Lukáš Gábor

2020

THE QUALITY OF SPATIAL DATA AND ITS EFFECT ON SPECIES DISTRIBUTION MODELS

Lukáš Gábor

DISSERTATION THESIS

THE QUALITY OF SPATIAL DATA AND ITS EFFECT ON SPECIES DISTRIBUTION MODELS

Czech University of Life Sciences Prague

Faculty of Environmental Sciences
Department of Applied Geoinformatics and Spatial Planning

May 2020

Czech University of Life Sciences Prague
Faculty of Environmental Sciences

The Quality of Spatial Data and its Effect on Species Distribution Models

Lukáš Gábor

This dissertation is submitted for the degree of *Doctor of Philosophy* at the Department of Applied Geoinformatics and Spatial Planning.

Prague May 2020

The Quality of Spatial Data and its Effect on Species Distribution Models

Dissertation Thesis

Author: Lukáš Gábor

Supervisor: Dr. Vítězslav Moudrý
Czech University of Life Sciences Prague

I hereby declare that the dissertation entitled *"The Quality of Spatial Data and its Effect on Species Distribution Models"* submitted for the Degree of Philosophy of Applied and Landscape Ecology is my original work guided by my supervisor. All sources of information, text, illustration, tables and images have been specifically cited.

.....

Acknowledgements

I would like to thank Vítězslav Moudrý for mentorship and brilliant guidance during the entire process of PhD studies. This dissertation would never have been written without his support, suggestions and never-ending reading of individual manuscripts. I am also thankful to Vincent Lecours, Tomáš Václavík, Helena Mitášová, Václav Petráš and Anna Petrášová for the opportunity to spend part of my research time abroad. Thanks to this, I got the opportunity to gain new experience and skills and to deepen my expertise in the field of species distribution modelling. Moreover, Vincent Lecours and Tomáš Václavík significantly contributed to the publications included in this dissertation. Big thanks go out also to Michal Fogl for preparation of LiDAR data used in some studies, to Michal Hnátek for invaluable help with all the paperwork and to Vojtěch Barták for improving my statistics and R skills.

Finally, I would like to thank my parents for endless and shoreless support throughout my entire studies. Without them, I would not be where I am.

THANK YOU ALL!

The scientific papers included in this thesis was funded by the following research projects:

- 17-17156Y
  Czech Science Foundation
- IGA 20174241
  Internal Grant Agency of the Faculty of Environmental Sciences, Czech University of Life Sciences Prague

Abstract

In the last decades, species distribution models (SDMs) have been widely applied to model species-environment relationships, often involving environmental variables (based on remotely sensed data) and species occurrences (based on field observations). Although these models are now routinely used, they still have critical limitations, especially those related to spatial data quality issues. However, studying the influence of varying spatial data quality on SDMs using real species would be extremely difficult as a real species itself brings additional (often unknown) uncertainties into the equation. For example, the real species prevalence or response to environmental gradients may be unknown or only approximate, as well as the magnitude of data error. Virtual species approach, on the other hand, allows researchers to isolate certain aspects of spatial data quality and to test its effect on SDMs. This thesis aims to test how different species characteristics and the quality of both species and environmental data affect SDMs and to answer the question whether species data or environmental data of high accuracy may be replaced by lower accuracy data. Specifically, the thesis addresses: 1) the influence of species characteristics (e.g. species response to environmental gradients, species prevalence, niche breadth); 2) the effect of species data quality (e.g. sample size, sampling bias, positional error); 3) the interactions between species characteristics and diverse data quality and 4) how different sources of environmental data (e.g. digital elevation models), its processing and subsequent use in modeling affect SDMs. Overall, the results showed that all these factors have a considerable effect on the output models. Therefore, it is always critical to evaluate the quality of input data with respect to their

source or the way of their processing. In the case of previously gathered species and environmental data where the accuracy is questionable or unknown, scientists should be cautious when interpreting their results. Where new surveys are undertaken, it is recommended to pay attention to data collection techniques to minimize the data error (e.g. positional error in species occurrences) and to help avoid its negative effect on SDMs. Additionally, it has been shown that there is a high level of interactions among individual species characteristics and the influence of various data quality on SDMs. Ignoring this may lead to misleading outcomes and conclusions.

Abstrakt (Czech)

Modely druhové distribuce jsou v posledních desetiletích běžně používány k modelování vztahů mezi druhem a prostředím, ve kterém se vyskytuje. Tyto modely jsou založeny na environmentálních proměnných (často získaných pomocí metod dálkového průzkumu země, DPZ) a datech o výskytu druhů získaných především při terénních pozorováních. Přestože jsou dnes tyto modely používány rutinně, stále narážejí na mnohá omezení, a to především na ta spojené s kvalitou vstupních dat. Studovat vliv různé kvality prostorových dat na modely druhové distribuce s použitím reálných druhových dat je nicméně velmi obtížné, jelikož reálné druhy vnášejí do celého procesu řadu dalších, často neznámých faktorů, které mohou výsledné modely ovlivnit. U reálných druhů se může stát, že například neznáme prevalenci nebo reakci na změny v prostředí, ve kterém se studovaný druh vyskytuje, stejně tak jako často neznáme chybovost použitých druhových dat. Oproti tomu použití virtuálního druhu umožňuje vědcům izolovat specifické aspekty spojené s kvalitou prostorových dat a studovat jejich vliv na výsledné modely. Cílem této disertační práce je testovat vliv různých druhových charakteristik a kvality prostorových dat (druhových i environmentálních) na modely druhové distribuce a zodpovědět otázku, zda a za jakých podmínek je možné nahradit přesná prostorová data těmi s nižší přesností. Specificky se tato disertační práce zabývá: 1) vlivem druhových charakteristik (např. reakcí druhu na environmentální gradient, druhovou prevalencí, šířkou niky); 2) vlivem různé kvality druhových dat (např. velikostí vzorku, nadměrným sběrem dat v určitých lokalitách, polohovou chybou); 3) interakcemi mezi druhovými charakteristikami a různou kvalitou prostorových dat a 4) otázkou,

jak různé zdroje environmentálních dat (např. digitálních výškových modelů), jejich zpracování a následné využití ovlivňují kvalitu modelů druhové distribuce. Výsledky této práce ukazují, že všechny tyto faktory mají na výsledné modely významný vliv. Proto je nezbytné před samotným modelováním kriticky zhodnotit kvalitu vstupních prostorových dat s ohledem na jejich zdroj nebo způsob zpracování. V případě využití druhových nebo environmentálních dat, která byla sbírána v minulosti a jejichž kvalita je neznámá nebo pochybná je nutné výsledné modely interpretovat obezřetně. V případě sběru nových dat, jak druhových, tak environmentálních, je nutné zaměřit se na správnou metodiku sběru s cílem minimalizovat jejich prostorovou chybovost (například polohovou chybu), která negativně ovlivňuje výsledné modely. Dále bylo prokázáno, že vliv různé kvality prostorových dat na modely druhové distribuce se mění v závislosti na odlišných druhových charakteristikách a ignorovaní těchto interakcí může vést k zavádějícím výsledkům a závěrům.


-NO PROB-
llama

Contents

I Introduction and Theory **19**

1 Thesis preface **21**

1.1 Foreword 21

1.2 Scientific motivation 22

1.3 Thesis structure 23

2 Objectives of the thesis **25**

3 Theoretical background **27**

3.1 Introduction 27

3.2 Species distribution models 29

3.3 Ecological niche 32

3.4 Species data and associated error 34

3.5 Environmental variables and associated error 38

3.6 Spatial scale 41

3.7 Virtual species spproach 43

II Research **47**

4 How do species and data characteristics affect species distribution models and when to use environmental filtering? 49

4.1 Introduction 51

4.2 Material and methods 54

4.2.1 Simulating ecological patterns with virtual species 54

4.2.2 Environmental filtering of sampled occurrences 58

4.2.3 Species distribution models 59

4.2.4 Assessment of model performance 60

4.3 Results 61

4.4 Discussion 65

4.4.1 Sample size 66

4.4.2 Species prevalence 67

4.4.3 Sampling bias 68

4.4.4 Environmental filtering 68

4.5 Conclusions 69

5 The effect of positional error on fine scale species distribution models increases for specialist species 71

5.1 Introduction 73

5.2 Material and Methods 77

5.2.1 LiDAR data acquisition, processing and variable selection 77

5.2.2 Simulating virtual species with different niche breadths 79

5.2.3 Simulating positional error in species occurrences 80

5.2.4 Model fitting and evaluation 81
5.3 Results 85
5.3.1 Unaltered models 85
5.3.2 Effect of positional error on models of species with different niche breadth 89
5.3.3 Comparison of the relative importance of individual predictors (R^2) 90
5.4 Discussion 91
5.5 Conclusions 95

6 On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs 97
6.1 Introduction 100
6.2 Materials and Methods 104
6.2.1 Study area and reference DTM 104
6.2.2 DEMs validation 105
6.2.3 Derived topographic indices 106
6.2.4 Virtual species generation 107
6.2.5 Model fitting and evaluation 108
6.3 Results and discussion 108
6.3.1 DEMs and derived attributes accuracy 108
6.3.2 SDM accuracy 110
6.4 Conclusions 115

7 Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies 117
7.1 Introduction 120

7.2 DTM-derived variables in SDM 124
7.2.1 Transparency in reporting information used to derive terrain attributes 125
7.2.2 Two common approaches to DTM scale alteration and how they produce different outcomes 127
7.3 Virtual species experiment 131
7.3.1 Study area 131
7.3.2 Calculating slope 131
7.3.3 Virtual species generation 132
7.3.4 Model fitting 132
7.3.5 Model evaluation 133
7.4 Results and discussion 134
7.5 Final remarks 134
7.6 Conclusions 137

8 Discussion and summary **139**
8.1 Influence of various species characteristics and species data quality on SDMs 140
8.2 Influence of various environmental data sources and their processing on SDMs 144
8.3 Virtual species approach 146
8.4 Conclusions 147
8.5 Further research 148
8.6 Afterword 149

References **151**

Curriculum Vitae & List of Publications **185**

Part I

Introduction and Theory

Chapter 1

Thesis Preface

1.1 Foreword

Biodiversity has been decreasing at a rate unprecedented in human history and its conservation should be the world´s highest priority. A proper understanding of the relationships between species and their environment represents a fundamental prerequisite for effective conservation actions. Species distribution models (SDMs), also known as species niche models or climatic models have, become a powerful tool helping scientists to understand such relationships. With the increasing availability of both species and environmental data in the last two decades, the implementation of SDMs has dramatically increased. However, despite this boom, the question of how such models are affected by the quality of spatial data has been only poorly investigated. We do neither know how SDMs based on data of poor quality may interact with species characteristics, nor how this may vary at different spatial scales. This lack of knowledge offers a broad field of research opportunities. Besides, handling spatial data through geographic information systems is often naive (e.g., different approaches to deriving terrain attributes derived from DEM are used), which in turn hampers the repeatability of studies. The presented thesis partially answers some unresolved issues related to interactions data and species characteristics and provides guidelines for appropriate spatial data handling.

Specifically, positional error in species occurrences, sampling bias and sample size across different species characteristics (i.e. prevalence, niche breadth), and the influence of various sources of environmental data and their processing on SDMs were explored.

1.2 Scientific motivation

Species distribution models (SDMs) use species occurrences and environmental data to produce a set of rules explaining the environmental space where species were collected or observed. In the last few decades, SDM-related methodological studies have been mostly focused on choosing appropriate modeling algorithms or evaluation metrics. However, to my surprise, the effect of varying data quality on SDMs has remained mostly uncharted for a long time, assuming that the input data were free of spatial error. Nonetheless, all spatial data inherently contain a certain level and type of spatial errors.

When I was going through prior studies focusing on the effect of varying quality of spatial data on SDMs, I realized several things. Firstly, studies that focused solely on the quality of species data often yielded contradictory conclusions. Secondly, prior studies mostly did not include interactions between various ecological characteristics of species and differences in data quality. Thirdly, I began to realize that many studies on SDMs use data non-critically, without the necessary GIS knowledge. Therefore, I personally believe that it is necessary to demonstrate, quantify and understand the consequences of using spatial data of varying quality in SDMs. Such research can lead to much-needed improvements in the current methodological standards for SDMs and its conclusions should be used for practical nature conservation and biodiversity protection.

1.3 Thesis structure

The thesis consists of four published studies and is divided into 2 Parts and 8 Chapters. The **Part I** contains a preface and general introduction into the field of species distribution modeling. The **Part II** consists of individual published studies:

- **Study I:** How do species and data characteristics affect species distribution models and when to use environmental filtering?

- **Study II:** The effect of positional error on fine scale species distribution models increases for specialist species

- **Study III:** On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs

- **Study IV:** Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies.

Chapter 2

Objectives of the Thesis

The aim of this thesis is to test how different species characteristics and quality of species and environmental data affect SDMs and to answer the question whether species and/or environmental data of high accuracy may be replaced by those of lower accuracy. Specifically, the thesis addresses: **1)** the influence of species characteristics (e.g. species response to the environmental gradient, species prevalence, niche breadth); **2)** the effect of species data quality (e.g. sample size, sampling bias, positional error); **3)** the interactions between species characteristics and diverse data quality and **4)** how different sources of environmental data (e.g. digital elevation models), its processing and subsequent use in modeling affect SDMs.

Chapter 3

Theoretical Background

3.1 Introduction

Biodiversity is declining throughout the world and understanding of how species are distributed in space and time is crucial for mitigating that decline and answering conservation questions at hand. Over the last few decades, this effort was facilitated by advances in various modeling techniques. The objective of modelling is to relate an in situ response variable (e.g. species occurrences) and explanatory variables (environmental variables, often derived by remote sensing) to describe the relationship between them or to predict unknown values of the response variables characterizing biodiversity (Ferrier et al. 2017). Species distribution models (SDMs) represent the most frequently used tool for such analyses. Although these models are now routinely used, they still have critical limitations, especially those related to spatial data quality (Araújo et al. 2019).

The use of accurate spatial data (e.g. Duputié et al. 2014, Guillera-Arroita et al. 2015, Araújo et al. 2019, Gábor et al. 2019, 2020) is an elementary prerequisite for creating a valid SDMs. Unfortunately, there are many inherent sources of uncertainty of (both species and environmental) input data that can affect the model performance. Species data may be affected for example by sampling bias (Isaac and Pocock 2015, Boria et al. 2014), by low sample size (Pearce and Ferrier 2000, Stockwell and Peterson 2002), by positional error (Johnson and

Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009) or limited by scale (Šímová et al. 2019). Similarly, environmental data can be negatively affected by positional error (Osborne and Leitão 2009), by its origin or type (Moudrý et al. 2018), by the way of data processing (Moudrý et al. 2019b) or resolution (Gottschalk et al. 2011, Turner et al. 2019). All these play a role when modelling species distribution and were more or less explored.

In addition, the performance of SDMs is complicated by various spatial (e.g. prevalence or range size) and ecological (e.g. niche breadth) characteristics of the studied species (Luoto et al. 2005, Bulluck et al. 2006, McPherson and Jetz 2007, Evangelista et al. 2008, Chefaoui et al. 2011, Connor et al. 2018). These characteristics are usually studied separately or in combinations of two or three (but see Thibaud et al. 2014, Fernandes et al. 2018, Liu et al. 2019) and rarely together with data quality issues. Therefore, studies addressing both are particularly valuable as they help understand interactions between ecological characteristics of the studied species and issues related to poor data quality.

Studying the influence of different spatial data quality and its interactions with various species characteristics on SDMs with real species is challenging. For example, species response to the environmental gradient, its prevalence or rarity rate is unknown. In addition, the magnitude of error in the data is often unknown or approximate. The use of a virtual species approach, on the other hand, facilitates the isolation of certain aspects of spatial data quality and species characteristics and testing their effects on SDMs (Zurell et al. 2010). Therefore, this approach is increasingly used for evaluating the effects of data inaccuracies on model performances (see for example Meynard et al. 2019).

The presented thesis focuses on answering the question of how differences in the quality of spatial data and in its processing, together with differences in species characteristics, affect species distribution models.

3.2 Species distribution models

Understanding of the relationship between species and its environment is an ongoing effort of ecologists and biogeographers, representing one of the cornerstones of these fields (see for example Humbold von and Bonpland 1807). For this purpose, species and environment data from in situ (i.e. field) observations are combined and used for explanatory studies. The first experiments on modeling species-environment relationship can be dated back to the end of 1950s (e.g. Hairston 1959, Neyman and Scott 1959, MacArthur 1960). The development of computing technology, geographic information systems, remote sensing and various statistical methods (with the greatest boom in the mid-1980s) led to attempts to use more complex species-environment combinations and to the first experiments with modeling species distribution (predictive studies; see review by Ferrier et al. 2017) (see Figure 3.1 for SDMs workflow).

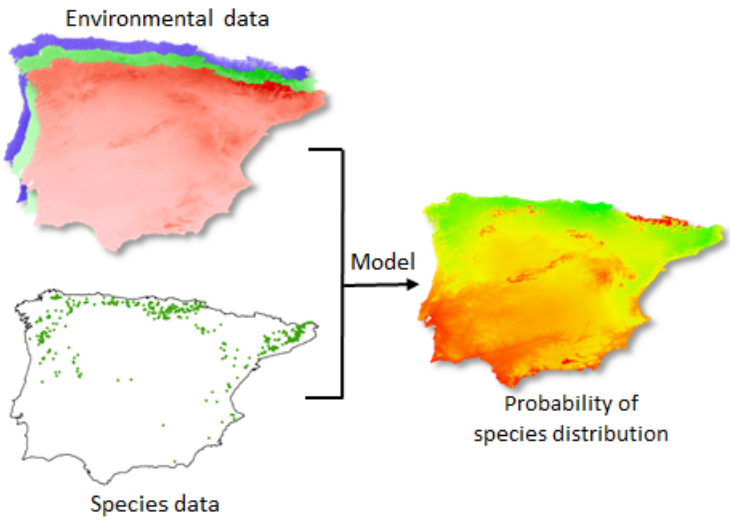

Figure 3.1 SDMs workflow. Species data from in situ field observations and various environmental data (for example derived from remote sensing) are combined to study species-environment relationships or to predict species distribution in geographical space.

As mentioned above, the objective of such modelling is to relate an in situ response variable (e.g. species distribution) and explanatory variables (e.g. remotely sensed environmental data) in order to describe the relationship between these two or to predict unknown values of the biodiversity response variable (Ferrier et al. 2017). Later, these models became known as ecological niche models (ENMs) or, more often, species distribution models (SDMs). They are widely used, for example, in determining locations potentially threatened by invasive species or for studying the impact of climate change on biodiversity (see Table 3.1 for more examples).

Table 3.1 Examples of various uses of species distribution models.

Assessing species invasion	Battini et al. 2019, Guan et al. 2020
Assessing the impact of climate changes on species distributions	Della Rocca et al. 2019, Sun et al. 2020
Modelling species assemblages from individual species predictions	Zurell et al. 2020, Norberg et al. 2019
Quantifying the environmental niche of species	Chen et al. 2019, Manzoor et al. 2020
Suggesting unsurveyed sites with a high potential of occurrence for rare species	Escalera-Vázquez et al. 2018, McCune 2019
Supporting conservation planning	Filer et al. 2020, Préau et al. 2020
Testing biogeographical, ecological and evolutionary hypotheses	Soley-Guardia et al. 2019, Dufresnes et al. 2020

Based on Guisan and Thuiller 2005

SDMs can be classified into three categories: mechanical, empirical and analytical models. When modeling species distribution, two out of three desirable model characteristics (reality, precision and generality) can be simultaneously optimized when a model is developed and refined (Guisan and Zimmermann 2000). This fact is still generally accepted because no model can simultaneously achieve a high performance

(precision), be based on natural processes (reality) and be universally applicable (see Figure 3.2).

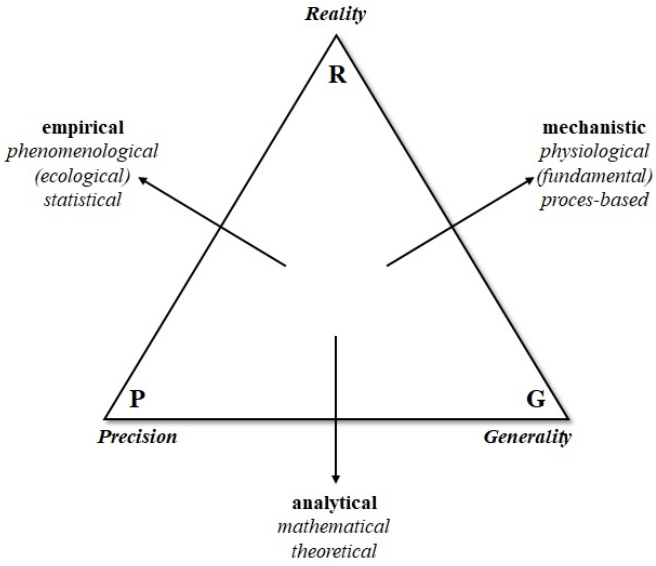

Figure 3.2 Classification of species distribution models based on their intrinsic properties (Guisan and Zimmermann 2000).

The process of modeling species distribution could be divided into three interconnected parts (Austin 2002, 2007, Williams et al. 2012):

- conceptual model based on ecological theory
- data model
- statistical model

The formulation of the conceptual model includes defining the objectives of modeling, postulating working hypotheses and, in particular, deciding what environmental data are relevant for studied species. The data model part should be focused on both species (response) and environmental (explanatory) data. For species data, their type (i.e. presence-only versus presence-absence) and source (e.g. systematic filed surveys, records from museums or global databases) are important. For

example, techniques appropriate for presence-absence data modeling include generalized linear models, generalized additive models or boosted decision trees whereas Bioclim, Maxent or multivariate distance are designed for presence-only (presence-background) techniques (Elith et al. 2006). For environmental data, their resolution and extent, source (in situ observations versus remote sensing) and the way of processing play a role. Finally, statistical modeling includes the elimination of poorly performing variables and selection of proper statistical methods of model fitting and evaluation (Williams et al. 2012).

Many studies focused on the conceptual model formulation and statistical model part (e.g. Austin 2007, Jiménez-Valverde et al. 2008, Elith and Graham 2009, Peterson and Soberón 2012). Nevertheless, the data model part has been neglected because the availability of both species and environmental data was limited. The increased data accessibility in the last few decades, however, changed the situation considerably. Unfortunately, this has led to combining data of different quality without any advanced knowledge of how this could affect SDMs (Bayraktarov et al. 2019, Isaac et al. 2020). Therefore, this dissertation aims to contribute to the knowledge related to the data model part (sensu Austin 2002) of modelling species distribution.

3.3 Ecological niche

The term ecological niche describes how, rather than just where, species live (Townsend et al. 2003). It is a quite old term that has undergone much development over the last century (see for example Grinnell 1917). The modern concept of the ecological niche was established by Hutchinson in 1957. Hutchinson (1957) defined the niche as a hypervolume n-dimensional area shaped by the environmental conditions under which species can exist indefinitely. It is easy to illustrate this relationship for a one-dimensional niche. An example of this is the change of the

probability of species occurrence along an environmental gradient (e.g. temperature; see Figure 3.3).

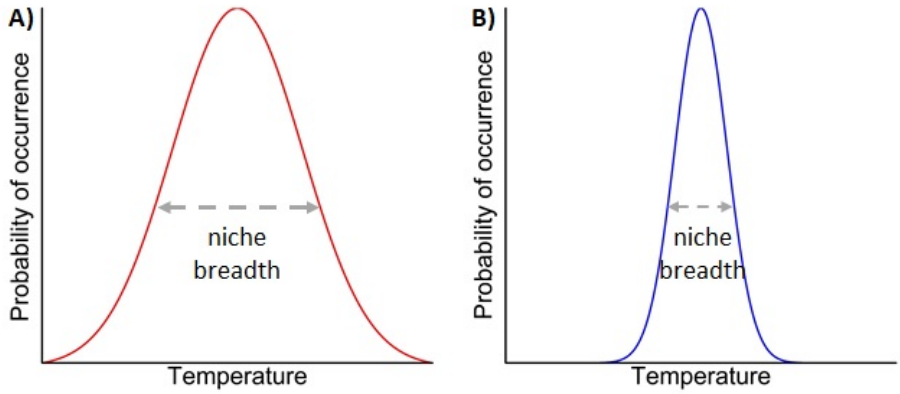

Figure 3.3 An example of a changing probability of species occurrence for species with wider (A – red line) and narrower (B – blue line) niche breadth along the temperature gradient (one-dimensional species niche).

Hutchinson (1957)´s definition of ecological niche forms the theoretical basis for describing the relationship between species and environmental variables, which is crucial for understanding ecological processes and has been used to frame SDMs studies (Franklin 2010).

This definition was further developed and in 1961, Hutchinson introduced new terms - the fundamental and the realized niche (Figure 3.4). The term "fundamental niche" describes the range of natural conditions where a species is naturally capable of living whereas the "realized niche" exemplifies the real distribution of species. According to Hutchinson (1961), the realized niche is made up of subsets of the fundamental niche as a result of biotic interactions (e.g. predation, symbiosis). Additionally, it is expected that the realized niche depends also on the biogeography, respectively on the historical occurrence of species. Although additional information on the co-occurrence of competing / host species and detailed historical species records is rarely available for modeling (but see Schweiger et al. 2012, Singer et al. 2018,

Zurell 2017), it is generally agreed that SDMs allow us to quantify the realized niche of the species (Guisan and Thuiller 2005).

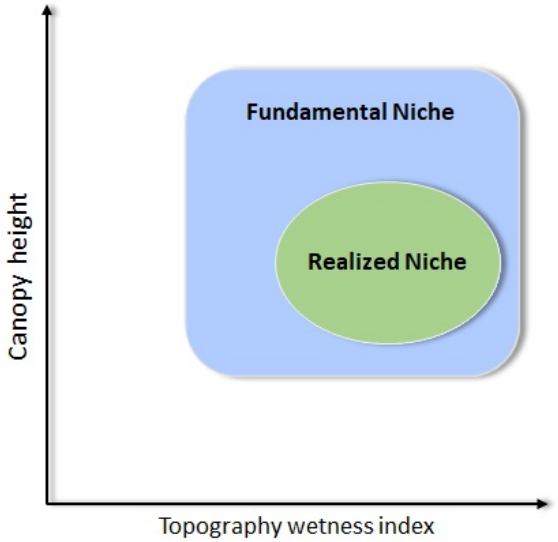

Figure 3.4 A conceptual diagram of the Hutchinson's (1961) fundamental and realized niche. The fundamental niche (blue) color illustrates the fundamental range of natural conditions where a hypothetical species is naturally capable of living whereas a realized niche (green) exemplifies the real distribution determined by biotic interactions (e.g. predation, symbiosis).

3.4 Species data and associated error

In terms of data collection, species data could be divided into structured and unstructured. Structured data are gathered from stratified, repeatable sampling designs, which are mostly geographically restricted (Kindsvater et al. 2018, Peterson and Soberón 2018), and it is expected that they are free of any kind of spatial bias. Unstructured data, on the other hand, suffer from various types of spatial bias; most of the currently available species occurrence data are of this type (Isaac and Pocock 2015, Isaac et al. 2020).

Both structured and unstructured data are increasingly combined in online global databases such as eBird (www.ebird.org), Vertnet (www.vertnet.org) or iNaturalis (www.inaturalist.org) where not only scientists but also general public throughout the world share their field observations. Data from these databases are usually freely available and together with data from other sources (e.g. museum records, atlases or natural history collections) easily available for download through global aggregators such as Global Biodiversity Information Facility (GBIF; www.gbif.org) or Ocean Biogeographic Information System (OBIS; www.obis.org). While the number of records in aggregated databases is constantly growing, their spatial quality varies and not all of them are, therefore, necessarily useful for modeling species distribution (Bayraktarov et al. 2019, Moudrý and Devillers 2020). Even though some ecologists argue that quality datasets are essential for decision-making processes (e.g. Bayraktarov et al. 2019), the majority of them may still be seduced by the idea that *"the more data, the better"*. However, existing studies addressing the question of whether a smaller sample size of more accurate data is better than a larger sample size with an inferior positional accuracy yielded inconsistent results (Reside et al. 2011, Mitchell et al. 2017, Bayraktarov et al. 2019, Gábor et al. 2020).

When modeling species distribution, high quality occurrence records were suggested to generate informative and accurate SDMs (Osborne and Leitão 2009, Duputié et al. 2014, Moudrý et al. 2017). Most of the species in global databases are however under-sampled, particularly rare and endangered species (i.e. those of the highest importance from a conservation perspective), resulting in a sample size that is too low to provide reliable models. The effects of the sample size on model performance have been studied extensively (e.g. Jiménez-Valverde et al. 2009, Moudrý and Šímová 2012), although no consensus has been reached; some studies concluded that even very small sample sizes can provide reliable models (Varela et al. 2014, Proosdij et al. 2016)

while others have shown the opposite (Wisz et al. 2008, Tessarolo et al. 2014). Therefore, additional studies focusing on this topic, especially in combination with additional data quality issues, are needed.

Furthermore, species occurrence records are often spatially biased (i.e., the sampling effort is uneven, typically higher in protected or easily accessible areas) (Isaac and Pocock 2015). Such sampling bias has been reported towards easily accessible areas (Reddy and Dávalos 2003), protected areas (Boakes et al. 2010), or more populated areas (Geldmann et al. 2016). In the case of global databases, the sampling bias can occur because species records are shared only by some countries (Beck et al. 2014). It is important to account for spatial bias in SDMs because it may affect model calibration and cause an overestimation of SDM performance (Leitão et al. 2011, Hijmans 2012, Boria et al. 2014). Various methods have been proposed to compensate for sampling bias in species occurrence records, including manipulation of background data (Phillips et al. 2009) and spatial filtering (Veloz 2009, Anderson and Raza 2010, Boria et al. 2014, Tessarolo et al. 2014). Recently, Varela et al. (2014) suggested that spatial filtering could fail because species occurrences with unique environmental conditions could be removed. Instead, they suggested the use of environmental filtering to down-weight repeated species occurrences in similar environmental conditions. However, filtering necessarily reduces the sample size, and although Varela et al. (2014) suggested that a filtered subsample of occurrences can be better than using all available records to calibrate models, the trade-off between lower sample size after filtering and higher sample size without filtering has yet to be tested.

Besides, most applications of SDMs naively assume that species occurrence data are free of positional error, even though it is inherently present in all datasets. The negative influence of positional error in spatial modelling is a long-known fact (Heuvelink 1998) and many studies addressed this issue. For example, several studies explored the limits and impacts of image registration errors in remote sensing (Townshend

et al. 1992, Wang and Ellis 2005), others suggested solutions for handling georeferencing errors and calculating uncertainty (e.g. Wieczorek et al. 2004). When modelling species distribution, it is often expected that the negative effects of positional error in species occurrence data are minimal or mainly associated with relatively older datasets that are often georeferenced from textual descriptions of their locations, which may cause errors of up to hundreds of meters (Wieczorek et al. 2004). However, it is also necessary to consider positional errors inherent to data georeferenced using global navigation satellite systems (GNSS). Moreover, species occurrence data often represent the position of the observer and not the actual position of the species (Zhang et al. 2018). Therefore, even though the accuracy of a standard GNSS is usually below 30 meters (Frair et al. 2010), the errors associated with such data may be much bigger. The number of studies focusing on the influence of positional error in species occurrences on the performance of SDMs has been growing, there is, however, still little consensus on how this influence is manifested. For instance, while Graham et al. (2008) or Mitchell et al. (2017) concluded that SDMs are robust to positional error, others argued that positional errors reduce the model performance (Johnson and Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009). Furthermore, prior studies used relatively coarse environmental data (but see Mitchell et al. 2017). Positional error considered in prior studies ranged from 50 m up to 50 km (see Table 5.1). While such error results in a shift over several cells in a coarse-resolution SDM (e.g. 1 × 1 km), it will cause a much greater shift in a fine-resolution SDM (e.g. 10 × 10 m). Therefore, with the increasing availability of fine-scale data, additional studies are needed (Osborne and Leitão 2009); it can be expected that SDMs at fine scales would be more sensitive to positional error. In addition, it is intuitive that positional error of a given magnitude might have a greater effect on specialist (narrow niche breadth) than generalist species (wide niche breadth), as it is more likely that occurrences get incorrectly shifted

into cells representing an unsuitable environment, i.e. an environment that is outside of the species' environmental niche. This, however, has never been thoroughly explored.

3.5 Environmental variables and associated error

Species distribution models rely on the availability of environmental variables in the form of spatial data (Goodchild 1996, Franklin 2010), which can be divided into three different categories, depending on their effect on species distribution (Austin 1980: resource, direct and indirect variables).

- Resource variables have a direct impact on the species growth (e.g. the amount of light, water, nutrients)
- Direct variables have a direct impact on species growth, but in a different way than resources (e.g. temperature, humidity, pH)
- Indirect variables don't have a direct impact on species growth but usually correlate with the resource or direct variables (e.g. altitude, slope)

The traditional way of gathering mainly resource and direct environmental variables were long-term in situ observations. As an example of such data sources, we could name the global weather and climate database WorldClim (www.worldclim.org). Available environmental variables from this database were created by interpolating data obtained from in situ weather station measurements (Hijmans et al. 2005, Fick and Hijmans 2017). However, the weather station network across the world is sparse and thus, the final resolution of these data is coarse. Advances in in-situ monitoring, such as the Temperature-Moisture-Sensor (TMS; allows to gather air, surface and soil temperatures or moisture) (Wild

et al. 2019) facilitated measuring resource and direct variables on a fine scale. Still, fine-scale in situ monitoring remains limited to small areas.

For large areas and repeated monitoring, an alternative option is to use remote sensing data. It has been shown that a wide range of environmental variables derived from various types of remote sensing can be used as explanatory variables for species distribution (Mahecha et al. 2017, Randin et al. 2020). For example, the use of data on land cover derived from passive remote sensing is common (Seoane et al. 2004, Venier et al. 2004, Verburg et al. 2011). Similarly, the climatic variables such as temperature and precipitation are increasingly based on remote sensing data (Naumann et al. 2012, Chen and Li 2016, Macharia et al. 2020).

An excellent example of how SDMs benefit from advances in remote sensing is the possibility to use 3D ecosystem structure variables derived from active remote sensing methods. 3D ecosystem structure was long ago suggested as an important variable that plays a role in species distribution (e.g. of birds) (Dunlavy 1935). Nevertheless, its standardized measurement was extremely problematic in the past (MacArthur and MacArthur 1961, Brown 1981). Nowadays, 3D vegetation structure is commonly measured using active remote sensing methods such as LiDAR (light detection and ranging - a remote sensing method that uses light in the form of a pulsed laser to measure ranges) and variables representing 3D vegetation structure were shown to be important variables of species distribution. As example, we can mention the canopy structural variability or understory density derived from airborne laser scanning data (see reviews by Davies and Asner 2014, Bakx et al. 2019.

Arguably one of the most common remote sensing products used in SDMs is the digital elevation model (DEM) and terrain attributes derived from it (e.g. slope, aspect, topographic wetness index). DEM's derived attributes can be used as surrogates for a variety of field-measured environmental variables such as air temperature, soil moisture and incoming solar radiation (Hengl and Reuter 2009). For example,

topographic wetness index is a surrogate for soil moisture, an environmental variable that affects the vegetation composition (Besnard et al. 2013, Reif et al. 2018). Slope is another example of such a variable. Slope affects the velocity of subsurface and surface flow and it is, therefore, an important variable in predictive vegetation mapping (e.g. Zhang et al. 2016).

Although highly accurate DEMs exist at both local and national levels (for example from airborne laser scanning missions), many studies rely on global space-borne DEMs that have lower spatial resolutions and accuracy (e.g. Zhang et al. 2016). Nowadays, global or near-global DEMs are available from several spaceborne missions: Shuttle Radar Topography Mission (SRTM), Advanced Spaceborne Thermal Emission Reflectometer (ASTER) onboard NASA's Terra satellite, Advanced Land Observing Satellite (ALOS), or TANDEM-X. Unfortunately, both interferometric (SRTM, TANDEM-X) and stereoscopic (ASTER, ALOS) DEMs suffer from local inaccuracies or errors due to limitations associated with the methods used for elevation measurements. It has been shown that such inaccuracies and errors can in turn influence the derived topographic indices (Van Niel et al. 2004, Oksanen and Sarjakoski 2005, Sofia et al. 2013, Lecours et al. 2017c, Moudrý et al. 2018) and various steps of the species distribution modelling process (e.g., the shape of response curves, prediction accuracy measures, spatial extent of predictions) (Van Niel and Austin 2007, Lecours et al. 2017b, Moudrý et al. 2019b).

SRTM DEM (1 arc-second resolution; approximately 30m at the equator) is one of the most commonly used global DEMs. However, an important but often misunderstood characteristic of the SRTM DEM is that it does not provide a "bare-earth" elevation: the measurements actually include a systematic positive bias due to the objects above the ground (such as canopy), the height of which is included into the model. This in turn produces considerable differences in accuracy between forested and open areas (e.g. Nelson et al. 2009). All

available versions of the SRTM DEM are impacted by vertical error, including one of the most – if not the most – cited versions of the SRTM DEM produced by the Consultative Group for International Agriculture Research Consortium for Spatial Information (CGIAR-CSI; http://www.cgiar-csi.org/data; e.g. Šímová et al. 2015, Kosicki 2017). By not acknowledging the vertical error, and more particularly its vegetation offset component, most studies use global DEMs as digital terrain models (DTM). If using the appropriate nomenclature, the original SRTM product and many of its subsequent alterations are actually digital surface models (DSM); they do not represent the bare ground elevation in vegetated areas and require further processing to remove vegetation height in order to create a proper 'bare-earth' DTM.

3.6 Spatial scale

Spatial scale is one of the most important characteristics of spatial data and requires a thorough consideration in any ecological study. Three types of spatial scale are relevant for modeling species distribution: (1) the ecological scale, which is the scale at which a pattern or process occurs, (2) the observational scale, which refers to the characteristics of the data, usually defined by the spatial resolution and extent of the data and (3) the analytical scale that refers to the methods used to analyse the environmental data (e.g., the neighbourhood size used in focal statistics or geomorphometry) (Dungan et al. 2002, Moudrý et al. 2019b). It is well known that the lack of explicit consideration of scale affects the outcomes of ecological analyses such as the assessment of species-environment relationships (Levin 1992, Mertes and Jetz 2018). Almost 30 years ago, Wiens (1989) argued that most ecological studies had been ignoring spatial scale and its effects. Most studies were performed as if patterns and processes were scale-independent and studies from different scales were often inappropriately compared. Since then, the role and importance of spatial scale have been extensively discussed in both the geographic and ecological literature (e.g., Schneider 2001,

Dungan et al. 2002, Goodchild 2011, Moudrý and Šímová 2012), and it is now widely recognized that ecological patterns and processes are scale-dependent and that no single scale is appropriate for the study of all natural phenomena.

One of the main challenges of using remotely sensed data in SDMs is that the original spatial resolution of different datasets included in the analysis may vary significantly (Cord et al. 2013). In terrestrial applications, the use of high-resolution remotely sensed environmental data is often limited by the resolution of species distribution data, which are usually available at much coarser scales (Jetz et al. 2012, Šímová et al. 2019). An opposite situation may occur in marine applications (e.g., in deep water environments), where the scale at which the species are observed can be much finer than that of available environmental data. In some cases, the highest available resolution of environmental data may not be required for the SDMs if the biological / ecological processes, or species distribution / abundance in SDMs, occur at a coarser scale or over a large area with limited observations. Therefore, finding a resolution representing a compromise between the resolution of available data and a resolution most suited to the application is often necessary. A common practice to ensure the valid integration of data from multiple scales (for example, to avoid an ecological fallacy or the modifiable areal unit problem; see Lecours et al. 2015), is to modify the resolution of some of the data so that it matches the resolution at which the study is to be performed (e.g., by averaging environmental variables within field plots; Gottschalk et al. 2011, Moudrý et al. 2017).

The effects of altering data resolution (i.e., matching the observational scales with ecological scales) (e.g. Lechner et al. 2012b, Svensson et al. 2013, Mateo Sánchez et al. 2014, Mertes and Jetz 2018) on the outcomes of ecological analyses (e.g., SDM) have recently received more attention because of a growing demand on users to provide more detailed methodologies (e.g., exact computer code for GIS analyses and data processing) to allow complete reproducibility of their results (e.g., Mich-

ener and Jones 2012, Rocchini and Neteler 2012, Meynard et al. 2019). However, details on the use of terrain attributes derived from digital terrain models (DTMs) using neighbourhood operations (e.g., slope, rugosity, orientation, curvature) in SDMs has received less attention. Deriving terrain attributes has become a routine operation, despite several potential pitfalls in the data processing workflow (Lecours et al. 2017c). Specifically, the approach used and the step at which the scale is altered are particularly important for DTM-derived attributes, as they will cause the analytical scale of a study to vary, thus providing different representations of the reality and potentially producing different outcomes. Therefore, selecting an inappropriate scale alteration technique could have an unforeseen impact on SDMs. However, this has not been demonstrated yet. Thus, there is a need to show how SDMs may be affected when the different approaches to altering data resolution are used.

3.7 Virtual species spproach

To study a compound effect of various species and data characteristics (both species and environmental) using real species could be challenging due to the above-mentioned data complexities. A virtual species, which is increasingly used in ecological studies (see Table 3.2), on the other hand, allows to ensure the full knowledge of the exact ecological and geographical characteristics of the species and to avoid unknown complexities associated with real data.

The process of simulation involves four steps: (i) generating a virtual species, (ii) projecting it into the landscape, (iii) converting its probability to presence-absence data and (iv) sampling occurrences (Figure 3.5; Meynard et al. 2019). The first step involves definition of the species response to environment (e.g. gaussian, linear, logistic or beta response) using one or more environmental variables and then combining these functions into a single suitability function. In this step, potential effects of environmental variables (i.e. their complexity and number), and

species characteristics (i.e. commonness or rarity and niche breadth) could be addressed.

Table 3.2 Examples of various virtual species studies in SDMs.

Environmental data resolution and extent effects	Moudrý et al. 2019b, Friedrichs-Manthey et al. 2020
SDMs performance metrics effects	Leroy et al. 2018, Warren et al. 2020
Species spatial data quality, sampling size or sampling bias effects	Fernandes et al. 2018, Liu et al. 2019
Species specialization, niche breadth effects	Proosdij et al. 2016, Connor et al. 2018
Effects of various SDMs modeling techniques, approaches	Zurell et al. 2016, Hallgren et al. 2019

Based on (Meynard et al. 2019)

In the second step, the simulated virtual species is projected into the landscape (real or virtual). As highlighted by Meynard et al. (2019), the use of real environmental data has the advantage of being simple and allowing a realistic set of explanatory environmental variables with collinearity and interactions that could be related to real case studies. On the other hand, a virtual landscape allows the use of environmental variables with various heterogeneity / homogeneity or with different spatial autocorrelation. At this step, the influence of the resolution and extent of environmental variables, of their processing strategies, or of climate change could be tested.

The next step is to convert the probability of virtual species occurrence (generated in the previous step) into a presence-absence distribution. Two main conversion methods for translating initial suitability (probability of virtual species occurrence) to presence/absence are currently used: a threshold approach (Hirzel et al. 2001) and a probability approach (Meynard and Kaplan 2012, 2013, Meynard et al. 2019). The threshold approach generates occurrences where species always occur above a given threshold and never below it. On the other hand,

the probability approach allows generating species presences and absences along the whole environmental gradients (i.e. the shape of the logit function used to transform the occurrence probability to presences/absences). Moreover, the probability approach allows to consider species prevalence.

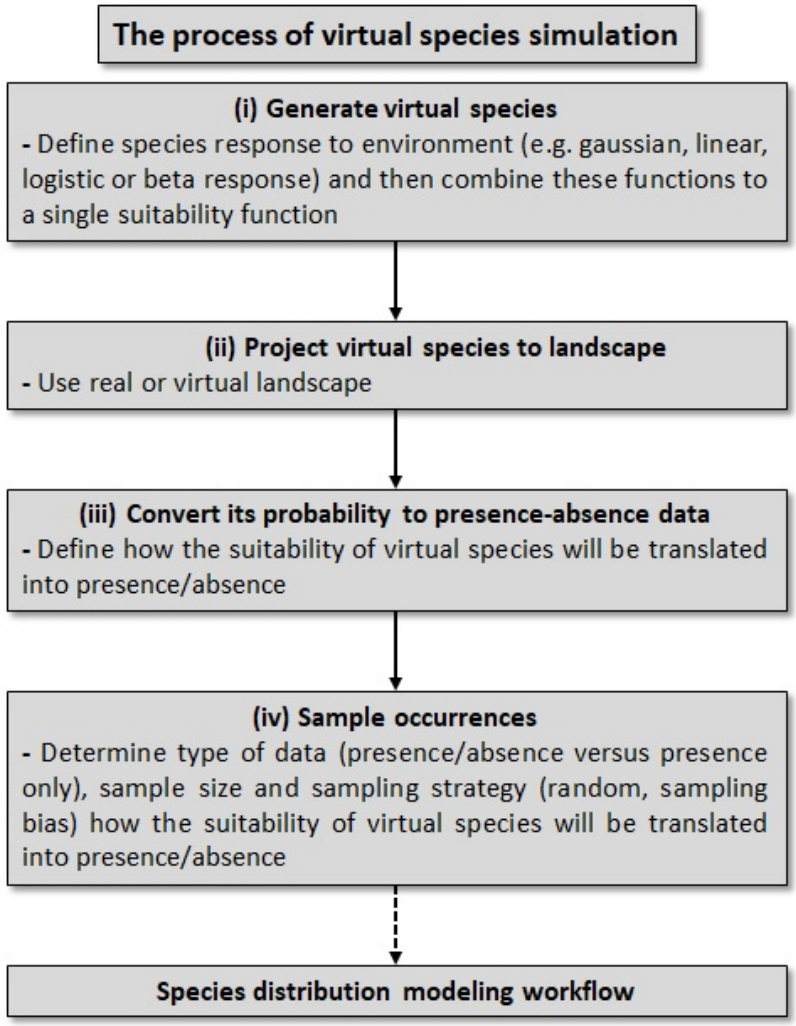

Figure 3.5 The virtual species approach simulation process (Meynard et al. 2019).

Thus, the probability approach is closer to ecological theories supporting the idea of dynamic occupancy patterns in space and time (see Hanski 1998, Meynard and Kaplan 2012, 2013). Therefore, the probability approach has been deemed more appropriate for generating virtual species than the threshold approach (Meynard and Kaplan 2012, 2013, Moudrý 2015, Meynard et al. 2019). In this step, the effect of species prevalence and dispersal constraints could be analyzed.

The last step is sampling species occurrences. Here, different data types (presence only versus presence/absence), sample size or sampling strategy could be assessed.

Part II

Research

Chapter 4

How do species and data characteristics affect species distribution models and when to use environmental filtering?

Lukáš Gábor, Vítězslav Moudrý, Vojtěch Barták, Vincent Lecours

International Journal of Geographical Information Science (2019): 1–18.

Publication metrics:
34 of 155 (Q1) rank in WOS category Computer Science, Information Systems
IF (2018) 3.545; AIS (2018) 0.707

Author's contribution: 60%

Abstract

Species distribution models (SDMs) are widely used in ecology and conservation. However, their performance is known to be affected by a variety of factors related to species occurrence characteristics. In this study, we used a virtual species approach to overcome the difficulties associated with testing of combined effects of those factors on performance of presence-only SDMs when using real data. We focused on the individual and combined roles of factors related to response variable (i.e. sample size, sampling bias, environmental filtering, species prevalence, and species response to environmental gradients). Results suggest that environmental filtering is not necessarily helpful and should not be performed blindly, without evidence of bias in species occurrences. The more gradual the species response to environmental gradients is, the greater is the model sensitivity to an inappropriate use of environmental filtering, although this sensitivity decreases with higher species prevalence. Results show that SDMs are affected to the greatest degree by the species response to environmental gradients, species prevalence, and sample size. Models' accuracy decreased with sample size below 300 presences. Furthermore, a high level of interactions among individual factors was observed. Ignoring the combined effects of factors may lead to misleading outcomes and conclusions.

Keywords: MaxEnt, Schoener´s D, Species rarity, Spatial data filtering, Virtual species

4.1 Introduction

Many of the modeling techniques developed in the last two decades are now recognized to play an important role in monitoring of biodiversity and its conservation (Guisan and Zimmermann 2000, Honrado et al. 2016). Species distribution models (SDMs) have become a common tool for the assessment of species-environment relationships. The objective of SDMs is to relate species occurrence data (i.e. response variable) and environmental data (i.e. predictor variables) in order to either describe relationships between them ('explanatory modeling') or predict probabilities of species occurrences at unsampled sites or times ('predictive modeling') (see review by Ferrier et al. 2017). SDMs are now routinely used, for example to assess the spread of invasive species (Gillard et al. 2017, Bazzichetto et al. 2018a), the impact of climate change on biodiversity (Sun et al. 2017), or species ranges (Williams and Crouch 2017). High-quality species occurrence records (i.e. unbiased, positionally accurate data without false presences and absences) are essential to generate informative and accurate SDMs (Osborne and Leitão 2009, Duputié et al. 2014, Moudrý et al. 2017). However, acquisition of such data is often challenging and the underlying challenge in SDMs is to derive response curves from incomplete and biased datasets.

In practice, the most commonly available species records are usually non-systematic observations (see Bino et al. 2014), such as collections of individual observations from various sources (e.g. museums, citizen science data) available through global databases (e.g. the Global Biodiversity Information Facility – GBIF; www.gbif.org). This type of species observations are referred to as presence-only records (presence-background records sensu Guillera-Arroita et al. 2015). Most of the species in global databases are however under-sampled, particularly rare and endangered species (i.e. those of the highest importance from a conservation perspective), resulting in a sample size that is too low to

provide reliable models (but see Breiner et al. 2015, 2018 for possibilities of overcoming limitations of modeling species with few occurrences). The effects of sample size on model performance have been studied extensively (e.g. Jiménez-Valverde et al. 2009, Moudrý and Šímová 2012), although no consensus has been reached; some studies concluded that even very small sample sizes can provide reliable models (Guisan et al. 2007, Varela et al. 2014, Proosdij et al. 2016) while others have shown the opposite (Wisz et al. 2008, Tessarolo et al. 2014).

Furthermore, species occurrence records are often spatially biased (Isaac and Pocock 2015), which is usually caused by uneven sampling efforts or data sharing. Such bias has been reported for data collected in easily accessible areas (Reddy and Dávalos 2003), protected areas (Boakes et al. 2010), or heavily populated areas (Geldmann et al. 2016). It is important to account for spatial bias in SDMs because it may affect model calibration and cause an overestimation of SDM performance (Leitão et al. 2011, Hijmans 2012, Boria et al. 2014). Various methods have been proposed to compensate for sampling bias in species occurrence records, including manipulation of background data (Phillips et al. 2009) and spatial filtering (Veloz 2009, Anderson and Raza 2010, Boria et al. 2014, Tessarolo et al. 2014). Spatial filtering is used to reduce the negative influence of sampling bias in geographic space. Recently, however, Varela et al. (2014) suggested that this approach could fail because species occurrences with unique environmental conditions could be removed. Instead, they suggested the use of environmental filtering to down-weight repeated species occurrences in similar environmental conditions, which we also adopted in this study. Increasing attention has also been given to comparison or evaluation of those methods (Kramer-Schadt et al. 2013, Varela et al. 2014, Ranc et al. 2016). Filtering necessarily reduces the sample size, and although Varela et al. (2014) suggested that a filtered subsample of occurrences can be better than using all available records to calibrate models, the tradeoff between lower sample size after filtering and higher

sample size without filtering has yet to be tested.

In addition to the quality of occurrence data (e.g. sample size, sampling bias) and methods used to filter the data (e.g. environmental filtering, geographic filtering), species characteristics also need to be considered. Studies have shown that commonness and rarity or prevalence may influence the ability to predict species distribution;models for rare species (i.e. species with low prevalence) tend to have higher prediction accuracy than models generated for more common species (i.e. with high prevalence; Syphard and Franklin 2009, Sor et al. 2017).

Species characteristics (e.g. prevalence, response to environmental gradients) and data characteristics (e.g. bias, filtering, sample size) are usually studied separately or in combinations of two or three factors (but see e.g. Thibaud et al. 2014, Fernandes et al. 2018, Liu et al. 2019). It is therefore difficult to determine a characteristic affecting SDMs performance the most, as well as to evaluate potential interactions between species and data characteristics. In this study, we used a virtual species approach to assess the effects of prevalence, response to environmental gradients, sampling bias, sample size, and samples filtering, as well as their interactions, on SDMs performance. The use of virtual species approach enables full control over the factors influencing models and the disentanglement of confounding effects (Zurell et al. 2010, Miller 2014). Consequently, this approach is increasingly used to evaluate SDMs performance (e.g. Václavík and Meentemeyer 2012, Qiao et al. 2015, Moudrý et al. 2018). To test how the five species and data characteristics affect SDMs performance, we produced SDMs for virtual species with different responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), different levels of prevalence (very rare, rare, common), different sample sizes, unbiased and biased samples, and non-filtered and environmentally filtered datasets (Figure 4.1). Our specific objectives were to (i) determine the role of the sample size and species prevalence in SDMs, (ii) assess whether environmental filtering improves models based on biased samples, and (iii) evaluate the

effect of species response to environmental gradients used to generate virtual species on factors under study (i.e. sample size, sampling bias, environmental filtering, species prevalence).

4.2 Material and methods

Figure 4.1 illustrates the workflow used in this study. First, virtual species distributions were modeled for the study area, encompassing to the Iberian Peninsula. The species distributions were modeled using various responses to environmental gradients and various species prevalences. Subsets of species occurrences were subsequently extracted from the species presence/absence distributions, using various sample sizes, sampling patterns and with/without application of filtering. Finally, SDMs were produced using those different subsets and their performance was evaluated and compared.

4.2.1 Simulating ecological patterns with virtual species

Data derived from Worldclim (www.worldclim.org) database are often adopted in SDMs (e.g. Moudrý and Šímová 2013). To build virtual species distributions we used the same variables downloaded from Wordclim that were adopted in the study by Varela et al. (2014) who first presented the idea of environmental filtering. However, their study used virtual species generated with threshold approach for evaluation, which was recently criticized (Meynard and Kaplan 2012, Moudrý 2015). In our study, we used a probability approach (see Meynard and Kaplan 2012, 2013) to generate virtual species (see the next paragraph). Using the same variables and study area allowed us to directly compare our results with theirs. The adopted variables included the maximum temperature of the warmest month (Bio5), minimum temperature of

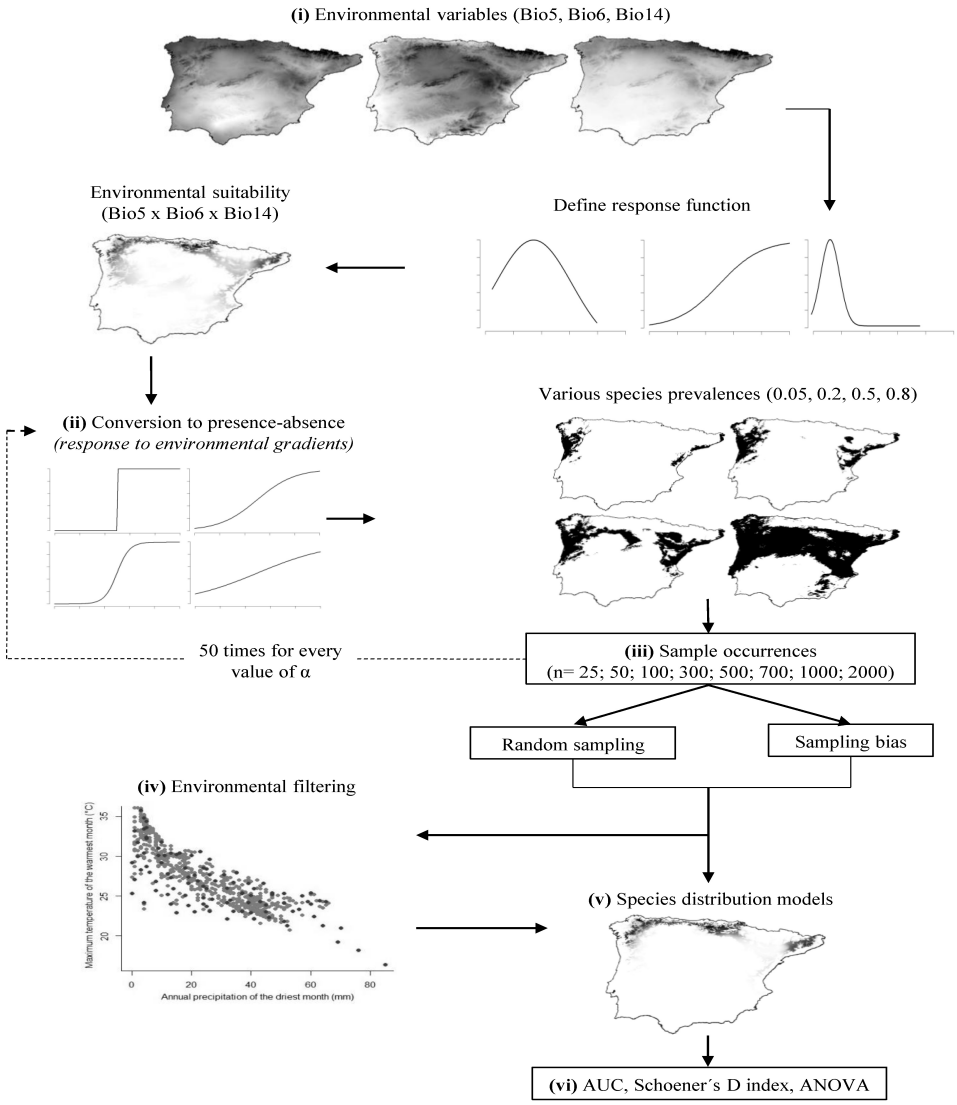

Figure 4.1 General modeling process. (i) Generating a map of probability occurrence for virtual species (environmental suitability map). (ii) Translating the probability of occurrence into a presence-absence map for various species prevalences. (iii) Sampling species occurrences randomly or with uneven sampling intensity and repeating the sampling 50 times for every a value and species prevalence. (iv) Applying environmental filter. (v) Creating models of species distribution with and without filtered occurrences. (vi) Quantifying SDMs performance using AUC and Schoener´s D index and performing ANOVA to statistically compare SDMs performance.

the coldest month (Bio6), and annual precipitation of the driest month (Bio14). Those were downloaded at a resolution of 30 arc seconds (approximately 1 km^2) from WorldClim and clipped to the extent of the Iberian Peninsula.

Various software and packages have been developed to facilitate the use of virtual species in SDMs (e.g. Duan et al. 2015, Leroy et al. 2016, Qiao et al. 2016). There are currently two main methods for generating virtual species: a threshold approach and a probability approach (Meynard and Kaplan 2012). The threshold approach generates occurrences where species always occur above a given threshold and never below it. On the other hand, the probability approach allows to generate species presences and absences along the whole environmental gradients (i.e. the shape of the logit function used to transform the occurrence probability to presences/absences). Moreover, the probability approach allows to take species prevalence into consideration. It is thus closer to ecological theories supporting the idea of dynamic occupancy patterns in space and time (see Hanski 1998, Meynard and Kaplan 2012, 2013). Therefore, the probability approach has been deemed more appropriate for generating virtual species than the threshold approach (Meynard and Kaplan 2012, 2013, Moudrý 2015). Virtual species distributions were created with the package *virtualspecies* (Leroy et al. 2016) in the statistical software R (version 3.4.4).

The virtual species were created in three steps (Leroy et al. 2016). First, we defined a relationship (i.e. the response function) between the artificial species and each variable, using a Gaussian distribution. Response functions were defined as follows (mean ± standard deviation): Bio5 (20 ± 10°C), Bio6 (10 ± 10°C), and Bio14 (20 ± 10 mm). The combination of these three response functions produced environmental suitability rasters for the Iberian Peninsula (function *generateSpFromFun*). Second, a probabilistic approach was used to convert environmental suitability rasters to binary presence-absence rasters (function *convertToPA*); a logistic function was applied to the

suitability rasters to model the response to environmental gradients. The logistic function had two parameters, α and β, where $1/\alpha$ corresponded to the slope of the curve at the inflection point and β to the position of the inflection point. Therefore, using α, one can control the steepness of the species response to the environmental gradients, and with a given value of α, the species prevalence can be controlled by β (see Figure 4.2).

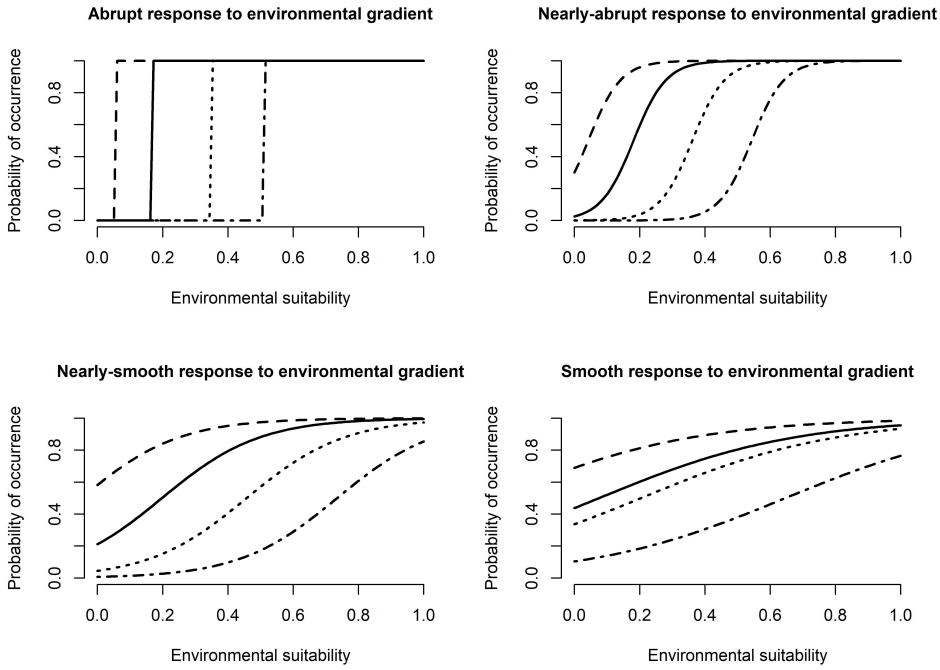

Figure 4.2 Contrasting examples of conversion curves for all species responses to environmental gradients and species prevalences (dotdash line = 0.05, dotted line = 0.2, solid line = 0.5, dashed line = 0.8).

We modeled four species types with respect to their response to environmental gradients: species with an abrupt response ($\alpha = -0.000001$), species with a nearly abrupt response ($\alpha = -0.05$), species with a nearly smooth response ($\alpha = -0.15$), and species with a smooth response ($\alpha = -0.3$). In addition, to evaluate the effect of prevalence for each type of response, four different species prevalence values were produced (0.05,

0.2, 0.5 and 0.8) by varying the parameter β (Figure 4.2). We generated species occurrences 50 times for each combination of α and β to produce multiple replications. For each replication, a valid estimation of the true species distribution was provided (Leroy et al. 2016). This approach contrasts with the threshold approach, which always generates the same distribution (presences/absences).

The last step consisted of sampling occurrences of virtual species from the modeled distributions using the *sampleOccurrences* function of the package.To test for sampling bias, two different sampling methods (survey designs) were used to generate presenceonly data: (i) random sampling across the entire area and (ii) a scheme that extracted samples 40 more times in the 50 largest protected areas of the Iberian Peninsula; this method has been used before to study sampling bias (e.g. Tessarolo et al. 2014, Varela et al. 2014). The protected areas of the Iberian Peninsula were downloaded from Protected Planet (www.protectedplanet.net). Eight different sample sizes were used to test sample size effects on SDMs (n = 25, 50, 100, 300, 500, 700, 1000, 2000).

4.2.2 Environmental filtering of sampled occurrences

We used the *gridSample* function of the *dismo package* to filter the sampled presence–only data (Hijmans et al. 2012). Based on a pre-defined grid, the function allows to eliminate repeated occurrences under similar environmental conditions. The environmental filters were defined using only two of the three environmental variables – the maximum temperature of the warmest month and the precipitation of the driest month. This enabled simulating a situation in which some of the environmental characteristics affecting species distribution were unknown, which is often the case in SDMs (Varela et al. 2014). The resulting

filters were applied to all versions of the generated virtual species (i.e. combination of response to environmental gradients and prevalence; 12 species), to the eight sample sizes, and to both survey designs (random and spatially biased). This totaled 384 unique combinations of model parameters, enabling an indepth comparison of the effects of the five different characteristics of SDMs under study. For each one of those combinations, 50 models were computed using the previously described probabilistic approach; each of those 50 repetitions can be viewed as a different run of the same stochastic process. A total of 19,200 different virtual species distributions were thus generated to be tested in SDMs.

4.2.3 Species distribution models

While there is currently no consensus on which SDM technique is best, it is widely recognized that every single technique has benefits and drawbacks (Elith et al. 2006, Elith and Graham 2009, Fernandes et al. 2019). For the purpose of this study, we needed a technique that could be kept consistent across the methodology to allow the comparison of outcomes. We selected the maximum entropy approach (MaxEnt), which is often adopted in ecological studies as a presence-only modeling technique, due to its good performance when compared to other techniques (Elith et al. 2006, Phillips et al. 2006). SDMs were built in R using the *dismo* package and the same three environmental variables that were used to generate virtual species distributions. To enable comparison of the different SDMs produced, we needed to maintain the parameters of the modeling technique unchanged. Although using MaxEnt with default settings is usually not recommended as it can overfit the models, it is not an issue when using virtually generated species as virtually generated data fit the pre-defined response perfectly and the risk of overfitting therefore is very low (we also employed hinge and linear feature classes and got the same results as with the default settings). Therefore, like many others before (e.g. Phillips et al. 2009, Beltrán et al. 2014, Ficetola et al. 2014, Fourcade et al. 2014, Franklin

et al. 2014, Varela et al. 2014, Beaumont et al. 2016, Holloway et al. 2016, Ranc et al. 2016, Tingley et al. 2018, Ye et al. 2018), we produced the models using the default settings, except for background points.

Since using background points that do not have the same bias as species occurrences (e.g. using random background points when species occurrences are spatially biased) has been shown to negatively affect SDMs performance (Phillips et al. 2009, Leroy et al. 2018), we did not use randomly generated background points. Instead, based on the artificially generated binary map of the virtual species illustrating true occupied and unoccupied areas, we generated a set of background points (i) across the entire area for randomly drawn species presence data (simulating unbiased dataset) and (ii) with higher sampling intensity in the 50 largest protected areas of the Iberian Peninsula (simulating biased dataset) and use those as background points. We used two times more background points than species occurrences as recommended by Liu et al. (2019). For each model replication, a new set of background points was generated. Similarly, as Thibaud et al. (2014), we had absence data available. Therefore, we generated background points in locations where species were absent and used Maxent in a nonstandard manner. Hence, the models can be viewed as presence-absence, allowing us to use the area under the receiver operating characteristic curve (AUC) as an appropriate measure for model performance. To evaluate the models, a fivefold cross-validation was used where the data were randomly divided into fifths. Four-fifths of the data were used to train the model and the remaining one fifth was used to quantify the performance.

4.2.4 Assessment of model performance

The AUC was calculated to quantify model performance. AUC indicates model performance based on predictions of presences/absences (Fielding and Bell 1997) and varies between 0 and 1 where values 0.9–1

indicate excellent models. In addition, we calculated Schoener's D index (Schoener 1968) to compare modeled probabilities of occurrence. Schoener's D is considered one of the best measures of evaluation of SDMs outputs (Rödder and Engler 2011). This metric measures the absolute spatial conformity between continuous predictions of the species as,

$$D = 1 - \frac{1}{2} \sum_{ij} |Z_{1ij} - Z_{2ij}|$$

where z_{1ij} is entity 1 occupancy (virtual reality) and z_{2ij} is entity 2 occupancy (model prediction) (Renkonen 1938). It varies between 0 (no overlap/agreement) and 1 (complete overlap/agreement). An analysis of variance (ANOVA) was used to assess the individual and combined effects of species response to environmental gradients, species prevalence, sample size, sampling bias and environmental filtering on SDM's performance. We fitted separate ANOVA models for AUC and Schoener's D index as a response, including all possible interactions among all five factors in both models.

4.3 Results

The ANOVA including all possible interactions explained 96% of Schoener's D and 89% of AUC variability (variance of the models explained by used characteristics). All analyses were highly significant given the large number of iterations (n = 19,200). For both Schoener's D and AUC, most of their variability was explained by the species response to environmental gradients (from abrupt to smooth), species prevalence, and sample size; these factors together (and disregarding their interactions) explained 53% of Schoener's D variability and 64% of AUC variability, with species prevalence being more influential for Schoener's D (17%) than for AUC (5%) (see Table 4.1).

Table 4.1 Degrees of freedom (Df), R^2 (%), and F statistics for ANOVA of Schoener's D index and AUC performance metrics.

	Schoener's D			**AUC**		
	Df	R^2	F	Df	R^2	F
Main effects:						
Sampling method	1	0.2	1127	1	0.2	334.3
Species (Spec)	3	19.9	32479.2	3	30.8	17153.6
Species prevalence (Prev)	2	17.0	41424.6	2	5.1	4227.1
Sample size (Sample)	7	16.3	11353.3	7	28.1	6697.3
Filter application (Filter)	1	2.4	11474.9	1	1.6	2665.9
Pair-wise interactions:						
Samp : Spec	3	0.3	452.7	3	0.2	124.3
Samp : Prev	2	0.0	96.6	2	0.1	81.3
Spec : Prev	6	8.9	7206.9	6	2.5	702.6
Samp : Size	7	0.1	58.7	7	0.0	10.0
Spec : Size	21	3.6	835.1	21	2.2	170.8
Prev : Size	14	2.0	699.5	14	1.0	116.3
Samp : Filter	1	1.2	5859.4	1	0.0	31.6
Spec : Filter	3	5.7	9280.1	3	8.0	4472.8
Prev : Filter	2	5.1	12.448	2	1.3	1093.8
Size : Filter	7	1.4	993.1	7	1.4	335.1
Higher-order interactions:	-	12.0	-	-	6.3	-
Total:	-	96.2	-	-	88.8	-

The effect of sample size on SDMs was relatively constant across other factors' (e.g. species prevalence, species response to environmental gradients) levels (see generally low R^2 values for its interaction terms in Table 4.1). Results show an initial steep increase in performance with increasing sample size, generally stabilizing around 300 samples after which more samples do not necessarily result in better models (see Figures 4.3 and 4.4). The initial increase was considerably steeper for AUC metric than for Schoener's D. The three exceptions to this general pattern were (1) almost constant values of Schoener's D across sample sizes for species with smooth or nearly smooth response to environmental gradients and with higher prevalence, (2) decreasing Schoener's D for

abrupt and nearly abrupt species with species prevalence 0.05 and non-filtered models (Figure 4.3) and (3) no stabilization for AUC values for species with nearly smooth or smooth response (Figure 4.4).

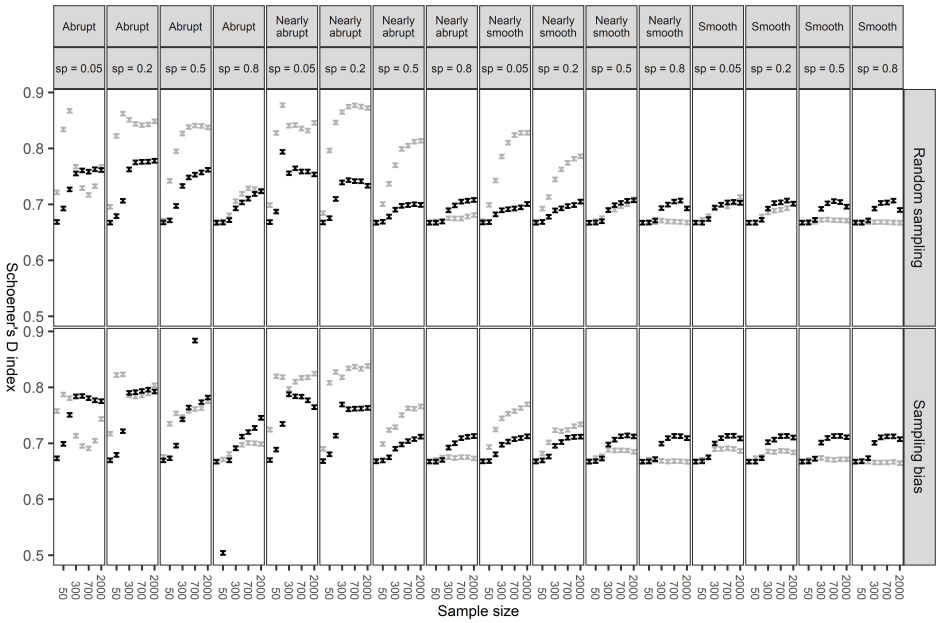

Figure 4.3 Resulting Schoener's D index values according to different species responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), species prevalence (sp = 0.05, 0.2, 0.5, 0.8), various methods of sampling occurrences (random, sampling bias) and different sample size (n = 25, 50, 100, 300, 500, 700, 1000, 2000). Gray color indicates results for non-filtered models, and the black color shows results for models where the environmental filter was applied.

The relatively low main-effect R^2 values of filter application term (see Table 4.1) resulted from a reverse effect of this factor in different species types (and also species prevalences in case of Schoener's D) (see Figures 4.3 and 4.4). Indeed, taking into account also its pair-wise interactions, filter application explained 16% of the Schoener's D and 12% of the AUC variability. For Schoener's D, these interactions can be summarized as follows (see Figure 4.3): the performance of the non-filtered SDMs was the highest for species with abrupt response to environmental gradients (about 0.85) and decreased to approximately 0.66 for those with smooth

response. On the contrary, the performance of filtered SDMs was much more stable, ranging from approx. 0.78 for species with abrupt response to 0.70 for those with smooth response.

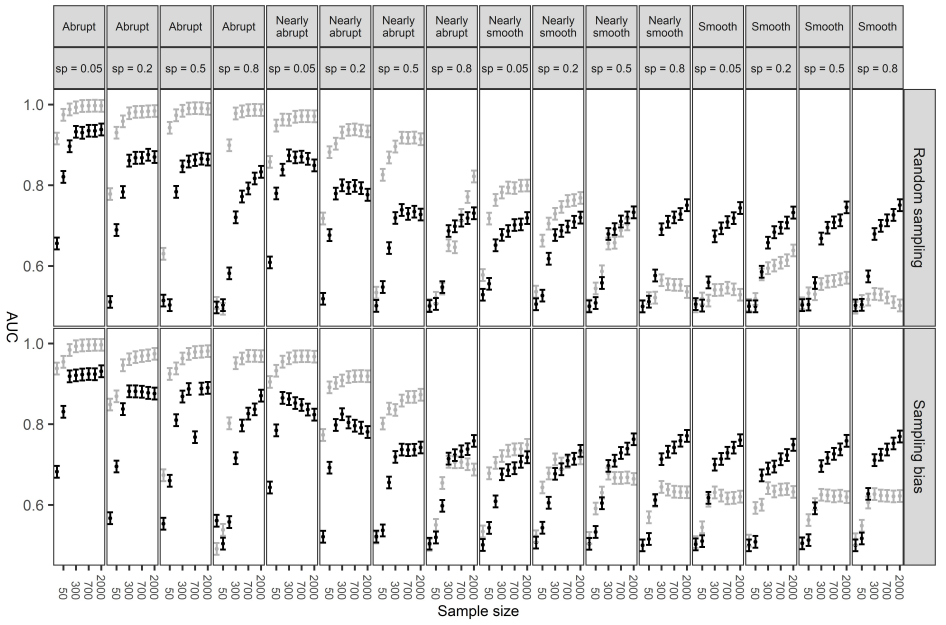

Figure 4.4 Resulting AUC values according to different species responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), species prevalence (sp = 0.05, 0.2, 0.5, 0.8), different methods of sampling occurrences (random, sampling bias) and different sample sizes (n = 25, 50, 100, 300, 500, 700, 1000, 2000). Gray color indicates results for non-filtered models, and the black color shows results for models where the environmental filter was applied.

This led to a significantly better performance of non-filtered SDMs for abruptly responding species but a slightly better performance of filtered SDMs for those responding smoothly. This relationship was further influenced by a significant decrease of non-filtered SDMs performance with increasing species prevalence, which was more striking for species with abrupt or nearly abrupt response. The exception from this general pattern was models with sample size higher or equal 100 for abrupt and nearly abrupt species and species prevalence 0.05. In this case, filtered models achieved better results than non-filtered models. Moreover,their

resulting Schoener's D was even lower in comparison to smoothly or nearly smoothly responding species. For AUC metric (Figure 4.4), the pattern was similar, with generally larger performance ranges for non-filtered SDMs (from almost 1.0 to approximately 0.5 for random sampling), which led to larger differences between non-filtered and filtered SDMs for species with smooth response. Sampling method (random vs. biased sampling) showed the least importance, both as main effect and in interactions with other effects (maximum R^2 being 1.2% but typically from 0.0% to 0.3%; see Table 4.1).

4.4 Discussion

Our results show that species prevalence and sample size have an equivalent effect on variability in model performance when using the MaxEnt modeling technique. Models performance increased with sample size (often up to a certain level), and where the sample size was constant, the model performance decreased with increasing prevalence. Moreover, our results show that both effects are independent of sampling bias. As opposed to what is often done in other studies, here we were changing the steepness of the response to environmental gradient (i.e. logistic curve) to create virtual species from abrupt (i.e. similar to what would be done with the threshold approach) to very smooth (see Figure 4.3). Generally, the more abrupt the response of species to the environmental gradient was, the greater the effect of species prevalence, sample size, and environmental filtering was. Since both measures (AUC and Schoener´s D) showed similar trends, the following discussion is based mostly on the behavior of the Schoener ´s D. We highlight the differences in AUC behavior where necessary.

4.4.1 Sample size

It has been shown many times that the performance of SDMs depends on sample size (see review by Moudrý and Šímová 2012). Prior studies examined sample sizes that varied from a few occurrences up to thousands of occurrences. While Guisan et al. (2007) or later Proosdij et al. (2016) have shown that a few occurrences may suffice to produce reliable models, other studies argued that it is best to use larger sample sizes (Pearson et al. 2007, Wisz et al. 2008, Tessarolo et al. 2014). Such opposing suggestions can be explained by the differences in data characteristics and model selection in these studies. One reason can be the complexity of species responses to environmental variables. It is clear that the more complex is the species response to environmental variables, the higher is the number of species occurrences required to achieve high model performance (e.g. Barry and Elith 2006). Studies using virtually generated data (e.g. Jiménez-Valverde et al. 2009, Varela et al. 2014, Proosdij et al. 2016) that have occurrences perfectly following the adopted response to environmental variables (e.g. Gaussian) suggested that reliable models can be developed with very small sample size (10 or even 5 samples). In contrast, studies with real species occurrence data (Wisz et al. 2008, Tessarolo et al. 2014) suggested the opposite. In addition, the effect of sample size could be also affected by species prevalence. Proosdij et al. (2016) concluded that increasing species prevalence decreases the influence of sample size. It has also been shown that some modeling techniques are less sensitive to sample size than others (Guisan et al. 2007, Tessarolo et al. 2014). Besides, Wisz et al. (2008) also show that the influence of sample size is changing across different spatial extents and resolutions of environmental variables (sites with resolution of 100 x 100 m performed better in comparison with those with resolution of 1000 × 1000 m). Our results show that a larger sample size has a significant positive effect on SDMs performance, although with a threshold after which

more samples do not necessarily improve performance. In our case, that threshold was usually at 300 or 500 samples. This effect, however, was only consistent when measured by AUC, which was expected due to the sensitivity of AUC to the ratio of sample prevalence and species prevalence (see Meynard and Kaplan 2012). Our results are, moreover, in accordance with prior studies by Thibaud et al. (2014) and Fernandes et al. (2018) who also tested the impact of various factors affecting SDMs using virtual species and concluded that sample size is one of the most important factors. For Schoener's D, the effect of sample size considerably varied with species response to the environmental gradients, species prevalence and the use of environmental filtering.

4.4.2 Species prevalence

Our results show that species prevalence is one of the most important factors affecting SDMs, having generally a negative effect on both model performance metrics (i.e. model performance was generally decreasing with increasing species prevalence). While this negative effect has been observed by a number of previous studies looking at AUC (e.g. Manel et al. 2001, Allouche et al. 2006, Lobo and Tognelli 2011, Meynard and Kaplan 2012, Syfert et al. 2013), it is to be noted that Proosdij et al. (2016) have found an opposite trend for Schoener's D. However, the authors did not provide any explanation or hypothesis for that trend, making its comparison with our study difficult. A potential explanation for that difference could be that their sample sizes (5 to 50 occurrences) were much smaller than the ones used in the current study (25 to 2000). In addition, our results show that this negative effect only applies to species with abrupt or nearly abrupt response to environmental gradients, a factor that was not specified in Proosdij et al. (2016).

4.4.3 Sampling bias

Sampling bias, caused by uneven sampling of species occurrences, is often considered as one of the major factors that have a negative impact on SDMs (e.g. Araújo and Guisan 2006, Leitão et al. 2011, Duputié et al. 2014, Guillera-Arroita et al. 2015). Prior studies have demonstrated that the presence of sampling bias decreases model performance (e.g. Loiselle et al. 2008, Leitão et al. 2011, Sánchez-Fernández et al. 2011, Fourcade et al. 2014, Ranc et al. 2016), While our results agree with that, they show that the contribution of sampling bias to the overall, combined effects of the different studied factors on SDMs is relatively low, explaining no more than 2% of the variability of the performance metrics. This demonstrates the importance of simultaneously studying multiple factors and their impacts on SDMs, whereas other studies focused solely on the effect of sampling bias and did not provide measures of explained variability, our study compared its effect with the effect of other factors, finding it statistically significant but relatively negligible. Our results are in agreement with the study by Tessarolo et al. (2014) who also concluded that sampling bias has rather minor effects on model performance compared to other factors (species characteristics, sampling method, sample size, SDMs technique). Interestingly, they used the same study area as our study (i.e. Iberian Peninsula), the difference lied in the use of 34 real species (amphibians, reptiles, mammals). Nevertheless, the effect of sampling bias may be related to autocorrelation in the predictor variables, which is relatively high in interpolated climate data (such as Worldclim used in both their and our study).

4.4.4 Environmental filtering

Another goal of our study was to test the applicability of environmental filtering on models generated with spatially biased data. According

to Varela et al. (2014), environmental filtering consistently improves model performance. Our results show that the measured effect of environmental filtering was significant, they however also showed that that effect was relatively unimportant when compared to other factors (see Table 4.1). Moreover, its positive or negative effect strongly depended on the type of species response to environmental gradients, species prevalence, and sample size. We only confirmed the positive effect for species with smooth or nearly smooth response, whereas for species with abrupt or nearly abrupt response the effect was negative (except models with species prevalence 0.05). This contradicts the results of Varela et al. (2014) as their positive effect was observed for species generated using a threshold approach (i.e. the equivalent of our abrupt-responding species). In addition, the positive effect was much stronger when assessed by AUC (up to more than 20% increase, see Figure 4.4) than by Schoener's D (only approx. 5% increase, see Figure 4.3). This is in accordance with previous concerns about using AUC as the only model performance measure (Jiménez-Valverde 2012, Moudrý 2015, Fernandes et al. 2019).

We recognize that SDMs may be affected by many other factors (see Thibaud et al. 2014, Fernandes et al. 2019). Thus, we recommend that further studies focus on interactions of environmental filtering with other factors, such as the effects of spatial scale (extent and resolution) (e.g. Connor et al. 2018, Šímová et al. 2019), spatial autocorrelation (Thibaud et al. 2014) or modeling technique (Fernandes et al. 2018).

4.5 Conclusions

We focused on several factors related to species occurrences (response variable) in SDMs (i.e. environmental filtering, sampling bias, sample size, species prevalence and species response to environmental gradient). We found that both sample size and species prevalence equivalently

affect performance (measured by AUC and Schoener´s D) of SDMs (in general, increasing sample size positively, increasing species prevalence negatively). Our results also highlighted the importance of using a probability approach to the generation of virtual species distribution, which allowed us to model species with different response to environmental gradient from abrupt to smooth, as opposed to a threshold approach, which is still commonly used. Indeed, our results showed that the response of a species to environmental gradients has a strong effect not only on the model performance itself but also on the effects of other factors. The unprecedented complexity of our study enabled us to recognize the importance not only of each of the factors themselves but also of their interactions. Ignoring such interactions, which is almost inevitable in studies focusing on one or two factors only, may lead to substantially misleading conclusions.

Our results suggest that environmental filtering is not always a good idea and should not be performed blindly without evidence of bias in species occurrences. Environmental filtering down-weights repeated observations of the same environmental conditions and reduces sample size. Therefore, sampling must be dense enough to characterize the curve and the algorithms must be able to uncover the true form of the relationship. Our results show that at least 300 presences are necessary for accurate predictions when using presence-only models fitted by MaxEnt. We suggest that models using original, unfiltered data should be always fitted. We highlight that the more gradual is the species response to environmental gradients (except species with prevalence 0.05), the greater is the model sensitivity to inappropriate use of environmental filtering, although the sensitivity decreases with higher species prevalence. Finally, we advocate that additional data and species characteristics (e.g. resolution, extent, positional error) should be evaluated using more complex virtual species (e.g. with more complex response curves) to improve SDM use in biodiversity monitoring and conservation.

Chapter 5

The effect of positional error on fine scale species distribution models increases for specialist species

Lukáš Gábor, Vítězslav Moudrý, Vincent Lecours, Marco Malavasi, Vojtěch Barták, Michal Fogl, Petra Šímová, Duccio Rocchini, Tomáš Václavík

Ecography (2020)

Publication metrics:
12 of 165 (Q1) rank in WOS category Ecology
IF (2018) 5.946; AIS (2018) 2.414

Author's contribution: 60%

Abstract

Species occurrences inherently include positional error. Such error can be problematic for species distribution models (SDMs), especially those based on fine-resolution environmental data. It has been suggested that there could be a link between the influence of positional error and the width of the species ecological niche. Although positional errors in species occurrence data may imply serious limitations, especially for modelling species with narrow ecological niche, it has never been thoroughly explored. We used a virtual species approach to assess the effects of the positional error on fine-scale SDMs for species with environmental niches of different widths. We simulated three virtual species with varying niche breadth, from specialist to generalist. The true distribution of these virtual species was then altered by introducing different levels of positional error (from 5 to 500 m). We built generalized linear models and MaxEnt models using the distribution of the three virtual species (unaltered and altered) and a combination of environmental data at 5 m resolution. The models' performance and niche overlap were compared to assess the effect of positional error with varying niche breadth in the geographical and environmental space. The positional error negatively impacted performance and niche overlap metrics. The amplitude of the influence of positional error depended on the species niche, with models for specialist species being more affected than those for generalist species. The positional error had the same effect on both modelling techniques. Finally, increasing sample size did not mitigate the negative influence of positional error. We showed that fine-scale SDMs are considerably affected by positional error, even when such error is low. Therefore, where new surveys are undertaken, we recommend paying attention to data collection techniques to minimize the positional error in occurrence data and thus to avoid its negative effect on SDMs, especially when studying specialist species.

Keywords: Data errors, Niche breadth, Spatial overlay, Virtual species

5.1 Introduction

Studying relationships between species and their environment is fundamental for understanding Earth's biodiversity. Species distribution models (SDMs) are a common tool used to study these relationships. They use species occurrence data and environmental data to produce a set of rules explaining the environmental space where species were collected or observed (Ferrier et al. 2017). All applications of SDMs, however, assume that species occurrence data are largely free of spatial error. Nonetheless, all spatial data inherently contain some level and type of spatial errors. These errors can be, for example, related to the use of inadequate spatial resolution (Gottschalk et al. 2011, Šímová et al. 2019), low sample size (Wisz et al. 2008, Moudrý et al. 2017), biased sampling (Hijmans 2012, Ranc et al. 2016) or occurrences with positional error (Graham et al. 2008, Osborne and Leitão 2009, Mitchell et al. 2017). Data quality (both for species occurrences and environmental variables) is currently considered a major factor limiting SDM accuracy (Araújo et al. 2019) and demonstrating, quantifying and understanding the consequences of these errors is therefore critical.

It is often assumed that the negative effects of positional error (i.e. inaccurate location of species occurrences) is minimal or mainly associated with relatively older datasets that are often georeferenced from textual descriptions of their locations (which may cause errors of up to hundreds of meters, Wieczorek et al. 2004). However, it is also necessary to consider positional errors inherent to data georeferenced using modern global navigation satellite systems (GNSS). The positional error of GNSS data may be caused by the use of outdated technology, by poor satellite signal reception (e.g. because of inappropriate site conditions), or by data processing (e.g. conversion between coordinate systems or rounding of coordinate values). Moreover, species occurrence data often represent the position of the observer and not the actual position of the species (Zhang et al. 2018). Additionally, where the marine environment

is concerned, species data are often acquired using underwater cameras, in which case the positional error can be affected for example by the camera depth; the deeper the camera is, the greater is the positional error (Rattray et al. 2014, Mitchell et al. 2017). Therefore, even though the accuracy of standard GNSS is usually below 30 m (Frair et al. 2010), the errors associated with such data may be much larger.

In addition, performance of SDMs is complicated by various spatial (e.g. prevalence or range size) and ecological (e.g. niche breadth) characteristics of the studied species (Luoto et al. 2005, Bulluck et al. 2006, McPherson and Jetz 2007, Evangelista et al. 2008, Chefaoui et al. 2011, Connor et al. 2018). It has been hypothesized that range size is positively correlated with niche breadth (i.e. the range of environments that the species can inhabit), in other words that species able to tolerate a wider range of conditions are typically more widespread (Brown 1984, Gaston et al. 1997, Arribas et al. 2012, Boulangeat et al. 2012). The niche breadth–range size relationship is one of the possible mechanisms explaining commonness and rarity. Modelling rare species (i.e. species with small geographical ranges) is particularly problematic and novel approaches have been adopted for this purpose (Breiner et al. 2015) to overcome the common problem of a low number of occurrences available for modelling that may not be sufficient to completely describe the species niche. Similar effects can be caused by a low positional accuracy of the occurrences (Johnson and Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009).

Although the magnitude of the niche breadth–range size relationship is still under debate, a recent meta-analysis of 64 studies found a significant positive relationship between the range size and niche breadth (Slatyer et al. 2013). Such a synergic relationship can increase the already high vulnerability of specialist species to environmental changes. In addition, Slatyer et al. (2013) suggested that specialist species might be particularly vulnerable to any environmental change due to synergistic effects of a narrow niche and small range size. Specialist species are

of high conservation concern, and SDMs might be the only tractable means of estimating their distribution and reaction to environmental change. However, confounding effects of inaccurate data on modelling species that utilize a narrow niche breadth (i.e. specialist) versus species that utilize a wide niche breadth (i.e. generalist) are unknown (Connor et al. 2018).

It is intuitive that positional error of a given magnitude might have a greater effect on specialist than generalist species, as it is more likely that occurrences get incorrectly shifted into cells representing an unsuitable environment, i.e. environment that is outside of the species' environmental niche. This, however, has never been thoroughly explored because it is extremely difficult, if not impossible, to estimate the true responses of a real species to the environment and, consequently, to be able to fully understand the true suitability of an area for the species in question.

In this study, we focused on Light Detection and Ranging (LiDAR)-derived variables that are being more and more often combined with species distribution data of unknown positional accuracy to study species–environment relationships at fine scales. Studies published so far have used real species to test the effect of positional error. However, real species distribution data are usually affected by a complex set of other uncertainties (e.g. sampling bias, incompleteness, inaccuracies). As a consequence, the isolation and identification of the effects of positional error can be very challenging, if not impossible. This is likely one of the reasons why little consensus exists on how the effect of positional error manifests in SDMs (Naimi et al. 2011, Mitchell et al. 2017). For example, Graham et al. (2008) concluded that SDMs are robust to positional error while others argued that positional errors reduce models' performance (Johnson and Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009).

Another aspect may be that positional errors of species occurrences were studied using relatively coarse environmental data (but see Mitchell et al. 2017). Positional error considered in prior studies ranged from 50 m up to 50 km (Table 5.1). While such error results in a shift over several cells in a coarse-resolution SDM (e.g. 1 × 1 km), it will cause a much greater shift in a fine-resolution SDM (e.g. 10 × 10 m). Therefore, with the increasing availability of fine-scale data, additional studies are needed (Osborne and Leitão 2009); it can be expected that SDMs at fine scales would be more sensitive to positional error.

Table 5.1 Overview of prior studies focused on the influence of positional error in species occurrence data on SDMs.

	Species data	**Environmental data**
Graham et al. 2008	observed	categorical,continuous
Johnson and Gillingham 2008	observed	categorical
Fernandez et al. 2009	observed	continuous
Naimi et al. 2011	artificial	continuous
Mitchell et al. 2017	observed	continuous
	Range of shifting occurrences	
Graham et al. 2008	0–5 km	0–50 pixels
Johnson and Gillingham 2008	50–1000 m (over 50 m)	1–34 pixels
Fernandez et al. 2009	5–10–25–50 km	1–5, 1–10, 1–25, 1–50 pixels
Naimi et al. 2011	x	1–30 (over 1 pixel)
Mitchell et al. 2017	5–25–50–20–400 m	1–2, 1–12, 1–80, 1–160 pixels
	Resolution of input environmental data (pixel size)	
Graham et al. 2008	100 × 100 m	
Johnson and Gillingham 2008	30 × 30 m	
Fernandez et al. 2009	1 × 1 km	
Naimi et al. 2011	artificial data	
Mitchell et al. 2017	2.5 × 2.5 m	

To ensure the full knowledge of the exact ecological and geographical characteristics of the species and to avoid unknown complexities associated with real data, we used a virtual species approach to test the effect of the positional error in species occurrences on fine-scale SDMs in the context of species niche breadth (i.e. specialist versus generalist species). We generated three virtual species that differed in characteristics related to the geographic distribution of the species, i.e. prevalence and relative occurrence area (ROA); the proportion of the

total study area occupied by the species (Lobo 2008).

The virtual species approach allowed us to control the experiment and to isolate the effects of positional error (Zurell et al. 2010). This approach is increasingly used to evaluate the effects of data inaccuracies on model performance (Barbet-Massin et al. 2012, Václavík and Meentemeyer 2012, Qiao et al. 2015, Ranc et al. 2016, Fernandes et al. 2018, Leroy et al. 2018, Moudrý et al. 2018, Gábor et al. 2019, Meynard et al. 2019), but has yet to be adopted for the study of positional error. In particular, we tested whether: 1) SDMs for specialist species are more affected by positional error than those for generalist species; 2) it is possible to compensate the assumed negative effect of a positional error with a higher sample size; and 3) the positional error has different effects when using a parametric (e.g. generalized linear model) versus a nonparametric (e.g. MaxEnt) modelling technique.

5.2 Material and Methods

5.2.1 LiDAR data acquisition, processing and variable selection

Discrete LiDAR data were collected in Krkonose Mountains National Park (KRNAP), Czech Republic (Supplementary material Appendix 1 Fig. A1) in 2012 using a small-footprint airborne LiDAR system (RIEGL LMS Q-680i). The average point density was approximately six points per square meter. The LiDAR point cloud was automatically classified into ground, vegetation, building, wire and transmission tower classes in the ENVI LiDAR software (ver. 5.3) and LAStools (ver. 171215). The terrain data points were used to produce a digital terrain model (DTM), and the vegetation data points were used to produce a canopy height model (CHM) (Khosravipour et al. 2016). Both models were generated from the point cloud at a 0.5 m resolution

and subsequently resampled to 5 m cell resolution for the analysis to improve processing time. A topographic wetness index (TWI) was derived from the DTM based on the equation

$$TWI = \ln(\frac{As}{tan\beta})$$

where As is the specific catchment area and tan β is the local slope in radians (Beven and Kirkby 1979). To calculate the specific catchment area, we used the multiple flow routing algorithm of Quinn et al. (1991), recommended by Kopecký and Čížková (2010), using SAGA-GIS (Conrad 2003).

The selection of these three variables (DTM, CHM, TWI) was motivated by the need to simulate a realistic situation that includes variables with various levels of spatial autocorrelation (Supplementary material Appendix 2 Fig. A2). CHM describes a horizontal structural variability of the vegetation and is known to affect species richness (Lefsky et al. 2002). For example, higher vegetation was found to be related to higher bird species richness (Davies and Asner 2014). TWI is a surrogate for soil moisture, an environmental variable that affects the vegetation composition and that has been previously used to predict bird occurrences (Besnard et al. 2013, Reif et al. 2018). The relationships between CHM and TWI on the one side and bird distribution and richness on the other side make our study relatable to applications with real species; our virtual species could theoretically be birds with specific habitat requirements in terms of terrain characteristic and vegetation structure. We also used the DTM as a surrogate for climatic variables and to restrict our virtual species to certain altitudes (Coops et al. 2010, Vogeler et al. 2014).

5.2.2 Simulating virtual species with different niche breadths

Virtual species were generated with the *virtualspecies* package (Leroy et al. 2016) in the statistical software R v.3.4.4 (R Development Core Team). The process involved three steps: a) generating the true distribution of the virtual species' environmental suitability, b) converting the environmental suitability into presences and absences and c) sampling species occurrences for further analysis and modelling.

Applying the *formatFunctions* function in R, we defined the species–environment relationships using normal distribution curves. To simulate species with different niche breadth, prevalence and ROA, we used the same means and varied standard deviations of the used environmental variables (Supplementary material Appendix 3 Table B1). Specifically, we simulated three distinct virtual species with varying ROAs and prevalence that represent realistic scenarios of species' extent of occurrence in the study area. The species with low ROA (4%) represents a specialist with low species prevalence (0.04), narrow niche breadth and small geographical range. The species with medium ROA (12%) may be described as an intermediate species (species prevalence = 0.12) with a wider niche breadth and medium geographical range. Finally, the species with high ROA (52%) can be perceived as a generalist with high species prevalence (0.47), wide niche breadth and wide geographical range (Futuyma and Moreno 1988, Devictor et al. 2010, Franklin 2010, Peers et al. 2012). Subsequently, we multiplied individual species' responses to environmental variables in order to acquire an environmental suitability raster (function *generateSpFromFun*). We opted for multiplication of the variables to assume irreplaceability of environmental conditions (i.e. we assumed that unsuitability of one condition causes a low probability of occurrence even though remaining conditions are in species' range of suitable values).

As noted in several studies (Meynard and Kaplan 2012, 2013, Moudrý

2015, Meynard et al. 2019), an appropriate setting of the whole simulation with respect to the research questions is crucial for obtaining reliable results. In addition, Meynard et al. (2019) highlighted that simulation studies based on the threshold approach fail in appropriately separating factors such as prevalence and niche breadth. Therefore, due to these concerns, we adopted a probabilistic simulation approach (logistic function with α = -0.05 and β = 0.3) to convert the environmental suitability rasters into probabilities of occurrences that were subsequently used to sample binary presence/absence rasters (function *convertToPA*). To sample species occurrences (function *sampleOccurrences*), we randomly generated, using a uniform random distribution, both presence-only and presence/absence data. Both types of occurrence datasets were generated in order to test different modelling techniques (cf. section Model fitting and evaluation). To test whether it is possible to compensate the assumed negative effect of positional error with a higher sample size, we generated four different sample sizes. Specifically, 30, 100, 500 and 1000 species presences were generated, complemented for the purpose of GLM modelling by twice as many absences.

5.2.3 Simulating positional error in species occurrences

It is generally assumed that the magnitude of the positional error in species occurrence varies based on the source of the error. The positional error associated with GNSS points (e.g. species occurrences) may range from a few centimetres up to several metres. Furthermore, in some species such as birds or big predators, it is usually impossible to record their accurate position and such data are shifted by tens or hundreds of meters. An even greater shift is sometimes observed in museum databases. Therefore, to evaluate the range of possible magnitudes of the positional error, we simulated the positional error by shifting the sampled locations (i.e. presences and, in case of GLM, also absences)

in a random direction according to six scenarios that corresponded to different distances ranging from 5–10 m up to 100–500 m. The error in the focal virtual species locations was 5–10 m for S1 scenario, 10–15 m for S2, 15–20 m for S3, 20–50 m for S4, 50–100 m for S5 and 100–500 m for S6 (Supplementary material Appendix 4 Table C1). Scenarios S1–S4 simulated realistic degrees of error if using modern monitoring technologies like GNSS, while scenarios S5–S6 simulated more extreme positional errors that could be associated with species observations recorded without GNSS, species difficult to pinpoint properly such as birds or big predators, or occurrences from museum databases. If the shifting of the original data points resulted in the points falling outside the study area, we recalculated the shift until the new coordinates were located within the boundaries of the study area. We provide a script of how we simulated virtual species and shifting occurrences in Supplementary material Appendix 2.

5.2.4 Model fitting and evaluation

We selected generalized linear models (GLM; Nelder and Baker 1972, Oksanen and Minchin 2002) as a presence/absence method and MaxEnt (Phillips et al. 2006) as a presence-background method that are often adopted in ecological studies (Moudrý and Šímová 2013, Linda et al. 2016, Malavasi et al. 2018, Gábor et al. 2020, Watts et al. 2019). In addition, Graham et al. (2008) showed that these two approaches were among the better performing modelling techniques when the data was affected by positional errors. Models were built in the statistical software R using the *dismo* (ver. 1.1.4) and *glm2* (ver. 1.2.1) packages. The GLM was run with a logit-link function and binomial distribution. The quadratic terms of the three environmental variables were included because of the known normal distribution curves of the response function. To enable the comparison of individual SDMs, we needed to maintain the parameters of MaxEnt unchanged, as done in many prior studies (Franklin et al. 2014, Fourcade et al. 2014, Holloway et al. 2016, Ranc

et al. 2016, Tingley et al. 2018, Ye et al. 2018). The default settings established by Phillips et al. (2009) were used with randomly drawn background data generated from the binary map of the true occurrences of the virtual species. The same three environmental variables (DTM, CHM and TWI) used in the process of generating virtual species were used in the SDMs. Fivefold cross-validation where the data were randomly divided into fifths was used to evaluate the models. Four fifths of the data were used to train the model and the remaining one fifth was used to assess the performance. Control models without positional error were calculated for all three species with different niche breadth, prevalence and ROA and for both modelling techniques, allowing an easy comparison of the effect of positional error on model performance.

The area under the receiver operating characteristic curve (AUC) (Fielding and Bell 1997, Jiménez-Valverde 2012) and the true-skill statistic (TSS; Allouche et al. 2006) were used to assess model performance (i.e. discrimination accuracy). AUC is widely used in ecological studies as a single threshold independent measure of model performance (Václavík and Meentemeyer 2012, Mitchell et al. 2017). The AUC ranges from 0 to 1 where a score of 1 indicates perfect discrimination, a score of 0.5 indicates random performance and values lower than 0.5 indicate a worse than random performance. TSS is a frequently used threshold dependent metric (Cianfrani et al. 2018, Eaton et al. 2018) taking both omission and commission errors into account. It ranges from -1 to +1 where +1 indicates perfect agreement and values of zero or less indicate random performance (Allouche et al. 2006).

To quantify differences between the true probability of occurrence of virtual species and the predicted distribution inferred from the models in geographical space, their niche overlap was compared using the I measure (Warren et al. 2008, Rödder and Engler 2011) and Spearman's rank correlation. The I ranges between 0 (no overlap) and 1 (perfect overlap). Following Rödder and Engler (2011), we used the following

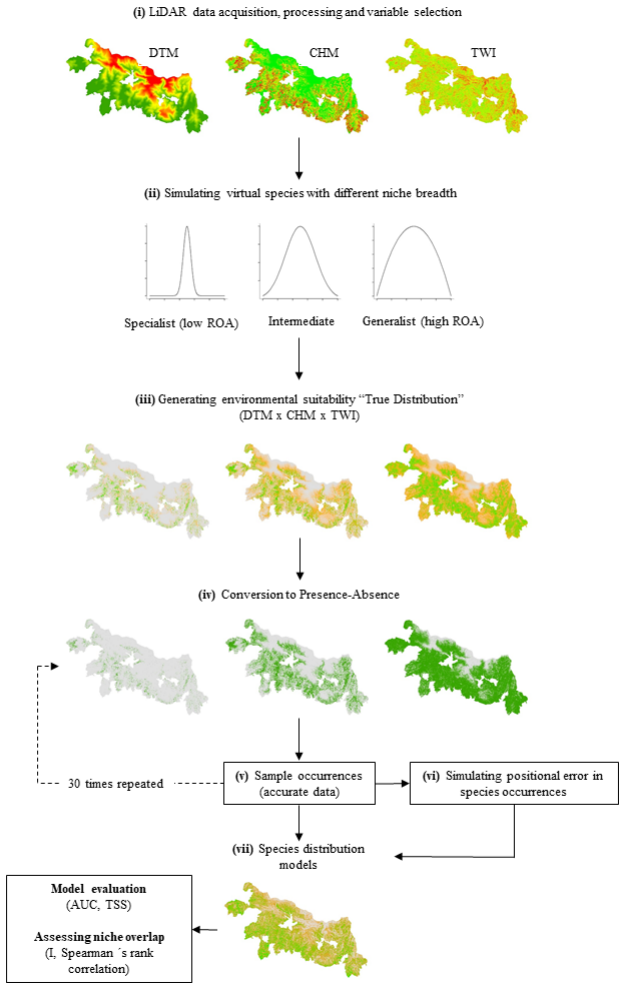

Figure 5.1 General modelling process. (i) We first acquired and processed LiDAR data and selected three fine-scale environmental predictors: DTM, CHM and TWI. (ii) We simulated virtual species with different niche breadths (ROA) by defining their response to environmental gradients for each environmental variable. (iii) We multiplied those variables to generate environmental suitability ('true' distribution of virtual species). (iv) We translated the probability of species occurrence to a presence–absence raster. (v) We sampled occurrences based on the presence–absence raster. (vi) We simulated the positional error in species occurrences. (vii) We generated SDMs with accurate as well as shifted occurrences, evaluated their performances (AUC, TSS) and assessed the niche overlap (I, Spearman´s rank correlation) in the geographical and environmental space.

classes to interpret the results: no or very limited overlap (0-0.2), low overlap (0.2-0.4), moderate overlap (0.4-0.6), high overlap (0.6-0.8) and very high overlap (0.8-1.0). Spearman's rank correlation ranges between -1 and +1, where -1 indicates that species responses to the environment are exactly negatively correlated (opposite) and +1 indicates perfectly positively correlated overlap (identical). The closer the values are to zero, the lower is the niche overlap.

The magnitude of the negative effect of the positional error on SDMs is dependent on the size of the positional error and distribution of species' suitable environment in the geographical space (Naimi et al. 2011). The positional data may be shifted in the geographical space and even a relatively low positional error in geographical space can have a profound effect on environmental niche estimates in environmental space and vice versa. Furthermore, we expected this would be related to the species niche breadth. Therefore, we were also interested in how the positional error is manifested in the environmental space and measured the niche overlap in the environmental space as well. We used I and Spearman's rank correlation implemented in *ENMTools* 0.2 (Warren 2019, Warren et al. 2020) to estimate overlap in the environmental space between models fitted with accurate occurrences without any positional error (hereafter unaltered models) and models fitted with shifted occurrences (i.e. scenarios S1–S6)

We ran the entire process from species generation to model evaluation 30 times (Figure 5.1). In addition, we used the analysis of variance (ANOVA) to assess the strength of the individual effects of the positional error, sample size, ROA and modelling technique, including all possible interactions. We compared the relative importance of individual predictors based on their contribution to the overall explained variation (R^2). Instead of formal testing, we plotted the effects (and their confidence intervals) of all predictors combinations and evaluated them qualitatively. Because both AUC and TSS values were highly heteroscedastic (e.g. the ratio between maximum and minimum standard

deviation across all factors combinations was 22 resp. 19 for AUC resp. TSS), we used robust variance–covariance matrix estimator suggested by Mackinnon and White (1985) for computation of confidence intervals. This was done using an R package *sandwich* (Zeileis 2006).

5.3 Results

5.3.1 Unaltered models

Both performance metrics (AUC and TSS) largely followed the same pattern and highlighted excellent model performance for all, i.e. specialist, intermediate and generalist, species (AUC ranged from 0.91 up to 0.97 for MaxEnt models and from 0.80 up to 0.85 for GLM models). The only exception were the MaxEnt models for generalist species where AUC achieved only good performance (mean AUC 0.73). MaxEnt models were more successful in modelling specialist and intermediate species while GLM models were more accurate for the generalist species (Figure 5.2)

Models achieved high or very high niche overlaps in geographical space according to both I and Spearman´s rank correlation. In general, the niche overlap decreased in the following order: generalist, specialists and intermediate species except for the Spearman´s rank correlation for specialists modelled by MaxEnt that achieved very high correlation. Comparison of modelling techniques showed that MaxEnt models achieved a higher niche overlap than GLM for all species with the most obvious differences in specialist species. An increase in the sample size of unaltered models led to none or negligible increase in niche overlap (Figure 5.3).

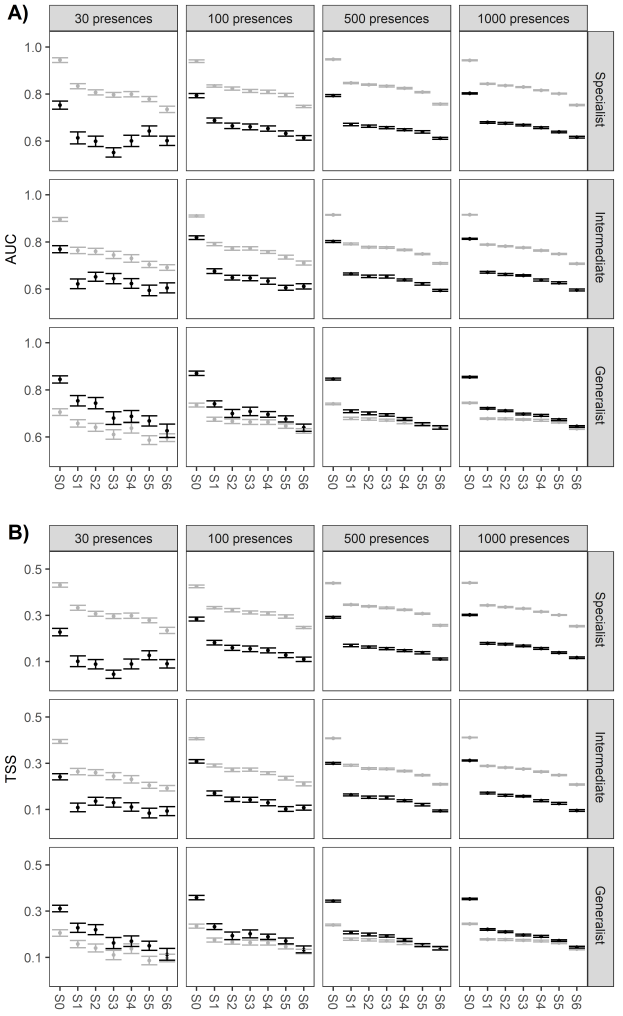

Figure 5.2 Resulting AUC (A) and TSS (B) scores according to different species niche breadth (specialist, intermediate, generalist), positional error (S0, unaltered models; S1, 5–10 m; S2, 10–15 m; S3, 15–20 m; S4 20–50 m, S5, 50–100 m; S6, 100–500 m) and sample size (number of presences = 30, 100, 500, 1000; note that for GLM models twice as many absences compared to presences were generated). Black colour shows results for GLM models while grey shows results for MaxEnt models.

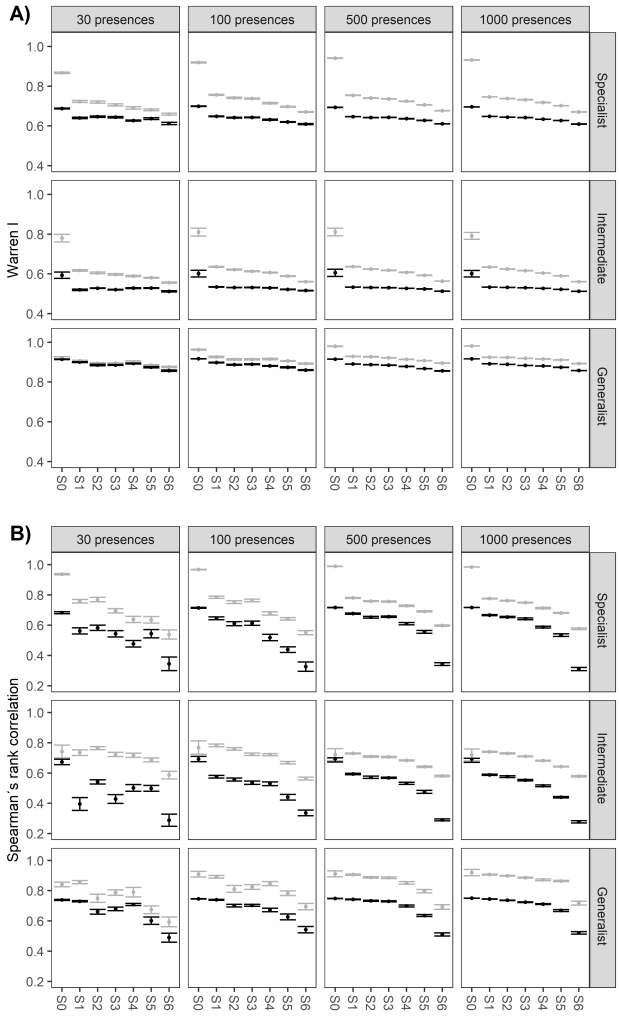

Figure 5.3 Resulting I (A) and Spearman's rank correlation (B) scores of niche overlap in geographical space according to different species niche breadth (specialist, intermediate, generalist), positional error (S0, unaltered models; S1, 5–10 m; S2, 10–15 m; S3, 15–20 m; S4, 20–50 m, S5, 50–100 m; S6, 100–500 m) and sample sizes (number of presences = 30, 100, 500, 1000; note that for GLM models twice as many absences compared to presences were generated). Black colour shows results for GLM models while grey shows results for MaxEnt models.

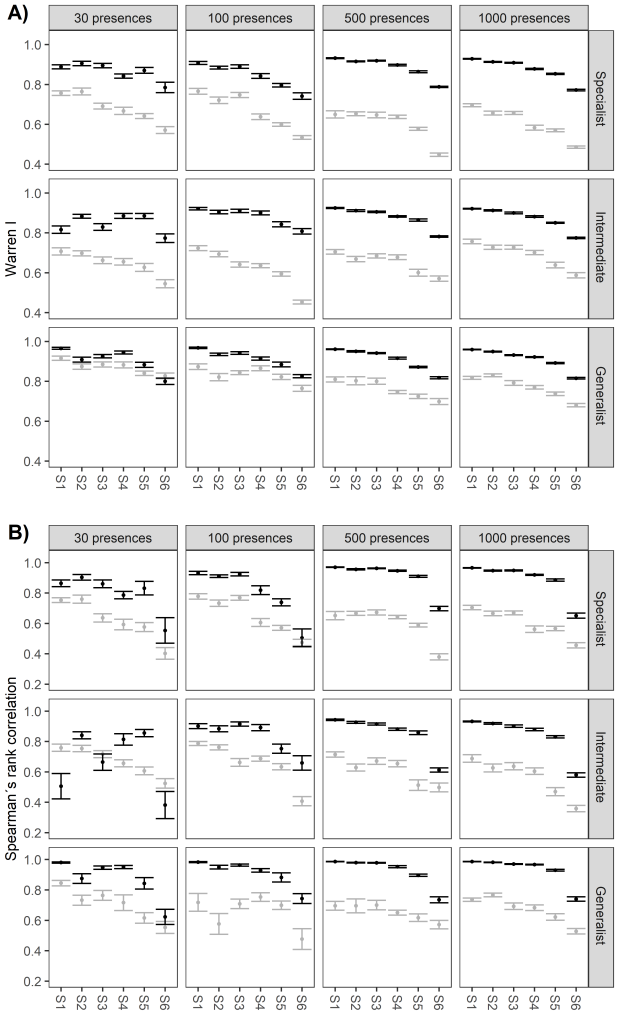

Figure 5.4 Resulting I (A) and Spearman's rank correlation (B) scores of niche overlap in the environmental space according to different species niche breadth (specialist, intermediate, generalist), positional error and sample size (number of presences = 30, 100, 500, 1000; note that for GLM models, twice as many absences as presences were generated). Also note that here we show the niche overlap between unaltered models and models affected by a specified positional error (and not a comparison with simulated probability of occurrences as in Figure 5.3). Thus, for example, S1 shows a comparison of niche overlap between unaltered models and models affected with positional error in the range of 5–10 m. Black colour shows results for GLM models while grey shows results for MaxEnt models.

5.3.2 Effect of positional error on models of species with different niche breadth

Results show, independently of the modelling technique, a clear trend of the positional error worsening model performance (both AUC and TSS). The highest drop is evident between unaltered models and models affected by the smallest simulated positional error (5–10 m).

Increasing the positional error further led to additional decrease in model performances; however, this decrease was minimal (positional error 10–50 m). Even the extreme cases of positional error (50–100 and 100–500 m) led to a relatively low decrease in models' performances in contrast to the drop caused by the 5–10 m error. For example, in the case of MaxEnt models for intermediate species, AUC dropped on average from 0.91 (unaltered models) to 0.79 for the positional error of magnitude inherent to any occurrence data (i.e. up to 10 m), and to 0.71 in the case of the extreme positional error (100–500 m), respectively (Figure 5.2). Nevertheless, the magnitude of the negative effect of positional error varied according to the species niche breadth. For both GLM and MaxEnt models the drop between unaltered models and the smallest simulated positional error (5–10 m) was higher for specialist and intermediate species (AUC dropped on average about 0.12) than for generalist species (AUC dropped on average about 0.05).

The results showed that the positional error in the occurrence data reduced the niche overlap in both the geographical and environmental space of both GLM and MaxEnt models. Niche overlap decreased gradually with the increasing positional error with an especially significant decrease in models' niche overlap at the extreme case of the positional error (100–500 m) (Figures 5.3, 5.4). However, the effect of the positional error on the niche overlap varied depending on species' niche breadth. Decrease in the niche overlap was higher for specialist and intermediate species than for generalist species, especially in the geographical space. For example, in case of MaxEnt models, Spear-

man´s rank correlation was reduced from 0.98 to 0.58 for the specialist and from 0.83 to 0.70 for the generalist species, respectively (Figure 5.3). However, the effect of the positional error was not that evident from I, especially for the generalist species in geographical space. For example, the decrease for generalist species and MaxEnt models was on average only from 0.96 to 0.9 and the GLM models appeared as not being affected at all.

Finally, independently of the validation metric, results showed that increasing the sample size cannot compensate for the effect of positional error (Figures 5.2, 5.3, 5.4). On the contrary, it is evident that a combination of low sample size of 30 samples with positional error led to erratic behaviour and generally low performance of the models.

5.3.3 Comparison of the relative importance of individual predictors (R^2)

The results show that the positional error and modelling technique had the highest relative importance (R^2) for the model performance (AUC, TSS). The relative importance of the sample size and niche breadth was much smaller and mutually comparable (Table 5.2). According to the niche overlap in geographical space assessed by I (model predictions), niche breadth had the greatest effect, followed by the positional error, modelling technique and sample size, the importance of which was almost negligible. In contrast, according to correlations, the modelling technique and positional error had the highest relative importance (R^2) followed by the niche breadth and by sample size, the importance of which was minimal. When assessing relative importance for niche overlap in the environmental space, the modelling technique and positional error showed the highest contribution followed by the niche breadth and by sample size, the importance of which was almost negligible, just like in the above metrics. All those factors significantly affected SDMs performance and predictions (p-value < 0.05).

5.4 Discussion

In this study, we focused on the effect of positional error in species occurrences on fine-scale SDMs. We simulated species with different levels of niche breadth to assess whether there was a link between the width of the environmental niche and the effect of the size of positional error. Our results showed that introducing positional error into species occurrence data led to a decrease in model performance and prediction accuracy in both the geographical and environmental space. However, the effect of the positional error varied with species niche breadth. The same positional error had a greater impact on specialist (low ROA and prevalence, narrow breadth of niche) than on generalist (high ROA and prevalence, wide breadth of niche) species. This is likely because in case of specialist species, occurrences could be easily shifted to inappropriate environments outside of the species' environmental niche. This could also explain the inconsistent conclusions of previous studies (Graham et al. 2008, Fernandez et al. 2009).

Table 5.2 Comparison of the relative importance of individual factors (R2, %) for ANOVA of performance metrics (AUC, TSS) and niche overlap in the geographical and environmental spaces (I, correlation).

Factor	**AUC**	**TSS**
ROA	4	4.14
Sample size	1.1	1.78
Modelling technique	18.7	21.35
Positional error	25.4	24.58

Factor	**I geographical space**	**Correlation geographical space**
ROA	75	11.2
Sample size	0.1	1
Modelling technique	8	24.7
Positional error	8.4	27.5

Factor	**I environmental space**	**Correlation environmental space**
ROA	9.7	1.7
Sample size	0.2	0.4
Modelling technique	45.4	21.5
Positional error	13.2	18.3

Higher sample sizes slightly improved unaltered models' accuracy; the results however showed that increasing the sample size could not compensate for the effect of positional error on models' accuracy (Figures 5.2, 5.3, 5.4). On the other hand, low sample sizes of positionally inaccurate data were especially problematic for modelling. These results are in general agreement with the study by Mitchell et al. (2017) who investigated the influence of sample size (ranging from 100 samples to 400) in conjunction with the positional error; their results showed that models based on smaller sample sizes were more affected by a positional error than those with higher numbers of species occurrences. However, it is difficult to conclude whether or not 100 records with positional error of 10 m are better or worse for modelling at the scale of 5 m than 500 records with positional error 25 m. For example, Moudrý and Šímová (2012) suggested that the spatial resolution of the environmental data should be coarser than the biggest positional error of the occurrence data and Naimi et al. (2011) showed that the effect of positional error is reduced by spatial autocorrelation in environmental variables. However, the trade-off between the scale and positional error has not been thoroughly studied.

The degree of decrease between unaltered and altered models (i.e. those with positional error) differed among adopted validation metrics and assuming a sufficiently large sample size, AUC and TSS provided clear evidence of decreasing model quality. The ability of evaluation metrics to identify the magnitude of error caused by positional inaccuracies was previously discussed by Osborne and Leitão (2009). Interestingly, they found that the use of AUC for the error quantification in models affected by positional error was limited as AUC did not decrease when compared to the control models. We hypothesize that this contradiction results from confounding effects of real data used in their study (i.e. they did not use virtual species). In Osborne and Leitão (2009), the modelling algorithms were allowed to choose the best combination of environmental variables from a set of twelve variables for scenarios

with different levels of positional error. Indeed, they showed that positional error led to alteration of the variables selected by the modelling algorithm. The selected variables however often failed to represent the conditions pertinent to the species during habitat selection. In contrast, here we used the same variables throughout, both to generate the virtual species and to model their distribution. Hence, our modelling approaches (GLM, MaxEnt) did not have the option to select variables that would provide a closer fit to the altered occurrence data but that were lacking ecological relevance and as a result did not lead to spurious increase in AUC and TSS values. We suggest that the effect of positional error on selection of environmental variables should be further investigated.

The effects discussed above raise serious concerns as it is possible that the use of positionally inaccurate data combined with an arbitrary selection of environmental variables that may lack ecological relevance results in seemingly accurate but entirely wrong models. For instance, Fourcade et al. (2018) successfully fitted SDMs with non-ecological variables such as paintings to demonstrate this point. While Osborne and Leitão (2009)) and Mitchell et al. (2017) suggested that useful predictions can still be generated from data affected by positional error, they warned that the ecological interpretation of such data and predictions was dangerous. Our results support the importance of assessing data in terms of fitness-for-use (Lecours 2017). Fitness-for-use is the concept of determining whether or not a dataset is of sufficient quality for a particular purpose (Goodchild 2006). Spatial scale is intrinsically linked to such assessment of fitness-foruse (Lecours et al. 2017a) as data accuracy is dependent on the spatial resolution of the environmental data. As indicated by Moudrý and Šímová (2012), the spatial resolution of the environmental data should always be coarser than the largest positional error associated with occurrence data.

In line with previous work (Van Niel and Austin 2007, Rocchini et al. 2011, Lecours et al. 2017a), we believe that attempts to predict species

distributions with data of unknown accuracy are potentially dangerous and as such, we highlight the necessity of quantifying the positional accuracy of data. If such assessment is limited by metadata availability, for example in case of historical data, we recommend to at least approximate the positional accuracy based on known information such as the collection methodology or the number of decimals recorded with coordinates. With a proper fitness-for-use assessment that includes data quality and scale, the resolution of environmental variables can be coarsened before they are integrated into a modelling exercise to minimize the adverse effects of the positional error of species occurrences. However, we are aware that this may involve altering the spatial resolution of data to a level that is no longer eligible for potentially optimal resolution(s), i.e. the scale at which species respond to the environment (Lecours et al. 2015, Moudrý et al. 2019b). As demonstrated in Lecours et al. (2017a), there is a tradeoff between spatial scale and data quality that needs to be evaluated as a part of the fitness-for-use assessment. While no experiments are currently available to help quantify which is more important for successful modelling (whether it is the data quality or scale), we suggest that pre-analyses be performed to test whether keeping a finer resolution is more important than minimizing positional error, or vice-versa. For new surveys, we suggest paying a close attention to measurement techniques to minimize positional error, for instance by using differential GNSS, especially for species with a narrow ecological niche as our results show that the positional error of species occurrence data has a profound effect on results of SDMs. Finally, we advocate for additional studies focused on the influence of positional error using more complex virtual species (e.g. with a higher number of environmental variables or with more complex response curves) to improve SDM use in ecology, macroecology and biogeography.

5.5 Conclusions

In this study, we explored how positional error in species occurrences affects fine-scale SDMs. We showed that the influence of positional error on SDMs differed according to the width of species' ecological niches and this effect was evident in both geographical and environmental space. The effect of the positional error on generalist species was much smaller than the effect on specialist species, which were affected the most. In addition, our results show that the negative effects of positionally inaccurate data entering SDMs cannot be mitigated by increasing the sample size. Therefore, a take away message of our study is that improving positional accuracy of data appears to be more effective than increasing sample size. We suggest that it is critical to evaluate the quality of data with respect to the spatial resolution of the environmental variables and to select occurrences with a low positional error (note that a low positional error can be even 1km if the spatial resolution of environmental variables is of similar size). Future research should be focused on the influence of positional error using more complex virtual species (e.g. with a higher number of environmental variables or with more complex response curves) and on how positional accuracy errors may affect the selection of variables in modelling species distribution to improve its future application in ecology, macroecology and biogeography.

Supplementary materials

Supplementary materials to this chapter (article) can be found online at http://www.ecography.org/appendix/ecog-04687.

Chapter 6

On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs

Vítězslav Moudrý, Vincent Lecours,Kateřina Gdulová, **Lukáš Gábor**, Lucie Moudrá, Jan Kropáček, Jan Wild

Adapted from Ecological Modelling 383 (2018): 3–9, with permission of corresponding author (V. Moudrý).

Publication metrics:
63 of 165 (Q2) rank in WOS category Ecology
IF (2018) 2.634; AIS (2018) 0.722

Author's contribution: 20%

Abstract

It is now widely acknowledged that the increasing availability of remotely sensed data facilitates ecological modelling. Digital elevation models (DEMs) are arguably one of the most common remote sensing products used in this context. Topographic indices (e.g. slope, orientation, rugosity) derived from DEMs are widely used as surrogates for field-measured environmental variables. Available global DEMs, such as those from the shuttle radar topography mission (SRTM), however, do not provide information on bare-earth elevation as they measure elevation of the highest objects above the ground (e.g. canopy). This affects the derived topographic indices and limits the use of global DEMs in ecological modelling. Unfortunately, most ecological studies ignore this limitation despite the fact that methods to remove the vegetation offset have been developed. We used high resolution LiDAR DTM to assess the accuracy of two newly available global bare-earth DEMs where such methods were applied and to compare them with the SRTM DEM. Furthermore, we assessed the effect of DEMs' vertical error on species distribution models (SDMs) by calculating slope and topographic wetness index (TWI) from these different models and evaluating their suitability for SDMs by adopting a virtual species approach. We simulated virtual species based on slope and TWI derived from accurate LiDAR DTM at three resolutions (30 m, 90m and 900 m) and developed univariate generalized models to assess the performance of the bare-earth and SRTM DEMs. Our results show that the vertical error in both newly available, vegetation-corrected global DEMs is indeed successfully reduced. The overall vertical root mean squared error (RMSE) was 10.52m for SRTM, while it was 6.80m and 6.25m for the two global bare-earth DEMs. The effect of the vertical error on SDMs was most significant at finer spatial resolutions. Using SRTM DEM, as opposed to a more accurate bare-earth DEM, led to a decline in area under curve (AUC) values from 0.94 to 0.77. SDMs

fitted with slope and TWI derived from new global bare-earth DTMs performed slightly better than SRTM. Since methods for vegetation-offset removal in DEMs exist and corrected DEMs are freely available, we argue that the vertical accuracy of DEMs should be more consistently considered. Local, high-accuracy DEMs should be used where available; in remaining instances, however, global DEMs where vertical bias was minimized should be used in ecological modelling. Further improvement of global DEMs at 30m and better resolutions are needed to enhance accuracy of derived indices and ecological models.

Keywords: Geomorphometry, Remote sensing, Scale, Species distribution model, Vertical error, Virtual species

6.1 Introduction

Understanding the principles that drive the spatial distribution of organisms and ecosystems is of central interest in ecology and the application of these principles to conservation and management problems is fundamental to the development of successful conservation and management strategies (Whittaker et al. 2005, Piroddi et al. 2015). Over the last few decades, this effort was facilitated by advances in modelling techniques. The objective of such modelling is either to relate a biodiversity response variable (e.g. the distribution of individual species or species richness) and explanatory variables to quantify their relationships ('explanatory modelling'), or to predict unknown values of the biodiversity response variable based on pre-established relationships with other variables ('predictive modelling') (Ferrier et al. 2017). Species distribution models (SDMs) are the most popular examples of such analysis (e.g. Alba-Sánchez et al. 2010, Reino et al. 2013, Piroddi et al. 2015, Zhang et al. 2016).

The improved accessibility of remotely sensed data facilitates ecological modelling (Geller et al. 2017). However, it may potentially bring hidden dangers emerging from the use of such data by users with limited understanding of data collection and processing methods, leading them to make erroneous conclusions (Lecours et al. 2017c). A joint effort from ecology and remote sensing practitioners is often needed to ensure a robust and valid use of available data and methods (Cord et al. 2013).

Arguably, one of the most common remote sensing product used in ecological modelling is the digital elevation model (DEM). Topographic indices derived from DEMs (e.g. slope, aspect, topographic wetness index) are routinely calculated using geographic information systems (GIS) and are widely used as surrogates for a variety of field-measured environmental variables such as air temperature, soil moisture and incoming solar radiation (Hengl and Reuter 2009). DEMs and their derived topographic indices have long been used for a vast range of

studies in ecology (Leempoel et al. 2015, Meineri et al. 2015, Lecours et al. 2016) and constitute a backbone of SDMs (see Franklin 1995, Mod et al. 2016). It is essential that a DEM used in a modelling exercise captures the relevant topographic details affecting species distribution (Lecours et al. 2015). Although highly accurate DEMs exist at both local and national level (for example from airborne laser scanning missions), many studies rely on global space-borne DEMs that have lower spatial resolutions and accuracy (e.g. Zhang et al. 2016).

Nowadays, global or near-global DEMs are available from several space-based data collection missions: Shuttle Radar Topography Mission (SRTM), Advanced Spaceborne Thermal Emission Reflectometer (ASTER) onboard NASA's Terra satellite, Advanced Land Observing Satellite (ALOS), or TANDEM-X. Unfortunately, both interferometric (SRTM, TANDEM-X) and stereoscopic (ASTER, ALOS) DEMs suffer from local inaccuracies or errors due to limits associated with the methods used for elevation measurements. Furthermore, the vertical accuracy of all space-borne DEMs strongly depends on the relief and ruggedness of the terrain as well as on the vegetation cover (Thomas et al. 2015). Importantly, it has been shown that such inaccuracies and errors can in turn influence the derived topographic indices (Van Niel et al. 2004, Oksanen and Sarjakoski 2005, Sofia et al. 2013, Lecours et al. 2017a) and various steps of the species distribution modelling process (e.g. shape of response curves, prediction accuracy measures, spatial extent of predictions) (Van Niel and Austin 2007, Lecours et al. 2017b).

SRTM DEM is one of the most commonly used global DEMs. The SRTM raw data were collected by C-band radar during an 11-days mission in February 2000 (Farr et al. 2007). The processed data were first released in June 2003. The SRTM DEM product was initially provided as SRTM-3 with a resolution of 3 arc-seconds (approximately 90m at the equator), but the United States Government recently released an updated version (SRTM-1) with a resolution of 1 arc-second

(approximately 30m at the equator) and a near-global coverage. The raw data, however, contain voids (areas for which no radar signal was returned), which reduces its usability in modelling. Since its initial release, the SRTM-3 was post-processed to fill data voids and is now available for free download, which greatly encouraged its widespread use.

The SRTM DEM contains various errors, the description of which is beyond the scope of this study (more information can be found in Rodríguez et al. 2006). However, an important but often misunderstood characteristic of the SRTM DEM is that it does not provide a "bare-earth" elevation: the measurements actually include a systematic positive bias due to the objects above the ground (such as canopy), the height of which is included into the model and this in turn produces considerable differences in accuracy between forested and open areas (e.g. Nelson et al. 2009). It is caused by the inability of the C-band radar signal to penetrate the vegetation canopy and to reach the bare ground: most of the incoming signals are reflected by various scatterers in the upper part of the canopy (e.g. leaves, branches) with the size similar to the relatively short wavelength of the C-band (5.6 cm). Consequently, the elevation values captured by the sensor are located somewhere between the ground and the top of the vegetation canopy (depending on vegetation structure). The theoretical vertical accuracy according to SRTM mission specifications is 16 m. Similarly, other available global DEMs (ASTER GDEM, ALOS DEM, TANDEM-X DEM) are also subject to the effects of vegetation offsets (e.g. Nelson et al. 2009, Thomas et al. 2015).

All available versions of the SRTM DEM are impacted by vertical error, including one of the most – if not the most – cited versions of the SRTM DEM produced by the Consultative Group for International Agriculture Research Consortium for Spatial Information (CGIAR-CSI; http://www.cgiar-csi.org/data; e.g. Moudrý and Šímová 2013, Reino et al. 2013, Šímová et al. 2015, Kosicki 2017). By not acknowledging

the vertical error, and more particularly its vegetation offset component, most studies use global DEMs as digital terrain models (DTM). If using the appropriate nomenclature, the original SRTM product and many of its subsequent alterations are actually digital surface models (DSM): they do not represent the bare ground elevation in vegetated areas and require further processing to remove vegetation heights in order to create a proper 'bare-earth' DTM.

Methods for vegetation offset removal rely on maps of tree cover and vegetation height from independent sources. Such data however must have an appropriate resolution and match the environmental conditions at the time of DSM acquisition. Consequently, most efforts to remove vegetation offset have been only applied locally. Gallant and Read (2016) developed a method consisting of three steps. First, a map of tree presence/absence yielding the best fit to the DEM offset is created from available global data on vegetation cover (e.g. www.globallandcover.com, www.earthenginepartners.appspot.com). Subsequently, the offsets near the edges of vegetation patches and in areas of continuous vegetation cover are estimated. Finally, the estimated offsets are subtracted from the DSM to produce bare-earth DTM. A similar method has been used by O'Loughlin et al. (2016) to develop the first near-global 'Bare-earth' DTM based on SRTM DEM at 3 arcseconds resolution. This DTM, hereafter referred to as the SGS-UB DTM, was made freely available for non-commercial use by the School of Geographical Sciences at University of Bristol, United Kingdom (https://data.bris.ac.uk/data/). More recently, Yamazaki et al. (2017) developed MERIT DTM, a high-accuracy global DTM at 3 arc-seconds resolution produced by eliminating multiple error components, including vegetation offset. MERIT DTM was also made freely available for noncommercial use by the Japan Agency for Marine-Earth Science and Technology (http://hydro.iis.u-tokyo.ac.jp/yamadai/MERITDEM/).

The general aim of this study is to promote the valid and robust use of global DEMs in ecological modelling by raising awareness about

the importance of using DEMs that are corrected for vegetation offset. Specific objectives are to (1) assess accuracy of SRTM DEM compared with newly available SGS-UB and MERIT DTMs with respect to land cover type (forested and non-forested areas); (2) evaluate whether topographic indices derived from newly available bare-earth DTMs perform better in SDMs than those derived from SRTM DEM; and (3) assess the role of spatial resolution for DEMs vertical error propagation to SDMs.

6.2 Materials and Methods

6.2.1 Study area and reference DTM

Our study area encompassed the Czech part of Krkonose mountains national park (KRNAP), located in Central Europe (15°25´–15°50´E and 50°38´–50°50´N). Krkonose is the highest mountain range in Czechia and constitutes an area significant for biodiversity on the regional level. The area is approximately 35 km in length, with the main ridges and valleys arranged in a northwest to southeast direction. The altitude range from 400 to 1600m is covered mostly by grasslands, pastures, and spruce monocultures with remnants of original broad-leaf and mixed mountain forests. The tree line traverses the altitudinal range of 1200–1350 m.

For the purpose of the comparison, a high-quality DTM derived from small-footprint airborne LiDAR data provided by KRNAP was used. LiDAR data were collected in 2012 and comprise an area of 478 km2 with the average pulse density of 5 pulses perm2. We classified the point cloud into "ground" and "non-ground" returns using *lasground_ new* with default setting for nature with following amendments: a) all returns were considered as possible ground and b) intensified search for initial ground points was set to fine (LAStools 2017). We filtered the ground

returns only and, using *blast2dem*, generated a DTM with a cell size of 1m (hereafter referred as LiDAR DTM). The horizontal coordinate system of LiDAR DTM is Datum of Uniform Trigonometric Cadastral Network (S-JTSK; EPSG: 5514) and the vertical coordinate system is Mean Sea Level (MSL; Baltic Vertical Datum – after adjustment; EPSG: 5705). The vertical datum of SRTM DEM, SGS-UB DTM and MERIT DTM is EGM96 (EPSG: 5171), which is a very close approximation of MSL (in Czechia, the differences should be below 1 m), and all models can therefore be directly compared. The data were horizontally referenced to WGS84 and projected to S-JTSK using the bilinear resampling method to 30m cell resolution for SRTM DEM and 90m resolution for SGS-UB DTM and MERIT DTM, respectively. There were no changes in the terrain height in the study area between the data acquisition for global DEMs and LiDAR data.

6.2.2 DEMs validation

The LiDAR DTM is considerably more accurate compared to other DEMs and can thus be used as the reference dataset (true elevation). To assess the accuracy of remaining DEMs, we first calculated vertical differences between LiDAR DTM and remaining models (SRTM DEM, SGS-UB DTM, and MERIT DTM) using pairwise combinations of all DEMs on cell-by-cell basis. We used the differences to calculate root mean square error (RMSE) and mean error (ME), expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(DEM_i - REF_i)^2}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(DEM_i - REF_i)$$

where DEM_i is the ith elevation from DEM surface, REF_i is the corresponding "true" measured elevation, and n is the number of elevation

points (cells) sampled. The RMSE expresses the dispersion of the frequency distribution of deviations between the true elevation (in this case represented by LiDAR DTM) and the DEM data. The ME tells us whether a set of measurements consistently overestimate (positive value) or underestimate (negative value) the true elevation. To evaluate the success of SGS-UB DTM and MERIT DTM in removing vegetation offset, we additionally assessed RMSE and ME in forested and non-forested areas. The information on forested areas was obtained from the vector CORINE land cover database (CORINE 3.1.1; 3.1.2; 3.1.3) for the year 2000.

6.2.3 Derived topographic indices

We aggregated LiDAR DTM using focal statistics (i.e. mean value) at 30m and 90m resolutions and derived commonly employed direct and indirect variable. The indirect variable we used was slope, calculated according to Zevenbergen and Thorne (1987). Slope (i.e. rate of change of elevation) affects the velocity of subsurface and surface flow and other surface processes (Gallant and Wilson 2000) and it is therefore an important variable in predictive vegetation mapping (e.g. Zhang et al. 2016). The direct variable was the topographic wetness index (TWI) which is a surrogate for soil moisture (Raduła et al. 2018). Soil moisture is among the most important environmental variables affecting vegetation composition (e.g. Kopecký and Čížková 2010). The TWI is defined as

$$TWI = \ln(\frac{As}{tan\beta})$$

where As is the specific catchment area and β is the local slope in radians (Beven and Kirkby 1979). To calculate the specific catchment area, we used the multiple flow routing algorithm of Quinn et al. (1991)

recommended by Kopecký and Čížková (2010). Slope and TWI were derived from four different datasets: (i) the LiDAR DTM used as a reference dataset in the simulation of virtual species, (ii) SRTM DEM as a model burdened with vegetation offset, (iii) the SGS-UB DTM and (iv) the MERIT DTM as models that were corrected for vertical error caused by vegetation.

6.2.4 Virtual species generation

In order to exemplify the effect of vertical error on SDMs and to assess whether the recently introduced "error-free" global MERIT DTM and SGS-UB improve model performance, we used the virtual species approach (Meynard and Kaplan 2012, Moudrý 2015). We used a simulated virtual species to ensure complete knowledge of species distribution in order to enable a proper assessment of model performance without confounding effects of real data (e.g. Moudrý et al. 2017). We simulated relationships (response function) between species and an environmental variable (i.e., slope and TWI derived from LiDAR DEM) to generate environmental suitability. The response to both variables was defined as the Gaussian response function (Slope: mean=15° and standard deviation=2°; TWI: mean=8 and standard deviation=1) (e.g. Varela et al. 2014). We adopted a probabilistic approach (logistic function with alpha= -0.15 and beta=0.65) to convert environmental suitability into probability of occurrence that is subsequently applied to randomly sampled presences and absences. Finally, we simulated the sampling of 400 presences and 400 absences of the virtual species. The use of relatively simple species-environment relationships allows us to highlight the effects of vertical error on the performance of SDMs. All simulations were undertaken in R v.3.2.2, using the recently developed package *virtualspecies* (Leroy et al. 2016), recommended by Moudrý (2015) to simulate virtual species.

6.2.5 Model fitting and evaluation

It is customary to fit Gaussian response with a polynomial model. We used generalized linear models (GLMs) with binomial error distribution and logit link function (McCullagh and Nelder 1989, Oksanen and Minchin 2002). The modelling was performed as univariate logistic regression. The linear and quadratic term of slope and TWI were included because of the known Gaussian shape of the response function. We modelled species distribution at 30 m, 90m and 900m to evaluate the effect of spatial resolution.

To evaluate the models, we split sampled presences-absences into test (50%) and training (50%) datasets. We run the entire process from species generation to model evaluation 100 times. Each repetition provided a different presence-absence distribution (Leroy et al. 2016). We evaluated model calibration by plotting the estimated environmental relationships and their discrimination capacity, assessed by means of computing the area under the curve (AUC) of the receiver operating characteristic plot (Fielding and Bell 1997). In addition, we calculated RMSE from differences between true (i.e. virtual) and predicted probability of occurrence. All spatial analyses were performed in ArcGIS 10.4.1. and Saga 2.1.4 (Conrad et al. 2015).

6.3 Results and discussion

6.3.1 DEMs and derived attributes accuracy

Among the DEMs studied, only slight differences of the mean, minimum and maximum values of elevation, slope and TWI were detected. The values of mean elevation for both error corrected DTMs were closer to the LiDAR DTM than those of the original SRTM DEM (Table 6.1). The overall RMSE for the SRTM DEM was 10.52 with a mean bias of

7.62 m. The vertical error of SRTM DEM significantly differed between forested and non-forested areas as expected due to limits associated with the methods used for DEM measurements and shown by numerous studies (e.g. Nelson et al. 2009). In the forested areas, the RMSE was 13.25 and elevations were, on average, 11.60m higher than the LiDAR DTM elevations. In the non-forested areas, the RMSE was 5.74 and the difference was 2.85m (Table 6.2).

Table 6.1 Descriptive statistics of elevation, slope, and topographic wetness index (TWI) estimated from different DEMs at 90m spatial resolution. Elevation is present as height above mean sea level.

	Elevation		Slope		TWI	
	Mean ± s.d. [m]	Range [m]	Mean ± s.d. [°]	Range [°]	Mean ± s.d.	Range
LiDAR	853 ± 235	396–1553	13.9 ± 6.5	0.1–48.4	7.7 ± 1.4	4.8–17.4
SRTM	860 ± 233	402–1549	3.1 ± 6.2	0.1–44.2	7.8 ± 1.3	5.4–15.9
SGS-UB	850 ± 233	397–1548	13.1 ± 6.2	0.1–44.7	7.8 ± 1.3	4.8–16.0
MERIT	856 ± 233	400–1552	13.3 ± 6.3	0.1–45.7	7.8 ± 1.3	4.8–16.3

The overall RMSE for the SGS-US DTM was 6.80 with a mean bias of -2.30 m. In the forested areas, the RMSE was 6.94 and elevations, on average, -1.08m lower than the LiDAR DEM elevations. This improvement was consistent with RMSE of 6m reported by O'Loughlin et al. (2016). The vertical error has significantly improved in forested areas, the terrain was however on average slightly underestimated as shown by our results (Table 6.2, Figure 6.1). This is likely due to overestimation of tree heights. O'Loughlin et al. (2016) admitted that artefacts may exist at the boundaries between forested and non-forested areas with the elevation in forested areas possibly being slightly lower than that of the adjacent non-forested area. This can be especially true in higher altitudes of our study area where tree heights can be significantly lower than those expected by the models. In addition, the results of SGS-UB DTM got worse in nonforested areas (Table 6.2, Figure 6.1) with the RMSE of 6.62 (compare to RMSE of 5.74m for the SRTM DEM) and the mean bias of -3.75 m.

The overall RMSE for the MERIT DTM was 6.25 with mean bias of 3.09 m. Compared to SRTM, the vertical error has significantly

improved in both forested and non-forested areas (Table 6.2, Figure 6.1). In the forested areas, the RMSE was 7.30 and elevations, on average, 4.58m higher than the LiDAR DTM elevations. In the non-forested areas, the RMSE was 4.70 and the difference 1.30 m. Our results suggest a better accuracy of MERIT DTM over SGS-UB DTM. This is likely due to multiple error components considered (i.e. speckle noise, stripe noise, absolute bias, and tree height bias) compared to study by O'Loughlin et al. (2016)who only removed tree height bias to construct their SGS-UB DTM. Furthermore, the accuracy is highly dependent on the estimation of tree heights and size of individual forested areas. SGS-UB DTM accuracy decline in non-forested areas is likely due to overestimation of the size of individual forested areas caused by limitations of adopted 250m MODIS Vegetation Continuous Field product DiMiceli et al. (2017), compared to Yamazaki et al. (2017) who adopted 30m resolution data (Hansen et al. 2013).

Table 6.2 RMSE and ME of elevation of SRTM, SGS-UB and MERIT in forested and nonforested areas.

	RMSE			ME		
	Overall	Forest	Non-forest	Overall	Forest	Non-forest
SRTM	10.52	13.25	5.74	7.62	11.60	2.85
SGS-UB	6.80	6.94	6.62	-2.30	-1.08	-3.75
MERIT	6.25	7.30	4.70	3.09	4.58	1.30

6.3.2 SDM accuracy

Our results show considerable differences in RMSE and AUC values between models fitted with DEM affected by vertical bias (SRTM DEM) and "error-free" DTMs. Models fitted with "error-free" DTMs (SGS-UB DTM and MERIT DTM) performed better than SRTM in most cases (and never worse). However, the improvement in AUC values was rather small and did not achieve the quality of models fitted with LiDAR DTM (which was used for virtual species generation).For instance, the

AUC of the model fitted with topographic wetness index derived from MERIT DTM was 0.83 as opposed to 0.79 with SRTM DEM and 0.88 with the LiDAR-derived reference DTM (Table 6.3).

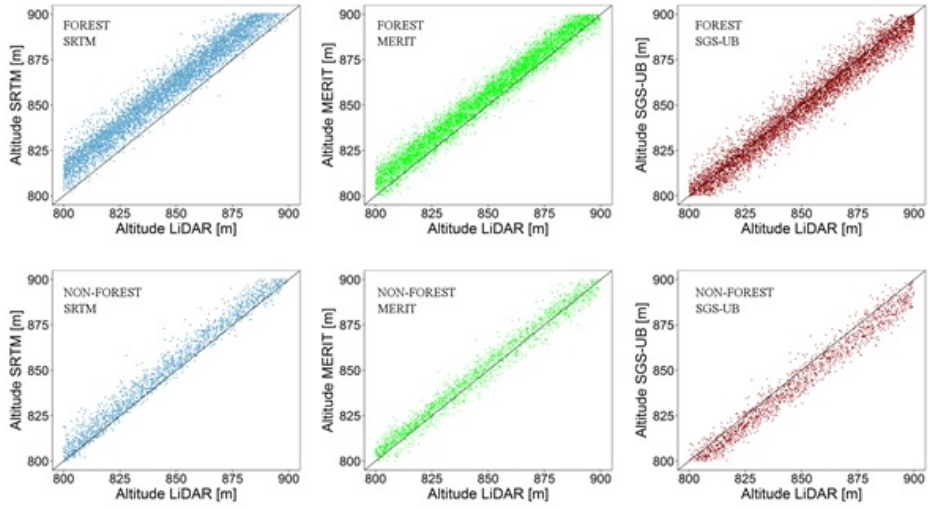

Figure 6.1 Scatter plots showing relationship between elevation at 90m resolution derived from SRTM DEM (blue), MERIT DTM (green), and SGS-UB DTM (red), respectively, and elevation derived from LiDAR DTM. Top row of figures show forested while bottom figures non-forested areas. The solid line indicates y=x. Average altitude of the study area is approximately 850m and only cells with elevation from 800m to 900m were selected as a representative example. Note that accuracy is affected by the accuracy of land use data used (CORINE land cover).

We also tested models with logistic response function, obtaining the same results (therefore we only present results for Gaussian response). However, the effect of vertical bias on performance of individual models was more evident for models with Gaussian response function than for simple logistic response. This suggests that the effect of the vegetation offset can be much more significant for more complex response functions (skewed, unimodal responses), often occurring in real ecological and biogeographical contexts (Oksanen and Minchin 2002, Dvorský et al. 2017). It is also likely that using additional explanatory variables would increase differences in models performance. When the multiple

topographic indices and the elevation surface are combined for species distribution modelling applications, the errors add up and can significantly impact the accuracy of the modelling (Van Niel and Austin 2007).

Table 6.3 Median AUC (and 2.5th–97.5th percentiles) and mean RMSE of probability of occurrence of the evaluated models for four datasets, two variables and three resolutions obtained from 100 simulated data sets with 400 training sites and 400 validation sites. For 900m resolution the sampling was 10 times lower.

DEM	30m resolution		90m resolution		900m resolution	
	Slope	TWI	Slope	TWI	Slope	TWI
AUC						
LiDAR DTM	0.94 (0.90–0.96)	0.90 (0.87–0.92) 0.93	(0.81–0.99) 0.88	(0.83–0.91)	0.92 (0.83–0.98)	0.81 (0.63–0.91)
SRTM DEM	0.77 (0.72–0.82)	0.77 (0.72–0.81)	0.85 (0.70–0.95)	0.79 (0.73–0.83)	0.90 (0.81–0.98)	0.78 (0.61–0.91)
SGS-UB DTM	*not available*	*not available*	0.86 (0.71–0.96)	0.81 (0.75–0.85)	0.90 (0.82–0.98)	0.78 (0.61–0.91)
MERIT DTM	*not available*	*not available*	0.87 (0.72–0.96)	0.83 (0.77–0.88)	0.92 (0.82–0.98)	0.80 (0.62–0.91)
RMSE						
LiDAR DTM	0.19	0.10	0.17	0.14	0.14	0.27
SRTM DEM	0.35	0.25	0.29	0.23	0.18	0.28
SGS-UB DTM	*not available*	*not available*	0.28	0.22	0.16	0.28
MERIT DTM	*not available*	*not available*	0.27	0.20	0.15	0.28

The ability to model species-environment relationships and to discriminate between presences and absences was also strongly affected by spatial resolution. In ecological studies, species distribution data are often available at a coarser resolution than data on elevation and topographic indices. The latter are thus often aggregated/resampled to match the coarser spatial resolution of the species data (e.g. Alba-Sánchez et al. 2010, Zhang et al. 2016, Kosicki 2017). It is well known that when DEMs contain vertical errors, the accuracy of derived topographic indices increases with lower resolution (e.g. Zhou and Liu 2004). In accordance with this, shifts in the modelled response functions and differences in AUC values (Table 6.3) were almost negligible at 900m resolution. However, models based on SRTM DEM at 30m resolution significantly underestimated the highest probability of occurrence and overestimated low probability of occurrence (Figure 6.2). This caused the highest drop in AUC values, decreasing from 0.94 and 0.90 for slope and TWI, respectively (reference LiDAR DTM) to 0.77 (SRTM DEM). As expected, the differences in TWI values are the highest at forest

boundaries where the SRTM DEM records false increases in elevation and the areas of highest TWI values are therefore shifted compared to values derived from LiDAR DTM (Figure 6.3). It is therefore likely that such vertical error in DEMs may also be the reason why some multi-scale studies reported better models at intermediate scales than at the highest resolution available. For example, Zhang et al. (2016) assessed performance of topographic indices derived from ASTER GDEM (which suffer from vegetation offset error) at five resolutions from 30m to 900m and reported the best models for *Abies faxoniana* and *Quercus aquifolioides* to be at 120m and 240 m. In contrast, Mohamedou et al. (2017) tested the effect of LiDAR derived (i.e., free of vertical bias) DTM resolution on TWI ability to predict tree growth. They adopted resolutions ranging from 1m to 30m and found the best models to be derived at 1m and 15m resolutions, respectively. Although different species may respond to environment at different resolutions (e.g. Lecours et al. 2015), it supports our hypothesis that vertical bias in DEMs can prevent from getting the best models in higher resolutions.

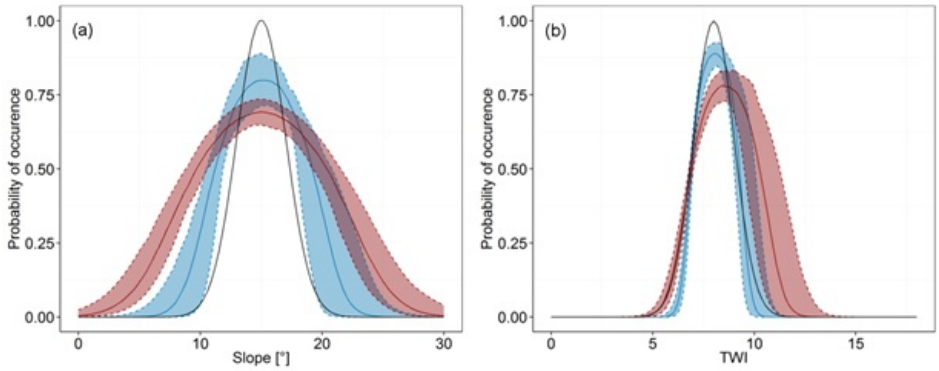

Figure 6.2 Environmental relationships between slope (a) and topographic wetness index (b) estimated with generalized linear models for LiDAR DTM (blue) and SRTM DEM (red) at 30m resolution. The shaded areas represent the regions delimited by the 5th–95th percentiles of the estimated probability of occurrence obtained from 100 simulations. Black lines show the "true" relationships.

According to Zhang et al. (2016), it is possible to use solely DEMs to predict vegetation distribution. In combination with the recent availability of global DEMs at higher spatial resolution (e.g. 12m TANDEMX DEM), this can provide better grounds for conservation and management actions. However, vertical accuracy bias can become even more problematic and using models burdened with it (albeit having good spatial resolution) may provide misleading results and must be carefully interpreted to avoid the risk of making incorrect decisions.

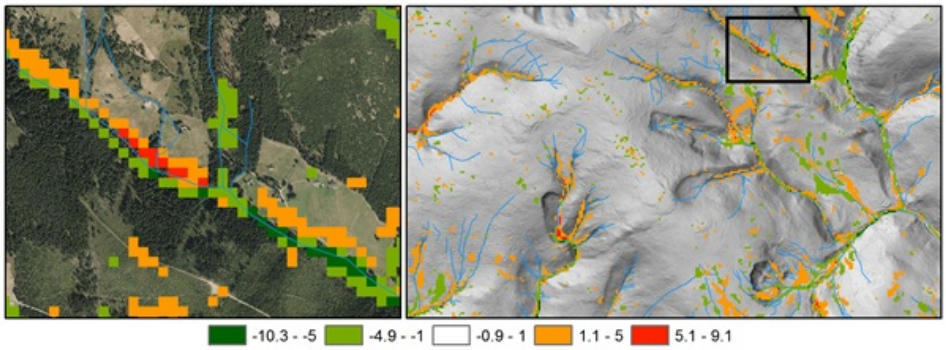

Figure 6.3 The difference between TWI calculated from SRTM DEM and LiDAR DTM at 30m resolution. On the left detail, the major effect of the forest boundary along the stream on TWI error is obvious. On the right, the differences over shaded relief are depicted. The most striking errors of TWI are evident along rivers in valleys. Positive values show areas of (incorrectly) higher TWI index (orange and red colours) while negative values (green) show areas of lower TWI index calculated from SRTM DEM compared to LiDAR DTM. Streams are shown in blue, transparent pixels of the model denote good fit (-0.9–+1).

SDMs are increasingly generated using high resolution data (Moudrý and Šímová 2012, Lecours et al. 2015), particularly DEMs (Pradervand et al. 2014, Matawa et al. 2016, Nezer et al. 2017, Bazzichetto et al. 2018a). It is therefore necessary to take into consideration the vertical accuracy of DEMs before including them and their derived topographic indices in the modelling process. Failure to do so can lead to misinterpretation of species-environment relationships and misidentification of areas important for species conservation. Local high-accuracy DTMs

should be used when available. For example, such models are available through governmental agencies in Europe (e.g. Fogl and Moudrý 2016). Where not available, the recently released global DTMs with reduced vertical error represent an adequate substitute. While our study area and the environmental conditions studied were limited in scope and extent, our results are in line with those of other studies that have achieved satisfactory global accuracy assessments. For instance, O'Loughlin et al. (2016) showed that improvement in DTM accuracy was consistent over all types of forest vegetation (evergreen forest, deciduous forest, etc.). In addition, Yamazaki et al. (2017) showed that the most significant improvement was in flat forested areas and that most residual errors were found in mountainous areas due to large subpixel topographic variability. While studies looking at the accuracy of newly created global DTMs will be required to validate our conclusions in other study areas, our results show that correcting for vertical bias in mountainous area significantly improved the DTMs' vertical accuracy and consequently the performance of SDMs. Based on those results, we recommend the adoption of the MERIT DTM in any area where forests are present.

6.4 Conclusions

We have shown that recently available vertical error-corrected MERIT DTM at 90m has a better accuracy in our study area than SRTM DEM. Furthermore, models developed with MERIT DTM performed slightly better than uncorrected SRTM DEM, thus being a more reliable alternative to DEMs affected by a vertical error. However, the improvement is lower than expected particularly due to limitations associated with estimates of tree heights and size of individual forested areas. Elimination or at least reduction of error components (e.g. vegetation offset) for DEMs available at better resolutions (e.g. TANDEMX) should be a priority for further enhancement of ecological modelling. Finally, users

must be aware that substantial vertical bias can still be present even in corrected DEMs, potentially propagating through the analysis and affecting the outcomes of ecological modelling.

Chapter 7

Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies

Vítězslav Moudrý, Vincent Lecours, Marco Malavasi, Benjamin Misiuk, **Lukáš Gábor**, Kateřina Gdulová, Petra Šímová, Jan Wild

Adapted from Ecological Informatics (2019) with permission of corresponding author (V. Moudrý).

Publication metrics:
81 of 165 (Q2) rank in WOS category Ecology
IF (2018) 2.31; AIS (2018) 0.645

Author's contribution: 25%

Abstract

Terrain attributes (e.g., slope, rugosity) derived in Geographic Information Systems (GIS) from digital terrain models (DTMs) are widely used in both terrestrial and marine ecological studies due to their potential to act as surrogates of species distribution. However, the spatial resolution of DTMs is often altered to match the scale at which species observations were collected. Here, we highlight the significance of adequately reporting the methods used to derive terrain attributes from DTMs and the consequences of their incorrect reporting in ecological studies. To ensure full repeatability of studies, they should report (i) the source and the resolution of the original DTM; (ii) the algorithm used to calculate terrain attributes; (iii) the method used for rescaling (e.g., aggregating or resampling, using the mean or maximum values); and (iv) the order in which these operations were performed. We contrast the effects of two common scale alteration approaches for the derivation of terrain attributes from DTMs. These two scale alteration methods differ in the step at which the change is performed: (i) the resolution alteration is performed after computing terrain attributes from the original DTM at the native resolution, or (ii) the resolution alteration is performed on the native DTM before computing terrain attributes. While these approaches conceptually do the same thing (i.e., change the resolution of the terrain attributes), we demonstrate that they produce two distinct sets of variables that are not interchangeable and describe different properties of the terrain. In a species distribution modelling (SDM) context, the first approach calculates terrain attribute values within the cell where a species is found, while the second approach calculates terrain attribute values with respect to neighbouring cells. A mutual substitution of the two approaches results in a decrease of models' discrimination ability and in misleading spatial predictions of species probability of occurrence. Regardless of the DTM-derived attribute, we argue that the choice of the approach should be care-

fully guided by both the ecological scale relevant to the question being asked and the performance of pre-analyses. We emphasize that selected methods be clearly described to encourage reproducibility and proper interpretation of results, thus enabling a better understanding of the role of scale in ecology.

Keywords: DTM, Geomorphometry, Remote sensing, Scale, Species distribution model, Terrain attributes

7.1 Introduction

In the last few decades, developments in ecology and biogeography have been fuelled by progress in tools, modelling techniques, and software such as Geographic Information Systems (GIS) (Rocchini et al. 2017), along with the growing availability of high-quality data from a variety of sources (Wiersma et al. 2011). Spatial data, particularly those acquired with remote sensing methods, have become a critical part of species-environment relationships studies across the terrestrial (e.g., Elith and Leathwick 2009), marine (e.g., Robinson et al. 2011) and aquatic (e.g., Domisch et al. 2015) realms. However, the improved accessibility of readily available tools and data may present pitfalls for non-critical users with a limited understanding of spatial data characteristics and of how they are collected and processed (Jarnevich et al. 2015, Lecours 2017, Moudrý et al. 2018).

One of the elements that often suffer a lack of explicit consideration is spatial scale (Wheatley and Johnson 2009, Lecours et al. 2015, Araújo et al. 2019). Spatial scale is one of the most important characteristics of spatial data (Zhang et al. 2014) and one that has been defined numerous times, with small variations depending on the context. Dungan et al. (2002) and more recently Lecours et al. (2015) discussed three types of spatial scale relevant for ecological studies that use spatial data: (1) the ecological scale, which is the scale at which a pattern or process occurs, (2) the observational scale, which refers to the characteristics of the data used to represent or describe natural phenomena and is usually defined by the spatial resolution and extent of the data, and (3) the analytical scale that refers to the methods used to analyse the data (e.g., the neighbourhood size used in focal statistics or geomorphometry).

It is well known in ecology that the lack of explicit consideration of scale may affect the outcomes of ecological analyses (e.g., species-environment relationships). Thirty years ago, Wiens (1989) argued that most ecological studies at the time were not considering spatial

scale and its effect on analyses. Most studies were performed as if patterns and processes were scale-independent, and studies performed at different scales were often compared while they should not have been. Since then, the role and importance of spatial scale have been extensively discussed in both the geographic and ecological literature (e.g., Schneider 2001, Dungan et al. 2002, Goodchild 2011, Moudrý and Šímová 2012, Lecours et al. 2015), and it is now widely recognized that ecological patterns and processes are scale-dependent and that no single scale is appropriate for the study of all natural phenomena (e.g., Levin 1992). Nevertheless, Wheatley and Johnson (2009) showed that 70% of the studies considered in their analysis arbitrarily selected the observational scales without consideration of whether they matched the ecological scale. Indeed, the issues highlighted by Wiens (1989) about overlooking spatial scale and its effect on analyses persist to some extent.

In recent decades, species distribution modelling (SDM) has been widely applied to model species-environment relationships, often involving the use of remotely sensed data as predictor variables. One of the main challenges of using remotely sensed data in SDM is that the original spatial resolution of different datasets included in the analysis may vary significantly (Cord et al. 2013). In terrestrial applications, the use of high-resolution remotely sensed environmental data is often limited by the resolution of species distribution data, which are usually available at much coarser scales (Jetz et al. 2012, Šímová et al. 2019). An opposite situation may occur in marine applications (e.g., in deep water environments), where the scale at which the species are observed can be much finer than that of available environmental data. In some cases, the highest available resolution of environmental data may not be required for the SDM, if the biological or ecological processes, or species distribution or abundance in SDMs, occur at a coarser scale or over a large area with limited observations. Therefore, finding a resolution representing a compromise between the resolution of available data

and a resolution most suited to the application is often necessary. A common practice to ensure the valid integration of data from multiple scales, for example to avoid an ecological fallacy or the modifiable areal unit problem (see Lecours et al. 2015), is to modify the resolution of some of the data so that it matches the resolution at which the study is meant to be performed (e.g., by averaging environmental variables within field plots; Gottschalk et al. 2011, Moudrý et al. 2017). The effects of altering data resolution (i.e., matching the observational scales with ecological scales) (e.g., Lechner et al. 2012b, Svensson et al. 2013, Mateo Sánchez et al. 2014, Lowen et al. 2016, Mertes and Jetz 2018) and of changing analytical scale (e.g., Dolan and Lucieer 2014, Wilson et al. 2007) on the outcomes of ecological analyses (e.g., SDM) have recently received more attention because of a growing demand on users to provide more detailed methodologies (e.g., exact computer code for GIS analyses and data processing) to allow complete reproducibility of their results (e.g., Michener and Jones 2012, Rocchini and Neteler 2012, Meynard et al. 2019). However, details on the use of terrain attributes derived from digital terrain models (DTMs) using neighbourhood operations (e.g., slope, rugosity, orientation, curvature) in SDM has received less attention. Deriving terrain attributes has become a routine operation, despite several potential pitfalls in the data processing workflow (Lecours et al. 2017c).

More specifically, the approach used and the step at which scale is altered are particularly important for DTM-derived attributes, as they will cause the analytical scale of a study to vary, thus providing different representations of the reality and potentially producing different outcomes (Figure 7.1). This paper aims (i) to illustrate which information related to scale alteration approach should be provided in ecological studies in order to facilitate their repeatability, and (ii) to demonstrate how addressing the effects of the process of altering data resolution in ecology can impact derived terrain attributes and subsequently SDMs. Accordingly, this paper also aims to provide some guidelines for a crit-

ical assessment of spatial data to be used in ecological studies. Our general aim is to increase awareness about this topic and promote better practice. Our examples and analyses will use slope as DTM-derived terrain attribute, as slope is one the most common variables calculated from DTM (Bradie and Leung 2017, Lucieer et al. 2018).

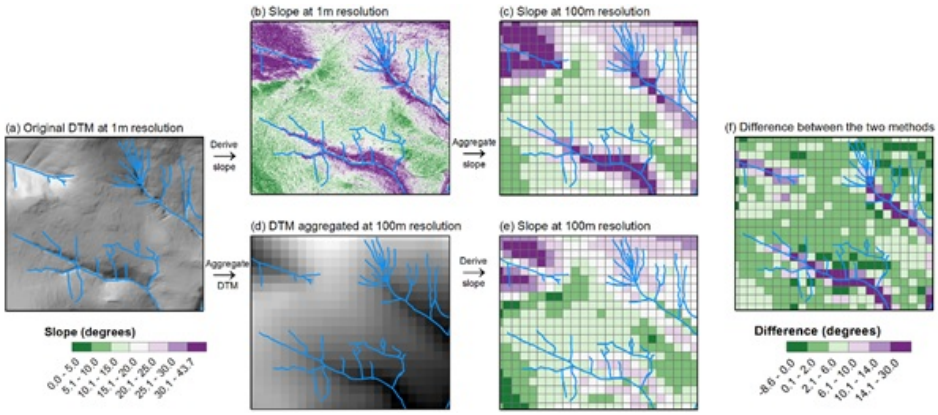

Figure 7.1 Multiple scale comparison of the two approaches. (a) Hill-shaded, LiDAR-based DTM at 1m resolution. (b) Slope derived from the 1m resolution LiDAR-based DTM. (c) Slope values resulting from the aggregation (mean value) at 100m resolution of the slope derived from the original DTM at 1m resolution. (d) DTM resulting from aggregation (mean value) at 100m resolution of the original DTM at 1m resolution (e) Slope values derived from the aggregation of the original DTM at 100m resolution. Note that both images (c, e) are at 100m resolution, but differ in slope values (f) because they were produced by different approaches (i.e., the timing of scale alteration within the data processing workflow was changed). Note the large differences along the valleys. Streams are shown in blue. Slope was computed using a 3×3 cell neighbourhood with Horn's (1981) algorithm.

7.2 DTM-derived variables in SDM

DTMs and their derived terrain attributes constitute a backbone of SDM (Franklin 1995, Mod et al. 2016) and are one of the most common types of datasets that undergo scale alteration in preparation for inclusion in SDM. Alterations induced to DTMs are known to affect terrain attributes (e.g., slope, orientation, rugosity) that describe and quantify terrain morphology (Pike et al. 2009, Lecours et al. 2017a) (Figure 7.1). DTMs and their derived terrain attributes are widely used in both terrestrial and marine ecological studies (Wilson et al. 2007, Moudrý and Šímová 2013, Pradervand et al. 2014, Bouchet et al. 2015, Lecours et al. 2016, Bazzichetto et al. 2018a,b, Walbridge et al. 2018) due to their potential to act as surrogates for species distribution. For instance, aspect (i.e., the orientation of the slope) can inform on exposure to sunlight or dominant currents that can respectively be important for vegetation (e.g., Li et al. 2010) and marine suspensionfeeders (e.g., Tong et al. 2016). Terrain attributes are especially relevant for marine studies where bathymetry is often one of the few continuous environmental variables available, particularly in deeper waters.

Terrain attributes are derived from DTMs using neighbourhood operations; the size of the neighbourhood that is selected for a particular analysis thus defines the analytical scale of the study. This makes the integration of DTMs and terrain attributes in ecology particularly relevant to the understanding of scale issues, including mismatches between types of scale, as it is one of the few applications that involve the simultaneous consideration of ecological, observational and analytical scales. Terrain attributes are highly scale-dependent (Tate and Wood 2001), and changes in native data resolution (observational scale) (Deng et al. 2007, Dolan and Lucieer 2014, Grohmann 2015) and extent of the analysis window (analytical scale) (Wilson et al. 2007, Dolan and Lucieer 2014) will generally produce different resulting terrain attribute surfaces. This variability, combined with the scale dependency of eco-

logical patterns and processes, makes it challenging to adequately link terrain attributes with observed ecological patterns, potentially leading to misidentified scale-dependent patterns (Lechner et al. 2012a).

7.2.1 Transparency in reporting information used to derive terrain attributes

In order to facilitate the repeatability of a study, a complete description of the DTM processing workflow for the derivation of terrain attributes should be provided. Studies should report (i) source and resolution of the original DTM; (ii) algorithm used to calculate the terrain attributes; (iii) method used for rescaling (e.g., aggregating or resampling, using the mean or maximum values); and (iv) order in which operations have been performed.

Of those five elements, the source and original resolution of DTMs are the most commonly reported, although there are still some studies that omit to explicitly share this information. For example, Hsu et al. (2012) highlighted that terrain-related factors must be incorporated when projecting species response to climate change, but did not provide the source and resolution of the original DTM they used to calculate slope and other terrain variables. This has implications for their argumentation since the prediction of species responses to climate change is directly dependent on the resolution and scale at which terrain attributes are calculated. Similarly, many studies do not provide information on the operations used for rescaling (i.e., statistical operation used to calculate new values such as average or maximum) their data or the interpolation method (e.g., nearest neighbour, bilinear, cubic) used to do so. Most often, only a statement about the decreasing or resampling of the resolution is provided (e.g., Brito et al. 2009, 2011, Rodríguez-Soto et al. 2011, Martínez-Gutiérrez et al. 2018). This often goes hand in hand with missing information about the order in which

the operations were performed (e.g., Convertino et al. 2012, Mateo-Tomás and Olea 2015). For example, Martínez-Gutiérrez et al. (2018) stated that, among other terrain attributes, slope was calculated from shuttle radar topography mission (SRTM) DTM at a spatial resolution of 30´´. However, the original resolution of SRTM is either 1´´ or 3´´, depending on the version (Farr et al. 2007), highlighting that information on the resampling method and the order of operations was omited. A reader not familiar with the SRTM DTM would likely not be aware that resampling had even occurred. In other cases, the methods are not transparent and all the necessary information for judicious use of terrain attributes in models are missing (i.e., i, ii, iii, iv) (e.g., Brambilla and Ficetola 2012, Abolmaali et al. 2018), thus preventing reproducibility.

Nevertheless, there has been a promising increase in studies who do report all the necessary information. Those studies should be used as a starting point for developing good practices. For example, Guisan and Hofer (2003) explicitly stated that they used the Swiss digital elevation model (DEM) (i.e., the source (i)) at 25m resolution (i.e., the resolution (i)), and provided a reference for accessing this DEM (OFT 2002). They also stated that they used the slope function implemented in ArcGIS (ESRI, CA, USA) (i.e., the algorithm (ii)), which is known to be Horn's (1981) algorithm. A possibly even better way of reporting the methods would have been to refer directly to the particular algorithm (Horn's) and neighbourhood shape and size used to calculate the terrain attributes. Finally, the authors also explicitly mentioned that slope was generated from the original DEM and then aggregated using the mean within the scale of the study (1 km resolution).

7.2.2 Two common approaches to DTM scale alteration and how they produce different outcomes

The derivation of terrain attributes from DTMs is a particular case in the broader context of altering data resolution as many different techniques can be used (see Dolan 2012, Dolan and Lucieer 2014). Here we present two main approaches that differ by the step at which the alteration is performed (e.g., Grohmann 2015) (Figure 7.1). In the first case (Figure 7.1c), terrain attributes are derived from the DTM at its native resolution, and the resolution of the resulting terrain attributes is then altered (e.g., Guisan and Hofer 2003). In the second case (Figure 7.1e), the DTM native resolution is altered before deriving the terrain attributes (e.g., Irl et al. 2015). Both approaches are commonly used in the literature, but the rationale behind the chosen approach is often left undocumented, casting doubt on whether it is haphazardly or intentionally selected, and on the quality of the analytical outcomes.

A typical example that involves altering data resolution is the use of terrain attributes to explain or model species distribution available as gridded data in distributional atlases (e.g., Krojerova-Prokesova et al. 2008, Šímová et al. 2015). Distributional atlases aim to provide information on distribution and abundance of species over a geographical area that may range from tens of square kilometers up to a whole continent. A common approach for sampling species occurrences is to cover the study area with a grid, the resolution of which typically ranges from hundreds of meters to tens of kilometers (e.g., Gibbons et al. 2007). Using distributional atlases as an example, Figure 7.2 shows how the two approaches to scale alterations capture different representations of reality (Figure 7.2c, g). The first approach (Figure 7.2a–d) calculates terrain attribute values within the hypothetical atlas grid cell, while the second one (Figure 7.2e–h) calculates terrain attribute values based on neighbouring atlas grid cells. The calculation of terrain attributes is

a focal (i.e., neighbourhood) operation and the output value for a given cell is a function of the cell values within a specified neighbourhood around that given cell. Note that the neighbouring cells used in terrain attribute calculations might differ depending on the adopted algorithm and its specifications. For example, to derive slope, Horn's (1981) algorithm uses eight neighbouring cells (queen's case) while Zevenbergen and Thorne's (1987) algorithm uses only four neighbouring cells (rook's case) to calculate each slope value. Other methods (e.g., Wood 1996) also include the center pixel in the calculations.

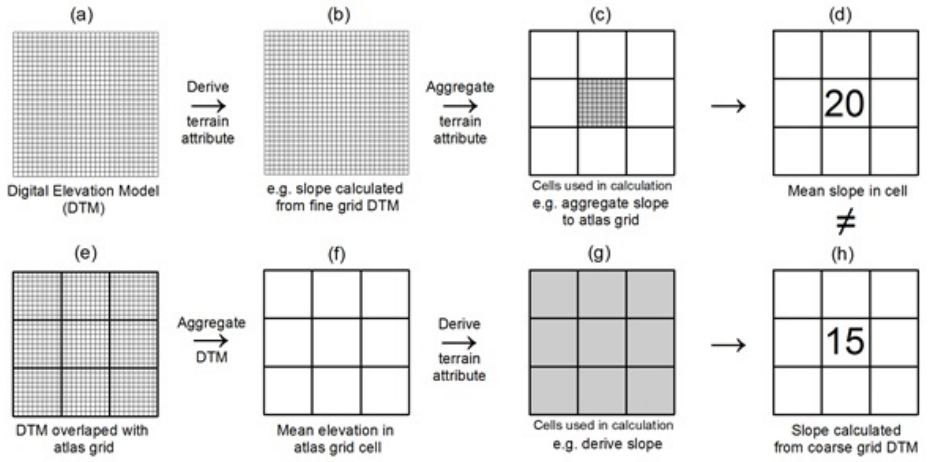

Figure 7.2 The upper row shows a (a) finer-resolution DTM (b) that is directly used to calculate a terrain attribute at the same resolution of the DTM. (c) The attributes are then aggregated into hypothetical atlas grid cells (d) and used to calculate mean terrain attribute values in the atlas grid cells. The lower row shows (e) the finer resolution DTM overlaid by a coarser-resolution species occurrence data or a resolution at which the study is meant to be performed. (f) The DTM is then aggregated or resampled to that coarser resolution before (g) deriving a terrain attribute resulting in terrain attribute values at the resolution of the (h) hypothetical atlas grid cells. The grey areas in (c) and (g) represent the cells that are used to calculate the resulting slope (i.e., different components of the reality).

Figure 7.2 demonstrates these concepts in a general context but can be translated into an ecological example to show the importance of differentiating between scale alteration methods. Mogotes are isolated

steep-sided hills surrounded by flat plains; the slopes alongside mogotes are thus very steep. If we were to survey a mogote and produce a DTM representation of it, the application of the first approach (i.e., derive terrain attributes first, then aggregate) would yield very high slope values along the sides of the mogote. On the contrary, the application of the second approach (i.e., aggregate first, then derive terrain attributes) could yield lower slope values by computing a DTM with a cell size that incorporates the elevation of the mogote with that of the surrounding terrain – effectively "smoothing" out the slopes when calculated at a coarser resolution (Figure 7.3). If an ecologist was to map habitat suitability for a bird species that is known to nest on the very steep sides of mogotes, the application of the first approach would appropriately capture the high slopes and identify them as suitable habitat, while the second approach could fail to do so. In general, the more complex the topography (e.g., mountainous regions with steep slopes and deep valleys), the higher the differences between the two approaches will be.

Although it is generally possible to always use both approaches, we recommend basing the methodological decision on existing ecological knowledge and research questions. We suggest that the first approach (i.e., derive terrain attributes first, then aggregate) is appropriate when characterising environment within the target cell size (e.g., atlas grid square) where the species was observed (e.g., bird nest is located on a steep rock wall). The first approach can depict the original features of the terrain more realistically than the second approach (Grohmann 2015) and should be considered any time when features of scale finer than the scale of the study are expected to affect modelled species. On the other hand, we recommend using the second approach (i.e., aggregate first, then derive terrain attributes) when the goal is to relate observation of species with topography at a scale broader than the target cell size (e.g., bird's home range can be related to variables around the mogote).

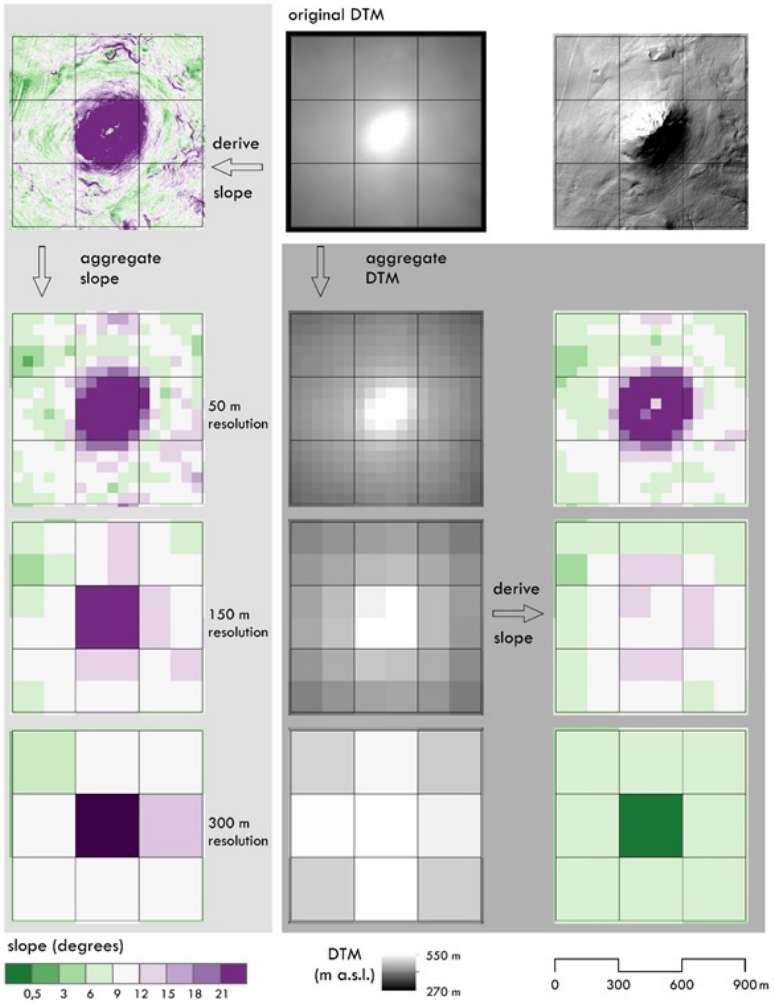

Figure 7.3 Different outcomes of the two scale alteration approaches can be illustrated on an example of slope calculation for an isolated steep-sided hill surrounded by flat plains. The first approach that derive terrain attributes first and then aggregate to coarser resolution yields relatively high slope values along the sides of the hill (left panel). The second approach that first aggregate DTM to coarser resolution and then derive terrain attributes yields lower slope values, especially with coarsening the resolution (right panel). Note how the difference between the two approaches progresses with changes in resolution. The resolution of the original DTM is 2 m, and its mean values were used to aggregate the slope and the original DTM, respectively. Slope was computed using a 3×3 cell neighbourhood with Horn's (1981) algorithm.

7.3 Virtual species experiment

7.3.1 Study area

Here we provide a virtual species example to demonstrate the effects of the substitution of two scale alteration approaches in a controlled environment on SDMs (Moudrý 2015, Gábor et al. 2019, Meynard et al. 2019). Our study area is a real landscape located in north-west Bohemia, Czech Republic (50°32´ N, 13°50´ E) and occupies an area of 35 km2 (Figure 7.4). The DTM for the study area was derived at 1m resolution from airborne LiDAR data collected in May 2017 with an average density of 8 points per square meter (see Moudrý et al. 2019a for details about the flight parameters, point cloud processing, and DTM generation).

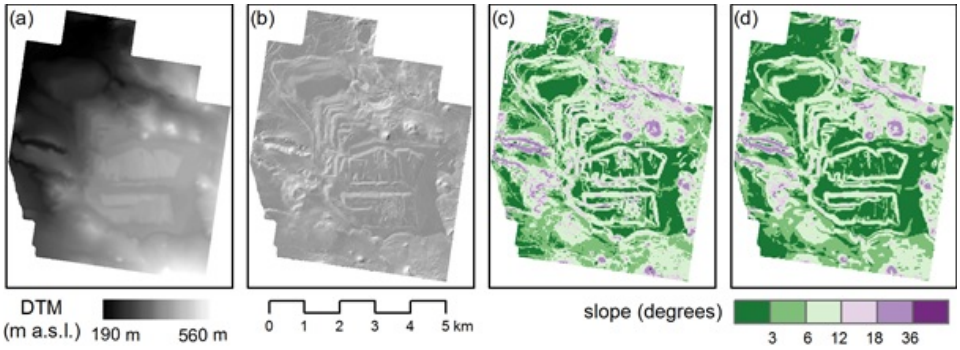

Figure 7.4 Study area. (a) LIDAR-derived DTM; (b) hillshaded terrain; (c) slope calculated at 1m resolution and subsequently aggregated using mean value to 30m resolution; (d) slope calculated from aggregated DTM (mean value) at 30m resolution.

7.3.2 Calculating slope

We used both approaches explained above to calculate slope from the DTM. The first approach (hereafter, Approach 1) calculated the slope at 1m resolution and subsequently aggregated to 30m resolution using

mean value. The other approach (hereafter, Approach 2) consisted of aggregating the DTM to 30m resolution using mean value and subsequently derive slope based on the aggregated DTM. In both instances, we used Horn's (1981) algorithm with the 3×3 cell neighbourhood to calculate the slope.

7.3.3 Virtual species generation

We generated a virtual species and calculated its distribution based on slopes acquired using both approaches using the following procedure. We first simulated a relationship (response function) between a virtual species and slope as a Gaussian response function (mean=8° and standard deviation=4°), which provided a map of environmental suitability. We then adopted a probabilistic simulation approach (logistic function with α= -0.10 and β=0.65) to convert environmental suitability into probabilities of occurrence that were subsequently used to sample presences and absences in each cell (i.e., we made a random draw of presence or absence weighted by the probability of occurrence). Finally, we simulated the sampling of presences and absences of the virtual species at 200 localities (i.e., cells). All simulations were conducted in R v.3.5.1, using the *virtualspecies* package (Leroy et al. 2016).

7.3.4 Model fitting

To illustrate the implications of confusing the two scaling approaches, we ran generalized linear models (GLMs) with slope generated using each of the approaches (resulting in creation of four different models, see Figure 7.5). It is customary to fit Gaussian responses with a polynomial model, so we used GLMs with binomial error distribution and logit link function (McCullagh and Nelder 1989). Both linear and quadratic terms of slope were included because of the known Gaussian shape of the response function.

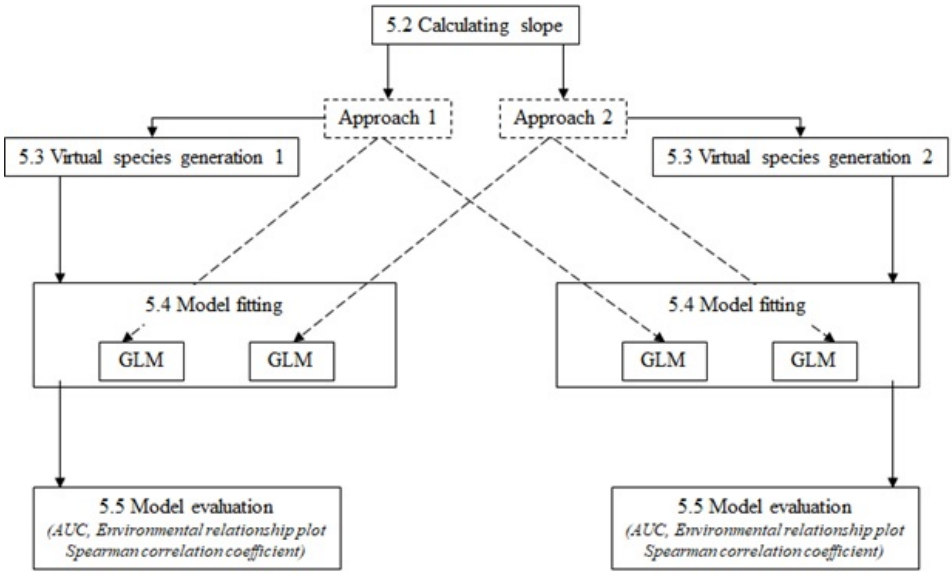

Figure 7.5 The experimental workflow: Virtual species distribution was acquired using slopes calculated from both approaches. Each distribution was subsequently used to create two models using slopes generated by each approach, and the models performances were tested.

7.3.5 Model evaluation

We evaluated each model and its performance For evaluation, we split both datasets of presences-absences acquired in the previous step into test (50%) and training (50%) datasets We evaluated the model calibration by plotting the estimated environmental relationships and their discrimination capacity, assessed by computing the area under the curve (AUC) of the receiver operating characteristic plot (Fielding and Bell 1997). In addition, we calculated Spearman correlation coefficients between the virtual species (i.e., true) and predicted probabilities of occurrence. We ran the entire process from species generation to model evaluation 50 times. Each repetition provided a different presence- absence distribution (Leroy et al. 2016). The entire workflow is illustrated in Figure 7.5.

7.4 Results and discussion

Results of this virtual experiment show considerable differences between fitted GLMs. GLMs based on the same slope calculation approach that was used for creating the virtual species distribution fit the "true" response function perfectly (Figure 7.6). However, when the GLM is based on a slope calculated using the other approach, a significant underestimation of the highest probability of occurrence and overestimation of the lowest probability of occurrence was observed (Figure 7.6), which resulted in a drop in median AUC values (from 0.94 to 0.84 in both cases) and Spearman correlation coefficients (from 0.99 to 0.74). These results show that the approaches are not interchangeable as their mutual substitution can result in a decrease of models' discrimination ability and in misleading spatial predictions of probabilities of occurrence.

7.5 Final remarks

Ecologists and biogeographers are facing an increasingly overwhelming amount and diversity of data and tools, often with little guidance to inform methodological and technical decisions (e.g., Lecours et al. 2017c, Michener and Jones 2012). In this paper, we illustrate that critical information related to scale alteration approach of DTM terrain attributes (e.g., slope) is often undocumented in ecological studies, and we emphasize the effects that two standard scale alteration approaches can have on an analytical output. While these approaches conceptually do the same thing (change data resolution to the same end resolution), they produce two distinct sets of variables and are therefore not interchangeable. In such a context, the data resulting from the alteration used for modelling must capture the environmental variability that influences species distribution (Goodchild 2011), thus matching the

observational and analytical scales to the ecological scale (Lecours et al. 2015). With questions of scale being key in ecology and biogeography (Schneider 2001), more consideration must be given to the impact of scale selection on analyses. Sufficient pre-analyses should be performed before making decisions about altering data resolution and selecting scales.

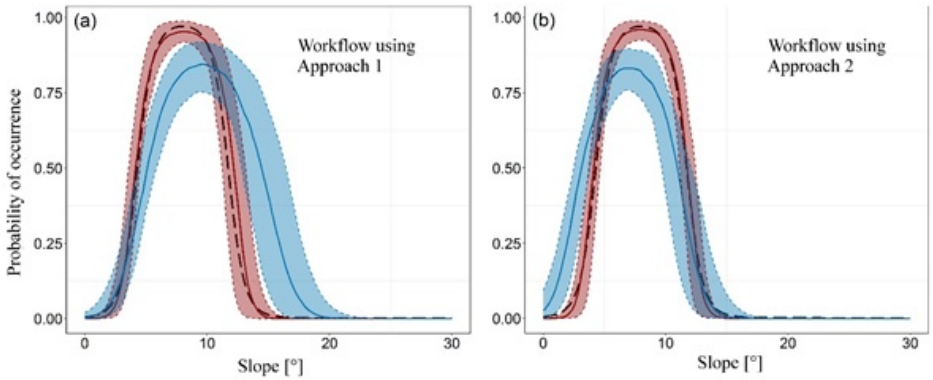

Figure 7.6 Estimation of the probability of virtual species occurrence based on generalized linear models. (a) Workflow using Approach 1, terrain attributes are derived from the DTM at its native resolution, and the resolution of the resulting terrain attributes is then altered (b) Workflow using Approach 2, the DTM native resolution is altered before deriving the terrain attributes (see Fig. 5). Black dashed line shows the "true" probability of occurrence of the virtual species. Red colour represents probabilities resulting from GLM based on the same slope calculation approach that was used for creating the virtual species while blue shows the results of GLM based on a slope calculated using the other approach. The shaded areas represent the regions delimited by the 5th–95th percentiles of the estimated probability of occurrence obtained from 50 simulations.)

In a geomorphological context, Grohmann (2015) compared the two approaches used in our example using 12 levels of aggregation (from 300m to 1850m resolution), and concluded that the two terrain attributes studied, slope and aspect, should be derived from the highest resolution available and then aggregated to a coarser resolution if needed, rather than aggregating the DTM before deriving terrain attributes. However, this may not always be the most appropriate course of action

(see Dolan and Lucieer 2014). We suggest starting from the best data resolution available in order to preserve more topographic details and test the performance of the DTM and its derived terrain attributes over a wide set of scales and using different approaches in pre-analyses. Vaze et al. (2010) showed that a higher-resolution DTM that is coarsened still provides more topographic details than a coarserresolution DTM. Consequently, more representative terrain attributes can also be expected from fine-scale data (e.g., LiDAR) than easily accessible coarser-resolution data (e.g., global elevation datasets; Moudrý et al. 2018), and it may often be beneficial to use a coarsened version of a high-resolution DTM instead of a coarser-resolution DTM. However, we are also aware that this criterion may not always hold in an ecological context (e.g., in SDM implementation), with the acquisition of finer-scale data being not necessarily cost-effective in relation to the species and process to be investigated, in particular when the scales at which specific terrain characteristics drive species distribution are already known. It is essential to consider that a combination of multiple scales is more likely to produce better, more representative models than a single scale (Lecours et al. 2015, Tong et al. 2016, Misiuk et al. 2018). A particular scale alteration approach is not necessarily superior to the others, and the appropriateness of terrain attribute scale will also depend on other non-topographic data used.

We also note that users must be aware of the DTMs they use and how they were produced. Most users often assume that publicly available DTMs are good and produced by valid methods, but not all DTMs are appropriate for all purposes. For instance, many DTMs (e.g., GEBCO 2019) are compiled datasets for which some areas already went through some process of resampling, interpolation, or aggregation, among others. However, because spatial scale is intrinsically related to spatial data quality (e.g., Sofia et al. 2013, Lecours 2017), it may not be adequate to use the finest resolution available to derive terrain attributes for inclusion in SDMs. Once again, tests must be performed during pre-

analyses and users must be on the look-out for different types of errors in the DTM they plan to use to determine whether the finest resolution available is appropriate for their purpose. While Lecours et al. (2017a) found that finer-resolution terrain attributes are more impacted by a poor DTM quality than broader-resolution attributes, Lecours et al. (2017b) established that this trend does not hold when a biological or ecological context is integrated into the analysis, for instance in habitat mapping and species distribution modelling. We recommend acknowledging known issues with data or, when available, focus on data with known provenance such as the LIDAR-derived DTM used in our examples.

7.6 Conclusions

Knowledge is often derived from aggregated information in ecology and biogeography. It is commonplace to superimpose a spatial grid over species occurrence data to summarize them over a defined extent, or to collect such data using a predefined sampling grid (e.g., atlases). Environmental data are often coarsened to match the scale of these data, or in some cases, like in the deep sea, species occurrence data need to be aggregated to match the scale of environmental data. While reductions in computer processing time and data size are among the recognized advantages of coarsening data resolution, it may not always be appropriate from an ecological perspective.

There is a need to replace what has previously been an arbitrary selection of scales and resampling or aggregating methods with an explicit selection of observational scale. Since the observational scale should preferably be grounded in ecological knowledge, the approach should be selected based on how well the representations of environment characteristics capture what is relevant for the species of interest. When this is unclear, it may be appropriate to test multiple methods. We

argue that there is no single correct approach, but rather that the most appropriate one(s) will depend on the objectives of a particular study. In all cases, the methods should be explicit enough to enable replication and meet fundamental norms of science (see conclusions by Goodchild 2011).

The problem of spatial scale is arguably one of the most important in ecology, biogeography, and many other disciplines that use geomorphometry. Because of its complexity, it is essential to appropriately address scale questions and generalize findings to infer ecological patterns and processes appropriately. We emphasize the importance of consistent choice of variables, of careful consideration of scales relative to the organisms being studied as well as resampling techniques utilized, and of the explicit and detailed description of methods and decisions made. We recommend that authors explicitly report (i) the source and the resolution of the original DTM(s) they use; (ii) the algorithm used to calculate terrain attributes; (iii) the method used for rescaling (e.g., aggregating or resampling, using the mean or maximum values); (iv) the order in which these operations were performed; and (v) a summary of the pre-analyses that informed these choices.

Chapter 8

Discussion and Summary

Thanks to the ever-developing computing technologies, geographic information systems, remote sensing and various statistical methods, SDMs became a powerful tool often used for exploring species-environment relationships. However, despite its development during the last two decades, the knowledge about how different quality of spatial data affects species distribution models is insufficient (Araújo et al. 2019). This dissertation offers additional insights into this research gap; nonetheless, we are still far from a full understanding of this problem.

The introduced dissertation described how different quality of spatial data and its processing affects species distribution models. This thesis includes research on both species and environmental data and helps to fill the research gap in the field of species distribution modelling. Specifically, the titles of the studies are as follows:

Study I – How do species and data characteristics affect species distribution models and when to use environmental filtering?

Study II – The effect of positional error on fine scale species distribution models increases for specialist species

Study III – On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs

Study IV – Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies

As full discussion and conclusions are included in the respective previous chapters, they will not be repeated here. Instead, this part aims to provide links among the individual presented studies, major points worth highlighting, suggestions for future research and an afterword.

8.1 Influence of various species characteristics and species data quality on SDMs

The fact that the performance of SDMs is complicated by various spatial (e.g. prevalence or range size) and ecological (e.g. niche breadth) characteristics of the studied species is well known (see for example McPherson and Jetz 2007) and has been repeatedly assessed (Foody 2011, Tessarolo et al. 2014, Connor et al. 2018, Liu et al. 2019). However, prior studies usually used a combination of two or three characteristics. The uniqueness of **Study I** lies in the use of combinations of five species characteristics in the same study. Specifically, different sample sizes (with and without sampling bias), sampling methods, species prevalence and species responses to environmental gradient were used together. One of the most important conclusions of this study is that there is a high level of interactions among individual characteristics (see Table 4.1) and ignoring this may lead to misleading outcomes and conclusions. For example, when studying sample size, the species prevalence and/or niche breadth should be also considered as it is expected that for species with low prevalence / narrow niche breadth, a smaller sample size could be potentially enough. Therefore, when assessing the influence of varying species data quality, as many of these characteristics as possible should be included into or at least considered in the analysis to ensure that the results are not affected by some hidden factor. For example, Varela et al. (2014) tested whether environmental filtering of species occurrences negatively affected by sampling bias could potentially increase SDMs

performance. They concluded that environmental filtering improves the model performance. Specifically, AUC increased from 0.91 to 0.98 when an environmental filter was applied. However, **Study I** showed that environmental filtering works only in a few specific cases, one of which was coincidentally used in the study by Varela et al. (2014). The experience from **Study I** taught us that the influence of varying data quality also depends on species characteristics led us to the **Study II** where species with different niche breadths (i.e. specialist, intermediate and generalist species) were simulated and studied in interactions with different magnitudes of positional error in species occurrences.

Positional error is one of the most common errors in species occurrence data. Despite this fact, few prior studies focused on this topic and, even more interestingly, they yielded contradictory results. In **Study II**, we followed a suggestion by Osborne and Leitão (2009) who noted that performing a comparative analysis of the effects of positional error in SDMs in relation to species ecological characteristics could be worthwhile and could potentially explain inconsistent results of prior studies. Therefore, **Study II** focused on the effect of positional error in species occurrences on fine-scale SDMs for species with different niche breadths. As expected, results showed that introducing positional error into species occurrence data led to a decrease in the model performance, the magnitude of which varied with the species niche breadth. The same positional error had a greater impact on species with a narrow niche breadth than on those with a wide niche breadth. This could partly explain the aforementioned inconsistent conclusions of prior studies. Interestingly, we also showed that increasing sample size did not mitigate the negative effect of the positional error in species occurrences. This result is in concordance with recent studies, which concluded that increasing the number of species occurrences in aggregated databases is not necessarily useful for SDMs (see for example Bayraktarov et al. 2019) and contradicts the widely accepted assumption that more data provide more reliable models.

A follow-up research should verify whether coarsening the resolution of environmental variables can compensate for the negative effect of positional error as suggested by Moudrý and Šímová (2012). They proposed that spatial resolution should be coarse enough to prevent a shift of species occurrences due to positional error to inappropriate environmental conditions. Nevertheless, Moudrý and Šímová (2012) did not consider the scale of the effect (i.e. the scale at which environmental variables determine species distribution). Additionally, the first results of an ongoing research of mine and my colleagues indicate the opposite, i.e, that coarsening resolution does not compensate for the negative effect of positional error. However, even if it is concluded that it does (compensate), an optimal trade-off between adopted spatial resolution of environmental data and positional accuracy of species occurrences will still have to be found. The reason is that the coarser is the resolution, the higher is the loss of information about the environment (see Figure 8.1), which may lead to misleading results. For example, the distribution of species may be dependent on rare environmental entities such as small or temporary water bodies (that can e.g. form a habitat for birds) that cannot be detected using coarse resolution data but which do exist in the natural world (Moudrý and Šímová 2012, Šímová et al. 2019).

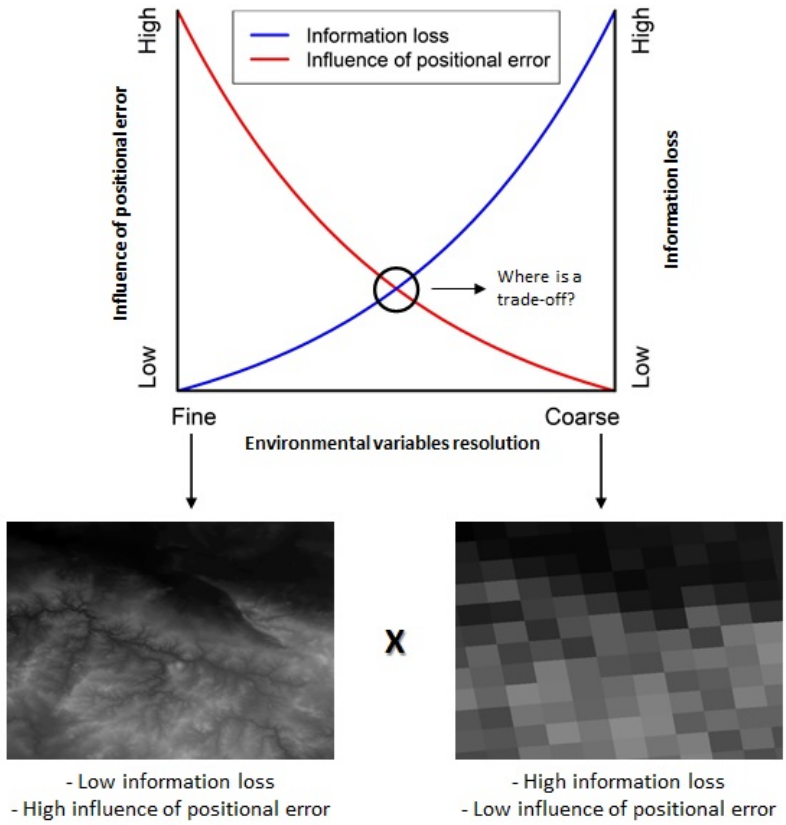

Figure 8.1 *Moudrý and Šímová (2012) proposed that spatial resolution should be coarse enough to prevent a shift of species occurrences due to the positional error to inappropriate environmental conditions. Nevertheless, the coarser is the resolution of environmental data, the higher is the loss of information about the environment. Therefore, if their assumption is confirmed, a trade-off between the adopted spatial resolution of environmental variables and positional accuracy of species occurrences has to be found, which could lead to a major improvement in the current methodological SDMs standards.*

8.2 Influence of various environmental data sources and their processing on SDMs

Environmental data possess the same data quality issues as species data. For example, a shift in environmental data (e.g. in DEMs) is also common (Nuth and Kääb 2011). Despite this, research on the influence of varying quality of environmental data in SDMs studies has remained on the sidelines of scientific interest for a long time (but see for example Seoane et al. 2004, Venier et al. 2004, Osborne and Leitão 2009, Gottschalk et al. 2011). Consequently, some ecologists and biogeographers have been processing and using various environmental data with insufficient background about methodological and technical decisions related to these processes (e.g. with spatial scale alteration; Lecours et al. 2017c). Additionally, a majority of SDMs studies failed to document at all (or just briefly mentioned) how they processed the environmental data they utilized, which has decreased their repeatability.

For instance, in prior ecological studies, SRTM DEM was naively used as a model representing terrain, although SRTM DEM contains also above-the-terrain features such as vegetation or buildings. Therefore, **Study III** aimed to promote the valid and robust use of global DEMs in species distribution modeling by raising awareness about the importance of using DEMs that are corrected for vegetation offset. This issue was addressed only in a few prior studies (e.g. Van Niel et al. 2004, Van Niel and Austin 2007) and so it was not entirely certain how significant are the effects of vertical error in DEMs on SDMs. The results of **Study III** showed that there are considerable differences in model performance between SDMs fitted with DEMs with and without vegetation offset. Hence, at the beginning of the species distribution modelling process, it is necessary to consider the vertical accuracy of DEMs and their derived topographic indices. Otherwise, the resulting models may

provide misleading results, which could lead to incorrect decisions and consequently to misleading conservation actions.

Nowadays, a more precise, fine-resolution global TanDEM-X DEM is available (0.4 arc-second resolution, approximately 12 m on the equator). Therefore, further research is needed because it is expected that in finer resolution, the differences in SDM performance between models fitted with DEMs with and without vegetation offset will be even greater. Additionally, only two selected terrain attributes (i.e. slope, twi) were used in **Study III**. However, it would be helpful to know which terrain attributes are affected by the vertical error of DEMs and which are not.

In SDMs studies, a decision about the spatial resolution of environmental data and (therefore) about the approach to its alteration is, no doubt, one of the most common decisions potentially affecting the performance of resulting models (see for example Šímová et al. 2019). However, as mentioned above, the methodological steps of the spatial resolution alteration are rarely documented. Therefore, **Study IV** aimed to determine which information related to the scale alteration approach should be provided in SDMs studies in order to facilitate their repeatability. Besides, we demonstrated the effects of the methods for altering data resolution on the derived terrain attributes and subsequently on SDMs. The results showed that the way of deriving terrain attributes from DEMs and the step in which the scale is changed, respectively, can result in a decrease of models' discrimination ability and in misleading spatial predictions of probabilities of occurrence. Despite this, the information on the approach used for DTM attributes rescaling often remains undocumented in SDM studies. Therefore, this study is important for future work as it promotes better practice and emphasizes the importance of documenting the approach of scale alteration. This allows to precisely repeat prior studies and facilitates further development of species distribution modeling.

8.3 Virtual species approach

One common thing for all individual studies presented here is the use of the virtual species approach. The assessment of the conceptual and methodological assumptions is common in SDMs literature. However, addressing these questions with real data is challenging (Meynard et al. 2019). For example, specific species characteristics or the scale of the effect relevant for studied species may be unknown. Therefore, the virtual species approach is, mainly in methodological testing, increasingly used (see review by Meynard et al. 2019). This approach enabled to precisely control both species characteristics (species prevalence, niche breadth, response to environmental gradient) and species data quality (sampling bias, sample size and the range of positional error), to combine them with environmental data of varying quality and to test the influence of these combinations on SDMs.

Our results indicate that in all future SDM studies, it is imperative to properly document and share scripts used for virtual species generation as highlighted by Meynard et al. (2019) because it makes the experiments easily repeatable and their conclusions verifiable. The methodological studies using virtual species approach proved this to be very important. Great examples were presented by Meynard et al. (2019). Guillera-Arroita et al. (2014) have repeated the previous study by Thibaud et al. (2014) using the original scripts (concluding that MAXENT often outperformed the other SDMs models) and they showed that conclusions by Thibaud et al. (2014) resulted from an artifact of the inappropriate use of virtual species.

In addition, the virtual species approach could be used as a convincing evidence of some issues related to SDMs. For example, it was hard to convince reviewers that **Study IV** represents an important contribution to the field of SDM. In the beginning, **Study IV** did not include the virtual species example and failed to be sufficiently convincing. Most of the reviewers thought that the scale alteration approach selection is not

a big issue or that its influence is negligible and that documenting it is not crucial. After a few unsuccessful submissions, we added the virtual species approach. Being able to illustrate the possible consequences with virtual species demonstrated the value of **Study IV** and it went through more smoothly.

8.4 Conclusions

As detailed conclusions can be found in previous chapters, only some major points are provided here:

- When modeling species distribution, there are many interactions among individual spatial and ecological species characteristics; ignoring this may cause misleading outcomes and conclusions. Therefore, as many of these characteristics as possible should be analyzed simultaneously as it could help understand interactions between various ecological characteristics of the studied species and different data quality.
  - For instance, when studying sample size, species prevalence and/or niche breadth should also be considered as it is expected that for species with low prevalence / narrow niche breadth, a smaller sample size could potentially suffice. On the other hand, such species will be more affected by erroneous data (e.g. by positional error).
- SDMs are affected by a variable quality of both species and environmental spatial data; species occurrences and/or environmental data of high accuracy can hardly be replaced by those of lower accuracy. Specifically:
  - Data accuracy is crucial as considerable differences in model performance between accurate and inaccurate data have been demonstrated. Interestingly, our results showed that

the negative effects of inaccurate data entering SDMs cannot be mitigated by increasing the sample size. The takeaway message is that improving species data accuracy appears to be more effective than increasing sample size.

  - The environmental data source, their processing and resolution are important in SDMs. Therefore, future studies should explicitly report the source and resolution of the original environmental data (not only DTMs), the algorithm used to calculate terrain attributes, the method used for rescaling, the order in which these operations were performed and, of course, a summary of the pre-analyses that resulted in these choices.

- Based on our results, it is advocated for not using data with an unknown level of accuracy as such models may lead to misleading outcomes and conclusions and to improper conservation actions.

8.5 Further research

Future research should be focused on the influence of data uncertainty using more complex virtual species and, consequently, real species data across various resolutions of environmental variables and extents of the study areas. Such studies will have the potential to further improve the use of SDMs in biodiversity monitoring and conservation. For example, at present, I participate in research on assessing the influence of the positional error in species occurrence data using real bird data (downloaded from eBird database) in a continental extent across varying resolutions of environmental variables. Based on results presented in chapters 6 and 7, future research should always consider the source of the utilized environmental data, be consistent in processing environmental data and always take into account both the scale of the

species response to the environmental gradient and the resolution of environmental data.

8.6 Afterword

Since 1960s, explanatory modeling (species-environment relationship modeling) has come a long way, which gave rise to a whole new field we call species distribution modeling. At the beginning, SDM studies on methodological aspects mostly focused on the selection of the appropriate modeling techniques (algorithms, see for example Elith et al. 2006). With the increasing number of such studies, it, however, became more and more clear that the quality of spatial data plays an important role in modeling processes (Araújo et al. 2019). Nevertheless, there still is a research gap and the knowledge about the influence of varying quality of spatial data on SDMs is not complete. It is remarkable how deeply ecologists have immersed themselves in studying the data issues in SDMs. This only underlines the importance of this topic – I am really curious how far this effort will lead. Will we be able to predict with pinpoint accuracy where the species occur and where not once we fully understand the influence of data quality or once we find methods how to overcome these issues? Is all this effort necessary or is it just a dead end? These are the questions that will be hopefully answered by the next decade(s) of research.

References

Abolmaali, S. M. R. et al. (2018). MaxEnt modeling for predicting suitable habitats and identifying the effects of climate change on a threatened species, Daphne mucronata, in central Iran. *Ecological Informatics*, 43:116–123.

Alba-Sánchez, F. et al. (2010). Past and present potential distribution of the Iberian Abies species: A phytogeographic approach using fossil pollen data and species distribution models. *Diversity and Distributions*, 16(2):214–228.

Allouche, O. et al. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6):1223–1232.

Anderson, R. P. and Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, 37(7):1378–1393.

Araújo, M. B. et al. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1):eaat4858.

Araújo, M. B. and Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10):1677–1688.

Arribas, P. et al. (2012). Dispersal ability rather than ecological tolerance drives differences in range size between lentic and lotic water beetles (Coleoptera: Hydrophilidae). *Journal of Biogeography*, 39(5):984–994.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2-3):101–118.

Austin, M. P. (1980). Searching for a model for use in vegetation analysis. *Vegetatio*, 42(1-3):11–21.

Austin, M. P. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1-2):1–19.

Bakx, T. et al. (2019). Use and categorization of light detection and ranging vegetation metrics in avian diversity and species distribution research. *Diversity and Distributions*, 25(7):1045–1059.

Barbet-Massin, M. et al. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 3(2):327–338.

Barry, S. and Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43(3):413–423.

Battini, N. et al. (2019). Staying ahead of invaders: using species distribution modeling to predict alien species' potential niche shifts. *Marine Ecology Progress Series*, 612:127–140.

Bayraktarov, E. et al. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, 6:239.

Bazzichetto, M. et al. (2018a). Modeling plant invasion on Mediterranean coastal landscapes: An integrative approach using remotely sensed data. *Landscape and Urban Planning*, 171:98–106.

Bazzichetto, M. et al. (2018b). Plant invasion risk: A quest for invasive species distribution modelling in managing protected areas. *Ecological Indicators*, 95:311–319.

Beaumont, L. J. et al. (2016). Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? *Ecological Modelling*, 342:135–146.

Beck, J. et al. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15.

Beltrán, B. J. et al. (2014). Effects of climate change and urban development on the distribution and conservation of vegetation in a Mediterranean type ecosystem. *International Journal of Geographical Information Science*, 28(8):1561–1589.

Besnard, A. G. et al. (2013). Topographic wetness index predicts the occurrence of bird species in floodplains. *Diversity and Distributions*, 19(8):955–963.

Beven, K. J. and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24(1):43–69.

Bino, G. et al. (2014). Identifying minimal sets of survey techniques for multi-species monitoring across landscapes: an approach utilising species distribution models. *International Journal of Geographical Information Science*, 28(8):1674–1708.

Boakes, E. et al. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, 8(6):e1000385.

Boria, R. A. et al. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–77.

Bouchet, P. J. et al. (2015). Topographic determinants of mobile vertebrate predator hotspots: Current knowledge and future directions. *Biological Reviews*, 90(3):699–728.

Boulangeat, I. et al. (2012). Niche breadth, rarity and ecological characteristics within a regional flora spanning large environmental gradients. *Journal of Biogeography*, 39(1):204–214.

Bradie, J. and Leung, B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, 44(6):1344–1361.

Brambilla, M. and Ficetola, G. F. (2012). Species distribution models as a tool to estimate reproductive parameters: A case study with a passerine bird species. *Journal of Animal Ecology*, 81(4):781–787.

Breiner, F. T. et al. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, 6(10):1210–1218.

Breiner, F. T. et al. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, 9(4):802–808.

Brito, J. et al. (2009). Biogeography and conservation of taxa from remote regions: An application of ecological-niche based models and GIS to North-African canids. *Biological Conservation*, 142(12):3020–3029.

Brito, J. C. et al. (2011). Biogeography and conservation of viperids from North-West Africa: An application of ecological niche-based models and GIS. *Journal of Arid Environments*, 75(11):1029–1037.

Brown, J. H. (1984). On the relationship beween abundance and distribution of species. *American Naturalist*, 124(2):255–279.

Brown, S. (1981). A Comparison of the Structure, Primary Productivity, and Transpiration of Cypress Ecosystems in Florida. *Ecological Monographs*, 51(4):403–427.

Bulluck, L. et al. (2006). Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. *Global Ecology and Biogeography*, 15(1):27–38.

Chefaoui, R. M. et al. (2011). Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, 34(2):229–247.

Chen, F. and Li, X. (2016). Evaluation of IMERG and TRMM 3B43 monthly precipitation products over mainland China. *Remote Sensing*, 8(6):1–18.

Chen, Y. C. et al. (2019). The niches of nuthatches affect their lineage evolution differently across latitude. *Molecular Ecology*, 28(4):803–817.

Cianfrani, C. et al. (2018). More than range exposure: Global otter vulnerability to climate change. *Biological conservation*, 221:103–113.

Connor, T. et al. (2018). Effects of grain size and niche breadth on species distribution modeling. *Ecography*, 41(8):1270–1282.

Conrad, O. (2003). Module topographic wetness index (SAGA).

Conrad, O. et al. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, 8(7):1991–2007.

Convertino, M. et al. (2012). Epistemic uncertainty in predicting shorebird biogeography affected by sea-level rise. *Ecological Modelling*, 240:1–15.

Coops, N. C. et al. (2010). Assessing the utility of lidar remote sensing technology to identify mule deer winter habitat. *Canadian Journal of Remote Sensing*, 36(2):81–88.

Cord, A. F. et al. (2013). Modelling species distributions with remote sensing data: Bridging disciplinary perspectives. *Journal of Biogeography*, 40(12):2226–2227.

Davies, A. and Asner, G. (2014). Advances in animal ecology from 3D-LiDAR ecosystem mapping. *Trends in ecology and evolution*, 29:681–691.

Della Rocca, F. et al. (2019). Identifying hotspots for rare species under climate change scenarios: improving saproxylic beetle conservation in Italy. *Biodiversity and Conservation*, 28(2):433–449.

Deng, Y. et al. (2007). DEM resolution dependencies of terrain attributes across a landscape. *International Journal of Geographical Information Science*, 21(2):187–213.

Devictor, V. et al. (2010). Defining and measuring ecological specialization. *Journal of Applied Ecology*, 47(1):15–25.

DiMiceli, C. et al. (2017). Annual global automated MODIS vegetation continuous fields (MOD44B) at 250 m spatial resolution for data years beginning day 65, 2000-2010.

Dolan, M. and Lucieer, V. (2014). Variation and Uncertainty in Bathymetric Slope Calculations Using Geographic Information Systems. *Marine Geodesy*, 37(2):187–219.

Dolan, M. F. J. (2012). Calculation of slope angle from bathymetry data using GIS-effects of computation algorithms, data resolution and analysis scale. *NGU Report*.

Domisch, S. et al. (2015). Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. *Fundamental and Applied Limnology*, 186(1-2):45–61.

Duan, R. Y. et al. (2015). SDMvspecies: A software for creating virtual species for species distribution modelling. *Ecography*, 38(1):108–110.

Dufresnes, C. et al. (2020). Are glacial refugia hotspots of speciation and cytonuclear discordances? Answers from the genomic phylogeography of Spanish common frogs. *Molecular Ecology*, 29(5):986–1000.

Dungan, J. L. et al. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, 25(5):626–640.

Dunlavy, J. (1935). Studies on the Phyto-Vertical Distribution of Birds. *JSTOR*, 52(4):425–431.

Duputié, A. et al. (2014). Where are the wild things? Why we need better data on species distribution. *Global Ecology and Biogeography*, 23(4):457–467.

Dvorský, M. et al. (2017). Niche asymmetry of vascular plants increases with elevation. *Journal of Biogeography*, 44(6):1418–1425.

Eaton, S. et al. (2018). Adding small species to the big picture: Species distribution modelling in an age of landscape scale conservation. *Biological Conservation*, 217:251–258.

Elith, J. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151.

Elith, J. and Graham, C. H. (2009). Do they? How do they? WHY do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, 32(1):66–77.

Elith, J. and Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):677–697.

Escalera-Vázquez, L. H. et al. (2018). Predicting Ambystoma ordinarium Habitat in Central Mexico Using Species Distribution Models . *Herpetologica*, 74(2):117–126.

Evangelista, P. H. et al. (2008). Modelling invasion for a habitat generalist and a specialist plant species. *Diversity and Distributions*, 14(5):808–817.

Farr, T. G. et al. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2):RG2004.

Fernandes, R. F. et al. (2018). How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. *Ecological Informatics*, 48:125–134.

Fernandes, R. F. et al. (2019). Effects of simulated observation errors on the performance of species distribution models. *Diversity and Distributions*, 25(3):400–413.

Fernandez, M. A. et al. (2009). Locality uncertainty and the differential performance of four common niche-based modeling techniques. *Biodiversity Informatics*, 6:36–52.

Ferrier, S. et al. (2017). Biodiversity modelling as part of an observation system. In Walters, M. and Scholes, R. J., editors, *The GEO Handbook on Biodiversity Observation Networks.*, pages 239–257. Springer International Publishing, michele wa edition.

Ficetola, G. F. et al. (2014). How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. *International Journal of Geographical Information Science*, 28(8):1723–1739.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1):38–49.

Filer, A. et al. (2020). Distribution mapping of specialized amphibian species in rare, ephemeral habitats: Implications for the conservation of threatened "acid" frogs in south-east Queensland. *Conservation Science and Practice*, 2(1):e143.

Fogl, M. and Moudrý, V. (2016). Influence of vegetation canopies on solar potential in urban environments. *Applied Geography*, 66:73–80.

Foody, G. M. (2011). Impacts of imperfect reference data on the apparent accuracy of species presence-absence models and their predictions. *Global Ecology and Biogeography*, 20(3):498–508.

Fourcade, Y. et al. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLOS ONE*, 9(5):1–13.

Fourcade, Y. et al. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, 27(2):245–256.

Frair, J. L. et al. (2010). Resolving issues of imprecise and habitat-biased locations in ecological analyses using GPS telemetry data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1550):2187–2200.

Franklin, J. (1995). Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, 19(4):474–499.

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge.

Franklin, J. et al. (2014). Linking spatially explicit species distribution and population models to plan for the persistence of plant species under global change. *Environmental Conservation*, 41(2):97–109.

Friedrichs-Manthey, M. et al. (2020). From topography to hydrology—The modifiable area unit problem impacts freshwater species distribution models. *Ecology and Evolution*, 10(6):2956–2968.

Futuyma, D. J. and Moreno, G. (1988). The evolution of ecological specialization. *Annual review of ecology and systematics. Vol. 19*, pages 207–233.

Gábor, L. et al. (2019). How do species and data characteristics affect species distribution models and when to use environmental filtering? *International Journal of Geographical Information Science*, pages 1–18.

Gábor, L. et al. (2020). The effect of positional error on fine scale species distribution models increases for specialist species. *Ecography*, 43(2):256–269.

Gallant, J. C. and Read, A. M. (2016). A near-global bare-Earth Dem from SRTM. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 41, pages 137–141. International Society for Photogrammetry and Remote Sensing.

Gallant, J. C. and Wilson, J. (2000). Primary topographic attributes. In Wilson, J. and Gallant, J., editors, *Terrain analysis: Principles and Applications*, chapter Primary topographic attributes, pages 51–96. Wiley, New York.

Gaston, K. et al. (1997). Interspecific abundance-range size relationships: an appraisal of mechanisms. *Journal of Animal Ecology*, pages 579–601.

GEBCO (2019). GEBCO. Technical report, British Oceanographic Data Centre.

Geldmann, J. et al. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11):1139–1149.

Geller, G. et al. (2017). Remote sensing for biodiversity. In Walters, M. and Scholes, R., editors, *The GEO Handbook on Biodiversity Observation Networks*, pages 187–210. Springer International Publishing.

Gibbons, D. W. et al. (2007). Mapping avian distributions: The evolution of bird atlases. *Bird Study*, 54(3):324–334.

Gillard, M. et al. (2017). Present and future distribution of three aquatic plants taxa across the world: decrease in native and increase in invasive ranges. *Biological Invasions*, 19(7):2159–2170.

Goodchild, M. (1996). The spatial data infrastructure of environmental modellin. *GIS and environmental modelling*, pages 11–15.

Goodchild, M. (2006). *Fundamentals of Spatial Data Quality*. ISTE, London.

Goodchild, M. F. (2011). Scale in GIS: An overview. *Geomorphology*, 130(1-2):5–9.

Gottschalk, T. K. et al. (2011). Influence of grain size on species-habitat models. *Ecological Modelling*, 222(18):3403–3412.

Graham, C. H. et al. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45(1):239–247.

Grinnell, J. (1917). The Niche-Relationships of the California Thrasher. *JSTOR*, 34(4):427–433.

Grohmann, C. H. (2015). Effects of spatial resolution on slope and aspect derivation for regional-scale analysis. *Computers and Geosciences*, 77:111–117.

Guan, B. et al. (2020). Shifting ranges of eleven invasive alien plants in China in the face of climate change. *Ecological Informatics*, 55(October 2019):101024.

Guillera-Arroita, G. et al. (2014). Maxent is not a presence-absence method: A comment on Thibaud et al. *Methods in Ecology and Evolution*, 5(11):1192–1197.

Guillera-Arroita, G. et al. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24(3):276–292.

Guisan, A. et al. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, 77(4):615–630.

Guisan, A. and Hofer, U. (2003). Predicting reptile distributions at the mesoscale: Relation to climate and topography. *Journal of Biogeography*, 30(8):1233–1243.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186.

Hairston, N. G. (1959). Species Abundance and Community Organization. *Ecology*, 40(3):404–416.

Hallgren, W. et al. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, 408:108719.

Hansen, M. C. et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396(6706):41–49.

Hengl, T. and Reuter, H. (2009). Geomorphometry: concepts, software, applications. *Developments in Soil Science*, 33:765.

Heuvelink, G. (1998). *Error propagation in environmental modelling with GIS*. CRC press.

Hijmans, R. J. (2012). Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, 93(3):679–688.

Hijmans, R. J. et al. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Hijmans, R. J. et al. (2012). "Package 'dismo'. Species distribution modeling.

Hirzel, A. H. et al. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2-3):111–121.

Holloway, P. et al. (2016). Incorporating movement in species distribution models: how do simulations of dispersal affect the accuracy and uncertainty of projections? *International Journal of Geographical Information Science*, 30(10):2050–2074.

Honrado, J. P. et al. (2016). Fostering integration between biodiversity monitoring and modelling. *Journal of Applied Ecology*, 53(5):1299–1304.

Horn, B. K. (1981). Hill Shading and the Reflectance Map. *Proceedings of the IEEE*, 69(1):14–47.

Hsu, R. C. et al. (2012). Simulating climate change impacts on forests and associated vascular epiphytes in a subtropical island of East Asia. *Diversity and Distributions*, 18(4):334–347.

Humbold von, A. and Bonpland, A. (1807). *Essai sur la géographie des plante*. Schoel & Co., Lyon.

Hutchinson, G. E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding remarks*, 22:415–427.

Hutchinson, G. E. (1961). The Paradox of the Plankton. *The American Naturalist*, 95(882):137–145.

Irl, S. D. et al. (2015). Climate vs. topography - spatial patterns of plant species diversity and endemism on a high-elevation island. *Journal of Ecology*, 103(6):1621–1633.

Isaac, N. J. et al. (2020). Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology and Evolution*, 35(1):56–67.

Isaac, N. J. and Pocock, M. J. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3):522–531.

Jarnevich, C. S. et al. (2015). Caveats for correlative species distribution modeling. *Ecological Informatics*, 29:6–15.

Jetz, W. et al. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology and Evolution*, 27(3):151–159.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4):498–507.

Jiménez-Valverde, A. et al. (2009). The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, 10(2):196–205.

Jiménez-Valverde, A. et al. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14(6):885–890.

Johnson, C. J. and Gillingham, M. P. (2008). Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. *Ecological Modelling*, 213(2):143–155.

Khosravipour, A. et al. (2016). Generating spike-free digital surface models using LiDAR raw point clouds: A new approach for forestry applications. *International journal of applied earth observation and geoinformation*, 52:104–114.

Kindsvater, H. K. et al. (2018). Overcoming the Data Crisis in Biodiversity Conservation. *Trends in Ecology and Evolution*, 33(9):676–688.

Kopecký, M. and Čížková, Š. (2010). Using topographic wetness index in vegetation ecology: Does the algorithm matter? *Applied Vegetation Science*, 13(4):450–459.

Kosicki, J. Z. (2017). Should topographic metrics be considered when predicting species density of birds on a large geographical scale? A case of Random Forest approach. *Ecological Modelling*, 349:76–85.

Kramer-Schadt, S. et al. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11):1366–1379.

Krojerova-Prokesova, J. et al. (2008). Species richness of vertebrates in the Czech Republic. *Folia Zoologica*, 57(4):452–464.

LAStools (2017). LAStools.

Lechner, A. M. et al. (2012a). Are landscape ecologists addressing uncertainty in their remote sensing data? *Landscape Ecology*, 27(9):1249–1261.

Lechner, A. M. et al. (2012b). Investigating species-environment relationships at multiple scales: Differentiating between intrinsic scale

and the modifiable areal unit problem. *Ecological Complexity*, 11:91–102.

Lecours, V. (2017). On the use of maps and models in conservation and resource management (Warning: Results may vary). *Frontiers in Marine Science*, 4(288):1–18.

Lecours, V. et al. (2015). Spatial scale and geographic context in benthic habitat mapping: Review and future directions. *Marine Ecology Progress Series*, 535:259–284.

Lecours, V. et al. (2016). A review of marine geomorphometry, the quantitative study of the seafloor. *Hydrology and Earth System Sciences*, 20(8):3207–3244.

Lecours, V. et al. (2017a). Artefacts in marine digital terrain models: A multiscale analysis of their impact on the derivation of terrain attributes. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5391–5406.

Lecours, V. et al. (2017b). Influence of artefacts in marine digital terrain models on habitat maps and species distribution models: a multiscale assessment. *Remote Sensing in Ecology and Conservation*, 3(4):232–246.

Lecours, V. et al. (2017c). Towards a framework for terrain attribute selection in environmental studies. *Environmental Modelling and Software*, 89:19–30.

Leempoel, K. et al. (2015). Very high-resolution digital elevation models: Are multi-scale derived variables ecologically relevant? *Methods in Ecology and Evolution*, 6(12):1373–1383.

Lefsky, M. et al. (2002). Lidar remote sensing for ecosystem studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies. *BioScience*, 52:19–30.

Leitão, P. J. et al. (2011). Effects of geographical data sampling bias on habitat models of species distributions: A case study with steppe birds in southern Portugal. *International Journal of Geographical Information Science*, 25(3):439–454.

Leroy, B. et al. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, 39(6):599–607.

Leroy, B. et al. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, 45(9):1994–2002.

Levin, S. A. (1992). The problem of pattern and scale in ecology. *Ecology*, 73(6):1943–1967.

Li, H. et al. (2010). Strawberry plant fruiting efficiency and its correlation with solar irradiance, temperature and reflectance water index variation. *Environmental and Experimental Botany*, 68(2):165–174.

Linda, R. et al. (2016). Developing a criterion for distinguishing tetraploid birch species from diploid and modelling their potential distribution on the Czech Republic. In Kacálek, D., Novák, J., Nováková, K., and Součková, J., editors, *Proceedings of Central European Silviculture*, pages 71–78, Dobruška. Výzkumný ústav lesního hospodářství a myslivosti,.

Liu, C. et al. (2019). The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, 42(3):535–548.

Lobo, J. M. (2008). More complex distribution models or more representative data? *Biodiversity Informatics*, 5:14–19.

Lobo, J. M. and Tognelli, M. F. (2011). Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, 19(1):1–7.

Loiselle, B. A. et al. (2008). Predicting species distributions from herbarium collections: Does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, 35(1):105–116.

Lowen, J. B. et al. (2016). Effects of spatial resolution on predicting the distribution of aquatic invasive species in nearshore marine environments. *Marine Ecology Progress Series*, 556:17–30.

Lucieer, V. et al. (2018). Charting the course for future developments in marine geomorphometry: An introduction to the special issue. *Geosciences*, 8(12):1–9.

Luoto, M. et al. (2005). Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, 14(6):575–584.

MacArthur, R. (1960). On the Relative Abundance of Species. *The American Naturalist*, 94(874):25–36.

MacArthur, R. H. and MacArthur, J. W. (1961). On Bird Species Diversity. *Ecology*, 42(3):594–598.

Macharia, J. M. et al. (2020). Comparison of satellite remote sensing derived precipitation estimates and observed data in Kenya. *Agricultural and Forest Meteorology*, 284:107875.

Mackinnon, J. G. and White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of econometrics*, 29(3):305–325.

Mahecha, M. et al. (2017). Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences*, 14(18):4255–4277.

Malavasi, M. et al. (2018). Plant invasions in Italy: An integrative approach using the European LifeWatch infrastructure database. *Ecological indicators*, 91:182–188.

Manel, S. et al. (2001). Evaluating presence-absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

Manzoor, S. A. et al. (2020). Evidence of ecological niche shift in Rhododendron ponticum (L.) in Britain: Hybridization as a possible cause of rapid niche expansion. *Ecology and Evolution*, pages 2040–2050.

Martínez-Gutiérrez, P. G. et al. (2018). Niche centrality and human influence predict rangewide variation in population abundance of a widespread mammal: The collared peccary (Pecari tajacu). *Diversity and Distributions*, 24(1):103–115.

Matawa, F. et al. (2016). Modelling the spatial-temporal distribution of tsetse (Glossina pallidipes) as a function of topography and vegetation greenness in the Zambezi Valley of Zimbabwe. *Applied Geography*, 76:198–206.

Mateo Sánchez, M. C. et al. (2014). Scale dependence in habitat selection: the case of the endangered brown bear (Ursus arctos) in the Cantabrian Range (NW Spain). *International Journal of Geographical Information Science*, 28(8):1531–1546.

Mateo-Tomás, P. and Olea, P. P. (2015). Livestock-driven land use change to model species distributions: Egyptian vulture as a case study. *Ecological Indicators*, 57:331–340.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman, London.

McCune, J. L. (2019). A new record of Stylophorum diphyllum (Michx.) Nutt. in Canada: A case study of the value and limitations of building species distribution models for very rare plants. *The Journal of the Torrey Botanical Society*, 146(2):119.

McPherson, J. M. and Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography*, 30(1):135–151.

Meineri, E. et al. (2015). Using Gaussian Bayesian Networks to disentangle direct and indirect associations between landscape physiography, environmental variables and species distribution. *Ecological Modelling*, 313:127–136.

Mertes, K. and Jetz, W. (2018). Disentangling scale dependencies in species environmental niches and distributions. *Ecography*, 41(10):1604–1615.

Meynard, C. N. et al. (2019). Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? *Ecography*, 42(12):2021–2036.

Meynard, C. N. and Kaplan, D. M. (2012). The effect of a gradual response to the environment on species distribution modeling performance. *Ecography*, 35(6):499–509.

Meynard, C. N. and Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, 40(1):1–8.

Michener, W. K. and Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, 27(2):85–93.

Miller, J. A. (2014). Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, 38(1):117–128.

Misiuk, B. et al. (2018). A multiscale approach to mapping seabed sediments. *PLOS ONE*, 13(2):e0193647.

Mitchell, P. J. et al. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution*, 8(1):12–21.

Mod, H. K. et al. (2016). What we use is not what we know: environmental predictors in plant distribution models. *Journal of Vegetation Science*, 27(6):1308–1322.

Mohamedou, C. et al. (2017). LiDAR-based TWI and terrain attributes in improving parametric predictor for tree growth in southeast Finland. *International Journal of Applied Earth Observation and Geoinformation*, 62:183–191.

Moudrý, V. (2015). Modelling species distributions with simulated virtual species. *Journal of Biogeography*, 42(8):1365–1366.

Moudrý, V. and Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics*, 56.

Moudrý, V. et al. (2017). Which breeding bird categories should we use in models of species distribution? *Ecological Indicators*, 74:526–529.

Moudrý, V. et al. (2018). On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs. *Ecological Modelling*, 383:3–9.

Moudrý, V. et al. (2019a). Comparison of leaf-off and leaf-on combined UAV imagery and airborne LiDAR for assessment of a post-mining site terrain and vegetation structure: Prospects for monitoring hazards and restoration success. *Applied Geography*, 104:32–41.

Moudrý, V. et al. (2019b). Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies. *Ecological Informatics*, 54:100987.

Moudrý, V. and Šímová, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: A review. *International Journal of Geographical Information Science*, 26(11):2083–2095.

Moudrý, V. and Šímová, P. (2013). Relative importance of climate, topography, and habitats for breeding wetland birds with different latitudinal distributions in the Czech Republic. *Applied Geography*, 44:165–171.

Naimi, B. et al. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, 38(8):1497–1509.

Naumann, G. et al. (2012). Monitoring Drought Conditions and Their Uncertainties in Africa Using TRMM Data. *Journal of Applied Meteorology and Climatology*, 51(10):1867–1874.

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models. In *Encyclopedia of statistical sciences*. Wiley, New York.

Nelson, A. et al. (2009). DEM production methods and sources. *Developments in Soil Science*, 33:65–85.

Neyman, J. and Scott, E. (1959). Stochastic Models of Population Dynamics. *Science*, 130:303–308.

Nezer, O. et al. (2017). High-resolution species-distribution model based on systematic sampling and indirect observations. *Biodiversity and Conservation*, 26(2):421–437.

Norberg, A. et al. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):e01370.

Nuth, C. and Kääb, A. (2011). Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *Cryosphere*, 5(1):271–290.

OFT (2002). Digital Height Model Level 2, Product Information. *Federal Office of Topography*.

Oksanen, J. and Minchin, P. R. (2002). Continuum theory revisited: What shape are species responses along ecological gradients? *Ecological Modelling*, 157(2-3):119–129.

Oksanen, J. and Sarjakoski, T. (2005). Error propagation of DEM-based surface derivatives. *Computers and Geosciences*, 31(8):1015–1027.

O'Loughlin, F. E. et al. (2016). A multi-sensor approach towards a global vegetation corrected SRTM DEM product. *Remote Sensing of Environment*, 182:49–59.

Osborne, P. E. and Leitão, P. J. (2009). Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, 15(4):671–681.

Pearce, J. and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133(3):225–245.

Pearson, R. G. et al. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34(1):102–117.

Peers, M. et al. (2012). Reconsidering the specialist-generalist paradigm in niche breadth dynamics: resource gradient selection by Canada lynx and bobcat. *PLOS ONE*, 7(12):e51488.

Peterson, A. T. and Soberón, J. (2012). Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservação*, 10(2):102–107.

Peterson, A. T. and Soberón, J. (2018). Essential biodiversity variables are not global. *Biodiversity and Conservation*, 27(5):1277–1288.

Phillips, S. et al. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259.

Phillips, S. J. et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.

Pike, R. J. et al. (2009). Geomorphometry: A brief guide. In Hengl, T. and Reuterm, H., editors, *Geomorphometry: Concepts, Software, Applications*, pages 3–30. Elsevier, Amsterdam.

Piroddi, C. et al. (2015). Using ecological models to assess ecosystem status in support of the European Marine Strategy Framework Directive. *Ecological Indicators*, 58:175–191.

Pradervand, J. N. et al. (2014). Very high resolution environmental predictors in species distribution models: Moving beyond topography? *Progress in Physical Geography*, 38(1):79–96.

Préau, C. et al. (2020). Niche modelling to guide conservation actions in france for the endangered crayfish austropotamobius pallipes in relation to the invasive pacifastacus leniusculus. *Freshwater Biology*, 65(2):304–315.

Proosdij, v. A. S. et al. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39(6):542–552.

Qiao, H. et al. (2015). Marble Algorithm: A solution to estimating ecological niches from presence-only records. *Scientific Reports*, 5(1):1–10.

Qiao, H. et al. (2016). NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, 39(8):805–813.

Quinn, P. et al. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes*, 5(1):59–79.

Raduła, M. W. et al. (2018). Topographic wetness index explains soil moisture better than bioindication with Ellenberg's indicator values. *Ecological Indicators*, 85:172–179.

Ranc, N. et al. (2016). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, 40(9):1076–1087.

Randin, C. F. et al. (2020). Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models. *Remote Sensing of Environment*, 239:111626.

Rattray, A. et al. (2014). Quantification of Spatial and Thematic Uncertainty in the Application of Underwater Video for Benthic Habitat Mapping. *Marine Geodesy*, 37(3):315–336.

Reddy, S. and Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30(11):1719–1727.

Reif, J. et al. (2018). Competition-driven niche segregation on a landscape scale: Evidence for escaping from syntopy towards allotopy in two coexisting sibling passerine species. *Journal of Animal Ecology*, 87(3):774–789.

Reino, L. et al. (2013). Does local habitat fragmentation affect large-scale distributions? The case of a specialist grassland bird. *Diversity and Distributions*, 19(4):423–432.

Renkonen, O. (1938). Statistisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore. *Annales Zoologici*, 6:1–231.

Reside, A. E. et al. (2011). Incorporating low-resolution historic species location data decreases performance of distribution models. *Ecological Modelling*, 222(18):3444–3448.

Robinson, L. M. et al. (2011). Pushing the limits in marine species distribution modelling: Lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6):789–802.

Rocchini, D. et al. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of. *Progress in Physical Geography*, 35(2):211–226.

Rocchini, D. et al. (2017). Spatio-ecological complexity measures in GRASS GIS. *Computers and Geosciences*, 104:166–176.

Rocchini, D. and Neteler, M. (2012). Let the four freedoms paradigm apply to ecology. *Trends in Ecology and Evolution*, 27(6):310–311.

Rödder, D. and Engler, J. O. (2011). Quantitative metrics of overlaps in Grinnellian niches: Advances and possible drawbacks. *Global Ecology and Biogeography*, 20(6):915–927.

Rodríguez, E. et al. (2006). A global assessment of the SRTM performance. *Photogrammetric Engineering and Remote Sensing*, 72(3):249–260.

Rodríguez-Soto, C. et al. (2011). Predicting potential distribution of the jaguar (Panthera onca) in Mexico: Identification of priority areas for conservation. *Diversity and Distributions*, 17(2):350–361.

Sánchez-Fernández, D. et al. (2011). Species distribution models that do not incorporate global data misrepresent potential distributions:

A case study using Iberian diving beetles. *Diversity and Distributions*, 17(1):163–171.

Schneider, D. (2001). The Rise of the Concept of Scale in Ecology. *BioScience*, 51(7):545.

Schoener, T. W. (1968). The Anolis Lizards of Bimini: Resource Partitioning in a Complex Fauna. *Ecology*, 49(4):704–726.

Schweiger, O. et al. (2012). Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecology and Biogeography*, 21(1):88–99.

Seoane, J. et al. (2004). Are existing vegetation maps adequate to predict bird distributions? *Ecological Modelling*, 175(2):137–149.

Šímová, P. et al. (2015). Refugial role of urbanized areas and colonization potential for declining Crested Lark (Galerida cristata) populations in the Czech Republic, Central Europe. *Journal of Ornithology*, 156(4):915–921.

Šímová, P. et al. (2019). Fine scale waterbody data improve prediction of waterbird occurrence despite coarse species data. *Ecography*, 42(3):511–520.

Singer, A. et al. (2018). Constructing a hybrid species distribution model from standard large-scale distribution data. *Ecological Modelling*, 373:39–52.

Slatyer, R. A. et al. (2013). Niche breadth predicts geographical range size: A general ecological pattern. *Ecology Letters*, 16(8):1104–1114.

Sofia, G. et al. (2013). Variations in multiscale curvature distribution and signatures of LiDAR DTM errors. *Earth Surface Processes and Landforms*, 38(10):1116–1134.

Soley-Guardia, M. et al. (2019). Sufficient versus optimal climatic stability during the Late Quaternary: using environmental quality to guide phylogeographic inferences in a Neotropical montane. *Journal of Mammalogy*, 100(6):1783–1807.

Sor, R. et al. (2017). Effects of species prevalence on the performance of predictive models. *Ecological Modelling*, 354:11–19.

Stockwell, D. R. and Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148(1):1–13.

Sun, J. et al. (2020). Modeling the potential distribution of Zelkova schneideriana under different human activity intensities and climate change patterns in China. *Global Ecology and Conservation*, 21:e00840.

Sun, Y. et al. (2017). Climatic suitability ranking of biological control candidates: A biogeographic approach for ragweed management in Europe. *Ecosphere*, 8(4):e01731.

Svensson, J. R. et al. (2013). Excessive spatial resolution decreases performance of quantitative models, contrary to expectations from error analyses. *Marine Ecology Progress Series*, 485:57–73.

Syfert, M. M. et al. (2013). The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLOS ONE*, 8(2):e55158.

Syphard, A. D. and Franklin, J. (2009). Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, 32(6):907–918.

Tate, N. and Wood, J. (2001). Fractals and scale dependencies in topography. In Tate, N. and Atkinson, P., editors, *Scale in Geographical Information Systems.*, chapter Fractals a, pages 33–51. Wiley, Chichester.

Tessarolo, G. et al. (2014). Uncertainty associated with survey design in species distribution models. *Diversity and Distributions*, 20(11):1258–1269.

Thibaud, E. et al. (2014). Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, 5(9):947–955.

Thomas, J. et al. (2015). Suitability of spaceborne digital elevation models of different scales in topographic analysis: an example from Kerala, India. *Environmental Earth Sciences*, 73(3):1245–1263.

Tingley, R. et al. (2018). Integrating transport pressure data and species distribution models to estimate invasion risk for alien stowaways. *Ecography*, 41(4):635–646.

Tong, R. et al. (2016). Quantifying relationships between abundances of cold-water coral Lophelia pertusa and terrain features: A case study on the Norwegian margin. *Continental Shelf Research*, 116:13–26.

Townsend, C. R. et al. (2003). *Essentials of ecology*. Blackwell Science, Oxford, 2 edition.

Townshend, J. R. G. et al. (1992). The Impact of Misregistration on Change Detection. *Geoscience and Remote Sensing*, 30(5):1054–1060.

Turner, J. A. et al. (2019). How does spatial resolution affect model performance? A case for ensemble approaches for marine benthic mesophotic communities. *Journal of Biogeography*, 46(6):1249–1259.

Václavík, T. and Meentemeyer, R. K. (2012). Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, 18(1):73–83.

Van Niel, K. and Austin, M. (2007). Predictive vegetation modeling for conservation: Impact of error propagation from digital elevation data. *Ecological Applications*, 17(1):266–280.

Van Niel, K. et al. (2004). Effect of error in the DEM on environmental variables for predictive vegetation modelling. *Journal of Vegetation Science*, 15(6):747–756.

Varela, S. et al. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11):1084–1091.

Vaze, J. et al. (2010). Impact of DEM accuracy and resolution on topographic indices. *Environmental Modelling and Software*, 25(10):1086–1098.

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, 36(12):2290–2299.

Venier, L. A. et al. (2004). Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Basin. *Journal of Biogeography*, 31(2):315–331.

Verburg, P. H. et al. (2011). Challenges in using land use and land cover data for global change studies. *Global Change Biology*, 17(2):974–989.

Vogeler, J. et al. (2014). Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. *Remote Sensing of Environment*, 147:13–22.

Walbridge, S. et al. (2018). Unified Geomorphological Analysis Workflows with Benthic Terrain Modeler. *Geosciences*, 8(3):94.

Wang, H. and Ellis, E. C. (2005). Image misregistration error in change measurements. *Photogrammetric Engineering and Remote Sensing*, 71(9):1037–1044.

Warren, D. (2019). ENMTools: initial beta release.

Warren, D. L. et al. (2008). Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution*, 62(11):2868–2883.

Warren, D. L. et al. (2020). Evaluating species distribution models with discrimination accuracy is uninformative for many applications. *Journal of Biogeography*, 47(1):167–180.

Watts, S. et al. (2019). Modelling potential habitat for snow leopards (Panthera uncia) in Ladakh, India. *PLOS ONE*, 14(1):e0211509.

Wheatley, M. and Johnson, C. (2009). Factors limiting our understanding of ecological scale. *Ecological Complexity*, 6(2):150–159.

Whittaker, R. J. et al. (2005). Conservation biogeography: Assessment and prospect. *Diversity and Distributions*, 11(1):3–23.

Wieczorek, J. et al. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International journal of geographical information science*, 18(8):745–767.

Wiens, J. A. (1989). Spatial Scaling in Ecology. *Functional Ecology*, 3(4):385.

Wiersma, Y. F. et al. (2011). Introduction, landscape modeling of species and their habitats: history, uncertainty, and complexity. In Drew, C., Wiersma, Y., and Huettmann, F., editors, *Predictive Species and Habitat Modeling in Landscape Ecology*, chapter Introducti, pages 1–6. Springer, New York.

Wild, J. et al. (2019). Climate at ecologically relevant scales: A new temperature and soil moisture logger for long-term microclimate measurement. *Agricultural and Forest Meteorology*, 268:40–47.

Williams, K. J. et al. (2012). Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, 26(11):2009–2047.

Williams, V. L. and Crouch, N. R. (2017). Locating sufficient plant distribution data for accurate estimation of geographic range: The relative value of herbaria and other sources. *South African Journal of Botany*, 109:116–127.

Wilson, M. F. et al. (2007). Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy*, 30(1-2):3–35.

Wisz, M. S. et al. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5):763–773.

Wood, J. (1996). *The geomorphological characterisation of digital elevation models*. PhD thesis, University of Leicester.

Yamazaki, D. et al. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853.

Ye, X. et al. (2018). Impacts of future climate and land cover changes on threatened mammals in the semi-arid Chinese Altai Mountains. *Science of the Total Environment*, 612:775–787.

Zeileis, A. (2006). Object-oriented Computation of Sandwich Estimators Object-oriented Computation of Sandwich Estimators. *Statistical Software*, 16(9):1–16.

Zevenbergen, L. W. and Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12(1):47–56.

Zhang, G. et al. (2018). A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. *Transactions in GIS*, 22(1):202–216.

Zhang, J. et al. (2014). *Scale in spatial information and analysis*. CRC Press.

Zhang, L. et al. (2016). Using DEM to predict Abies faxoniana and Quercus aquifolioides distributions in the upstream catchment basin of the Min River in southwest China. *Ecological Indicators*, 69:91–99.

Zhou, Q. and Liu, X. (2004). Analysis of errors of derived slope and aspect related to DEM data properties. *Computers and Geosciences*, 30(4):369–378.

Zurell, D. (2017). Integrating demography, dispersal and interspecific interactions into bird distribution models. *Journal of Avian Biology*, 48(12):1505–1516.

Zurell, D. et al. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, 119(4):622–635.

Zurell, D. et al. (2016). Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology*, 22(8):2651–2664.

Zurell, D. et al. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, 47(1):101–113.

Curriculum vitae

& list of publications

Personal

Name: Lukáš Gábor
Date of birth: January 1st, 1991 (*Czech Republic*)
E-mail: gabor@fzp.czu.cz
ORCID: https://orcid.org/0000-0001-6137-0994
RG: https://www.researchgate.net/profile/Lukas_Gabor

Affiliations

2017 – present
Department of Applied Geoinformatics and Spatial Planning,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
Member of Spatial Science in Ecology and Environment Research Group

Education

2016 – present
Department of Applied Geoinformatics and Spatial Planning,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
PhD studies in Applied and Landscape Ecology
Thesis topic: The Quality of Spatial Data and its Effect on Species Distribution Models

2014 – 2016
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
Master's degree in Applied Ecology
Thesis topic: Do Environmental Filters Improve Predictions of Species Distribution Models?

Participation in Research Projects

2017 – 2019
The quality of spatial data and its effect on species distribution models
Grant Agency of the Czech University of Life Sciences Prague
project Leader

Internships

10/2019 – 12/2019
Center for Geospatial Analytics
NC State University, USA
Internship

6/2018 – 8/2018
Marine Geomatics Lab
University of Florida, USA
Internship

10/2017 – 12/2017
Department of Computational Landscape Ecology
Helmholtz Centre for Environmental Research, Germany
Internship

International Conferences

10/2019
Does Positional Error Affect Fine-Scale Species Distribution Models?
Gábor L., Moudrý V., Lecours V., Malavasi M., Barták V. & Václavík T.
In: The International Society for Ecological Modelling Global Conference.
Salzburg, Austria.

5/2019
Contrasting Fine-Scale Environmental Preferences of Cold-Water Coral Species in the Northwest Atlantic
Lecours V., **Gábor L.**, Edinger E. & Devillers R.
In: 18th International Symposium GeoHab.
St. Petersburg, Russia.

4/2017
The Quality of Spatial Data and its Effect on Species Distribution Models
Gábor L. & Moudrý V.
In: Doctoral Consortium on 3rd International Conference on Geographical Information Systems Theory, Applications and Management, INSTICC.
Porto, Portugal.

Publications in Journals with Impact Factor (J_{imp})

The effect of positional error on fine scale species distribution models increases for specialist species (2020). **Gábor L.**, Moudrý V., Lecours V., Malavasi M., Barták V., Fogl M., Šímová P., Rocchini D. & Václavík T. *Ecography*, Oikos, vol.43, p. 256-269.

How do species and data characteristics affect species distribution models and when to use environmental filtering? (2019). **Gábor L.**, Moudrý V., Barták V. & Lecours V. (2019). *International Journal of Geographical Information Science*, Taylor & Francis, p. 1-18.

Potential pitfalls in rescaling Digital Terrain Model-derived attributes for ecological studies (2019). Moudrý V., Lecours V., Malavasi M., Misiuk B., **Gábor L.**, Šímová P. & Wild J. *Ecological Infomatics*, Elsevier, 100987.

On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs (2018). Moudrý V., Lecours V., Gdulová K., **Gábor L.**, Moudrá L., Kropáček J. & Wild J. *Ecological Modelling*, Elsevier, vol. 383, p. 3-9.

Other Publications

Fine-scale habitat characterization of The Gully, the Flemish Cap, and the Orphan Knoll, Northwest Atlantic, with a focus on cold-water corals (2020). Lecours V., **Gábor L.**, Edinger E. & Devillers R. In *Seafloor Geomorphology as Benthic Habitat*, Elsevier, p. 735-751.

Developing a criterion for distinguishing tetraploid birch species from diploid and modelling their potential distribution on the Czech Republic (2016). Linda R., **Gábor L.**, Kunes I., Balas M., Rasakova N. & Gallo J. In *Proceedings of Central European Silviculture*, Kacálek D., Novák J., Součková J. & Škopová J. (Eds.), p. 71-77.

Active commuting of the inhabitants of Liberec city in low and high walkability areas (2015). Rubín L., Mitáš J., Dygrýn J., Šmída J., **Gábor L.** & Pátek A. Active commuting of the inhabitants of Liberec city in low and high walkability areas. *Acta Gymnica*, vol. 45, p. 195-202.

Teaching Experience

2016 – present
Department of Applied Geoinformatics and Spatial Planning
Seminars of GIS; Computer Technology Utilization; Cartography and Mapping

2016 – present
Department of Applied Geoinformatics and Spatial Planning
Final thesis supervisor or consultant; 6 successfully defended Bachelor's or Master's Thesis