	mm/month	RCP 4.5	-6,18*	6,27*	-	-	-	-	-	-	-
Environmental Drivers	co2	Mean deviations co2 from standard scenario RCP 4.5 per site	ppm	RCP 4.5	-95,4	237	-	-	-	-	-	-	-
Environmental Drivers	ndep	Mean deviations nitrogen deposition from standard scenario RCP 4.5 per site	g/mm^2/year	RCP 4.5	5,30E-07*	-4,22E-07*	-	-	-	-	-	-	-

3.2.4 *Sensitivity analysis and uncertainty analysis*

LPJ-GUESS predicts a substantial number of output variables, which could all be examined regarding their sensitivities and uncertainties. Here, we concentrate on carbon outputs (gross primary production GPP, total standing biomass TSB and net biome productivity NBP), because of forests' role for carbon cycling (Bonan, 2008), their large contribution to the land carbon sink (Pugh et al., 2019) and the economic importance of tree growth for forest owners (Pearce, 2001).

Sensitivities and uncertainties were calculated by Monte-Carlo sampling from the assumed multivariate parameter and climate uncertainty. For the monospecific / mixed simulations, we drew respectively 10.000 / 50.000 parameter and climate combinations randomly from the prespecified uncertainty ranges, and ran the model based on these combinations for each of the 200 sites. Note, that for mixed simulations, for each simulation we individually drew parameter combinations for each species, i.e. the same parameter could be different for different species. In total, this means that 200 x (50.000 + 3 x 10.000) = 16 million LPJ-GUESS simulations were run.

We quantified sensitivity and uncertainty indices by running multiple linear regressions with the model output averaged over time as response, and parameters and drivers as well as their second order interactions as predictors. With 200 sites, each having three monospecific and one mixed stands setup, we overall ran 200 x (3 +1) = 800 linear regressions. This analysis corresponds to a global SA/UA in the context of regression analysis and has been applied to other system models (e.g. Sobie, 2009). The estimated effects from the regression can be interpreted as sensitivities, as the effect of a unit change of the driver on the response (model output) is estimated. By scaling the predictors to the range [-0.5, 0.5], we obtained the corresponding uncertainties. To check whether we missed non-linear effects, we additionally applied a random forest and extracted the variable importance (following Augustynczik et al., 2017, see Supporting Information S1, Chapter 1.3.). To calculate mean sensitivities/uncertainties for each species, we averaged site-specific sensitivities over all sites with an average annual biomass production greater than 2 tC/ha. We have chosen this threshold because smaller values indicate that the environment is not suitable for the species, however, for each site at least one species was able to establish. For the mixed stands, we first averaged the three species-specific sensitivities/uncertainties per site and then averaged over all sites. Mean percentual sensitivities were calculated by dividing by the mean model output, while mean uncertainty contributions were calculated by dividing by the entire uncertainty budget. Thereby, positive values mean that the respective output increases with increasing parameter values, while negative values mean that it decreases.

It is important to note that uncertainties and sensitivities have different interpretations, and which of these two is more relevant strongly depends on the purpose. The calculated percental sensitivities can be interpreted as percentage change in the corresponding output, when changing a parameter value 1% in the prespecified range. The calculated uncertainties per parameter/driver can be interpreted as relative proportion of the overall uncertainty budget coming from environmental drivers and parameters. For scenario-analysis, e.g. comparing different cut intervals of forests, sensitivities provide a direct estimate of the model response, e.g. how much biomass changes when the cut interval is changed. For a comparison of different model forecasts, uncertainties are usually more relevant. If a reduction of uncertainty via a model-data comparison is the purpose, both measures are important, as parameters with high sensitivities can contribute more or less predictive uncertainty, depending on their input uncertainty.

3.3 RESULTS

3.3.1 *Mean sensitivities over Europe*

Regardless of the output variable, LPJ-GUESS was most sensitive to photosynthesis-related parameters (respcoeff, lambda_max), parameters controlling the wood turnover (turnover_sap) and tree allometry (k_rp), water-related parameters (emax), mortality-related parameters (greffmin) and environmental drivers (temperature, CO_2 and solar radiation) (Fig. 3.1). When looking at differences in the strength of sensitivities for different outputs, TSB was most sensitive to the respiration coefficient (respcoeff), the growth suppression mortality threshold (greff_min) and solar radiation while NBP projections showed negative sensitivity to wood turnover rates (turnover_sap) and longevity and positive sensitivity to temperature, CO_2 and the ratio of intercellular to ambient CO_2 (lambda_max). GPP was negatively sensitive to the respiration coefficient (respcoeff), growth suppression mortality threshold (greffmin), tree allometry (k_rp) and temperature and positive to CO_2, solar radiation and the maximum transpiration rate (emax). Establishment and nitrogen showed the smallest sensitivities for all three carbon-related projections (Fig.1). Note also that NBP had higher percentual sensitivities than GPP and TSB.

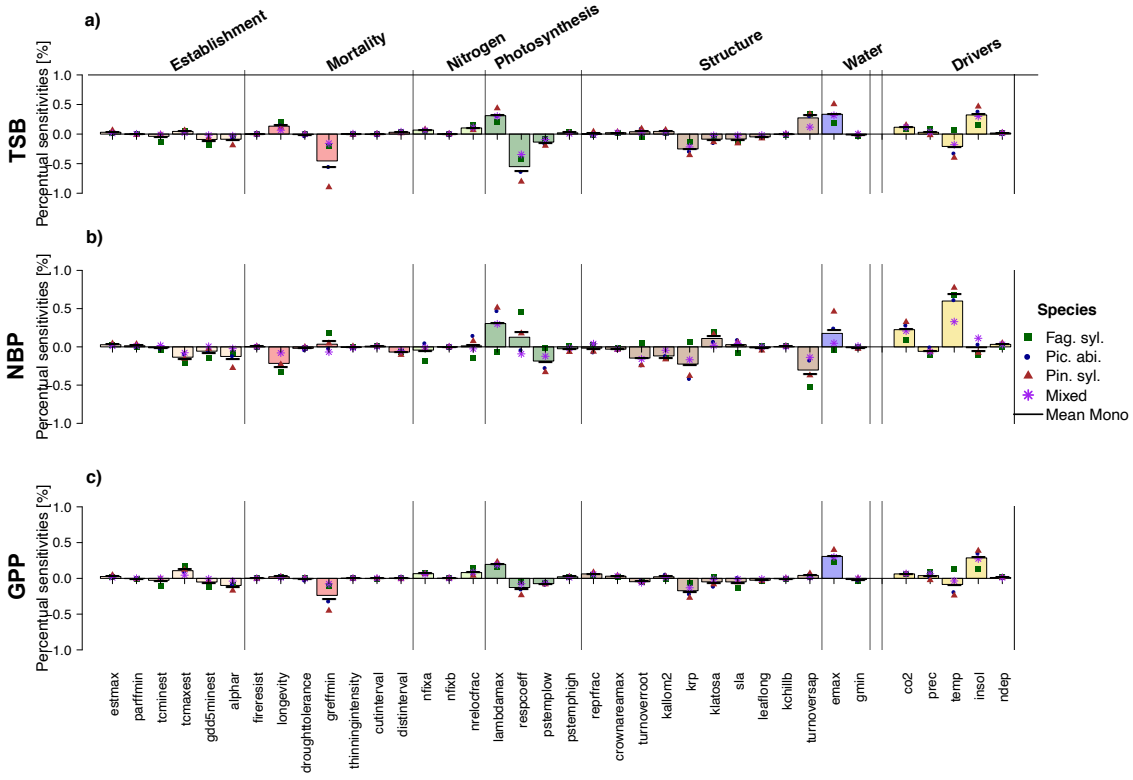


Fig. 3.1: Relative sensitivities (percent output change per percent parameter change) of the individual parameters and environmental drivers regarding a) total standing biomass, b) net biome productivity and c) gross primary production. Sensitivities were not substantially different between Fag. syl. (green squares), Pic. abi. (blue circles) and Pin. syl. (red triangles), but parameter sensitivities were stronger for mono-specific stands than mixed stands (purple asterisks). The height of the bar reflects the mean over mono and mixed stands. Positive values for points and bars indicate a positive and negative values a negative relationship with the corresponding output.

Mixed stands were less sensitive to changes in parameters than mono-specific stands (Fig. 3.1). For monospecific simulations, species sometimes showed different magnitudes and even directions of sensitivities, especially Fag. syl. was more strongly affected by bioclimatic limits and Pin. syl. showed higher sensitivity to environmental drivers (temperature and solar radiation) than the other species. Moreover, TSB and GPP are negatively sensitive to temperature except for Fag. syl. For NBP, the direction of sensitivities changes between species for the non-water-stressed ratio of intercellular to ambient CO_2 (lambdamax), the respiration coefficient (respcoeff), the root turnover (turnoverroot), an allometric constant (krp) and the maximum evapotranspiration rate (emax).

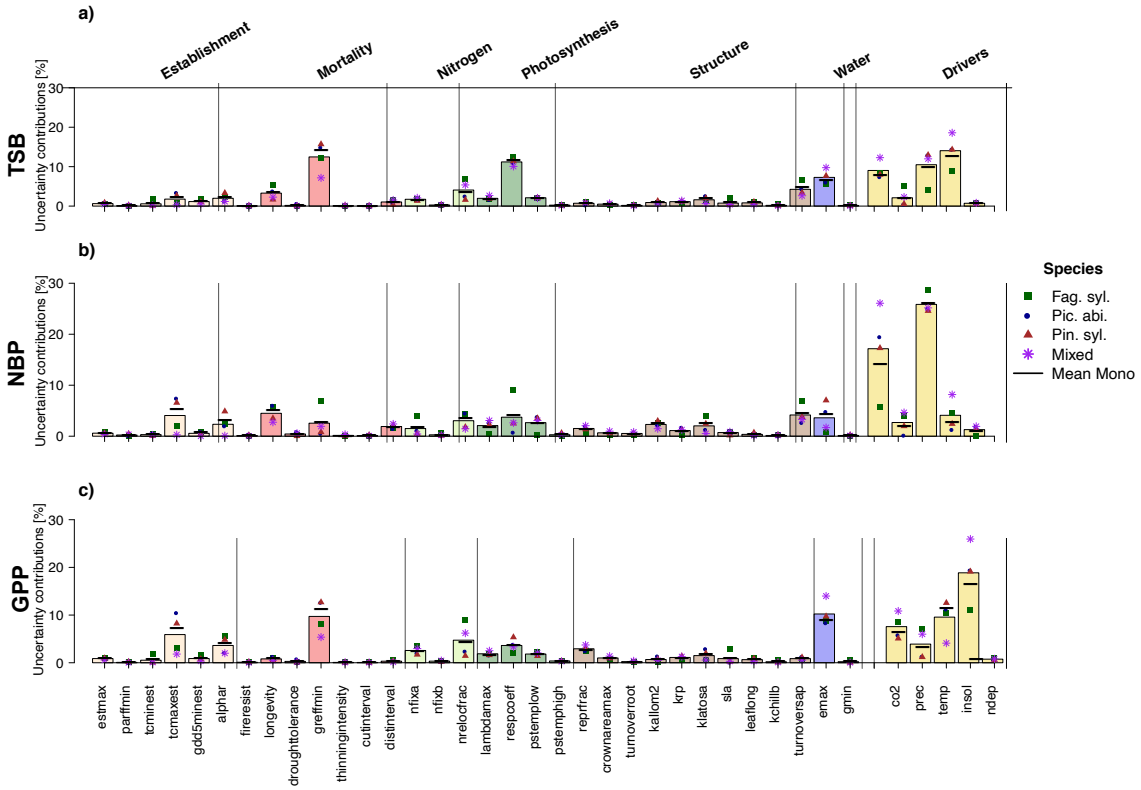


Fig. 3.2: Uncertainty contributions in percent of the individual parameters and environmental drivers regarding a) total standing biomass, b) net biome productivity and c) gross primary production showed no strong differences between Fag. syl. (green squares), Pic. abi. (blue circles) and Pin. syl. (red triangles) and were stronger for mono-specific stands than mixed stands (purple asterisks). The height of the bars reflects the mean over mono and mixed stands. Positive values for points and bars indicate a positive and negative values a negative relationship with the corresponding output.

3.3.2 *Mean uncertainties over Europe*

Looking at uncertainties, we found that environmental drivers contributed most of all processes/drivers to the predictive uncertainty (Fig. 3.2), regardless of the considered model output. For TSB projections, CO_2, solar radiation and temperature contributed substantial uncertainty (Fig. 3.2a). Additionally, large uncertainty contributions arose from growth suppression mortality thresholds (greffmin) and the respiration coefficient (lambda_max). Uncertainty in NBP projections was substantially affected by model parameters (longevity (Mortality process), tcmax_est (Establishment process), turnover_sap (Tree structure process), greffmin (Mortality process) and emax (Water process)), additionally to the high contributions of temperature and CO_2 (Fig. 3.2b). For GPP projections, solar radiation and CO_2 contributed most to climate induced uncertainty, while the threshold for growth suppression mortality (greffmin) and maximum evaporation rate (emax) contributed most to parameter induced uncertainty (Fig. 3.2c). Notably, also nitrogen-fixation induced uncertainty was substantial (7-9%) for TSB and GPP. Most tree structure related parameters except the sapwood to heartwood turnover rate (turnoversap) and the fraction of NPP allocated to reproduction (repfrac) contributed only small uncertainties (Fig. 3.2). Uncertainty contributions analyzed by a random forest are similar to linear regression results (see Supporting Information S1, Chapter 1.3.).

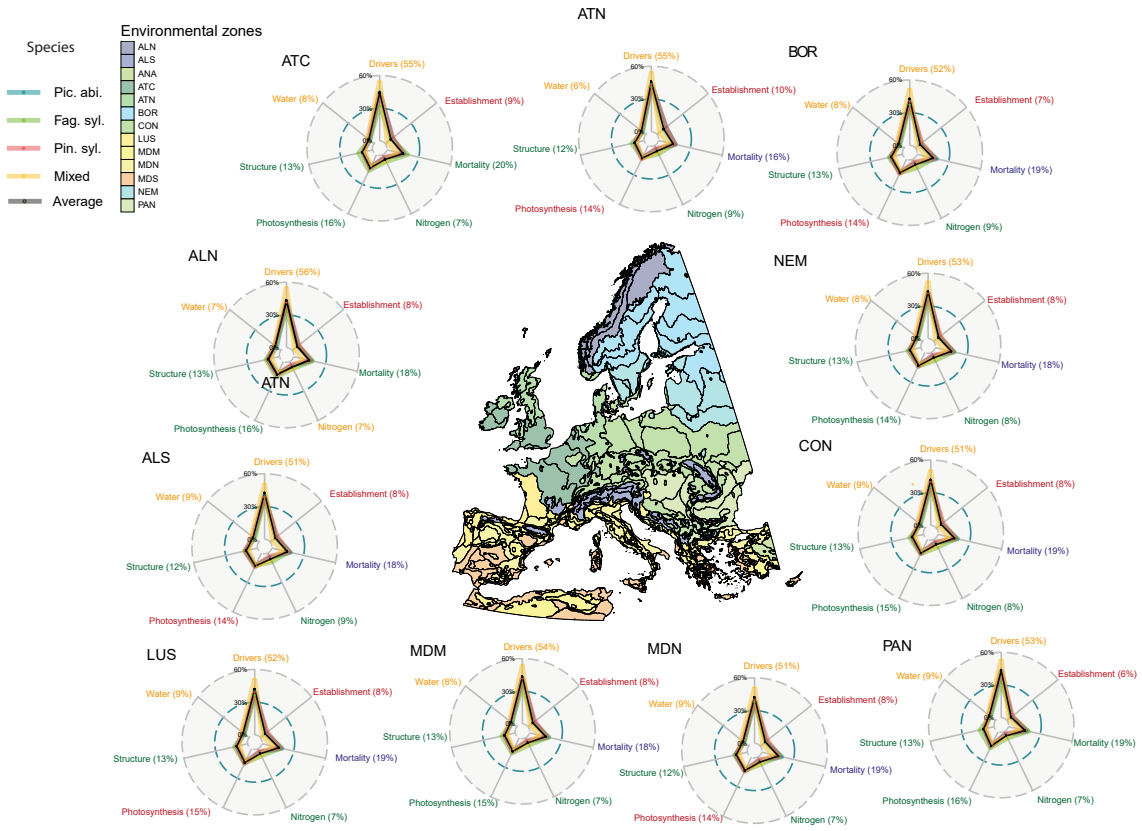


Fig. 3.3: The aggregated relative uncertainties of total standing biomass per environmental zone (with more than five sites) show a higher importance of drivers in the south than in the north. The environmental zones are from Metzger et al. (2005): ALN–Alpine North; ALS – Alpine South; ANA - Anatolian; ATC – Atlantic Central; ATN– Atlantic North; BOR–Boreal; CON–Continental; LUS – Lusitanian; MDM – Mediterranean Mountains; MDN – Mediterranean North; MDS – Mediterranean South; NEM – Nemoral; PAN – Pannonian. For each environmental zones the color and percentage value of the process label indicates which simulation setup (monospecific with corresponding species or mixed) has contributed most uncertainty and how much.

By analyzing uncertainty contributions on a species level, a more diverse picture emerged. Fag. syl. was more affected by temperature and less by solar radiation than the other species. Additionally, we found that uncertainty contributions of environmental drivers were substantially higher for mixed than for mono-specific stands.

3.3.3 *Geographic variation in uncertainties of TSB across Europe*

To project the uncertainties of TSB (for GPP and NBP see Supporting Infromation Fig. S1.4.) into the European environmental space, we filtered stands according to environmental zones, then calculated mean uncertainties per environmental zone and aggregated these per process.

The broad pattern of TSB uncertainty contributions for all three monospecific and mixed stands remains similar in all environmental zones. On average across all environmental zones, stands and species about 45% of the uncertainty was due to environmental drivers, 15% due to mortality-, 14% due to photosynthesis-, 12% due to structure-, 7% due to water- and 7% due to nitrogen-related parameters (Fig. 3.3).

For the individual environmental zones, however, there were subtle differences. In the Mediterranean mountain (MDN) and Pannonian (PAN) zone, environmental driver induced uncertainty was higher than on average especially for monospecific stands (Fig. 3.3). In the Boreal (BOR), Atlantic central (ATC), and Atlantic north (ATN) zone, tree structure- related uncertainty increased compared to the average pattern (Fig. 3.3). In the Atlantic central (ATC) and Atlantic north (ATN) zones nitrogen related uncertainty increased for all species and stands (Fig. 3.3).

To examine this spatial pattern further, we investigated the change of uncertainties across a temperature gradient. To this end, we aggregated the uncertainties per site and process/driver and then fitted a linear regression with the process/driver as predictor and the aggregated uncertainties as dependent variables.

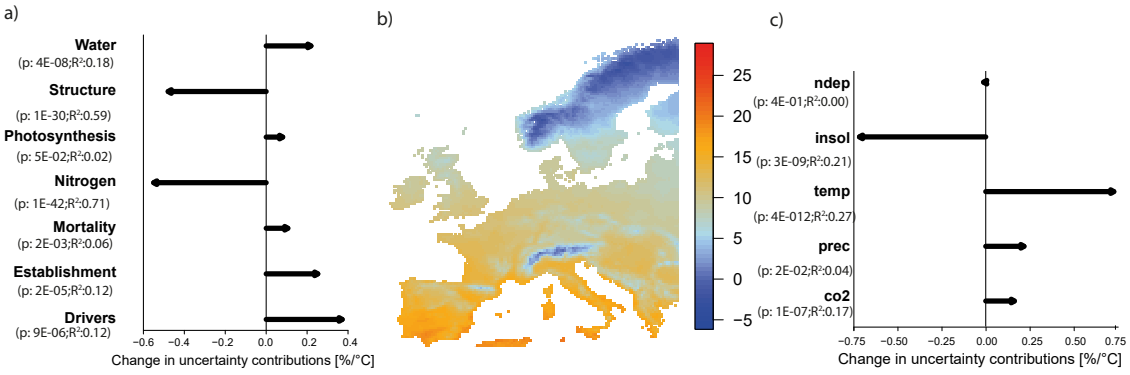


Fig. 3.4: The uncertainty contributions to total standing biomass projections of parameters and environmental drivers change across a mean annual temperature gradient across Europe from north to south (with p-values and R^2 for the processes/drivers). With increasing temperature, the importance of drivers and establishment became higher for total standing biomass, while the uncertainty contributions from nitrogen and structure declined (a). The uncertainty contributions due to temperature increased on the temperature gradient and the contributions from solar radiation decreased (c).

For TSB, we found that increasing mean annual temperature increased the uncertainty contributions of environmental drivers, water- and establishment-parameters, while the uncertainty due to nitrogen- and tree structure- related parameters decreased (Fig. 3.4a). Thereby, the uncertainty contributions of environmental drivers (≈ 0.4%/°C) increased the most (measured in percentage points per °C) and uncertainty contributions of nitrogen fixation decreased most (≈ -0.5%/°C). Mortality and photosynthesis stayed approximately constant on the gradient (Fig. 3.4b).

Looking in more detail at the environmental drivers, temperature (≈+0.75%/°C) as well as CO_2 (≈+0.2%/°C) and precipitation (≈+0.25%/°C) induced uncertainty increased with mean annual temperature, while the uncertainty contribution of solar radiation (≈-0.75%/°C) decreased with mean annual temperature (Fig. 3.4c). Nitrogen deposition induced uncertainty contributions stayed approximately constant on a mean annual temperature gradient.

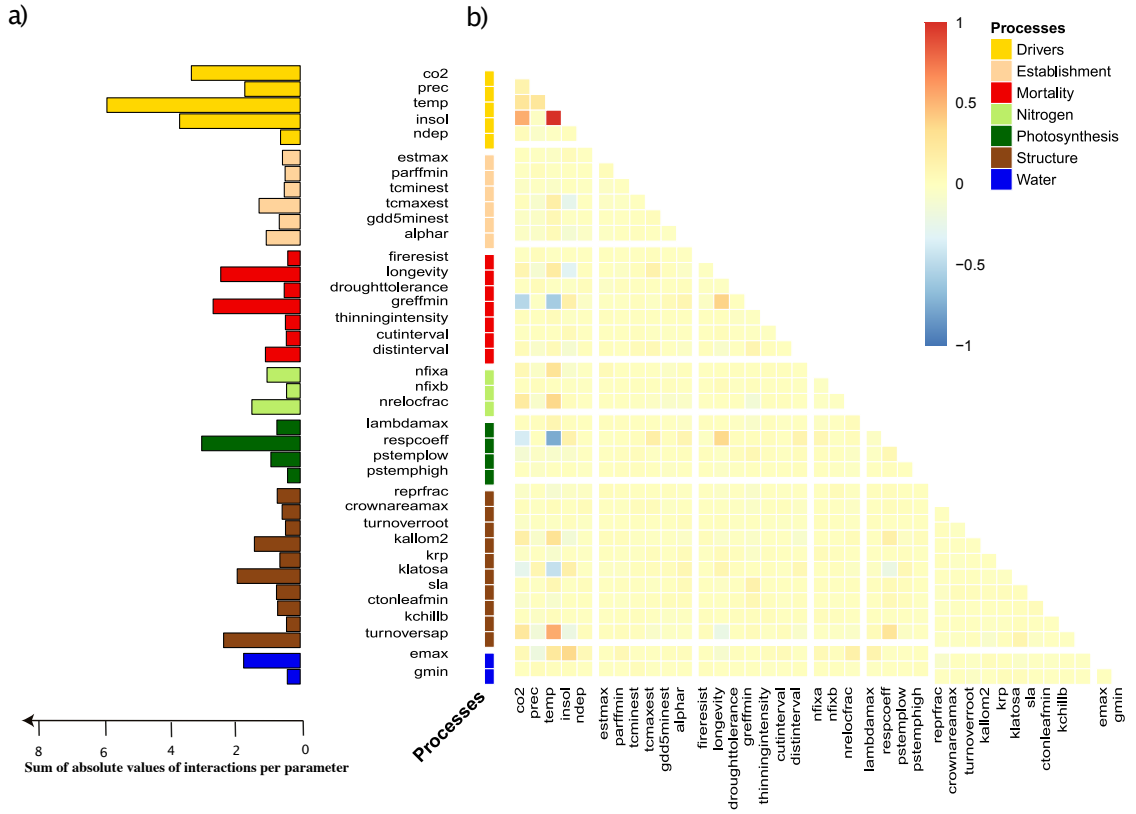


Fig. 3.5: The induced uncertainty of environmental drivers, mortality- and photosynthesis-related parameters changed the most depending on other parameters (a). Strong individual interactions between parameters and environmental drivers in monospecific projections of total standing biomass were rare (b). If strong interactions occurred, these were mainly between two environmental drivers or environmental drivers and parameters and only rarely between two parameters (b).

The above geographical and correlative observations of changing uncertainties across Europe receive further support when looking at the interactions between uncertainties of different drivers/parameters (Fig. 3.5). Interaction indices were calculated by averaging the interactions found in the linear regression over all sites and species (Fig. 3.5b). Moreover, to investigate the overall influence on other parameters or drivers we summed the absolute individual interaction indices

of each parameter with each other (Fig. 3.5a).

We found that environmental drivers (temperature, solar radiation, CO_2 and precipitation) had the highest sum of interactions for TSB (Fig. 3.5a). Moreover, the respiration coefficient (respcoeff), the growth suppression mortality threshold (greffmin), longevity, the sapwood to heartwood turnover rate (turnover_sap) and maximum evaporation rate (emax) had a lower, but still high sum of interactions (Fig. 3.5a). Establishment and nitrogen related parameters had only a few weak interactions (Fig. 3.5). Strong interaction effects occurred mostly with environmental drivers (Fig. 3.5b). A main part of these interactions was between the different environmental drivers themselves (solar radiation - CO_2 and solar radiation - temperature). Additionally, we found interactions of parameters and environmental drivers (temperature-sapwood to hardwood turnover (turnover_sap), temperature – threshold for growth suppression mortality (greffmin) and temperature-respiration coefficient (respcoeff) (Fig. 3.5b) and moderate parameter-parameter interactions (longevity (Mortality process) - greffmin (Mortality process), respcoeff (Water process) – longevity (Mortality process) (Fig. 3.5b). Similar patterns were present for the other two carbon outputs (see Supporting Information S1, Fig. S1.4.).

3.4 DISCUSSION

In this study, we analyzed sensitivities and uncertainties of the LPJ-GUESS vegetation model due to environmental driver and parameter variations across European forests. We found that the model is most sensitive to relative (percentage) changes in photosynthesis-related parameters, structure-related parameters controlling the wood turnover and tree allometry, water-related parameters, mortality-related parameters, and environmental drivers (Fig.1), irrespective of the considered output variable. When considering the different uncertainties (i.e. the entire plausible range) in these parameters and the environmental inputs, we found that environmental drivers and parameters controlling evapotranspiration, background mortality and nitrogen cycling contribute most to predictive uncertainty (Fig. 3.2). When correlated against a temperature gradient and thus geographically from north to south, uncertainty contributions to TSB increased for environmental drivers and decreased for tree structure and nitrogen-related parameters (Figs. 3.3 and 3.4). Interactions between the uncertainty contributions were mainly between different drivers or between model parameters and drivers, whereas only a few parameter-parameter interactions were present (Fig. 3.5).

Our finding that average sensitivities of carbon-related projections across European forests were highest for photosynthesis-related parameters amplifies the evidence from earlier studies (Pappas et al., 2013; Zaehle et al., 2005), although we have used different parameter ranges. In addition, the finding about high sensitivity of LPJ-GUESS to parameters controlling tree structure and especially carbon turnover (turnover_sap) (Fig. 3.1) is in line with results reported for a previous version of LPJ-GUESS (Pappas et al., 2013) and its important role for carbon allocation in trees found in empirical studies (e.g. Aza et al., 2011). The finding that carbon-related projections are very sensitive to mortality-related parameters (greffmin) is also supported by previous studies on the sensitivity of vegetation models and underlines the importance of improving mortality submodules for generating precise projections of vegetation dynamics (Bugmann et al., 2019; Hardiman et al., 2011). Moreover, sensitivities in mixed stands were lower than in monospecific stands for NBP and GPP (Fig. 3.1) (in line with Wramneby et al., 2008). The reason for that imbalance may be that other species can dampen and even benefit from non-optimal life-history strategies of an individual species (Loehle, 2000). Another reason might be, that for mixed simulations we sampled parameters for each species individually, which reduces the influence of each parameter on stand-level carbon projections.

We found that uncertainty contributions of environmental drivers were comparable to the uncertainty contributions of all parameters together (Figs. 3.2 to 3.5, see also Snell et al., 2018,

for the FLMs model; but see Petter et al., 2020, who found that most uncertainty is induced by the choice of the forest model). Especially high uncertainty contributions arose from temperature (negative effect for TSB, GPP positive for NBP), CO_2 (positive effect for all variables) and solar radiation (positive effect for all variables). These results are supported by the earlier studies on the effect of environmental drivers in DGVMs (Barman et al., 2014; Wu et al., 2017; Wu et al., 2018). The positive effect of CO_2 could be explained by increased water-use efficiency and the CO_2 fertilization effect (also found for other DGVMs Keenan et al., 2011b; Galbraith et al., 2010), which in LPJ-GUESS is an emerging property of the formulation of photosynthesis and respiration (see Hickler et al., 2008). However, empirical studies do not find such an effect (Körner, 2006), which could be link to the fact that LPJ-GUESS does not model phosphor cycling which could be the limiting nutrient under increasing CO_2 and no fertilization effect occurs (for a DVGM study see Fleischer et al., 2019). We speculate that the negative effect of temperature (also found for multiple DGVMs, see Galbraith et al., 2010) arises from decreased photosynthetic efficiency and increased respiration rates with higher temperatures (see the empirical study of Gustafson et al. (2018), here confirmed by the negative relationship between temperature and the respiration coefficient). This effect, however, differed in magnitude and direction between tree species (Fig. 3.2) - while there was a strong effect for Pic. abi. and Pin. syl., Fag. syl. was less affected, which could be a sign of its higher resistance to increasing drought (Buras and Menzel, 2019; Tegel et al., 2014; but see Charru et al., 2010). From the parameters, especially water-, nitrogen- and mortality-related parameters contributed a substantial amount of uncertainty. The uncertainty contributions from mortality parameters (see Bugmann et al., 2019, for a variety of DGVMs) and water (Pappas et al., 2013, with different parameter ranges for LPJ-GUESS) were already highlighted by earlier studies.

3.4.1 *Geographical and environmental patterns in sensitivities and uncertainties*

Several of our results suggest that environmental context influences the sensitivity of LPJ-GUESS model parameters. First, we found changing uncertainties across different vegetation zones (Fig. 3.3) and on an environmental gradient (Fig. 3.4) and that most interactions occurred with environmental drivers (Fig. 3.5). Moreover, uncertainty contributions analyzed by a random forest were similar to the linear regression results, but assign higher importance to environmental drivers (see Supporting Information S1, Fig. S1.3). All these findings indicate that environmental context can change the importance of different processes in the model, which is in line with the biological expectation that the environment affects the physiology of organisms directly and thus indirectly the fitness and biotic interactions (e.g. Seebacher and Franklin, 2012; Tylianakis et al., 2008), and that environmental responses can be particularly nonlinear (e.g. Burkett et al., 2005) or show higher order interactions.

Interestingly, our results of decreased uncertainty contributions of structure- related parameters and increased contributions of environmental drivers on the temperature gradient (Fig. 3.4) also seem in line with the stress-gradient hypothesis (Maestre et al., 2009), an empirically-observed pattern which states that in stressful environments, positive interactions should occur more often than in benign environments (e.g. Callaway, 2007). For the ecosystem that we consider, we interpret increasing temperature as increasing stress (e.g. Ruiz-Pérez and Vico, 2020), and structure as the best indicator for competitive interactions as the structure dictates resource allocation (e.g. bigger crown, but identical stem diameter leads to more photosynthesis; more sapwood to heartwood turnover requires less NPP). With this interpretation, one would conclude that under increasing stress, the importance of competition-related parameters decreases in the model, as expected from the stress-gradient hypothesis. We acknowledge that a fair amount of interpretation is needed to arrive at this conclusion, and we do not claim that this result lends evidence to the empirical discussion about the generality of the stress-gradient hypothesis, but we find it noteworthy that such a large-scale pattern emerges in the model from lower-level pro-

cesses, without having been imposed (see Levin, 1992).

3.4.2 *Associated uncertainties of previous changes in model structure and implications for future model development*

The management and the nitrogen cycling module are the most recent improvements of the LPJ-GUESS model (Smith et al., 2014; Lindeskog et al., 2021). Compared to previous sensitivity and uncertainty analysis, the high contributions of the nitrogen fixation to the predictive uncertainty of TSB and GPP (Fig. 3.2 a,c) are novel, though not surprising, as nitrogen is an important factor for the productivity of most temperate and boreal ecosystems (Vitousek and Howarth, 1991). The main reason why few earlier studies report those uncertainties is that vegetation models have only recently begun to integrate nitrogen cycling and limitation (e.g. Smith et al., 2014). The management module showed only small uncertainties, which could be due to the narrow parameter ranges for the cut interval and thinning intensity reflecting typical forest owners' choices. As forest owners usually try to maximize their profits (Johansson, 1986; but see Brazee and Amacher, 2000) and thus biomass production, low sensitivities of the management module are not surprising. A more suitable and important test case and application of the management module is a historical reconstruction of foliage projective cover data or similar outputs of the LPJ-GUESS model.

Our study helps to guide the model application, discussion of uncertainties and model development of LPJ-GUESS and other DGVMs. First, future model applications and model comparisons should focus on mortality as these processes contributes high uncertainties for carbon-related projections (see Figs. 3.1 to 3.3). Thereby, it should be investigated if these uncertainties stem from the intra-specific variability of the parameters itself (Bolnick et al., 2011), parameters are just not identifiable (see Marsili-Libelli et al., 2014), or if a model data comparison could reduce uncertainties in the parameters (e.g. Hartig et al., 2012). Using time series inventory data might help as it is informative for constraining mortality modules (Cailleret et al., 2020). Second, small sensitivities of establishment related parameters are surprising as we know that not all three investigated species can effortlessly establish across all of Europe, e.g. Fag. syl. can only establish on locations with no extreme drought and heat and no extreme winter frosts (Bolte et al., 2007). Thus, either we missed important parameters of this module, or the parametrization of the model needs to be updated. Third, when introducing new processes or coupling with other models (Forrest et al., 2020) calculating interactions helps to get a first impression where these new processes influence other model processes and potentially detect missing links. Moreover, future model applications can interpret their results with regard to the sensitivities in different factors (Saltelli et al., 2019) and discuss uncertainties and the causing factors, when used in policy advice (Laberge, 2013).

3.4.3 *Limitations*

We caution that our results regarding the importance of different factors for predictive uncertainties (but not sensitivities) depend on the a priori defined uncertainty range of the contributing factors (see Wallach and Genard, 1998), as well as on several other technical choices in our study. For determining uncertainty ranges of the drivers, we used RCP scenarios; however, these were not created as probabilistic min / max ranges. For the model parameters, we relied on expert guesses, reducing subjectivity as far as possible by following the SHELF expert elicitation protocol (Gosling, 2018). Future studies could include more experts and their opinion on parameter distributions to reduce variability in this protocol. As the model is sensitive to parameters and environmental drivers, and because these influence each other, we treated them in a combined sensitivity and uncertainty analysis (Saltelli et al., 2019), however, when interpreting it should be kept in mind that the one group relates to uncertainties in the model, while the other is external,

so the two are conceptually very different. A certain ambiguity also arises from the definition of the indicators: here, we calculated sensitivities and uncertainties by capturing only linear components and second-order interactions, and we may therefore miss highly non-linear (and in particular hump-shaped) responses in LPJ-GUESS (Roux et al., 2021). However, our comparison to uncertainties calculated with random forest variable importance, a method that would also capture nonlinearities, did not reveal any qualitative differences in the ranking of parameter importance (Supporting Information S1, Fig. S1.3). Overall, while we acknowledge that a certain amount of subjectivity exists in the choice of input uncertainty and calculation of indices, we believe that our results are quantitatively robust to those choices.

Moreover, we acknowledge that LPJ-GUESS is known to be sensitive to the scaling parameters alpha_a and alpha_C3 (Pappas et al., 2013; Zaehle et al., 2005), which we have omitted from our analysis. These parameters, however, are not accessible in the parameter input file. Instead, they are hard coded in the model's source code and therefore a normal user would not change them. We argue that these parameters should thus be counted towards the more general and here neglected contribution of structural uncertainty (i.e. the uncertainty regarding the functional form of processes or even to entire modules) to the joint model uncertainty. Several previous studies suggest that the sensitivity of vegetation models to structural changes can be large, often larger than to parameters (e.g. Bugmann et al., 2019), and it would certainly be useful (although very complicated) to explore these uncertainties together with the here considered factors in a joint analysis. In the present study, however, we considered only the parameters that would be accessible to normal LPJ-GUESS users, and neglect structural uncertainty that could be explored by changing the source code.

3.5 CONCLUSIONS

Our findings highlight the relative importance of parametric uncertainties in different processes and their interactions with uncertainties in environmental drivers for carbon projections with LPJ-GUESS. Our results demonstrate that environmental context changes uncertainty contributions of other processes across the European environmental gradient. The pattern of decreasing importance of competition towards the warmer areas is in line with the stress-gradient hypothesis, which posits that the importance of competition decreases with increasing environmental stress. Our findings improve our understanding of forest ecosystem models, enable pathways for future ecosystem model development and thus builds a basis for more realistic projections. In the future, parametric uncertainties could be reduced by model-data fusion (e.g. Trotsiuk et al., 2020a) of LPJ-GUESS, concentrating on the parameters contributing most uncertainty in each geographic region (see Fig. 3.3). Reducing uncertainties in the drivers is more difficult. To some extent, environmental drivers are themselves influenced by the vegetation (Strengers et al., 2010), so model-data fusion on a fully coupled model including feedback loops between vegetation and climate, as well as a general improvement of climate models, could reduce driver uncertainty to some degree. Effectively, however, much of the uncertainty in this section arises from potential greenhouse gas emission trajectories, for which a probabilistic assignment is difficult due to their dependency on human decision-making.

4

A SITE-SPECIFIC BAYESIAN CALIBRATION OF A PHYSIOLOGICAL FOREST MODEL SHOWS INTRASPECIFIC FUNCTIONAL VARIATION IN TREE SPECIES ACROSS EUROPE

Status: Submitted to *Ecography*
Authors: Johannes Oberpriller, Volodymyr Trotsiuk, Lukas Heiland, Lisa Hülsmann, Florian Hartig
Author Contributions: JO and FH conceived and designed the study and wrote a first draft. JO implemented the case studies,ran the experiments, and analyzed the results. LHe and JO prepared the data. VT advised regarding 3-PG. All authors contributed to discussing and interpreting the results, and to the preparation of the manuscript.

ABSTRACT Intraspecific variation within plant species plays an important role for their persistence and resilience under environmental stress. Accounting for intraspecific variation may thus be crucial for accurately forecasting ecosystem responses to global and climatic change. So far, however, comprehensive data of functional variation in plant species is scarce, and only few models explicitly account for intraspecific functional variation in their predictions. A possible solution for this problem is to infer functional phenotypic variation across species ranges by fusing process-based physiological models with range-wide observation data. Here, we take this approach and fuse the physiological forest model 3-PG for three European tree species (*Fagus sylvatica* L., *Picea abies* (L.) Karst. and *Pinus sylvestris* L.) to national forest inventory data across Germany and Sweden, using a hierarchical Bayesian approach. In the calibration, we allow for spatial variation in the optimal growth temperature, maximum photosynthesis rate and response to vapor pressure deficit. Our results show evidence for substantial intraspecific variation in model parameters that correlated with environmental conditions and reduced the predictive error on forests biomass. Moreover, looking at model predictions for a hypothetic climate warming scenario, we find that the model calibrated with intraspecific variation differed substantially from a model with a homogenous parameterization. In conclusion, our results show the potential of inferring intraspecific functional variation by using hierarchical Bayesian calibration approaches together with process-based forest or ecosystem models.

4.1 INTRODUCTION

It is widely accepted that most plant species show substantial intraspecific variation in their traits or functions (e.g. Bolnick et al. 2011, Siefert et al. 2015). The origin of this intraspecific functional variation (IV, see Albert et al., 2010), which can be observed via the variation in traits or plant performance across climatic or geographic clines (see e.g. Woods et al., 2012) or in common garden experiments (see George et al., 2017, for an example), may be heritable (i.e. genetic e.g. Savolainen et al., 2007; or epi-genetic, e.g. Bose et al., 2020) as well as plastic (e.g. leaf trait plasticity, see Henn et al., 2018).

The existence of IV has important consequences for the persistence and resilience of species to global change (Des Roches et al., 2018). For example, IV allows species to adapt to and survive under extreme environmental stress (Bijlsma and Loeschcke, 2005; Henn et al., 2018), influences their distribution (Benito Garzón et al., 2011) and can buffer against climate change impacts (Oney et al., 2013; Münzbergová et al., 2017). For example, trees have a lower specific leaf area in warm and dry compared to cold and wet areas (Esperon-Rodriguez et al., 2020). For a more detailed review on the likely consequences of IV on ecosystem resilience, see Moran et al. (2016).

Despite the widely accepted importance of IV for community (e.g. Jung et al., 2010; Gravel et al., 2011; Violle et al., 2012) and range dynamics (e.g. Bestion et al., 2015) in plants, most ecological models essentially treat individuals of a species as a monolithic entity that will look and perform identical across space and time (Moran et al., 2016). This is true both for statistical models (e.g. species distribution models, which assume a constant niche for a species, but see Oney et al., 2013), but also for process-based and eco-physiological models (but see Berzaghi et al., 2020).

One of the main reasons for this "homogeneous species (HS) assumption" is the difficulty to obtain empirical or experimental data on functional variation (e.g. Chevin et al., 2013). For example, physiological parameters such as the optimal growth temperature or the maximum photosynthesis rate are difficult to measure, especially across larger gradients. As in many other parts of functional ecology, ecologists therefore often resort to measuring (functional) traits. Those are assumed to correlate with ecological performance (e.g., diameter and height growth may serve as proxies for fitness Alberto et al., 2013), however, the relationship between such measurable functional traits and plant performance can be complex (Violle et al., 2007) and it has proven difficult to translate measured trait variation into physiological or demographic parameters that could be used by process-based vegetation models (Yang et al., 2018; Berzaghi et al., 2020).

Computational difficulties, i.e. how to represent IV (see discussion below, for proposed modeling attempts see Snell et al., 2014; see Morin et al., 2021, for IV in gap-models), and a lack of suitable data explain why IV is often ignored in modeling studies (Moran et al., 2016), but that does not alleviate concerns about possible inconsistencies of ecological models resulting from the HS assumption. Ignoring IV may also bias forecasts, potentially leading to wrong predictions of community assembly (Jung et al., 2010), functional diversity (Cianciaruso et al., 2009), ecosystem stability (Barabás and D'Andrea, 2016) and range dynamics (Atkins and Travis, 2010; Bocedi et al., 2013). Thus, to make robust projections of forest dynamics, it is of great importance to find operational means to account for intraspecific variation in models that describe tree species' functional responses to climatic variability (Van Bodegom et al., 2012; Sakschewski et al., 2015; Berzaghi et al., 2020).

A possible solution to this problem is to bypass the measurements of traits and infer IV directly by calibrating process-based or eco-physiological models to extensive spatial data. By allowing species parameters to vary in space and time or via relationships with a priori defined environmental factors (for prescribed trait-climate relationships in a DGVM see Verheijen et al., 2013), we directly represent functional IV in the modeling assumptions. Examples for such an approach

are, e.g. Dietze et al. (2008), who fitted variations in allometric equations or Fer et al. (2021), who fitted variations in the SIPNET model to eddy covariance data along 12 sites. Vanderwel et al. (2017) fitted spatial variation in demographic processes and Needham et al. (2018) inferred variations in growth rates. So far, however, there has been no calibration of intraspecific variation with a complex process-based models across the entire range of a species.

With a more complex process-based model, the challenge for conducting such spatially variable model calibrations is to solve both computational and statistical hurdles associated with such an approach. On the computational side, a continent-wide statistical model calibration with locally varying parameters requires considerably more model evaluations than a calibration to a single or a few local sites. On the statistical side, one must find a balance on a flexibility-rigidity gradient between two extremes: Site-specific and joint multisite calibration (Fer et al., 2021). Site-specific calibrations are most flexible assuming species parameters across sites are completely independent of each other (Minunno et al., 2016). This approach may overfit to site-specific data and thus calibrated models often perform poorly when extrapolating (Basler, 2016). The most rigid approach is a joint multi-site calibration assuming no IV and thus species reacting in the same way to environmental change across space and time (Tian et al., 2020). As discussed, however, this assumption will not fit the underlying biological reality, and the resulting structural error can lead to wrong or overconfident predictions (Oberpriller et al., 2021b; Fer et al., 2021). Thus, when calibrating models, it seems important to find model structures of intermediate flexibility allowing parameters to vary, however, without making them too flexible.

There are essentially two options with intermediate flexibility that allow IV: the first is to make prior assumptions about the dependence of model parameters to certain (typically environmental) predictors (see Verheijen et al., 2013, for an example) . The second is to impose spatial or global regularization on model parameters so that site-specific parameter estimates are attracted to the overall species mean (Cressie et al., 2009). Both approaches can be biologically defended, the first by directed environmental selection of species traits, and the second by gene flow, which tends to homogenize species parameters.

Here, we take the second approach, meaning that we implement a calibration that allows for IV in model parameters while site-specific species parameters are at the same time attracted to the overall species mean. Technically, this is implemented using site-specific random effects (Clark, 2005) for selected species parameters in a hierarchical Bayesian model (HBM) (for an example of this approach for allometric equations see Vieilledent et al., 2010). By regressing the fitted site-specific parameters against the site-specific climate, we then test our IV hypotheses (Webb et al., 2010). The approach to test hypotheses while using HBM to calibrate an ecosystem model is summarized in Fig. 4.1.

Using this approach, we calibrate the 3-PG forest model, a physiological model of forest stand development (Landsberg and Waring, 1997; Sands and Landsberg, 2002), with a model accounting for intraspecific variation (IV-model) and a "homogeneous-species" (HS-model, technically a joint multi-site model, i.e. fixed species parameters) model for the three of the most common species in Europe: *Picea abies* (L.) Karst., *Pinus sylvestris* L. and *Fagus sylvatica* L., using national forest inventory data from Germany and Sweden spanning 2171 km of latitudinal gradient and 10.5 °C of a mean annual temperature (MAT) gradient. We hypothesize that in the IV-model (1) optimal growth temperature increases across a temperature gradient, (2) trees close their stomata stronger in dry areas and (3) the maximum photosynthesis rate correlates with drought and temperature.

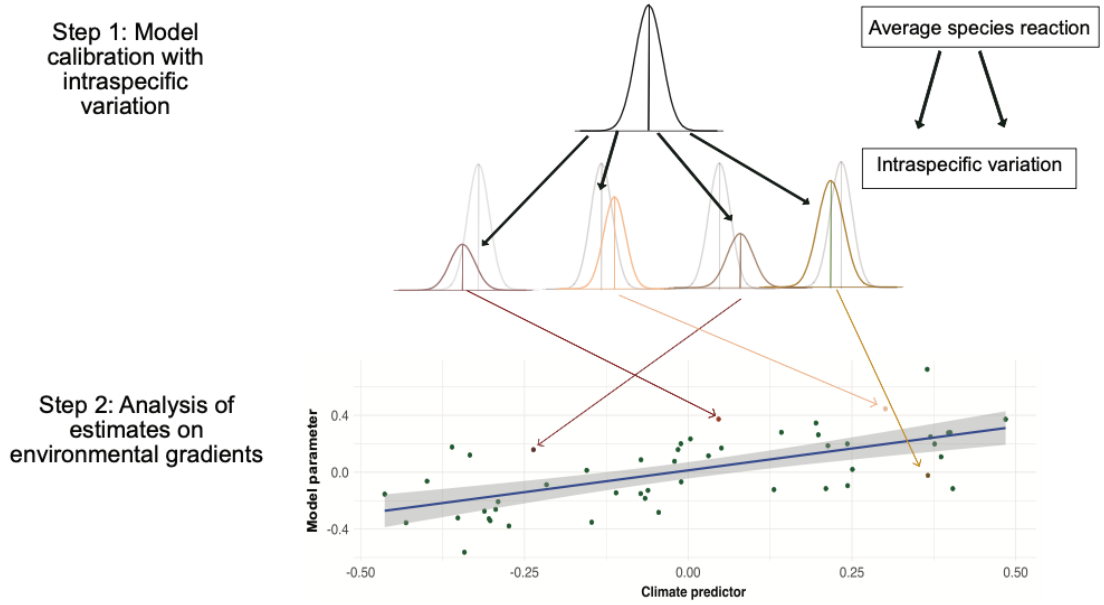


Fig. 4.1: A framework to account for intraspecific variation in forest ecosystem models using hierarchical Bayesian modeling (HBM) and to test ecological hypotheses on environmental gradients. First one uses hierarchical Bayesian calibration to estimate site-specific parameters (Step 1) and second one tests if estimates correlate with climatic variables (Step 2).

4.2 METHODS

4.2.1 *Forest and climate data*

To test for IV across a climatic gradient, we selected three of the most common and widely distributed tree species: *F. sylvatica*, *P. abies* and *P. sylvestris*. We chose these species because of good data availability and because previous evidence for IV in either of them (e.g. Carsjens et al., 2014, for beech, Kapeller et al., 2012, for spruce and; Laforest-Lapointe et al., 2014, for pine).

We aggregated national forest inventory data from Germany and Sweden (for a description see Lawrence et al., 2010; Gschwantner et al., 2016) covering a large climatic gradient. In both countries, three forest inventories were recorded (Germany: 1986-1989, 2001-2003, 2011-2012; Sweden 2001-2002, 2009-2010, 2013-2015). From the data, we sampled 245 monospecific, unmanaged, and undisturbed (no tree died between first and last inventory) sites geographically and environmentally stratified (see Fig. 4.2).

As climate data, we used mean monthly minimum and maximum temperature, monthly rainfall, the number of frost days per month calculated from daily climate records at .0083 x .0083 degrees resolution from Moreno and Hasenauer (2016). Monthly solar radiation data was extracted from the world climate data Fick and Hijmans (2017). Soil class and initial, minimum, and maximum soil water were extracted from Tóth et al. (2017). As a drought index, we calculated SPEI with the SPEI package (Vicente-Serrano et al., 2010).

4.2.2 *3-PG model*

To simulate forest dynamics, we used the r3PG Fortran implementation (Trotsiuk et al., 2020b) of the physiological forest model 3-PG (Landsberg and Waring, 1997; Sands and Landsberg, 2002). Forest dynamics in 3-PG are based on five monthly updated submodules (light, productivity, water, allocation and mortality). In the light submodule, light absorption is simulated based

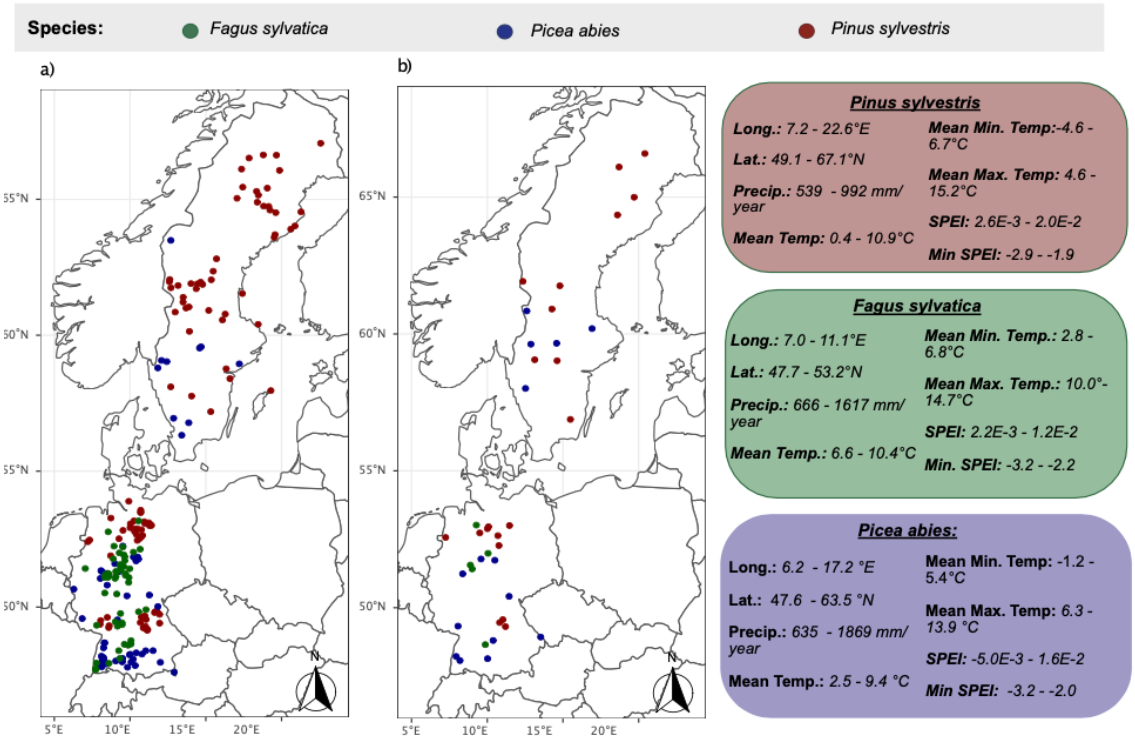


Fig. 4.2: The selected sites for a) calibration and b) evaluation located across Germany and Sweden. The boxes display the (climatic) range of the investigated species [*F. sylvatica* (green dots), *P. sylvestris* (red dots) and *P. abies* (blue dots)] in our study area.

on species-specific light extinction coefficients and leaf area index accounting for the horizontal canopy structure with fractional ground cover of the canopy (Landsberg and Sands, 2011). Accounting for canopy structure enables to calculate vertical gradients in radiation, vapour pressure deficit and aerodynamic conductance. To calculate gross primary production (GPP), the productivity submodule accounts for species-specific canopy quantum efficiency, which is influenced by temperature, frost, VPD, soil moisture, soil nutrient status, atmospheric CO2 and stand age (Landsberg and Waring, 1997; Sands and Landsberg, 2002; Almeida et al., 2004). Net primary production (NPP) is a constant fraction of GPP (Waring et al., 1998). In the water submodule, tree transpiration and soil evaporation, which together with canopy interception predict evapotranspiration, are calculated by the Penman–Monteith equation. In the allocation submodule, NPP is distributed to stems, roots and foliage based on soil nutrient status, VPD and soil moisture. In the mortality submodule, self-thinning is based on density-dependent mortality (-3/2 self-thinning law by Yoda (1963)) and density-independent mortality (e.g. by pests, diseases or drought).

To run the 3-PG model on the selected forest inventory sites, we extracted latitude, altitude, and the planting time from the data. The model was initialized with different age classes according to the planting time. For each age class, we then converted inventory data to stem-, leave- and aboveground- biomass using allometric equations (Forrester et al., 2017). The first inventory and the derived quantities of stem-, foliage-, aboveground biomass and stems per hectare were used to initialize the 3-PG model on each site, while the consecutive observations of stem biomass were used in the calculation of the objective function during the parameter calibration (see below).

4.2.3 *Selection of parameters displaying IV*

Based on a literature review of IV in tree species, we identified three key parameters in the model that could be expected to display IV across the species' range. First, we hypothesize that the optimal growth temperature (Topt) increases with MAT (although there are differences between different functional groups, see Way and Oren, 2010). This is based on the substantial role of optimal growth temperature for biomass production and that trees show their transgenerational memory under heat stress (Shanker et al., 2020). Second, we hypothesize that the maximum canopy quantum efficiency (alphaCx), i.e. the maximum rate of photosynthesis in the canopy, decreases with drought stress (Kattge and Knorr, 2007) and temperature. Both climatic variables enhance transpiration and thus cause water stress. One would therefore expect that trees reduce photosynthesis to prevent water loss trees in these climates. Third, we hypothesize that trees have stronger stomatal response (CoeffCond) to vapour pressure deficit when they experience more drought stress to prevent water loss (Novick et al., 2016).

4.2.4 *Sensitivity analysis and hierarchical Bayesian Model calibration*

Prior to calibrating the model, we performed a sensitivity analysis to select the most sensitive parameters of 3-PG. We used Morris screening (r = 200, levels = 10, grid jumps = 5) because of its computational efficiency (Morris, 1991). Parameter ranges were based on the estimates of Forrester et al. (2021) with +-10% around the best parameterization for parameters that were not zero and [-0.05,0.05] for parameters that were zero. As the target function for the sensitivity analysis, we used the same likelihood function as in the calibration: a gaussian distribution of the difference between predicted and observed stem biomass.

We then selected the five most sensitive parameters (see Supporting Information S2, Fig. S2.1) and the three parameters hypothesized to show IV for calibration. For the fixed species parameters and the means of the site-specific parameters, we generated priors based on the calibration of Forrester et al. (2021) (see Supporting Information S2, Table S2.1). Since this study was conducted in Switzerland and thus in a different climate, we increased prior uncertainties by a factor of two. As the maximum photosynthesis rate can only be positive, we logit-transformed it for the calibration but report back-transformed estimates.

We assumed a gaussian distributed error on the difference between predicted and observed stem biomass as likelihood, with an inverse gamma prior on the standard deviation. To implement site-specific individual estimates of Topt, alphaCx and CoeffCond the IV-model uses a HBM framework. This framework assumes that the site-specific parameters originate from a normal distribution with a calibrated average species response (mean of the distribution) and a certain amount of IV, which is fitted as a parameter with a uniform prior on the standard deviation of the normal distribution. Note that with this approach, we allow the data to decide about how much species differ across sites and environmental space (Van Oijen, 2017). The HS-model had the same parameters with the same likelihood and prior on the standard deviation, but estimated no site-specific parameters for Topt, alphaCx and CoeffCond.

To estimate posterior parameter distributions for both models, we used the Differential-Evolution Markov-Chain Monte Carlo (DEzs MCMC, see Ter Braak and Vrugt, 2008) sampler implemented in the BayesianTools package (Hartig et al., 2019). For each species, we ran six independent DEzs MCMCs, each with three internal chains, and tested convergence by visual inspection and Gelman–Rubin diagnostics (Gelman and Rubin, 1992). For mean parameter estimates for both models see Supporting Information S2, Table S2.1.

4.2.5 *Analysis of IV and model performance*

From the IV-model, we obtained species mean and standard deviation, as well as site-specific values for all IV parameters. We analyzed the magnitude of IV for each species and parameter by is standard deviation. Because the model's sensitivity to parameter variation depends also on the parameter's mean and thus the numeric estimates for the standard deviation cannot be directly compared between species, we additionally calculated an effective biological influence of the estimated IV by comparing the mean relative difference of average posterior predicted biomass to average predicted biomass when shuffling random effect estimates randomly across sites.

To test for a systematic climatic influence on the site-specific parameter estimates (see Selection of parameters displaying IV), we regressed the site-specific maximum a posteriori (MAP) estimates of the IV-model against the environmental variables specified in our hypotheses. We then extracted the R^2 of the regression and the p-value and confidence interval of the slope.

To judge the performance of prior and posterior parameter estimates, we drew each 1000 samples from the prior and posterior (HS-model and IV-model) distributions and calculated the relative root mean square error between model predictions and observations. Confidence intervals (CIs) of mean scaled errors were calculated as the standard error of the scaled error per site. To run the IV-model on the validation sites, we predicted IV parameters based on the fitted relationships between IV and climate.

To exemplify the implications of including IV into models under climate change, we extrapolated forest dynamics for *P. abies* (the most affected species in terms of biological influence, see Results) into a future scenario with a constant 2°C temperature increase. We used plots with similar MAT (excluding one site because of its strongly different climatic conditions). In this experiment, we differentiate between two scenarios: trees change their functional responses to new climatic conditions (consistent with a plastic origin of IV) or behave the same as before (consistent with a heritable origin of IV). We then simulated 3-PG with these parameters and calculated relative changes in predictions of biomass and NPP compared to no temperature increase. To see trends in the response, we then linearly regressed the relative changes against MAT.

4.3 RESULTS

The intraspecific variation model estimated substantial IV with strong biological influence on biomass predictions (Table 4.1). All three parameters displaying IV had standard deviation estimates with credible intervals not overlapping with zero (Table 4.1). For the optimal growth temperature, the numerical variation in parameter estimates was largest for *F. sylvatica*, followed by *P. abies* and *P. sylvestris* (Table 4.1a). Variation in the optimal growth temperature had the strongest biological influence (see methods for definition) on *P. abies* (7% biomass variation) and *P. sylvestris* (3.8%), while the influence on the biomass predictions of *F. sylvatica* was small (0.52%) (Table 4.1a). The variation for the maximum photosynthesis rate was small (Table 4.1b), both numerically and in terms of their biological influence. For the parameter controlling stomatal response to VPD, variation was relatively similar across species (Table 4.1c), but the biological influence was strongest for *F. sylvatica* (3.9%) and smaller for *P. sylvestris* (1.1%) and *P. abies* (0.92%).

Table 4.1: Estimates and credible intervals (CIs) for the magnitude (sd of the assumed normal random effect) and the biological influence (= impact on predicted biomass, calculated by the relative difference of average posterior biomass and average biomass when shuffling random effect estimates across the sites) of intraspecific variation in model parameters. The values were obtained by fitting a hierarchical structure for the a) optimal growth temperature, b) maximum photosynthesis rate and c) response to VPD independently for the three species *P. abies*, *F. sylvatica* and *P. sylestris*.

	P. abies			F. sylvatica			P. sylvestris		
	Estimate	CI	Bio. Infl. [%]	Estimate	CI	Bio. Infl. [%]	Estimate	CI	Bio. Infl. [%]
a) Variations in the optimal growth temperature									
Topt	1.30	(0.73–2.4)	7.000	1.70	(0.84–3.4)	0.520	1.200	(0.54–2.6)	3.800
b) Variations in the maximum photosynthesis rate									
alphaCx	0.11	(0.06–0.19)	0.077	0.14	(0.064–0.31)	0.046	0.099	(0.046–0.2)	0.053
c) Variations in the response to VPD									
CoeffCond	0.27	(0.14–0.48)	0.920	0.34	(0.16–0.68)	3.900	0.250	(0.11–0.5)	1.100

Next, we asked if site-specific parameter estimates correlate with the environmental factors that we identified as potential drivers for IV. We found that the optimal growth temperature significantly increased with MAT for *P. abies*, while there was no significant effect for *F. sylvatica* and *P. sylvestris* (Table 4.2a). The R^2, i.e. the share of variance explained by the response variables, was 0.081 for *P. abies*, while for *P. sylvestris* and *F. sylvatica* they were almost negligible (0.02 and 0.00016). For the maximum photosynthesis rate there were no significant effects of the minimum SPEI and maximum temperature for all three species and the R^2 were between 0.01 and 0.02 (Table 4.2b). For the response to vapor pressure deficit, there was a significant negative effect for *P. abies* (with an R^2 of 0.087) and no evidence for *F. sylvatica* ($p = 0.08$, $R^2 = 0.058$) and *P. sylvestris* ($p = 0.69$, $R^2 = 0.003$).

Table 4.2: Estimates, confidence intervals (CIs) and R2 of a linear model, regressing the locally variable model parameters (for a) optimal growth temperature, b) maximum photosynthesis rate and c) response to VPD) estimated by the IV-model against a priori selected environmental predictors. Independent regressions were fitted for the three species *P. abies*, *F. sylvatica* and *P. sylestris*.

	P. abies			F. sylvatica			P. sylvestris		
	Estimate	CI	R^2	Estimate	CI	R^2	Estimate	CI	R^2
a) Variations in the optimal growth temperature									
MAT	0.26 *	(0.026–0.5)	0.081	–0.011	(–0.26–0.24)	0.00016	–0.047	(–0.12–0.023)	0.020
b) Variations in the maximum photosynthesis rate									
Min. SPEI	4.5e–05	(–0.00036–0.00045)	0.020	–0.00019	(–0.00085–0.00047)	0.012	–4.3e–05	(–0.00029–2e–04)	0.022
Max. Temp.	–0.00021	(–0.00062–0.00019)	0.020	0.00021	(–0.00045–0.00087)	0.012	–0.00016	(–4e–04–8.8e–05)	0.022
c) Variations in the response to VPD									
SPEI	–0.046 *	(–0.085––0.006)	0.087	0.035·	(–0.0061–0.076)	0.058	0.0049	(–0.014–0.024)	0.0030

Signif. codes: <0.001*** ; <0.01**; <0.05*; <0.1 .

To understand if site-specific parameters affect model performance, we compared biomass predictions of 3-PG run with prior and posterior parameter estimates for both the HS and the IV-model (all calculated as mean error, samples from the respective distributions). We find that the HS-model had 10% error and the IV-model 7%, compared to 104% of the prior parameter estimates (Fig. 4.3a). The error of the IV strategy is smaller than the error using multi-site calibration for calibration as well as validation (Fig. 4.3a,b). The improvements were significant on the calibration sites, but not significant on the validation sites. The average error on the validation sites is slightly higher than on calibration sites.

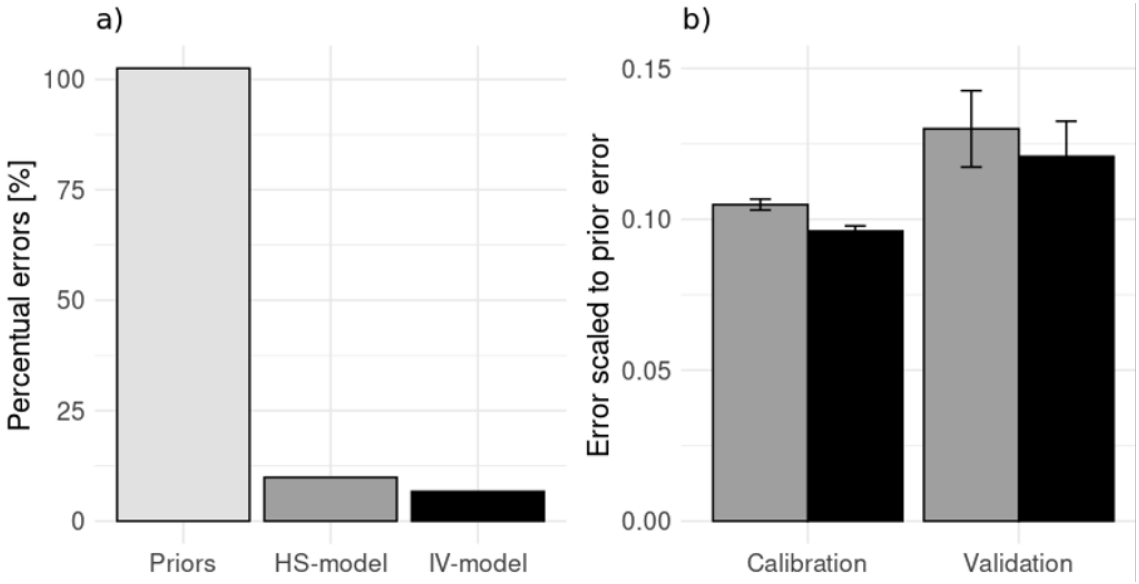


Fig. 4.3: Comparison of the different calibration strategies (no calibration, HS and IV) regarding a) relative RMSE on the calibration sites and b) relative RMSE scaled to no calibration error for the calibration and validation sites.

It was often suggested that modeling IV is important for correctly predicting ecosystem responses under global change (e.g. Moran et al. 2016). To test this prediction, we calculated relative changes of biomass/NPP per site for a 2°C temperature warming and regressed them against MAT. The overall pattern was that the NPP/biomass change under warming was more pronounced and variable in the IV-model compared to the HS-model (Fig. 4.4). For the HS-model, warming had little influence on biomass projections (Fig. 4.4a,c), but increased NPP by 10% (Fig. 4.4 b,d). With the IV-model, NPP/biomass increased for colder sites and decreased for warmer sites with a higher average NPP/biomass for no change in the functional response (Fig. 4.4). Moreover, note that the variability of the predicted responses was significantly higher for the IV-model.

4.4 DISCUSSION

In this study, we calibrated the physiological 3-PG model for three European species (*F. sylvatica*, *P. abies* and *P. sylvestris*) across a climatic gradient spanning from southern Germany to northern Sweden. Thereby, we allowed for IV in key model parameters (optimal growth temperature, maximum photosynthesis rate and response to vapor pressure deficit). We found that site-specific parameter values showed substantial and biological meaningful variation, in particular for the optimal growth temperature and the response to VPD (Table 4.1). A part of this variation could be explained by climatic predictors. More specifically, we found a significant correlation of MAT with the estimated optimal growth temperature, and of drought with the response to VPD for *P. abies*. Other correlations of site-specific parameters with climatic predictors were non-significant

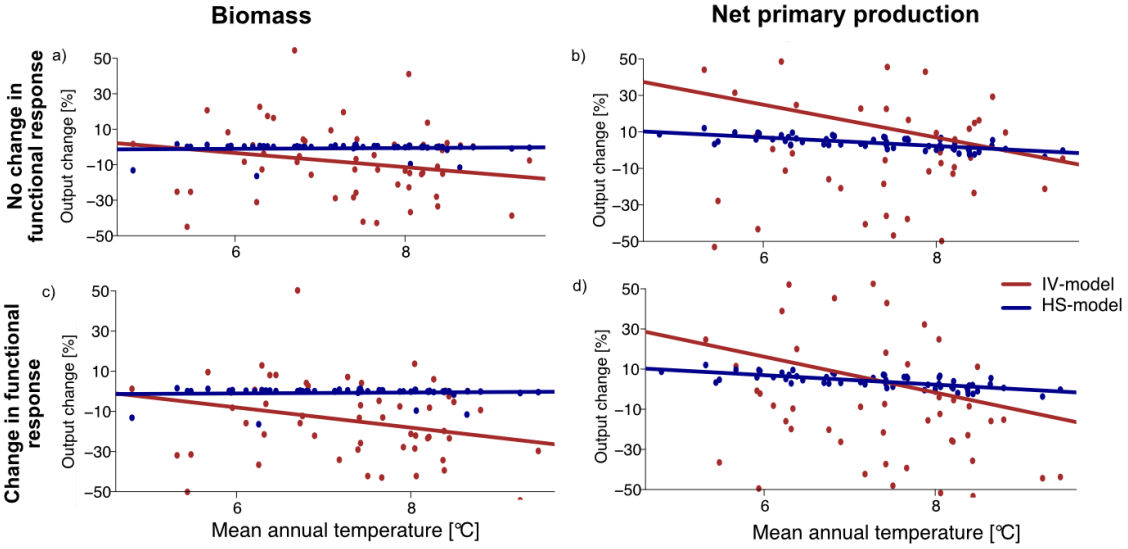


Fig. 4.4: Relative changes in biomass (a,c) and NPP (b,d), when predicting with the calibrated HS-model (blue) and IV-model (red) for a constant 2°C temperature warming are regressed against the MAT (red and blue lines) for heritable (upper row, meaning that local IV remains unchanged under climate change) and plastic (lower rows, meaning that local IV changes with the climate) trees.

(Table 4.2). The IV-model reduced calibration error compared to HS-model and using the derived climate-parameter correlations on new sites additionally also reduced validation error (Fig. 4.3). When testing the response of the fitted models to a temperature increase, we found that productivity and biomass responded more negatively towards warmer conditions in the IV-model compared to the HS-model (Fig. 4.4).

When considering the IV estimates in more detail, we first want to note again that the estimated variation induces substantial biological effects (Table 4.1). We judge these values to be in a plausible range for IV effects from heritable or plastic intraspecific trait variation (e.g. Aitken and Bemmels, 2016, found in a meta-analysis 3.6%/°C height growth increase). If these variations are heritable or plastic is not clear from this analysis. To distinguish between these options, we would need data of genetic similarity.

Looking in more detail at the patterns of inferred IV, we find them broadly consistent with empirical studies. *F. sylvatica* showed most IV in the optimal growth temperature (i.e. highest standard deviation), but with small biological influence, which might be a sign for its higher plasticity and greater climatic capacity (e.g. Hajek et al., 2016). The substantial biological effect of variations in VPD response for *F. sylvatica*s indicates an effective stomatal closure in line with an empirical study during the 2003 drought (Hentschel et al., 2016), increased protection strategy as climax species (Durrant et al., 2016a) and higher stomatal conductance of broad-leaved compared to conifer species (Ouyang et al., 2022). For conifer species, i.e. *P. abies* and *P. sylvestris*, IV in the optimal growth temperature had much stronger biological influence which can be a sign that increasing temperature is already challenging for these species (Hanewinkel et al., 2013). Alternatively it could be explained by them being pioneer specie (Caudullo et al., 2016; Durrant et al., 2016b) and thus investing into and relying on fast growth.

It is usually assumed that IV is in part a response to environmental conditions (e.g. Moran et al., 2016) and we found evidence for this (R^2 values up to 0.09; however, significant only for *P. abies*). The significant positive correlation of optimal growth temperature and MAT for *P. abies* could be epigenetic. Studies show that *P. abies* "remembers" the temperature at seed maturation (Johnsen et al., 2005), which then influences adaptive traits like optimum growth temperature (Kvaalen and Johnsen, 2008). This memory effect is stronger in *P. abies* than in other

conifer species (Schmidtling and Hipkins 2004), which may explain the non-significant correlation for *P. sylvestris*. The second significant effect, the negative correlation between drought and response to VPD, means that *P. abies* populations experiencing more severe drought more strictly regulate their stomatal conductance to reduce transpiration. This finding is in line with higher drought sensitivity of *P. abies* (Zang et al., 2014) compared to the other investigated species. Non-significant effects of the maximum photosynthesis rate (Table 4.2b) could indicate a trade-off in the variability between maximum photosynthesis rate and stomatal response to VPD (see Kattge and Knorr, 2007).

Accounting for IV decreased error compared to the HS-model both on calibration and validation sites (Fig. 4.3), although we acknowledge that the reduction on the validation sites was not statistically significant (Fig. 4.3b). We are not overly concerned by this lack of significance, however, because based on the residual error (R^2 values in Table 4.2) of the IV with the environment, we had to expect a much lower improvement of fit on the validation sites, and moreover, the number of validation sites was relatively low. We assume that with a larger number of validation sites, we may have found significance also for the validation.

Nevertheless, the lack of significance in the validation highlights the more general question on how to decide and check if or when including IV will improve predictions. We divide this question into two parts: the spatial predictability of IV, and the stability of this variation under changing conditions.

The first part, the spatial predictability of IV, means that to improve model performance on extrapolation sites, we require that IV is spatially predictable. In our case, the highest R^2 between model parameters and the environment was 0.09, meaning that the predictability of IV was relatively low and performance gains of the IV-model on the validation sites was expected to be lower by a factor 10 at least. We note, however, that we made predictions only with the climate predictors that were chosen based on our ecological hypotheses - it is likely that including more climatic predictors and more sophisticated modelling approaches could have improved the ability to extrapolate IV in space.

The second issue for IV improving predictive performance is related to the origin of the inferred spatial variation in parameter estimates. For model predictions under climate change, IV would not only make a difference if the origin is plastic or heritable (see Fig. 4.4), but there is the additional possibility that the IV-model compensates for interactions and processes that are not represented in the model (Medlyn et al., 2015, see also our projections under warming). Ignoring the unmodeled processes (HS-model) decreases the model-data fit (Oberpriller et al., 2021b), but incorporating too many processes might also bias projections. Thus, the plausibility of consequences of inferred parameter-climate relationships should be checked by comparing simulations to expectations. If the consequences are implausible, this may indicate that the relationship is rather a process error than the result of true IV.

The issues discussed in the previous paragraphs are exemplified by our results on model projections under climate change, where we found that the IV-model for *P. abies* responded more negatively compared to the HS-model at the pre-adopted warm edge of the distribution (Fig. 4.4). The result was contrary to our expectation that IV should buffer against climatic stress and particularly surprising because the site-specific calibration showed an increase in optimal temperature towards the warm edge of the distribution, which is in line with the original expectation. While it is certainly possible that the nonlinearities in the model lead to these unexpected outcomes, it is also possible that the IV-model uses its increased flexibility to compensate for physiological processes that are not well represented in the HS-model. Specifically, we speculate that the IV-model can better fit the observed productivity-temperature relationship for *P. abies* in Europe (more productivity for colder sites than for warmer sites, see Levanič et al., 2008). Consequently, we then also observe this effect in our climate change simulations (Fig. 4.4a). This

is also consistent with our results of decreased average projected biomass, when we assume that *P. abies* trees adapt to new climatic conditions (Fig. 4.4b). Whether nonlinearities or un-modeled processes are responsible for the response, and if the predictions are realistic, should be investigated further. Nevertheless, our results demonstrate that predictions from the IV and the HS model differ substantially.

A certain limitation of our study is that we considered IV in only three model parameters. Although this choice was based on carefully selected ecological considerations, there is the danger that the calibration did transfer variation of other parameters or functions, as well as model error, to the three parameters that were allowed to vary between sites (similar to excluding parameters from the calibration, see Minunno et al. 2013). In theory, we should thus allow IV in all parameters ("let the data speak", see Van Oijen, 2017), but in practice, this would require far more data and computational resources and, in case of too little data, could lead to imprecise estimates (similar to variance estimates in mixed-effect models with a low number of levels; see Oberpriller et al., 2021a). Thus, although we acknowledge the ambiguity associated with allowing IV only in selected parameters, we defend the choice of priori specifying a hypothesis regarding where IV occurs. Another potential issue leading to spurious results are errors in or a lack of environmental data. For example, soil fertility is an input to the 3-PG model, which can hardly be constrained, but influences the predictions. However, as long as these drivers are not systematically biased, our estimates are conservative (for a detailed discussion see Hutcheon et al., 2010).

4.5 CONCLUSION

We show suitability and applicability of the hierarchical Bayesian modeling to incorporate intraspecific variability into a dynamic ecosystem model, letting the data inform the magnitude and patterns of variation. The estimated IV in model parameters also allows testing ecological hypotheses which could not be addressed in field studies without high financial effort (e.g. in provenance studies) and long temporal duration. Our findings highlight the potential benefits of representing IV for making accurate site-specific predictions (e.g. for soil stability predictions Ali et al., 2017; or functional tradeoffs underlying biodiversity patterns He et al., 2021), and support the idea that IV may reduce some of the unexplained residual variance in forest predictions at large scales (Dietze et al., 2008). While our approach offers a way forward to quantify intraspecific functional variation from field data, an important remaining issue for ecological understanding and practical responses to climate change is to understand the source of these variations (Moran et al., 2016). Given such an understanding, incorporating relevant information about IV into forest models will improve model prediction and thus making informed choices about potential strategies to buffer climate change impacts on forests (Di Sacco et al., 2021).

5

FIXED OR RANDOM? ON THE RELIABILITY OF MIXED-EFFECTS MODELS FOR A SMALL NUMBER OF LEVELS IN GROUPING VARIABLES

Status: Provisionally accepted in *Ecology and Evolution*
Authors: Johannes Oberpriller, Melina de Souza Leite, Maximilian Pichler
Author Contributions: MP, JO and MSL designed the study. MP and JO ran the simulations, analyzed the results and wrote a first draft. All authors contributed equally to revising the manuscript and interpreting and discussing results.

ABSTRACT: Biological data are often intrinsically hierarchical (e.g., species from different genera, plants within different mountain regions) which made mixed-effects models a common analysis tool in ecology and evolution because they can account for the non-independence. Many questions around their practical applications are solved but one is still debated: Should we treat a grouping variable with a low number of levels as a random or fixed-effect? In such situations, the variance estimate of the random effect can be imprecise, but it is unknown if this affects statistical power and type I error rates of the fixed-effects of interest. Here, we analyzed the consequences of treating a grouping variable with 2-8 levels as fixed- or random-effect in correctly specified and alternative models (under- or overparametrized models). We calculated type I error rates and statistical power for all model specifications and quantified the influences of study design on these quantities. We found no influence of model choice on type I error rate and power on the population-level effect (slope) for random intercept only models. However, with varying intercepts and slopes in the data-generating process, using a random slope and intercept model, and switching to a fixed-effects model, in case of a singular fit, avoids overconfidence in the results. Additionally, the number and difference between levels strongly influences power and type I error. We conclude that inferring the correct random-effect structure is of great importance to obtain correct type I error rates. We encourage to start with a mixed-effects model independent of the number of levels in the grouping variable and switch to a fixed-effects model only in case of a singular fit. With these recommendations, we allow for more informative choices about study design and data analysis and make ecological inference with mixed-effects models more robust for small number of levels.

5.1 INTRODUCTION

Many biological data from experimental or observational studies have hierarchical grouping (or blocking, or clustering) structures that introduces dependencies among observations (McMahon and Diez, 2007; Bolker et al., 2009; Zuur et al., 2009; Harrison et al., 2018). A statistical analysis must account for these dependencies to ensure consistency of statistical properties (e.g. type I error rate Arnqvist, 2020), a task for which linear and generalized mixed-effects models (LMMs or GLMMs) were designed (Laird and Ware, 1982; Chen and Dunson, 2003). Mixed-effects models have replaced ANOVAs as the common tool for variance analysis (Wainwright et al., 2007; Bolker et al., 2009; Boisgontier and Cheval, 2016) because they allow simultaneous analysis of variance at different hierarchical levels (Krueger and Tian, 2004; Boisgontier and Cheval, 2016), handle unbalanced study designs better (Swallow and Monahan, 1984; Lindstrom and Bates, 1988; Pinheiro and Bates, 1995; Littell, 2002), and have better statistical properties for missing data (Baayen et al., 2008).

Mixed-effects models have the ability to adapt to different data structures, but the flexibility (see Box 1; Wainwright et al., 2007) that comes with them also leads to discussions about their challenging application (Nakagawa and Schielzeth, 2013; Dixon, 2016). This includes data-related properties such as the best way to handle overdispersion (Harrison, 2014; Harrison, 2015), small sample sizes in the individual blocks (Gelman and Hill, 2007), technical aspects such as robustness to wrong distributional assumptions of the random effects (Schielzeth et al., 2020), and to questions about how to compare different mixed-effects models (e.g. using R^2, Nakagawa and Schielzeth, 2013). Additionally, there are application-oriented issues (Harrison et al., 2017; Meteyard and Davies, 2020) such as the question about the complexity of the random-effect structure (Barr et al., 2013; but see Matuschek et al., 2017), the interpretation of random-effects (e.g. Dixon, 2016), or when a grouping variable should be treated as random or fixed-effect (Harrison et al., 2018).

A priori, modeling a grouping variable as fixed- or random-effect are for balanced study designs equally well suited for multilevel analysis (Kadane, 2020). There are no strict rules because the best strategy generally depends on the goal of the analysis (Gelman and Hill, 2007, see Box 2), however, for unbalanced designs there are some subtilities. For instance, random-effect estimates incorporate between and within group information whereas the corresponding fixed-effects model (grouping variable is specified as a fixed-effect) only within group information which leads to different weighting of the individual level estimates (not in balanced study designs) (McLean et al., 1991; Dixon, 2016; Shaver, 2019; ,but see Giesselmann and Schmidt-Catran, 2020). This is important when one is interested in the actual level effects themselves (narrow-sense inference analysis), but also when only interested in the population-level effect (broad-sense inference analysis). If we want to analysis the average fixed-effect over groups for a non-linear model, the naive weighted average gives a wrong result, because of the nonlinearity.

The different inferential conclusions that result from fixed and random effect modeling are due to the different assumptions underlying these two options (Millar and Anderson, 2004). Modeling a grouping variable as random-effect implicitly assumes that the individual levels of the grouping variable are realizations of a common distribution, usually a normal distribution, for which the variance and the mean (the population-level effect) need to be estimated (e.g. DerSimonian and Laird, 1986). As random effects are commonly parametrized so that the random-effect has a zero mean, this assumption shrinks the estimates of each random-effect level to zero. In contrast, treating a grouping variable as a fixed-effect makes no distributional assumptions about the individual level estimates (i.e., treating the levels separately of each other and thus no between level information is used to estimate the level effects). The random-effect model has fewer effective parameters than the fixed-effects model because of the shrinkage (e.g. Gelman and Hill, 2007), which can lead in balanced designs to higher statistical power to detect significant population-level effects at the cost of higher computational and numeric demand (Bolker et

al., 2009), discussions on how to correctly calculate p-values in unbalanced designs (Bolker et al., 2009; see Nugent and Kleinman, 2021) and a bias towards zero of the random-effect estimates (Johnson et al., 2015).

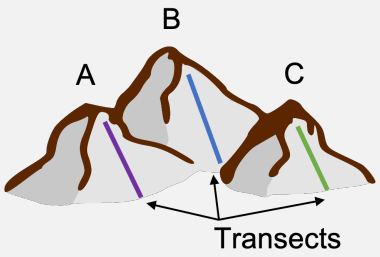
So, if we are not interested in each individual level effect (broad-sense inference), random-effect modeling seems preferable over fixed-effects modeling. It is, however, unclear if these advantages remain when the number of levels in the grouping variable is small (cf. also Harrison et al., 2018), because this might cause an imprecise and biased random-effects' variance estimate (Harrison et al., 2018), which then could influence the population-level effect estimate of the mixed-effects model (Hox et al., 2017).

The ecological literature suggests as a rule of thumb that an approximately precise estimate of the random-effect' variance requires at least five, sometimes eight, levels (Bolker, 2015; Harrison, 2015; Harrison et al., 2018). With four or fewer levels in the grouping variable, the preferred alternative is to include it as a fixed-effect (Gelman and Hill, 2007; Bolker et al., 2009; Bolker, 2015). But this threshold seems to be arbitrary chosen as it varies by discipline, e.g. 10-20 in psychology (McNeish and Stapleton, 2016) or 30-50 in sociology (Maas and Hox, 2005). To our knowledge, however, none of these values were based on a systematic analysis of how the modeling choice of the grouping variable affects statistical properties such as the type I error rate and power of the estimated population-level effects (i.e., the weighted average slope or intercept over a grouping variable).

Box 1: Scenario of an ecological study design with grouping/blocking variables

Sampling design. Suppose we want to understand the population-level effect of temperature on the height of a plant species that grows in different mountains. We hypothesize that higher temperature (lower altitude) increases the height of flowering plants. To do so, we establish altitudinal transects in many mountains and collect information from a certain number of plants. In this idealized scenario, we assume that the temperature predictor variable is colinear with altitude and not confounded with any other predictors like soil type, moisture, or ph.
Problem. The transects are not in the same geographical alignment, the type of soil varies in each mountain, and the plants are genetically very distinct among populations. All these factors introduce differences among populations that are not exactly of our interest (given our hypotheses), but statistically, plants of the same mountain are non-independent observations. The mountains can be considered as grouping, blocking or control variable.
Modeling options. We may use a mixed-effects model with a random intercept and slope (Box 2) for mountain to account for the differences among populations (grey lines in Fig. I while still modeling the relationship of interest as fixed-effects (blueline). An alternative may be to use a fixed-effects model, i.e., to include mountain as a categorical predictor (Box 2).

Sampling design



Hypothesis - The height of flowering plants increases with temperature:

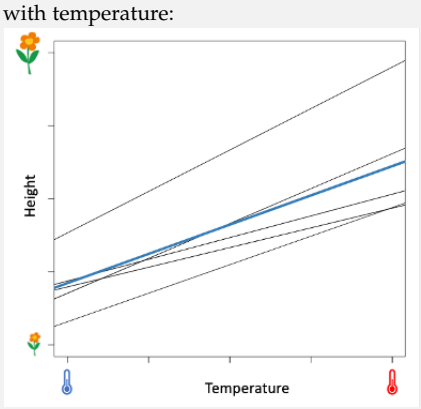


Figure I: Individual realizations of the height dependence on temperature (grey lines) and the overall realization (blue line).

Here, we analyze a situation where an analyst wants to infer the population-level effect and decided to use a mixed-effects model but is confronted with a low number of levels in the grouping variable. For this scenario, we simulated an unbalanced study design on the height of a plant on a temperature gradient to compare empirical power and type I error with a varying number of levels (two to eight mountains). To represent the challenge of correctly specifying the model structure and the consequences if the structure is not correctly specified, we additionally tested mis-specified models (overparametrized or underparametrized versions of the fixed and mixed-effects models). To quantify the effect of these modeling choices on the population-level effect, we compared: type I error rates and statistical power. Based on our results and in the context of broad-sense inference, we give practical recommendations on when to include grouping variables as random-effect or as fixed-effect.

Box 2: **Modeling a grouping variable as random or fixed-effect**

Fixed or random effect? The question of whether to include a grouping (blocking) variable as random or fixed-effect in the analysis depends on several factors. Fixed-effects are usually used when the analysts are interested in the individual level estimates of a grouping variable (Bolker et al., 2009) and these are independent, mutually exclusive, and completely observed (e.g. control and treatment in experiments, male and female when analyzing differences between sex) (e.g. Hedges and Vevea, 1998; Gunasekara et al., 2014). Random-effects are modeling choices when the variance between the different levels (Bolker et al., 2009) and not the exact estimates of the different levels are of interest (e.g. DerSimonian and Laird, 1986). Additionally, random-effects can be used when not every realization of the underlying mechanism can be observed (e.g. species across a number of observational sites in different geographic areas) but the analysts want to control for its influence (i.e. pseudo-replication, see Arnqvist, 2020). The two options differ in their interpretation, mixed-effects models use between- and within-group information whereas fixed-effects models use only within-group information. This subtle difference is important when for instance treatment or group differences are the goal of the analysis. Another important difference is that when modeling the categorical variable as fixed-effect conclusions apply to the levels used in the study, while when modeling as random-effect conclusions apply to the population of levels from where the studied levels were randomly sampled. However, in our example (Box 1), we are mainly interested in the population-level effect and not in the group differences which makes the inferential distinction negligible. See Gelman (2005) or Gelman and Hill (2007) for more decision criteria for whether an effect is random or fixed.

Technical differences between random and fixed-effects. When specifying a grouping variable as fixed-effect, the model with a default contrast in R estimates the effect of one reference level (see Schielzeth, 2010) differences between the reference level and possible linear combinations of other levels. Thus, it is not possible for fixed-effects models to estimate mean effect over groups (i.e., the population-level effect), but it can be calculated using e.g. bootstrapping (see Supporting Information S3), with sum-to-zero contrasts, or follow-on packages such as emmeans (Lenth, 2021). Mixed-effects models estimate the population-level effect and its variance and from a Bayesian perspective each individual level effect or from a frequentist perspective predict future realizations of the individual random-effect levels - Best Unbiased Linear Predictor (Fig. II b, d). Blocking variables may not only imply different intercepts (Fig II a, b), but also different slopes (Fig II c, d - the temperature "ecological" effect).
In fixed-effects models, this is done by introducing an interaction between the population level effect and the grouping variable. With mixed-effects models the choice of modeling different slopes and their correlation to intercepts for each group is related to the study design and may have impact on modeling structure and inference. Such correlations between random slopes and random intercepts are fitted by default but can be disabled.

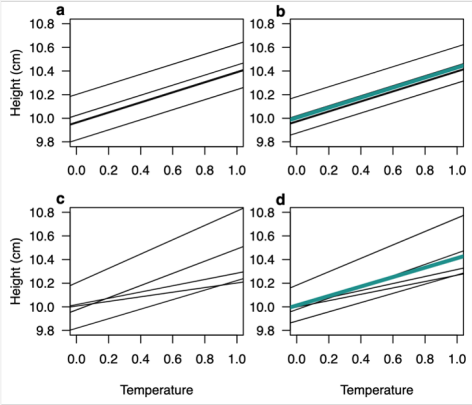


Figure II: Fixed- and mixed-effects models fit to simulated data with random intercept (a,b) and random intercept and slope (c,d) for each mountain in the example from H2 Box 1. Lines represent the individual estimates for each mountain. The blue line is the estimated population-level effect of mixed-effects models.

5.2 METHODS

5.2.1 *Simulated example and scenarios of data and model complexity*

To compare random- and fixed-effects modeling of a grouping variable with small number of levels, we simulated data based on our hypothetical example from Box 1. We hypothesized, that higher temperatures increase the average height of plants. We simulated an unbalanced study design – a common scenario in ecology and evolution (Schielzeth et al., 2020) - with two to eight mountains and a varying number of plants for each mountain (expected range between 40-360 plants per mountain) while keeping the overall number of plants constant (on average 200 plants per mountain) along altitudinal transects. For each case, we simulated 5000 datasets.

Table 5.1: Data-generating and tested models for each scenario: Scenario A random intercept for each mountain and B random intercept and slope for each mountain. For the fixed-effects models, we used R syntax for model formula in lm() function and for the mixed-effects models we used syntax from lmer functions from lme4. The response variable is height of flowering plants (H1, Box 1) and T is the temperature effect.

	Scenario A Random intercept only		**Scenario B** Random intercept and slope		**Description**
Data-generating model	(M1)	Height ~ T + (1|mountain)	(M6)	Height ~ T + (1|mountain) + (0 + T|mountain)	Effect of intercept (and slope in B) vary across mountains
Tested models					
Fixed-effects models	(M2)	Height ~ T	(M7)	Height ~ T	Temperature only main effect – underparametrized model
	(M3)	Height ~ 0 + T + mountain	(M8)	Height ~ 0 + mountain + T:mountain	Main effects of temperature and mountain (and interaction in B) – slightly more complex model
Mixed-effects models			(M9)	Height ~ T + (1|mountain)	Temperature and mountain both vary – underparametrized models
	(M4)	Height ~ T + (1|mountain)	(M10)	Height ~ T + (1|mountain) + (0 + T|mountain)	Effect of intercept (and uncorrelated slope temperature in B) vary across mountains – correctly specified models
	(M5)	Height ~ T + (1|mountain) + (0 + T|mountain)	(M11)	Height ~ T + (T|mountain)	Effect of intercept and slope temperature (correlated effects in B) across mountains – overparametrized models

5.2.2 *Scenario A - random intercepts per mountain*

In scenario A, we assumed mountains only differ in their intercepts (mean height) and the effect of temperature (slope) is the same for each mountain (constant slope over the levels of the grouping variable, Table 5.1, Eq. M1). We tested two different mixed-effects model structures: a correctly specified model which corresponds to the data generating process (Table 5.1, Eq. M4) and an overparametrized model (Table 5.1, Eq. M5) with an additional random slope for each mountain. Since in real studies the true underlying data generating process is unknown, it is useful to understand if an overparametrized model correctly estimates the variances of the random effects to zero and predicts all random slope levels to zero (or nearly zero) and, thus, approximate the data generating process (Table 5.1, Eq. M1). As fixed-effect alternatives, we tested the correctly specified model with mountain as fixed intercept together with temperature as slope (Table1, Eq. M3), and an underparametrized model omitting mountain at all (Table 5.1,

Eq. M2). This last model corresponds to a mixed-effects model that estimates the variances of the random effect to be zero and thus predicts the random effects to be zero.

5.2.3 *Scenario B - random intercepts and random slopes per mountain*

In scenario B, we assumed the data generating process contained a random intercept and a random slope (without correlation among the random slopes and intercepts) for each mountain (Table 5.1, Eq. M6). Here, the population-level effect (temperature) differs among levels of the grouping variable (mountain). We tested three different mixed-effects model structures: a correctly specified model corresponding to the data generating process (Table 5.1, Eq. M 10), an overparametrized model containing an extra term for the correlation of the random intercept and random slope (Table 5.1, Eq. M 11), and an underparametrized model with only a random intercept for each mountain (Table 5.1, Eq. M 9). We used the underparametrized model to test the effect of not accounting for important contributions to the data-generating process. Note, however, only in case of balanced designs and linear models the population-level effect estimate from the underparametrized model is consistent with the full model, because of different weighting schemes (for unbalanced designs) and the fact that the expected value of a non-linear transformation of estimates is not the same as the non-linear transformation of the expected value of these estimates. As fixed-effect alternatives, we tested the correctly specified model with the main effects of temperature, mountain and their interaction (Table 5.1, Eq. M 8), and the under-parametrized model without mountain as predictor (Table 5.1, Eq. M 7). We tested the last model because mixed-effects models that estimate zero variance for both random-effects are virtually the same as fixed-effects models that omit the grouping variable.

5.2.4 *Model fitting*

We fitted linear mixed-effects models to our simulated data with the lme4 R package (Bates et al., 2014) together with the lmerTest (Kuznetsova et al., 2017) package, which uses the Kenward-Rogers approximation to get the p-values of the fixed-effects. For fixed-effects models, we used the lm() function of the R stats package (Version 4.1, R Core Team 2021). For fixed-effects models in scenario A, we extracted p-values from the summary() function and, for scenario B, we used the fitted variance-covariance matrix and the individual level effects to bootstrap the population-level effect and its standard error (see Supporting Information S3). Obtaining p-values for mixed-effects models is intensively discussed in the statistical community and they are only exact for simple designs and balanced data (Kuznetsova et al., 2017). One reason is that in order to calculate p-values in mixed-effects models denominator degrees of freedom must be calculated, which generally can only be approximated (Kuznetsova et al., 2017). For best practice in which situations one should use which approximation see (Bolker et al., 2009; see also Nugent and Kleinman, 2021). The lmerTest package uses the Satterthwaite method to approximate the degree of freedoms of the fixed-effects in the linear mixed-effect model. We used the restricted maximum likelihood estimator (REML) (for a comparison of REML and MLE see Supporting Information S3). All results of mixed-effects models presented in scenario A and B are for the datasets without singular fits (see section Variances of random-effects and singular fits). Technically, singular fits occur when at least one of the variances (diagonal elements) in the Cholesky decomposition of the variance-covariance matrix are exactly zero or correlations between different random-effects are estimated close to -1 or 1. We repeated the analysis for the glmmTMB R-package because it uses a different approach to estimate mixed-effect models, see Supporting Information S3.2 for methods and results.

5.2.5 *Statistical properties and simulation setup*

We used type I error rate and statistical power of the population-level effects (average height and temperature) to compare the modeling options. For example, type I error rate for the temperature (slope) is the probability to identify a temperature effect as statistically significant although the

effect is zero. Statistical power in this case is the probability to detect the temperature effect as significant if the effect is truly greater than zero. For a correctly calibrated statistical test, the type I error is expected to be equal to the alpha-level (in our case 5%). To investigate type I error rates of the models on the intercept (average height) and average slope (temperature effect), we simulated data with no effects, i.e., the effects of temperature and mountain on height is zero. To additionally investigate statistical power, we simulated an example with a weak effect which corresponds to an average increase in size per unit step of the standardized temperature (linear scale) of 0.4 cm. For scenarios A and B, the individual effects for each mountain were drawn from a normal distribution with variance of 0.01 and 0.25 around the average effects: 0.4 cm average height (intercept), and 0.4 cm average increase in size with temperature (slope). We chose to run and compare simulations with these two values for the variance of the random effects to understand better how a larger or smaller variance may interfere in type I and power.

5.2.6 *Variances of random-effects and singular fits*

To understand how the number of levels affected random-effects variance estimates, we compared the variance estimates for random intercepts and slopes from the correctly specified mixed-effects model in scenario B (Table 5.1, Eq. M10). We also compared optimization routines (REML and MLE) in terms of estimating zero variances (singular fits, see below) (see Supporting Information S3). For bounded optimizations, which most R packages apply for the variance, it has been shown that the null distribution of a random effect's variance is a combination of a point mass at zero and a chi-squared distribution (Stram and Lee, 1994). For the sampling distribution with a true variance unequal to zero there are no proofs, but one would expect a similar distribution. While singular fits do not signal a convergence issue, the consensus is that the results of such models are not reliable. However, we decided to use non-singular fits and additionally non-singular and singular fits combined for calculating power and type I error for the mixed-effects models and to infer the effect of singular fits on the averaged statistical properties. We classified a dataset as singular or non-singular if the mixed-effects model ran in lme4 reported a singular fit warning message. For fixed-effects models, we used estimates from non-singular and singular datasets combined. Using only non-singular fits for calculating power and type I error impacts these statistical properties (e.g., type I error) because they are conditional on this selection and thus likely not to be at the nominal level (e.g., 5% for type I error rate). However, as our main intention is to report the type I error rates from the point of the analyst who may adjust the model structure to dispose of the singular fit, our reported rates represent empirical type I error rates.

5.2.7 *Quantifying the influences of study design on power and type I error*

Power and type I error of the population-level effect may depend not only on the number of levels (mountains) but also on the random-effect variance, the overall number of observations and the balance of observations among levels. To further quantify the impact of these study design factors on statistical power and type I error rate of the population-level effect, we additionally ran 1,000 iterations (each with 1,000 non-singular model fits) with the data generating process from scenario B for our ecological example. Thereby, we sampled the number of mountains from 2 to 20 with equal probability for each number, the random-effects variances from 10-4 to 4, the overall number of observations from 10 to 500 times the number of mountains. Additionally, to create different degrees of unbalance in data, we sampled for each mountain the average share of total observations from 0.1 to 0.9, which corresponds to at least 3 observations per mountain. We used the difference between the largest and the lowest proportion as proxy for the degree of unbalance. For the so generated data, we fitted the correctly specified linear mixed-effects and fixed-effects models from scenario B (Table 5.1, Eq 8) and calculated type I error rate and statistical power of the population-level effect. We then fitted a quantile regression using the qgam R-package (Fasiolo et al., 2020), with the statistical property (power and type I error rate) as

response and variance, number of levels, total number of observations and the unbalance proxy as splines. We used a quantile regression with splines as we expect a non-linear relationship.

5.3 RESULTS

5.3.1 *Scenario A - random intercepts per mountain*

When the effect of the temperature predictor was the same among mountains, irrespectively of the number of levels (mountains), all models except for the overparametrized model (random intercept and slope) showed an average type I error rate of 5% (Fig. 5.1a-d). Average power increased (Fig. 5.1e-h) with the number of mountains from 90% (2 mountains) to 100% (5 to 8 mountains). Note that the model omitting the grouping variable presented similar properties as the other models for small variances in the random effect. However, when increasing the variance of the random intercept in the simulation, the model omitting the grouping variable showed lower power (Fig. 5.1 g, h). For the overparametrized model, we found on average a lower type I error rate of less than 5% (Fig. 5.1a-d), and lower average statistical power to detect the temperature effect for a small number of mountains (Fig. 5.1e-h). When combining singular and nonsingular fits the overparametrized model had more average power compared to only non-singular fits and an average type I error closer to the nominal level (Fig. 5.1). The results for the intercept for the different models (see Supprting Information S3, Fig. S3.9) are similar to the results for the slope in scenario B (see below).

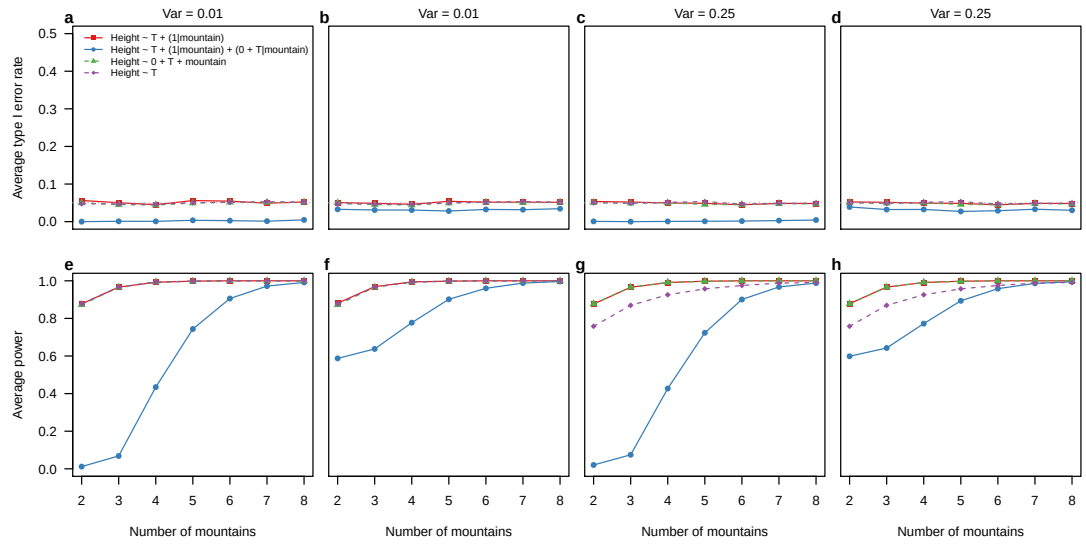


Fig. 5.1: Average type I error rates and average power for linear fixed and mixed-effects models fitted to simulated data with 2-8 mountains (random intercept for each mountain - Scenario A). For each scenario, 5000 simulations and models were tested. (a, b, e, f) show results for simulated data with a variance of 0.01 in the random effects. (c, d, g, h) show results for simulated data with a variance of 0.25 in the random effects. (a, c, e, g) show results for mixed-effects models only from datasets in which mixed-effects models converged without presenting singular fit problems and (b, d, f, h) results for mixed-effects models for all datasets. Results for fixed-effects (a-h) model are from all datasets. (a-d) the dotted line represents the 5% alpha level.

5.3.2 *Scenario B - random intercepts and slopes per mountain*

In scenario B, where the effect of the temperature differed among levels, the modeling decision influenced the average power and average type I error (Fig. 5.2). We found that average type I error rate of the correctly specified mixed-effects model (Table 5.1, Eq. M10) slightly increased (Fig. 5.2a) with the number of levels towards the nominal value (0.05) (Fig. 5.2a). The increase was stronger for larger variances (0.25) in the random effects (Fig. 5.2c). With singular fits, the mixed-effects models showed a higher average type I error rate than the nominal level for lower number of mountains (Fig. 5.2b, d). With a higher variance in the random effects, the average type I error rate was only increased for two levels (Fig. 5.2d). The overparametrized model with correlated random intercept and random slope (Table 5.1, Eq. M11) presented similar properties, but with decreased average power (Fig. 5.2e-h).

For the correctly specified fixed-effects model, average type I error ($\approx$ 2%) stayed constant with the number of levels (Fig. 5.2c) and a low variance in the random effects but increased stronger to the nominal level with a higher variance (Fig. 5.2d). Average power increased with the number of mountains (Fig. 5.2e-h). The mixed-effects model showed higher average power than the fixed-effects model irrespective of the number of mountains (Fig. 5.2 e-h).

The underparametrized model without the grouping variable had a higher average type I error rate (0.2) and higher average power than the other models (Fig. 5.2e-h). With a higher variance, the average type I error rate was even higher (0.8; Fig. 5.2c, d).

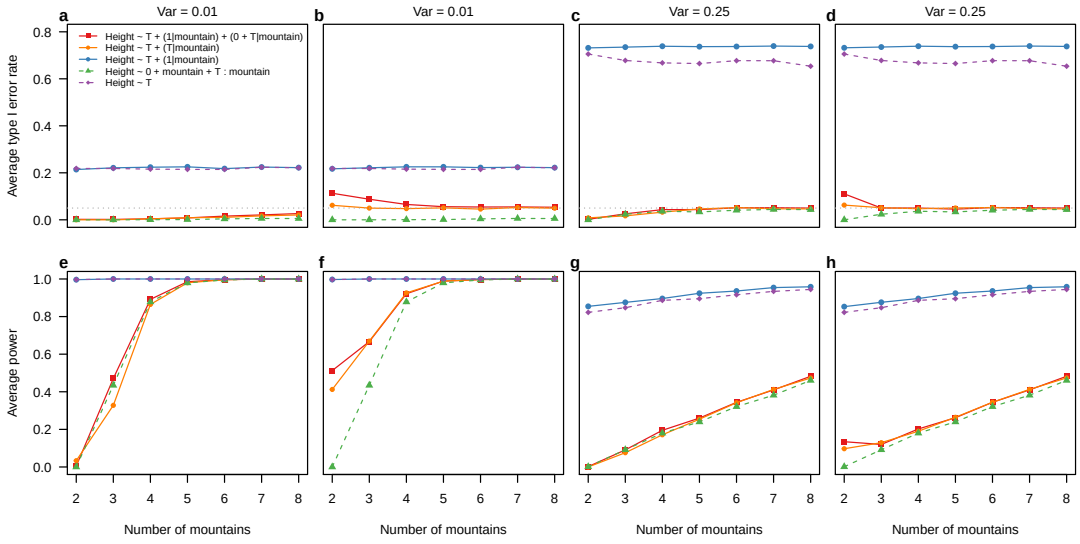


Fig. 5.2: Average type I error rates and average power for linear (mixed-effect) models fitted to simulated data with 2-8 mountains for scenario B (random intercept and random slope for each mountain range). For each scenario, 5.000 simulations and models were tested. (a, b, e, f) show results for simulated data with a variance of 0.01 in the random effects. (c, d, g, h) show results for simulated data with a variance of 0.25 in the random effects. (a, c, e, g) show results for mixed-effects models only from datasets in which mixed-effects models converged without presenting singular fit problems and (b, d, f, h) results for mixed-effects models for all datasets. Results for fixed-effects (a-h) model are from all datasets. In (a-d) the dotted line represents the 5% alpha level.

5.3.3 *Variance estimates of random-effects and singular fits*

We found for the models (singular and non-singular fit results combined) in Scenario B (random intercept and slope) that random-effects' variance estimates of the correctly specified model (Table 5.1, Eq. M10) approximately distributed as a chi-squared distribution around the correct value (0.01) and a point mass at zero (Fig. 5.3a, b median is near to zero). The point mass at zero decreased in height with increasing number of levels, i.e., less models estimated a variance of zero with an increasing number of mountains (Fig. 5.3a, b, see also Table S1). There was smaller bias for the random intercept variance estimates than for the random slope variance estimates, which were still biased for eight levels. When looking at models without singular fits, the variance estimates were chi-square distributed (Fig. 5.3c, d). The bias towards larger values was stronger compared to estimates with singular fits, especially for the random slope estimates (Fig. 5.3d).

By comparing the fitting algorithms, we found that using MLE led to more zero-variance estimates, i.e., singular fits, (see Supprting Information S3, Fig. S3.3, S3.4) than REML. Additionally, using MLE, non-singular variance estimates were strongly biased (Supprting Information S3, Fig. S3.3, S3.4), but the bias decreases with increasing number of levels. As expected, for both optimization routines, increasing the number of levels reduced the number of singular fits (Table S1). We found that singular fits led to different type I error rate and statistical power (Fig. 5.4) in

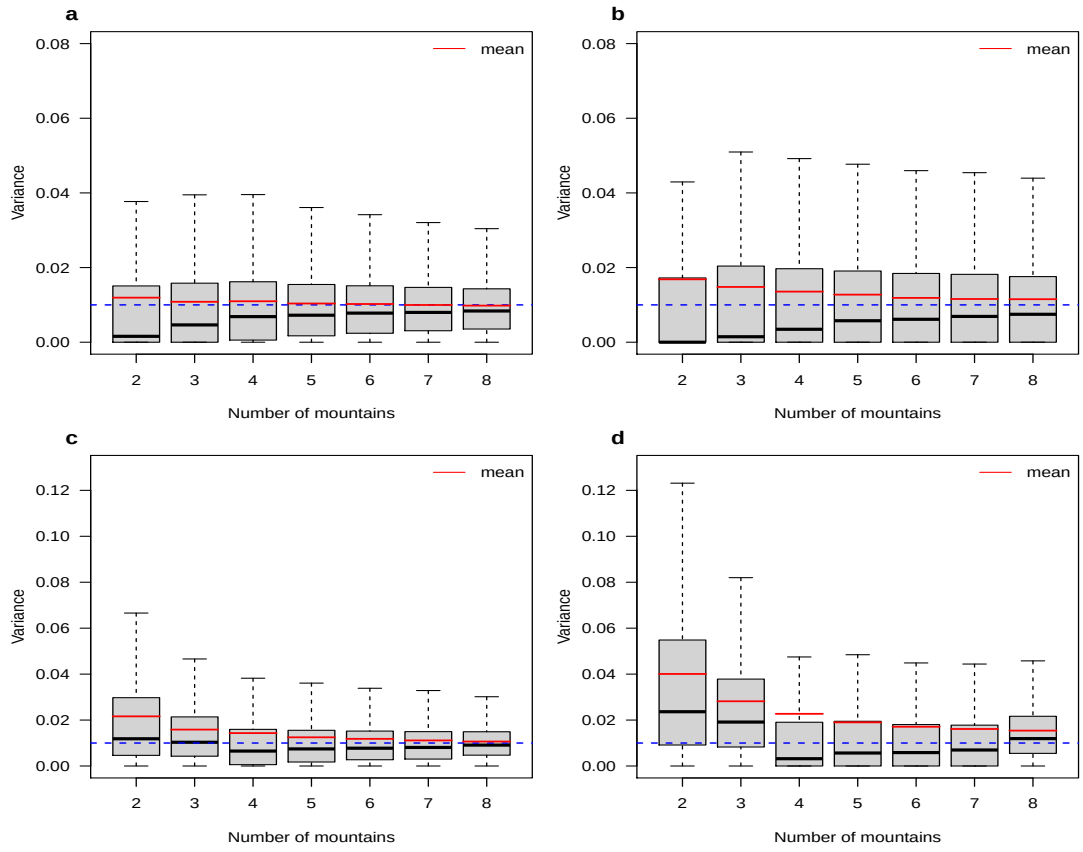


Fig. 5.3: Variance estimates of random intercepts (a, c) and random slopes (b, d) for linear mixed-effects models (LMM, Table 5.1. Eq. M10) in Scenario B, fitted with lme4 using REML to simulated data with 2-8 mountains. Figures (a) and (b) show the results for all models (singular and non-singular fits) and figures (c) and (d) show the results for only non-singular fits. For each scenario, 5.000 simulations and models were tested. The blue dotted lines represent the true variance used in the simulation (0.01) and the red lines the average variance estimates.

mixed- and fixed-effects models. For singular fits, the type I error rate of the correctly specified mixed-effects model was constant around 10% (like the model omitting the grouping variable), while with non-singular fits it was 1% for two levels and increased towards 3% with eight levels (Fig. 5.4a). In comparison, the fixed-effects model had similar type I error rates (no distinction between singular and non-singular fits because fixed-effects models don't estimate the variance of the individual level estimates), both increasing from 0% (two levels) towards 1% (eight levels) (Fig. 5.4c). We also found differences in power for the mixed-effects models between singular and non-singular fits (Fig. 5.4b, d). The power of the mixed-effects model with correct structure was higher for singular than non-singular fits especially for a low number of mountains (Fig. 5.4b).

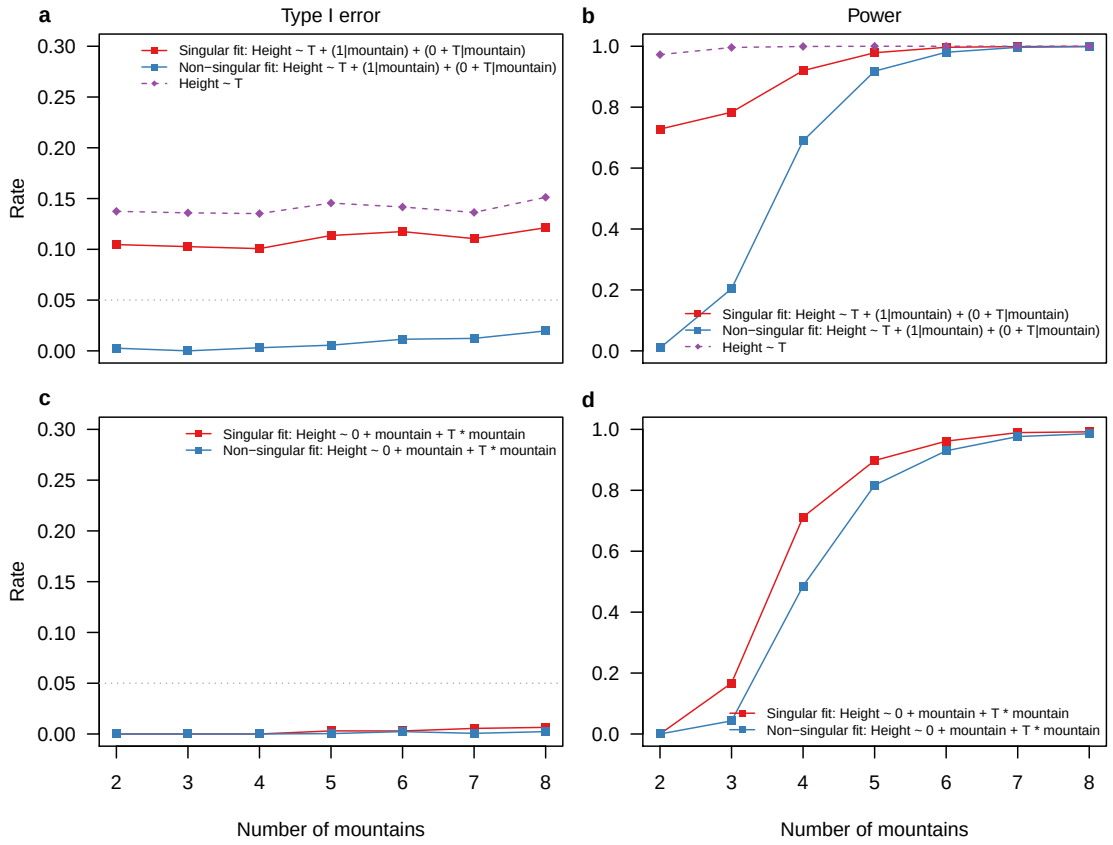


Fig. 5.4: Type I error rate and power of the correctly specified linear fixed and mixed-effects models in scenario B. We separated the datasets based on if when fitted they presented a singular fit (red lines) or non-singular fit (blue lines) warning. Figure (a) and (b) are results for the linear mixed-effects models, and (c) and (d) for the linear fixed-effects models. For comparisons, we show also results for the fixed-effects model that omits the grouping variable (mountain).

5.3.4 *Quantifying the influences of study design on power and type I error*

We found that the average type I error of mixed-effects models is slightly closer to the nominal value than its fixed-effect counterpart (Fig. 5.5a). Additionally, we found that the number of levels most strongly influences the type I error rate for mixed- as well as fixed-effects model (Fig. 5.5c). With five or more levels, however, the influence of the number of levels becomes negligible. Differences between the mixed- and fixed-effects models arose for the variance and the total number of observations. Here, the mixed-effects model was less influenced by a small random-effects' variance and a low number of total observations than the fixed-effects model (Fig. 5.5 b, d). Balance, following our definition, (see Methods) did not influence the population-

level effect in both models (Fig. 5.5e). For power we found no difference between fixed- and mixed-effects model (Fig. 5.5 f-j). For both models, an increase in variance decreased the power, while increasing the number of levels increased the power (Fig. 5.5 g, i). The total number of observations and the balance between groups had less influence (Fig. 5.5 h, j).

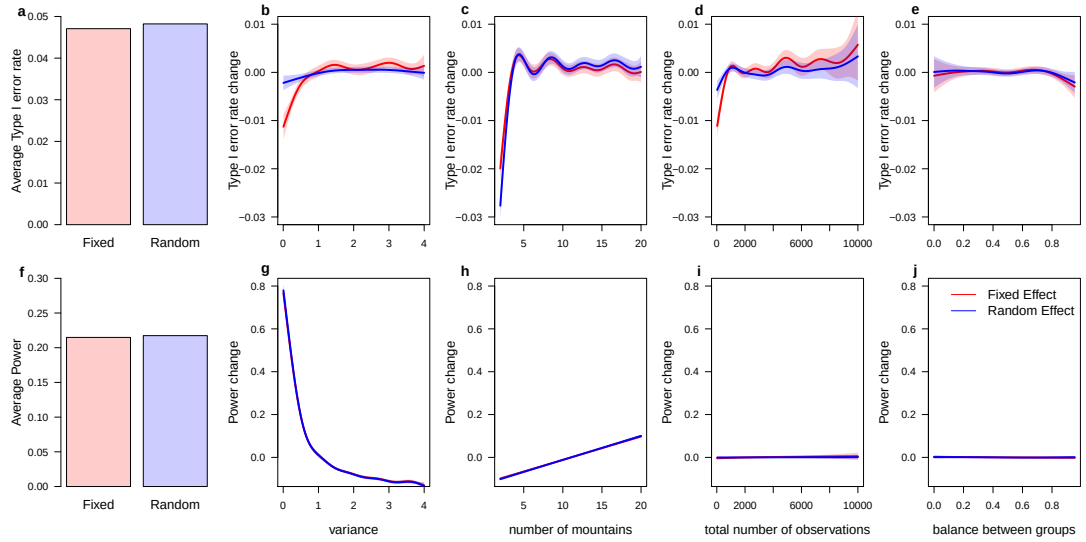


Fig. 5.5: Comparing the influence of study design factors on the type I error rate (b - e) and power (g - j) of linear mixed- (blue lines) and fixed-effects models (red lines) with their respective average values (a, f). We found that the variance of the random-effects and the number of levels (number of mountains) are the most important values to get correct type I error. For this analysis, we used the plant height example for Scenario B (random intercept and random slope). Results for mixed-effects models are only from datasets in which mixed-effects models converged without presenting singular fit problems, while results for fixed-effects model are from all datasets.

5.4 DISCUSSION

Ecological data collections or experiments produce data with grouping structures and mixed-effects models can account for these dependencies. The main questions we explored in this article was: "should analysts stick to the mixed-effects model or fall back to a fixed-effects model, when the grouping variable has few levels?", and "how does this decision influence statistical power and type I error rate of the population-level effect?" Here, we showed with simulations that mixed-effects models with small number of levels in the grouping variable are technically robust (Fig. 5.1) and that the decision between random and fixed-effect matters most when the effect size of the ecological predictor variable differs among levels (Fig. 5.2).

When the effect of the ecological predictor is the same for each level of the grouping variable (scenario A, random intercept model), almost all models presented the same average power and average type I error (see also Gomes, 2021) (Fig. 5.1a-d). The only exception was the over-parametrized model that presented too low average type I errors and lower average power (Fig. 5.1). We speculate that the model was unable to correctly predict the additional random effects to zero. Notably, for scenario A, the underparametrized model omitting the grouping variable presented correct average type I error rate (Fig. 5.1a-d). However, this is illusive because average power decreased with increasing effect sizes of the random effects (Fig. 5.1g, h). This confirms that the grouping variable needs to be included to correctly partition the variance among the different predictors (Gelman, 2005; Gelman and Hill, 2007; Bell et al., 2019). Also including the grouping variable is mandatory if one is interested in the average intercept, otherwise it would cause inflated average type I error rates (see Supprting Information S3, Fig. S3.1; see the

following section).

When the effect size of the ecological predictor differs for each level of the grouping variable (scenario B; random intercept and random slope model), the average type I error and power were influenced by both model choice and the presence of singular fit warnings. The mixed-effects models had a better average type I error than the fixed-effects models, especially for a larger number of mountains (Fig. 5.2). Power was comparable between mixed and fixed-effects models. But with non-singular and singular fits combined, the mixed-effects model had higher type I error rates and power than the fixed-effects models. In both cases, the mixed-effects models showed good type I error rates (about more or less than 5%) for a small number of levels.

Overparametrized mixed-effects models presented in both scenarios slightly lower average type I error and average power compared to the correctly parameterized mixed-effects model (Fig. 5.1,Fig. 5.2). This trade-off between type I error and power is in line with Matuschek et al. (2017) for different model complexities. Overall, the overparametrized models are more conservative but have less power than the simplified models. We think these more conservative estimates are preferable over anti-conservative estimates because some analysists tend to try a variety of analyses and only report significant ones (Simmons et al., 2011) and more conservative average type I error counteract this procedure.

However, dropping the correlation structure between random-effects should be carefully considered. It is possible that the type I error rate increases when no correlation in the model is assumed although there is one in data-generating process. Group-mean centering of the population-level effect may mitigate the requirement of assuming a correlation, but it also changes the interpretation of the model because the individual levels are not referenced to the population level effect anymore (they are now independent). In scenario B, underparametrized models exhibited inflated type I errors (in line with Schielzeth and Forstmeier, 2009; Barr et al., 2013; Bell et al., 2019) but very high average power (Fig. 5.2). We speculate that additional variance coming from the difference between levels in the grouping variable, which is not accounted, is attributed to the population-level effect and causes overconfident estimates.

5.4.1 *Variances of random-effects and singular fits*

The rate of singular fits was very high for small number of levels (Fig. 5.3, Table S1). In our simulations, singular fits corresponded to zero variance estimates of the random effects. The resulting distribution of variance estimates consisted of a right skewed chi-squared distribution and a point mass at zero (many zeros corresponding to the singular fits) as expected (see Stram and Lee, 1994). The variance estimates were biased and imprecise with a small number of levels, but the bias decreased with the number of levels towards zero (McNeish, 2017). Removing the singular fits led to even more bias in the variance estimates (Fig. 5.3c, d).

The biased variance estimates are caused by ensuring positive variances in the optimization routines (Bates et al., 2014; Brooks et al., 2017). In case of a singular fit, the correctly specified mixed-effects model had similar power and type I error as a fixed-effects model dropping the grouping variable (Fig. 5.4): no difference between the levels, which corresponds to a fixed-effects model without the grouping variable. However, the models still differed in their number of parameters (and degrees of freedom) which might explain the slight differences in power and type I error (Fig. 5.4).When switching to fixed-effects models for singular fits in the random-effect, the type I error rate and power were similar to the random-effect model with non-singular fits (Fig. 5.4).

5.4.2 *Connection to study design*

Earlier studies reported mixed recommendations about important study design factors. While some studies only stressed the importance of the total number of observations (Martin et al., 2011; Pol, 2012), we found, in accordance with Aarts et al. (2014), that the number of levels and the variance between levels have a strong influence on type I error rates and power. Due to our simulation design, which automatically increases the number of observations when increasing the number of levels, we however, cannot perfectly separate the effects of number of observations and levels from each other. The influence of the variance on power and type I error is mixed. On the one hand, increasing the variance had a positive effect on the type I error for both models but the fixed-effects model was more strongly affected (Fig. 5.5). The different distributional assumptions might explain this different behavior: the mixed-effects model assumes the levels to be normally distributed and estimates the variance of the levels flexibly, whereas the fixed-effects model makes no distributional assumptions. We speculate that the mixed-effects model benefits from this informative distribution assumption in this edge case with less than 5 levels. On the other hand, increasing the variance over a certain value (Fig. 5.5g) decreased the power of both models because more variance is explained by the difference between levels, and this increases the uncertainty of the slope effect estimate.

Given the strong influence of the number of mountains on type I error rates, we encourage to design a study with at least 8 levels because with more than 8 levels, the type I error rate was approximately not affected by the number of levels (Fig. 5.5c). In our scenarios, the influence of the unbalanced number of observations between levels was small (Fig. 5.5) confirming the robustness of mixed-effects to unbalanced data (Swallow and Monahan, 1984; Pinheiro and Bates, 1995; Schielzeth et al., 2020). However, if possible one should try to balance the groups because despite the robustness of mixed-effect models to an unbalanced design, it impacts the interpretation of the random effects and balanced studies create the least problems regarding the model option (Dixon, 2016). Moreover, the impact of study design on type I error and power stresses the importance of pre-experiments and power analyses (e.g. Johnson et al., 2015; Green and MacLeod, 2016; Brysbaert and Stevens, 2018) to maximize the meaningfulness and efficiency of a study.

5.4.3 *Practical suggestion*

Before giving practical advice, we must recall the exact situation in which this manuscript acts. We assume that an analyst is interested in a population-level effect, and that they have already decided to use a mixed-effects model (broad-sense analysis, not interested in the individual levels effects), but faces a small number of levels, so that our recommendations only apply to such situations. In this situation, the variance estimates of the random effects stabilizes in a reasonable manner with at least five levels in a grouping variable (Fig. 5.2). With less than five levels, variance estimates are biased to zero (Fig. 5.3) though without an effect on the observed average type I error rates of the population-level effect (Figs. 5.1 and 5.2). We rather found that the question of how to deal with a singular fit in the mixed-effects model is more crucial than the actual number of levels. If there is a singular fit warning, switching to the fixed-effects model leads to more conservative average type I error rates (Fig. 5.2). Acknowledging that most singular fits occur with a small number of levels (Table S1), this might also explain the common rule of thumb to do not fit a grouping variable as random-effect if it has fewer than 5 levels (Gelman and Hill, 2007; Bolker et al., 2009; Bolker, 2015).

Our recommendations are summarized in Fig. 5.6. We recommend starting with the mixed-effects model, regardless of the number of levels, and switching to a fixed-effects model only in case of a singular fit warning. How to deal with singular fits is a topic of ongoing discussion. While Barr et al. (2013) states to start with the maximum model and simplify the model in case of convergence issues and singular fits, Matuschek et al. (2017) suggests to think a priori about

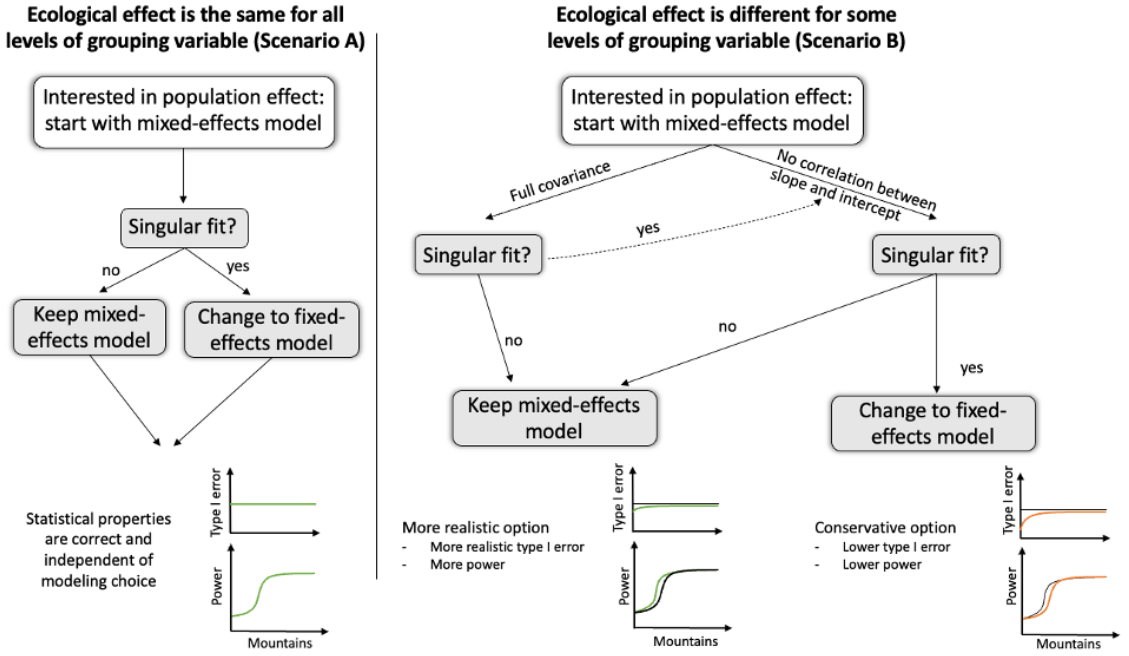


Fig. 5.6: Consequences and recommendations for mixed-effects models with a small number of levels in the random-effect. When the ecological effect (population-level effect) does not differ between different levels of the grouping variable (left side) all modeling options, which include the grouping variable, lead to the same results and thus only a singular fit requires a change to a fixed-effects model. If the ecological effect (population-level effect) differs among levels (middle to right side), starting with the mixed-effects model and only changing to the fixed-effects model in case of a singular fit is recommended.

using simpler models because of higher power in return of increased type I error rate. However, we disagree with the view of Matuschek et al. (2017) that trading a small increase in type I error rate for higher power is favorable, even though it could still be an interesting solution with the often-small number of observations in ecological studies, when the increase in power prevails the increase in type I error rate. We follow the position of Barr et al. (2013), and thus recommend starting with correlated random slope and intercept, when the population-level effect differs among levels. If obtaining a singular fit, switch to uncorrelated random-effects (following Matuschek et al., 2017) and in case of another singular fit, switch to a fixed-effects model.

Our recommendations assume that the random effect structure (e.g., random slope or not) is known a priori, which is often difficult in practice. Although model selection is theoretically possible for random effects (e.g. simulated (restricted) LRTs Wiencierz et al., 2011; or by residual checks as faciliated by Hartig, 2019), the frequentist point of view recommends sticking closely to the a priori derived hypothesis, otherwise the risks such as they arise from multiple testing increase. Moreover, if the grouping variable was included as a confounder, this erroneous omission can cause a high type I error and wrong estimates. If there is uncertainty about the random-effect structure or concern about the statistical power, more time should be invested up front in hypothesis design and appropriate power analyses for mixed-effects models (e.g. Green and MacLeod, 2016; Brysbaert and Stevens, 2018).

5.5 CONCLUSION

In conclusion, we showed that mixed-effects models are more robust than previously thought, despite the biased variance estimates for low number of levels in the grouping variable. We found that power and type I error of the population-level effect are robust against the model choice when the ecological effect is the same among the levels of the grouping variable, however,

the model matters when the ecological effect differs among levels. When in doubt about the data-generating process, we encourage starting with a simplified model (random intercept only) and consult model diagnostics and simulated LRTs to check for evidence of random slope effects. When finding evidence for random slopes in these tests, we recommend starting with the mixed-effects model and switching only to a fixed-effects model in case of a singular fit problem. With this work, we provide a practical guideline, which helps analysts in the study design, the data analysis and thus making ecological inference more informative and robust.

6

TOWARDS ROBUST STATISTICAL INFERENCE FOR COMPLEX COMPUTER MODELS

Status: Published in *Ecology Letters*, 2021
Authors: Johannes Oberpriller, David R. Cameron, Michael C. Dietze, Florian Hartig
Author Contributions: FH and JO conceived and designed the study. JO implemented the case studies, ran the experiments, and analyzed the results. DC advised regarding implementing errors in the BASFOR model. All authors contributed equally to discussing and interpreting the results, and to the preparation of the manuscript.

ABSTRACT Ecologists increasingly rely on complex computer simulations to forecast ecological systems. To make such forecasts precise, uncertainties in model parameters and structure must be reduced and correctly propagated to model outputs. Naively using standard statistical techniques for this task, however, can lead to bias and underestimation of uncertainties in parameters and predictions. Here, we explain why these problems occur and propose a framework for robust inference with complex computer simulations. After having identified that model error is more consequential in complex computer simulations, due to their more pronounced nonlinearity and interconnectedness, we discuss as possible solutions data rebalancing and adding bias corrections on model outputs or processes during or after the calibration procedure. We illustrate the methods in a case study, using a dynamic vegetation model. We conclude that developing better methods for robust inference of complex computer simulations are vital for generating reliable predictions of ecosystems responses.

6.1 INTRODUCTION

Ecological systems are often complex and interdependent (Levin, 1998). To understand these systems, and to forecast their dynamics under changing conditions, ecologists rely increasingly on complex computer simulations (CCS, near synonymous terms include: process-based models, mechanistic models, system models; see e.g. Evans et al., 2012; Briscoe et al., 2019; Thompson et al., 2020), for example to predict ecosystem responses to climate change (e.g. Cheaib et al., 2012; Rahn et al., 2018). The trend towards an increasing use of complex computer simulations mirrors similar developments in other scientific fields, for example galaxy formation (Somerville and Davé, 2015), macroevolutionary dynamics (Rangel et al., 2018) or epidemiological disease control (Drake et al., 2015).

For any of these models, precise forecasts and correct estimates of predictive uncertainty is paramount, both for their scientific interpretation (Petchey et al., 2015), and for decision making and governmental actions (Dietze et al., 2018). The IPCC report, for example, uses a combination of different earth system models to simulate future behavior of the atmosphere, ocean, land surface and fluxes (Bindoff et al., 2013). Using computer simulations for decision making is only sensible, however, if their predictions are sufficiently precise, and if their uncertainties are correctly communicated (Budescu et al., 2009).

Achieving these goals depends on correctly determining model structure, parameters, and their uncertainties. Where parameters and model structure cannot be determined directly by measurement or theory, they have to be estimated by comparing model predictions to data (model calibration and selection, e.g. Hartig et al., 2012; Dietze, 2017a). In recent years, the field has moved from informal methods for model calibration to established statistical methods such as maximum likelihood estimation (MLE, e.g. Castiglioni et al., 2010) or Bayesian inference (e.g. Harrison et al., 2012; Luke et al., 2017) . Superficially, it would seem that parameter calibration and uncertainty propagation in CCS is no different from the statistical regression models familiar to most ecologists, and that no special statistical theory is needed for these models (at least as long as model outputs are approximately deterministic, for stochastic simulation models see Hartig et al., 2011).

In practice, however, there are important differences between calibrating simple statistical models and CCS. One trivial difference is the sheer computational challenge of constraining large models to big data (e.g. Fer et al., 2018). Another, more fundamental disparity arises through the model structure. Compared to statistical models, CCS are characterized by having a higher level of interconnectedness and nonlinearity, as well as multiple variables and outputs. Moreover, CCS typically make a large number of structural assumptions based on prior knowledge (Dormann et al., 2012). As a consequence, they are often less flexible in terms of what outputs or patterns can be produced, despite having a large number of parameters (Fatichi et al., 2016).

These traits lead to certain problems when calibrating CSS that are less common in statistical models. A particularly important example are trade-offs when calibrating to multiple data streams. It has been argued that using multiple data streams is desirable, because information from different biological levels of organization (e.g. daily carbon fluxes and yearly inventory data) contains more complementary information than a single data stream (e.g. Grimm, 2005; Medlyn et al., 2015). However, the combination of internal constraints (e.g. mass- or energy balance) with structural error will often make it impossible for a CCS to fit all data streams simultaneously (for a list of examples see MacBean et al., 2016). Moreover, the information or observation density of data at different organizational levels can differ substantially, leading to unbalanced data (substantial differences in the number of observations of different data streams) for the calibration. This means that the calibration cannot avoid a systematic misfit (bias) in some of the model outputs, and additionally faces a conflict between the information provided by different, possibly unbalanced data streams, both situations that are less common in statistical

models.

The goal of this paper is to explore these problems in more detail and provide an overview of strategies for robust statistical inference with CCS. In the remainder of the text, we first explain the the problems that may occur when calibrating CCS with structural error, illustrated with the example of a complex forest ecosystem model. Based on our results, we test a range of suggested remedies, and finally provide practical recommendations for using statistical inference with CCS in ecology and evolution.

6.2 WHY DOES MODEL ERROR AFFECT STATISTICS DIFFERENTLY IN COMPLEX COMPUTER SIMULATIONS?

To start our discussion, it will be helpful to further clarify how conventional statistical models differ from CCS. Models exist on a continuum between these two classes (Dormann et al., 2012), but considering the ends of this spectrum, we see clear distinctions between models typically used for statistical data analysis (e.g. GLMMs, see Bolker et al., 2009) and CCS (e.g. Trotsiuk et al., 2020a). One key difference is that CCS usually connect a sizeable number of state variables via processes that aim to represent our scientific understanding of the natural system, often with submodels that are calculated at different time steps (e.g. daily, weekly and annual, see as an example the LPJ-GUESS model, Smith et al., 2001). It has often been argued that their mechanistic nature makes CCS more appropriate than regression models for forecasting far into the future, because, at least in principle, they should be able to predict into domains for which no previous data exists (e.g. Kearney et al., 2010; Rastetter, 2017; Radchuk et al., 2019).

These benefits of CCS, however, come along with larger structural complexity, which exacerbates challenges in identifying the correct model structure and correcting possible model-data discrepancies (Peng et al., 2011). For example, their typically high interconnectedness hampers the localization of structural errors. Moreover, while their mechanistic underpinning grants better inclusion of prior understanding of the processes driving system dynamics (Dietze et al., 2013), it can become a liability when mechanisms or parameters are unknown and have to be guessed. A final point is that CCS have to apply certain simplifications and discretizations for computational reasons (e.g. discrete soil layers Tiktak and Bouten, 1992). As a result of these and many more challenges, most CCS display certain structural errors, which are difficult to fix immediately (e.g. Richardson et al., 2012).

These structural errors (including observational bias as part of the statistical model) and their associated uncertainties increase the uncertainties in the calibration process (Bayarri et al., 2007; see also Beven, 2005; Trucano et al., 2006). To address this issue, the field has moved towards using formal, statistical methods for model calibration and uncertainty propagation. These methods, however, infer parameters and uncertainties conditional on the assumed model structure being correct. Statistical modellers are usually not overly concerned about these assumptions, because their models flexibly adjust to data, and thus their main concerns are distributional assumptions (e.g. Warton et al., 2015). In CCS, however, this assumption will not hold, and structural errors will interact with the inference, in particular when nonlinearities are large, and when the model is fit to imbalanced data (Abramowitz et al., 2008), i.e. when one data stream has much more observations than another.

A statistical calibration will respond to this problem by compensating structural error through adjusting parameters to values that differ from the true values of the underlying process (Bell and Schlaepfer, 2016). The resulting model may still display acceptable performance in the domain for which data is available, but parameter estimates may be biased, and their uncertainties may be underestimated. Moreover, when extrapolating beyond the data domain, which is considered an important strength of CCS, biases and underestimation of uncertainty can become

substantial (He et al., 2014), especially when the model is calibrated to multiple unbalanced data streams (an example dealing with these issues is Richardson et al., 2010). If the model is not able to fit both data streams at the same time, the calibration algorithm will face a conflict (MacBean et al., 2016). In this situation, the calibration will tend to use parameters adjustments to compensate the error in the more data-rich outputs, at the cost of increased error and too narrow confidence intervals (Sargsyan et al., 2019) particularly in the data-poor model outputs (Fig. 6.1).

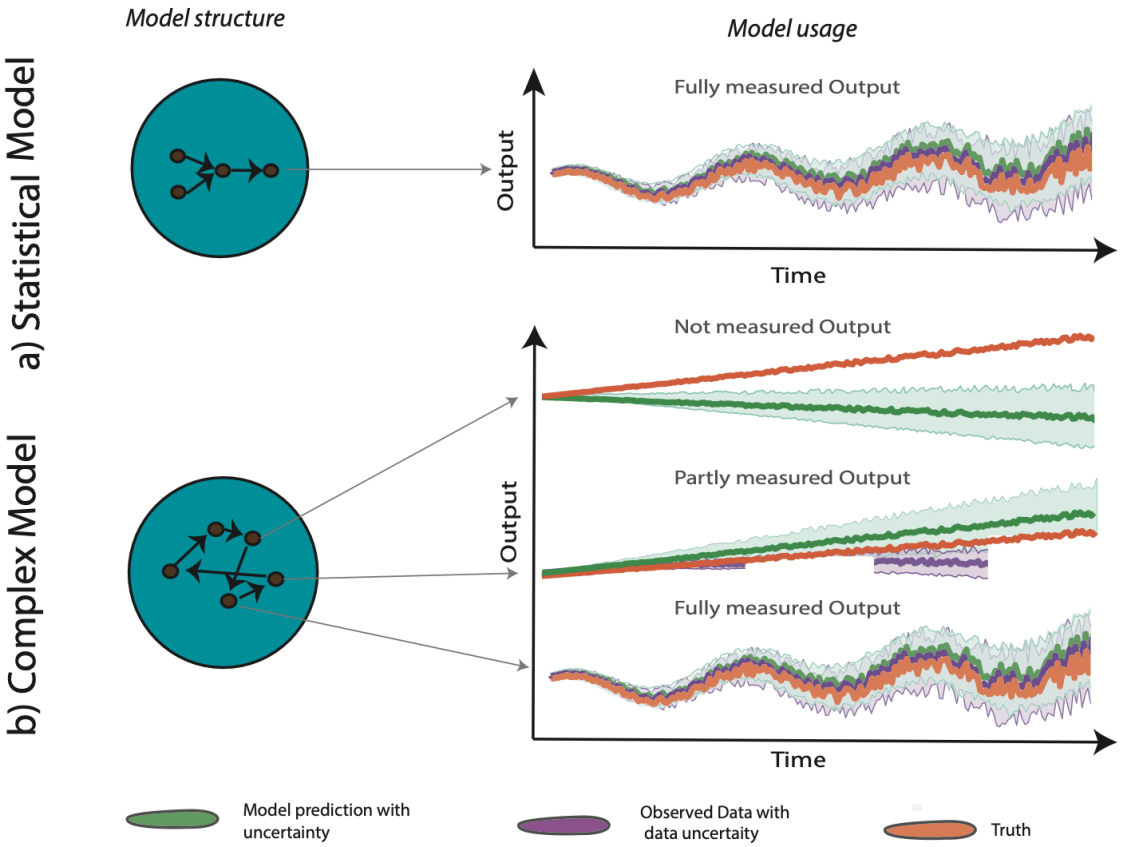


Fig. 6.1: A visualization of differences between complex computer simulations and statistical models. While statistical models are generally fit to only one response variable, complex computer simulations often predict multiple response variables and thus can be fit to multiple data sources, which may vary in sample size, and can be used to extrapolate to unobserved variables. Moreover, complex computer simulations typically have more variables that are in a more nonlinear and connected dependence structure. From these differences, we hypothesize that 1) Biased complex models will lead to biased parameter estimates and wrong predictions. 2) Standard calibration underestimates uncertainty and 3) both of these problems increase when calibrating against unbalanced data sets.

6.2.1 *Case study*

To provide a practical example of these problems, we examine the influence of structural model error when calibrating to multiple balanced or unbalanced data streams on predictions, parameters and uncertainty estimation in the calibration of the Basic forest model (BASFOR).

Model structure and introduced structural error

BASFOR simulates horizontal homogeneous forest stands by representing three biogeochemical cycles (carbon, nitrogen and water) as well as soil environment-interaction. It is driven by environmental data (atmospheric CO_2 concentration, solar radiation, air temperature, precipitation,

wind speed and humidity) to derive 17 state variable (nine tree-related and eight soil-related).

To examine the implications of structural error, we modified several key processes in BASFOR. Firstly, we changed the temperature dependence of NPP allocation (higher optimal temperature, fewer allowed deviations). Secondly, we made decomposition of litter temperature dependent. Thirdly, we changed dependence of water runoff to leaf-area-index (exponential quadratic instead of exponential linear). Fourthly, we weighted nitrogen allocation to tree components with their nitrogen use efficiency. Lastly, we made nitrogen leaching root-depth dependent. Although the exact location and nature of these modifications was somewhat arbitrary, we think of those modifications as realistic for structural errors that could also occur in real ecosystem models.

Statistical inference

We then used the original BASFOR model (henceforth called the "true" model) to simulate synthetic data with random observation errors (0.2) for daily observations of Gross Primary Production (GPP) and daily (balanced data streams) or ten-day (imbalanced data streams, so called because of an unbalance between the number of observations of GPP and ET) measurements of evapotranspiration (ET). Drivers for the simulation were climate data from 1920-2005 from Hyytiala (Finland) (Reyer et al., 2020).

Prior to the calibration, we conducted a sensitivity analysis of BASFOR. Based on the results, we removed insensitive parameters and three parameters that showed very high trade-offs with other parameters from the calibration by fixing them to their true values (the goal of this procedure is to speed up MCMC computations; see, e.g. Minunno et al., 2013). Because the true parameter values were known, no model error was introduced by this procedure, and validity of our further results is thus not affected by the parameter screening. In a real application, where "true" parameter values would be unknown, this procedure could introduce additional model error, which would further motivate the need to find methods to compensate for model error, such as the ones we present in this study.

We applied Bayesian inference (e.g. Van Oijen et al., 2011) to infer the values and uncertainties of the remaining six model parameters and the two standard deviation parameters of the observation model from the synthetic data. We specified flat (uniform) priors on the model parameters and vague gamma priors for the standard deviation parameters. We estimated posteriors with the Differential-Evolution Markov-Chain Monte-Carlo (Ter Braak and Vrugt, 2008) algorithm, implemented in the R package BayesianTools, (Hartig et al., 2019). To speed up computations, we generated initial values and the Z matrix with a differential evolution optimizer, (DEoptim, Ardia et al., 2016). We applied this procedure to both the "true" model and the model with structural error.

Quantification of the error in inference

To assess the effect of model error on the inference, we calculated the average error of parameter estimates by averaging the percentage difference between the "true" parameter and the calibrated parameter over the posterior (numerically calculated through N = 10000 samples from the posterior), the different parameters (P) and the five replicates (M).

$$\text{Parameter error} = \frac{1}{P}\sum_{i}^{P}\left|\frac{1}{M}\sum_{j}^{M}\frac{1}{N}\sum_{k}^{N}\frac{p_{i,j,k}-p^{\star}}{p^{\star}}\right| \tag{1}$$

Moreover, to assess the error of model predictions (also called time series error), we calculated the mean absolute error of data d_i and model prediction $m_i(x, \theta_j)$ (driven with climatic drivers

x and parameters θ_j) averaged over time (T), the posterior distribution (through N = 120 samples from the posterior) and five calibration replicates (M).

$$\text{Error} = \frac{1}{M}\sum_{j}^{M}\frac{1}{N}\sum_{k}^{N}\frac{1}{T}\sum_{i}^{T}|d_i - m_i(x,\theta_{j,k})| \tag{2}$$

Note that in most cases with structural model error, the error in the parameters and predictions was systematic, meaning that it can be interpreted as bias.

To relate the error to the estimated uncertainties, and thus examine if uncertainty estimates were reliable, we calculated error scaled to estimated uncertainty (ESEU) by dividing the mean error per day by the posterior standard deviation $\sigma_i(m_i(x,\theta_j))$, averaged over time, the posterior distribution and the five replicates.

$$\text{ESEU} = \frac{1}{T}\sum_{i}^{T}\frac{|d_i - \frac{1}{M}\sum_{j}^{M}\frac{1}{N}\sum_{k}^{N}m_i(x,\theta_j)|}{\sigma_i(m_i(x,\theta_j))} \tag{3}$$

A mean absolute error the same magnitude as the estimated uncertainty (standard deviation) will result in an ESEU of 1. Values substantially larger than one suggest that the estimation or prediction error is larger than the estimated uncertainty. For the model outputs and uncertainties, we differentiated between calibration and extrapolation domain.

Comparison between calibrating a "true" model and a model with structural error

The results of the calibration with the "true" model (without structural error) show that the error of the inferred parameters was virtually zero (< 0.02 %) for balanced and unbalanced data sets (Fig. 6.2a). In both of these cases, extrapolation and calibration error was small with narrow uncertainties (ESEU = 0.1) (Fig. 6.2b).

For the model with structural error, inferential errors were much larger (Fig. 6.2a). In particular, the parameter error was three times larger for the unbalanced data (~5%) compared to the balanced data (~1.7%) (Fig. 6.2a). Higher parameter error for the model with structural error led in all cases to higher time series errors compared to the correct model (Fig. 6.2b). For the balanced data set, the error for calibration was smaller than for extrapolation, while for the unbalanced data set this only was true for the high-resolution data (GPP). Moreover, GPP error was slightly smaller for the unbalanced than the balanced data set, but ET error otherwise. These errors led to a very high ESEU (Fig. 6.2b). This effect was stronger for the unbalanced data, especially for the undersampled data (ET) in the calibration domain (Fig. 6.2b).

These results support our theoretical expectations that calibrating with a correct structural model leads to unbiased parameter estimates, correct predictions and reliable uncertainty estimates, regardless whether the data is balanced or unbalanced. Introducing structural model error, however, led to erroneous parameter estimations (Fig. 6.2a), caused erroneous time series predictions and high ESEU (Fig. 6.2b), and these effects are intensified by unbalanced data sets (Fig. 6.2a,b).

6.3 A TOOLBOX FOR STATISTICAL INFERENCE IN COMPLEX COMPUTER SIMULATIONS

After having confirmed our intuition that statistical calibrations of CSS are highly susceptible to structural error, we turn our attention to possible solutions. Few general treatments of the problem exist in literature, but there are certain strategies and suggestions that are frequently used in practice. To deal with the problem of imbalanced data, many studies rebalance or reweight data streams. Remaining model-data discrepancies (bias) have sometimes been addressed by introducing data-driven models to the process-model after or during the calibration. In the following, we will discuss these potential solutions and test their applicability in our case study.

6.3.1 *Weighting of data streams*

The strategy of rebalancing and reweighting data addresses the issue that standard statistical methods weight the importance of each data stream principally by its content of independent observations. While the latter is perfectly sensible for a correct model, it will lead to distortions towards the model output with more data when structural error makes it impossible to fit both data streams at the same time.

Case study - weighting of data streams

To examine possible benefits of weighting for our case study, we down-weighted the likelihood for the GPP data with 1/10, the ratio of ET to GPP observations, thus giving both data streams the same weight (for details, see Supporting Information S4, section 4.1). Weighting the data streams increased the error for the estimated parameters of the correct model (Fig. 6.3a) by a small amount, which propagates through the model into a small error in predictions and a higher ESEU (Fig. 6.3b). For the model with structural error, introducing weights in the likelihood decreased parameter error leading to smaller ET error, but slightly increased GPP error (Fig. 6.3b). Moreover, the ESEU of ET in the calibration domain is smaller due to a reduction of ET error. Overall, we can thus conclude that weighting slightly decreased the inferential performance for the correct model, but dramatically improved the performance for the model with structural error.

6.3.2 *Bias correction after calibration*

Another option to deal with model error is statistical bias correction. The simplest approach is to fit flexible statistical or machine learning models post-hoc (i.e. after the model has been fit) to the residual errors (but see Beyer et al., 2019). The logic here is that if the model makes the same error under similar conditions, we can learn this error and apply corrections to future predictions (called "time invariance" Ehret et al., 2012). Obviously, this method only corrects predictions and not the parameter estimates, as the actual inference remains unchanged.

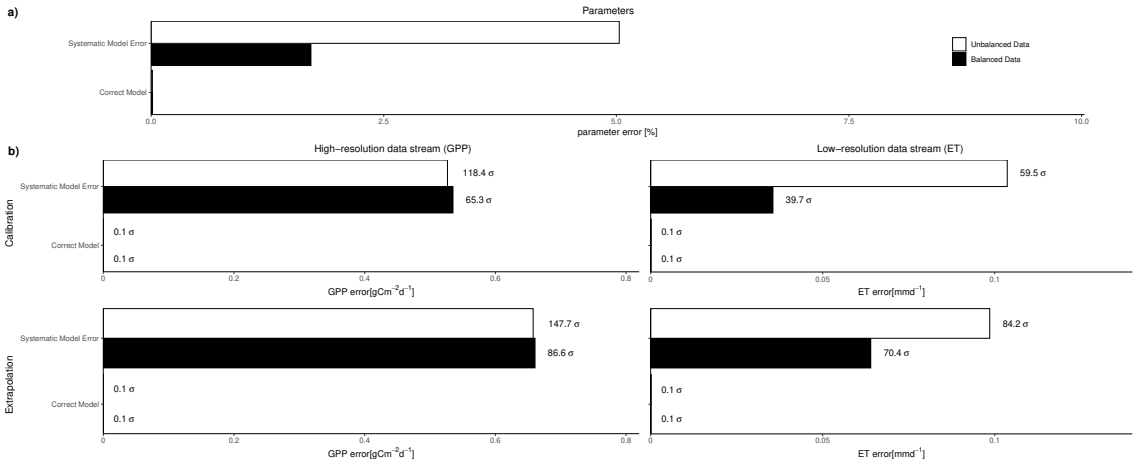


Fig. 6.2: Performance of the model with and without structural model error for balanced and unbalanced data. The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty (i.e. the error of the model which can be explained due to a high estimated uncertainty). The case study indicates that structural model bias leads to a) parameters with serious errors, b) erroneous model outcomes and high error scaled to estimated uncertainty.

Case study - bias correction after calibration

To test this method, we used a flexible Gaussian process (GP) model from the kernlab package (Karatzoglou et al., 2004) with a distance-based covariance structure (for details see Supporting Information S4, section 4.2). We fitted the model to approximately six years of residual errors as a response, and the corresponding model drivers (e.g. temperature and humidity) and CCS output as predictors, and extrapolated the error to future predictions. Our results show that this approach decreased the predictive GPP error of the model with structural error by similar amounts in the calibration and extrapolation periods (Fig. 6.4). ET error was approximately the same between the corrected and uncorrected versions of the model with a structural error, but there was a large decrease in ESEU (Fig. 6.4), not only caused by reduced error, but mostly by the variance coming from the explicitly modeled model error. Applying the same method to the true model introduced a slightly larger error in the time series and increased ESEU (Fig. 6.4). We speculate that this is due to the GP overfiting on random error.

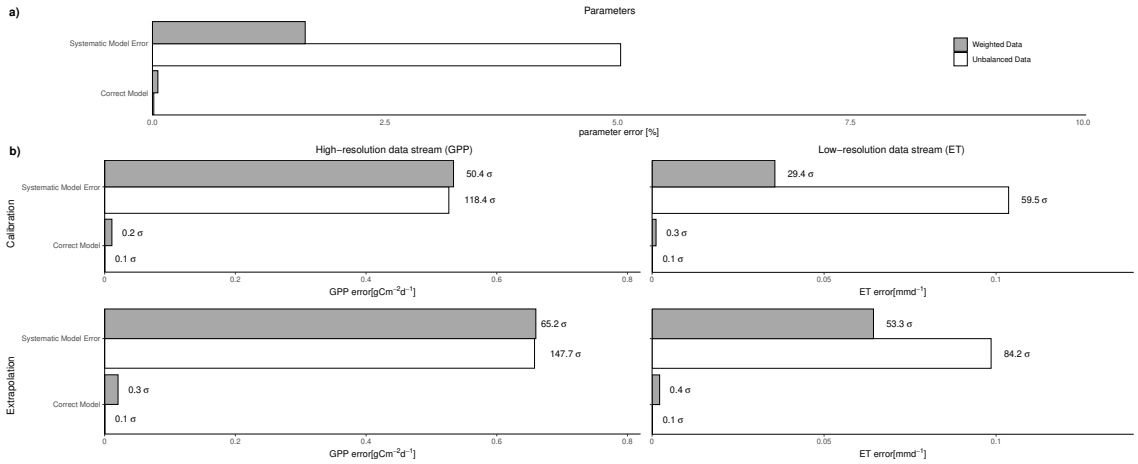


Fig. 6.3: Comparison of the performance of the model with structural model error and the correct model for weighted and unbalanced data. The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty. The case study indicates that weighting the data streams decreases a) parameters error, b)shifts error in model outcomes and improves ESEU.

6.3.3 *Bias correction during calibration*

A second option is to perform the bias correction within the calibration. A common example of this is the Kennedy-O'Hagan (KOH) approach (Kennedy and O'Hagan, 2001). In this approach, we fit again a GP for the bias together with the other model parameters in the same likelihood:

$$L(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}[d - (m(\theta, x) + GP(x, m))]^2\} \quad (4)$$

Here, σ is the standard deviation of the observational error, GP a Gaussian process. While the advantage of this approach is that the bias correction can also improve the inference on the model's parameters, the drawback is that it may suffer from an identifiability issue between parameters and model error. Whether this problem occurs depends on how distinct the structure of the process and the error model are. Note also that multiple data streams can be helpful in this regard, because they would typically impose independent constraints on the process model. Moreover, it has been shown, that incorporating suitable prior knowledge about the model error (e.g. smooth with respect to some predictor variables) allows the KOH method to separate between parameters and model error (Brynjarsdóttir and O'Hagan, 2014). Because of these attractive properties, there are a sizeable number of studies which have tested and modified this approach (e.g. Higdon et al., 2004; Higdon et al., 2008; Goldstein and Rougier, 2009; Tuo, 2017; Tuo and Wu, 2016).

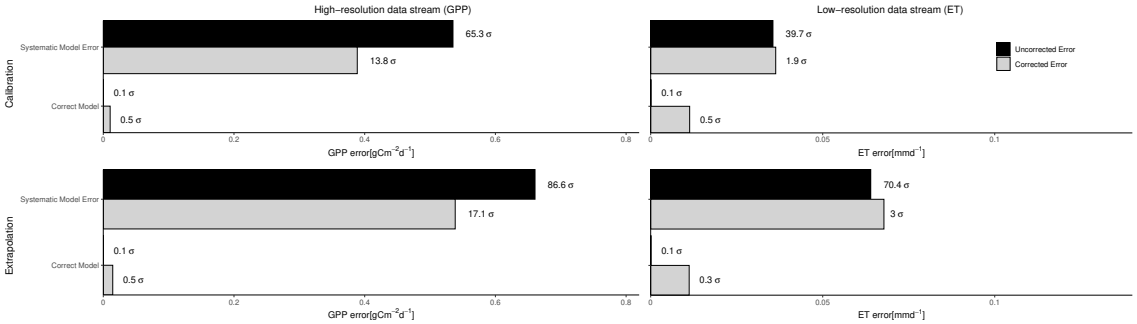


Fig. 6.4: Comparison of the performance of the model with structural error and the correct model fitting a correction term after calibration.The bars reflect error in absolute values and numbers reflect the error scaled to estimated uncertainty. The case study indicates that correcting the data streams decreases error in model outcomes and decreases ESEU.

Case study - bias correction during calibration

In its original version, the KOH method fits the GP against all calibration data with all drivers and state variables as predictors. However, as the computational cost of GP fitting and evaluation scale unfavourable with the number of data points, this makes it more difficult for typical environmental model calibrations. The computational problems occur because the calculation of the GP requires an inversion of a large covariance matrix. Moreover, the KOH method assumes having enough observational data of model determining variables (model state and external drivers) to fully constrain the Gaussian process (Kennedy and O'Hagan, 2001), which for typical ecological models is not a realistic assumption (in our case study we do not have virtual measurements of any state variables, we measured only the fluxes GPP and ET).

For our case study, we propose an alternative variant of the KOH method, which makes three changes to decrease computational cost. Firstly, we only use the drivers and the observed values as predictors. Secondly, we calibrate against a subsample of data (in our case we subsample to 10% of the data, the last 8 years of data and drivers as best proxies for future drivers). We do so because, typically models' systematically predict a GPP that is too small on warm summer days and ET that is too high when humidity is low. Thirdly, we avoid the costly inversion of the covariance matrix that is only needed to match GP parameters to their prior by approximating the inverse covariance by its diagonal, while still infering the full covariance matrix (rbfdot kernel) in the likelihood. To code a preference for explaining the data by the process-model, we apply a regularizing gamma(2,0.1) prior with high probability weight near zero on the diagonal. Based on the GP predictions, we calculate model-data discrepancies for the rest of the time series (a detailed tutorial is given in Supporting Information S4, section 4.2).

When applying bias correction during calibration, parameter error stayed near zero for the correct model, and decreased for the model with structural error (Fig. 6.5a). However, whereas time series error decreased in both outputs, for the model with structural error, for the true model, error increased (Fig. 6.5b), with an almost identical pattern to the post-hoc GP (Fig. 6.5b). For the model with structural error, the calibration resulted in higher estimated uncertainty and thus lower ESEU compared to a calibration without an explicit model error term (Fig. 6.5b). Overall, the method improves parameters, predictions and ESEU for the model with structural error, but decreases the performance for the correct model.

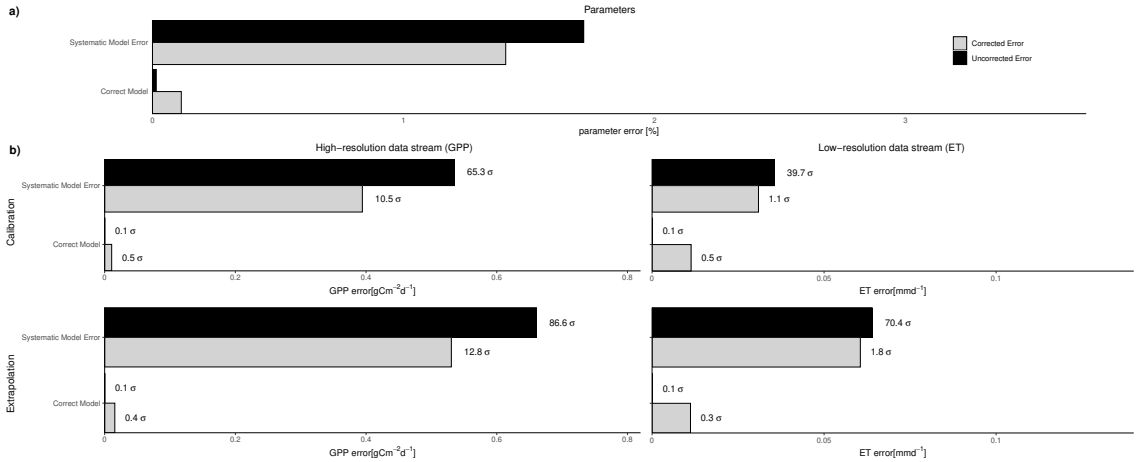


Fig. 6.5: Comparison of the performance of the model with structural model error and the correct model for a correction during calibration. The bars reflect error in absolute values and numbers reflect the uncertainty in units of standard deviation. The case study indicates that correcting error during the calibration decreases a) parameters error for the wrong model, b) reduces error in the model outcomes and improves uncertainty estimation.

6.3.4 *Correcting processes rather than outputs*

We have seen so far that correcting bias on the model outputs can improve predictions and inference. The true error, however, is not on the outputs, but in the model processes themselves. It therefore seems obvious to explore if the processes themselves could be bias-corrected. For simple population models, this idea has been suggested under the name "partially specified ecological models" (Wood, 2001). The drawback of this approach for CCS is that the complexity of the error term and therefore the issue of identifiability increases significantly if errors in all possible subprocess are considered. For our case study we attempted to correct process-errors directly via a state-space approach (details see supporting information S4, section 4.3), but did not succeed in improving the statistical inference in this way. Nevertheless, we believe that this is worthwhile for further research, in particular because it would not only correct errors, but also allow to identify their location.

6.4 DISCUSSION

CCS are increasingly used in ecology, evolution and earth system sciences. Our ability to confront these models with data and to estimate uncertainties in parameters and predictions is critical for their utility.

In this contribution, we highlighted that certain issues emerge when using standard statistical methods to calibrate CCS. Most importantly, our theoretical explanations as well as our case study demonstrated that naive applications of standard calibration methods to imperfect computer simulations can lead to biased parameter estimates and predictions, and to underestimated uncertainties (Fig. 6.2), and that these biases are more pronounced than in flexible statistical models. These issues are particularly severe when calibrating against unbalanced data (Fig. 6.2). Weighting of data streams can reduce the aggravating effect of unbalanced data (Fig. 6.3). Data-driven models can be used to describe and remove the remaining bias after or during the calibration. In our case study, fitting model bias with a GP after calibration improved time series predictions (Fig. 6.4). Thus, our results show that robust methods exist for ameliorating negative consequences of structural model for making predictions with calibrated CCS.

Using a GP during calibration can additionally improve parameter inference (Fig. 6.5). We acknowledge that the interpretation of parameter values across structurally different models is tricky, because those parameters have different meanings in the respective models, and thus, one could argue that both the true and the model with structural error have parameters that are correct under their respective assumptions. This view neglects, however, that researchers will tend to interpret parameter values as if their models were structurally unbiased, and representation of the true process. Our comparison of the estimated and the true parameter therefore measures to what extent this interpretation is justified and shows that explicitly modeling structural error increases the chances of model parameters representing their real values (Goldstein and Rougier, 2009).

Our results regarding the consequences of model error are qualitatively supported by the few earlier studies that have looked at the problem (for balanced data by White et al., 2014; and for unbalanced data sets by Abramowitz et al., 2008). In general, however, this topic seems surprisingly underappreciated in the statistical literature. We speculate that most statisticians do not operate with large system models, and the modelers that do are not primarily interested in statistical methods. Nevertheless, a good understanding of these issues is urgently needed, as many important forecasts rely on correct identification of parameters and their uncertainties. In the next subsections, we summarize our conclusions from existing literature and our new simulations, provide practical guidance for their use, and delineate a statistical research program to develop a theory of robust inference for CCS.

6.4.1 *Which methods work to improve inference for biased system models?*

To achieve a more balanced impact of the different data on the calibration, many modelling studies weigh data streams. Despite its popularity, very few studies have examined the justification for this practice. Contrary to Wutzler and Carvalhais (2014), who only found minor improvements, we found that weighting improved all considered performance measures (Fig. 6.3). Different CCS and a different severity of model error may explain the differences of the two studies. In general, benefits from weighting likely depend on the statistical context, the weighting strategy, and the model error. Overall, however, we believe that weighting is a useful and conservative strategy if structure model error is suspected.

One open question that would profit from more research is how the weighting of different data streams should be performed. Creating balance by upweighting the less abundant data stream, which essentially corresponds to the common practice of oversampling in machine learning, could lead to a serious underestimation of uncertainties as it is equivalent to using the same data multiple times. Downweighting, the far more common approach in studies calibrating CCS, is more conservative, but it also artificially decreases the information in the more abundant data stream to the level of the less abundant stream, which can hardly be optimal to get realistic uncertainties. In general, these two options represent the extremes of a broad spectrum of possibilities, and more research is required to understand how an optimal weighting could be justified. An option to avoid the problem would be to calibrate against patterns, as suggested by the POM (Grimm, 2005), to independently update subsets of parameters against different data streams (Wutzler and Carvalhais, 2014), or to set up subjective likelihoods (White et al., 2014), as in the GLUE approach (Beven and Binley, 1992). The downside, however, is that these approaches could be considered even more subjective than weights on the data streams.

A complementary class of methods directly addresses the issue of model error, by identifying and correcting structural biases from model's predictions. In our case study, this approach (via the KOH method) improved parameters, predictions and uncertainty quantification (in line with Brynjarsdóttir and O'Hagan, 2014). However, the standard KOH method has two main challenges - high computational complexity (Conti and O'Hagan, 2010) and possible identifiabil-

ity issues between model parameters and model error (Brynjarsdóttir and O'Hagan, 2014). We addressed the first problem by only using a fraction of the available data to fit the GP and extrapolated to the remaining calibration domain. We speculate subsampling works for models with mechanistic structure, as long as the learned discrepancy will behave similarly in the future. We appreciate that using a fraction of the calibration data potentially disregards useful information, and that our additional numerical approximations could further reduce the method's performance. The fact that we reduced the model error, however, suggests that these problems are probably mild. Still, in situations where computational costs are not limiting, it would be better to use the original method suggested in Kennedy and O'Hagan (2001). The issue of identifiability is important, but arises in many statistical situations, and several strategies exist to deal with it, for example regularization or informative priors (Brynjarsdóttir and O'Hagan, 2014). Thus, we think these methods can lead to better predictions for ecological CCS and modellers should be using them.

A limitation of our case study is that it tested validity and effectiveness for one specific model, with one specific error structure. While we do think that the chosen example is typical and representative for the field, it would be useful to explore the generality of our results in future studies and their robustness to observation errors and uncertainties, which can be expected to exacerbate statistical problems.

Finally, all our successful examples used bias corrections on model outputs. In particular when making predictions, these implicitly assume that the model error is stationary, which is unlikely to be true (Chen et al., 2015). It would therefore be preferable to move bias corrections directly inside the modelled processes. In our case study, we attempted such a correction with a state-space approach, but could not achieve an increase in inferential performance. It is possible that idiosyncrasies of our setup were responsible for this negative result, but it seems equally plausible that corrections on the outputs are already at the limit of what can be sensibly inferred from data. Either way, these considerations suggest that bias corrections are currently no panacea, and that careful improvements of the model structure, if possible, are still the preferable solution.

6.4.2 *Practical Suggestions*

As famously noted by Box (1976): "All models are wrong, but some are useful". Accepting this fact, the question for CCS is what type of error is dominant. If statistical error dominates the structural error (this can be checked by an analysis of residuals, see Supporting information S4, section 4.2), all standard statistical techniques work fine, regardless of the balance of data. In this case, using methods that accounting for possible structural model errors tends to somewhat increase uncertainties (Figs. 6.4 and 6.5, see recommendations Fig. 6.6). When structural error dominates, however, severe statistical problems can arise, in particular for imbalanced data. In this case, weighting of data streams or adding bias correction to the CCS can improve the outcomes of a model calibration dramatically. Our recommendation for modelers with little statistical background is that downweighting imbalanced data is a simple, conservative approach that can alleviate some of issues created by structural error. Although it is somewhat ad-hoc, it improved results in our case study, and it makes uncertainty estimates (e.g. confidence intervals) more conservative. For more experienced modellers, we propose to consider additional bias corrections after or during calibration, or even consider if bias corrections can be moved inside the processes, which would not only improve the inference, but also model understanding. For all these purposes we provide sample code (https://github.com/JohannesOberpriller/Oberpriller-et-al-2021).

Cases	Statistical error dominates, balanced or unbalanced data	Structural error dominates, balanced data	Structural error dominates, unbalanced data
Naive use of standard methods	Standard methods are sufficient for parameters, predictions and uncertainty quantification	Standard methods lead to biased predictions and parameters.	Standard methods lead to a higher bias in parameters, predictions and uncertainty estimation
Our recommendations	1. Standard methods are sufficient	1. Bias correction after calibration improves predictions 2. Bias correction during calibration additionally improves parameters	1. Weighting reduces bias by a lot 2. Bias correction after calibration improves predictions 3. Bias correction during calibration additionally improves parameters
Remarks	-	Bias correction has high computational costs	Bias correction has high computational costs

Fig. 6.6: The different situations in environmental model calibration and our suggestions for improving model performance. The two main factors, which need to be taken into account are the data situation (balanced or unbalanced) and the sources of error (random or structural). This general advice can slightly change in different situations as model complexity and computational demand strongly depend on the CCS, domain of extrapolation and number of data streams. Overall performance will become worse with increasing observational error for all methods including standard calibration.

6.4.3 *Towards a statistical theory for robust inference in complex computer simulations*

More broadly, our paper highlights that structural model error raises specific problems for statistical inference with complex computer simulations. This should alert the ecological community that model error is a real problem for the calibration of CCS, and naively applying standard statistical methodologies does not always lead to the desired results.

Although we did a step into the direction of robust inference in CCS by reviewing proposed solutions, explaining their theoretical justification and providing practical guidance for their application, further work is required to arrive at a general solution for robust statistical inference. For example, we have no good theory about how to set weights for different data streams. When considering a data stream with only one observation, it becomes clear that downweighting to the least common data stream is likely not always optimal. Moreover, it would be interesting to extend bias corrections also to methods that use simulation-based inference, such as Approximate Bayesian Computing (ABC) or synthetic likelihood (Hartig et al., 2011; Csilléry et al., 2010).

A last point is that statistical bias corrections are important for improving the inference, but the correct model still consistently performed best in our case study, and we should thus also think about how to develop methods to track down the location of the error. To localize errors, one could start by analyzing model discrepancies for patterns, and use those to attempt a rough localization of the structural error. Moreover, we speculate that when a dramatic change of a parameter value between KOH and standard calibration happens, this gives a hint that model error affects this specific parameter and thus that model error is "near" to this parameter. Then using time-dependent parameters (instead of constant, see Reichert and Mieleitner, 2009) could be an option to get a better localization of the error. Another idea (Wood, 2001) goes a step further, by saying that flexible models (generalized additive models) should account for the processes, or Reichstein et al. (2019), which propose to learn entire submodels. These approaches should be tested in practice to finally improve model performance.

7 DISCUSSION

Our research focused on three main questions around the use and application of forest ecosystem models for ecological forecasting and inference. The first one was: What are the main contributors of uncertainty in forest ecosystem models? Can we use uncertainty analysis and calibration of forest ecosystem models to analyze ecological patterns on environmental gradients? The next question deals with the remaining variance: To what extent can random effects be used to represent ecological variation and how much data points are required to estimate these variations precisely? And the last question investigates if the findings above are robust when we have structural model error: What are the consequences and solutions of calibrating of and projecting with models with structural errors?

Based on a short summary of our main results I will elaborate on their relevance and significance. To conclude this chapter I will point out further directions of research that are arising from our findings and possible approaches to address them.

7.1 MAIN RESULTS

The main results of this thesis are:

Calibration and uncertainty analysis over large environmental gradients reveal ecological patterns: In chapter 3, we show that environmental drivers are the main contributors of predictive uncertainty in projections of forest ecosystem models and that environmental drivers also modifiy sensitivities and thus uncertainties of model parameters along environmental gradients and thereby find evidence for ecological principles and hypotheses like the stress-gradient hypothesis. In chapter 4, we show that with a calibration of a forest ecosystem model we can infer intraspecific variation across environmental gradients in line with experimental studies. These findings demonstrate the ability of forest ecosystem models for testing ecological principles and hypotheses along large environmental gradients.

Random effects can absorb unexplained variance, but require sufficient data points: In chapter 4, we show that accounting for intraspecific variation via a random effect on the sites in the calibration of an ecosystem model improves the model-data fit. In chapter 5, we show that with a low number of levels in a grouping variable the decision to model the grouping variable as fixed or random effect leads to different statistical properties of the estimator. Moreover, we show that to precisly estimate the variance of the random effect 5-7 levels in the grouping variable are required. These findings underline the importance of random effects to absorb unexplained variance, but that estimating the variation requires a sufficient number of data points.

Accounting for the structural model error of the model itself has a significant impact on model projections: In chapter 6 we show that structural model uncertainty leads to increased errors and overconfidence in predictions. These effects are amplified when the model is fitted to unbalanced data. Accounting for structural model error and data imbalance leads to a reduction in error, but more importantly to a drastic improvement in the quantification of uncertainty.

7.2 DISCUSSION OF RESULTS

In the following I will discuss how our main results can help to clarify our research questions.

7.2.1 *Environmental conditions as contributors and modifiers of sensitivity and uncertainty along environmental gradients allow to test ecological patterns with numerical methods applied to forest ecosystem models*

Forest ecosystem models are usually used to project forest ecosystems under climate change into the future (e.g. Reyer, 2015). For these projections, environmental conditions are not only the biggest uncertainty contributors, but they also modify the uncertainty of other processes (chapter 3). We can use this fact to test big-scale patterns along large environmental gradients without the need to collect additional data. Collecting data along large environmental gradients for specific questions would be a high financial and time-consuming effort. We can avoid at least a part of these efforts by applying numerical experiments and looking at models behavior at large environmental gradients. For example, we found evidence for the stress-gradient hypothesis with the LPJ-GUESS model in chapter 3 or that variations of the optimal growth temperature in *P. abies* correlate with mean annual temperature in chapter 4. Overall, the potential of numerical experiments with forest ecosystem models to test and infer ecological patterns instead of purely projecting is not yet exhausted.

The idea that we can use mechanistic models describing microscopic processes and upscale these to macroscopic patterns that are applicable on much larger scales is not a new one as the entire field of statistical physics is built on it (e.g. Landau and Lifshitz, 2013). A related idea in ecology is pattern oriented modeling (Grimm, 2005), which proposes to calibrate models such that they are able to reproduce patterns instead of rigorously optimizing a likelihood. Having these approaches in mind, it is surprising that the primary use of forest ecosystem models is projection instead of inference of hard to measure characteristics of the described systems. With already existing mechanistic models and an increasing amount of data, we can calibrate forest ecosystem models and infer specific ecological information (via hierarchical Bayesian modeling, also proposed by Laubmeier et al., 2020) along large gradients without additional costs.

7.2.2 *Random effects to account for biological variation in forest ecosystem models*

Models of biological systems often describe hierarchies of processes and patterns (e.g. Levin, 1998), but ignore a fundamental driver of evolution: that processes and pattern vary in time and space. For example forest ecosystem models often have as a smallest unit species or plant functional types, which act as homogeneous entity, but they do not account for intraspecific variation. We can represent these variations via random effects. This has the benefit, that, when data is coupled to models, the data decides about the amount of variation (Van Oijen, 2017). When using random effects to reflect intraspecific variation (see chapter 4), we were able to decrease the error in the validation as well extrapolation domain, which indicates that random effects can explain additional unexplained biological variation (Dietze et al., 2008).

From a technical Bayesian view, we can interpret random effects as additional prior on the similarity of parameters (e.g. Higgins et al., 2009), because random effects are usually parametrized to pull parameter estimates to the same mean. In our example from chapter 4, the effect that parameter estimates should be similar reflects the species concept as individuals from the same species have similar, but not exactly the same traits. Because random effects let the data rather flexibly decide about the amount of variation, random effects transform a part of the non-parametric structural uncertainy (the missing representation of intraspecific variation) into parametric uncertainty and thereby allow to represent a priori ecological knowledge about forest ecosystems.

However, with a small data set it might be better not to model the biological variation because of the additional complexity of estimating the amount of variation from the data (e.g. the standard deviation of the normal distribution in the random effect). We have analyzed this specific situation for (generalized) linear regression models in chapter 5. These findings indicate that

we should only model biological variation as random effect with at least 5-7 data points per organizational level to get a precise estimate of the amount of variation. With many levels of organization, a typical situation for forest ecosystem models (e.g. individuals per plot, plots per species, species per plant function type), there might be too little data. To decide when there is enough data to model biological variations as random effects in forest ecosystem models, existing rules of thumb from mixed-effect models should be used (for an overview of related challenges see Silk et al., 2020).

7.2.3 *Sources, consequences and solutions of projecting with structural model error in ecological modeling*

Forest ecosystem models are, due to a lack of knowledge and computational restrictions, inevitably approximations of the underlying natural system. These approximations lead, as discussed in chapter 6, to structural model uncertainty and can have severe consequences when not accounted for.

The practical question for ecological forecasting and inference, however, is how influential model structural error is for the purpose of the study. In our numerical experiments, we found mixed results. The sensitivities, uncertainties and their interactions on environmental gradients in chapter 3 and the quantification of intraspecific variation in chapter 4 were in agreement with experimental studies and thus would suggest that practically applied forest ecosystem models are structurally realistic for the purpose we have used them. However, the biomass and net primary production projections in chapter 4 were unexpected and could be a sign that additional flexibility in the model structure is used to compensate for structural model error. Overall, our results suggest that models approximate observed patterns in the data domain well (regardless of whether manually (LPJ-GUESS, chapter 3) or automatically (3-PG, chapter 4) calibrated), but structural model error becomes a more severe problem when extrapolating to non-analogous situations.

A particular example for such a non-analogous situation are climate change impact projections. For such projections, the most alarming result of chapter 6 is that without accounting for structural model error, we are overconfident in our predictions. This is particularly worrying, because many studies rely on just one model (for an example see Ahlström et al., 2012). When the uncertainty in such a study is underestimated, we might miss potential irreversible effects such as tipping points (Scheffer, 2010). The IPCC tackles this issue by using ensemble projections, which deal with parametric and structural uncertainty at the same time (IPCC, 2014). When we have models with a quite different structure, this will reduce bias and lead to an improved quantification of uncertainty (Schomaker, 2012). However, such a model average approach can only lead to unbiased results, when the available models sample the full space of structural uncertainty (e.g. Fragoso et al., 2018), an elusive situation in practical applications.Thus, we have to work on a better representation of structural uncertainty for a single model, similar to what we discussed in chapter 6, or think about ways to better represent the space of structural uncertainty for ensemble predictions.

7.3 CONCLUSIONS AND OUTLOOK - TOWARDS MORE RELIABLE PROJECTIONS AND INFERENCE IN FOREST ECOSYSTEM PROJECTIONS

In conclusion, we have shown that most uncertainty about future projections with ecosystem models are induced by environmental conditions (chapter 3) and that we can use forest ecosystem models to infer patterns along large environmental gradients without the need for further costly data collections (chapter 3, chapter 4). When there is a sufficient amount of data (chapter 5), we can also infer biological variation via random effects (chapter 4). However, when the for-

est ecosystem model has structural error we are overconfident in the results and thus need to account for structural error during calibration (chapter 6).

When we want to have reliable and robust projections with a correct quantification of uncertainty, we have to work on all parts of projecting with forest ecosystem models.

7.3.1 *Focus on likely input scenarios*

Forest ecosystem models are driven by climate scenarios, which often only represent best (RCP 2.6) and worst case (RCP 8.5), but have no probabilities assigned. This is understandable because of the high uncertainty in never-observed and hard to estimate processes (e.g. the CO_2 and methane released from the tundra permafrost), the inherent uncertainty in anthropogenic CO_2 emission and the difficulty in the communication with stakeholders (Hausfather and Peters, 2020). Additionally, a probabilistic assignment would require cooperations of social science, political science, policy makers as well as industry and climate science and thus be an immense effort.

However, in chapter 3 it is shown that environmental conditions induce most uncertainty into forest ecosystem projections. Thus, it is necessary that ecological actors dealing with practical consequences of climate change, e.g. forest owners or farmers, can plan ahead with the most likely future conditions instead of scenarios with unknown probabilities. How is a forest owner supposed to assess probabilities in scenarios and decide based on these probabilities which species he should plant under ongoing climate change, when science does not want to do so? One argument often given for not trying to estimate probabilities is that uncertainty reduces public faith in science, however, there are also contrary studies (e.g. Bles et al., 2020). When probabilities are communicated, best practices for forest owners or farmers can be provided, which can enhance the number of people taking appropriate actions (Clar et al., 2013) to maintain forest ecosystem services under global and climate change.

7.3.2 *Avoiding structural error due to more model flexibility*

Increasing data availability and computational power allow ecologists to build forest ecosystem models in another and new way compared to the time when modelers began to build the current state-of-the-art models, 20-30 years ago (Farley et al., 2018). At that time computational restrictions forbid to assume a flexible model and let the data decide which flexibility and processes best fit the underlying reality. By now it is possible to do this with machine- and deep learning and automatically collected data helps to constrain these models (Reichstein et al., 2019).

In this context, physics-based deep learning (Thuerey et al., 2021) is especially interesting as it allows to connect deep learning with already existing knowledge on ecological processes. Particularly, I can imagine that state variables, a conceptual cornerstone of most forest ecosystem models, are connected to each other with known functional relationships that are already part of the state-of-the art models while their propagation to new states are simulated with neural networks. With this combination of process-based and flexible networks, a more precise forecast of forests under climate change might be possible.

However, identifying the necessary and sufficient processes that we must provide to make these models transferable to new environmental conditions is a difficult task (Reichstein et al., 2019). In a top-down approach we might start to exchange a small number of equations or submodules and then test the resulting model and only move forward to new modules if the model fits the historic data well enough. As this process will take a while, we also have to improve calibration and uncertainty quantification in the meantime.

7.3.3 *Outlook: Further steps in model-data integration*

Although there are methods allowing to account for structural model error during calibration there is still no completely satisfying solution that makes projections with forest ecosystem models totally reliable. A first step towards such a solution would be to allow for more flexibility in the design of calibration algorithms, especially in a hierarchical framework. To make this more convenient for ecologists, a crucial step would be to provide a user-friendly interface like it is done for GLMMs (for GLMMs see Bates et al., 2014; Brooks et al., 2017). Another important step would be to go into the model itself instead of correcting the outputs. We have discussed this issue in more detail in chapter 6. A third step, which requires additional research, is to more extensively use near-term forecasting (Dietze et al., 2018). In the best case one could do this on a data-model platform, which automatically updates all registered models and their predictions as soon as new data becomes available. The update could for example be done via recursive particle filters (for an example see Clark et al., 2022). A last important step would be to design forest ecosystem models modular, such that individual modules from one model can be plugged into another model. With such a data-model platform and a modular interchangeable design of the existing ecosystem models one could also build surrogate models that possibly allow for an realistic sampling of "the structural uncertainty space", i.e. the space of plausible model structures. This would then lead to realistic estimates of the structural uncertainty in forest ecosystem models and via model averaging to more precise forest projections under climate change.

When all these steps are implemented, forest ecosystem response to global and climate change can be more precisely predicted and uncertainties can be more realistically assessed. Realistic predictions with quantified uncertainties are necessary for the fight against climate change and so are advances in forest ecosystem modeling.

BIBLIOGRAPHY

Aarts, Emmeke et al. (2014). "A solution to dependency: using multilevel analysis to accommodate nested data." In: *Nature Neuroscience* 17.4, pp. 491–496.

Abramowitz, Gab et al. (2008). "Evaluating the Performance of Land Surface Models." In: *Journal of Climate* 21.21, pp. 5468–5481.

Ahlström, A. et al. (2012). "Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections." In: *Environmental Research Letters* 7.4, p. 044008.

Aitken, Sally N. and Jordan B. Bemmels (2016). "Time to get moving: assisted gene flow of forest trees." In: *Evolutionary Applications* 9.1, pp. 271–290.

Albert, Cécile Hélène et al. (2010). "Intraspecific functional variability: extent, structure and sources of variation." In: *Journal of Ecology* 98.3, pp. 604–613.

Alberto, Florian J. et al. (2013). "Potential for evolutionary responses to climate change – evidence from tree populations." In: *Global Change Biology* 19.6, pp. 1645–1661.

Ali, Hamada E. et al. (2017). "Effects of plant functional traits on soil stability: intraspecific variability matters." In: *Plant and Soil* 411.1, pp. 359–375.

Almeida, Auro C et al. (2004). "Parameterisation of 3-PG model for fast-growing Eucalyptus grandis plantations." In: *Forest Ecology and Management*. Synthesis of the physiological, environmental, genetic and silvicultural determinants of the growth and productivity of eucalypts in plantations. 193.1, pp. 179–195.

Andrieu, Christophe et al. (2003). "An Introduction to MCMC for Machine Learning." In: *Machine Learning* 50.1, pp. 5–43.

Ardia, David et al. (2016). "'DEoptim': Differential Evolution in 'R'." Version 2.2-4. In.

Arnqvist, Göran (2020). "Mixed Models Offer No Freedom from Degrees of Freedom." In: *Trends in Ecology & Evolution* 35.4, pp. 329–335.

Atkins, K. E. and J. M. J. Travis (2010). "Local adaptation and the evolution of species' ranges under climate change." In: *Journal of Theoretical Biology* 266.3, pp. 449–457.

Augustynczik, Andrey L. D. et al. (2017). "Productivity of Fagus sylvatica under climate change – A Bayesian analysis of risk and uncertainty using the model 3-PG." In: *Forest Ecology and Management* 401, pp. 192–206.

Aza, Celia Herrero de et al. (2011). "Carbon in heartwood, sapwood and bark along the stem profile in three Mediterranean Pinus species." In: *Annals of Forest Science* 68.6, p. 1067.

Baayen, R. H. et al. (2008). "Mixed-effects modeling with crossed random effects for subjects and items." In: *Journal of Memory and Language*. Special Issue: Emerging Data Analysis 59.4, pp. 390–412.

Bagnara, Maurizio et al. (2019). "An R package facilitating sensitivity analysis, calibration and forward simulations with the LPJ-GUESS dynamic vegetation model." In: *Environmental Modelling & Software* 111, pp. 55–60.

Balaman, Şebnem Yılmaz (2019). "Chapter 5 - Uncertainty Issues in Biomass-Based Production Chains." In: *Decision-Making for Biomass-Based Production Chains*. Ed. by Şebnem Yılmaz Balaman. Academic Press, pp. 113–142.

Barabás, György and Rafael D'Andrea (2016). "The effect of intraspecific variation and heritability on community pattern and robustness." In: *Ecology Letters* 19.8, pp. 977–986.

Barman, Rahul et al. (2014). "Climate-driven uncertainties in modeling terrestrial gross primary production: a site level to global-scale analysis." In: *Global Change Biology* 20.5, pp. 1394–1411.

Barr, Dale J. et al. (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal." In: *Journal of Memory and Language* 68.3, pp. 255–278.

Basler, David (2016). "Evaluating phenological models for the prediction of leaf-out dates in six temperate tree species across central Europe." In: *Agricultural and Forest Meteorology* 217, pp. 10–21.

Bastos, A. et al. (2020). "Sources of Uncertainty in Regional and Global Terrestrial CO2 Exchange Estimates." In: *Global Biogeochemical Cycles* 34.2, e2019GB006393.

Bates, Douglas et al. (2014). "Fitting linear mixed-effects models using lme4." In: *arXiv preprint arXiv:1406.5823*.

Batjes, Niels H (2005). "ISRIC-WISE global data set of derived soil properties on a 0.5 by 0.5 degree grid (ver. 3.0)." In: p. 24.

Bayarri, Maria J et al. (2007). "A Framework for Validation of Computer Models." In: *Technometrics* 49.2, pp. 138–154.

Bell, Andrew et al. (2019). "Fixed and random effects models: making an informed choice." In: *Quality & Quantity* 53.2, pp. 1051–1074.

Bell, David M. and Daniel R. Schlaepfer (2016). "On the dangers of model complexity without ecological justification in species distribution modeling." In: *Ecological Modelling* 330, pp. 50–59.

Benito Garzón, Marta et al. (2011). "Intra-specific variability and plasticity influence potential tree species distributions under climate change." In: *Global Ecology and Biogeography* 20.5, pp. 766–778.

Berzaghi, Fabio et al. (2020). "Towards a New Generation of Trait-Flexible Vegetation Models." In: *Trends in Ecology & Evolution* 35.3, pp. 191–205.

Bestion, Elvire et al. (2015). "Dispersal response to climate change: scaling down to intraspecific variation." In: *Ecology Letters* 18.11, pp. 1226–1233.

Beven, K. (2005). "On the concept of model structural error." In: *Water Science and Technology* 52.6, pp. 167–175.

Beven, Keith and Andrew Binley (1992). "The future of distributed models: Model calibration and uncertainty prediction." In: *Hydrological Processes* 6.3, pp. 279–298.

Beyer, Robert et al. (2019). *A systematic comparison of bias correction methods for paleoclimate simulations*. preprint. Climate Modelling/Modelling only/Pleistocene.

Bijlsma, R. and V. Loeschcke (2005). "Environmental stress, adaptation and evolution: an overview." In: *Journal of Evolutionary Biology* 18.4, pp. 744–749.

Bindoff, Nathaniel L et al. (2013). "Detection and Attribution of Climate Change: from Global to Regional." In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 867–952.

Bles, Anne Marthe van der et al. (2020). "The effects of communicating uncertainty on public trust in facts and numbers." In: *Proceedings of the National Academy of Sciences* 117.14, pp. 7672–7683.

Bocedi, Greta et al. (2013). "Effects of local adaptation and interspecific competition on species' responses to climate change." In: *Annals of the New York Academy of Sciences* 1297.1, pp. 83–97.

Boisgontier, Matthieu P. and Boris Cheval (2016). "The anova to mixed model transition." In: *Neuroscience & Biobehavioral Reviews* 68, pp. 1004–1005.

Bolker, Benjamin M (2015). "Linear and generalized linear mixed models." In: *Ecological Statistics: Contemporary theory and application*, pp. 309–333.

Bolker, Benjamin M. et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution." In: *Trends in Ecology & Evolution* 24.3, pp. 127–135.

Bolnick, Daniel I. et al. (2011). "Why intraspecific trait variation matters in community ecology." In: *Trends in Ecology & Evolution* 26.4, pp. 183–192.

Bolte, A. et al. (2007). "The north-eastern distribution range of European beech a review." In: *Forestry* 80.4, pp. 413–429.

Bonan, Gordon B. (2008). "Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests." In: *Science* 320.5882, pp. 1444–1449.

Bose, Arun K et al. (2020). "Memory of environmental conditions across generations affects the acclimation potential of scots pine." In: *Plant, Cell & Environment* 43.5, pp. 1288–1299.

Box, George E. P. (1976). "Science and Statistics." In: *Journal of the American Statistical Association* 71.356, pp. 791–799.

Brazee, Richard J. and Gregory S. Amacher (2000). "Duality and Faustmann: Implications for the Evaluation of Landowner Behavior." In: *Forest Science* 46.1, pp. 132–138.

Brilli, Lorenzo et al. (2017). "Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes." In: *Science of The Total Environment* 598, pp. 445–470.

Briscoe, Natalie J. et al. (2019). "Forecasting species range dynamics with process-explicit models: matching methods to applications." In: *Ecology Letters* 22.11, pp. 1940–1956.

Brockerhoff, Eckehard G. et al. (2017). "Forest biodiversity, ecosystem functioning and the provision of ecosystem services." In: *Biodiversity and Conservation* 26.13, pp. 3005–3035.

Brooks E., Mollie et al. (2017). "glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling." In: *The R Journal* 9.2, p. 378.

Brynjarsdóttir, Jenný and Anthony O'Hagan (2014). "Learning about physical parameters: the importance of model discrepancy." In: *Inverse Problems* 30.11, p. 114007.

Brysbaert, Marc and Michaël Stevens (2018). "Power Analysis and Effect Size in Mixed Effects Models: A Tutorial." In: *Journal of Cognition* 1.1.

Budescu, David V. et al. (2009). "Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change." In: *Psychological Science* 20.3, pp. 299–308.

Bugmann, Harald (2001). "A Review of Forest Gap Models." In: *Climatic Change* 51.3, pp. 259–305.

Bugmann, Harald et al. (2019). "Tree mortality submodels drive simulated long-term forest dynamics: assessing 15 models from the stand to global scale." In: *Ecosphere* 10.2, e02616.

Buras, Allan and Annette Menzel (2019). "Projecting Tree Species Composition Changes of European Forests for 2061–2090 Under RCP 4.5 and RCP 8.5 Scenarios." In: *Frontiers in Plant Science* 9.

Buras, Allan et al. (2018). "Are Scots pine forest edges particularly prone to drought-induced mortality?" In: *Environmental Research Letters* 13.2, p. 025001.

Burgman, Mark (2005). *Risks and Decisions for Conservation and Environmental Management*. Cambridge University Press. 516 pp.

Burkett, Virginia R. et al. (2005). "Nonlinear dynamics in ecosystem response to climatic change: Case studies and policy implications." In: *Ecological Complexity* 2.4, pp. 357–394.

Cailleret, Maxime et al. (2020). "Bayesian calibration of a growth-dependent tree mortality model to simulate the dynamics of European temperate forests." In: *Ecological Applications* 30.1, e02021.

Callaway, Ragan M. (2007). *Positive Interactions and Interdependence in Plant Communities*. Springer Netherlands.

Cariboni, J. et al. (2007). "The role of sensitivity analysis in ecological modelling." In: *Ecological Modelling*. Special Issue on Ecological Informatics: Biologically-Inspired Machine Learning 203.1, pp. 167–182.

Carsjens, Caroline et al. (2014). "Intra-specific variations in expression of stress-related genes in beech progenies are stronger than drought-induced responses." In: *Tree Physiology* 34.12, pp. 1348–1361.

Castiglioni, Simone et al. (2010). "Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach." In: *Advances in Water Resources*. Special Issue on Novel Insights in Hydrological Modelling 33.10, pp. 1235–1242.

Caswell, Hal (2019). "Introduction: Sensitivity Analysis – What and Why?" In: *Sensitivity Analysis: Matrix Methods in Demography and Ecology*. Ed. by Hal Caswell. Demographic Research Monographs. Cham: Springer International Publishing, pp. 3–12.

Caudullo, Giovanni et al. (2016). "Picea abies in Europe: distribution, habitat, usage and threats." In: *Caudullo, Giovanni; Tinner, Willy; de Rigo, Daniele (2016). Picea abies in Europe: distribution, habitat, usage and threats. In: San-Miguel-Ayanz, J.; de Rigo, D.; Caudullo, G.; Houston Durrant, T.; Mauri, A. (Hg.) European Atlas of Forest Tree Species (S. 114-116). Luxembourg: Publication Office of the European Union*. Ed. by J. San-Miguel-Ayanz et al. In collab. with Giovanni Caudullo et al. Luxembourg: Publication Office of the European Union, pp. 114–116.

Charru, M. et al. (2010). "Recent changes in forest productivity: An analysis of national forest inventory data for common beech (Fagus sylvatica L.) in north-eastern France." In: *Forest Ecology and Management* 260.5, pp. 864–874.

Chauliac, Christian et al. (2011). "NURESIM – A European simulation platform for nuclear reactor safety: Multi-scale and multi-physics calculations, sensitivity and uncertainty analysis." In: *Nuclear Engineering and Design*. Seventh European Commission conference on Euratom research and training in reactor systems (Fission Safety 2009) 241.9, pp. 3416–3426.

Cheaib, Alissar et al. (2012). "Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty." In: *Ecology Letters* 15.6, pp. 533–544.

Chen, Jie et al. (2015). "Assessing the limits of bias correcting climate model outputs for climate change impact studies." In: *Journal of Geophysical Research: Atmospheres* 120.

Chen, Zhen and David B. Dunson (2003). "Random Effects Selection in Linear Mixed Models." In: *Biometrics* 59.4, pp. 762–769.

Chevin, Luis-Miguel et al. (2013). "Phenotypic plasticity and evolutionary demographic responses to climate change: taking theory out to the field." In: *Functional Ecology* 27.4, pp. 967–979.

Cianciaruso, M. V. et al. (2009). "Including intraspecific variability in functional diversity." In: *Ecology* 90.1, pp. 81–89.

Clar, Christoph et al. (2013). "Barriers and guidelines for public policies on climate change adaptation: A missed opportunity of scientific knowledge-brokerage." In: *Natural Resources Forum* 37.1, pp. 1–18.

Clark, James S. (2005). "Why environmental scientists are becoming Bayesians." In: *Ecology Letters* 8.1, pp. 2–14.

Clark, James S. (2020). *Models for Ecological Data: An Introduction*. Princeton University Press.

Clark, Nicholas J. et al. (2022). "Near-term forecasting of companion animal tick paralysis incidence: An iterative ensemble model." In: *PLOS Computational Biology* 18.2, e1009874.

Cleveland, Cory C. et al. (1999). "Global patterns of terrestrial biological nitrogen (N2) fixation in natural ecosystems." In: *Global Biogeochemical Cycles* 13.2, pp. 623–645.

Collins, M. (2002). "Climate predictability on interannual to decadal time scales: the initial value problem." In: *Climate Dynamics* 19.8, pp. 671–692.

Collins, Matthew and Myles R. Allen (2002). "Assessing the Relative Roles of Initial and Boundary Conditions in Interannual to Decadal Climate Predictability." In: *Journal of Climate* 15.21, pp. 3104–3109.

Conti, Stefano and Anthony O'Hagan (2010). "Bayesian emulation of complex multi-output and dynamic computer models." In: *Journal of Statistical Planning and Inference* 140.3, pp. 640–651.

Courbaud, B. et al. (2015). "Applying ecological model evaludation: Lessons learned with the forest dynamics model Samsara2." In: *Ecological Modelling* 314, pp. 1–14.

Cramer, Wolfgang et al. (2001). "Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models." In: *Global Change Biology* 7.4, pp. 357–373.

Cressie, Noel et al. (2009). "Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling." In: *Ecological Applications* 19.3, pp. 553–570.

Csilléry, Katalin et al. (2010). "Approximate Bayesian Computation (ABC) in practice." In: *Trends in Ecology & Evolution* 25.7, pp. 410–418.

Cui, Erqian et al. (2019). "Vegetation Functional Properties Determine Uncertainty of Simulated Ecosystem Productivity: A Traceability Analysis in the East Asian Monsoon Region." In: *Global Biogeochemical Cycles* 33.6, pp. 668–689.

DerSimonian, Rebecca and Nan Laird (1986). "Meta-Analysis in Clinical Trials." In: *Controlled Clinical Trials* 7.3, pp. 177–188.

Des Roches, Simone et al. (2018). "The ecological importance of intraspecific variation." In: *Nature Ecology & Evolution* 2.1, pp. 57–64.

Devenish, B. J. et al. (2012). "Sensitivity analysis of dispersion modeling of volcanic ash from Eyjafjallajökull in May 2010." In: *Journal of Geophysical Research: Atmospheres* 117 (D20).

Di Sacco, Alice et al. (2021). "Ten golden rules for reforestation to optimize carbon sequestration, biodiversity recovery and livelihood benefits." In: *Global Change Biology* 27.7, pp. 1328–1348.

Dietze, Michael C. (2017a). *Ecological Forecasting*. Princeton University Press. 284 pp.

Dietze, Michael C. (2017b). "Prediction in ecology: a first-principles framework." In: *Ecological Applications* 27.7, pp. 2048–2060.

Dietze, Michael C. et al. (2008). "Capturing diversity and interspecific variability in allometries: A hierarchical approach." In: *Forest Ecology and Management* 256.11, pp. 1939–1948.

Dietze, Michael C. et al. (2013). "On improving the communication between models and data." In: *Plant, Cell & Environment* 36.9, pp. 1575–1585.

Dietze, Michael C. et al. (2018). "Iterative near-term ecological forecasting: Needs, opportunities, and challenges." In: *Proceedings of the National Academy of Sciences* 115.7, pp. 1424–1432.

Dixon, Philip (2016). "Should blocks be fixed or random?" In: *2016 Conference on Applied Statistics in Agriculture Proceedings*, pp. 23–39.

Dormann, Carsten F. et al. (2012). "Correlation and process in species distribution models: bridging a dichotomy." In: *Journal of Biogeography* 39.12, pp. 2119–2131.

Drake, John M. et al. (2015). "Ebola Cases and Health System Demand in Liberia." In: *PLOS Biology* 13.1, e1002056.

Dufresne, J.-L. et al. (2013). "Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5." In: *Climate Dynamics* 40.9, pp. 2123–2165.

Durrant, Tracy et al. (2016a). "Fagus sylvatica in Europe: distribution, habitat, usage and threats." In.

Durrant, Tracy et al. (2016b). "Pinus sylvestris in Europe: distribution, habitat, usage and threats." In.

Ehret, U. et al. (2012). "Should we apply bias correction to global and regional climate model data?" In: *Hydrology and Earth System Sciences Discussions* 9.4, pp. 5355–5387.

Esperon-Rodriguez, Manuel et al. (2020). "Functional adaptations and trait plasticity of urban trees along a climatic gradient." In: *Urban Forestry & Urban Greening* 54, p. 126771.

Evans, Matthew R. et al. (2012). "Predictive ecology: systems approaches." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1586, pp. 163–169.

Evans, Matthew R. et al. (2013). "Do simple models lead to generality in ecology?" In: *Trends in Ecology & Evolution* 28.10, pp. 578–583.

Farley, Scott S et al. (2018). "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions." In: *BioScience* 68.8, pp. 563–576.

Fasiolo, Matteo et al. (2020). "qgam: Bayesian non-parametric quantile regression modelling in R." In: *arXiv preprint arXiv:2007.03303*.

Fatichi, Simone et al. (2016). "An overview of current applications, challenges, and future trends in distributed process-based models in hydrology." In: *Journal of Hydrology* 537, pp. 45–60.

Fer, Istem et al. (2018). "Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation." In: *Biogeosciences* 15.19, pp. 5801–5830.

Fer, Istem et al. (2021). *Capturing site-to-site variability through Hierarchical Bayesian calibration of a process-based dynamic vegetation model*. preprint. Ecology.

Fick, Stephen E. and Robert J. Hijmans (2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas." In: *International Journal of Climatology* 37.12, pp. 4302–4315.

Fisher, Rosie A. et al. (2018). "Vegetation demographics in Earth System Models: A review of progress and priorities." In: *Global Change Biology* 24.1, pp. 35–54.

Fleischer, Katrin et al. (2019). "Amazon forest response to CO2 fertilization dependent on plant phosphorus acquisition." In: *Nature Geoscience* 12.9, pp. 736–741.

Fontes, Luis et al. (2011). "Models for supporting forest management in a changing environment." In: *Forest Systems* 3.4, p. 8.

Forrest, Matthew et al. (2020). "Including vegetation dynamics in an atmospheric chemistry-enabled general circulation model: linking LPJ-GUESS (v4.0) with the EMAC modelling system (v2.53)." In: *Geoscientific Model Development* 13.3, pp. 1285–1309.

Forrester, David I. et al. (2017). "Generalized biomass and leaf area allometric equations for European tree species incorporating stand structure, tree age and climate." In: *Forest Ecology and Management* 396, pp. 160–175.

Forrester, David I. et al. (2021). "Calibration of the process-based model 3-PG for major central European tree species." In: *European Journal of Forest Research* 140.4, pp. 847–868.

Fragoso, Tiago M. et al. (2018). "Bayesian Model Averaging: A Systematic Review and Conceptual Classification." In: *International Statistical Review* 86.1, pp. 1–28.

Galbraith, David et al. (2010). "Multiple mechanisms of Amazonian forest biomass losses in three dynamic global vegetation models under climate change." In: *New Phytologist* 187.3, pp. 647–665.

Gelman, A and J Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. 1–651.

Gelman, Andrew (2005). "Analysis of variance—why it is more important than ever." In: *The Annals of Statistics* 33.1, pp. 1–53.

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences." In: *Statistical Science* 7.4, pp. 457–472.

Gelman, Andrew et al. (2020). "Bayesian Workflow." In.

George, Jan-Peter et al. (2017). "Genetic variation, phenotypic stability, and repeatability of drought response in European larch throughout 50 years in a common garden experiment." In: *Tree Physiology* 37.1, pp. 33–46.

Gerten, Dieter et al. (2004). "Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model." In: *Journal of Hydrology* 286.1, pp. 249–270.

Giesselmann, Marco and Alexander W. Schmidt-Catran (2020). "Interactions in Fixed Effects Regression Models." In: *Sociological Methods & Research*, p. 0049124120914934.

Giglio, L. et al. (2010). "Assessing variability and long-term trends in burned area by merging multiple satellite fire products." In: *Biogeosciences* 7.3, pp. 1171–1186.

Goldstein, Michael and Jonathan Rougier (2009). "Reified Bayesian modelling and inference for physical systems." In: *Journal of Statistical Planning and Inference* 139.3, pp. 1221–1239.

Gomes, Dylan G. E. (2021). "Including random effects in statistical models in ecology: fewer than five levels?" In: *bioRxiv*, p. 2021.04.11.439357.

Gosling, John Paul (2018). "SHELF: The Sheffield Elicitation Framework." In: *Elicitation: The Science and Art of Structuring Judgement*. Ed. by Luis C. Dias et al. International Series in Operations Research & Management Science. Cham: Springer International Publishing, pp. 61–93.

Gravel, Dominique et al. (2011). "Species coexistence in a variable world." In: *Ecology Letters* 14.8, pp. 828–839.

Green, Peter J. et al. (2015). "Bayesian computation: a summary of the current state, and samples backwards and forwards." In: *Statistics and Computing* 25.4, pp. 835–862.

Green, Peter and Catriona J. MacLeod (2016). "SIMR: an R package for power analysis of generalized linear mixed models by simulation." In: *Methods in Ecology and Evolution* 7.4, pp. 493–498.

Grimm, V. (2005). "Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology." In: *Science* 310.5750, pp. 987–991.

Gschwantner, Thomas et al. (2016). "Comparison of methods used in European National Forest Inventories for the estimation of volume increment: towards harmonisation." In: *Annals of Forest Science* 73.4, pp. 807–821.

Gunasekara, Fiona Imlach et al. (2014). "Fixed effects analysis of repeated measures data." In: *International Journal of Epidemiology* 43.1, pp. 264–269.

Gustafson, Eric J. (2013). "When relationships estimated in the past cannot be used to predict the future: using mechanistic models to predict landscape ecological dynamics in a changing world." In: *Landscape Ecology* 28.8, pp. 1429–1437.

Gustafson, Eric J. et al. (2018). "Can Future CO_2 Concentrations Mitigate the Negative Effects of High Temperature and Longer Droughts on Forest Growth?" In: *Forests* 9.11, p. 664.

Hajek, Peter et al. (2016). "Intraspecific Variation in Wood Anatomical, Hydraulic, and Foliar Traits in Ten European Beech Provenances Differing in Growth Yield." In: *Frontiers in Plant Science* 7, p. 791.

Hamby, D. M. (1994). "A review of techniques for parameter sensitivity analysis of environmental models." In: *Environmental Monitoring and Assessment* 32.2, pp. 135–154.

Hanewinkel, Marc et al. (2013). "Climate change may cause severe loss in the economic value of European forest land." In: *Nature Climate Change* 3.3, pp. 203–207.

Hardiman, Brady S. et al. (2011). "The role of canopy structural complexity in wood net primary production of a maturing northern deciduous forest." In: *Ecology* 92.9, pp. 1818–1827.

Harrison, Kenneth W. et al. (2012). "Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques." In: *Water Resources Research* 48.11.

Harrison, Xavier A. (2014). "Using observation-level random effects to model overdispersion in count data in ecology and evolution." In: *PeerJ* 2, e616.

Harrison, Xavier A. (2015). "A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution." In: *PeerJ* 3, e1114.

Harrison, Xavier A. et al. (2017). "Best practice in mixed effects modelling and multi-model inference in ecology." In.

Harrison, Xavier A. et al. (2018). "A brief introduction to mixed effects modelling and multi-model inference in ecology." In: *PeerJ* 6, e4794.

Hartig, Florian (2019). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*.

Hartig, Florian et al. (2011). "Statistical inference for stochastic simulation models – theory and application." In: *Ecology Letters* 14.8, pp. 816–827.

Hartig, Florian et al. (2012). "Connecting dynamic vegetation models to data – an inverse perspective." In: *Journal of Biogeography* 39.12, pp. 2240–2252.

Hartig, Florian et al. (2019). "BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics." Version 0.1.7. In.

Hausfather, Zeke and Glen P. Peters (2020). "Emissions – the 'business as usual' story is misleading." In: *Nature* 577.7792, pp. 618–620.

He, Dong et al. (2021). "The importance of intraspecific trait variability in promoting functional niche dimensionality." In: *Ecography* 44.3, pp. 380–390.

He, Yujie et al. (2014). "Uncertainty in the fate of soil organic carbon: A comparison of three conceptually different decomposition models at a larch plantation." In: *Journal of Geophysical Research: Biogeosciences* 119.9, pp. 1892–1905.

Hedges, Larry V. and Jack L. Vevea (1998). "Fixed- and random-effects models in meta-analysis." In: *Psychological Methods* 3.4, pp. 486–504.

Henn, Jonathan J. et al. (2018). "Intraspecific Trait Variation and Phenotypic Plasticity Mediate Alpine Plant Species Response to Climate Change." In: *Frontiers in Plant Science* 9, p. 1548.

Hentschel, Rainer et al. (2016). "Stomatal conductance and intrinsic water use efficiency in the drought year 2003: a case study of European beech." In: *Trees* 30.1, pp. 153–174.

Hickler, Thomas et al. (2004). "USING A GENERALIZED VEGETATION MODEL TO SIMULATE VEGETATION DYNAMICS IN NORTHEASTERN USA." In: *Ecology* 85.2, pp. 519–530.

Hickler, Thomas et al. (2008). "CO2 fertilization in temperate FACE experiments not representative of boreal and tropical forests." In: *Global Change Biology* 14.7, pp. 1531–1542.

Higdon, Dave et al. (2004). "Combining Field Data and Computer Simulations for Calibration and Prediction." In: *SIAM Journal on Scientific Computing* 26.2, pp. 448–466.

Higdon, Dave et al. (2008). "Computer Model Calibration Using High-Dimensional Output." In: *Journal of the American Statistical Association* 103.482, pp. 570–583.

Higgins, Julian P. T. et al. (2009). "A re-evaluation of random-effects meta-analysis." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1, pp. 137–159.

Hill, Mary C and Claire R Tiedeman (2007). "EFFECTIVE GROUNDWATER MODEL CALIBRATION: With Analysis of Data, Sensitivities," in: p. 44.

Howden, S. Mark et al. (2007). "Adapting agriculture to climate change." In: *Proceedings of the National Academy of Sciences* 104.50, pp. 19691–19696.

Hox, Joop J et al. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huang, Shongming et al. (1992). "Comparison of nonlinear height–diameter functions for major Alberta tree species." In: *Canadian Journal of Forest Research*.

Huber, Nica et al. (2018). "Global sensitivity analysis of a dynamic vegetation model: Model sensitivity depends on successional time, climate and competitive interactions." In: *Ecological Modelling* 368, pp. 377–390.

Huntzinger, D. N. et al. (2017). "Uncertainty in the response of terrestrial carbon sink to environmental drivers undermines carbon-climate feedback predictions." In: *Scientific Reports* 7.1, p. 4765.

Hutcheon, Jennifer A. et al. (2010). "Random measurement error and regression dilution bias." In: *BMJ* 340, p. c2289.

IPCC, 2014 (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 [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp." In.

Jiang, Yueyang et al. (2012). "Uncertainty analysis of vegetation distribution in the northern high latitudes during the 21st century with a dynamic vegetation model." In: *Ecology and Evolution* 2.3, pp. 593–614.

Johansson, P. O. (1986). "The economics of forestry and natural resources." In.

Johnsen, Øystein et al. (2005). "Climatic adaptation in Picea abies progenies is affected by the temperature during zygotic embryogenesis and seed maturation." In: *Plant, Cell & Environment* 28.9, pp. 1090–1102.

Johnson, Paul C. D. et al. (2015). "Power analysis for generalized linear mixed models in ecology and evolution." In: *Methods in Ecology and Evolution* 6.2, pp. 133–142.

Jung, Vincent et al. (2010). "Intraspecific variability and trait-based community assembly." In: *Journal of Ecology* 98.5, pp. 1134–1140.

Jørgensen, Sven Erik and G. Bendoricchio (2001). *Fundamentals of Ecological Modelling*. Elsevier. 544 pp.

Kadane, Joseph B (2020). *Principles of uncertainty*. Chapman and Hall/CRC.

Kapeller, Stefan et al. (2012). "Intraspecific variation in climate response of Norway spruce in the eastern Alpine range: Selecting appropriate provenances for future climate." In: *Forest Ecology and Management* 271, pp. 46–57.

Karatzoglou, Alexandros et al. (2004). "kernlab - An S4 Package for Kernel Methods in R." In: *Journal of Statistical Software* 11.1, pp. 1–20.

Kattge, Jens and Wolfgang Knorr (2007). "Temperature acclimation in a biochemical model of photosynthesis: a reanalysis of data from 36 species." In: *Plant, Cell & Environment* 30.9, pp. 1176–1190.

Kearney, Michael R. et al. (2010). "Correlative and mechanistic models of species distribution provide congruent forecasts under climate change." In: *Conservation Letters* 3.3, pp. 203–213.

Keenan, Trevor F. et al. (2011a). "The model–data fusion pitfall: assuming certainty in an uncertain world." In: *Oecologia* 167.3, pp. 587–597.

Keenan, Trevor et al. (2011b). "Predicting the future of forests in the Mediterranean under climate change, with niche- and process-based models: CO2 matters!" In: *Global Change Biology* 17.1, pp. 565–579.

Kennedy, Marc C. and Anthony O'Hagan (2001). "Bayesian calibration of computer models." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 425–464.

Knight, Frank H. (1921). *Risk, Uncertainty and Profit*. SSRN Scholarly Paper ID 1496192. Rochester, NY: Social Science Research Network.

Krause, A. et al. (2019). "Multimodel Analysis of Future Land Use and Climate Change Impacts on Ecosystem Functioning." In: *Earth's Future* 7.7, pp. 833–851.

Kremer, James N. (1983). "Ecological implications of parameter uncertainty in stochastic simulation." In: *Ecological Modelling* 18.3, pp. 187–207.

Krueger, Charlene and Lili Tian (2004). "A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points." In: *Biological Research For Nursing* 6.2, pp. 151–157.

Kuznetsova, Alexandra et al. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." In: *Journal of Statistical Software, Articles* 82.13.

Kvaalen, Harald and Øystein Johnsen (2008). "Timing of bud set in Picea abies is regulated by a memory of temperature during zygotic and somatic embryogenesis." In: *New Phytologist* 177.1, pp. 49–59.

Körner, Christian (2006). "Plant CO2 responses: an issue of definition, time and resource supply." In: *The New Phytologist* 172.3, pp. 393–411.

Laberge, Yves (2013). "Simulating nature: a philosophical study of computer-simulation uncertainties and their role in climate science and policy advice." In: *Journal of Applied Statistics* 40.4, pp. 919–920.

Laforest-Lapointe, Isabelle et al. (2014). "Intraspecific variability in functional traits matters: case study of Scots pine." In: *Oecologia* 175.4, pp. 1337–1348.

Laird, Nan M. and James H. Ware (1982). "Random-Effects Models for Longitudinal Data." In: *Biometrics* 38.4, pp. 963–974.

Lamarque, J.-F. et al. (2013). "Multi-model mean nitrogen and sulfur deposition from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): evaluation of historical and projected future changes." In: *Atmospheric Chemistry and Physics* 13.16, pp. 7997–8018.

Landau, L. D. and E. M. Lifshitz (2013). *Statistical Physics: Volume 5*. Elsevier. 563 pp.

Landsberg, J. J. and R. H. Waring (1997). "A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning." In: *Forest Ecology and Management* 95.3, pp. 209–228.

Landsberg, Joe and Peter Sands (2011). "Chapter 9 - The 3-PG Process-Based Model." In: *Terrestrial Ecology*. Vol. 4. Physiological Ecology of Forest Production. Elsevier, pp. 241–282.

Larcher, Walter (1983). "Ökophysiologische Konstitutionseigenschaften von Gebirgspflanzen." In: *Berichte der Deutschen Botanischen Gesellschaft* 96.1, pp. 73–85.

Latif, M. (2011). "Uncertainty in climate change projections." In: *Journal of Geochemical Exploration*. Sustainability of Geochemical Cycling 110.1, pp. 1–7.

Laubmeier, Amanda N. et al. (2020). "Ecological Dynamics: Integrating Empirical, Statistical, and Analytical Methods." In: *Trends in Ecology & Evolution* 35.12, pp. 1090–1099.

Lawrence, Judy et al. (2020). "Climate change: making decisions in the face of deep uncertainty." In: *Nature* 580.7804, pp. 456–456.

Lawrence, Mark et al. (2010). "Comparisons of National Forest Inventories." In: *National Forest Inventories: Pathways for Common Reporting*. Ed. by Erkki Tomppo et al. Dordrecht: Springer Netherlands, pp. 19–32.

Lenth, Russell V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*.

Levanič, Tom et al. (2008). "The climate sensitivity of Norway spruce [Picea abies (L.) Karst.] in the southeastern European Alps." In: *Trees* 23.1, p. 169.

Levin, Simon A. (1992). "The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture." In: *Ecology* 73.6, pp. 1943–1967.

Levin, Simon A. (1998). "Ecosystems and the Biosphere as Complex Adaptive Systems." In: *Ecosystems* 1.5, pp. 431–436.

Lindeskog, M. et al. (2013). "Implications of accounting for land use in simulations of ecosystem carbon cycling in Africa." In: *Earth System Dynamics* 4.2, pp. 385–407.

Lindeskog, Mats et al. (2021). "Accounting for forest management in the estimation of forest carbon balance using the dynamic vegetation model LPJ-GUESS (v4.0, r9333): Implementation and evaluation of simulations for Europe." In: *Geoscientific Model Development Discussions*, pp. 1–42.

Lindstrom, Mary J and Douglas M Bates (1988). "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data." In: *Journal of the American Statistical Association* 83.404, pp. 1014–1022.

Littell, Ramon C. (2002). "Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS." In: *Journal of Agricultural, Biological, and Environmental Statistics* 7.4, pp. 472–490.

Liu, Peiran R. and Adrian E. Raftery (2021). "Country-based rate of emissions reductions should increase by 80% beyond nationally determined contributions to meet the 2 °C target." In: *Communications Earth & Environment* 2.1, pp. 1–10.

Loehle, Craig (2000). "Strategy Space and the Disturbance Spectrum: A Life-History Model for Tree Species Coexistence." In: *The American Naturalist* 156.1, pp. 14–33.

Luke, Adam et al. (2017). "Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States." In: *Water Resources Research* 53.7, pp. 5469–5494.

Luyssaert, Sebastiaan et al. (2008). "Old-growth forests as global carbon sinks." In: *Nature* 455.7210, pp. 213–215.

Maas, Cora J. M. and Joop J. Hox (2005). "Sufficient Sample Sizes for Multilevel Modeling." In: *Methodology* 1.3, pp. 86–92.

MacBean, Natasha et al. (2016). "Consistent assimilation of multiple data streams in a carbon cycle dataassimilation system." In: *Geoscientific Model Development* 9.10, pp. 3569–3588.

Makela, A. et al. (2000). "Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation." In: *Tree Physiology* 20.5, pp. 289–298.

Marsili-Libelli, Stefano et al. (2014). "Practical identifiability analysis of environmental models." In: *Proceedings - 7th International Congress on Environmental Modelling and Software: Bold Visions for Environmental Modeling, iEMSs 2014*. Conference Organising Committee.

Martin, Julien G. A. et al. (2011). "Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models." In: *Methods in Ecology and Evolution* 2.4, pp. 362–374.

Matott, L. Shawn et al. (2009). "Evaluating uncertainty in integrated environmental models: A review of concepts and tools." In: *Water Resources Research* 45.6.

Matuschek, Hannes et al. (2017). "Balancing Type I error and power in linear mixed models." In: *Journal of Memory and Language* 94, pp. 305–315.

Maxim, Laura and Jeroen P. van der Sluijs (2011). "Quality in environmental science for policy: Assessing uncertainty as a component of policy analysis." In: *Environmental Science & Policy* 14.4, pp. 482–492.

McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 1453000/9bac68d1409cff87869308d509d855a2.

McLean, Robert A. et al. (1991). "A Unified Approach to Mixed Linear Models." In: *The American Statistician* 45.1, pp. 54–64.

McMahon, Sean M. and Jeffrey M. Diez (2007). "Scales of association: hierarchical linear models and the measurement of ecological systems." In: *Ecology Letters* 10.6, pp. 437–452.

McNeish, Daniel M. and Laura M. Stapleton (2016). "The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration." In: *Educational Psychology Review* 28.2, pp. 295–314.

McNeish, Daniel (2017). "Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction." In: *Multivariate Behavioral Research* 52.5, pp. 661–670.

Medlyn, Belinda E. et al. (2015). "Using ecosystem experiments to improve vegetation models." In: *Nature Climate Change* 5.6, pp. 528–534.

Meinshausen, Malte et al. (2011). "The RCP greenhouse gas concentrations and their extensions from 1765 to 2300." In: *Climatic Change* 109.1, p. 213.

Meteyard, Lotte and Robert A. I. Davies (2020). "Best practice guidance for linear mixed-effects models in psychological science." In: *Journal of Memory and Language* 112, p. 104092.

Metzger, M. J. et al. (2005). "A climatic stratification of the environment of Europe." In: *Global Ecology and Biogeography* 14.6, pp. 549–563.

Millar, Russell B. and Marti J. Anderson (2004). "Remedies for pseudoreplication." In: *Fisheries Research* 70.2, pp. 397–407.

Minunno, F. et al. (2013). "Selecting Parameters for Bayesian Calibration of a Process-Based Model: A Methodology Based on Canonical Correlation Analysis." In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1, pp. 370–385.

Minunno, F. et al. (2016). "Calibration and validation of a semi-empirical flux ecosystem model for coniferous forests in the Boreal region." In: *Ecological Modelling* 341, pp. 37–52.

Moran, Emily V. et al. (2016). "Intraspecific trait variation across scales: implications for understanding global change responses." In: *Global Change Biology* 22.1, pp. 137–150.

Moreno, Adam and Hubert Hasenauer (2016). "Spatial downscaling of European climate data." In: *International Journal of Climatology* 36.3, pp. 1444–1458.

Morin, Xavier et al. (2021). "Beyond forest succession: A gap model to study ecosystem functioning and tree community composition under climate change." In: *Functional Ecology* 35.4, pp. 955–975.

Morris, Max D. (1991). "Factorial Sampling Plans for Preliminary Computational Experiments." In: *Technometrics* 33.2, pp. 161–174.

Mäkelä, Jarmo et al. (2020). "Sensitivity of 21st century simulated ecosystem indicators to model parameters, prescribed climate drivers, RCP scenarios and forest management actions for two Finnish boreal forest sites." In: *Biogeosciences* 17.10, pp. 2681–2700.

Münzbergová, Zuzana et al. (2017). "Genetic differentiation and plasticity interact along temperature and precipitation gradients to determine plant performance under climate change." In: *Journal of Ecology* 105.5, pp. 1358–1373.

Nakagawa, Shinichi and Holger Schielzeth (2013). "A general and simple method for obtaining R2 from generalized linear mixed-effects models." In: *Methods in Ecology and Evolution* 4.2, pp. 133–142.

Needham, Jessica et al. (2018). "Inferring forest fate from demographic data: from vital rates to population dynamic models." In: *Proceedings of the Royal Society B: Biological Sciences* 285.1874, p. 20172050.

Novick, Kimberly A. et al. (2016). "The increasing importance of atmospheric demand for ecosystem water and carbon fluxes." In: *Nature Climate Change* 6.11, pp. 1023–1027.

Nugent, Joshua R. and Ken P. Kleinman (2021). "Type I error control for cluster randomized trials under varying small sample structures." In: *BMC medical research methodology* 21.1, p. 65.

Oberpriller, Johannes et al. (2021a). *Fixed or random? On the reliability of mixed-effects models for a small number of levels in grouping variables*. bioRxiv, p. 2021.05.03.442487.

Oberpriller, Johannes et al. (2021b). "Towards robust statistical inference for complex computer models." In: *Ecology Letters* 24.6, pp. 1251–1261.

Olin, S. et al. (2015). "Modelling the response of yields and tissue C : N to changes in atmospheric CO_2 and N management in the main wheat regions of western Europe." In: *Biogeosciences* 12.8, pp. 2489–2515.

Oney, Brian et al. (2013). "Intraspecific variation buffers projected climate change impacts on Pinus contorta." In: *Ecology and Evolution* 3.2, pp. 437–449.

Ouyang, Lei et al. (2022). "Interpreting the water use strategies of plantation tree species by canopy stomatal conductance and its sensitivity to vapor pressure deficit in South China." In: *Forest Ecology and Management* 505, p. 119940.

O'Hagan, Anthony (2012). "Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux." In: *Environmental Modelling & Software*. Thematic issue on Expert Opinion in Environmental Modelling and Management 36, pp. 35–48.

Pappas, Christoforos et al. (2013). "Sensitivity analysis of a process-based ecosystem model: Pinpointing parameterization and structural issues." In: *Journal of Geophysical Research: Biogeosciences* 118.2, pp. 505–528.

Pearce, David W. (2001). "The Economic Value of Forest Ecosystems." In: *Ecosystem Health* 7.4, pp. 284–296.

Peng, Changhui et al. (2011). "Integrating models with data in ecology and palaeoecology: advances towards a model–data fusion approach." In: *Ecology Letters* 14.5, pp. 522–536.

Petchey, Owen L. et al. (2015). "The ecological forecast horizon, and examples of its uses and determinants." In: *Ecology Letters* 18.7, pp. 597–611.

Petter, Gunnar et al. (2020). "How robust are future projections of forest landscape dynamics? Insights from a systematic comparison of four forest landscape models." In: *Environmental Modelling & Software* 134, p. 104844.

Pianosi, Francesca et al. (2016). "Sensitivity analysis of environmental models: A systematic review with practical workflow." In: *Environmental Modelling & Software* 79, pp. 214–232.

Pinheiro, José C. and Douglas M. Bates (1995). "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." In: *Journal of Computational and Graphical Statistics* 4.1, pp. 12–35.

Pol, Martijn van de (2012). "Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models." In: *Methods in Ecology and Evolution* 3.2, pp. 268–280.

Porcar-Castell, Albert and Sari Palmroth (2012). "Modelling photosynthesis in highly dynamic environments: the case of sunflecks." In: *Tree Physiology* 32.9, pp. 1062–1065.

Prestele, Reinhard et al. (2016). "Hotspots of uncertainty in land-use and land-cover change projections: a global-scale model comparison." In: *Global Change Biology* 22.12, pp. 3967–3983.

Pretzsch, H. et al. (2015). "Growth and yield of mixed versus pure stands of Scots pine (Pinus sylvestris L.) and European beech (Fagus sylvatica L.) analysed along a productivity gradient through Europe." In: *European Journal of Forest Research* 134.5, pp. 927–947.

Pugh, Thomas A. M. et al. (2019). "Role of forest regrowth in global carbon sink dynamics." In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4382–4387.

Pullen, Nick and Richard J. Morris (2014). "Bayesian Model Comparison and Parameter Inference in Systems Biology Using Nested Sampling." In: *PLOS ONE* 9.2, e88419.

Rabin, Sam S. et al. (2017). "The Fire Modeling Intercomparison Project (FireMIP), phase 1: experimental and analytical protocols with detailed model descriptions." In: *Geoscientific Model Development* 10.3, pp. 1175–1197.

Rabinovich, Semyon G. (2006). *Measurement Errors and Uncertainties: Theory and Practice*. Springer Science & Business Media. 313 pp.

Radchuk, Viktoriia et al. (2019). "Transferability of Mechanistic Ecological Models Is About Emergence." In: *Trends in Ecology & Evolution* 34.6, pp. 487–488.

Rahn, Eric et al. (2018). "Exploring adaptation strategies of coffee production to climate change using a process-based model." In: *Ecological Modelling* 371, pp. 76–89.

Rangel, Thiago F. et al. (2018). "Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves." In: *Science* 361.6399.

Rastetter, Edward B. (2017). "Modeling for Understanding v. Modeling for Numbers." In: *Ecosystems* 20.2, pp. 215–221.

Reichert, Peter and Johanna Mieleitner (2009). "Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters." In: *Water Resources Research* 45.10.

Reichstein, Markus et al. (2019). "Deep learning and process understanding for data-driven Earth system science." In: *Nature* 566.7743, pp. 195–204.

Reyer, Christopher P. O. et al. (2016). "Integrating parameter uncertainty of a process-based model in assessments of climate change effects on forest productivity." In: *Climatic Change* 137.3, pp. 395–409.

Reyer, Christopher P. O. et al. (2020). "The PROFOUND Database for evaluating vegetation models and simulating climate impacts on European forests." In: *Earth System Science Data* 12.2, pp. 1295–1320.

Reyer, Christopher (2015). "Forest Productivity Under Environmental Change—a Review of Stand-Scale Modeling Studies." In: *Current Forestry Reports* 1.2, pp. 53–68.

Richardson, Andrew D. et al. (2010). "Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints." In: *Oecologia* 164.1, pp. 25–40.

Richardson, Andrew D. et al. (2012). "Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis." In: *Global Change Biology* 18.2, pp. 566–584.

Robichaud, Edgar and Ian R. Methven (1992). "The applicability of the pipe model theory for the prediction of foliage biomass in trees from natural, untreated black spruce stands." In: *Canadian Journal of Forest Research*.

Roux, Sébastien et al. (2021). "Cluster-based GSA: Global sensitivity analysis of models with temporal or spatial outputs using clustering." In: *Environmental Modelling & Software* 140, p. 105046.

Ruiz-Pérez, Guiomar and Giulia Vico (2020). "Effects of Temperature and Water Availability on Northern European Boreal Forests." In: *Frontiers in Forests and Global Change* 3, p. 34.

Sakschewski, Boris et al. (2015). "Leaf and stem economics spectra drive diversity of functional plant traits in a dynamic global vegetation model." In: *Global Change Biology* 21.7, pp. 2711–2725.

Saltelli, Andrea (2002). "Sensitivity Analysis for Importance Assessment." In: *Risk Analysis* 22.3, pp. 579–590.

Saltelli, Andrea et al., eds. (2008). *Global sensitivity analysis: the primer*. Chichester, England ; Hoboken, NJ: John Wiley. 292 pp.

Saltelli, Andrea et al. (2019). "Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices." In: *Environmental Modelling & Software* 114, pp. 29–39.

Sands, P. J. and J. J. Landsberg (2002). "Parameterisation of 3-PG for plantation grown Eucalyptus globulus." In: *Forest Ecology and Management* 163.1, pp. 273–292.

Saraiva, Sofia et al. (2019). "Uncertainties in Projections of the Baltic Sea Ecosystem Driven by an Ensemble of Global Climate Models." In: *Frontiers in Earth Science* 6.

Sargsyan, Khachik et al. (2019). "Embedded Model Error Representation for Bayesian Model Calibration." In: *International Journal for Uncertainty Quantification* 9.4.

Savolainen, Outi et al. (2007). "Gene Flow and Local Adaptation in Trees." In: *Annual Review of Ecology, Evolution, and Systematics* 38.1, pp. 595–619.

Scheffer, Marten (2010). "Foreseeing tipping points." In: *Nature* 467.7314, pp. 411–412.

Schenk, H. Jochen (1996). "Modeling the effects of temperature on growth and persistence of tree species: A critical review of tree population models." In: *Ecological Modelling* 92.1, pp. 1–32.

Schielzeth, Holger (2010). "Simple means to improve the interpretability of regression coefficients." In: *Methods in Ecology and Evolution* 1.2, pp. 103–113.

Schielzeth, Holger and Wolfgang Forstmeier (2009). "Conclusions beyond support: overconfident estimates in mixed models." In: *Behavioral Ecology* 20.2, pp. 416–420.

Schielzeth, Holger et al. (2020). "Robustness of linear mixed-effects models to violations of distributional assumptions." In: *Methods in Ecology and Evolution* 11.9.

Scholze, Marko et al. (2006). "A climate-change risk analysis for world ecosystems." In: *Proceedings of the National Academy of Sciences* 103.35, pp. 13116–13120.

Schomaker, Michael (2012). "Shrinkage averaging estimation." In: *Statistical Papers* 53.4, pp. 1015–1034.

Schurer, Andrew P. et al. (2013). "Separating Forced from Chaotic Climate Variability over the Past Millennium." In: *Journal of Climate* 26.18, pp. 6954–6973.

Seebacher, Frank and Craig E. Franklin (2012). "Determining environmental causes of biological effects: the need for a mechanistic physiological dimension in conservation biology." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1596, pp. 1607–1614.

Shanker, Arun K. et al. (2020). "Epigenetics and transgenerational memory in plants under heat stress." In: *Plant Physiology Reports* 25.4, pp. 583–593.

Shaver, J. Myles (2019). "Interpreting Interactions in Linear Fixed-Effect Regression Models: When Fixed-Effect Estimates Are No Longer Within-Effects." In: *Strategy Science* 4.1, pp. 25–40.

Silk, Matthew J. et al. (2020). "Perils and pitfalls of mixed-effects regression models in biology." In: *PeerJ* 8, e9522.

Simmons, Joseph P. et al. (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." In: *Psychological Science* 22.11, pp. 1359–1366.

Sitch, S. et al. (2003). "Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model." In: *Global Change Biology* 9.2, pp. 161–185.

Smith, B. et al. (2014). "Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model." In: *Biogeosciences* 11.7, pp. 2027–2054.

Smith, Benjamin et al. (2001). "Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space." In: *Global Ecology and Biogeography* 10.6, pp. 621–637.

Smith, Benjamin et al. (2011). "A model of the coupled dynamics of climate, vegetation and terrestrial ecosystem biogeochemistry for regional applications." In: *Tellus A: Dynamic Meteorology and Oceanography* 63.1, pp. 87–106.

Smith, Leonard A. (2001). "Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems." In: *Nonlinear Dynamics and Statistics*. Ed. by Alistair I. Mees. Boston, MA: Birkhäuser Boston, pp. 31–64.

Snell, R. S. et al. (2014). "Using dynamic vegetation models to simulate plant range shifts." In: *Ecography* 37.12, pp. 1184–1197.

Snell, Rebecca S. et al. (2018). "Importance of climate uncertainty for projections of forest ecosystem services." In: *Regional Environmental Change* 18.7, pp. 2145–2159.

Sobie, Eric A. (2009). "Parameter Sensitivity Analysis in Electrophysiological Models Using Multivariable Regression." In: *Biophysical Journal* 96.4, pp. 1264–1274.

Solomon, Susan et al. (2009). "Irreversible climate change due to carbon dioxide emissions." In: *Proceedings of the National Academy of Sciences* 106.6, pp. 1704–1709.

Somerville, Rachel S. and Romeel Davé (2015). "Physical Models of Galaxy Formation in a Cosmological Framework." In: *Annual Review of Astronomy and Astrophysics* 53.1, pp. 51–113.

Speagle, Joshua S. (2020). "A Conceptual Introduction to Markov Chain Monte Carlo Methods." In: *arXiv:1909.12313 [astro-ph, physics:physics, stat]*.

Stram, Daniel O. and Jae Won Lee (1994). "Variance Components Testing in the Longitudinal Mixed Effects Model." In: *Biometrics* 50.4, pp. 1171–1177.

Strengers, Bart J. et al. (2010). "Assessing 20th century climate–vegetation feedbacks of land-use change and natural vegetation dynamics in a fully coupled vegetation–climate model." In: *International Journal of Climatology* 30.13, pp. 2055–2065.

Swallow, William H. and John F. Monahan (1984). "Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components." In: *Technometrics* 26.1, pp. 47–57.

Tebaldi, Claudia and Reto Knutti (2007). "The use of the multi-model ensemble in probabilistic climate projections." In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2053–2075.

Tegel, Willy et al. (2014). "A recent growth increase of European beech (Fagus sylvatica L.) at its Mediterranean distribution limit contradicts drought stress." In: *European Journal of Forest Research* 133.1, pp. 61–71.

Ter Braak, Cajo J. F. and Jasper A. Vrugt (2008). "Differential Evolution Markov Chain with snooker updater and fewer chains." In: *Statistics and Computing* 18.4, pp. 435–446.

Thompson, Patrick L. et al. (2020). "A process-based metacommunity framework linking local and regional scale community ecology." In: *Ecology Letters*.

Thuerey, Nils et al. (2021). "Physics-based Deep Learning." In.

Tian, Shiying et al. (2014). "Global sensitivity analysis of DRAINMOD-FOREST, an integrated forest ecosystem model: GLOBAL SENSITIVITY ANALYSIS OF DRAINMOD-FOREST." In: *Hydrological Processes* 28.15, pp. 4389–4410.

Tian, Wei et al. (2018). "A review of uncertainty analysis in building energy assessment." In: *Renewable and Sustainable Energy Reviews* 93, pp. 285–301.

Tian, Xianglin et al. (2020). "Extending the range of applicability of the semi-empirical ecosystem flux model PRELES for varying forest types and climate." In: *Global Change Biology* 26.5, pp. 2923–2943.

Tiktak, A. and W. Bouten (1992). "Modelling soil water dynamics in a forested ecosystem. III: Model description and evaluation of discretization." In: *Hydrological Processes* 6.4, pp. 455–465.

Tomlin, Alison S. (2013). "The role of sensitivity and uncertainty analysis in combustion modelling." In: *Proceedings of the Combustion Institute* 34.1, pp. 159–176.

Trotsiuk, Volodymyr et al. (2020a). "Assessing the response of forest productivity to climate extremes in Switzerland using model–data fusion." In: *Global Change Biology* 26.4, pp. 2463–2476.

Trotsiuk, Volodymyr et al. (2020b). "r3PG – An r package for simulating forest growth using the 3-PG process-based model." In: *Methods in Ecology and Evolution* 11.11, pp. 1470–1475.

Trucano, T. G. et al. (2006). "Calibration, validation, and sensitivity analysis: What's what." In: *Reliability Engineering & System Safety*. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) 91.10, pp. 1331–1357.

Tuo, Rui (2017). "Adjustments to Computer Models via Projected Kernel Calibration." In: *arXiv:1705.03422 [stat]*.

Tuo, Rui and C. F. Jeff Wu (2016). "A Theoretical Framework for Calibration in Computer Models: Parametrization, Estimation and Convergence Properties." In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1, pp. 767–795.

Tylianakis, Jason M. et al. (2008). "Global change and species interactions in terrestrial ecosystems." In: *Ecology Letters* 11.12, pp. 1351–1363.

Tóth, Brigitta et al. (2017). "3D soil hydraulic database of Europe at 250 m resolution." In: *Hydrological Processes* 31.14, pp. 2662–2666.

Uusitalo, Laura et al. (2015). "An overview of methods to evaluate uncertainty of deterministic models in decision support." In: *Environmental Modelling & Software* 63, pp. 24–31.

Van Bodegom, P. M. et al. (2012). "Going beyond limitations of plant functional types when predicting global ecosystem–atmosphere fluxes: exploring the merits of traits-based approaches." In: *Global Ecology and Biogeography* 21.6, pp. 625–636.

Van Oijen, M. et al. (2011). "A Bayesian framework for model calibration, comparison and analysis: Application to four models for the biogeochemistry of a Norway spruce forest." In: *Agricultural and Forest Meteorology* 151.12, pp. 1609–1621.

Van Oijen, Marcel (2017). "Bayesian Methods for Quantifying and Reducing Uncertainty and Error in Forest Models." In: *Current Forestry Reports* 3.4, pp. 269–280.

Vanderwel, Mark C. et al. (2017). "Predicting the abundance of forest types across the eastern United States through inverse modelling of tree demography." In: *Ecological Applications* 27.7, pp. 2128–2141.

Verheijen, L. M. et al. (2013). "Impacts of trait variation through observed trait–climate relationships on performance of an Earth system model: a conceptual analysis." In: *Biogeosciences* 10.8, pp. 5497–5515.

Vicente-Serrano, Sergio M. et al. (2010). "A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index." In: *Journal of Climate* 23.7, pp. 1696–1718.

Vieilledent, Ghislain et al. (2010). "Individual variability in tree allometry determines light resource allocation in forest ecosystems: a hierarchical Bayesian approach." In: *Oecologia* 163.3, pp. 759–773.

Violle, Cyrille et al. (2007). "Let the concept of trait be functional!" In: *Oikos* 116.5, pp. 882–892.

Violle, Cyrille et al. (2012). "The return of the variance: intraspecific variability in community ecology." In: *Trends in Ecology & Evolution* 27.4, pp. 244–252.

Vitousek, Peter M. and Robert W. Howarth (1991). "Nitrogen limitation on land and in the sea: How can it occur?" In: *Biogeochemistry* 13.2, pp. 87–115.

Wainwright, Patricia E. et al. (2007). "Advantages of mixed effects models over traditional ANOVA models in developmental studies: A worked example in a mouse model of fetal alcohol syndrome." In: *Developmental Psychobiology* 49.7, pp. 664–674.

Walentowski, Helge et al. (2017). "Assessing future suitability of tree species under climate change by multiple methods: a case study in southern Germany." In: *Annals of Forest Research* 60.1, pp. 101–126.

Walker, Stephen G. (2013). "Bayesian inference with misspecified models." In: *Journal of Statistical Planning and Inference* 143.10, pp. 1621–1633.

Wallach, D. and M. Genard (1998). "Effect of uncertainty in input and parameter values on model prediction error." In: *Ecological Modelling* 105.2, pp. 337–345.

Wang, Fugui et al. (2013). "Global sensitivity analysis of a modified CENTURY model for simulating impacts of harvesting fine woody biomass for bioenergy." In: *Ecological Modelling* 259, pp. 16–23.

Waring, R. H. (1983). "Estimating Forest Growth and Efficiency in Relation to Canopy Leaf Area." In: *Advances in Ecological Research*. Ed. by A. MacFadyen and E. D. Ford. Vol. 13. Academic Press, pp. 327–354.

Waring, R. H. et al. (1998). "Net primary production of forests: a constant fraction of gross primary production?" In: *Tree Physiology* 18.2, pp. 129–134.

Warton, David I. et al. (2015). "Model-based thinking for community ecology." In: *Plant Ecology* 216.5, pp. 669–682.

Watson, Ty A. et al. (2013). "Parameter and predictive outcomes of model simplification." In: *Water Resources Research* 49.7, pp. 3952–3977.

Way, Danielle A. and Ram Oren (2010). "Differential responses to changes in growth temperature between trees from different functional groups and biomes: a review and synthesis of data." In: *Tree Physiology* 30.6, pp. 669–688.

Webb, Colleen T. et al. (2010). "A structured and dynamic framework to advance traits-based theory and prediction in ecology." In: *Ecology Letters* 13.3, pp. 267–283.

White, Jeremy T. et al. (2014). "Quantifying the predictive consequences of model error with linear subspace analysis." In: *Water Resources Research* 50.2, pp. 1152–1173.

Wieder, William R et al. (2015). "Effects of model structural uncertainty on carbon cycle projections: biological nitrogen fixation as a case study." In: *Environmental Research Letters* 10.4, p. 044016.

Wiencierz, Andrea et al. (2011). "Restricted likelihood ratio testing in linear mixed models with general error covariance structure." In: *Electronic Journal of Statistics* 5 (none), pp. 1718–1734.

Wood, Simon N. (2001). "Partially Specified Ecological Models." In: *Ecological Monographs* 71.1, pp. 1–25.

Woods, Ellen C. et al. (2012). "Adaptive geographical clines in the growth and defense of a native plant." In: *Ecological Monographs* 82.2, pp. 149–168.

Wramneby, Anna et al. (2008). "Parameter uncertainties in the modelling of vegetation dynamics—Effects on tree community structure and ecosystem functioning in European forest biomes." In: *Ecological Modelling* 216.3, pp. 277–290.

Wu, Jianguo and Harbin Li (2006). "UNCERTAINTY ANALYSIS IN ECOLOGICAL STUDIES:AN OVERVIEW." In: *SCALING AND UNCERTAINTY ANALYSIS IN ECOLOGY*. Ed. by JIANGUO WU et al. Dordrecht: Springer Netherlands, pp. 45–66.

Wu, Zhendong et al. (2017). "Climate data induced uncertainty in model-based estimations of terrestrial primary productivity." In: *Environmental Research Letters* 12.6, p. 064013.

Wu, Zhendong et al. (2018). "Effect of climate dataset selection on simulations of terrestrial GPP: Highest uncertainty for tropical regions." In: *PLOS ONE* 13.6, e0199383.

Wutzler, T. and N. Carvalhais (2014). "Balancing multiple constraints in model-data integration: Weights and the parameter block approach." In: *Journal of Geophysical Research: Biogeosciences* 119.11, pp. 2112–2129.

Xia, Jianyang et al. (2013). "Traceable components of terrestrial carbon storage capacity in biogeochemical models." In: *Global Change Biology* 19.7, pp. 2104–2116.

Yang, Jie et al. (2018). "Why Functional Traits Do Not Predict Tree Demographic Rates." In: *Trends in Ecology & Evolution* 33.5, pp. 326–336.

Yoda, K. (1963). "Self-thinning in overcrowded pure stands under cultivated and natural conditions (Intraspecific competition among higher plants. XI)." In: *J. Inst. Polytech. Osaka City Univ. Ser. D.* 14, pp. 107–129.

Yousefpour, Rasoul and Marc Hanewinkel (2016). "Climate Change and Decision-Making Under Uncertainty." In: *Current Forestry Reports* 2.2, pp. 143–149.

Zaehle, S. et al. (2005). "Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics." In: *Global Biogeochemical Cycles* 19.3.

Zang, Christian et al. (2014). "Patterns of drought tolerance in major European temperate forest trees: climatic drivers and levels of variability." In: *Global Change Biology* 20.12, pp. 3767–3779.

Zeide, Boris (1993). "Analysis of Growth Equations." In: *Forest Science* 39.3, pp. 594–616.

Zuur, Alain et al. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

ACKNOWLEDGEMENTS

I would like to use the opportunity to say thank you to all people who were supporting me throughout the entire time of my PhD.

First and foremost, thank you Florian Hartig for being my supervisor and for retaining a good working atmosphere during a global pandemic. I enjoyed the discussions we had about science.

Thank you Björn Reineking and Rainer Spang for willing to support me during the PhD as mentors.

Thank you Maximilian Pichler and Lukas Heiland for being my office mates at the beginning of my PhD. Our discussion about many aspects of science, ecology and life and your mental support were really encouraging.

Thank you Felix Gottschlich for sharing the office with me for half of a year. Your open and funny manner made the work more enjoyable.

In addition, I want to thank all the other lab members during my time in the group: Lisa Hülsmann, Magdalena Mair, Loic Chalmandrier and Amaël Le Squin. It has been a pleasure to work with all of you.

A big thank you goes to the various open source projects I have used during this thesis and without which science would not be possible: LaTex, R and RStudio, Python and zotero.

Finally, I am so grateful for all the emotional support of my family and my friends that I had throughout the entire time. To my parents: Thank you for everything. Without your constant support none of this would have been possible.

ELECTRONICAL SUPPORTING INFORMATION

The electronical appendix contains for each research paper a supporting information.

Chapter 3:
Supporting Information S1

Chapter 4:
Supporting Information S2

Chapter 5:
Supporting Information S3

Chapter 6:
Supporting Information S4