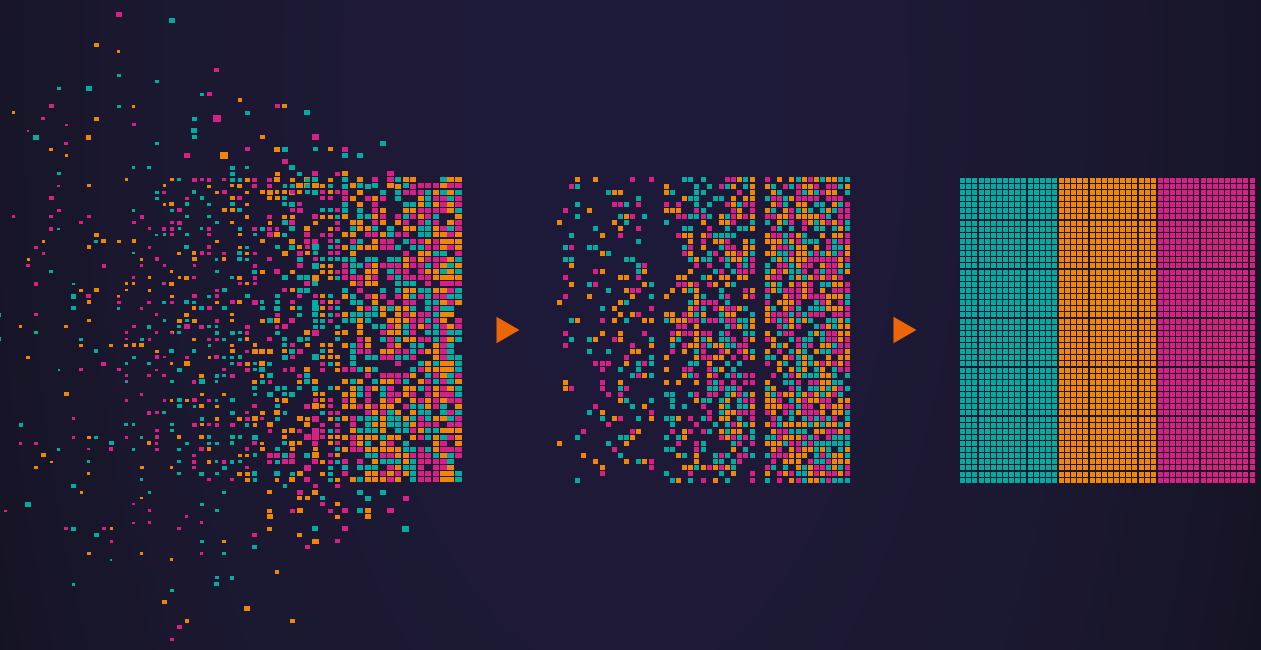
BEYOND ASSUMPTIONS:

Unraveling Data Limitations In Predictive Ecology



Lukáš Gábor

To Vítězslav Moudrý

My exceptional supervisor, mentor and more importantly friend.

This dissertation entitled *"Beyond Assumptions: Unraveling Data Limitations in Predictive Ecology"* is submitted to the Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences in Prague in July 2024 for the degree of *Doctor of Philosophy*. All sources of information, including text, illustrations, tables, and images, have been duly acknowledged and cited.

Essentially, all models are wrong, but some are useful.
-George Box

Acknowledgements

I would like to express my sincere gratitude to several people who have played a pivotal role in completing my dissertation.

First, I would like to recognize my supervisor, Vítek Moudrý, for your invaluable guidance, unwavering support, and exceptional mentorship throughout my Ph.D. journey. Your expertise, dedication, and encouragement have been instrumental in shaping the direction of my research and pushing me to reach my full potential.

I am also deeply grateful to the Fulbright Commission, Petr Keil, and Walter Jetz for the incredible opportunity to pursue a significant portion of my Ph.D. studies at Yale University. Yale's resources, intellectual environment, and collaborative opportunities have greatly enriched my research experience. Walter, I am truly indebted to you for your generosity and belief in my abilities.

Furthermore, I would like to thank my family for their love, encouragement, and relentless support. Mum and Dad, your patience, understanding, and certainty in my abilities have been an immense source of strength throughout this academic journey.

Special appreciation goes to Ruchika C. Rana for your unwavering support. You appeared unexpectedly in my life, but your presence has been a source of solace during the ups and downs of this, not only, academic endeavor. I am grateful for your insightful discussions, encouragement, and love you provided, which propelled me forward during challenging moments.

Lastly, I want to express my profound appreciation to all my friends, colleagues, and mentors who have supported me along the way. Your constructive feedback, stimulating discussions, and shared experiences have been invaluable in shaping my ideas and strengthening my research.

The scientific papers included in my dissertation were made possible through the generous support of the below-listed scientific grants. I am sincerely grateful for the financial support provided by these grants, as they enabled me to conduct the necessary research and contribute to advancing knowledge in my field.

• Technological grant agency of the Czech Republic (grant no. SS02030018 DivLand)

• OP RDE Improvement in Quality of the Internal Grant Scheme at CZU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016944 (grant no. 43/2021).

• Internal Grant Agency of Faculty of Environmental Sciences, Czech Univ. of Life Sciences Prague (grant no. 2020B0009)

• Internal grant agency of the faculty of environmental sciences, Czech Univ. of Life Sciences, Prague (grant no. 2021B0009)

Abstract

Context

Species distribution models (SDMs) have proven valuable in filling gaps in our knowledge of species occurrences. However, despite their broad applicability, SDMs exhibit critical shortcomings due to limitations in species occurrence data and in environmetal variables. Typical example of such limitation in species records is spatial uncertainty in their location, ranging from a few meters to tens of kilometers (e.g., positional uncertainty in the GBIF database can exceed 300 kilometers). Similarly, environmental variables may be limited by how much detailed information about the environment they provide. For example, land cover types typically indicate the amount of habitat within a spatial unit. However, it is possible that a simple binary presence/absence of suitable habitat may be the only information available or even more important aspect than area, determining species distribution.

Objectives

This dissertation had three main objectives. Firstly, I explored the potential of an alternative approach to incorporate environmental variables into models (i.e., binary versus continuous habitat information). Secondly, I investigated the influence of positional uncertainty in species records on the ecological interpretability of models. Thirdly, I evaluated whether the appropriateness of using binary or proportional (continuous) type of variables, and the influence of positional error, is affected by the scale of the analyses. Specifically, I addressed the following research questions: a) Can binary land cover predictors provide models of higher accuracy than traditionally used proportional variables? b) If so, what is the role of spatial grain in determining the usability of binary land cover predictors? c) To what extent does positional uncertainty in species occurrence data affect model parameter estimation and the ecological interpretability of species distribution models? d) What are the trade-offs between analysis grain and positional uncertainty in modeling species distributions?.

Results

Results indicated that models' performances were not affected by the type of the adopted habitat variable (proportional or binary but the usability of binary variables decreased with coarsening the resolution (i.e., binary representation of habitat is useful at finer grain sizes of approx. 1km^2. Results confirmed that model performance decrease with increasing positional error in species records, as demonstrated in prior studies. However, I have shown that coarsening the analysis grain to compensate for positional error did not improve model performance as was widely assumed. This, however, doesn't mean we should exclude species records with high positional uncertainty from our studies, because the negative consequences of positional uncertainty on model performance did not extend as strongly to the ecological interpretability of the models.

Conclusions

These findings are encouraging for practitioners using SDMs to reveal relationships between species occurrences and its environmental drivers as such relationship can be to some degree estimated using positionally uncertain data and simple environmental variables describing presence or absence of a habitat. On the other hand, my findings show that positional uncertainty in species data can cause inaccurate spatial predictions leading to inaccurate maps of species distributions, especially in heterogeneous environments and when using fine resolution environmental data. Therefore, such models are not suitable for tasks like setting up protected areas or prioritizing conservation efforts.

Abstract in Czech

Souvislosti

Modely druhové distribuce (SDMs) jsou důležitým nástrojem při doplňování mezer v našich znalostech o výskytech druhů. Navzdory tomu, že se tyto modely často používají v ekologických studiích, mají zásadní nedostatky kvůli nepřesnostem v datech o výskytech druhů a environmentálních prediktorech. Typickým příkladem takového omezení je polohová nejistota v záznamech druhů, která může být od několika metrů až po desítky kilometrů (např. polohová nejistota v databázi GBIF může přesáhnout 300 kilometrů). Environmentální prediktory pak mohou být omezeny tím, jak přesnou informaci o prostředí, ve kterém se druh nalézá, poskytují. Například proměnné krajinného pokryvu obvykle udávají rozlohu nebo podíl habitatu v rámci určité oblasti. Avšak co když, jednoduchá binární informace (přítomnost/absence) o vhodném habitatu může být jedinou dostupnou informací nebo dokonce důležitějším aspektem při určení druhové distribuce než informace o celkové rozloze habitatu?

Cíle

Tato disertační práce měla tři hlavní cíle. Za prvé jsem zkoumal možnosti použití nového typu environmentálních prediktorů (binárních dat), které obsahují pouze informaci o přítomnosti nebo absenci vhodného habitatu. Za druhé jsem zkoumal vliv polohové nejistoty v druhových záznamech na ekologickou interpretovatelnost modelů. Třetím cílem pak bylo posouzení role prostorového měřítka na modely, které používali binární prediktory a druhová data s různou polohovou nejistotou. Konkrétní výzkumné otázky byly: a) Mohou binární environmentální prediktory krajinného pokryvu zvýšit přesnost modelů? b) Pokud ano, jaká je role použitého prostorového měřítka? c) Do jaké míry ovlivňuje polohová nejistota v druhových datech ekologickou interpretovatelnost modelů? d) Jak spolu souvisí a polohová chyba druhových dat?

Výsledky

Výsledky ukázaly, že přesnost modelů nebyla významně ovlivněna typem použitých environmentální prediktorů (proporcionální nebo binární infomrace o vhodném habitatu). Je nicméně důležité říct, že ale použitelnost binárních prediktorů klesala s hrubším prostorovým rozlišením. To znamená, že binární reprezentace habitatu je užitečná především při použití prediktorů s vyšším prostorovým rozlišením (cca od 1km^2). Výsledky kromě toho potvrdily, že přesnost modelů klesá se zvyšující se polohovou nejistotou v záznamech druhů, tak jak bylo prokázáno v předchozích studiích. Důležitým závěrem práce je fakt, že zhoršení prostorového měřítka nekompenzuje negativní vliv polohové nejistoty, jak se všeobecně předpokládalo. To však neznamená, že bychom neměli při modelování záznamy druhů s vysokou polohovou nejistotou používat.

Závěry

Jak ukázaly výsledky, polohově neurčitá data a binární informaci o přítomnosti habitatu lze za určitých podmínek využít pro studium vztahů mezi organismy a prostředím. Na druhou stranu mapy druhové distribuce vycházející z modelů založených na datech s vysokou polohovu nejistotou, jsou nepřesné (zejména v heterogenním prostředí a při použití environmentálních prediktorů s vysokým prostorovým rozlišením) a nevhodné pro aplikace v ochraně přírody.

Contents

Preface

1.1. Foreword

I am filled with a profound sense of accomplishment and gratitude upon completing this dissertation. The journey leading up to this moment has been filled with rigorous exploration, tireless research, and unwavering dedication. This work represents the culmination of years of academic pursuit and a deep passion for understanding the intricacies of ecological predictive modeling.

The driving force behind this dissertation stems from my enduring fascination with data and its potential to unravel the complex tapestry of ecological systems across various spatial scales. From the earliest stages of my academic pursuits, I recognized the importance of spatial data quality and the influence it wields over the outcomes of predicting species distribution. It became evident that while a wealth of studies focused on modeling methodologies and theoretical frameworks, the critical aspects of spatial data quality and the range of spatial data types applicable in predicting species distribution deserved deeper exploration.

Embracing this challenge, I embarked on a meticulous investigation to bridge the gap between theory and practical application. The research contained within these pages attempts to shed light on the impact of spatial data quality and their types and the validity of commonly held assumptions, offering insights that can enhance the accuracy, reliability, and robustness of predictive models.

As we find ourselves in an age where data is increasingly accessible in unprecedented quality and volume, these findings take on heightened significance. The ever-expanding availability of data holds great promise for advancing our understanding of ecological systems and informing effective conservation and management strategies. I sincerely hope that the research presented herein will contribute to this collective pursuit and inspire further investigations into the realm of predicting species distribution.

With great pride and a sense of anticipation, I offer this dissertation to the academic community, hoping that it will contribute to the knowledge and serve as a catalyst for future research endeavors. May the ideas presented within these pages ignite curiosity, inspire collaboration, and foster a deeper appreciation for the intricate wonders of predictive modeling in ecology.

1.2. Scientific motivation

From my early years at high school, I became captivated by how data quality and availability can shape our lives and influence the world around us. Therefore, many years after, when the time came to select a topic for my dissertation, I deliberately chose to explore the realm of spatial data quality and types in predictive ecology. This area of investigation carries substantial importance, given that while numerous studies have concentrated on modeling methodologies and theoretical frameworks, only a limited number have delved into the critical aspects of data quality and data types applicable in predictive modeling. More importantly, such research takes on heightened significance in the contemporary era, where there is an exponentially growing wealth of spatial data accessible to us.

In addition, during an extensive literature review, I observed a recurring recommendation to downscale the resolution of environmental predictors to match the highest positional uncertainty of species records. However, there appeared to be a notable lack of empirical

testing to validate this assumption. This sparked my curiosity, prompting me to investigate the validity of this prevalent and customary practice. I was eager to determine the extent to which this assumption held true and whether it could influence the accuracy and reliability of modeling outcomes.

I firmly believe that investigating spatial data quality and diverse types within ecological predictive modeling is of utmost significance. By addressing these crucial aspects, we can enhance predictive models' accuracy, reliability, and robustness, ultimately contributing to a more comprehensive understanding of ecological systems and facilitating informed decision-making for environmental conservation and management.

1.3. Dissertation structure

The dissertation comprises of four published papers, which collectively contribute to the body of knowledge in predicting species distribution. It is divided into two distinct parts. The first part of the dissertation beginning with a preface and a comprehensive general introduction to predicting species distribution. These chapters provide a foundational understanding of the subject matter and set the stage for subsequent studies wherein specific research challenges and questions will be examined in greater detail.

The second part of the thesis consists of four chapters, which delve into individual studies conducted within the realm of species predictive modeling. These chapters present the specific research endeavors undertaken, each contributing to the broader understanding of the field and offering unique insights and findings.

Chapter 3.1: Habitats as predictors in species distribution models: Shall we use continuous or binary data?

Chapter 3.2: Assessing the applicability of binary land-cover variables to species distribution models across multiple grains.

Chapter 3.3: Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable.

Chapter 3.4: Positional errors in species distribution modeling are not overcome by the coarser grains of analysis.

Together, these parts form a cohesive body of work that reflects the culmination of extensive research and analysis, ultimately contributing to advancing the knowledge in predicting species distribution.

1.4. Dissertation objectives

The objectives of my dissertation were threefold. Firstly, I explored the potential of a novel environmental data type to improve the accuracy of species distribution predictions. Secondly, I investigated the influence of positional uncertainty in species records on the model's ecological interpretability and assessed our abilities to compensate for these data shortcomings. Thirdly, I assessed the role of spatial grain for the usability of binary variables, and for the effect of positional uncertainty. Specifically, I aimed to address the following research questions:

a) Can binary land cover predictors (i.e., information about the presence or absence of a habitat) enhance the precision of species distribution models?

b) If so, what is the role of spatial scale in determining the effectiveness of binary land cover predictors?

c) To what extent does positional uncertainty affect model parameter estimation? Specifically, what is the influence of positional uncertainty on the ecological interpretability of species distribution models?

d) What trade-offs exist between analysis grain and positional uncertainty when modeling species distributions? Is it advisable to coarsen the analysis grain to mitigate the impact of positional error, or should the analysis grain align as closely as possible with the assumed response grain, irrespective of positional error?

Theoretical background

Predictive modeling in ecology has become an indispensable tool in studying various ecological phenomena, including species-environment relationships, species interactions, and population dynamics; it is also crucial in facilitating nature conservation efforts and in informing decision-making processes (Guisan and Zimmermann 2000, Anderson et al. 2003, Tan et al. 2006, Drew and Perera 2010). Its significance has become even more pronounced considering climate changes, as it provides valuable insights and predictions to guide proactive measures for the protection and sustainable management of our natural environment (Alitzer et al. 2013, Urban et al. 2016, Nogués-Bravo et al. 2018, Schleuning 2020). Species distribution models (SDMs) are widely recognized as powerful tools in predictive ecological modeling. The primary objective of SDMs is to establish relationships between species records and environmental variables which allow for describing these relationships and predicting species distribution in an environmental or geographical space (Elith and Leathwick 2009, Miller 2010, Ferrier et al. 2017, Franklin 2023). While SDMs have become a routine part of ecological research, it is important to acknowledge their inherent limitations, particularly those associated with input data (Araújo et al. 2019).

To achieve accurate SDMs, it is crucial to ensure the accuracy of both species' records and environmental predictors (Robinson et al. 2011, Aubry et al. 2017, Ahmad Suhaimi et al. 2021, Arenas-Castro et al. 2022). However, attaining such high accuracy is often challenging in real-world scenarios. Additionally, the choice of data type also plays a significant role. When constructing SDMs, one must make decisions regarding utilizing species records from museum databases, species atlases, or global databases (e.g., eBird or GBIF). Furthermore, selecting appropriate environmental predictors presents additional complexities. For one thing, determining the optimal grain size (i.e., resolution of environmental predictors) is essential (Pearson and Dawson 2003, Guisan et al. 2007, Kaliontzopoulou et al. 2008, Seo et al. 2009, Miguet et al. 2016, Mertes and Jetz 2018). In addition, it is equally important to choose what predictors should be used (e.g., climate, topography, soil, land cover, canopy height, elevation; Elith and Leathwick 2009, Miller 2010, Franklin 2010). Moreover, the source of these data (e.g., LiDAR, satellite images) needs to be considered (Austin and Van Niel 2011, Bucklin et al. 2015, Moudrý et al. 2023a), as well as the potential incorporation of environmental predictors exhibiting high correlation (Franklin 2010, Dormann et al. 2013, De Marco and Nóbrega 2018, Sillero and Barbosa 2021). In my dissertation thesis, I strived to transcend common assumptions in predicting species distribution by offering an innovative approach and challenging established practices.

2.1. Species distribution models

Species distribution models (SDMs) have diverse applications in ecology and conservation. By establishing the relationship between species records and environmental conditions, SDMs can generate spatial predictions of species distributions (Ferrier et al. 2017; see Figure 2.1). This information is valuable for understanding species' habitat requirements, elucidating niche dynamics, and uncovering the underlying mechanisms that drive species distributions. In addition, these models have significant implications for conservation planning. They can provide insights into species vulnerability and help identify areas with high conservation value (Araújo and Guisan 2006, Austin 2007, Pearman et al. 2008, Elith and Leathwick 2009, Franklin 2010, Miller 2010, Merow et al. 2014, D'Amen and Azzurro 2020, Liu et al. 2022, Lo Parrino et al. 2023). Furthermore, by combining SDMs' predictions with information on protected areas, land use patterns, and other conservation priorities, decision-makers can ef-

fectively prioritize conservation interventions, establish wildlife corridors for species movement, and guide land-use planning to minimize detrimental impacts on biodiversity from a local to global scale (Johnson and Gillingham 2005, Domisch et al. 2019, Sutton et al. 2023, Van Moorter 2023). More importantly, SDMs can project the potential impacts of climate change on species distributions and habitats. By integrating future climate scenarios into the models, SDMs can predict how species distributions might shift in response to changing environmental conditions (Synes and Osborne 2011, Stanton et al. 2012, Gotelli 2015, Booth 2018, Briscoe et al. 2023, Festa et al. 2023). This information is valuable for assessing species' vulnerability to climate change, identifying areas where conservation efforts may need to be intensified, and informing adaptation strategies.

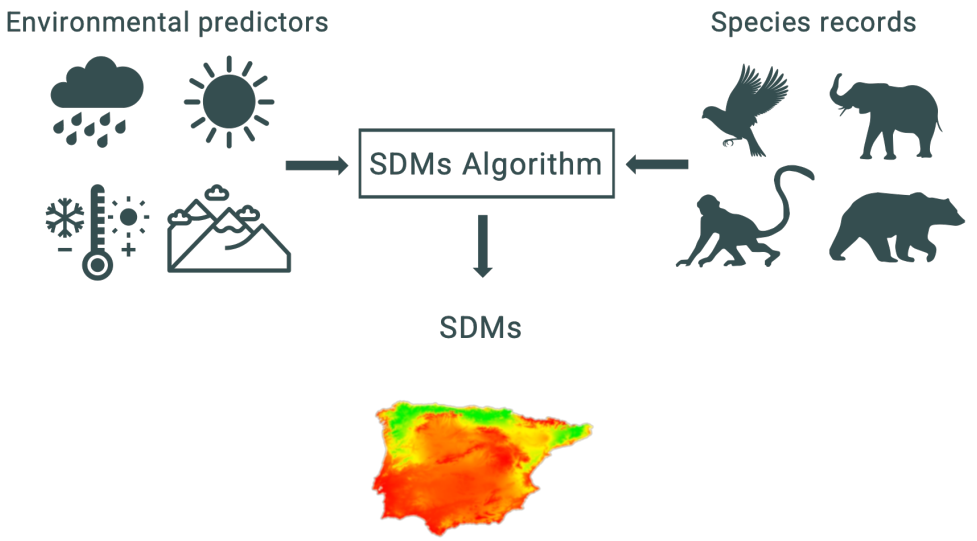


Figure 2.1: The general process of species distribution modeling. SDMs combine environmental predictors and species records to quantify species-environment relationships. Once the model is fitted, we can predict species distribution in environmental and geographical spaces or explore species–environment relationship inferences.

Species distribution modeling encompasses various modeling techniques. The most common techniques used in SDMs include Boosted Regression Trees (BRT), Generalized Additive Models (GAM), Generalized Linear Models (GLMs), Random Forests (RF), and Maximum Entropy Models (MaxEnt). These techniques can be divided into linear models (GAM, GLM) and machine learning models (MaxEnt, RF, BRT). GLMs provide a flexible framework for modeling the relationship between species occurrence or abundance and environmental variables, enabling the incorporation of various functional forms and link functions (Nelder and Baker 1972, Oksanen and Minchin 2002). GAM, a flexible regression technique, extends GLMs by incorporating non-linear relationships through smoothing functions. GAMs allow for more flexible and nuanced modeling of species-environment relationships, capturing complex response patterns that linear relationships may not adequately represent (Yee and Mitchell 1991).

MaxEnt, a widely used machine learning approach in SDMs, predicts species distributions by maximizing the entropy of the model while adhering to constraints imposed by environmen-

tal predictors. This technique is particularly suitable for handling presence-only data and capturing complex interactions among environmental predictors (Phillips et al. 2006). RF, a machine learning technique, is extensively employed in SDMs due to its ability to combine multiple decision trees, considering the importance of different environmental variables and their interactions (Breiman 2001). BRT, another machine learning technique, sequentially builds decision trees and iteratively adjusts their weights to enhance model predictions. This technique is known for effectively capturing complex interactions and non-linear relationships in species distributions (Friedman et al. 2000). The choice of modeling technique depends on the specific research question, available data, and characteristics of the species under investigation (Elith and Elith and Leathwick 2009, Norberg et al. 2019, Valavi et al. 2021).

In addition to employing separate modeling techniques, researchers can utilize an ensemble modeling approach. Ensemble modeling, also referred to as model stacking or model averaging, entails integrating multiple modeling techniques to enhance the accuracy and robustness of SDMs. This approach offers two main benefits. First, it reduces the risk of relying solely on the strengths and limitations of a single modeling technique. Second, ensemble modeling can improve predictive accuracy by capitalizing on the complementary strengths of the previously mentioned models. By combining the outputs of different techniques, ensemble modeling can capture a broader range of ecological processes, handle different data types effectively, and produce more robust and reliable predictions (Bates and Granger 1969, Makridakis and Winkler 1983, Araújo and New 2007, Mateo et al. 2012, Parker 2013). While writing about ensemble modeling, I am intrigued by the possibility of utilizing this approach to mitigate the negative influence of positional uncertainty in species records. Considering the challenges posed by inaccuracies in species occurrences, exploring the potential of ensemble modeling to address positional uncertainty is an avenue worth investigating, particularly as it could contribute to advancing SDMs methodologies. Whether pursued by myself or other researchers, exploring the application of ensemble modeling to mitigate the effects of positional uncertainty holds great potential for further refining the accuracy and applicability of SDMs in ecological studies.

2.2. Input data

Elementary input data for SDMs consist of two main components: environmental predictors and species records. Environmental predictors encompass a wide range of, for example, biophysical and climatic variables that characterize the environmental conditions of the study area. Integrating species records and environmental predictors allows for exploring the relationships between species occurrences and environmental conditions, forming the basis for predicting species distributions (Elith and Graham 2009, Elith and Leathwick 2009, Franklin 2010). Species records, also known as occurrence data, provide information on the presence or absence of a species in specific locations.

The accuracy of environmental predictors and species records is essential in determining the reliability and robustness of models' outcomes. Any inaccuracies or uncertainties in the input data can introduce noise and distort the relationships between species records and environmental conditions, leading to less accurate model results (Araújo et al. 2019). Environmental predictors should be accurately measured, derived, calculated, and appropriately scaled to the spatial extent of the study (Osborne and Leitao 2009, Moudrý et al. 2018, Moudrý et al. 2019). In terms of species records, it is crucial to ensure that the recorded species are correctly identified, the record's coordinates are properly georeferenced, and that species are

evenly gathered across a study area (Kamino et al. 2011, Syfert et al. 2013, Costa et al. 2015, Mitchell et al. 2017, Oleas et al. 2019, Gábor et al. 2020a, Rocchini et al. 2023).

In addition to species and environmental data, several other data types can be incorporated into SDMs to enhance their predictive accuracy and ecological insights (for example abundance data, genetic data, landscape connectivity data, or species interactions) Abundance data quantifies a species' relative abundance or density at different locations. This data type provides additional information beyond presence or absence, allowing for a more nuanced understanding of species distributions and population dynamics (Howard et al. 2014, Yu et al. 2020, Waldock et al. 2021). Genetic information, such as DNA sequences or markers, can provide valuable insights into population structure, gene flow, and adaptive genetic variation. Incorporating genetic data into SDMs can help identify genetic factors influencing species distributions (Termansen et al. 2006, Gotelli and Stanton-Geddes 2015, Marcer et al. 2016). Measures of landscape connectivity, such as habitat corridors or landscape resistance, can be incorporated into SDMs to account for the influence of landscape structure on species dispersal and connectivity (Maiorano et al. 2019, Shipley et al. 2021, Curd et al. 2022). Species interactions, such as predator-prey relationships, competition, or mutualistic interactions, can significantly impact species distributions. Incorporating data on ecological interactions into SDMs can improve the understanding of species-environment relationships and enhance the predictive power of the models (Wisz et al. 2013, Trainor et al. 2014, Dormann et al. 2018, Windsor et al. 2022).

The scope of my dissertation requires a specific focus on utilizing species records and environmental predictors. While other introduced data types can certainly contribute valuable insights to species distribution, their inclusion fell outside the purview of my dissertation. As such, those data types will not be further examined within this context, and the subsequent sections will remain focused on species records and environmental predictors.

2.2.1. Environmental predictors

Various environmental predictors can be incorporated into SDMs, each capturing different aspects of the species-environment relationship. Thus, selecting appropriate predictors is a major methodological challenge (Dormann et al. 2007, Williams et al. 2012, Misiuk et al. 2018, Smith and Santos 2020, Zurell et al. 2020).

Climate data, such as temperature, precipitation, and seasonality, are frequently used to model the broad-scale distribution of species, as climate exerts a significant influence on species' physiological and ecological tolerances (Fick and Hijmans 2017). Such data are often acquired from meteorological stations or interpolated from global climate databases, such as WorldClim.

Soil data, encompassing properties like soil type, pH, and nutrient content, contribute to understanding the edaphic conditions that shape species distributions and can be acquired from soil surveys, remote sensing techniques, or global soil databases like the SoilGrids system (Schröder 2008, Walthert and Meier 2017).

Topographic data, including elevation, slope, and aspect, provide information about the terrain and microclimate (Miller 2010, Carlson et al. 2022). Topographic data can be obtained from satellite (e.g., Tandem-X, Copernicus Digital Elevation Model) or airborne remote sensing platforms (e.g., National Center for Geographic Information). Vegetation data, such as vegetation indices (e.g., canopy height) or land cover classifications, offer insights into the

structure and composition of habitats, reflecting the resources available to the species (Davies and Asner 2014, Moudrý et al. 2023a). Vegetation data can be derived from satellite imagery, such as those provided by the Moderate Resolution Imaging Spectroradiometer (MODIS) or from land cover datasets like the CORINE Land Cover.

The remarkable progress in remote sensing technology allows us to derive environmental data at increasingly finer scales, enabling the integration of fine-resolution information in predicting species distribution, with resolutions as fine as a couple of meters. Remote sensing techniques, such as satellite imagery and airborne sensors (e.g., LiDAR), provide information about the Earth's surface and its characteristics, encompassing land cover, vegetation indices, topography data, and even climatic predictors. This advancement is of utmost importance, as it has been demonstrated that fine-grain data are essential for comprehending the impacts of climate change on biodiversity, spanning from local to global scales (Lembrechts et al. 2019a, Lembrechts et al. 2019b, Zellweger et al. 2019, Stark and Fridley 2022).

Land Cover: Continuous or Binary Predictor?

In SDMs, land cover predictors or habitat types are typically incorporated by representing the proportion of specific land cover types within individual sites, such as grid cells or atlas mapping squares (Chauvier et al. 2020, Randin et al. 2020, Cánibe et al. 2022, Yang et al. 2023, Venter et al. 2023). But what if, for some species, the total habitat area is less relevant than the simple fact that a particular habitat is present or absent? We can hypothesize that the presence of a species may be influenced by the binary occurrence of a habitat, whereby the quantity of habitat within a given spatial unit becomes irrelevant, and the mere presence of the habitat becomes more crucial (Figure 2.2). However, to establish this hypothesis, it is necessary to determine the minimum or maximum percentage of habitat required to sustain a viable species population. This hypothesis assumes the existence of a threshold amount of habitat, below which the species is unlikely to occur, and above which the species is expected to persist.

Furthermore, we might assume that an increase in habitat area beyond the threshold size will not further enhance species presence (noting that habitat data resolution influences the habitat amount threshold). Such a threshold has been theorized (Andrén 1994, Fahrig 2001) and empirically observed in bird species (Melo et al. 2018). In the field of conservation biology, this concept is closely related to the notion of critical habitat area (Fahrig 2001, Melo et al. 2018), which proposes that there is a specific threshold of habitat amount below which a species cannot survive, resulting in a step-like, rather than continuous, response of species occurrence probability to habitat area. To the best of my knowledge, this possibility has not been theoretically or empirically explored within SDMs. Therefore, based on those assumptions, I investigated this possibility in my dissertation. For further details and in-depth analysis, I invite readers to refer to Chapters 3.1 and 3.2 of my dissertation.

At the end of this chapter, it is important to highlight one additional aspect. Despite the wide range of environmental predictor sources available, using these sources in predicting species distribution is not immune to uncertainties. These uncertainties can stem from various factors, such as measurement errors, limitations in spatial and temporal resolution, interpolation methods employed, or the inherent intricacies of environmental processes (Osborne and Leitao 2009, Moudrý et al. 2018, Moudrý et al. 2019, Marešová et al. 2021). Consequently, when incorporating such data into modeling frameworks, these uncertainties have the potential to propagate through the model, affecting the accuracy and precision of species

distribution projections. It is, therefore, imperative to acknowledge and quantify these uncertainties to ensure the robustness and reliability of SDMs.

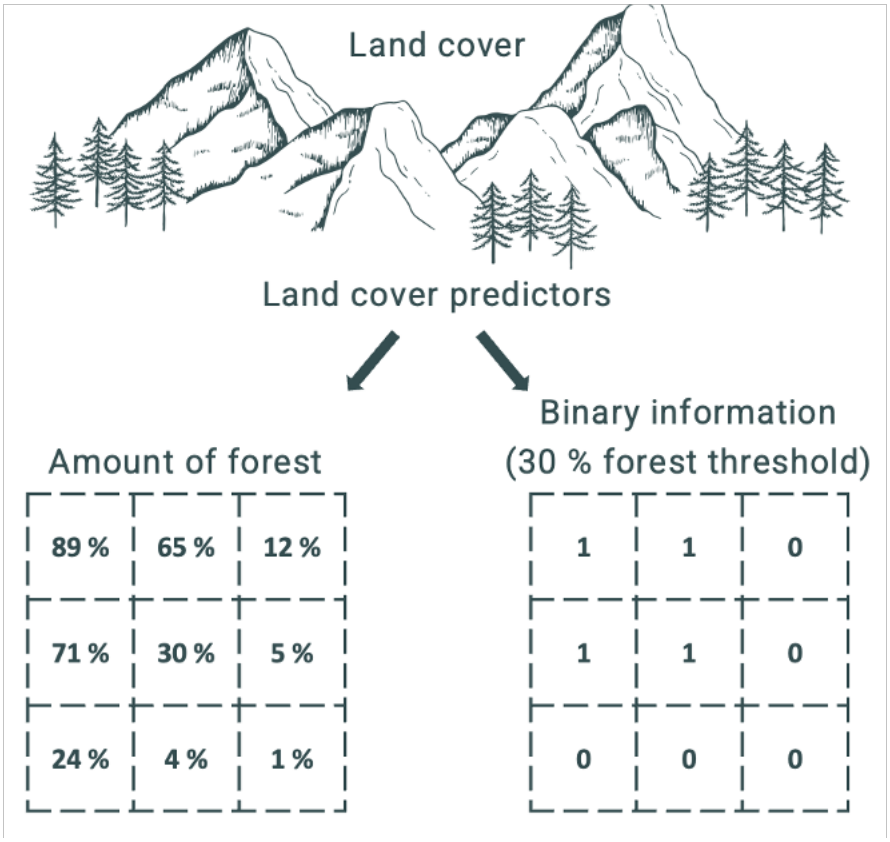


Figure 2.2: Graphical representation of the hypothesis. Land cover predictors are typically incorporated by representing the amount of specific land cover types within individual sites. But what if, for some species, the total habitat area is less relevant than the simple fact that a particular habitat is present or absent?

2.2.2. Species records

In SDMs, species records are typically represented as presence-only or presence-absence data. Presence-only data include records that only indicate the presence of a species at a particular location without information on its absence. On the other hand, presence-absence data provide information on the presence and absence of a species at specific sites. While presence-absence data are considered more informative for modeling species distributions, they are often scarcer and require careful sampling design to ensure adequate representation of absence locations (Franklin 2010, Guillera-Arroita et al. 2015). Species records can be obtained from various sources, including field surveys, citizen science initiatives, museum collections, and literature reviews, or generated based on expert knowledge.

Field surveys involve systematic sampling or observations researchers conduct to record species' presence or absence at specific locations. These surveys often employ various sampling techniques, such as transect surveys, point counts, quadrat sampling, or trapping meth-

ods to gather species data. Field surveys provide direct and reliable data collected specifically for the study. They allow for standardized data collection protocols and controlled sampling efforts, enabling researchers to gather accurate and comparable information. Additionally, field surveys offer the advantage of collecting additional ecological information, such as habitat characteristics or species behavior, which can provide valuable insights into species-environment relationships (Legendre 2002). However, there are some disadvantages associated with field surveys. They can be time-consuming and resource-draining (expensive), particularly when conducting large-scale studies requiring extensive data collection. Additionally, field surveys have limited spatial coverage, as they are often undertaken in specific study areas or sampling sites, which may not capture the full extent of species distributions. This can result in an incomplete representation of species occurrences, particularly for rare or elusive species that may be missed during surveys (Monk 2014, Wheater et al. 2020).

Citizen science initiatives engage the public in scientific data collection, allowing participants, often with limited scientific training, to contribute observations of species records through dedicated platforms or mobile apps designed for citizen science projects (e.g., eBird). They offer valuable advantages, including public involvement and education in scientific research and the ability to cover extensive geographic areas and gather data on common species (Kosmala 2016). Nonetheless, there are certain considerations associated with citizen science initiatives engage. One important factor is the potential variation in data quality due to observer skills and expertise differences. Participants may have varying levels of knowledge and experience, which can impact the accuracy and reliability of their observations. Additionally, there is a potential for biased sampling, as participants may be more likely to report charismatic, easily recognizable species or species in easily accessible locations. It is also worth noting that citizen science data may lack detailed ecological information or standardized data collection protocols, which can challenge data consistency and comparability (Fraisl 2022).

Natural museums have extensive collections of specimens collected over many decades. These collections include species occurrence information and preserved specimens, such as skins, bones, or DNA samples, which can provide valuable insight into extinctic or living species. Such data can be digitized and used as valuable input for SDMs. One advantage of utilizing museum collections is their inclusion of extinctic or living specimens, providing access to long-term data that can shed light on species distributions over time. These collections also offer taxonomically verified species records, ensuring the accuracy and reliability of the data. By incorporating museum specimens into SDMs, researchers can gain insights into the historical distribution patterns and understand how they have evolved over time (Brooke 2000, Graham et al. 2004). On the other hand, it is important to acknowledge certain limitations associated with museum collections. These collections are limited to the species represented within the museum holdings, which may not encompass the full range of species present in a study region. Additionally, the spatial and temporal coverage of museum collections can be patchy or biased towards specific regions or time periods, potentially impacting the representation of certain species or timeframes. Furthermore, the precise geolocation and comprehensive ecological information associated with museum specimens may be lacking, which can present challenges in accurately capturing the environmental context of species occurrences (Lister 2011).

Scientific publications, including research articles, reports, and species atlases, often report species occurrence data collected through various studies or monitoring programs. One

advantage of utilizing scientific publications is accessing data from multiple studies and reports, providing a wide range of species occurrence information. These publications can offer historical or regional perspectives on species distributions, allowing researchers to examine changes over time or variations across specific geographic areas. Furthermore, scientific publications may provide detailed ecological context and supporting information, enriching the understanding of species occurrences (Hampton et al. 2015, Michener 2015). Nevertheless, it is essential to consider certain limitations when utilizing data from scientific publications. Data availability and quality may vary across different publications, as some studies may provide more comprehensive and reliable data than others. Additionally, the scope and focus of the reviewed literature may limit the breadth and depth of available species records. It is crucial to acknowledge the potential for publication bias, as negative or non-significant results may not be published (Culina et al. 2018).

Understanding Challenges in Species Data

Each source of species data has its strengths and limitations, and the selection of data sources should be carefully considered based on the research objectives, spatial scale, and data availability. Combining multiple data sources can help mitigate the limitations and enhance the overall robustness of SDMs.

Species records obtained from the aforementioned sources are frequently consolidated and aggregated in online, public, global databases such as GBIF, eBird, or iNaturalist. These databases play a critical role in enhancing species records' spatial and taxonomic coverage. They serve as centralized and accessible repositories, providing researchers with a valuable resource for studying species distributions. One of the key advantages of these databases is their ability to aggregate data from multiple sources, thereby increasing the overall comprehensiveness and representativeness of the species records. By bringing together data from various contributors and organizations, these databases facilitate data sharing and collaboration among researchers, promoting a broader understanding of species distributions on a global scale. However, it is important to acknowledge that data quality can vary across sources and contributors. Including duplicate records is not uncommon, necessitating careful data cleaning and validation procedures (Zizka et al. 2020).

Furthermore, spatial bias can arise due to uneven sampling efforts or data gaps in certain regions (Rocchini et al. 2023). In addition, many occurrences may exhibit inherent positional uncertainties (Moudrý and Devillers 2020). Positional uncertainty can introduce bias in model predictions, as inaccurate or imprecise occurrence locations can lead to incorrect estimations of species-environment relationships.

The detrimental impact of positional uncertainty in geographical coordinates on spatial modeling is well-established. As spatial modeling tools have become increasingly utilized in ecological studies, the question of the utility of species occurrences with positional uncertainty has emerged. This has sparked a longstanding debate among ecologists regarding the value and reliability of such data. Fascinatingly, previous studies investigating the influence of positional uncertainty on species distribution predictions concluded conflicting findings, thereby adding to the issue's complexity. While some studies have reported that positional uncertainty leads to a decrease in model performance (Johnson and Gillingham 2008, Osborne and Leitao 2009, Lash et al. 2012, Tulowiecki et al. 2015, Mitchell et al. 2017, Soultan and Safi 2017, Zhang et al. 2018, Fernandes et al. 2019), others have reached contrasting conclusions (Graham et al. 2008, Fernandez et al. 2009). These opposing results can be explained using environmental predictors with different heterogeneity (Naimi et al. 2011,

2014, Gábor et al. 2023) or species with varying niche widths (Gábor et al. 2020a, Gábor 2023). Although previous studies thoroughly assessed the effect of positional uncertainty on the model predictive performance, the question of how positional uncertainty in species records affects models' parameter estimation (species–environment relationships inference) remains largely unexplored. Therefore, in my dissertation, I explored the extent to which parameter estimation is affected by positional uncertainty. Specifically, collaborating with my esteemed co-authors, we explored the impact of positional uncertainty on key aspects such as variable importance and the shape of response curves (see Chapter 3.3).

Researchers commonly try to mitigate or minimize the negative impact of positional uncertainty in species records. One approach involves georeferencing species records based on the adopted analysis grain, ensuring that the occurrence locations align with the desired spatial scale of analysis (Ballesteros-Mejia et al. 2017). Another strategy might be integrating imprecise species and high-accuracy records (Reside et al. 2011; Figure 2.3). Alternatively, when working with already georeferenced records, researchers may opt to remove imprecise occurrences, such as records with lower latitude and longitude precision (e.g., less than three decimal places) or those known to have high positional uncertainty (Gueta and Carmel 2016, Watcharamongkol et al. 2018, Ellis-Soto et al. 2021; Figure 2.3). However, removing positionally inaccurate records can result in a reduced sample size and potentially diminish the model's explanatory power (Smith et al. 2023). Another strategy employed to mitigate the negative influence of positional uncertainty is to coarsen the analysis grain (Moudrý and Šímová 2012, Keil et al. 2014, Sillero and Barbosa 2021; Figure 2.3).

Nevertheless, it is important to note that this approach also reduces the sample size. In addition, species' response to the environment might be better captured at a finer grain (Guisan et al. 2006, Merow et al. 2014). Both factors can have undesirable effects on the performance of SDMs. It is worth noting that while this approach has a theoretical basis, it has not been extensively tested in practical applications. Therefore, I decided to investigate this approach to bridge a gap in our understanding of SDMs (for more details, see Chapter 3.4).

2.3. Spatial scale

Scale is a fundamental concept in species distribution modeling, influencing model outcomes' accuracy and ecological interpretation. Understanding and addressing scale-related issues in SDMs is crucial for obtaining reliable predictions and gaining insights into species-environment relationships (Pearson and Dawson 2003, Elith and Leathwick 2009, Seo et al. 2009, Bradter 2012, Conor et al. 2018, Moudrý et al. 2023b). In modeling species distribution, three specific types of spatial scale are important. Firstly, the response grain refers to the scale at which patterns or processes occur. Secondly, the observational scale, which relates to the characteristics of the data, is usually defined by the spatial resolution and extent of the data. Lastly, the analysis scale relates to the methods for analyzing environmental data, including factors such as the neighborhood size used in focal statistics or geomorphometry (Dungan et al. 2002).

There is an ongoing debate revolving the spatial scales (grains) at which ecological processes underlying species distribution patterns operate (Miguet et al. 2016, Mertes and Jetz 2018). SDMs can be developed across a broad range of grains, from a few meters up to thousands of kilometers. Prior studies have documented the influence of analysis grain on the performance of SDMs (Guisan et al. 2007, Kaliontzopoulou et al. 2008). Besides this, it has been

demonstrated that species exhibit stronger responses to environmental factors at certain spatial scales than at others (Holland et al. 2004, McGarigal et al. 2016). This phenomenon is often called ecological scale or response grain (Holland et al. 2004, Mertes and Jetz 2018, Moudrý et al. 2023b).

Ideally, the analysis grain should align with the species' ecology and the study's objectives, ensuring compatibility with the response grain (Mertes and Jetz 2018). However, this may be influenced by the sampling processes of species occurrences (Hurlbert and Jetz 2007, Chase and Knight 2013), as well as the spatial extent of the study area. The spatial extent and resolution of the response variable play a crucial role in determining which explanatory variables are expected to contribute to species distribution (Pearson and Dawson 2003). Typically, climate is assumed to define species distribution at broad spatial scales, such as the extent of an entire continent. As the resolution becomes finer and at regional extents, topography or biotic interactions may emerge as the primary factors influencing species occurrence. Further, at even finer resolutions, factors such as vegetation structure or the presence of specific land cover categories (e.g., water bodies) may come into play. Generally, the significance of environmental factors varies with the chosen resolution and extent of the study, and factors that are important at one resolution and extent may lose their significance at others (Corsi et al. 2000).

The choice of the analysis grain in species distribution modeling typically falls into two scenarios: (a) when the response grain is known and fine-scale data is available, and (b) when the response grain is unknown, and the analysis grain is determined based on data availability rather than species ecology (Martin and Fahrig 2012, Stuber and Fontaine 2019, Mertes et al. 2020). As the choice of analysis grain directly affects our ability to detect species' responses to environmental variables, various factors need to be taken into consideration, including positional errors of species occurrences, resolution of available environmental data, and the expected response grain at which species are anticipated to respond to the environment (Lechner et al. 2012, Lecours et al. 2015).

Because, in theory, species distribution is driven by environmental variables at a range of scales (Levin 1992), and there is no correct spatial grain at which to characterize species-environment associations (Wiens 1989), some studies suggest developing SDMs using multi-grain environmental predictors (Mertes et al. 2020, Riva et al. 2023, Silveira et al. 2023). By conducting multi-grain analyses, researchers can better understand the scale dependence of species-environment relationships and identify the most appropriate spatial grains for modeling a particular species as such approach allows the selection of the best scale separately for each predictor variable significantly affecting species occupancy (Meyer 2007, Mazziotta et al. 2024).

Although I support the suggested approach, it is important to acknowledge that previous research has yielded varied conclusions regarding the utility of multi-grain analysis approaches. While some researchers have reported improved performance of models utilizing multiple environmental predictor grains compared to single-grain models (Mertes et al. 2020), others have not reached similar conclusions (Martin and Fahrig 2012). It is important to recognize that the observed improvements in multi-grain models are often relatively modest (Graf et al. 2005, Mateo Sánchez et al. 2014, Moudrý et al. 2023).

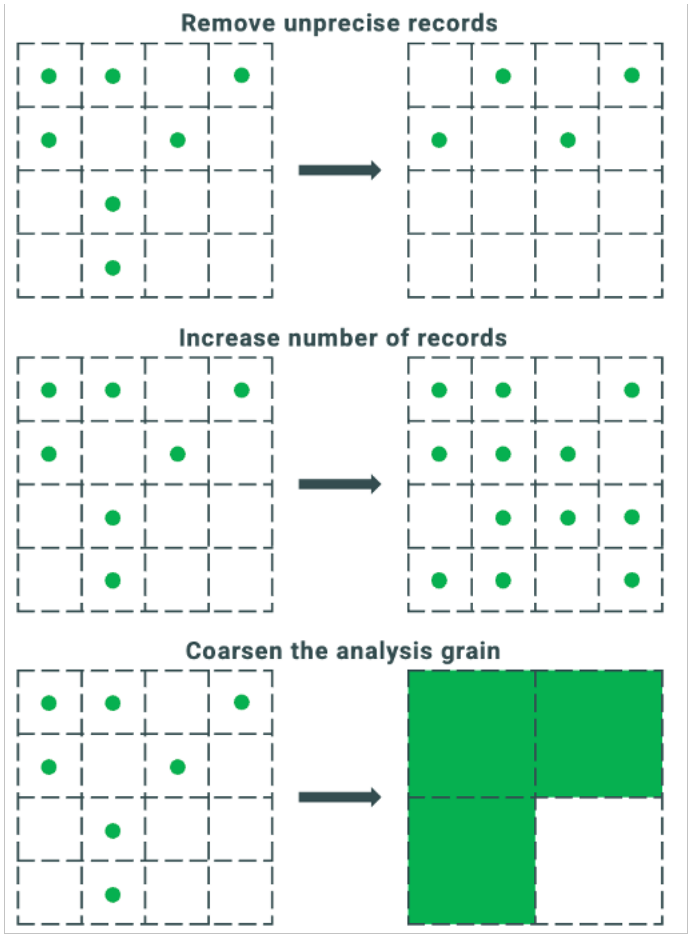


Figure 2.3: Graphical representation of common techniques to mitigate the negative effect of positional uncertainty. Previous studies have proposed several strategies to reduce the negative impact of positional uncertainty in species records. These strategies include, for example, removing imprecise occurrences, increasing the number of occurrences by combining highly accurate records with imprecise ones and coarsening the analysis grain to account for maximal positional uncertainty in the species dataset.

The scale at which species data uncertainties, for instance, positional uncertainty or uneven sampling, are examined plays a crucial role in understanding their negative effects in SDMs. For example, the impact of uneven sampling, where species occurrences are not evenly distributed across the study area, can vary depending on the scale of analysis. At larger scales, uneven sampling may have less impact on model results, as the aggregated species data can still provide meaningful insights into species-environment relationships. However, uneven sampling can lead to biased model outcomes at smaller scales, as the limited representation of certain habitats or regions may skew the results (Fithian et al. 2014, Dubos et al. 2022). Similarly, positional uncertainty can have different implications at different scales. At broad spatial scales, positional uncertainty may have limited effects on the overall model outcomes, as the coarse resolution of the analysis grain might overshadow the impact of positional

uncertainties. However, at fine spatial scales, positional uncertainties can significantly influence model predictions, especially in areas with complex or heterogeneous landscapes (Naimi et al. 2011, Moudrý and Šímová 2012, Naimi et al. 2014, Gábor et al. 2022b, Gábor et al. 2023).

Therefore, it is crucial to consider the scale at which species data uncertainties are examined and to investigate interactions of species data uncertainties with various scales. By studying these interactions, we can better understand the relative importance of species data uncertainties and their scale-dependent effects on SDMs.

2.4. A brief advocacy for the virtual species approach

In this chapter, I aim to briefly introduce the virtual species approach and express my strong support for its use. I firmly believe it offers the best possible means to investigate the effects of data inaccuracies on SDM outcomes comprehensively. By creating virtual species with predetermined traits, such as preferred habitat requirements, dispersal abilities, or sensitivity to environmental factors, researchers can gain a better understanding of how these traits influence species responses to changing environmental conditions and, in addition, how they affect models' performance (Meynard and Kaplan 2013, Zurell et al. 2010, Miller 2014, Moudrý 2015, Leroy et al. 2016, Meynard et al. 2019, Gábor et al. 2020b).

Additionally, the virtual species approach has emerged as a useful tool for investigating the impact of input data inaccuracies on SDMs. Input data inaccuracies encompass a range of errors, uncertainties, and limitations associated with species records and environmental predictors. These inaccuracies can be designed to simulate real-world scenarios as errors in species occurrence records, such as species misidentifications, incomplete or biased sampling, or imprecise geolocation (Mitchell et al. 2017, Gábor et al. 2020a, Gábor et al. 2020b, Inman et al. 2021, Collart and Guisan 2023, Gábor et al. 2023, Marsh et al. 2023). By constructing SDMs using artificial data that incorporates these inaccuracies, researchers can evaluate the propagation of such imperfections throughout the modeling process and assess their influence on model accuracy and ecological interpretations of the results.

In this dissertation, I employed the virtual species approach to simulate positional uncertainty in species records, enabling a comprehensive examination of the implications of species data positional inaccuracies on the ecological interpretability of SDMs. Additionally, a key focus of this research was to investigate how coarsening the resolution of environmental predictors could mitigate the negative effects of positional uncertainty in species records. By integrating the virtual species approach and exploring this specific aspect, a deeper understanding was gained regarding the interplay between input data uncertainties, environmental predictors, and the resulting accuracy and reliability of SDMs (see Chapters 3.3, 3.4).

It is important to note that while the virtual species approach presents a valuable tool, it is not intended to replace using real species. For example, in Chapter 3.3, virtual species simulations showed a rapid decrease in model performance with increasing positional error. In contrast, the real species models only slightly reduced model performance. Therefore, I strongly recommend that future studies follow a growing trend and combine simulations and real species data when studying methodological questions (Fithian et al. 2015, Guélat and Kéry 2018, Mertes and Jetz 2018, Renner et al. 2019, Rocchini et al. 2023).

Research studies

3.1. Habitats as predictors in species distribution models: Shall we use continuous or binary data?

Lukáš Gábor, Petra Šímová, Petr Keil, Alejandra Zarzo-Arias, Charles J. Marsh, Duccio Rocchini, Marco Malavasi, Vojtěch Barták and Vítězslav Moudrý

Adapted from Ecography, 2022(7), e06022.

Publication metrics:
Quartile (2023): Ecology, Evolution, Behavior and Systematics (Q1)
Impact Factor (2023): 6.08
SCImago Journal Rank (2023): 2.37

The first author contributed to the study as follows: study conception and design (lead), data curation (lead), analysis and interpretation of results (lead), visualization (lead), writing – original draft (lead), writing – review and editing (equal), overall study supervision (equal), funding acquisition (lead).

The link to the published article and supplementary materials can be found here:
https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.06022

3.1.1. Abstract

The representation of a land cover type (i.e., habitat) within an area is often used as an explanatory variable in species distribution models. However, it is possible that a simple binary presence/absence of the suitable habitat might be the most important determinant of the presence/absence of some species and, thus, be a better predictor of species occurrence than the continuous parameter (area). We hypothesize that the binary predictor is more suitable for relatively rare habitats (e.g., wetlands) while for common habitats (e.g., forests) the amount of the focal habitat is a better predictor. We used the Third Atlas of Breeding Birds in the Czech Republic as the source of species distribution data and CORINE Land Cover inventory as the source of the landcover information. To test our hypothesis, we fitted generalized linear models of 32 water and 32 forest bird species. Our results show that for water bird species, models using binary predictors (presence/absence of the habitat) performed better than models with continuous predictors (i.e., the amount of the habitat); for forest species, however, we observed the opposite. Thus, future studies using habitats as predictors of species occurrences should consider the prevalence of the habitat in the landscape, and the biological role of the habitat type in the particular species' life history. In addition, performing a preliminary comparison of the performance of the binary and continuous versions of habitat predictors (e.g., using information criteria) prior to modelling, during variable selection, can be beneficial. These are simple steps that will improve explanatory and predictive performance of models of species distributions in biogeography, community ecology, macroecology, and ecological conservation.

Keywords: Binary data, Continuous data, Land cover, Niche models, Variable selection

3.1.2. Introduction

Species distribution models (SDMs) are an important tool in macroecology, biogeography, and wildlife management. The goal of SDMs is to map species distributions or to estimate species niches, and there is an ongoing effort to improve their reliability (Araújo et al. 2019, Zurell et al. 2020, Merow et al. 2022). Selecting appropriate environmental predictors is a major methodological challenge of species distribution modeling (Dorman et al. 2007, Austin and Van Niel 2011, Williams et al. 2012, Mod et al. 2016, Misiuk et al. 2018, Smith and Santos 2020, Zurell et al. 2020). These environmental predictors, such as landcover or habitat type, are most often included in SDMs as the area or percentage of a particular land cover type within the individual sites (e.g., grid cells or atlas mapping squares; see for example Milanesi et al. 2017, Halstead et al. 2019, Lecours et al. 2020, Tessarolo et al. 2021).

But what if, for some species, the total area of habitat is less relevant than the simple fact that a particular habitat is present or absent? To our knowledge, this possibility has been considered neither theoretically, nor empirically. In conservation biology, this is somehow related to the concept of critical habitat area (Fahrig 2001, Melo et al. 2018), i.e., to the idea that there is a certain habitat amount (threshold) below which a species cannot survive, leading to a step-like, rather than continuous, response of species probability of occurrence to habitat area. To our knowledge this has not been explored in the context of SDMs. Further, a guideline on whether habitat predictors should be included in SDMs as continuous, or binary variables would be directly applicable in many subfields of biogeography and community ecology.

In this study, we evaluate the effect of using forest and water habitats as binary or continuous predictors in species distribution modelling of 64 forest and water specialist bird species. Specifically, we propose two alternative hypotheses linking the probability of occurrence (P) of a species to either (a) the amount or (b) the presence/absence of a particular habitat within a spatial unit (e.g., grid cell).

The first hypothesis (H1) assumes that P is driven by continuous areas (see Figure 3.1a), i.e., that P increases continuously with the increase in the habitat area within a spatial unit. Reasoning supporting H1 is as follows: (i) Larger habitat areas support larger populations due to their carrying capacity and food and shelter availability, so that populations are less susceptible to stochastic extinctions, competition, predation, and inbreeding depression (Hanski 1999, Lande et al. 2003), and (ii) larger habitat areas are bigger targets for colonizing individuals from the surrounding habitat matrix (Buckley and Knedlhans 1986), increasing the probability of rescue effects after extinction events (Brown and Kodric-Brown 1977). We propose that these mechanisms will operate in the most common habitat types. H1 will also apply to species specializing in these common habitats. In Central Europe, forests can be considered an example of such habitats, with forest specialist species such as the Long-tailed tit (*Aegithalos caudatus*), the Goldcrest (*Regulus regulus*) or the Crested tit (*Parus cristatus*).

An alternative hypothesis (H2) is that P is driven by binary presence or absence of a habitat (Figures 3.1b, 3.2b). In other words, the amount of habitat within a spatial unit is irrelevant, and what matters for the species is that the habitat is simply there.

However, we first need to know how small, or large such habitat needs to be to be able to support a viable population of the species. Therefore, H2 assumes that there is a threshold of habitat amount, (e.g., 20 % as in Figures 3.1b, 3.2b), below which the species is unlikely to occur, and above the threshold the species will persist. We assume that an increase of the

habitat area beyond the threshold size will not increase P (note that the threshold of habitat amount is affected by resolution of the habitat data). The presence of such threshold has been predicted both theoretically (Andrén 1994, Fahrig 2001) and documented empirically for birds (Melo et al. 2018).

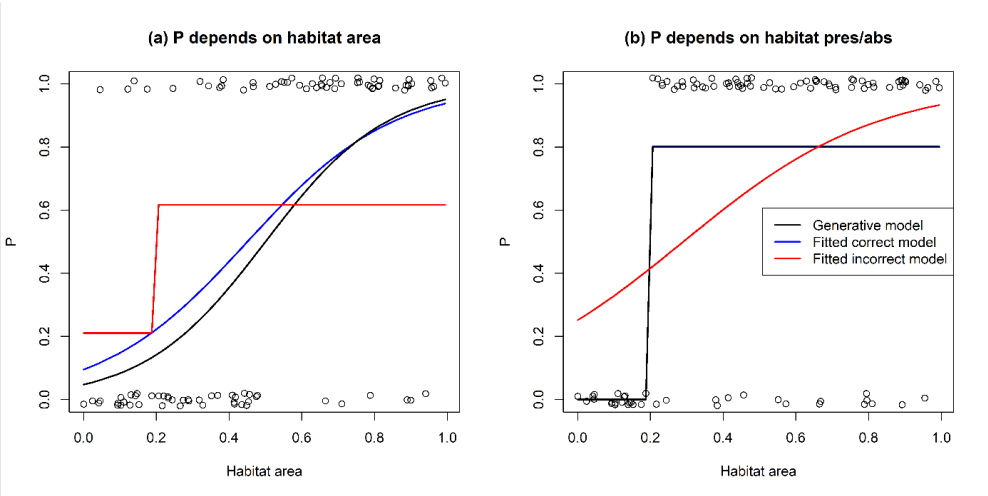


Figure 3.1: Two alternative hypotheses for the effect of the habitat area on the species' probability of occurrence (P), illustrating the theoretical possibility that habitat area be an accurate or inaccurate predictor of species distributions depending on if it is fitted as a continuous or binary variable and the process that generated the data. In the left panel we modelled P as a sigmoidal curve (generative model, black line) to generate 100 presences/absences of a species, drawn from a Bernoulli distribution with parameter P (jittered points). In right panel we used a binary habitat classification and a step function to generate the data. We then fitted binomial GLM with either continuous area, or binary area, as predictor (red and blue lines).

We propose that H2 applies particularly to species specializing in rare (i.e., less prevalent) habitats, and species with good dispersal abilities and ability to readily identify the habitats in the landscape. Consequently, if a fragment of suitable habitat (irrespectively of its area) appears in the landscape, it will quickly attract a population of the species, thus causing high P. In Central Europe, water bodies can be considered an example of such habitat for water specialist species such as the Common teal (*Anas crecca*), the Great-crested grebe (*Podiceps cristatus*), or the Black tern (*Chlidonias niger*).

3.1.3. Material and Methods

Study area and bird distribution data

The study area was the territory of the Czech Republic, a central European country covering almost 79,000 km^2 (see Figure 3.2a). Data on bird species were obtained from the Third Atlas of Breeding Bird Distribution in the Czech Republic (Šťastný et al. 2006). The study area is divided into 628 grid squares of approx. 134 km^2 (10' east longitude × 6' north latitude; hereafter referred to as mapping squares) to which bird

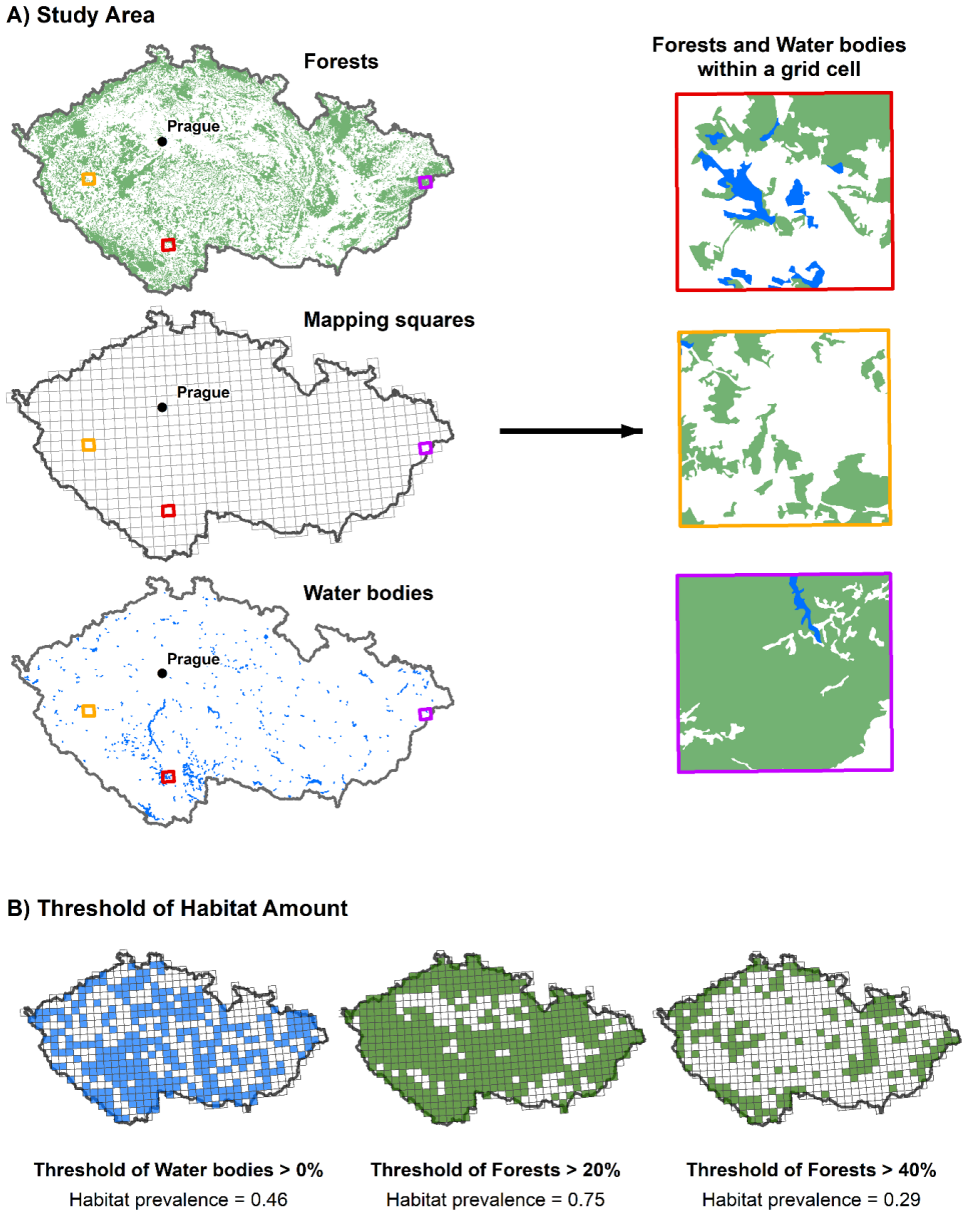


Figure 3.2: A) The study region covers the territory of the Czech Republic, Europe. The grid consists of cells of 10' east longitude × 6' north latitude (approximately 12 × 11.2 km, n = 628), as used in the breeding birds atlases of the Czech Republic (Šťastný et al. 2006). Water bodies shown on the left side have a 500 m buffer for better visibility. B) Representation of binary variables based on different threshold of habitat amount. We considered any amount of the water habitat in a cell as presence (i.e., the proportion of the cell occupied by one hectare set to > 0 %). In addition, we tested several thresholds (e.g., 20 % and 40 %,) to derive the binary predictor for forest variables.

occurrences and environmental predictors are referred. The fieldwork for the atlas was conducted by volunteers between 2001 and 2003 where the breeding status of all species was recorded in each mapping square. Field observations of the bird species occurring in each mapping square were recorded using 17 numerical breeding codes (Hagemeijer and Blair 1997). Breeding occurrence of each bird species within a given mapping square was classified into one of the following categories: 0 – Non-breeding (where no observations of the species were made, or where the species was observed but no breeding evidence was found), A – Possible breeding, B – Probable breeding or C – Confirmed breeding. For the purpose of our study, all breeding categories (A, B and C) were used as presences whereas category 0 was used as absences. We prepared data for 85 bird species, 36 of them nesting in wetlands and surrounding habitats (e.g., standing water, littoral zones of ponds, swamps), and 49 species nesting in forests, following classification of Reif et al. (2006). Nevertheless, we had to remove 21 species with relatively small (less than 30 presence cells out of 628 cells), and relatively high occupancy (more than 598 presence cells out of 628 cells). Therefore, 32 water and 32 forest bird species were included in the study (see Supplementary material Table A1).

Habitat variables

We derived four habitat predictors from the CORINE Land Cover database at 100 m resolution (Feranec et al., 2010). Specifically, within mapping squares, we derived the area of *agricultural areas* (CORINE class 2), *artificial surfaces* divided into four classes (0, 0–20, 20–40, > 40 km^2; CORINE class 1), *continuous area of water bodies* (CORINE class 5.1.2) and *area of forest* (CORINE class 3.1). In addition, binary factors representing presence or absence of *water bodies* and *forests*, respectively, were calculated. In order to generate binary habitat maps, it is necessary to determine an area threshold that defines the presence-absence of the habitat. An appropriate threshold should consider the prevalence of the habitat across the region of interest, the grain size at which the variable is being considered (i.e., the size of the grid cells at which the species are recorded) and the original grain size that the habitat variable is being aggregated from (i.e., the size of the grid cells of the original land-cover data, which is then aggregated to the larger modelling grain size). Due to the uncommonness of water habitats as well as due to the coarse resolution of CORINE Land Cover, we considered any amount of the water habitat in a cell as presence (i.e., the proportion of the cell occupied by one hectare set to > 0 %). Forest pixels are, on the other hand, present in all mapping squares across the study region and, for this reason, we tested several thresholds (10 %, 20 %, 30 %, 40 %, and 50 %) to derive the binary predictor.

Other environmental variables

Although the habitat predictors were our main focus, other predictors, such as climate, may also be important in determining the distributions of species. As climatic predictors, we used current climatic data from WorldClim (Hijmans et al., 2005). Following previous studies, we used two predictors: mean temperature and mean precipitation during the breeding season, i.e. in April–June (e.g., Moudrý and Šímová 2013, Venne and Currie 2021). We downloaded these at a resolution of 30 arc seconds (approximately 1 km^2) and averaged them inside each mapping square to match the grid resolution of the species distribution data (approximately 100 km^2). We also considered usage of elevation predictors such as maximum, minimum, and range of elevation derived from Shuttle radar topography mission (SRTM, Farr et al. 2007; Moudrý et al. 2018) as they might be ecologically important to birds (e.g., Kosicky 2017). However, as these variables were highly correlated with the mean temperature in April–

June, we eventually decided not to include them. The data were processed in ARCGIS 10.7.1 (ESRI, CA, USA) and R (R Development Core Team) software.

Species distribution models

We fitted SDMs for each species using the climate variables and the water (or forest cover) variable for water (or forest) species; modelling was always performed separately for the continuous as well as binary water (or forest) variable. We did not use the information about forest areas for wetland species and vice versa (see Supplementary material Table A2 for used formulas). In addition, we also considered using forest and wetland area transformed with arcsine, log, or square root, all of which are often adopted in ecological studies for areal predictors (for more details, see Roberts 1986 and Palmer 1993). However, these transformations did not improve the models and were not further considered in this study. We used generalized linear models (GLMs; McCullagh and Nelder, 1989), with binomial error distribution and a logit link function implemented in the R function glm. Environmental predictors were used as monotonic sigmoidal functions on the probability scale of the response (i.e., linear in logit space).

Model calibration and evaluation

We assessed the performance of the models using calibration and discrimination metrics where calibration refers to the accuracy of description of the environmental relationships, and discrimination refers to the ability to separate presences from absences (Lawson et al. 2014). We used five-fold cross-validation where the data were randomly divided into fifths 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. To assess model calibration, we used likelihood-based McFadden's pseudo R^2 (Smith and McKenna 2013), which indicates the proportion of the deviance in the dependent variable that is explained by the model (Agresti 2003). To assess the model discrimination ability, we used the area under the curve of the receiver operating characteristic plot (AUC). The AUC is a threshold independent measure of model performance that ranges from 0 to 1, where a score of 1 indicates perfect discrimination, and a score of 0.5 indicates random performance (Fielding and Bell 1997).

3.1.4. Results

The occurrence of water birds was better modelled using the binary variable (prevalence of water habitat = 0.4; see Figures 3.3, 3.4, Supplementary material Table A3), suggesting that their distribution is driven simply by presence rather than the area of water habitat. In contrast, the models with continuous environmental variables outperformed those using binary predictors in modelling forest birds (Figures 3.3, 3.4). This result was observed independently of the forest amount threshold for most of the species (Supplementary material Table A4); however, to maintain clarity, we present results of models fitted with a 40 % threshold (prevalence of forest habitat = 0.29; see Figures A1 – A4 in the Supplementary material for results using remaining tested thresholds).

For forest species, models fitted using the area of forest (i.e., a continuous habitat variable) achieved poor to excellent model calibration (R^2: min = 3.34 %, max = 42 %, mean = 16.2 %) and discrimination performances (AUC: min = 0.60, max = 0.91, mean = 0.74; see Table A4 in the Supplementary material for the performance of individual models). Both model calibration (R^2: min = 2.69 %, max = 38.42 %, mean = 14.3 %) and discrimination (AUC: min = 0.6, max = 0.89, mean = 0.72) were lower in models using the forest presence (i.e., a binary

habitat variable) in 28 out of 32 species, although differences in model performances were relatively small. The differences in model calibration between models fitted using the area of forest and forest presence were negligible, except for six species where R^2 increased by up to 7 % (Figure 3.4). Similarly, the difference in AUC was <0.02 for 22 out of 32 species. The highest AUC differences (0.07) were recorded for Long-tailed tit (*Aegithalos caudatus*), Black redstart (*Phoenicurus ochruros*), and Spotted flycatcher (*Muscicapa striata*).

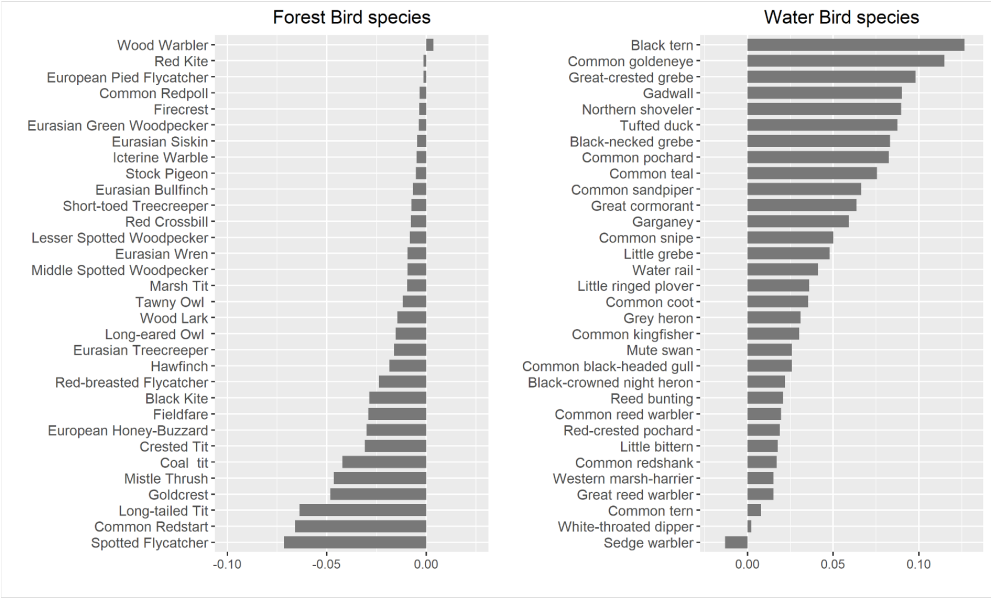


Figure 3.3: Differences in AUC between models fitted with binary habitat presence/absence vs. continuous area as predictors. Positive values indicate that models with binary habitat predictors performed better than those with continuous predictors and vice versa.

Models fitted for water species using the water presence performed better in both calibration (R^2: min = 9.62 %, max = 38.62 %, mean = 22.1 %) and discrimination (AUC: min = 0.69, max = 0.91, mean = 0.79; see Supplementary material Table A3 for the performance of individual models) compared to those using area of water (R^2: min = 5.05 %, max = 37.29 %, mean = 16.96 %; AUC: min = 0.63, max = 0.89, mean = 0.75) in nearly all cases. R^2 was on average 5 %, and up to almost 15 %, higher when using the water presence (Figure 4.3). For 13 out of 32 water bird species, model discriminations (AUC) were increased by >0.05 when considering the water presence compared to the area of water. In three cases (Black tern, *Chlidonias niger*; Common goldeneye, *Bucephala clangula*; and Great-crested grebe, *Podiceps cristatus*) the improvement in AUC was close to or even greater than 0.1. The model using the area of water was superior to that using the water presence only for a single species (Sedge warbler, *Acrocephalus schoenobaenus*).

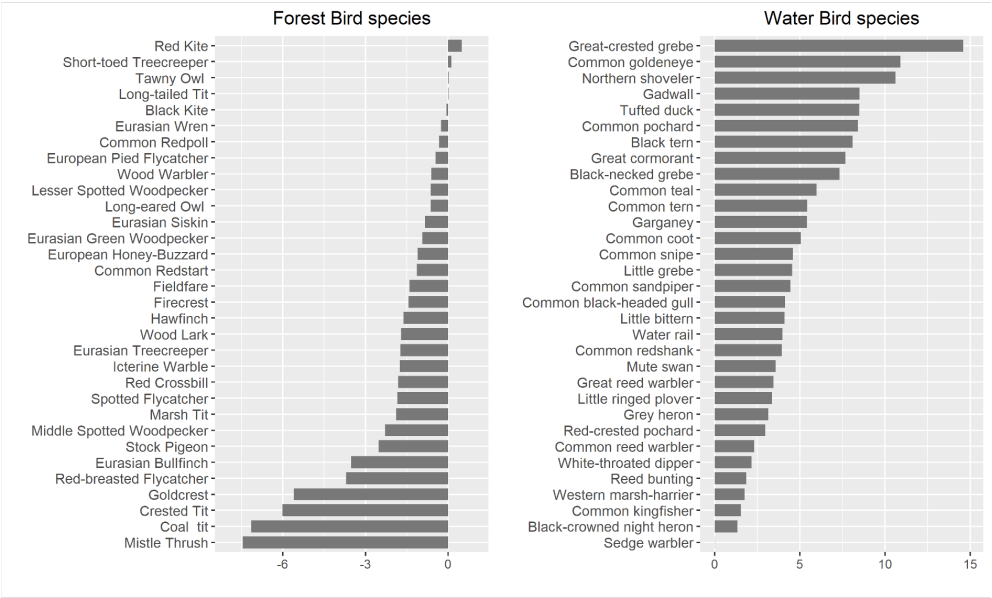


Figure 3.4: Differences in R^2 (%) between models fitted with binary habitat presence/absence vs. continuous area as predictors. Positive values indicate that models with binary habitat predictors performed better than models with continuous predictors and vice versa.

3.1.5. Discussion

Our results are in line with the hypotheses presented in the introduction. As expected, for species for which a widespread habitat (e.g., forest) is sufficient, models discriminated presences from absences better and explained more variability when a continuous, not binary, measure of the forest (habitat) area was used. On the other hand, the opposite was found for species specializing in a relatively rare habitat—water. In this case, models using water as a binary predictor outperformed those with water as a continuous area. As we have suggested, there are biological reasons for this: The relationship between species biology and specific habitat (and its rarity) determines how a binary habitat predictor stands out against continuous one. For instance, in waterbirds, the presence or absence of wetland or water habitats, which worked well in binary models, is directly related to food and shelter availability (Wiens 1989, Weller 1999, Gatto et al. 2008). Moreover, highly mobile waterbird species such as Common redshank (*Tringa totanus*), Common tern (*Sterna hirundo*), or Gadwall (*Anas strepera*) are able to spot such habitat and colonize it, even if the habitat is rare and isolated in an otherwise dry landscape matrix. Thus, a patch of isolated wetland within a grid cell is almost guaranteed to host the species, despite the habitat being rare. We argue that such biological reasoning should precede any decisions about the specific form (binary or continuous) of predictors in SDMs. However, future studies are needed to show if this explanation based on habitat rarity applies to other environments, habitats, and taxa.

In addition, we propose that the relative merit of continuous vs binary predictors depends on the interplay between spatial resolution of the habitat data (Domisch et al. 2015, Friedrichs-Manthey et al. 2019), spatial grain at which habitats are aggregated for modelling (response grain; see for example Seoane et al. 2004, Venier et al. 2004, Convertino et al. 2011, Tuanmu and Jetz 2014, Šímová et al. 2019), as well as the home range size of the species, and its

degree of specialisation to the habitat (Jedlikowski et al. 2016, Mertes et al. 2020). For highly specialised species, the ratio between the home range size and the grain size of the response variable may be particularly important (Jedlikowski et al. 2016), as it determines whether the species can gather resources from multiple grid cells, or whether it is confined to a single cell. For example, if the area of a single cell classified as water is larger than the home range of a highly specialised species, the binary predictor (water presence) should be used. However, if that area is smaller than the species home range, considering the habitat area is preferable, as the higher is the representation of water within a grid cell, the higher is the probability that the cell contains the habitat area necessary for species persistence. In addition, our habitat predictors were derived from the CORINE database (Feranec et al. 2010), with a minimum mapping unit of 25 ha, and it may be that forest species require habitats larger than that. Indeed, SDM studies using common land use categories as predictors, such as the proportion of forests, reported low improvement using finer resolution data (Seoane et al. 2004; Venier et al. 2004). In contrast, atlas squares with binary presence of water almost always contain a substantial area of water bodies, possibly enough to support a persistent breeding population of a waterbird species, leading to the good performance of the binary water predictor. In line with this, Šímová et al. (2019) showed that the area of water bodies derived from high-resolution (30 m) datasets explain distributions of waterbirds better than predictors derived from coarser water datasets (including CORINE Land Cover). This may be a reason why Tuanmu and Jetz (2014) found the Global Consensus Land cover (1 km resolution) performed worse for water species than for species that from other environments. In addition, Seoane et al. (2004) found considerable improvement of models for riparian species when finer-resolution data were used. However, this should be further validated, especially using finer resolutions than ours. Moreover, our results suggest that a hectare of wetland may be enough to be used by many water bird species and thus in future studies water habitats with equal or larger area then one hectare can be used as presence of habitat.

Other reason for the better performance of binary wetland predictor is that the threshold for absence was 0 %, whereas for forest predictors 40 % of habitat cover. In the threshold chosen in the results, the binary predictor is higher or lower than 40 % cover of forest, which could be expected to have lower discrimination capacity than absolute absence of forest versus presence of forest. Note, however, that with the resolution of response variable approx. 12 x 11 km (used in our study), the absolute absence of forest in most of the Europe is unlikely.

It is fair to point out that only few species show considerable difference between models fitted using binary and continuous variables (i.e., the differences are relatively small for most of the species). Thus, models adopting traditional continuous variables will likely produce useful predictions. Nevertheless, in terms of practical recommendations, we advocate for testing both types of such variables during variable selection; this could be done using model selection criteria (e.g., AIC, BIC, DIC), cross-validation, or measures of model fit (R^2, AUC). In addition, some of the biological mechanisms outlined in the introduction can also help with the decision on the preferable form of the predictor, and the importance of selecting biologically meaningful habitat predictors prior to modelling cannot be overstated. To summarize, categorical predictors would be preferred for (i) highly mobile species where even a small fragment of habitat is sufficient, (ii) species for which one can expect a threshold response to the environment (e.g., at least a 20 % coverage of the habitat within a mapping square is needed for the occurrence); and (iii) highly mobile species specializing in less prevalent habitats, which could be quickly identified in the landscape and colonized.

The fact that habitat variables can, depending on the commonness of the habitat, perform

best either as a binary or a continuous predictor, has relevance beyond simple species distribution models. After all, the information on the probability of species presence is sought after in many fields, from epidemiology to metacommunity ecology. Particularly in the latter, estimation of species responses to environmental conditions (including habitats) is at the core of the assessment of the relative role of niche vs. spatial processes structuring ecological communities (Cottenie 2005, Leibold and Chase 2018). Our results suggest that if an inappropriate response of species to habitat amount (threshold vs. continuous) is used, it can lead to underestimation of the importance of the niche processes. Furthermore, our results are relevant for conservation. Specifically, it is encouraging that, for many species, the presence of a (rare) habitat above a certain threshold (Radford et al. 2005, Melo et al. 2018) is important irrespectively of its area. If the critical threshold is low, even the protection of small areas is meaningful.

3.2. Assessing the applicability of binary land-cover variables to species distribution models across multiple grains

Lukáš Gábor, Jeremy Cohen, Vítězslav Moudrý, Walter Jetz

Adapted from Landscape Ecology, 39(3), 66.

Publication metrics:
Quartile (2024): Environmental Science - Ecology (Q1)
Impact Factor (2024): 5.05
SCImago Journal Rank (2024): 1.47

The first author contributed to the study as follows: study conception and design (lead), data curation (equal), analysis and interpretation of results (equal), visualization (lead), writing – original draft (lead), writing – review and editing (equal), overall study supervision (equal), funding acquisition (lead).

The link to the published article and supplementary materials can be found here:
https://link.springer.com/article/10.1007/s10980-024-01813-3

3.2.1. Abstract

Context
Species distribution models are widely used in ecology. The selection of environmental variables is a critical step in SDMs, nowadays compounded by the increasing availability of environmental data.

Objectives
To evaluate the interaction between the grain size and the binary (presence or absence of water) or proportional (proportion of water within the cell) representation of the water cover variable when modeling water bird species distribution.

Methods
eBird occurrence data with an average number of records of 880,270 per species across the North American continent were used for analysis. Models (via Random Forest) were fitted for 57 water bird species, for two seasons (breeding vs. non-breeding), at four grains (1 km^2 to 2,500 km^2) and using water cover as a proportional or binary variable.

Results
The models' performances were not affected by the type of the adopted water cover variable (proportional or binary) but a significant decrease was observed in the importance of the water cover variable when used in a binary form. This was especially pronounced at coarser grains and during the breeding season. Binary representation of water cover is useful at finer grain sizes (i.e., 1km^2).

Conclusions
At more detailed grains (i.e., 1km^2), the simple presence or absence of a certain land-cover type can be a realistic descriptor of species occurrence. This is particularly advantageous when collecting habitat data in the field as simply recording the presence of a habitat is significantly less time-consuming than recording its total area. For models using coarser grains, we recommend using proportional land-cover variables.

Keywords: eBird,land-cover, scale, SDM, variable selection

3.2.2. Introduction

The relationship between species occurrence and their environment is fundamental to ecology (Cadotte et al. 2011, Wisz et al. 2013, Schmeller et al. 2018, Young et al. 2023) and its importance is growing in the face of the ongoing global change (Butchart et al. 2010, Barnosky et al. 2011, Carlson et al. 2022). Understanding species' responses to the environment is intrinsically linked to the possibility of predicting species distribution patterns and, hence, is useful for their conservation and management. Species distribution models (SDMs) are widely used for such purposes (e.g. Václavík et al. 2012, Cohen et al. 2016, Ellis-Soto et al. 2021, Lindegren et al. 2022, Mohammadi et al. 2022, Cogliati et al. 2023).

Even though SDMs are now commonly adopted, ecologists still face challenges. These are in particular related to the quality of the input data, which can significantly impact the fitted models (Araújo et al. 2019, Gábor et al. 2020, Bazzichetto et al. 2023, Smith et al. 2023; Wang and Jackson 2023). Such challenges include, among other issues, the selection of the appropriate scale/grain (Miguet et al. 2016, Wunderlich, et al. 2022, Zarzo-Arias et al. 2022) and environmental variables (Williams et al. 2012, Moudrý et al. 2019, Smith and Santos 2020).

The selection of environmental variables is a critical step in SDMs, nowadays compounded by the increasing availability of environmental data (e.g., Cord et al. 2013, 2014, Šímová et al. 2019, Howard et al. 2020; Moudrý et al. 2023a). This is especially true for land-cover data, the availability and quality of which increases with the number of remote sensing data (e.g. Prošek et al. 2020; Karra et al. 2021; Zhang et al. 2021; Hopkins et al. 2022). Land cover types are among the most common predictors that enter the SDMs, especially in the context of land-cover changes (Coppée et al. 2022, Peng et al. 2022). Land cover type, which typically describes the habitat availability within a spatial unit, is commonly included as a continuous variable – for example, as the area or proportion of a specific land-cover type within the study area (e.g., Moudrý and Šímová 2013, Rose et al. 2020, Koma et al. 2022). The underlying rationale is based on an assumption that the probability of occurrence of a species increases continuously with the increase in land cover (or habitat) area within a given spatial unit. This could be attributed to the fact that larger habitat areas sustain larger populations due to their higher carrying capacity and shelter and food availability, so that populations are less susceptible to stochastic extinctions, competition, predation and inbreeding depression (Hanski 1999, Lande et al. 2003). In addition, larger habitat areas are bigger targets for colonizing individuals from the surrounding habitat matrix (Buckley and Knedlhans 1986), thereby increasing the probability of rescue effects following extinction events (Brown and Kodric-Brown 1977). However, in some cases, the mere presence or absence of a habitat may be more important than the total habitat area. In a recent study, Gábor et al. (2022a) demonstrated that for species specializing in rare habitats, such as water bodies in Central Europe, the amount of habitat within a spatial unit is sometimes less important than its simple presence or absence. This is related to the concept of critical habitat area (Andrén 1994, Fahrig 2001, Melo et al. 2018), which hypothesizes that there is a threshold in habitat amount below which a species cannot survive; for example, loons (order Gaviiformes) are physiologically constrained from foraging on land and require water habitat for survival (see Figure 3.5). While there are limited examples of such an approach in SDMs (e.g., Pearson et al. 2004, Romero et al. 2013, Zhang and Vincent 2018), in a recent study, Gábor et al. (2022a) demonstrated that for species specializing in rare habitats, such as water bodies in Central Europe, the amount of habitat within a spatial unit is sometimes less important than its simple presence or absence. This could be particularly advantageous when collecting

habitat data in the field as simply recording the presence of a habitat is significantly less time-consuming than recording its total area.

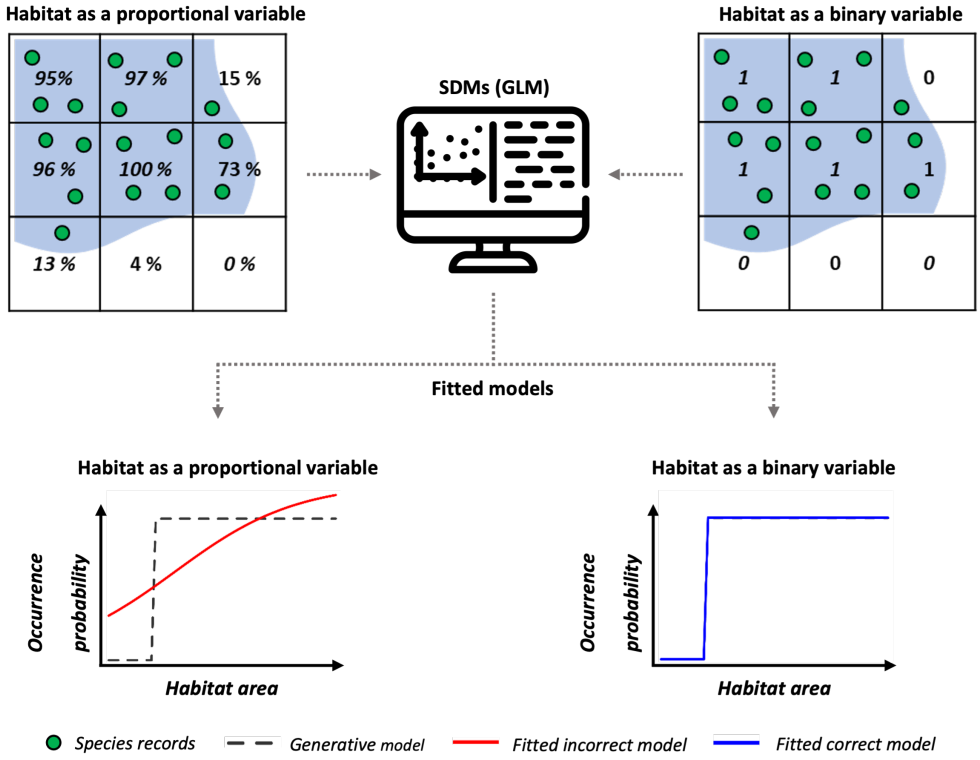


Figure 3.5: Graphical rationale for the hypothesis on binary variables. We modelled species occurrence probability as a step function (generative model, black dashed line) to generate 100 presences/absences of a species that needs at least 20% of land-cover within a spatial unit, drawn from a Bernoulli distribution with species occurrence probability as a parameter. Then, we fitted species distribution models (GLM) with either proportional (incorrect model, red line) and binary coverage (correct model, blue line), as a variable.

It is well-recognized that SDMs are grain-dependent (Elith and Leathwick 2009) and empirical evidence has shown that species exhibit stronger responses to their environment at certain grains than others (see reviews by Miguet et al. 2016 and Moudrý et al. 2023b). Therefore, the choice of the grain constitutes an important part of the modeling process as it can affect the ability to detect the response (Mertes and Jetz 2018, Luebert et al. 2022, Wunderlich, et al. 2022, Lu and Jetz 2023). In addition, the choice of the grain is often determined by data availability rather than study goals. This results in large variability of grains adopted in existing studies, from a few meters (e.g., Bazzichetto et al. 2018, Lecours et al. 2020, Casanelles-Abella et al. 2022, Stark and Fridley 2022) to many kilometers (e.g., Kleisner et al. 2017, Norberg et al. 2019, Zarzo-Arias et al. 2022). The grain size is also essential when selecting environmental variables (Pearson and Dawson 2003; Moudrý et al. 2023b), and it is likely to be critical when deciding whether to use land-cover as a binary or continuous variable when modelling species occurrence. However, this has never been thoroughly tested.

We hypothesize that at finer grains, binary variables are sufficient to fit a useful model, as a single habitat type is more likely to be dominant or completely absent. However, at coarse grains, the presence of habitat alone cannot be enough to sustain a viable population that requires a certain percentage of habitat in a given area, and the variables representing the habitat proportion may play a greater role. Further, habitat specialists may only require the presence of a given habitat type during winter or migration but, require that habitat to be the dominant type on the landscape during breeding season (Zuckeberg et al. 2016), emphasizing the importance of understanding seasonal moderators on the selection of environmental variables.

Therefore, in this study, we examined the interaction between the grain size and the representation of a land-cover variable (binary or continuous) when modeling species distribution. We used citizen science data downloaded from eBird to fit species distribution models for 57 water bird species in North America. The eBird dataset consists of extensive and often spatially detailed occurrence data (an average number of records per species in this study is 880,270 records), making it ideal to model species distributions down to fine spatial grains and across large spatial extents. For each species, we fit models with water cover represented as a proportional or binary variable and using multiple grains (1 x 1 km, 5 x 5 km, 10 x 10 km, and 50 x 50 km). In addition, we fit models for both breeding and non-breeding seasons because many of these species occupy highly distinct ranges between these seasons, with specialization to landcover type often greatest during breeding season (Zuckerberg et al. 2016). Specifically, we address the following questions: (1) Do models built with binary land-cover variables perform equally well or even better than those built with proportional variables? (2) Is the model performance affected by the adopted gain? (3) Does breeding and non-breeding season affect model performance? (4) Do grain and season affect the usability of the binary water cover variable?

3.2.3. Material and Methods

Modeling region and species selection

We modeled species distributions across the North American continent, in a box between 179.99°W, 42.68°W, 10.53°N, and 87.11°N that included Canada, the United States, and the majority of Central America. Although we only modeled species native to the United States and Canada and only present the biodiversity estimates for this region, modeling each species' entire continental range was required to ensure accurate predictions.

Based on the American Birding Association birding codes (2008) updated to the Clements bird taxonomy as of 2021, we compiled a list of 197 water-associated native bird species that breed or overwinter in the United States and Canada each year (Clements, 2007). These codes are widely used to distinguish regularly occurring species (code 1 or 2 species, which we use) from vagrants occurring irregularly (code 3+). We excluded species with almost entirely marine ranges or with insufficient data points and those for which records are only available for part of the season (i.e., breeding, non-breeding). We also avoided modeling Hawaiian endemics because these species were restricted to islands that were smaller than some of our spatial grain sizes.

Species data acquisition and filtering

We gathered data from eBird, a global citizen science initiative in which users submit checklists containing bird observations, widely used for understanding species distributions (Sul-

livan et al., 2014). eBird provides vast amounts of avian biodiversity data and the number of records used in this study is literally unique, ranging from tens of thousands to units of millions, with an average of 880,270 records per species (see Table A1 in Supplementary material). Users can indicate whether or not all observed species were recorded in the checklist ("complete" checklists), allowing for absence inference and presence-absence modeling. Users also indicate the level of effort involved in each observation by providing the distance traveled, time spent birding, and the number of observers (hereafter, effort indicators).

We compiled all eBird data for all species separately during the breeding season (June-August; June-July for shorebirds or Charadriiformes, which migrate early) and non-breeding season (December-February). We removed subspecies information from all checklists and summarized all data at the species level. Following established eBird data modeling protocols, we initially applied several filters to the data to reduce bias and improve data quality (Johnston et al., 2019; Kelling et al., 2018). First, we eliminated checklists with extremely long durations (> 3 hours), large numbers of observers (>5), or protocols other than "stationary" or "traveling," as these are incomparable with the majority of eBird's data. Second, to reduce observational positional error, we eliminated checklists covering more than 1 km because they are likely to result in greater spatial uncertainty (Gábor et al. 2022b, 2023). Third, data prior to 2004 (< 0.1% of points) were removed because there is insufficient data from earlier years to adequately control for long-term temporal trends in regional bird abundances, as recommended by eBird (Fink et al. 2010, 2020). Finally, data from users with fewer than five contributions were removed to reduce bias (e.g., false absence) caused by inexperienced birders.

Before modeling, the data was further filtered at the species level. Checklists were limited to those falling within a 200 km buffer of the species' seasonal expert range boundary to limit overprediction outside of the species' range extent (via Cornell's spatial boundaries - https://ebird.org/science/status-and-trends/download-data, accessed July 2021). When modeling shorebirds (order Charadriiformes), we excluded August data because many species are already migrating long distances by this time. We limited checklists to one per 5 km grid cell per week to reduce site selection and temporal bias in data collection. To reduce the imbalance between presence and absence points, we utilized "case-control sampling" repeating this filtering for populations of checklists where the species is present and absent as recommended by eBird's best practices (Fink et al. 2019). We ended up modeling 57 species (see Table A1 in Supplementary material). R 4.1.0 was used to complete all data compilation, analyses, and visualizations (R Core Team, 2021). Coordinates, polygons, and grids used in the study operated under a conical equal area projection. For spatial geoprocessing, the raster (Hijmans et al., 2015), rgdal (Bivand et al., 2015), and sf (Pebesma, 2018) packages were used.

Environmental data

In total, we considered 12 land-cover variables, 4 topographic/habitat variables, and 4 climatic variables (Table 1). We obtained land-cover data from the European Space Agency (Climate Change Initiative; "ESA. Land Cover CCI Product User Guide Version 2. Technical Report.," 2017) at a 300m resolution. This product provides proportional land-cover variables (i.e. they provide the proportion of the area covered by a particular land-cover type) and includes the following land-cover types: mixed forest, mosaic, shrubland, grassland, lichens/-mosses, sparse, flooded/freshwater, flooded/saltwater, flooded/shrub, urban, barren, and ice. We coarsened land-cover variables to a grain of 1 km by mean-aggregating the percentages

of individual land-cover types.

Table 3.1: Environmental covariates included in species distribution models.

Predictor	Definition	Source
Bio1	Annual mean temperature	CHELSA v2.1
Bio12	Annual precipitation	CHELSA v2.1
Bio15	Precipitation seasonality	CHELSA v2.1
Cloudsd	Intra-annual variation in cloud cover	EarthEnv
Evisum	Mean enhanced vegetation index	hydroSHEDS
TWI	Topographic wetness index	hydroSHEDS
TRI	Terrain roughness index	EarthEnv
elev	Elevation	EarthEnv
Mixed_forest	Percent land cover	ESA CCI
Mosaic_herbacious	Percent land cover	ESA CCI
Shrubland	Percent land cover	ESA CCI
Grassland	Percent land cover	ESA CCI
Lichens/Mosses	Percent land cover	ESA CCI
Sparse	Percent land cover	ESA CCI
Flooded (freshwater)	Percent land cover	ESA CCI
Flooded (saltwater)	Percent land cover	ESA CCI
Flooded (shrub)	Percent land cover	ESA CCI
Urban	Percent land cover	ESA CCI
Barren	Percent land cover	ESA CCI
Ice	Percent land cover	ESA CCI

To examine how water cover (flooded/freshwater category) influences model performance, we summarized this variable in five ways: as a proportional representation and as a binary variable classified using thresholds of 1%, 10%, 20%, or 50%. For example, when using the 1% threshold, any cell with water cover of 1% or more is considered to contain water in the binary representation and any cell with less than 1% is not.

While our primary focus in this study was on the influence of water cover on model performance and output, each SDM included several classes of covariates to account for numerous other factors that influence species distributions. Our topographic/habitat suite of variables included the mean elevation (from EarthEnv; Robinson et al., 2014), mean enhanced vegetation index (EVI; MODIS; https://lpdaac.usgs.gov/products/mod11a1v006/), topographic wetness index (TWI; hydroSHEDS; Marthews et al., 2015), and terrain roughness index (TRI; EarthEnv). Climatic covariates included the mean annual temperature (bio1), mean annual precipitation (bio12), precipitation seasonality (bio15; all bio variables from CHELSA v2.1; Karger et al., 2021), and intra-annual variation in cloud cover (EarthEnv). We selected environmental variables to reduce collinearity, although collinearity is less of a concern within random forest (Farrell et la. 2019). All variables (topographic, climatic, and land-cover) that were available at finer resolution were spatially aggregated to 1 km using the mean value of all cells falling within the 1x1km cell. We also used temporal covariates in our models, such as year, date, and time of day, to account for temporal variability in bird activity and long-term population trends. Finally, as covariates, we included all effort indicators (distance travelled, duration of checklist, and number of observers).

Coarsening the spatial grain

To test the role of representing the water cover variable as binary or proportional in different spatial grains, we coarsened the spatial grain of our analyses by aggregating all land-cover, topographic, and climatic variables. We aggregated from 1 x 1 km to 5 x 5 km, 10 x 10 km, and 50 x 50 km (hereafter 'test grains') using the means of all 1 km cells within the larger cell. The test grains were chosen based on their frequent use in published SDM studies (see review by Moudrý et al. 2023b). The species observation data were subsequently associated to the each coarser resolution grids. This approach leads to the fact that multiple species records may occur in any cell of the coarser grid. Following Guisan et al. (2007), we decided to keep the sample size constant. If any coarse cell contained more than one species record, all these records were retained rather than reducing these to single record per cell. This allowed us to keep the sample size (i.e. presence: absence ratio) consistent between all grains and avoid mistaking the effects of change in resolution for those caused by the change in sample size and prevalence (Leroy et al. 2018). Temporal and effort covariates were not scaled because they were assigned at the checklist (point) level.

Model fitting and evaluation

We used Random Forest to model the breeding and non-breeding distributions of each species separately. Random Forest is a machine learning method designed to analyse large datasets with many covariates and is frequently found to produce the most accurate SDMs (e.g., Mi et al., 2017, Valavi et al. 2023). Random forests are adaptable, automatically adjusting to complex, nonlinear relationships, and consider high-order interactions between environmental variables (Evans et al., 2011). Random Forest models were conducted in R version 4.1.0 using the ranger package (Wright et al., 2018).

We randomly divided the data into training and testing. The testing set was used for model validation. Models were parameterized to 100 trees and 7 threads. We compiled metrics of predictive model performance including the area under the ROC curve (AUC), true skill statistic (TSS), sensitivity, and specificity. We masked portions of the predictions that fell outside of the buffered range extent when estimating the predictive performance. To diagnose overfitting, we examined test-set calibration plots for each model. To compare the effect of binary and proportional water cover variables, we assessed their relative importance. Overall, we fit models for each of the 57 analyzed species during each season (breeding vs. non-breeding) at four test grains (1 x 1 km to 5 x 5 km, 10 x 10 km, and 50 x 50 km) and five types of water cover representation (proportional, and binary using four thresholds), resulting in 2,280 separate models. The computational time for our models was approximately a month using high-performance computing tools, which vividly demonstrates the volume of scenarios we ran to test our hypothesis.

3.2.4. Results

All performance metrics followed similar patterns, hence we focused on TSS for simplicity. In general, SDMs fitted at the finest grain (1 x 1 km) performed very well (breeding TSS = 0.8; non-breeding TSS = 0.79; Figure 3.6). The decrease in model performance was observed between original models at 1 x 1 km grain and models at 5 x 5 km grain. Further coarsening of grain resulted in minimal additional decrease in models performance (Figure 3.6). The comparison of model performance (TSS) between models fitted with water cover as a proportional variable and as a binary variable revealed minimal differences in model performance at all grains (Figure 3.7). At 1 x 1 km grain, the performance decreased with the

increasing threshold used to derive the binary water cover variable. At coarser grains, the threshold had no effect on model performance. The results showed that the choice of water cover representation mattered most in species with fewer observations, which are generally less common (Figure 3.7B). The drop in model performance was more pronounced for non-breeding season models (Figure 3.7B), while TSS dropped on average by 0.12 from 0.82 (1 x 1 km) to 0.7 (50 x 50 kilometres) for breeding season models; the drop for non-breeding season models was about 0.2. In the models fitted using binary water cover representation, the importance of the water cover variable significantly decreased compared to the models using its proportional representation (Figure 3.8). The importance of the water cover variable decreased with coarsening the grain and increasing the threshold. For example, when models were fitted using a 1x1 km grain and a 1% threshold, the mean drop in water cover importance was approx. 20% compared to the proportional representation, whereas for models fitted using the 50 x 50 km grain and 50% threshold, the drop was over 80% (Figure 3.8).

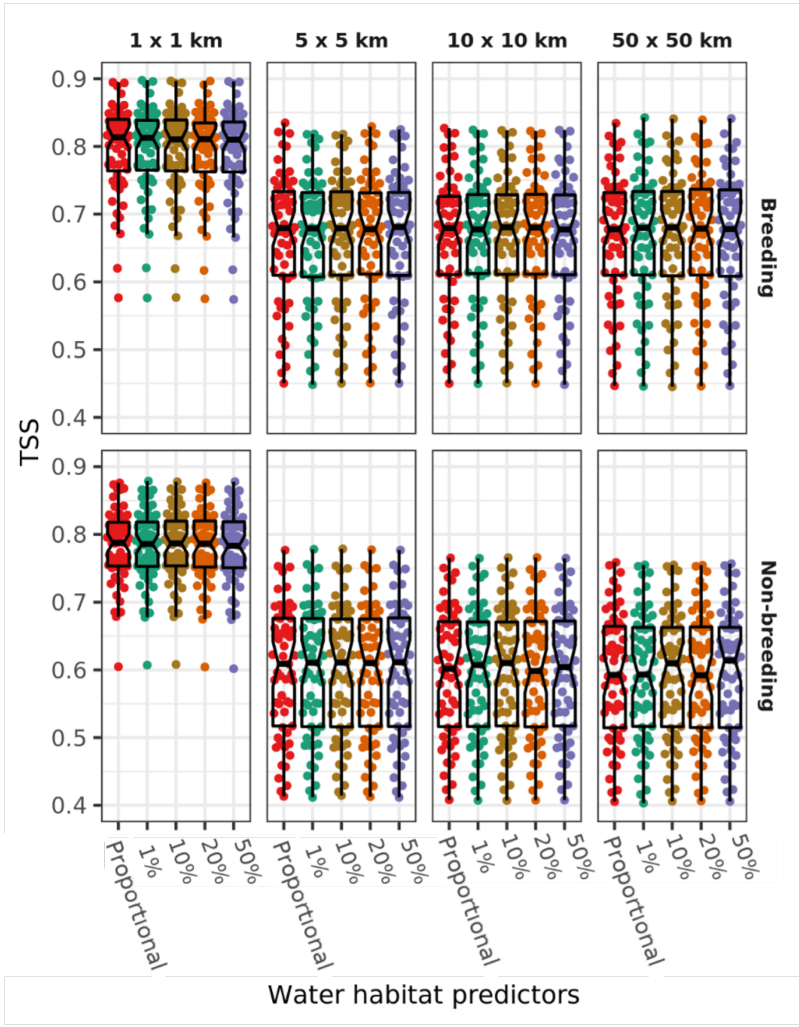


Figure 3.6: Predictive performance represented by TSS. Columns show results for different grain sizes, sorted by the thresholds (%) used to generate a binary land-cover variable. Rows show results for different seasons. Boxplot central lines represent median for each scenario, points represent individual species.

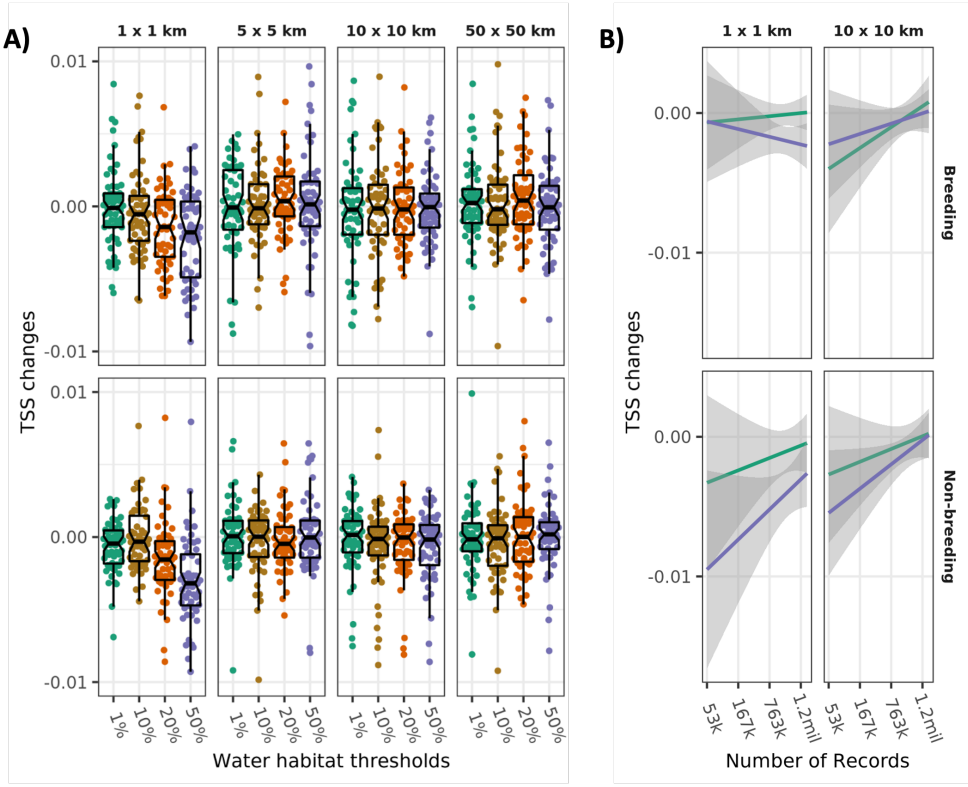


Figure 3.7: Variation of predictive performance across summarization method, sample size, and spatial grain. A: difference in TSS between models fitted with binary land-cover variable compared to a reference model using the proportional land-cover. Boxplot central lines represent the median for each scenario, points represent individual species. B: variation in TSS across different sample sizes (green and purple colour represent 1% and 50% thresholds, respectively). Columns show results for different grain sizes. Rows show results for different seasons. Positive values indicate that models with binary habitat predictors performed better than those with proportional predictors and vice versa.

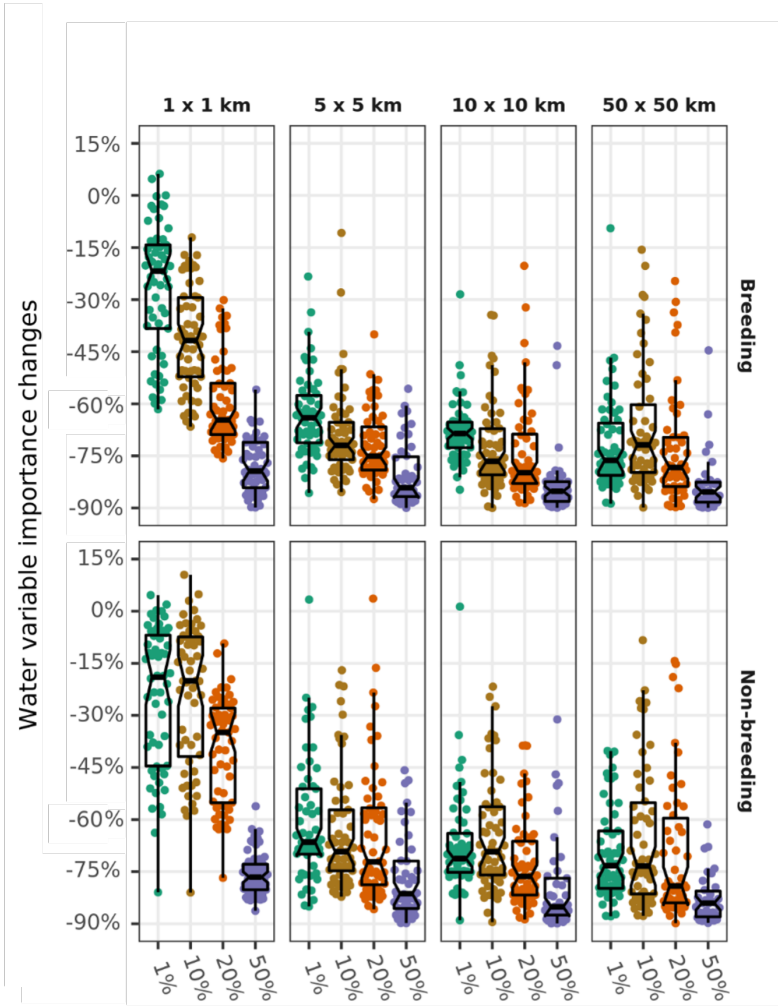


Figure 3.8: Differences in the importance of the water cover variable in models fitted binary land-cover (with four thresholds) compared to the importance resulting from the corresponding model using the proportional representation of this variable. Boxplot central lines represent the median for each scenario, points represent individual species. Negative values indicate that models with binary water cover predictors had lower water variable importance than models with proportional predictors and vice versa.

3.2.5. Discussion

To fully utilize the potential of the fact that the availability of data suitable for SDM keeps increasing, we must understand how to process and use these data effectively. This includes determining the optimal grain size and appropriate selection of environmental variables. In this study, we evaluated the usefulness of binary land-cover variables for SDMs at multiple grains. We computed species distribution models for 57 water bird species in North America across several grain sizes (i.e., from 1 km^2 up to 2500 km^2) for breeding and non-breeding seasons. We used the proportional and binary representations of land-cover variables derived with various thresholds (1%, 10%, 20%, 50%).

Our results show that the use of binary land-cover variables should be approached with caution. Models created with binary variables performed equally well as those with proportional variables (Figure 3.6), but there was a significant decrease in the importance of the water cover variable (Figure 3.8). The use of binary water cover variable often led to the situation when this variable was not identified as the most important land-cover variable for water bird species which is, to say the least, unexpected. In addition, we observed that with coarser grain and greater thresholds used to derive the binary water cover variable, the water cover became less important, with a drop of up to 90% at the coarsest grain used compared to the model using the proportional water cover variable (Figure 3.8). This contrasts with the study by Gábor et al. (2022a) who showed that for water bird species, models at approximately 10 x 10 km grain using binary predictors performed better than models with proportional predictors. The contrasting results may be attributed to the fact that in our study, the modeling algorithms were allowed to choose the best combination of environmental variables, i.e., the lower predictive power of the binary form of the water cover variable is compensated for by another variable. For example, flooded tree or shrubland habitats often occur near water cover and may become predictive of a species occurrence. In contrast, Gábor et al. (2022a) used only water cover variables for model fitting and, hence, their models did not have the option to select variables that would provide a closer fit.

Our results show that as the grain size becomes coarser, the model performance decreases, which is consistent with prior studies (Guisan et al. 2007, Seo et al. 2009, Chauvier et al. 2022, Zarzo-Arias et al. 2022) and supports the common recommendation to first try the most detailed grain that the data allows (Moudrý et al. 2023b). However, it's important to note that a finer grain may not always be better, as organisms respond to their environment more strongly at some grains than at others. These grains have been referred to as 'response grains' or 'ecological scales' and depend on the species ecology (e.g., its home-range; Mertes and Jetz 2018). Consequently, the choice of grain in models can strongly influence our ability to detect and measure species' response to the environment. The grain at which a species is expected to respond to the environment should be always considered (Manzoor et al. 2020, Wunderlich, et al. 2022, Moudrý et al. 2023b).

Importantly, our results suggest that the grain size can impact the applicability of binary land-cover variables. The large decline in the importance of the binary water cover variable relative to its proportional representation observed at 25km^2 resolution and coarser suggests that at grains coarser than 1km^2, the binary representation of water cover is not applicable. Therefore, the proportion or total perimeter of water bodies (e.g., Virkkala et al. 2005, Moudrý and Šímová 2013) should be preferred at coarser grains. However, at a 1km^2 grain, the decrease in the importance of the water cover variable was relatively small and models using binary predictors (presence/absence of the habitat) might be useful.

The threshold for deriving the binary water cover variable is another important aspect to consider. For example, a 1% threshold used in this study requires a water area equivalent to 2 American football fields at a 1 x 1 km resolution, 185at a 10 x 10 km resolution, and 4,638 at a 50 x 50 km resolution. For comparison, the area of the lake of the Great Salt Lake (Utah, USA) is 456 football fields. Thus, it is apparent that especially at coarser grains, the use of an inappropriate threshold can lead to a substantial drop in the number of cells identified as containing water cover, even though many of them provide enough carrying capacity for smaller water bird species populations (Hanski 1999, Melo et al. 2018). Indeed, our results show that using higher than 1% threshold leads to a decline in the importance of the binary water cover variable compared to its proportional representation, making it unusable.

Despite our expectations, we did not find any significant difference in model performance or in importance of the water cover variable between breeding and non-breeding seasons. We expected that models would perform better during the breeding season as water is more critical to the biology of many birds during this season. This expectation was based on the fact that birds have offspring that often cannot leave the water or fly, and many species remain stationary at a specific nesting site.

3.2.6. Conclusions

The appropriate selection of environmental variables and grain size in species distribution modeling is of considerable importance. In this study, we demonstrated that (1) the performance of the models was not significantly affected by the type of the adopted water cover variable (proportional or binary) but a significant decrease was observed in the importance of the water cover variable when used in a binary form, (2) the performance of the models was affected by the adopted grain, with models at a 1 km^2 grain performing considerably better than models at coarser grains, (3) models for the breeding and non-breading season performed equally well, and (4) the importance of the binary water cover variable declined relative to its proportional representation with coarser grains and during the breeding season. We highlight that the use of binary land-cover variables should be approached with caution. Binary representation of water cover might be useful at finer grain sizes (i.e., 1km^2), which is a promising result as at more detailed grains, the simple presence or absence of a certain land-cover type can be a realistic descriptor and can save time in fieldwork. However, binary representation of water cover is unsuitable for models at grains coarser than 1km^2. For such grains, the use of proportional land-cover variables produces more reliable models.

3.3. Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable

Lukáš Gábor, Walter Jetz, Alejandra Zarzo-Arias, Kevin Winner, Scott Yanco, Stefan Pinkert, Charles J. Marsh, Matthew S. Rogan, Jussi Mäkinen, Duccio Rocchini, Vojtěch Barták, Marco Malavasi, Petr Balej and Vítězslav Moudrý

Adapted from Ecography, 2023, e06358.

Publication metrics:
Quartile (2023): Ecology, Evolution, Behavior and Systematics (Q1)
Impact Factor (2023): 6.08
SCImago Journal Rank (2023): 2.37

The first author contributed to the study as follows: study conception and design (lead), data curation (lead), analysis and interpretation of results (lead), visualization (lead), writing – original draft (lead), writing – review and editing (equal), overall study supervision (equal), funding acquisition (lead).

The link to the published article and supplementary materials can be found here:
https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.06358

3.3.1. Abstract

Species distribution models (SDMs) have become a common tool in studies of species-environment relationships but can be negatively affected by positional uncertainty of underlying species occurrence data. Previous work has documented the effect of positional uncertainty on model predictive performance, but its consequences for inference about species-environment relationships remain largely unknown. Here we use over 12,000 combinations of virtual and real environmental variables and virtual species, as well as a real case-study, to investigate how accurately SDMs can recover species-environment relationships after applying known positional errors to species occurrence data. We explored a range of environmental predictors with various spatial heterogeneity, species' niche widths, sample sizes and magnitudes of positional error. Positional uncertainty decreased predictive model performance for all modeled scenarios. The absolute and relative importance of environmental predictors and the shape of species-environmental relationships co-varied with a level of positional uncertainty. These differences were much weaker than those observed for overall model performance especially for homogenous predictor variables. This suggests that, at least for the example species and conditions analyzed, the negative consequences of positional uncertainty on model performance did not extend as strongly to the ecological interpretability of the models. Although the findings are encouraging for practitioners using SDMs to reveal generative mechanisms based on spatially uncertain data, they suggest greater consequences for applications utilizing distributions predicted from SDMs using positionally uncertain data, such as conservation prioritization and biodiversity monitoring.

Keywords: Birds, Ecological modeling, Location error, Niche models, Species-environment relationship

3.3.2. Introduction

Species occurrences are increasingly being recorded in online, public, global databases such as GBIF (www.gbif.org), eBird (www.ebird.org), or iNaturalist (www.inatura-list.org), where scientists and the general public worldwide share field observations. However, whereas the number of records in these databases is constantly growing, many observations are characterized by substantial uncertainty in the occurrence location (Moudrý and Devillers 2020). Such uncertainty poses problems for analyses aimed at revealing species-environment relationships because the environmental conditions at recorded sites could differ from those at true locations.

SDMs are a widely used class of ecological models that use occurrence data to estimate species-environment relationships (Ferrier et al. 2017) and allow researchers to predict the relative probability of occurrence across unsampled areas of a study region. SDMs have broad utility in ecology (Elith and Leathwick 2009, Franklin 2010, Guisan et al. 2013, Zurell et al. 2019) and have been successfully used to identify critical habitats (Volis et al. 2021), delineate suitable locations for relocations (Segal et al. 2021), or assess the potential impacts of climate change (Santini et al. 2021). SDMs are also frequently used to infer the importance of environmental variables defining the species niche (e.g., Moudrý and Šímová 2013, Bradie and Leung 2017, Lecours et al. 2020, Li and Kou 2021, Smith and Santos 2020) and to determine the shapes of species responses to the environment (Austin et al. 2006, Hargreaves et al. 2014, Lee-Yaw et al. 2016, Dvorský et al. 2017, Bazzichetto et al. 2018). However, despite methodological advances improving the performance of SDMs over the last two decades (e.g., Phillips et al. 2006, Varela et al. 2014, Graham et al. 2019, Tessarolo et al. 2021), they remain sensitive to the spatial accuracy of occurrence data used in model fitting (Visscher 2006, Moudrý and Šímová et al. 2012, Moudrý et al. 2017, Araújo et al. 2019, Byaraktarov et al. 2020, Isaac et al. 2020, Etherington et al. 2021, Gábor et al. 2022b).

Maximum Entropy-based SDMs estimate a response curve in environmental space which discriminates between observed occurrences and "background" samples that do not contain occurrence information (Figure 3.9). Positional uncertainty describes the magnitude of error in the locations of occurrence records. In some cases, it quantifies the likelihood of a mismatch between the true environmental variables' values and the assigned value. Even if a spatial error does not lead to directional bias in environmental space, increased sampling error can decrease predictive model performance and even bias the slope of the response curves; or the estimations of variable importance (Figure 3.9; e.g., Johnson and Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009, Hefley et al. 2014, Fernandes et al. 2019). The magnitude of positional error in environmental variables measurements may be amplified in highly heterogeneous or structured landscapes where spatial autocorrelation in environmental variables is relatively low (Naimi et al. 2011, Naimi et al. 2014). Moreover, even uniform spatial error can create persistent bias in measurements of environmental variables depending on the spatial structure of the relevant variable. For example, uniformly-distributed spatial error for occurrences of a mountaintop-dwelling species would always lead to estimates of elevation that are biased to lower elevations than reality. Such a bias in even one environmental variable could reduce overall model predictive performance and bias the estimated response curve and variable importance (Figure 3.9).

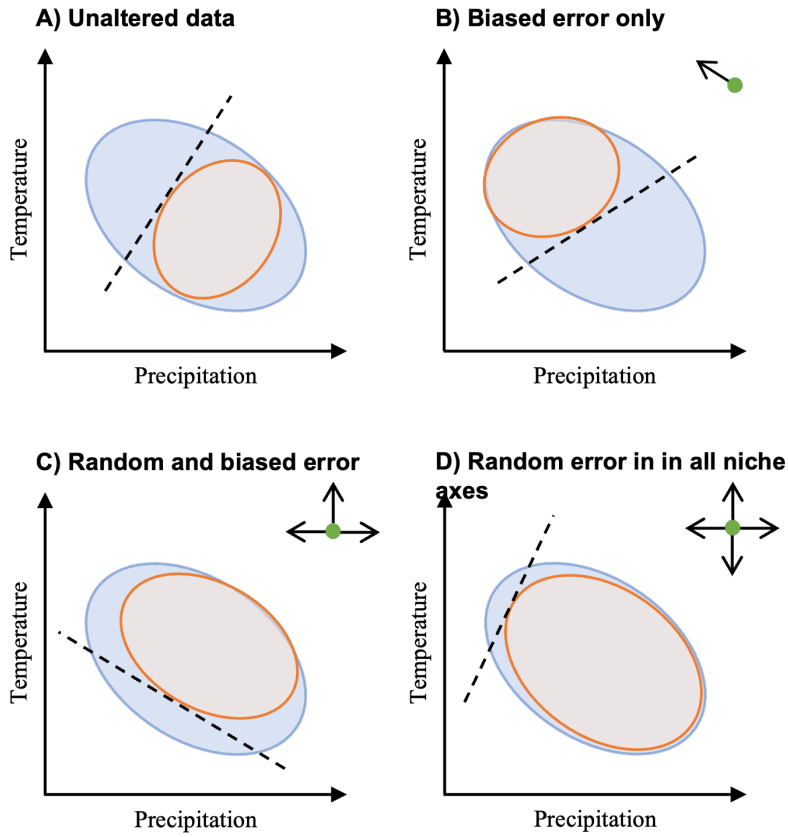


Figure 3.9: Heuristic illustration of the effect of positional error (depicted in environmental space) on response curve estimation. Panel (A) shows the area of species occurrences (orange) relative to area of the sampled background points (blue) without positional uncertainty. The 'true' response curve that would be estimated from these data would result in an approximate discrimination threshold which is represented by the dashed line. The response curve would differ from the 'true' response curve when spatial error leads to (B) persistent bias in environmental space, (C) persistent bias and unbiased sampling error in environmental space, and even (D) random sampling error in environmental space (i.e. without directional bias). Arrows in the upper-right of each panel indicate the directions of shift for the presences (green point).

Even more troubling, common strategies for mitigating the effects of positional uncertainty on SDMs have recently been shown to be ineffective. For example, Gábor et al. (2020) demonstrated that increased sample sizes do not reduce the negative effects of positional uncertainty. Similarly, Smith et al. (2022) showed that discarding data with high positional uncertainty limits our ability to determine species' distribution and climatic niche tolerances properly. Particularly, they demonstrated that using only accurate data dramatically reduces range size estimates and overestimates exposure to climate change. Recently, Gábor et al. (2022) concluded that coarsening the analysis grain to compensate for positional error did not improve model performance and recommended to develop models with the finest possible analysis grain and as close to the response grain as possible even when available species occurrences suffer from positional errors.

Although previous studies confirmed the effect of positional error on the model predictive performance (e.g., Graham et al. 2008, Johnson and Gillingham 2008, Fernandez et al. 2009, Osborne and Leitão 2009, Naimi et al. 2011, 2014, Hefley et al. 2014, Tulowiecki et al. 2015, Gueta and Carmel 2016, Mitchell et al. 2017, Soultan and Safi 2017, Fernandes et al. 2019, Gábor et al. 2022b), especially when specialist species are modeled (Visscher 2006, Gábor et al. 2020a), the question about how positional uncertainty in species occurrences affects models' parameter estimation (species-environment relationships inference) remain largely unexplored. Therefore, in this study, we explored the extent to which parameter estimation is affected by positional uncertainty. Specifically, we investigated the influence of positional error on variable importance and the shape of the response curves. We hypothesized that increasing positional uncertainty would lead to decreased model predictive performance and imprecise variable importance and response curves with more pronounced effects for species with narrow niche and heterogeneous variables.

3.3.3. Methods

We used a virtual species approach across two workflows (Figure 3.10), which allowed us to know the true underlying occurrence location and thus enabled us to characterize relative bias in parameter estimates (Zurell et al. 2010, Moudrý 2015, Meynard et al. 2019), as well as specify various spatial autocorrelation levels (SAC; Naimi et al. 2011, 2014) in the environmental variables.

We have simulated 12 560 combinations of virtual and real environmental data and virtual species to investigate our assumptions and have fitted over 628 000 models. Simulations were divided into two workflows and a third workflow investigated a real species. In Workflow 1, we combined virtual variables with different levels of spatial autocorrelation (SAC) and virtual species with varying widths of niches and number of occurrences. For these scenarios, models were fitted with only one variable (see Figure 3.10). Thanks to this, we got a simplified yet detailed insight into how various levels of positional uncertainty affects model's ability to properly detect species response to the environment across various SAC, niche width, and sample size.

Additionally, to mimic real SDMs situation, we have combined real environmental variables with virtual species with different niche widths and sample sizes and fitted models with multiple environmental variables (Workflow 2; Figure 3.10). This allowed us to explore our assumptions with more model complexity. Moreover, using numerous environmental variables to fit the models, we tested how positional uncertainty affects models' ability to properly detect the most influential variables (i.e., those used to generate virtual species).

Finally, we tested our assumptions using real environmental variables and real species (Band-tailed pigeon; Workflow 3; Figure 3.10). Our simulations showed that the model parameter estimation is negatively affected across various species niches (note, however, that the magnitude varies). Therefore, we hypothesized, considering the number of occurrences (n = 111) and the fact that the species is widely spread across the western part of the USA that positional uncertainty will bias model response curves and variable importance.

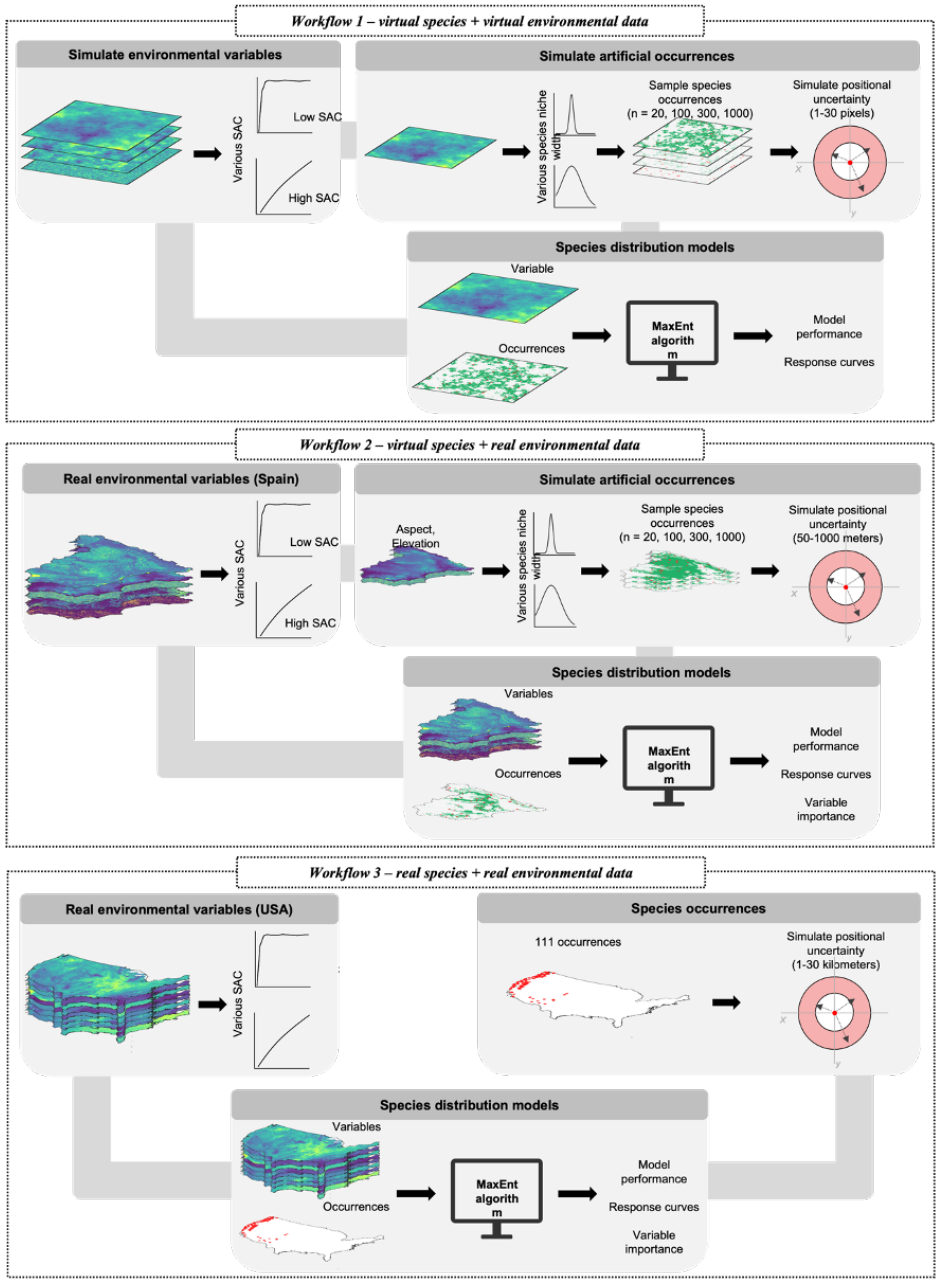


Figure 3.10: General modeling process for all three workflows. Each experiment was repeated 50 times

In Workflow 1, the artificial study area was given by the extent of the virtual landscape (200 x 200 cells; see details below). Virtual species in Workflow 2 used Spain (except islands) as a study area, whereas the band-tailed pigeon was modeled for the United States of America (Figure 3.10; Workflow 3).

Occurrence data

Workflow 1

We generated artificial occurrences using the '*virtualspecies*' package (Leroy et al. 2016, ver. 1.5) in the statistical software R (ver. 4.1.0, R Development Core Team 2021) with three steps: (i) define (virtual) species-environment relationships, (ii) project range into geographic space, and (iii) sample occurrence data from simulated range. We used a normal distribution to define the response of virtual species to the virtual environmental variable. To simulate species with different niche widths, we used the same mean (0.005) and varied standard deviation from 0.005 up to 0.2 using a logarithmically spaced sequence. In total, we generated 25 species with various niche widths. We then projected habitat suitability across our study area to define the probability distribution of occurrences. In the final step, we used a probabilistic simulation approach and logistic function with $\alpha = -0.05$ (controls the slope of the logistic curve) and $\beta = 0.3$ (the point of inflection of the logistic curve, i.e., the value of the environmental gradient at which the probability of occurrence is 50%), as recommended in prior studies, to convert the habitat suitability raster to a randomized binary presence-absence raster (see Meynard and Kaplan 2012, 2013, Meynard et al. 2019 for more details). This allowed us to generate virtual species with gradual responses to the environment that mimic the real species, as demonstrated by Meynard and Kaplan (2012, 2013). Subsequently, we sampled 20, 100, 300, and 1000 species occurrences.

Workflow 2

To generate virtual species occurrences for workflow 2, we used two environmental variables with various SAC (elevation - high SAC, aspect - low SAC). To simulate species with different niche widths, we used normal distributions with mean of 1000m and standard deviation of 100-500m for elevation, and a mean of 100° and standard deviation of 10-100° for aspect. This allowed us to generate three species with various niche widths (narrow, medium, wide). We then projected habitat suitability across our study area to define the probability distribution of occurrences. We used a logistic function with $\alpha = -0.05$ and $\beta = 0.3$ to convert the habitat suitability raster to a randomized binary presence-absence raster. We sampled 20, 100, 300, and 1000 species occurrences.

Workflow 3

Occurrences for the band-tailed pigeon, a species with high detection probability (Keppie and Braun 2000), were extracted from the North American Breeding Bird Survey (BBS, Sauer et al., 2015), a long-term collection of over 4,800 survey routes distributed across North America. Each survey route consists of 50 points count locations distributed 0.8 kilometers apart and sampled for 3 minutes. We considered only routes from the study region (the USA, see Figure 3.10) and retained only those routes where we assumed there was high certainty that the species was present. First, we discarded data sampled before the year 2000 and then kept only those routes with at least ten years of samples post-2000. We considered occurrences from a minimum of 5 years of samples as presences. Routes where the species was detected but on fewer occasions, and therefore presence status was unsure, were removed from the

analysis. The final dataset contained 111 presences.

Environmental variables

Naimi et al. (2011, 2014) showed that spatial autocorrelation in environmental variables affects the degree to which positional uncertainty creates mismatches between true and measured environmental variables values. Therefore, we generated artificial environmental variables and selected real environmental variables that spanned different degrees of spatial autocorrelation.

Workflow 1

We generated artificial environmental variables using the R '*gstat*' package (ver. 2.0-9) and unconditional simulation over a regular grid of 200 x 200 cells. Unconditional simulation allows for a generation of environmental variables with different SAC, where the level of SAC is defined by a variogram (Dungan 1999, Naimi et al. 2011). We used an exponential variogram with the same sill parameter of 0.025 for all simulations. To simulate variables across different SAC levels, we scaled the range parameters from 1 (low SAC, high heterogeneity) to 49 (high SAC, low heterogeneity) by increments of 2 to total 25 virtual environmental variables. Only one variable was used to generate virtual species and subsequently model the species distribution (Figure 6.10).

Workflow 2

We chose five environmental variables to construct models for Workflow 2 (Supplementary materials Table A1). Two of the variables were related to habitat characteristics: grassland coverage and forest coverage (http://centrodedescargas.cnig.es/; Spain national geographic center); and three were related to topography: topography wetness index, aspect and elevation (http://centrodedescargas.cnig.es/; Spain national geographic center). We used elevation and aspect that serve as a proxy for temperature (see for example Müller and Brandl 2009, Coops et la. 2010, Vierling et al. 2011, Work et al. 2011, Vogeler et al. 2014), and topography wetness index that is a proxy for water availability (e.g., Petroselli et al. 2013, Reif et al. 2018, Title and Bemmels 2018). Topography wetness index was derived from the elevation model (SAGA-GIS v. 2.1.4; Conrad et al. 2015). All environmental variables were resampled from an original resolution of 25 x 25 m (elevation) or 20 x 20 m (all other variables) to 50 x 50 m cell resolution using the mean values of the original data (Moudrý et al. 2019) for modeling purposes (see Supplementary materials Table A1). Only elevation and aspect were used to generate virtual species, while all variables were used to fit models.

Workflow 3

For the band-tailed pigeon, we selected nine variables that reflect fine to coarse-scales of spatial and temporal variation. Four variables were related to climate: mean annual temperature, seasonality of precipitation, growing season precipitation (CHELSA v1.2; Karger et al., 2017), and the inter-annual variation of cloud cover (EarthEnv; Wilson and Jetz 2016). Two variables were related to vegetation productivity: spatial heterogeneity of EVI (EarthEnv; Tuanmu and Jetz 2015) and the mean enhanced vegetation index (EVI) for winter, derived from MODIS (Didan 2015). Two variables were related to soil characteristics: proportion of soil silt content and soil clay content (SoilGrids v2; Poggio et al., 2021). The final variable, the terrain ruggedness index (EarthEnv; Amatulli et al., 2018), represented topographical variation. All variables were resampled to a 1 x 1 km cell size from their native projections (see Supplementary materials Table A1 for further details and provenance).

We used variance inflation factor analysis (VIF; '*usdm*' package, ver. 1.1-18) to identify potential multicollinearity issues between our environmental variables. Multicollinearity between predictors can negatively affect SDMs by causing unstable parameter estimates and biased test statistics (Belsley 1991, Chatfield 1995, Dormann et al. 2013). All VIF values indicated low multicollinearity (<3). Thus we did not exclude any variables on this basis (Zuur et al. 2010).

Simulating positional uncertainty in occurrence data

Positional error in species occurrences may vary depending on the data source and original collection method (e.g., geographic coordinates or written description). Whereas for occurrences gathered with GNSS (Global Navigation Satellite System), the positional uncertainty may range from a couple up to tens of meters, occurrences gathered with older technologies or those georeferenced from museum databases may have positional uncertainty of up to tens of kilometers (see for example Moudrý and Devillers 2020). Therefore, to mimic the range of positional uncertainty in real datasets, we shifted occurrences in a (uniform) random direction according to four different scenarios. As the resolution of environmental variables used in SDMs was different for both virtual species (1 x 1 pixel respectively 50 x 50 m) and for Band-tailed pigeon (1 x 1 km), we shifted occurrences in a random direction by drawing a shift distance from a uniform distribution from the following distances: S1: 1 – 2 pixels, S2: 2 – 5 pixels, S3: 5 – 10 pixels, S4: 10 – 30 pixels (Workflow1); S1: 50 – 100 m, S2: 100 – 250 m, S3: 250 – 500 m, S4: 500 – 1500 m (Workflow2); S1: 1 – 2 km, S2: 2 – 5 km, S3: 5 – 10 km, S4: 10 – 30 km (Workflow3; see Supplementary materials Table A2). If the original data points were shifted outside of the study area, the shift was recalculated until the new coordinates were located within the boundaries of the study area.

Model fitting and evaluation

We built species distribution models in the statistical software R (package '*sdm*' ver. 1.0-98; Naimi and Araújo 2016) using the MaxEnt modeling method (Phillips et al. 2006), a presence-background method often adopted in ecological studies (Linda et al. 2016, Rodríguez et al. 2019, Santamarina et al. 2019, Ancillotto et al. 2020, El-Gabbas et al. 2020, Boral and Moktan 2021, Ellis-Soto et al. 2021, Venne and Curie 2021, Gábor et al. 2022b). We used 10,000 randomly sampled background points and default model settings (Phillips and Dudík 2008), except we set the beta parameter to 0.5 and restricted used features. Only hinge features were allowed for virtual species (Workflow 1 and 2). Although hinge features might lead to model overfitting, we used them as our virtual species response to the environment was defined using a normal distribution (Elith et al. 2010). For band-tailed pigeon, we sampled background points only in the extent of species occurrences (western coast of USA; VanDerWal et al. 2009, Barve et al. 2011, Merow et al. 2013) and used quadratic features to avoid overfitting (Austin 2007).

We used a variety of discrimination metrics to evaluate predictive model performance. We used the Sorensen index (SI), recommended for SDMs evaluation using presence-only occurrences (Li and Guo 2013, Leroy et al. 2018). SI ranges from 0 to 1, where 0 means that none of the predictions matched any observation, and 1 means that predictions perfectly fit observations without any false positive or false negative (Leroy et al. 2018). In addition, we also calculated overprediction (OPR, Barbosa et al. 2013) and underprediction (UPR, Fielding and Bell 1997) rates to explore whether positional uncertainty led to a consistent over-/underprediction bias. The OPR measures the percentage of predicted presences corre-

sponding to false presences, whereas UPR measures the percentage of actual presences not predicted by the model (Fielding and Bell 1997, Barbosa et al. 2013, Leroy et al. 2018). In addition, we also computed the true skill statistic (TSS, Allouche et al. 2006), despite recent criticisms about its use due to prevalence dependency (Lobo et al. 2008, Jiménez Valverde 2012, Leroy et al. 2018). We explored TSS in addition to SI as it is still widely applied in ecological studies (e.g., Fern et al. 2020, Holder et al. 2021, Eduardo et al. 2022, Sanguet et al. 2022). TSS ranges from −1 to +1, where +1 indicates perfect agreement and values of zero or less indicate random performance (Allouche et al. 2006).

We ran SDMs using five-fold cross-validation (Merow et al. 2013), where species occurrences and background points were divided randomly into five-folds, and each fold was retained for model testing while the other four folds were used for model training. We repeated each experiment 50 times, and evaluations represent averages of the 50 repetitions.

We evaluated each predictor variable's importance and visualized predicted responses to the environmental variables to explore the effect on inference about generative mechanisms. To estimate variable importance, we used a leave-one-out sensitivity analysis method which calculates the improvement in the model performance with the inclusion of each variable compared to when the variable is excluded (AUCtest; Murray and Conner 2009). Response curves were automatically generated by the '*sdm*' package (Naimi and Araújo 2016) using the "evaluation strip" approach. This approach visualizes species responses for used environmental variables by including data frames that show the distribution of observed presence point locations within the environmental range investigated by the evaluation strips (Kindt 2018; detailed in Elith et al. 2005).

3.3.4. Results

Model predictive performance

Note that here we present only results for the Sorensen index to simplify the presentation of the results, but results for the true skill statistic (TSS) qualitatively followed the same pattern.

In general, in Workflow 1 where points were not shifted (hereafter unaltered), models achieved excellent model performances for species with narrow niche widths (SI > 0.9, OPR and UPR < 0.03; see Figure 3.11 and Supplementary material Figure A3). However, predictive performance generally decreased with increasing species niche width (SI decreased on average by 0.53, while OPR and UPR increased on average by 0.57, respectively by 0.5). Predictive performance generally decreased with increasing positional error in occurrence data. Where the level of SAC was low and the sample size small, the more pronounced was the negative effect of positional error in species occurrences (Figure 3.11 and Supplementary material Figure A3). In Workflow 2, unaltered models achieved very good model performances (SI > 0.86, OPR < 0.04, UPR < 0.19, Figure 3.11 and Supplementary material Figure A3). Again, performance decreased with increasing niche width and with introducing positional error (Figure 3.11 and Supplementary material Figure A3).

Unaltered models for band-tailed pigeon (Workflow 3) achieved very good model performance (SI achieved on average 0.86, OPR 0.13, and UPR 0.15), and once more, positional uncertainty led to decreases in model performance (Figure 3.12). Compared to virtual species data, the decrease in model performance was, however, lower (SI for Workflow 2 virtual species decreased on average over 0.3, vs. an average of 0.03 for the real species; Figures 3.11, 3.12 and Supplementary material Figure A3).

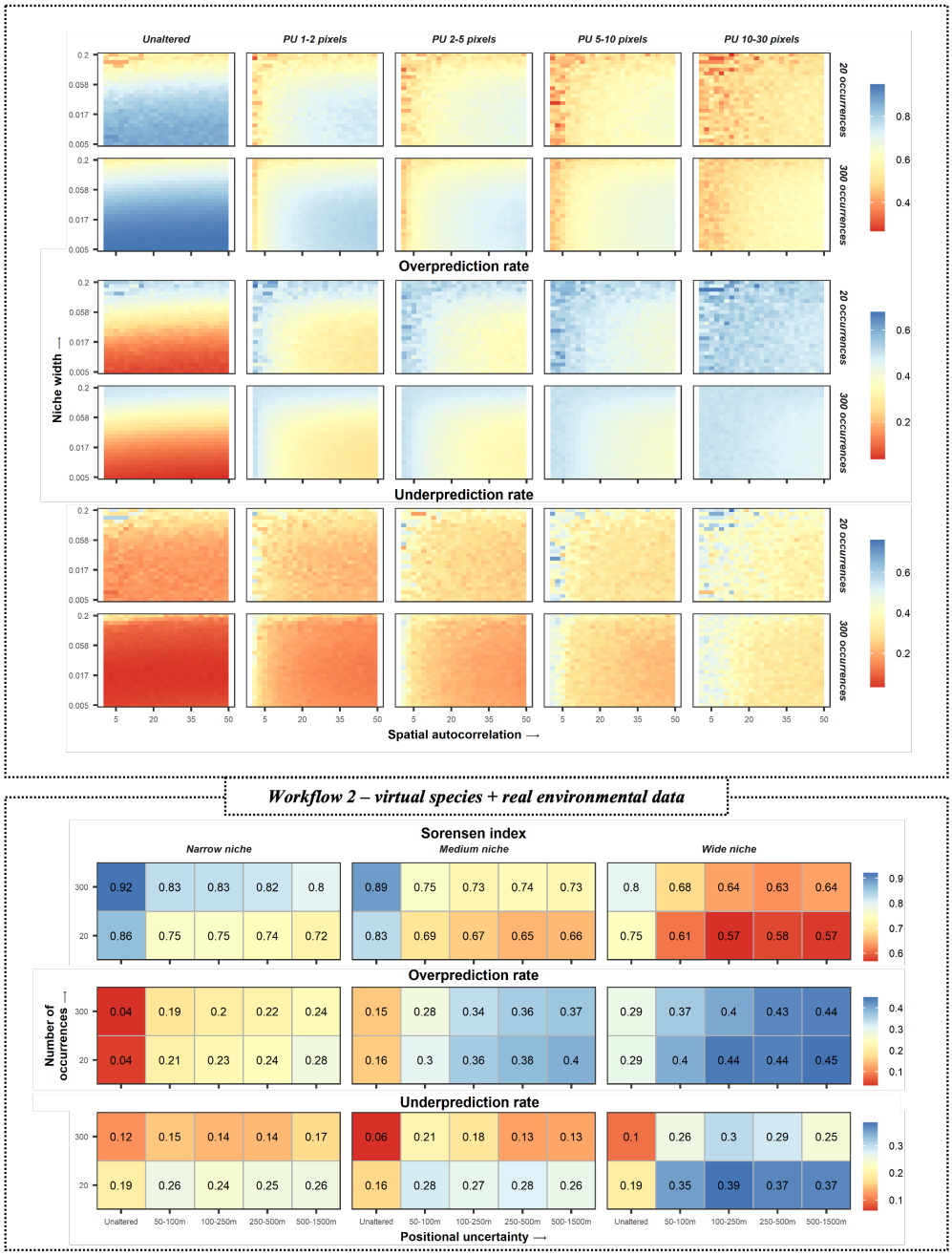


Figure 3.11: Change in model performance, measured through the Sorensen index, overprediction rate and underprediction rates for workflow 1 and 2. Resulting changes for all scenarios are plotted in Supplementary materials Figures A3. Values represent averages of the 50 repetitions.

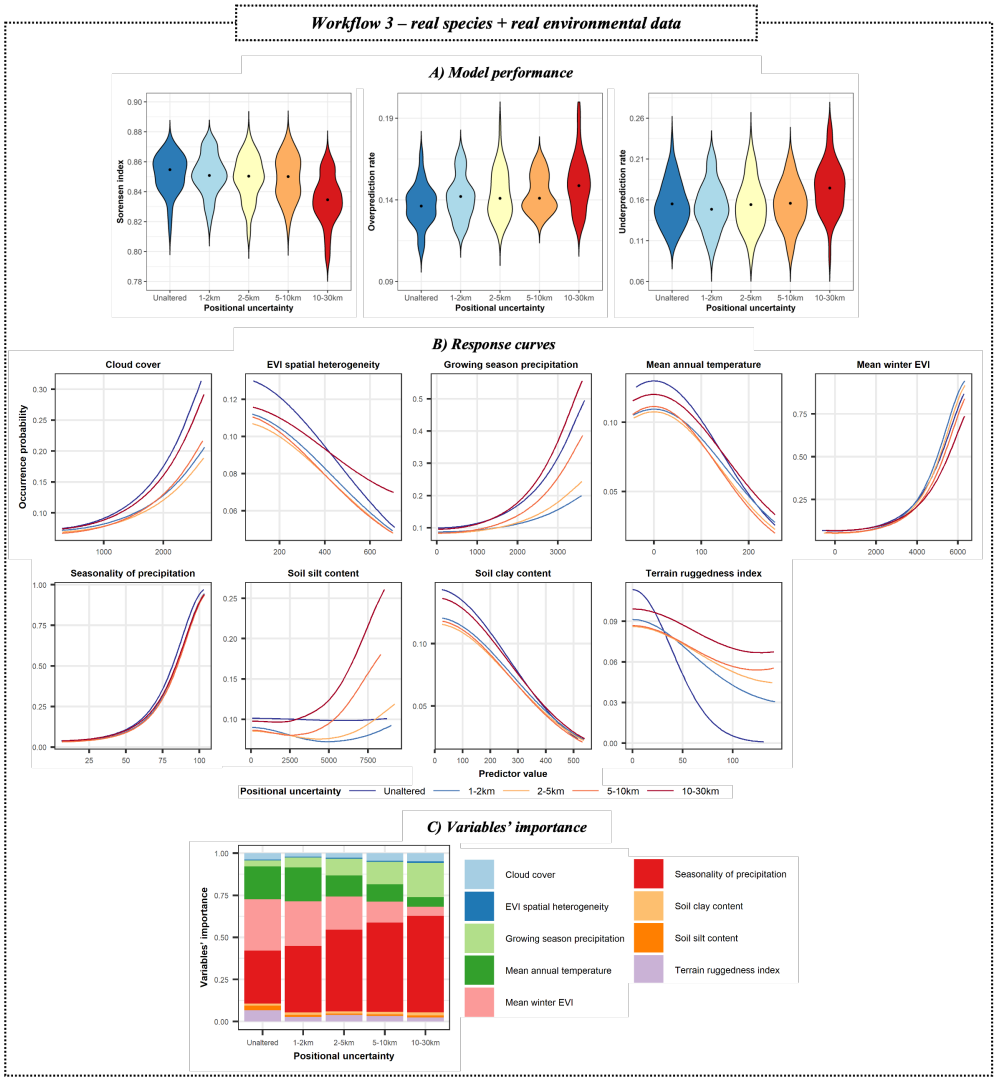


Figure 3.12: Resulting performance metrics (A), response curves (B) and variables' importance (C) of unaltered and altered models for all scenarios with real species and real environmental data. Values represent averages of the 50 repetitions.

The general increase in over and underprediction rates across all workflows implies that models fit to data with positional error tended to overpredict and, at the same time, underpredict species habitat suitability. Therefore, using positionally uncertain data might be highly risky for some ecological applications (e.g., nature conservation).

Variable importance

For Workflow 2 models correctly, across all modeled scenarios, detected the aspect and elevation, which were used to generate virtual species, as the most influential variables. (only these variables were used to generate virtual species; Figure 3.13 and Figure A4 in supplementary material). Increasing sample size increased the estimated importance of aspect and

elevation. On the other hand, as niche width increased, models estimated greater importance of other variables.

For band-tailed pigeon (Workflow 3), the most influential variables were mean winter EVI (16.5%) and seasonality of precipitation (17%), followed by mean annual temperature (10.1%), with other variables below 10% (terrain ruggedness, soil clay content, growing season precipitation, EVI spatial heterogeneity, soil silt content, cloud cover; Figure 3.12).

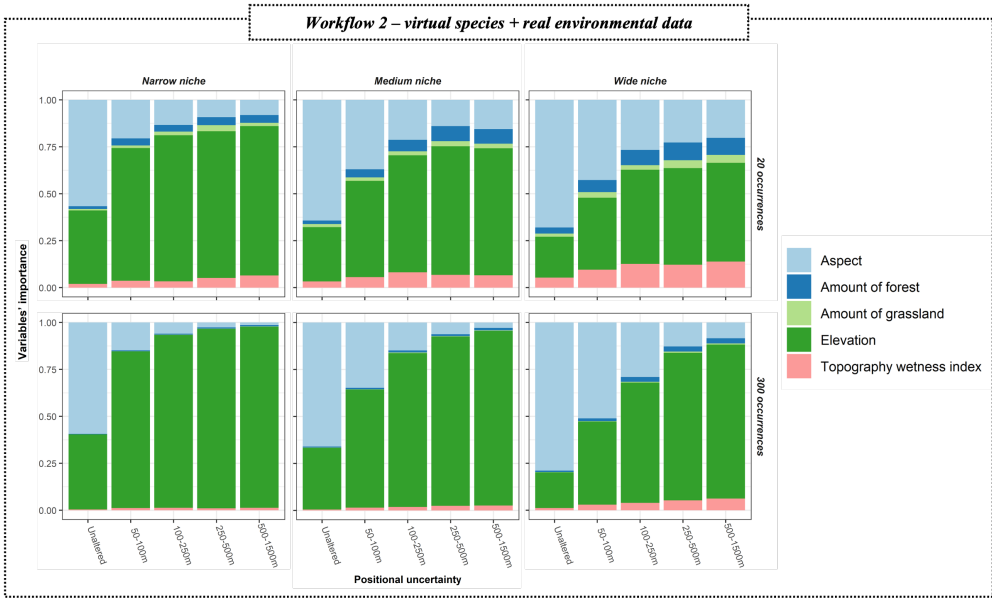


Figure 3.13: Comparison of the change in variables' importance between models generated with positionally accurate presences (Unaltered) and models built with various positional error in the data across various sample sizes. Values represent averages of the 50 repetitions. Variables' importance for all scenarios are plotted in Supplementary material Figure A4.

Positional errors led to changes in variable importance. Workflow 2 models correctly inferred the most influential variables regardless of the degree of positional error, although in the high-error scenario, the absolute importance of aspect decreased in importance by 41,2% while the importance of elevation increased by almost 32% (Figure 3.13 and Figure A4 in supplementary material). For variables with minor importance, we generally observed only small changes (i.e., < 4.4% change) to their importance.

In workflow 3, models correctly inferred the seasonality of precipitation as the most influential variable independently of positional error. As error increased, we observed a decrease in the importance of mean winter EVI and mean annual temperature (by 13.6% and 6.1% respectively), and an increase in the importance of seasonality of precipitation (by 13.1%), as well as growing season precipitation (by 10.8%), becoming the second most influential variable (Figure 3.12).

Response curves

In Workflow 1, unaltered models accurately recovered the true mean response except for species with wide niches and 20 occurrences., where models failed to recover the response

mean (see Figure 3.14 and Supplementary material Figure A5). The estimated standard deviation, however. Varied considerably across different scenarios. Where sample size was lowest (20 occurrences), models overestimated the standard deviation and thus tended to overpredict the probability of suitable habitat. The standard deviation estimation was significantly improved with increasing sample size and was, on average better for scenarios with higher levels of SAC (homogenous variables). This pattern was independent of species niche width (Figure 3.14 and Supplementary material Figure A5).

Therefore, where there was positional error and small sample sizes (20 occurrences), models were unable to accurately estimate either the response mean or standard deviation across all SAC types. This ability was improved with increasing sample size and level of SAC (Figure 3.14 and Supplementary material Figure A5). Across all species niche widths, where positional error was pronounced, models were better able to estimate response means and standard deviations when sample sizes were large and SAC levels high. However, even with the largest sample size (1000 occurrences) and the highest level of SAC, the models overestimated the standard deviation (Figure 3.14 and Supplementary material Figure A5).

For workflow 2, the unaltered models were able to recover responses to both aspect and elevation, which were used to generate virtual species. This was independent of modeled scenario (Figure 3.14 and Supplementary material Figure A5). When introducing positional error, models could still capture the approximate response to the elevation with a relatively high SAC level. However, although models recovered the correct shape of the response curve, standard deviation increased. In contrast, for aspect (low SAC level), the models developed with positional uncertainty failed to recover the correct response curve, even when larger sample sizes were used (Figure 3.14 and Supplementary material Figure A5). Note that Workflow 2 models could estimate the response even with the smallest sample size. This is in contrast to Workflow 1, potentially due to greater model complexity. These results support our assumptions that models developed with data containing positional uncertainty might be able to detect species responses for variables with high SAC levels, but fail to detect meaningful responses for variables with low SAC levels.

The positional error also affected response curves for the band-tailed pigeon. The largest changes to response curves and over- and underprediction tended to occur with the most heterogeneous variables, for example, the terrain ruggedness index (see Figure 3.12).

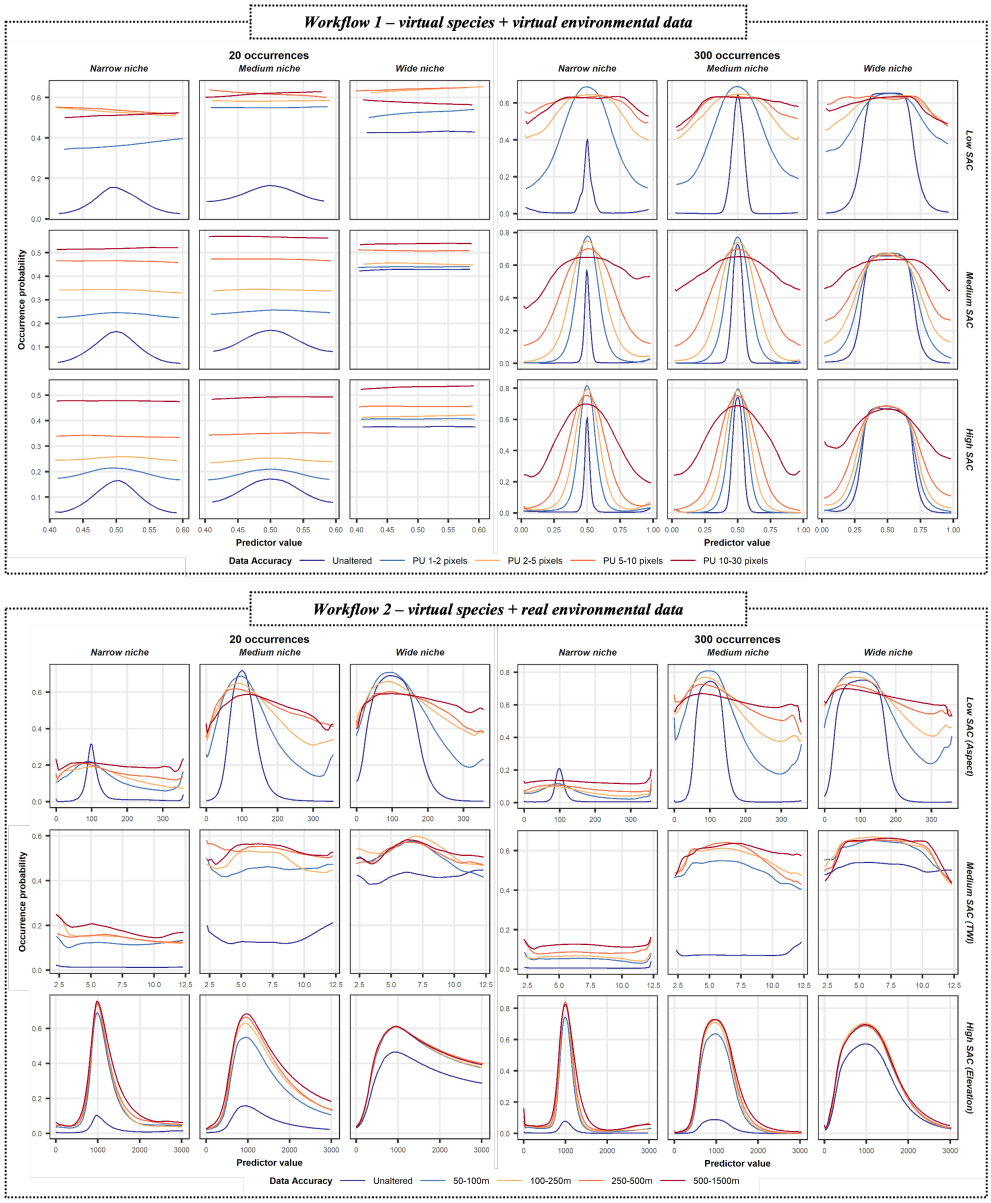


Figure 3.14: Variation of selected environmental response curves across models with unaltered and altered data, and various sample sizes for workflow 1 and 2. Resulting response curves for all scenarios are plotted in Supplementary materials 2. Response curves represent averages of the 50 repetitions.

3.3.5. Discussion

In our study, we used combinations of virtual and real environmental variables with different SAC levels, sets of virtual species with a variety of niche widths, and one real species to explore how positional error in species occurrence data can affect model performance and its ecological interpretability. Specifically, we investigated the ability of SDMs to appropri-

ately detect species' responses to the environment and variable importance using various scenarios with artificially applied positional error.

Our results show that positional uncertainty in species occurrences leads to a decrease in model predictive performance across all combinations of species niche widths, sample sizes, and SAC levels of the environmental variables, but that the magnitude of the negative impact of positional uncertainty varied for different combinations and depending on the distance that points were shifted. The negative influence was most pronounced for species with a narrow niche and scenarios with more heterogenous environmental variables. This is consistent with previous studies, which concluded that more accurate occurrence data generally yielded better-performing SDMs (Visscher 2006, Johnson and Gillingham 2008, Osbourne and Leitao 2009, Tulowiecki et al. 2015, Mitchell et al. 2017, Soultan and Safi 2017, Fernandes et al. 2019). It is important to highlight, that the magnitude of the negative effect of positional uncertainty varied across prior studies. This can be explained by using environmental variables with different heterogeneity (Naimi et al. 2011, 2014) or by using species with varying niche width (Gábor et al. 2020a).

Our models for real species were less affected by positional uncertainty than models for virtual species. This could possibly be explained by the spatial error already embedded in the real species data; meaning the "unaltered" scenario actually presents some minimal, but unknown, level of error, as well as errors in the environmental layers. Additionally, observations of transient individuals merely passing through unsuitable habitats could also contribute to this finding, although we attempted to filter out such cases. There may also be spatial and/or environmental biases in the data, such as disproportionate sampling efforts in locations where specific behaviors take place (e.g., water sources), where species detectability is increased (e.g., open areas), or areas with greater accessibility (e.g., near roads, walking trails; Kramer-Schadt et al. 2013, Fourcade et al. 2014).

On the other hand, our results showed that models built with even positionally inaccurate data may still be ecologically interpretable. The absolute and relative importance of environmental predictors and the shape of species-environmental relationships co-varied with the level of positional uncertainty. But these differences were much weaker than those observed for overall model performance. This indicates that low model performance doesn't necessarily lead to low capacity to infer which variables drive species distributions and the strength of those drivers. It is important to note that sample size and the SAC level of environmental variables play an important role here. In general, the higher the sample size and the lower the level of SAC in environmental variables the better were models able to recover response curves and detect the importance of the environmental variables (Figures 3.12, 3.13, 3.14).

On the other hand, in the case of environmental variables with low SAC level (high heterogeneity), positional error obscured the main patterns (e.g., aspect in Workflow 2 or terrain ruggedness index in Workflow 3; see Figures 3.12, 3.14). Our results suggest that, at least for the example species, positionally inaccurate records may still prove useful for assessing the relative importance of environmental variables in generating species distributions and for determination of the shapes of species responses. Thus, for some purposes positionally inaccurate records need not be discarded (as is common practice; e.g., Watcharamongkol et al. 2016, Gueta and Carmel 2016). This finding is particularly fortuitous because discarding positionally uncertain occurrence data can limit our ability to estimate range sizes and overestimates exposure to climate change (Smith et al. 2022).

Drawing methodological conclusions based on real data is difficult since the true underlying

population distribution is unknown, as are data deficiencies that could potentially affect results (Winner et al. 2018, Meynard et al. 2019, Šímová et al. 2019, Vollering et al. 2019, Mendes et al. 2020, Somveille et al. 2020, Yanco et al. 2020, Grimmett et al. 2021, Inman et al. 2021, Jiménez-Valverde 2021). On the other hand, simulated datasets simplify the real world, and their results should be interpreted cautiously (e.g., Wunder et al. 2008, Zurell et al. 2010, Meynard et al. 2019). Indeed, our results show that a virtual species approach may show different results than those using a real species. For example, our virtual species simulations showed a rapid decrease in model performance with increasing positional error, whereas the band-tailed pigeon showed only a slight decrease in model performance. We strongly recommend that future studies should follow a growing trend and combine simulations and real species data when studying methodological questions (see, for example, Fithian et al. 2015, Guélat and Kéry 2018, Mertes and Jetz 2018, Renner et al. 2019).

Although this study provides extensive insights that are optimistic about the potential utility of SDMs, caution is warranted in generalizing these results, and further research is needed. For example, future studies could explore whether our findings are robust to different MaxEnt settings, various modeling techniques, response, and analysis grain and different types of data uncertainty (e.g., spatial bias rather than a random error). In addition, within global aggregation databases, spatial uncertainty may not be uniformly distributed. Analyses that characterize the patterns of spatial uncertainty within these databases would allow researchers to identify situations wherein models are likely to fail.

3.4. Positional errors in species distribution modelling are not overcome by the coarser grains of analysis

Lukáš Gábor, Walter Jetz, Muyang Lu, Duccio Rocchini, Anna Cord, Marco Malavasi, Alejandra Zarzo-Arias, Vojtěch Barták and Vítězslav Moudrý

Adapted from Methods in Ecology and Evolution, 2022, 13(10), 2289-2302.

Publication metrics:
Quartile (2023): Agricultural and Biological Sciences - Ecology, Evolution, Behavior and Systematics (Q1)
Impact Factor (2023): 8.33
SCImago Journal Rank (2023): 2.99

The first author contributed to the study as follows: study conception and design (equal), data curation (lead), analysis and interpretation of results (equal), visualization (lead), writing – original draft (lead), writing – review and editing (lead), overall study supervision (equal), funding acquisition (lead).

The link to the published article and supplementary materials can be found here:
https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13956

3.4.1. Abstract

1. The performance of species distribution models (SDMs) is known to be affected by analysis grain and positional error of species occurrences. Coarsening of the analysis grain has been suggested to compensate for positional errors. Nevertheless, this way of dealing with positional errors has never been thoroughly tested. With increasing use of fine-scale environmental data in SDMs, it is important to test this assumption. Models using fine-scale environmental data are more likely to be negatively affected by positional error as the inaccurate occurrences might easier end up in unsuitable environment. This can result in inappropriate conservation actions.

2. Here, we examined the trade-offs between positional error and analysis grain and provide recommendations for best practice. We generated narrow niche virtual species using environmental variables derived from LiDAR point clouds at 5 x 5 m fine-scale. We simulated the positional error in the range of 5 m to 99 m and evaluated the effects of several spatial grains in the range of 5 m to 500 m. In total, we assessed 49 combinations of positional accuracy and analysis grain. We used three modelling techniques (MaxEnt, BRT and GLM) and evaluated their discrimination ability, niche overlap with virtual species and change in realized niche.

3. We found that model performance decreased with increasing positional error in species occurrences and coarsening of the analysis grain. Most importantly, we showed that coarsening the analysis grain to compensate for positional error did not improve model performance. Our results reject coarsening of the analysis grain as a solution to address the negative effects of positional error on model performance.

4. We recommend fitting models with the finest possible analysis grain and as close to the response grain as possible even when available species occurrences suffer from positional errors. If there are significant positional errors in species occurrences, users are unlikely to benefit from making additional efforts to obtain higher resolution environmental data unless they also minimize the positional errors of species occurrences. Our findings are also applicable to coarse analysis grain, especially for fragmented habitats, and for species with narrow niche breadth.

Keywords: Georeferencing, Grain size, Resolution, Scale, SDM, Virtual species

3.4.2. Introduction

Species distribution models (SDMs) use species occurrence data and environmental explanatory variables to infer species-environment relationships and predict species distribution ranges (Ferrier et al. 2017). Despite their routine use and relatively well-established practices (Simões et al. 2020) and standards (Araújo et al. 2019, Merow et al. 2019), some methodological considerations still require further investigation. With the increasing availability of heterogeneous data from a multitude of sources of varying quality, careful assessment of uncertainties and purpose-built methodologies are becoming more important (Wüest et al. 2019). Indeed, recent recommendations and methodological improvements are particularly relevant to data quality issues such as positional error, sampling bias, sample size and scale. Specialised tools have been developed for the identification of positionally inaccurate records (e.g., Robertson et al. 2016, Zizka et al. 2019). Similarly, development and testing of sampling bias correction methods continue (Gábor et al. 2020b, Inman et al. 2021) as well as the research into the effects of sample size (McPherson et al. 2004, McPherson and Jetz 2007, Hallman and Robinson 2020, Jiménez-Valverde 2020) and of changing the grain of response and explanatory variables (Mertes and Jetz 2018, Šímová et al. 2019).

Additionally, a key question, namely at which spatial scales (grains) the ecological processes underlying species distribution patterns operate, continues to be debated (Pearson and Dawson 2003, Miguet et al. 2016, Mertes and Jetz 2018). SDMs can be developed on a very wide range of grains (e.g. from 1 m^2 to 10,000 km^2 or more), and several studies (e.g. Guisan et al. 2007; Kaliontzopoulou et al. 2008; Seo et al. 2009) reported effects of the analysis grain on the performance of SDMs. At some spatial scales, species respond more strongly to their environment than at others (Holland et al. 2004, Mayor et al. 2009, McGarigal et al. 2016). This is often referred to as ecological scale, scale of effect, response grain or response scale (Holland et al. 2004, Wu and Li 2006, Mertes and Jetz 2018). Here, we follow Mertes and Jetz (2018) and use the term "response grain" to indicate the theoretical scale at which individuals of a species respond to environmental factors and "analysis grain" to describe the spatial unit (grain) at which the species occurrence is modelled. As the chosen analysis grain affects our ability to detect the species' response to environmental factors (variables), factors such as positional errors of species occurrences, resolution of available environmental data, and the response grain on which species are expected to respond to the environment need to be considered (Schneider 2001, Dungan et al. 2002, Lechner et al. 2012, Lecours et al. 2015).

It is increasingly recognized that positional uncertainty (associated with the location of species observations) is an important factor to consider during the modelling process. Positional errors cause problems in modelling, as environmental conditions at the recorded locations might differ from those at actual locations, which (as was demonstrated) can have a significant impact on SDM results. For example, Visscher (2006) showed that positional error can bias inferences about species-environment relationships. Similarly, Johnson and Gillingham (2008) concluded that positional errors have a significant effect on model quality, and Osbourne and Leitao (2009) recommended minimising positional errors through careful study design and data processing. More recently, Hefley et al. (2014) pointed out that positional errors can lead to biased estimates of regression coefficient. Indeed, the Darwin Core Standard (https://dwc.tdwg.org/) has proven to be useful for recording positional uncertainty of species occurrences (Wieczorek et al. 2012), and the importance of georeferencing accuracy has been highlighted by many studies (e.g., Moudrý and Devillers 2020), including a report on the suitability of Global Biodiversity Information Facility (GBIF) data for use in SDMs (Anderson et al. 2016).

Notably, with the increasing use of fine-scale resolution data in SDM, such as variables derived from LiDAR with a resolution of a few meters (e.g. Pradervand et al. 2014, Simonson et al. 2014, Sillero and Gonçalves-Seco 2014, Lecours et al. 2020, Wüest et al. 2020, Moudrý et al. 2021), the negative effects of positional error in species occurrence data are no longer associated only with relatively old datasets (e.g. from herbarium or museum collections) but it is also necessary to consider positional errors inherent to data georeferenced using global navigation satellite systems. Indeed, Gábor et al. (2020b) used a 5 x 5 m analysis grain and reported that the largest drop in model performance was observed at the smallest simulated positional error of 5-10 m (they simulated errors up to 500 m).

Both positional error and adopted analysis grain have been intensively studied; however, despite their interconnectedness, their interactions and trade-offs are rarely systematically addressed (but see Engler et al. 2004, Montgomery et al. 2011, Cheng et al. 2021). Particularly, the trade-off between the adopted analysis grain and positional error of species occurrence data is poorly acknowledged. Typically, studies try to balance these interconnected issues based on available data and metadata (i.e., users might know the positional error of occurrences but do not know the optimal grain and vice versa). For example, researchers aim to georeference species occurrences with respect to adopted analysis grain (Ballesteros-Mejia et al. 2017) or, when using already georeferenced data, they remove imprecise occurrences (e.g., records with latitude and longitude precision lower than three decimal places or with known high positional uncertainty; Gueta and Carmel 2016, Watcharamongkol et al. 2018, Ellis-Soto et al. 2021). Alternatively, coarsening the analysis grain can be used for correcting georeferencing errors (Engler et al. 2004, Moudrý and Šímová 2012, Keil et al. 2014, Vollering et al. 2016, Sillero and Barbosa 2021). These techniques, however, have a drawback: removing positionally inaccurate records or coarsening the analysis grain reduce the sample size. Moreover, the latter approach can lead to the loss of explanatory power of the model (as the grain at which species respond to the environment might be better represented by a finer grain). This may indeed limit our ability to observe how species respond to the environment (Mertes and Jetz 2018).

All in all, it is evident that both analysis grain and positional accuracy are important and interacting factors affecting SDM results (i.e., environmental niches and spatial distributions of modelled species). However, the knowledge of how they interact and the implications for modelling practice is lacking. It is crucial to have this knowledge, especially with increasing availability of fine-scale environmental data (e.g., Haesen et al. 2021, Li at al. 2021), and their use in predictive models developed for conservation and climate change studies (see for example Lembrechts et al. 2019a, 2019b, Zellweger et al. 2019, Stark and Fridley 2022). Therefore, we here address the following questions: (i) What are the trade-offs between analysis grain and positional error when modelling species distributions? (ii) Is it advisable to coarsen the analysis grain to minimize the effect of the positional error, or should the analysis grain be kept as close as possible to the assumed response grain, regardless of the positional error?

3.4.3. Materials and Methods

LiDAR data and derived environmental variables

We used a point cloud from airborne laser scanning of Krkonose Mountains National Park, Czech Republic, that covers over 370 square kilometres (approximately 30 km in west/east direction and 13 km in south/north direction), to derive three fine-scale environmental vari-

ables. It has been shown that the negative effect of positional error varies according to the degree of spatial autocorrelation in environmental variables. The lower is the spatial autocorrelation in environmental variables the more pronounced is the negative effect of positional error in species occurrences (Naimi et al. 2011, 2014). Therefore, we chose environmental variables with various levels of spatial autocorrelation to mimic a real modeling situation (Supplementary materials Figure A1). Note, that spatial autocorrelation is a function of resolution and may change as the analysis grain is coarsened (see Mertes and Jetz 2018). However, this is not our case, as the environmental variables maintained similar spatial autocorrelation across all used response grains (see Supplementary materials Figure A1). Specifically, we used the canopy height model (CHM) representing structural variability of the canopy, topographic wetness index (TWI) as a surrogate for soil moisture, thus affecting vegetation composition, and altitude in the form of a digital terrain model (DTM) as a sur¬rogate for microclimatic conditions. All these variables have been used in other studies for modelling species distributions, e.g. of birds (e.g. Vogeler et al. 2014, Reif et al. 2018, Bakx et al. 2019). Hence, our virtual species might represent a bird with certain habitat requirements in terms of vegetation structure, climate and terrain characteristics. To derive the three environmental variables at a resolution of 5 x 5 m, first the point cloud was classified into vegetation, building, and ground classes in the ENVI and LAStools software (Klápště et al. 2020). Second, following Khosravipour et al. (2016), we used points classified as vegetation to produce the CHM; points representing ground were used to create the DTM, which was subsequently used to derive the TWI.

Generating virtual species

We adopted the virtual species approach, which is increasingly used to answer methodological questions related to SDMs (Zurell et el. 2010). This popularity is due in particular to the fact that it is difficult to draw clear methodological conclusions with real data, since the actual distribution as well as data deficiencies that might influence the results are unknown (Moudrý 2015, Meynard et al. 2019, Grimmett et al. 2021, Inman et al. 2021). We used the *virtualspecies* package (ver. 1.5.1) in the statistical software R (R Core Team 2021) to generate virtual species (Leroy et al. 2016). To begin, we defined the response of virtual species to the environmental gradient at a resolution of 5 x 5 m (i.e., the finest resolution at which environmental variables were available). We used a normal distribution with the following parameters: (i) mean canopy height of 9 m and standard deviation of 4 m; (ii) mean altitude of 846 m and standard deviation of 100 m; and (iii) mean topographic wetness index of 8 and standard deviation of 0.4 m. These parameters allowed us to simulate virtual species with a narrow niche breadth as it has been suggested that SDMs of such species are more prone to positional error (Visscher 2006; Gábor et al 2020). We then multiplied the responses to obtain an environmental suitability raster. We applied the probabilistic approach (logistic function with $\alpha = -0.05$ and $\beta = 0.3$) to convert the environmental suitability raster into probabilities of occurrences that were subsequently used to sample binary presence-absence rasters. We developed both presence-only and presence-absence models (see below), using 99 presence sites and 200 absence sites (i.e., sample prevalence of 0.33), and a uniform random distribution for sampling species presences and absences. The virtual species could be recreated using the ‘vs‘ object and R script that is available via the Dryad repository (please see the link provided at the beginning of this chapter).

Simulating positional error and coarsening the analysis grain

Positional error in species occurrence data may range from a few metres up to hundreds of metres, depending on the data gathering technique and the source of the error. Here, we simulated the positional error in the range of 5 m to 99 m. We shifted each occurrence point in a random direction by a specified distance according to 6 scenarios. Each scenario is associated with a different shift, as follows: S1: 5 – 9 m, S2: 10 – 19 m, S3: 20 – 29 m, S4: 30 – 39 m, S5: 40 – 49 m and S6: 90 – 99 m. The scenario with the original, i.e., not shifted, data is referred to as "unaltered" hereafter. The R functions we used to simulate positional error in species occurrences are available in the R script via the Dryad repository. To test the effect of coarsening the analysis grain and, in particular, to assess whether the coarsening of the analysis grain can compensate for the negative effect of the positional error, we ran models at seven analysis grains representing two distinct situations, namely: (i) the response grain is known and relatively fine-scale data are available (5 x 5 m, 20 x 20 m, 40 x 40 m, 60 x 60 m, 80 x 80 m, and 100 x 100 m) and (ii) the analysis grain is selected on the basis of data availability (500 x 500 m). In the first situation, we used small steps (changes) and multiple scales to capture any minor changes, while in the second situation, the analysis was conducted with a grain considerably coarser than the response grain (a hundred times coarsened grain), which is undoubtedly a situation prevalent in current modelling practice. Thus, a total of 49 combinations of positional accuracy of species occurrences and analysis grains were evaluated. All environmental variables were resampled to coarser grains using the mean values of the original data (Moudrý et al. 2019). Note that coarsening the analysis grain results in multiple sampling sites ending up in the same cell (e.g., Engler et al. 2004, Guisan et al. 2007). When absences and presences occurred in the coarser grain cell after aggregation, the cell was considered a "presence" cell, resulting in a small decrease in the number of absences.

We did not observe multiple presences aggregated into a single cell (note that the largest analysis grain also limited the maximum number of background points for MaxEnt; see Table A1;). It is intuitive that the quality of the models is related to sample size. Indeed, prior studies showed that sample size play an important role in SDMs. In particular, they mostly concentrated on the effects of available presences on the development of accurate presence-only models (e.g., Wisz et al. 2008; van Proosdij et al. 2016). Recently, Liu et al. (2019) used virtual species approach and recommended that hundreds of presences are needed to reach the plateau where increasing the sample size adds little to the model performance. Therefore, we keep constant number of 99 presences for all scenarios. McPherson et al. (2004) evaluated the effects of sample size on the development of presence-absence models and shown that models trained with sample size of 300 (presences and absences) perform better than those trained with 100. In addition, Jiménez-Valverde et al. (2009) found that the effect of the sample size becomes apparent for models trained with less than 70 samples. Therefore, for presence-absence models we keep the constant number of 99 presences, and we let the absences to slightly vary between 150 and 200 (Table A1). Such minimal changes in number of absences certainly did not affect our results.

Model fitting

Three common modeling methods were used to fit species occurrence to environmental predictors: generalized linear model (GLM), boosted regression tree (BRT) and the maximum entropy model (MaxEnt). GLM, implemented in the R package *glm2* (ver. 1.2.1, Nelder and Baker 1972, Oksanen and Minchin 2002), and boosted regression trees (BRT), implemented

in the *gbm* package (ver. 2.1.5, Friedman et al. 2000), represented presence-absence methods, and MaxEnt, implemented in the *dismo* package (ver. 1.1-4, Phillips et al. 2006; ver. 3.4.3 of maxent.jar file, Phillips et al. 2020), a presence-background method. Using both presence-absence and presence-background methods allowed us to assess whether they are equally affected by positional errors and by coarsening of the analysis grain. The GLM was run with a logit link function and a binomial distribution. The quadratic terms of the environmental variables were included based on the known normal distribution curves of the response function. For BRT, we used Bernoulli distribution, shrinkage (learning rate) of 0.01, tree complexity of 1 (i.e., without interaction terms), bag fraction (the proportion of data used when selecting optimal tree number) of 0.5, and the maximum number of trees of 5,000. MaxEnt was used with default settings (i.e., auto features, logistic output format) and 10,000 backgrounds points. The only exception was for models with an analysis grain of 500 x 500 m, where the number of grids / cells was not sufficient to sample 10,000 background points, so we ended up with a smaller number of background points (see Tab. A1). The same three environmental variables (CHM, DTM and TWI) that were used in the process of generating virtual species were also used to fit the models in seven analysis grains (see the previous section).

Model evaluation

We used several discrimination metrics to evaluate the performance of the models. First, we used the Sørensen index (SI), which has been recommended for the evaluation of experiments testing SDM methodologies using virtual species (Li and Guo 2013, Leroy et al. 2018). We also aimed to determine whether predictions using erroneous/altered data tend to over- or underpredict species occurrences. Thus, we calculated the overprediction and underprediction rates. Overprediction refers to the proportion of observed absences in the predicted presence area, and underprediction measures the proportion of actual presences that were not predicted by the model (Barbosa et al. 2013, Leroy et al. 2018). However, these metrics use only three components (true positives, false positives and false negatives) of the confusion matrix and neglect the prediction of true negatives (Leroy et al. 2018). Because we manipulated the input data (i.e., introduced the positional error and changed the analysis grain), we were concerned that this might also affect the true negatives. Therefore, we added the area under the receiver operating characteristic curve (AUC; Fielding and Bell 1997; despite recent criticisms of this metric, see for example Lobo et al. 2008, Jiménez Valverde 2012) and the true skill statistics (TSS; Allouche et al. 2006), which are commonly used to assess the discriminatory power of models.

In addition, we took advantage of the virtual species approach and compared differences between the predicted distribution inferred from the models and the true probability of occurrence of virtual species in geographical space. However, it has been stressed that metrics used for niche comparison are seriously affected by the inclusion of large number of cells where the species are absent (i.e., with low occurrence probabilities) and it has been recommended to remove such cell from the evaluation (Rödder and Engler 2011). Therefore, for this evaluation, we extract occurrence probability only for occurrence data, which were used in the models. We used Spearman's rank correlation to quantify the differences. See Supplementary materials Figure A2 for visual comparison between virtual species true distribution and predicted probability of all modelled scenarios. Note that this comparison was performed using the same resolution for all models' predictions (i.e., 500m).

The model performance was evaluated at the analysis grain at which the individual models were fitted, which is a common practise in studies evaluating effect of analysis grain on the performance of SDM (e.g., Guisan et al. 2007, Kaliontzopoulou et al. 2008, Seo et al. 2009, Mertes and Jetz 2018, Lembrechts et al. 2019a, Zellweger et al. 2019, Stark and Fridley 2022). Performance metrics for each model were calculated using five-fold cross-validation for which 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 assess the performance. We performed the entire process from species generation to model evaluation 50 times and calculated average values and confidence intervals (MacKinnon and White 1985) of validation metrics from all replications. See Figure 3.15 for an overview of the general modelling process. Besides comparison of models' performance, we used linear regression to quantify how introducing positional error and coarsening of environmental variables affects species realized niche.

3.4.4. Results

Effects of positional error and analysis grain on species realized niche

Figure 3.16 shows linear regression line plots of species realized niche for unaltered and altered occurrence data across various analysis grains and all combinations of environmental data. It is obvious, that both introducing positional error and coarsening the analysis grain led to changes in species realized niche. More notably, the coarsening of analysis grain did not help to reconstruct the original niche. The change in realized niche is more pronounced for combination of environmental variables with lower spatial autocorrelation (i.e., TWI versus CHM; see Supplementary material Figure A1).

Overall model performance

All metrics largely followed the same pattern. Therefore, we focus only on SI and Spearman's rank correlation (for AUC TSS, overprediction rate and underprediction rate values, see supporting information Figures A3, A4). BRT and MaxEnt performed very well while GLM performed slightly worse using unaltered data and resolution of environmental variables (5 x 5 m). The SIs of the unaltered models were 0.76 for MaxEnt, 0.74 for BRT and 0.67 for GLM (Figure 3.17). Spearman's rank correlation indicates that MaxEnt and BRT models using unaltered data have high niche overlap with virtual species. They reached Spearman's rank correlation of 0.95 and 0.9, respectively. In contrast GLM achieved lower niche overlap and Spearman's rank correlation of 0.6 (Figure 3.17).

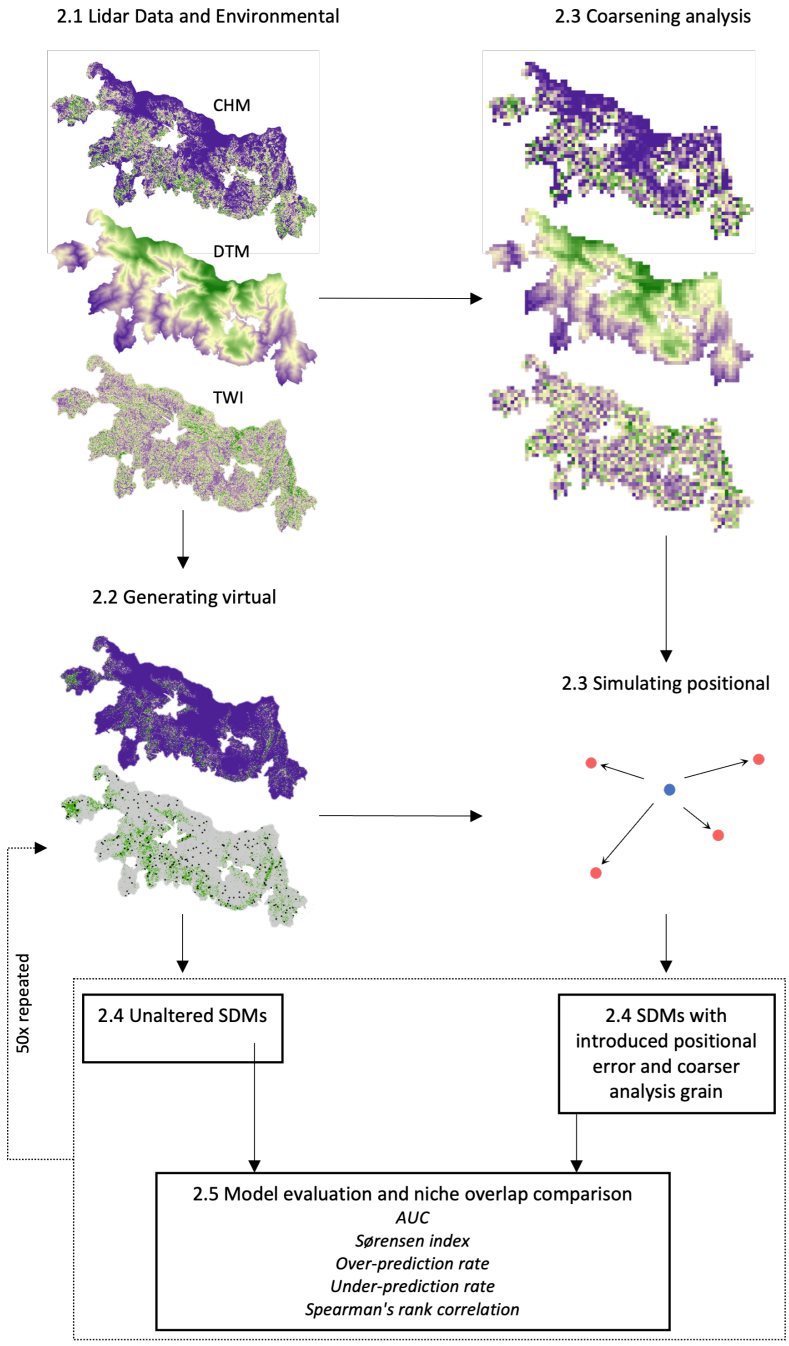


Figure 3.15: Overview of the modelling process. We first acquired and processed LiDAR data and selected three fine-scale environmental predictors (canopy height model, topographic wetness index, digital terrain model; Section 2.1). Further, we generated virtual species (2.2), simulated positional error in species occurrences, and coarsened analysis grain (2.3). We modelled species distribution with unaltered data as well as with shifted occurrences at various analysis grain sizes (2.4). In the last step, we evaluated models and compared their performance (2.5).

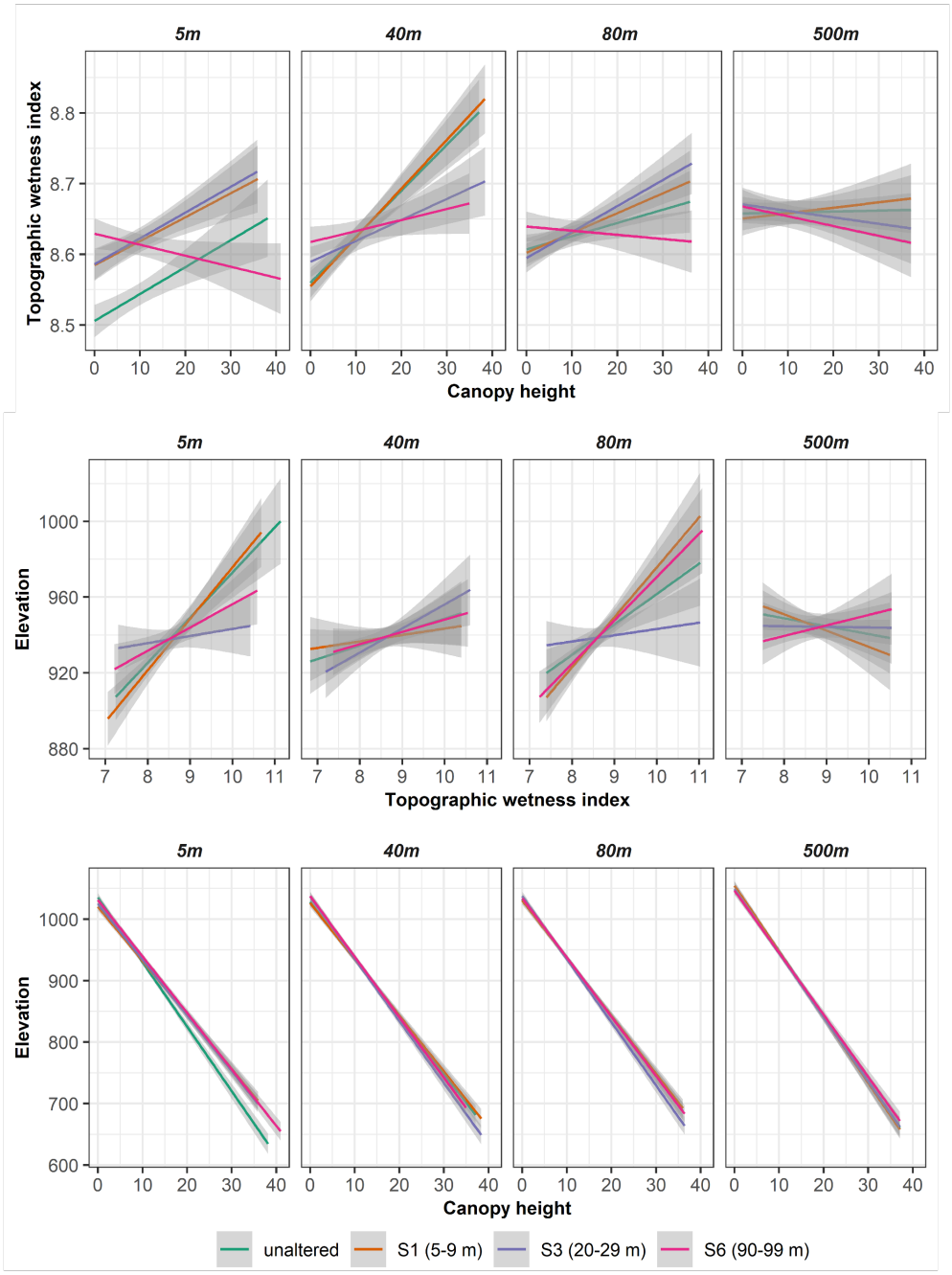


Figure 3.16: Comparison of changes in realized niche as a result of positional error in species occurrences and coarsening the analysis grain. Different colours show various levels of positional uncertainty while columns show different analysis grain. The line is obtained by linear regression and grey colour shows 95% confidence interval.

Effects of positional error and analysis grain

The performance of all modeling methods was negatively affected by the positional error in species occurrences. Results show a clear trend of decreasing model performance and increasing overprediction and underprediction rate with increasing positional error (Figure 3.17 and Supplementary material Figure A3), with the largest drop in performance occurring once positional error was introduced (i.e., between the no-error and 5 - 9 m error categories). For example, SI dropped from 0.76 to 0.72 and from 0.74 to 0.67 for MaxEnt and BRT, respectively (Figure 3.17). As the position error continued to increase, a slow but gradual decline in model performance was observed. The exception from this pattern is GLM modeling method where the negative effect of positional error is noticeable only for scenarios with more pronounced positional error (i.e., 40 metres and higher). The SI dropped from 0.67 (unaltered models) to 0.64 (90 – 99 m error). Regardless of modeling technique introducing positional error led to decrease in niche overlap between true and predicted species distribution probability. For example, Spearman's rank correlation dropped from 0.96 to 0.76 for MaxEnt respectively from to 0.6 to 0.34 for GLM (Figure 3.17).

The results also show a clear trend of decreasing model performance as the analysis grain is coarsened compared to the response grain (i.e., from the original resolution at which the virtual species were generated; 5 x 5 m). The largest decrease was observed between the unaltered models (5 m) and the models with the smallest change in the analysis grain (20 m). For example, SI decreased from 0.76 to 0.72 and from 0.74 to 0.67 for MaxEnt and BRT, respectively (Figure 3.17). Further coarsening of the analysis grain resulted in an additional decrease in models' performance; however, the overall decrease in performance between 20 m and 500 m was less than the decrease caused by the initial change in analysis grain (Figure 3.17). The same pattern shows also niche comparison assessed by Spearman's rank correlation (Figure 3.17). Note that the observed trends were independent of the validation metric.

Trade-off between positional error and analysis grain

Finally, and most importantly, our results clearly showed that coarsening the analysis grain cannot compensate for the effect of positional error (Figure 3.18). For each scenario positional error (S1-S6), we can observe that models with an analysis grain coarser than the initial grain (5 m) performed, at best, equally well, but never better than those with initial grain (i.e. response grain). In addition, models with a positional error of 20 - 29 m (S3) and higher perform almost equally well regardless of the analysis grain. This applies to all used performance metrics and Spearman's rank correlation used to assess the species niche overlap (Figure 3.18, Supplementary material Figure A4).

3.4.5. Discussion

In this study, we focused on the trade-off between the analysis grain and positional error in fine-scale SDMs. We simulated virtual species at 5 m resolution, coarsened the analysis grain (5 – 500 m) and introduced positional error (5 – 99 m) to evaluate their individual effects and potential trade-offs between them. Our results showed a negative effect of coarsening the analysis grain on SDMs performance. All modelling techniques were sensitive to the change in analysis grain (see also Guisan et al. 2007 for an analysis of the sensitivity of ten modelling techniques to the change in grain size). Although this could be perceived as a negative, we believe that this is actually a positive characteristic, as it means that these models are sensi-

tive to the use of an (in)appropriate resolution of the analysis grain. Similarly, introducing positional error led to a decrease in the discriminative ability of all modelling methods; yet, and importantly, coarsening the analysis grain did not offset for the effects of positional error.

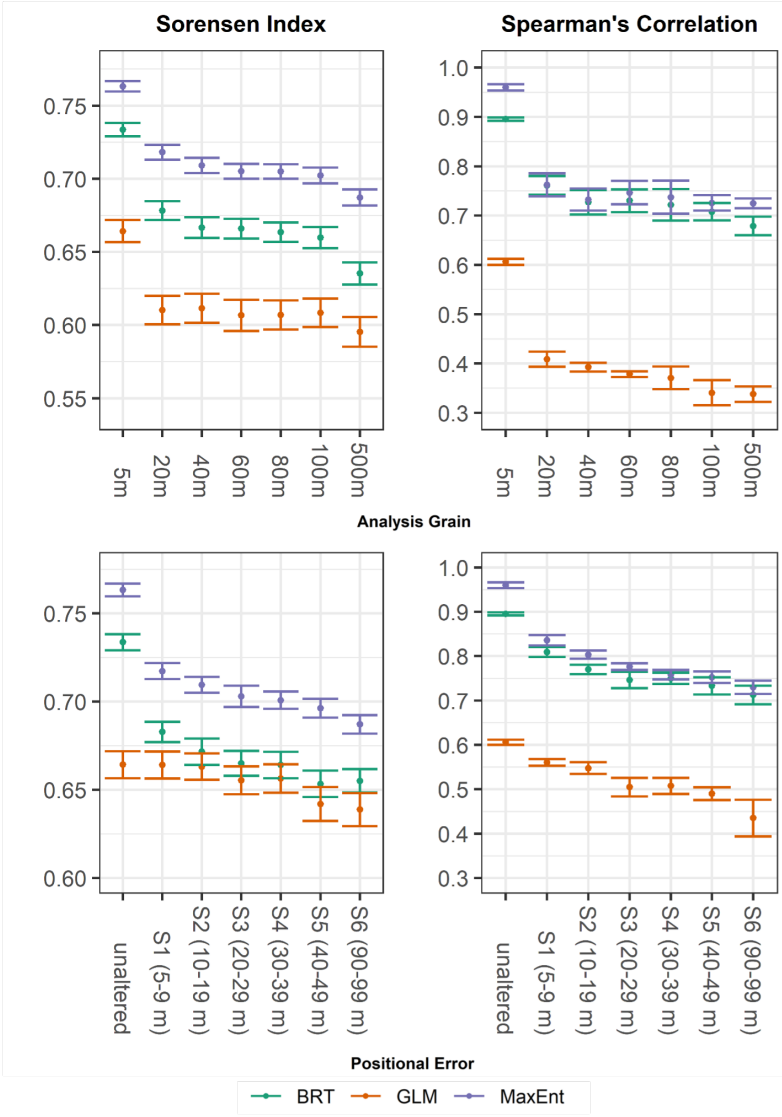


Figure 3.17: Sørensen index and Spearman's rank correlation scores of the different models. The first row shows results for models fitted with different analysis grains. The second row shows results for models fitted with an analysis grain of 5 m, but with positionally shifted species occurrences.

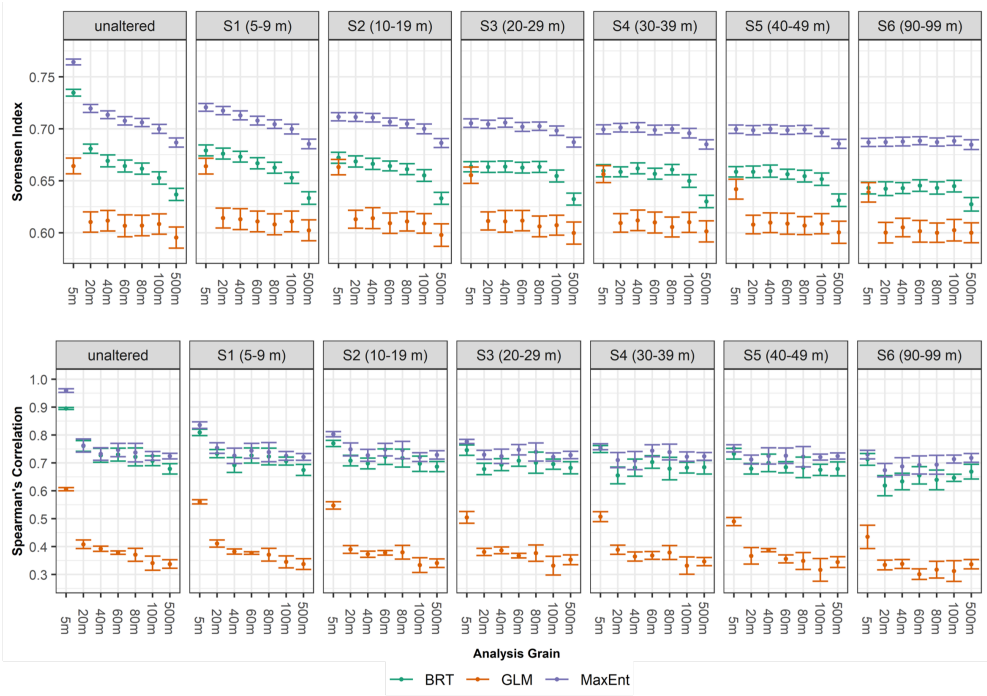


Figure 3.18: Sørensen index and Spearman's rank correlation scores according to different analysis grains and positional error scenarios (unaltered and S1-S6).

The correct choice of the analysis grain is an important part of the overall modelling process and is affected by several other modelling choices. Ideally, the analysis grain is dictated by the species ecology and the objectives of the study, i.e., it must match the response grain (Mertes and Jetz 2018) but it could be also affected by sampling processes of species occurrences (Rahbek 2005, Hurlbert and Jetz 2007, Chase and Knight 2013) and by the spatial extent of the study area. The spatial extent and resolution of the response variable govern what explanatory variables can be expected to act in determining species distribution (Pearson and Dawson 2003). Typically, it is assumed that climate defines the distribution of species at very broad spatial scales (e.g., an extent of a whole continent and resolution of 100 km^2). At successively finer resolutions and at regional extents, topography or biotic interactions may become the most important factors controlling species occurrence, whereas at even finer resolutions, vegetation structure or presence of individual land cover categories (e.g., water bodies) can play a role (Gábor et al. 2022a). However, some studies suggest that biotic interactions may shape species distribution across all spatial extents (Wisz et al. 2013; Alexander et al. 2015). Generally speaking, the importance of environmental factors varies with the adopted resolution and extent of the study, and factors that are important at one resolution and extent can lose their importance at others (Corsi et al. 2000).

There are two typical situations regarding the choice of the analysis grain in species distribution modelling: (i) we know the response grain and have fine-scale data available, or (ii) we do not know the response grain and/or the analysis grain is chosen based on data availability rather than species ecology (Holland et al. 2004, Graf et al. 2005, Lechner et al. 2012, Martin and Fahrig 2012, Stuberand Fontaine 2019, Mertes et al. 2020). The first situation is

represented in this study by the range of analysis grains from 5 m to 100 m, and the second by the 500 m grain. It should be noted that models are regularly built using an even coarser analysis grain than those tested in this study (e.g., 5 km or 10 km when using atlas data; Jetz et al. 2012). However, several studies have already tested the general effect of changing the grain of the response variable on modelling the species distribution in situations where the spatial resolution of the response variable was considerably coarser than the assumed response grain. For example, Seo et al. (2009) examined SDMs dynamics across a 64-fold (1 km to 64 km) change in the grain of the response variable and found that model performance decreased with increasing resolution. Similarly, Kaliontzopoulou et al. (2008) reported decreasing model performance at the 10 km response variable resolution compared to 1 km resolution.

Our results show that compensating position errors by coarsening the analysis grain does not lead to an improvement of the model performance in any of the scenarios investigated (Figure 3.17, A4). This is true even for very coarse analysis, i.e., an analysis grain several orders of magnitude larger than the expected response grain. Therefore, based on our results and the results of the above-mentioned studies, we recommend using an analysis grain as fine as possible (or, in other words, as close to the response grain as possible), even if the available species occurrences suffer from positional error. This is consistent with recent findings by Mertes and Jetz (2018), who showed that coarsening the analysis grain can negatively affect intrinsic fine-scale heterogeneity in environmental variables (i.e., the pattern of spatial autocorrelation inherent in an environmental variable) and lead to variables that strongly influence distribution patterns being discarded simply because of their low explanatory power at such coarsened resolution. On the other hand, this contradicts the widely held assumption that coarsening the analysis grain can compensate for the negative effect of positional errors on model performance (Engler et al. 2004, Moudrý and Šímová 2012, Keil et al. 2014, Vollering et al. 2016, Sillero and Barbosa 2021), but this has never been thoroughly tested. Our results show that above a certain level of positional error (approximately five times higher than the response grain), models perform almost the same regardless of the analysis grain. Therefore, if there is considerable positional error in species occurrence data, users are unlikely to gain anything from making additional efforts to obtain higher resolution data (but see Šímová et al. 2019) unless they also minimize the positional error.

Our findings and recommendations, however, do not mean that negative effects of the positional error can be ignored. On the contrary, the inability to compensate for the positional error by coarsening the analysis grain underscores the importance of careful georeferencing of species occurrence data. Our results show that the largest decrease in model performance occurs in the smallest simulated positional error (i.e., as soon as an error is introduced). This is consistent with previous studies and their conclusions that more accurate georeferencing approaches generally produce better performing SDMs (Lash et al. 2012, Tulowiecki et al. 2015, Zhang et al. 2018, Gábor et al. 2020b). For example, Lash et al. (2012) have shown that using less accurate automated georeferencing methods is problematic in mapping the occurrence of monkeypox and modelling its transmission risk in Africa. The same limitations have been reported by Tulowiecki et al. (2015) for pre-settlement land survey records in North America that are useful for modelling the past distribution of tree species (e.g., Tulowiecki 2020). On the other hand, it is fair to point out that Graham et al. (2008) concluded that SDMs are generally robust to positional errors. Similarly, Fernandez et al. (2009) concluded that while the models are somewhat sensitive to positional error, this sensitivity is considerably less than the sensitivity to the modelling method.

However, accurate georeferencing is an extremely time-consuming and labour-intensive process. In particular, georeferencing historical records can be challenging because in some parts of the world it is difficult to find suitable reference data with which to match place names. Guidelines for georeferencing exists (Wieczorek et al. 2004), and some heuristic approaches have been proposed to improve models created with poorly georeferenced data. These methods are applicable depending on the source of positional error and the available auxiliary data. For example, Hefley et al. (2014) used regression calibration to reduce the bias in coefficient estimates caused by the positional error. However, this approach requires that at least part of the data has locations recorded without error. Recently, Zhang et al. (2018) proposed a different approach to mitigate positional error in fine analysis grains (e.g., errors of tens of meters caused, for example, by the difference in position of the species and the observer). They narrowed down possible locations of species occurrences using auxiliary data such as the presence of habitat preferred by the species (e.g., forest), the assumed minimum and maximum distance (i.e., minimum distance the species keeps from the observer and the maximum distance at which the observer can see the species), and the observer's field of view (i.e., visibility analysis using a digital terrain model; Lagner et al. 2018).

We intentionally developed our models with fine-scale environmental data that are increasingly adopted for SDMs (e.g., Mitchell et al. 2017, de Vries et al. 2021, Guillaume et al. 2021). Although so far, such data are typically used in models developed to assess species-environment relationships at a landscape scale, it has been highlighted that they can be crucial for understanding species distributions at global scales (Lembrechts et al. 2019a, Lembrechts et al. 2019b, Zellweger et al. 2019, Stark and Fridley 2022). Moreover, such fine-scale environmental data tend to be more heterogeneous, and hence species occurrences might easier end up in unsuitable environment, which can negatively affect SDMs (see Naimi et al. 2011, 2014). Therefore, understanding the interaction of analysis grain and positional error at fine-grain is crucial for future development of SDMs for conservation and climate change studies.

It is important to note that the effect of analysis grain and positional error is dependent on the magnitude of the potential change of the analysis grain (not the grain itself) and similarly the effect of positional error depends on the ratio between the magnitude of the positional error and the analysis grain. In addition, the magnitude of the effect will be affected by other characteristics. For example, it has been shown that the magnitude of the negative effect of positional error is related to species characteristics, such as niche (Visscher 2006; Tulowiecki et al. 2015; Gabor et al. 2020b) and heterogeneity in environmental variables (i.e., spatial autocorrelation; Naimi et al. 2011, 2014). For instance, models for species with relatively wide niche breadth and a region dominated by highly autocorrelated environmental variables or a single habitat will be relatively unaffected by positional error. On the contrary, the models for a region with abrupt changes (e.g., fragmented habitats) and for species with narrow niche breadth will be negatively affected with positional error in species data (see Visscher 2006 Naimi et al. 2011, 2014). Therefore, our conclusions are also applicable into analysis using relatively coarse analysis grain, especially for SDMs developed for a region with abrupt changes in environment (e.g., fragmented habitats) and for species with narrow niche breadth (see Naimi et al. 2011, 2014, Gábor et al. 2020a,b).

In this study, we examined how, in a species distribution modelling context, analysis grain and positional error in species occurrences interact. Our particular objective was to answer the question of whether the analysis grain is best kept close to the response grain or whether it should instead be coarsened to minimize the negative effects of positional errors in species

occurrences on model performance, as suggested by several authors. We showed that a coarsened analysis grain is not able to compensate for the effects of positional errors. Thus, for data with unknown positional accuracy, we recommend keeping the analysis grain as close as possible to the response grain (i.e., usually as fine as possible) rather than coarsening the variables. We highlight that positional error in species occurrence cannot be overlooked and that great attention needs to be paid to the measurement and georeferencing techniques used to minimize positional error.

Key Findings

Since a comprehensive discussion and conclusions have already been provided in the preceding Chapter 3, they will not be reiterated here. Instead, this section connects the individual studies presented, emphasizes key findings, offers suggestions for future research, and provides a closing statement.

This dissertation addressed the challenges associated with predicting species distribution, particularly regarding the input data quality (species records and environmental predictors). The research focused on three main objectives. Firstly, my co-authors and I introduced a novel method for incorporating landcover data into species SDMs, expanding the range of variables considered and enhancing the models' predictive accuracy (Chapters 3.1, 3.2). Secondly, we investigated the influence of positional uncertainty in species data on the ecological interpretability of models (Chapter 3.3) and assessed our ability to compensate for these data limitations (i.e., by coarsening the resolution of environmental predictors to maximum positional inaccuracy in species data; Chapter 3.4). Finally we, together with my amazing colleagues, assessed the role of spatial grain for the usability of binary variables, and for the effect of positional uncertainty. There is no standalone chapter for this topic because it has been thoroughly covered throughout the various chapters.

In Chapter 3.1, our collaborative study aimed to examine if the binary information of presence or absence of a habitat can drive species distribution. We suggested that the amount of habitat (continuous land cover predictors) within a spatial unit might be irrelevant, and what matters for some species at some scales is that the habitat is simply there. We demonstrated that there might be a threshold of habitat percentage below which the species is unlikely to occur, and above which the species will persist. This threshold concept has been both theoretically predicted (Andrén 1994, Fahrig 2001) and empirically documented for bird species (Melo et al. 2018). Our results suggested that models using binary land cover predictors performed better for specialist species and fine-grained environmental variables. However, as we pointed out, further research is needed to validate our conclusions.

Thus, in Chapter 3.2, we investigated if the above-introduced hypothesis could be applied to species that rely on prevalent habitats. Beyond this, we determined which method of summarizing environmental features produces the best-performing models and examined how treatment of features influences range size estimates. More importantly, we tested how the effect of landcover summarization (continuous or binary) is moderated by spatial grain size, long known to influence model performance and estimate range size. Although the results indicated that the model performance was not significantly affected by the type of land cover predictors, using binary variables greatly impacted the models' ability to detect water bodies as the most important land cover predictor. For example, flooded tree cover or shrubland flooded tree cover often occurs near water cover and may predict a species occurrence when the species is also observed from these habitats. More importantly, our results suggest that grain size can impact the applicability of binary land cover variables for species' models specializing in prevalent habitats. The subtle difference in the importance of water bodies between models fitted with continual and binary variables, derived using a 1% threshold, indicates that the hypothesis proposed in Chapter 3.2 may be particularly useful at finer grain sizes. The fact that habitat variables can, depending on the commonness of the habitat or the used grain size, perform best as a binary or a continuous predictor has relevance beyond simple SDMs. After all, information on the probability of species' presence is sought in many fields, from epidemiology to metacommunity ecology. Our findings indicate that using an

inappropriate representation of the species-habitat relationship can lead to underestimating the importance of niche processes. This highlights the broader implications of our research in illuminating the complexities of ecological dynamics and the need to carefully consider habitat variables in diverse fields of study.

During the literature review conducted for Chapter 3.3, it became evident that previous studies addressing positional uncertainty in species occurrences consistently highlighted the influence of positional error on the predictive performance of SDMs (e.g., Graham et al. 2008, Johnson and Gillingham 2008, Fernandez et al. 2009, Naimi et al. 2014, Hefley et al. 2014, Tulowiecki et al. 2015, Mitchell et al. 2017, Soultan and Safi 2017, Fernandes et al. 2019). However, there remained a significant gap in understanding the effects of positional uncertainty on parameter estimation in species-environment relationship inference. Therefore, with my amazing co-authors, I aimed to investigate the influence of positional uncertainty on the model's parameter estimation. Specifically, we examined how positional error affects variable importance and the shape of response curves. We hypothesized that increasing positional uncertainty would decrease the models' predictive performance, imprecise variable importance estimation, and distorted response curves. We anticipated these effects would be more pronounced for species with narrow niches and heterogeneous variables. Surprisingly, our findings demonstrated that positionally inaccurate species records can still provide valuable insights into the relative importance of environmental variables and the shape of species responses. Therefore, removing positionally inaccurate records from the dataset may not always be necessary. This finding is particularly significant because discarding occurrences with positional uncertainty can limit our ability to estimate species range sizes and result in overestimations of exposure to climate change (Smith et al. 2023).

Furthermore, such conclusions hold promise for integrating species data from citizen science initiatives into SDMs, as these datasets often exhibit higher levels of positional uncertainty. This enables local communities to participate in monitoring and conserving species and their habitats. These practical implications can potentially enhance the performance of SDMs and advance their application in diverse fields such as biogeography, community ecology, macroecology, and ecological conservation. However, it is crucial to emphasize the significance of not overlooking positional uncertainty in species records. Attention also must be given to measurement and georeferencing techniques to minimize positional uncertainty.

Another significant outcome from conventional wisdom in the presented research is rejecting the prevailing assumption that environmental predictors should be coarsened to accommodate the maximum positional uncertainty of species records (Chapter 3.4). Positional uncertainty is related to almost all species records and can range from a couple of meters up to tens of kilometers, and prior studies concluded that it has a negative effect on SDMs' performance (see, for example, Lash et al. 2012, Tulowiecki et al. 2015, Zhang et al. 2018, Gábor et al. 2020a). Coarsening of the analysis grain has been suggested to compensate for positional uncertainty (see Moudrý and Šímová 2012). However, this widely accepted approach has never been thoroughly tested, and I wondered for a long time if this approach was flawed in some way. I had two main concerns related to prior studies. It is a well-known fact that coarsening the analysis grain can negatively affect the performance of SDMs (e.g., Guisan et al. 2007, Kaliontzopoulou et al. 2008, Seo et al. 2009). Besides, Mertes and Jetz (2018) showed that coarsening the analysis grain can negatively affect intrinsic fine-scale heterogeneity in environmental variables. With the increasing use of fine-grain environmental predictors, it became increasingly imperative to test this assumption. Our results showed that the performance of all models was negatively affected by the positional uncertainty in species records

and by coarsening the analysis grain.

Additionally, results showed that coarsening the analysis grain cannot compensate for the effect of positional uncertainty. Therefore, we recommend using an analysis grain as fine as possible (i.e., as close to the response grain as possible), even if the available species records suffer from positional uncertainty. By recognizing the potential value of incorporating fine-grain environmental data and questioning the need for coarsening, we can refine our modeling approaches, enhance accuracy, and ultimately contribute to more effective conservation strategies and management decisions (see, for example, Lembrechts et al. 2019a, Lembrechts et al. 2019b, Stark and Fridley 2022).

Future Research

I firmly believe that future studies can explore more thrilling questions directly related to the research presented. Firstly, there is a compelling need to investigate the use of binary variables in predicting species distribution. Additionally, it is crucial to address the challenge of mitigating the negative influence of positional uncertainty in species records on SDMs' performance. A complex interplay of various factors determines the magnitude of the negative influence of positional uncertainty in species records. Chapters 3.3 and 3.4 of this study have shed light on the importance of spatial scale and spatial autocorrelation in environmental predictors. Other factors like sample size or species traits also play a role (Mitchell et al. 2017; Gábor et al. 2020b; Figure 5.1). However, previous research has primarily focused on combining one or two factors. This limits our comprehensive understanding of the impact of positional uncertainty on SDMs. To achieve a more thorough understanding, future studies should explore the effects of multiple factors simultaneously.

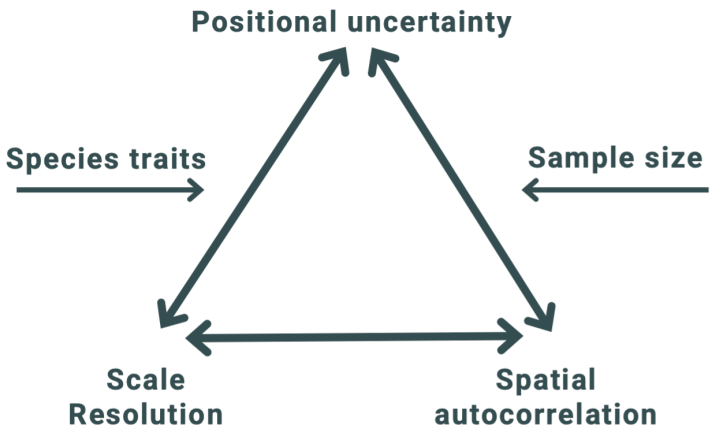


Figure 5.1: A complex interplay of various factors determines the magnitude of the negative influence of positional uncertainty in species records, specifically by spatial scale and spatial autocorrelation in environmental predictors. Factors like sample size or species traits can also play an important role. To achieve a more thorough understanding, future studies should explore the effects of multiple factors simultaneously.

Furthermore, there is still a lack of an established tool to effectively overcome the negative influence of positional uncertainty in species records. While methods for mitigating positional error do exist, their utilization is often constrained. Typically, they necessitate awareness of the error's magnitude, and a portion of the dataset must be recorded without positional error (Hefley et al., 2014; Velásquez-Tibatá et al., 2016; Zhang et al., 2018). In addition, its limited adoption and usage within the modeling community is evident, as reflected by the low number of citations since their publication (only ninety-one references since the publication of those papers!). Therefore, future studies must focus on developing a robust and accessible tool that effectively tackles the challenge of positional uncertainty while remaining user-friendly and widely applicable. A possible and relatively straightforward solution could involve calculating an average of environmental variables considering positional uncertainty (e.g., generate a buffer around known positional uncertainty of species occurrences

and derive an average value within that buffer). While this approach may be direct and user-friendly, its applicability could be constrained. For instance, species inhabiting mountain peaks would consistently yield lower altitude values. Therefore, if such an approach is developed, one must be mindful of these limitations.

Besides this, further important questions arise as we contemplate the future of predictive modeling in ecology. How can we navigate the delicate balance between embracing uncertainty and striving for accuracy in our models? Given the increasing availability of big data and advancements in remote sensing technologies, how can we harness these resources to improve the accuracy and resolution of our predictive models, particularly in data-sparse regions and for poorly sampled species? How can we address the potential biases and limitations in species occurrence data, especially in the context of citizen science and crowdsourced data collection, to ensure the robustness and reliability of our models? What are the ethical implications of using predictive models in decision-making processes, such as conservation prioritization? How can we ensure that SDMs are transparent and inclusive and contribute to equitable and sustainable outcomes? How can we effectively communicate the uncertainties associated with SDMs to stakeholders, policymakers, and the general public to foster informed decision-making and public engagement in conservation efforts?

By addressing these questions and embracing the interdisciplinary nature of ecological modeling, we can strive towards more accurate, reliable, and actionable predictions, ultimately contributing to the conservation and sustainable management of our planet's rich biodiversity.

Afterword

As I reflect upon the future of research and the intriguing questions that lie ahead, I am filled with both optimism and a tinge of sadness. I have no doubts that the collective brilliance of exceptional minds will eventually unravel the mysteries we seek to solve. In academia, there are remarkable individuals whose intellectual prowess and ingenuity will undoubtedly lead to groundbreaking discoveries. However, as I write these words, I find myself stepping away from academia for a while.

Over a decade immersed in the academic world has left me yearning for a respite, a chance to take a breath and explore different paths. Although my passion for scientific inquiry burns brightly, I recognized the need to pause and recharge. Though marked with a touch of melancholy, this decision is not a permanent farewell but rather a temporary divergence from the academic journey.

I eagerly anticipate the day when I can return to the vibrant halls of academia, where ideas flourish, collaborations thrive, and pursuing knowledge is an exhilarating adventure. I envision a future where I can once again contribute to the collective quest for understanding and make meaningful contributions to the field that has captivated me for so long.

So, it is not goodbye, but rather a ”see you soon” on this ever-evolving journey. I am grateful for the experiences, the lessons learned, and the connections forged during my time in academia. I am thankful for the opportunity to embark on this research journey, supported by mentors, colleagues, and the invaluable contributions of fellow researchers in the field. With a renewed spirit and a fresh perspective, I eagerly await the day when I can resume my scientific endeavors and again be part of the extraordinary community of researchers shaping the future of ecological modeling and conservation.

Until then, I embarked on a new chapter, fueled by curiosity, open to new experiences, and eager to explore diverse realms outside academia. I carry with me the knowledge that the future holds remarkable possibilities, and perhaps, one day, our paths will cross again in the pursuit of knowledge and the noble endeavor of safeguarding our planet’s precious ecosystems.

References

Agresti, A. (2003) Categorical data analysis. Vol. 482. John Wiley and Sons.

Ahmad Suhaimi, S. S., Blair, G. S., and Jarvis, S. G. (2021). Integrated species distribution models: A comparison of approaches under different data quality scenarios. Diversity and Distributions, 27(6), 1066-1075.

Alexander, J. M., Diez, J. M., and Levine, J. M. (2015). Novel competitors shape species' responses to climate change. Nature, 525(7570), 515-518.

Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of applied ecology, 43(6), 1223-1232.

Altizer, S., Ostfeld, R. S., Johnson, P. T., Kutz, S., and Harvell, C. D. (2013). Climate change and infectious diseases: from evidence to a predictive framework. Science, 341(6145), 514-519.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific data, 5(1), 1-15.

Ancillotto, L., Bosso, L., Smeraldo, S., Mori, E., Mazza, G., Herkt, M., ... and Russo, D. (2020). An African bat in Europe, Plecotus gaisleri: Biogeographic and ecological insights from molecular taxonomy and Species Distribution Models. Ecology and evolution, 10(12), 5785-5800.

Anderson, R. P., Lew, D., and Peterson, A. T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological modelling, 162(3), 211-232.

Anderson, R. P., Araújo, M., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., and Soberón, J. (2016). Final report of the task group on GBIF data fitness for use in distribution modelling. Global Biodiversity Information Facility.

Andren, H. (1994). Effects of habitat fragmentation on birds and mammals in landscapes with different proportions of suitable habitat: a review. Oikos, 355-366.

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

Araújo, M. B., and New, M. (2007). Ensemble forecasting of species distributions. Trends in ecology and evolution, 22(1), 42-47.

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., ... and Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. Science Advances, 5(1), eaat4858.

Arenas-Castro, S., Regos, A., Martins, I., Honrado, J., and Alonso, J. (2022). Effects of input data sources on species distribution model predictions across species with different distributional ranges. Journal of Biogeography, 49(7), 1299-1312.

Association, A.B. (2008). American Birding Association Checklist: Birds of the Continental United States and Canada.

Aubry, K. B., Raley, C. M., and McKelvey, K. S. (2017). The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. PLoS One, 12(6), e0179152.

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological modelling, 200(1-2), 1-19.

Austin, M. P., Belbin, L., Meyers, J. A. A., Doherty, M. D., and Luoto, M. (2006). Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. ecological modelling, 199(2), 197-216.

Austin, M. P., and Van Niel, K. P. (2011). Improving species distribution models for climate change studies: variable selection and scale.

Bakx, T. R., Koma, Z., Seijmonsbergen, A. C., and Kissling, W. D. (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.

Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., and Beck, J. (2017). Putting insects on the map: Near-global variation in sphingid moth richness along spatial and environmental gradients. Ecography, 40(6), 698-708.

Barbosa, M.A., Real, R., Muñoz, A. R., and Brown, J. A. (2013). New measures for assessing model equilibrium and prediction mismatch in species distribution models. Diversity and Distributions, 19(10), 1333-1338.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., ... and Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived?. Nature, 471(7336), 51-57.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., ... and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological modelling, 222(11), 1810-1819.

Bates, J. M., and Granger, C. W. (1969). The combination of forecasts. Journal of the operational research society, 20(4), 451-468.

Bayraktarov, E., Ehmke, G., O'connor, J., Burns, E. L., Nguyen, H. A., McRae, L., ... and Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge?. Frontiers in Ecology and Evolution, 6, 239.

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., ... and Sperandii, M. G. (2022). Effect of sampling strategies on the response curves estimated by plant species distribution models.

Bazzichetto, M., Malavasi, M., Barták, V., Acosta, A. T. R., Moudrý, V., and Carranza, M. L. (2018). Modeling plant invasion on Mediterranean coastal landscapes: An integrative approach using remotely sensed data. Landscape and Urban Planning, 171, 98-106.

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., ... and Sperandii, M. G. (2023). Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. Global Ecology and Biogeography.

Belsley, D. A. (1991). A guide to using the collinearity diagnostics. Computer Science in Economics and Management, 4(1), 33-50.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., ... and Bivand, M. R. (2015). Package 'rgdal'. Bindings for the Geospatial Data Abstraction Library. Available online: https://cran. r-project. org/web/packages/rgdal/index. html (accessed on October 15th 2017), 172.

Booth, T. H. (2018). Species distribution modelling tools and databases to assist managing forests under climate change. Forest Ecology and Management, 430, 196-203.

Boral, D., and Moktan, S. (2021). Predictive distribution modeling of Swertia bimaculata in Darjeeling-Sikkim Eastern Himalaya using MaxEnt: current and future scenarios. Ecological Processes, 10(1), 1-16.

Boyd, R., Harvey, M., Roy, D., Barber, T., Haysom, K., Macadam, C., ... and Preston, C. (2022). Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. EcoEvoRxiv.

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.

Bradter, U., Kunin, W. E., Altringham, J. D., Thom, T. J., and Benton, T. G. (2013). Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods in Ecology and Evolution, 4(2), 167-174.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. Briscoe, N. J., Morris, S. D., Mathewson, P. D., Buckley, L. B., Jusup, M., Levy, O., ... and Kearney, M. R. (2023). Mechanistic forecasts of species responses to climate change: the promise of biophysical ecology. Global Change Biology, 29(6), 1451-1470.

Brown, J. H., and Kodric-Brown, A. (1977). Turnover rates in insular biogeography: effect of immigration on extinction. Ecology, 58(2), 445-449.

Buckley, R. C., and Knedlhans, S. B. (1986). Beachcomber biogeography: interception of dispersing propagules by islands. Journal of biogeography, 13(1), 69-70.

Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romanach, S. S., ... and Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. Diversity and distributions, 21(1), 23-35.

Butchart, S. H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P., Almond, R. E., ... and Watson, R. (2010). Global biodiversity: indicators of recent declines. Science, 328(5982), 1164-1168.

Cadotte, M. W., Carscadden, K., and Mirotchnick, N. (2011). Beyond species: functional diversity and the maintenance of ecological processes and services. Journal of applied ecology, 48(5), 1079-1087.

Cánibe, M., Titeux, N., Domínguez, J., and Regos, A. (2022). Assessing the uncertainty arising from standard land-cover mapping procedures when modelling species distributions. Diversity and Distributions, 28(4), 636-648.

Carlson, C. J., Albery, G. F., Merow, C., Trisos, C. H., Zipfel, C. M., Eskew, E. A., ... and Bansal, S. (2022). Climate change increases cross-species viral transmission risk. Nature, 1-1.

Casanelles-Abella, J., Müller, S., Keller, A., Aleixo, C., Alós Orti, M., Chiron, F., ... and Moretti, M. (2022). How wild bees find a way in European cities: Pollen metabarcoding unravels multiple feeding strategies and their effects on distribution patterns in four wild bee species. Journal of Applied Ecology, 59(2), 457-470.

Chase, J. M., and Knight, T. M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. Ecology letters, 16, 17-26.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A (Statistics in Society), 158(3), 419-444.

Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., ... and Zimmermann, N. E. (2021). Influence of climate, soil, and land cover on plant species distribution in the European Alps. Ecological monographs, 91(2), e01433.

Chauvier, Y., Descombes, P., Guéguen, M., Boulangeat, L., Thuiller, W., and Zimmermann, N. E. (2022). Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity. Ecography, 2022(10), e05973.

Cheng, Y., Tjaden, N. B., Jaeschke, A., Thomas, S. M., and Beierkuhnlein, C. (2021). Using centroids of spatial units in ecological niche modelling: Effects on model performance in the context of environmental data grain size. Global Ecology and Biogeography, 30(3), 611-621.

Clements, J.F., 2007. Clements checklist of birds of the world. Comstock Pub. Associates/Cornell University Press.

Cogliati, M., Arikan-Akdagli, S., Barac, A., Bostanaru, A. C., Brito, S., Cerikcioglu, N., ... and Brandão, J. (2023). Environmental and bioclimatic factors influencing yeasts and molds distribution along European shores. Science of the Total Environment, 859, 160132.

Cohen, J. M., Civitello, D. J., Brace, A. J., Feichtinger, E. M., Ortega, C. N., Richardson, J. C., ... and Rohr, J. R. (2016). Spatial scale modulates the strength of ecological processes driving disease distributions. Proceedings of the National Academy of Sciences, 113(24), E3359-E3364.

Collart, F., and Guisan, A. (2023). Small to train, small to test: Dealing with low sample size in model evaluation. Ecological Informatics, 75, 102106.

Conrad, O., et al. (2015): System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007.

Coops, N. C., Duffe, J., and Koot, C. (2010). Assessing the utility of lidar remote sensing technology to identify mule deer winter habitat. Canadian Journal of Remote Sensing, 36(2), 81-88.

Coppée, T., Paquet, J. Y., Titeux, N., and Dufrêne, M. (2022). Temporal transferability of species abundance models to study the changes of breeding bird species based on land cover changes. Ecological Modelling, 473, 110136.

Cord, A. F., Meentemeyer, R. K., Leitão, P. J., and Václavík, T. (2013). Modelling species distributions with remote sensing data: bridging disciplinary perspectives. Journal of Biogeography, 40(12), 2226-2227.

Cord, A. F., Klein, D., Mora, F., and Dech, S. (2014). Comparing the suitability of classified land cover data and remote sensing variables for modeling distribution patterns of plants. Ecological Modelling, 272, 129-140.

Corsi, F., De Leeuw, J., and Skidmore, A. (2000). Modeling species distribution with GIS. Research techniques in animal ecology, 389-434.

Costa, H., Foody, G. M., Jiménez, S., and Silva, L. (2015). Impacts of species misidentification on species distribution modeling with presence-only data. ISPRS International Journal of Geo-Information, 4(4), 2496-2518.

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics. Ecology letters, 8(11), 1175-1182.

Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S., and Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. Nature ecology and evolution, 2(3), 420-426.

Curd, A., Chevalier, M., Vasquez, M., Boyé, A., Firth, L. B., Marzloff, M. P., ... and Dubois, S. F. (2023). Applying landscape metrics to species distribution model predictions to characterize internal range structure and associated changes. Global Change Biology, 29(3), 631-647.

D'Amen, M., and Azzurro, E. (2020). Lessepsian fish invasion in Mediterranean marine protected areas: a risk assessment under climate change scenarios. ICES Journal of Marine Science, 77(1), 388-397.

Davies, A. B., and Asner, G. P. (2014). Advances in animal ecology from 3D-LiDAR ecosystem mapping. Trends in ecology and evolution, 29(12), 681-691.

de Brooke, M. (2000). Why museums matter. Trends in Ecology and Evolution, 15(4), 136-137.

De Marco, P., and Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PloS one, 13(9), e0202403.

de Vries, J. P. R., Koma, Z., WallisDeVries, M. F., and Kissling, W. D. (2021). Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning. Diversity and Distributions.

Didan, K. (2015). MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC.

Domisch, S., Jähnig, S. C., Simaika, J. P., Kuemmerlen, M., and Stoll, S. (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, 45-61.

Dormann, F. C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., ... and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30(5), 609-628.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., ... and Kraan, C. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. Global ecology and biogeography, 27(9), 1004-1016.

Drescher, M., Perera, A. H., Johnson, C. J., Buse, L. J., Drew, C. A., and Burgman, M. A. (2013). Toward rigorous use of expert knowledge in ecological research. Ecosphere 4: 1–26.

Drew, C. A., and Perera, A. H. (2010). Expert knowledge as a basis for landscape ecological predictive models. In Predictive species and habitat modeling in landscape ecology: Concepts and applications (pp. 229-248). New York, NY: Springer New York.

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Montsarrat, S., Denelle, P., … and Luque, S. (2021). Assessing the effect of sample bias correction in species distribution models. arXiv. arXiv preprint arXiv:2103.07107.

Dungan, J. (1999) Conditional simulation: an alternative to estimation for achieving mapping objectives. Spatial statistics for remote sensing (ed. by A. Stein, F. Meer and B. Gorte), pp. 135–152. Springer, Dordrecht.

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M. J., … and Rosenberg, M. (2002). A balanced view of scale in spatial statistical analysis. Ecography, 25(5), 626-640.

Dvorský, M., Macek, M., Kopecký, M., Wild, J., and Doležal, J. (2017). Niche asymmetry of vascular plants increases with elevation. Journal of Biogeography, 44(6), 1418-1425.

Eduardo, A. A., Liparini, A., Martinez, P. A., Gouveia, S. F., and Riul, P. (2022). Assessing multi-temporal calibration for species distribution models. Ecological Informatics, 71, 101787.

El-Gabbas, A., Van Opzeeland, I., Burkhardt, E., and Boebel, O. (2021). Static species distribution models in the marine realm: The case of baleen whales in the Southern Ocean. Diversity and Distributions.

Elith, J., Ferrier, S., Huettmann, F., and Leathwick, J. (2005). The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. Ecological modelling, 186(3), 280-289.

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, 677-697.

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. Methods in ecology and evolution, 1(4), 330-342.

Ellis-Soto, D., Merow, C., Amatulli, G., Parra, J. L., and Jetz, W. (2021). Continental-scale 1 km hummingbird diversity derived from fusing point records with lateral and elevational expert information. Ecography, 44(4), 640-652.

Engler, R., Guisan, A., and Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of applied ecology, 41(2), 263-274.

ESA. Land Cover CCI Product User Guide Version 2. Technical Report., 2017

Evans, J. S., Murphy, M. A., Holden, Z. A., and Cushman, S. A. (2011). Modeling species distribution and change using random forest. Predictive species and habitat modeling in landscape ecology: Concepts and applications, 139-159.

Fahrig, L. (2001). How much habitat is enough?. Biological conservation, 100(1), 65-74.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., … and Alsdorf, D. (2007). The shuttle radar topography mission. Reviews of geophysics, 45(2).

Farrell, A., Wang, G., Rush, S. A., Martin, J. A., Belant, J. L., Butler, A. B., and Godwin, D. (2019). Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. Ecology and evolution, 9(10), 5938-5949.

changes and flows in European landscapes 1990–2000 using CORINE land cover data. Applied geography, 30(1), 19-35.

Fern, R. R., Morrison, M. L., Grant, W. E., Wang, H., and Campbell, T. A. (2020). Modeling the influence of livestock grazing pressure on grassland bird distributions. Ecological Processes, 9(1), 1-11.

Fernandes, R. F., Scherrer, D., and Guisan, A. (2019). Effects of simulated observation errors on the performance of species distribution models. Diversity and Distributions, 25(3), 400-413.

Fernandez, M., Blum, S., Reichle, S., Guo, Q., Holzman, B., and Hamilton, H. (2009). Locality uncertainty and the differential performance of four common niche-based modeling techniques. Biodiversity Informatics, 6.

Ferrier, S., Jetz, W., and Scharlemann, J. (2017). Biodiversity modelling as part of an observation system. In The GEO Handbook on Biodiversity Observation Networks (pp. 239-257). Springer, Cham.

Festa, F., Ancillotto, L., Santini, L., Pacifici, M., Rocha, R., Toshkova, N., ... and Razgour, O. (2023). Bat responses to climate change: A systematic review. Biological Reviews, 98(1), 19-33.

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.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., ... and Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications, 20(8), 2131-2147.

Fink, D., Auer, T., Ruiz-Gutierrez, V., Hochachka, W. M., Johnston, A., La Sorte, F. A., and Kelling, S. (2018). Modeling avian full annual cycle distribution and population trends with citizen science data. bioRxiv 251868.

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., and Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. Ecological Applications, 30(3), e02056.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in ecology and evolution, 6(4), 424-438.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (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), e97122.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P. Y., Danielsen, F., ... and Haklay, M. (2022). Citizen science in environmental and ecological sciences. Nature Reviews Methods Primers, 2(1), 64.

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

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. Journal of Biogeography.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals of statistics, 28(2), 337-407.

Gábor, L., Moudrý, V., Barták, V., and Lecours, V. (2020a). How do species and data characteristics affect species distribution models and when to use environmental filtering?. International Journal of Geographical Information Science, 34(8), 1567-1584.

Gábor, L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M., ... and Václavík, T. (2020b). The effect of positional error on fine scale species distribution models increases for specialist species. Ecography, 43(2), 256-269.

Gábor, L. et al. (2022), Positional errors in species distribution modelling are not overcome by the coarser grains of analysis, Dryad, Dataset, https://doi.org/10.5061/dryad.79cnp5hx3.

Gábor, L., Šímová, P., Keil, P., Zarzo-Arias, A., Marsh, C. J., Rocchini, D., ... and Moudrý, V. (2022a). Habitats as predictors in species distribution models: Shall we use continuous or binary data?. Ecography, e06022.

Gábor, L., Jetz, W., Lu, M., Rocchini, D., Cord, A., Malavasi, M., ... and Moudrý, V. (2022b). Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. Methods in Ecology and Evolution, 13(10), 2289-2302.

Gábor, L., Jetz, W., Zarzo-Arias, A., Winner, K., Yanco, S., Pinkert, S., ... and Moudrý, V. (2023). Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable. Ecography, e06358.

Gatto, A., Quintana, F., and Yorio, P. (2008). Feeding behavior and habitat use in a waterbird assemblage at a marine wetland in coastal Patagonia, Argentina. Waterbirds, 31(3), 463-471.

Gotelli, N. J., and Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. Journal of Biogeography, 42(9), 1577-1585.

Graf, R. F., Bollmann, K., Suter, W., and Bugmann, H. (2005). The importance of spatial scale in habitat models: capercaillie in the Swiss Alps. Landscape Ecology, 20(6), 703-717.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., and Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. Trends in ecology and evolution, 19(9), 497-503.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., and NCEAS Predicting Species Distributions Working Group. (2008). The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45(1), 239-247.

Graham, L. J., Spake, R., Gillings, S., Watts, K., and Eigenbrod, F. (2019). Incorporating fine-scale environmental heterogeneity into broad-extent models. Methods in ecology and evolution, 10(6), 767-778.

Grimmett, L., Whitsed, R., and Horta, A. (2021). Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes. Ecography, 44(5), 753-765.

Guélat, J., and Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution, 9(6), 1614-1625.

Gueta, T., and Carmel, Y. (2016). Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. Ecological Informatics, 34, 139-145.

Guillaume, A. S., Leempoel, K., Rochat, E., Rogivue, A., Kasser, M., Gugerli, F., ... and Joost, S. (2021). Multiscale very high resolution topographic models in Alpine ecology: Pros and cons of airborne LiDAR and Drone-based stereo-photogrammetry technologies. Remote sensing, 13(8), 1588.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., ... and Wintle, B. A. (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., and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. Ecological modelling, 135(2-3), 147-186.

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., and NCEAS Species Distribution Modelling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. Diversity and distributions, 13(3), 332-340.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., ... and Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. Ecology letters, 16(12), 1424-1435.

Haesen, S., Lembrechts, J. J., De Frenne, P., Lenoir, J., Aalto, J., Ashcroft, M. B., ... and Van Meerbeek, K. (2021). ForestTemp–Sub-canopy microclimate temperatures of European forests. Global Change Biology, 27(23), 6307-6319.

Hagemeijer, W. J., and Blair, M. J. (1997). The EBCC atlas of European breeding birds. Poyser, London, 479.

Hairston, N. G. (1959). Species abundance and community organization. Ecology, 40(3), 404-416.

Hallman, T. A., and Robinson, W. D. (2020). Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. Diversity and Distributions, 26(3), 315-328.

Halstead, K. E., Alexander, J. D., Hadley, A. S., Stephens, J. L., Yang, Z., and Betts, M. G. (2019). Using a species-centered approach to predict bird community responses to habitat fragmentation. Landscape Ecology, 34(8), 1919-1935.

Hampton, S. E., Anderson, S. S., Bagby, S. C., Gries, C., Han, X., Hart, E. M., ... and Zimmerman, N. (2015). The Tao of open science for ecology. Ecosphere, 6(7), 1-13.

Hanski, I. (1999). Metapopulation ecology. Oxford University Press.

Hardin, G. (1960). The competitive exclusion principle. science, 131(3409), 1292-1297.

Hargreaves, A. L., Samis, K. E., and Eckert, C. G. (2014). Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. The American Naturalist, 183(2), 157-173.

Hefley, T. J., Baasch, D. M., Tyre, A. J., and Blankenship, E. E. (2014). Correction of location errors for presence-only species distribution models. Methods in Ecology and Evolution, 5(3), 207-214.

Heikkinen, R. K., Marmion, M., and Luoto, M. (2012). Does the interpolation accuracy of species distribution models come at the expense of transferability?. Ecography, 35(3), 276-288.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology: A Journal of the Royal Meteorological Society, 25(15), 1965-1978.

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., ... and Hijmans, M. R. J. (2015). Package 'raster'. R package, 734, 473.

Holder, A. M., Markarian, A., Doyle, J. M., and Olson, J. R. (2020). Predicting geographic distributions of fishes in remote stream networks using maximum entropy modeling and landscape characterizations. Ecological Modelling, 433, 109231.

Holland, J. D., Bert, D. G., and Fahrig, L. (2004). Determining the spatial scale of species' response to habitat. Bioscience, 54(3), 227-233.

Hopkins, L. M., Hallman, T. A., Kilbride, J., Robinson, W. D., and Hutchinson, R. A. (2022). A comparison of remotely sensed environmental predictors for avian distributions. Landscape Ecology, 37(4), 997-1016.

Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D., and Willis, S. G. (2014). Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution, 5(6), 506-513.

Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D., Butchart, S. H., and Willis, S. G. (2020). Disentangling the relative roles of climate and land cover change in driving the long-term population trends of European migratory birds. Diversity and Distributions, 26(11), 1442-1455.

Humbold von, A. and Bonpland, A. (1807).Essai sur la géographie desplante. Schoel and Co., Lyon

Hurlbert, A. H., and Jetz, W. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences, 104(33), 13384-13389.

Inman, R., Franklin, J., Esque, T., and Nussear, K. (2021). Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere, 12(3), e03422.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., ... and O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. Trends in ecology and evolution, 35(1), 56-67.

Jaureguiberry, P., Titeux, N., Wiemers, M., Bowler, D. E., Coscieme, L., Golden, A. S., ... and Purvis, A. (2022). The direct drivers of recent global anthropogenic biodiversity loss. Science advances, 8(45), eabm9982.

Jetz, W., McPherson, J. M., and Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. Trends in ecology and evolution, 27(3), 151-159.

Jiménez-Valverde, A., Lobo, J. M., and Hortal, J. (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. (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. (2020). Sample size for the evaluation of presence-absence models. Ecological Indicators, 114, 106289.

Jiménez-Valverde, A., Aragón, P., and Lobo, J. M. (2021). Deconstructing the abundance–suitability relationship in species distribution modelling. Global Ecology and Biogeography, 30(1), 327-338.

Johnson, C. J., and Gillingham, M. P. (2005). An evaluation of mapped species distribution models used for conservation planning. Environmental Conservation, 32(2), 117-128.

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.

Johnson, C. J., Hurley, M., Rapaport, E., and Pullinger, M. (2012). Using expert knowledge effectively: lessons from species distribution models for wildlife conservation and management. Expert knowledge and its application in landscape ecology, 153-171.

Johnston, A., Hochachka, W., Strimas-Mackey, M., Gutierrez, V. R., Robinson, O., Miller, E., ... and Fink, D. (2019). Best practices for making reliable inferences from citizen science data: case study using eBird to estimate species distributions. BioRxiv, 574392.

Kaliontzopoulou, A., Brito, J. C., Carretero, M. A., Larbes, S., and Harris, D. J. (2008). Modelling the partially unknown distribution of wall lizards (Podarcis) in North Africa: ecological affinities, potential areas of occurrence, and methodological constraints. Canadian Journal of Zoology, 86(9), 992-1001.

Kamino, L. H., Stehmann, J. R., Amaral, S., De Marco Jr, P., Rangel, T. F., de Siqueira, M. F., ... and Hortal, J. (2012). Challenges and perspectives for species distribution modelling in the neotropics. Biology letters, 8(3), 324-326.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., ... and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. Scientific data, 4(1), 1-20.

Karger, D. N., Wilson, A. M., Mahony, C., Zimmermann, N. E., and Jetz, W. (2021). Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. Scientific Data, 8(1), 307.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P. (2021, July). Global land use/land cover with Sentinel 2 and deep learning. In 2021 IEEE international geoscience and remote sensing symposium IGARSS (pp. 4704-4707). IEEE.

Keil, P., Wilson, A. M., and Jetz, W. (2014). Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. Diversity and Distributions, 20(7), 797-812.

Kelling, S., Johnston, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Bonn, A., ... and Guralnick, R. (2018). Finding the signal in the noise of Citizen Science Observations. bioRxiv, 326314.

Keppie, D. M., and Braun, C. E. (2000). Band-tailed Pigeon. American Ornithologists' Union.

Khosravipour, A., Skidmore, A. K., and Isenburg, M. (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.

Kindt, R. (2018). Ensemble species distribution modelling with transformed suitability values. Environmental Modelling and Software, 100, 136-145.

Klápště, P., Fogl, M., Barták, V., Gdulová, K., Urban, R., and Moudrý, V. (2020). Sensitivity analysis of parameters and contrasting performance of ground filtering algorithms with UAV photogrammetry-based and LiDAR point clouds. International Journal of Digital Earth, 13(12), 1672-1694.

Kleisner, K. M., Fogarty, M. J., McGee, S., Hare, J. A., Moret, S., Perretti, C. T., and Saba, V. S. (2017). Marine species distribution shifts on the US Northeast Continental Shelf under continued ocean warming. Progress in Oceanography, 153, 24-36.

Koma, Z., Seijmonsbergen, A. C., Grootes, M. W., Nattino, F., Groot, J., Sierdsema, H., ... and Kissling, W. D. (2022). Better together? Assessing different remote sensing products for predicting habitat suitability of wetland birds. Diversity and Distributions, 28(4), 685-699.

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.

Kosmala, M., Wiggins, A., Swanson, A., and Simmons, B. (2016). Assessing data quality in citizen science. Frontiers in Ecology and the Environment, 14(10), 551-560.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., ... and Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and distributions, 19(11), 1366-1379.

Lagner, O., Klouček, T., and Šímová, P. (2018). Impact of input data (in) accuracy on overestimation of visible area in digital viewshed models. PeerJ, 6, e4835.

Lande, R., Engen, S., and Saether, B. E. (2003). Stochastic population dynamics in ecology and conservation. Oxford University Press on Demand.

Lash, R. R., Carroll, D. S., Hughes, C. M., Nakazawa, Y., Karem, K., Damon, I. K., and Peterson, A. T. (2012). Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk. International journal of health geographics, 11(1), 1-12.

Lawson, C. R., Hodgson, J. A., Wilson, R. J., and Richards, S. A. (2014). Prevalence, thresholds and the performance of presence–absence models. Methods in Ecology and Evolution, 5(1), 54-64.

Lechner, A. M., Langford, W. T., Jones, S. D., Bekessy, S. A., and Gordon, A. (2012). Investigating species–environment relationships at multiple scales: Differentiating between intrinsic scale and the modifiable areal unit problem. Ecological Complexity, 11, 91-102.

Lecours, V., Devillers, R., Schneider, D. C., Lucieer, V. L., Brown, C. J., and Edinger, E. N. (2015). Spatial scale and geographic context in benthic habitat mapping: review and future directions. Marine Ecology Progress Series, 535, 259-284.

Lecours, V., Gábor, L., Edinger, E., and Devillers, R. (2020). Fine-scale habitat characterization of The Gully, the Flemish Cap, and the Orphan Knoll, Northwest Atlantic, with a focus on

cold-water corals. In Seafloor Geomorphology as Benthic Habitat (pp. 735-751). Elsevier.

Lee-Yaw, J. A., Kharouba, H. M., Bontrager, M., Mahony, C., Csergő, A. M., Noreen, A. M., ... and Angert, A. L. (2016). A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. Ecology letters, 19(6), 710-722.

Legendre, P., Dale, M. R., Fortin, M. J., Gurevitch, J., Hohn, M., and Myers, D. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography, 25(5), 601-615.

Leibold, M. A., and Chase, J. M. (2017). Metacommunity Ecology, Volume 59 (Vol. 59). Princeton University Press.

Lembrechts, J. J., Lenoir, J., Roth, N., Hattab, T., Milbau, A., Haider, S., ... and Nijs, I. (2019). Comparing temperature data sources for use in species distribution models: From in-situ logging to remote sensing. Global Ecology and Biogeography, 28(11), 1578-1596.

Lembrechts, J. J., Nijs, I., and Lenoir, J. (2019). Incorporating microclimate into species distribution models. Ecography, 42(7), 1267-1279.

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. Ecography, 39(6), 599-607.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., and Bellard, C. (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.

Li, Q., and Kou, X. (2021). WiBB: an integrated method for quantifying the relative importance of predictive variables. Ecography.

Li, R., Ranipeta, A., Wilshire, J., Malczyk, J., Duong, M., Guralnick, R., ... and Jetz, W. (2021). A cloud-based toolbox for the versatile environmental annotation of biodiversity data. PLoS biology, 19(11), e3001460.

Li, W., and Guo, Q. (2013). How to assess the prediction accuracy of species presence–absence models without absence data?. Ecography, 36(7), 788-799.

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. et al. (eds), Proceedings of central European silviculture, pp. 71–77.

Lindegren, M., van Deurs, M., Maureaud, A., Thorson, J. T., and Bekkevold, D. (2022). A spatial statistical approach for identifying population structuring of marine fish species: European sprat as a case study. ICES Journal of Marine Science, 79(2), 423-434.

Lister, A. M. (2011). Natural history collections as sources of long-term datasets. Trends in ecology and evolution, 26(4), 153-154.

Liu, C., Newell, G., and White, M. (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.

Liu, C., Wolter, C., Courchamp, F., Roura-Pascual, N., and Jeschke, J. M. (2022). Biological invasions reveal how niche change affects the transferability of species distribution models. Ecology, 103(8), e3719.

Lo Parrino, E., Falaschi, M., Manenti, R., and Ficetola, G. F. (2023). All that changes is not shift: methodological choices influence niche shift detection in freshwater invasive species. Ecography, e06432.

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), 145-151.

Loreau, M., Cardinale, B. J., Isbell, F., Newbold, T., O'Connor, M. I., and de Mazancourt, C. (2022). Do not downplay biodiversity loss. Nature, 601(7894), E27-E28.

Lu, M., and Jetz, W. (2023). Scale-sensitivity in the measurement and interpretation of environmental niches. Trends in Ecology and Evolution.

Luebert, F., Fuentes-Castillo, T., Pliscoff, P., García, N., Román, M. J., Vera, D., and Scherson, R. A. (2022). Geographic Patterns of Vascular Plant Diversity and Endemism Using Different Taxonomic and Spatial Units. Diversity, 14(4), 271.

MacArthur, R. (1960). On the relative abundance of species. The American Naturalist, 94(874), 25-36.

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.

Maiorano, L., Chiaverini, L., Falco, M., and Ciucci, P. (2019). Combining multi-state species distribution models, mortality estimates, and landscape connectivity to model potential species distribution for endangered species in human dominated landscapes. Biological Conservation, 237, 19-27.

Makridakis, S., and Winkler, R. L. (1983). Averages of forecasts: Some empirical results. Management science, 29(9), 987-996.

Manzoor, S. A., Griffiths, G., and Lukac, M. (2018). Species distribution model transferability and model grain size–finer may not always be better. Scientific reports, 8(1), 7168.

Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C., and Picó, F. X. (2016). Tackling intraspecific genetic structure in distribution models better reflects species geographical range. Ecology and evolution, 6(7), 2084-2097.

Marešová, J., Gdulová, K., Pracná, P., Moravec, D., Gábor, L., Prošek, J., ... and Moudrý, V. (2021). Applicability of data acquisition characteristics to the identification of local artefacts in global digital elevation models: Comparison of the copernicus and tandem-x dems. Remote Sensing, 13(19), 3931.

Marsh, C. J., Gavish, Y., Kuemmerlen, M., Stoll, S., Haase, P., and Kunin, W. E. (2023). SDM profiling: A tool for assessing the information-content of sampled and unsampled locations for species distribution models. Ecological Modelling, 475, 110170.

Marthews, T. R., Dadson, S. J., Lehner, B., Abele, S., and Gedney, N. (2015). High-resolution global topographic index values for use in large-scale hydrological modelling. Hydrology and Earth System Sciences, 19(1), 91-104.

Martin, A. E., and Fahrig, L. (2012). Measuring and selecting scales of effect for landscape predictors in species–habitat models. Ecological Applications, 22(8), 2277-2292.

Marzialetti, F., Bazzichetto, M., Giulio, S., Acosta, A. T., Stanisci, A., Malavasi, M., and Carranza, M. L. (2019). Modelling Acacia saligna invasion on the Adriatic coastal landscape: An integrative approach using LTER data. Nature Conservation, 34, 127.

Mateo Sanchez, M. C., Cushman, S. A., and Saura, S. (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, R. G., Felicísimo, Á. M., Pottier, J., Guisan, A., and Munoz, J. (2012). Do stacked species distribution models reflect altitudinal diversity patterns?. PLoS One, 7(3), e32586.

Mayor, S. J., Schneider, D. C., Schaefer, J. A., and Mahoney, S. P. (2009). Habitat selection at multiple scales. Ecoscience, 16(2), 238-247.

Mazziotta, A., Lindén, A., Eyvindson, K., Bianchi, S., Kangas, A., Melin, M., ... and Forsman, J. T. (2024). Unraveling the characteristic spatial scale of habitat selection for forest grouse species in the boreal landscape. Forest Ecology and Management, 563, 122008.

McCluskey, E. M., Matthews, S. N., Ligocki, I. Y., Holding, M. L., Lipps Jr, G. J., and Hetherington, T. E. (2018). The importance of historical land use in the maintenance of early successional habitat for a threatened rattlesnake. Global Ecology and Conservation, 13, e00370.

McCullagh, P., and Nelder, J. A. (1989). Generalized linear models 2nd edition chapman and hall. London, UK.

McGarigal, K., Wan, H. Y., Zeller, K. A., Timm, B. C., and Cushman, S. A. (2016). Multi-scale habitat selection modeling: a review and outlook. Landscape ecology, 31(6), 1161-1175.

McPherson, J. M., Jetz, W., and Rogers, D. J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact?. Journal of applied ecology, 41(5), 811-823.

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

Melo, I., Ochoa-Quintero, J. M., de Oliveira Roque, F., and Dalsgaard, B. (2018). A review of threshold responses of birds to landscape changes across the world. Journal of Field Ornithology, 89(4), 303-314.

Mendes, P., Velazco, S. J. E., de Andrade, A. F. A., and Júnior, P. D. M. (2020). Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy. Ecological Modelling, 431, 109180.

Merow, C., Smith, M. J., and Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography, 36(10), 1058-1069.

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., ... and Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models?. Ecography, 37(12), 1267-1281.

Merow, C., Maitner, B. S., Owens, H. L., Kass, J. M., Enquist, B. J., Jetz, W., and Guralnick, R. (2019). Species' range model metadata standards: RMMS. Global Ecology and Biogeography, 28(12), 1912-1924.

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

Mertes, K., Jarzyna, M. A., and Jetz, W. (2020). Hierarchical multi-grain models improve descriptions of species' environmental associations, distribution, and abundance. Ecological Applications, 30(6), e02117.

Meyer, C. B. (2007). Does scale matter in predicting species distributions? Case study with the marbled murrelet. Ecological Applications, 17(5), 1474-1483.

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.

Meynard, C. N., Leroy, B., and Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?. Ecography, 42(12), 2021-2036.

Mi, C., Huettmann, F., Guo, Y., Han, X., and Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ, 5, e2849.

Michener, W. K. (2015). Ecological data sharing. Ecological informatics, 29, 33-44.

Miguet, P., Jackson, H. B., Jackson, N. D., Martin, A. E., and Fahrig, L. (2016). What determines the spatial extent of landscape effects on species?. Landscape ecology, 31, 1177-1194.

Milanesi, P., Herrando, S., Pla, M., Villero, D., and Keller, V. (2017). Towards continental bird distribution models: environmental variables for the second European breeding bird atlas and identification of priorities for further surveys. Vogelwelt, 137, 53-60.

Miller, J. (2010). Species distribution modeling. Geography Compass, 4(6), 490-5.

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., Lecours, V., and Bell, T. (2018). A multiscale approach to mapping seabed sediments. PLoS One, 13(2), e0193647.

Mitchell, P. J., Monk, J., and Laurenson, L. (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., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. Journal of Vegetation Science, 27(6), 1308-1322.

Mohammadi, A., Almasieh, K., Nayeri, D., Adibi, M. A., and Wan, H. Y. (2022). Comparison of habitat suitability and connectivity modelling for three carnivores of conservation concern in an Iranian montane landscape. Landscape Ecology, 37(2), 411-430.

Monk, J. (2014). How long should we ignore imperfect detection of species in the marine environment when modelling their distribution?. Fish and Fisheries, 15(2), 352-358.

Montgomery, R. A., Roloff, G. J., and Hoef, J. M. V. (2011). Implications of ignoring telemetry error on inference in wildlife resource use models. The Journal of Wildlife Management, 75(3), 702-708.

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.

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

Moudrý, V., Komárek, J., and Šímová, P. (2017). Which breeding bird categories should we use in models of species distribution?. Ecological Indicators, 74, 526-529.

Moudrý, V., Lecours, V., Gdulová, K., Gábor, L., Moudrá, L., Kropáček, J., and Wild, J. (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., Lecours, V., Malavasi, M., Misiuk, B., Gábor, L., Gdulová, K., ... and Wild, J. (2019). Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies. Ecological Informatics, 54, 100987.

Moudrý, V., and Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Informatics, 56, 101051.

Moudrý, V., Moudrá, L., Barták, V., Bejček, V., Gdulová, K., Hendrychová, M., ... and Šálek, M. (2021). The role of the vegetation structure, primary productivity and senescence derived from airborne LiDAR and hyperspectral data for birds diversity and rarity on a restored site. Landscape and Urban Planning, 210, 104064.

Moudrý, V., Cord, A. F., Gábor, L., Laurin, G. V., Barták, V., Gdulová, K., ... and Wild, J. (2023a). Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. Diversity and Distributions, 29(1), 39-50.

Moudrý, V., Keil, P., Gábor, L., Lecours, V., Zarzo-Arias, A., Barták, V., ... and Šímová, P. (2023b). Scale mismatches between predictor and response variables in species distribution modelling: A review of practices for appropriate grain selection. Progress in Physical Geography: Earth and Environment, 47(3), 467-482.

Müller, J., and Brandl, R. (2009). Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. Journal of Applied Ecology, 46(4), 897-905.

Murray, K., and Conner, M. M. (2009). Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology, 90(2), 348-355.

Naimi, B., Skidmore, A. K., Groen, T. A., and Hamm, N. A. (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.

Naimi, B., Hamm, N. A., Groen, T. A., Skidmore, A. K., and Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling?. Ecography, 37(2), 191-203.

Naimi, B., and Araújo, M. B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. Ecography, 39(4), 368-375.

Nelder, J. A., and Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

Neyman, J., and Scott, E. L. (1959). Stochastic Models of Population Dynamics: A few simple chance mechanisms may combine to reproduce many manifestations of a complex phenomenon. Science, 130(3371), 303-308.

Nogués-Bravo, D., Rodríguez-Sánchez, F., Orsini, L., de Boer, E., Jansson, R., Morlon, H., ... and Jackson, S. T. (2018). Cracking the code of biodiversity responses to past climate change. Trends in ecology and evolution, 33(10), 765-776.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., ... and Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological monographs, 89(3), e01370.

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.

Oleas, N. H., Feeley, K. J., Fajardo, J., Meerow, A. W., Gebelein, J., and Francisco-Ortega, J. (2019). Muddy boots beget wisdom: Implications for rare or endangered plant species distribution models. Diversity, 11(1), 10.

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.

Palmer, M. W. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. Ecology, 74(8), 2215-2230.

Paradinas, I., Illian, J., and Smout, S. (2022). Understanding spatial effects in species distribution models. Authorea Preprints.

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. Wiley interdisciplinary reviews: Climate change, 4(3), 213-223.

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

Pearman, P. B., Guisan, A., Broennimann, O., and Randin, C. F. (2008). Niche dynamics in space and time. Trends in ecology and evolution, 23(3), 149-158.

Pearson, R. G., and Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful?. Global ecology and biogeography, 12(5), 361-371.

Pearson, R. G., Dawson, T. P., and Liu, C. (2004). Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography, 27(3), 285-298.

Pebesma, E. J. (2018). Simple features for R: standardized support for spatial vector data. R J., 10(1), 439.

Peng, S., Zhang, J., Zhang, X., Li, Y., Liu, Y., and Wang, Z. (2022). Conservation of woody species in China under future climate and land-cover changes. Journal of Applied Ecology,

59(1), 141-152.

Petroselli, A., Vessella, F., Cavagnuolo, L., Piovesan, G., and Schirone, B. (2013). Ecological behavior of Quercus suber and Quercus ilex inferred by topographic wetness index (TWI). Trees, 27(5), 1201-1215.

Phillips S.J., Dudík, M., Schapire. R.E. [Internet] Maxent software for modeling species niches and distributions (Version 3.4.3).

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological modelling, 190(3-4), 231-259.

Phillips, S. J., and Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31(2), 161-175.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil, 7(1), 217-240.

Pradervand, J. N., Dubuis, A., Pellissier, L., Guisan, A., and Randin, C. (2014). Very high resolution environmental predictors in species distribution models: moving beyond topography?. Progress in Physical Geography, 38(1), 79-96.

Prošek, J., Gdulová, K., Barták, V., Vojar, J., Solský, M., Rocchini, D., and Moudrý, V. (2020). Integration of hyperspectral and LiDAR data for mapping small water bodies. International Journal of Applied Earth Observation and Geoinformation, 92, 102181.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahbek, C. (2005). The role of spatial scale and the perception of large-scale species-richness patterns. Ecology letters, 8(2), 224-239.

Randin, C. F., Ashcroft, M. B., Bolliger, J., Cavender-Bares, J., Coops, N. C., Dullinger, S., ... and Payne, D. (2020). Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models. Remote sensing of environment, 239, 111626.

Reif, J., Voříšek, P., Šťastný, K., and Bejček, V. (2006). Population trends of birds in the Czech Republic during 1982–2005. Sylvia, 42, 22-37.

Reif, J., Reifová, R., Skoracka, A., and Kuczyński, L. (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.

Renner, I. W., Louvrier, J., and Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. Methods in Ecology and Evolution, 10(12), 2118-2128.

Reside, A. E., Watson, I., VanDerWal, J., and Kutt, A. S. (2011). Incorporating low-resolution historic species location data decreases performance of distribution models. Ecological Modelling, 222(18), 3444-3448.

Reside, A. E., Critchell, K., Crayn, D. M., Goosem, M., Goosem, S., Hoskin, C. J., ... and Pressey, R. L. (2019). Beyond the model: expert knowledge improves predictions of species' fates under climate change. Ecological Applications, 29(1), e01824.

Riva, F., Barbero, F., Balletto, E., and Bonelli, S. (2023). Combining environmental niche models, multi-grain analyses, and species traits identifies pervasive effects of land use on butterfly biodiversity across Italy. Global change biology, 29(7), 1715-1728.

Roberts, D. W. (1986). Ordination on the basis of fuzzy set theory. Vegetatio, 66(3), 123-131.

Robertson, M. P., Visser, V., and Hui, C. (2016). Biogeo: an R package for assessing and improving data quality of occurrence record datasets. Ecography, 39(4), 394-401.

Robinson, L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., Possingham, H. P., and Richardson, A. J. (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.

Robinson, N., Regetz, J., and Guralnick, R. P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. ISPRS Journal of Photogrammetry and Remote Sensing, 87, 57-67.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A. M., Bazzichetto, M., ... and Malavasi, M. (2023). A quixotic view of spatial bias in modelling the distribution of species and their diversity. npj Biodiversity, 2(1), 10.

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, L., García, J. J., Carreño, F., and Martínez, B. (2019). Integration of physiological knowledge into hybrid species distribution modelling to improve forecast of distributional shifts of tropical corals. Diversity and Distributions, 25(5), 715-728.

Romero, D., Olivero, J., and Real, R. (2013). Comparative assessment of different methods for using land-cover variables for distribution modelling of Salamandra salamandra longirotris. Environmental conservation, 40(1), 48-59.

Rose, J. P., Halstead, B. J., and Fisher, R. N. (2020). Integrating multiple data sources and multi-scale land-cover data to model the distribution of a declining amphibian. Biological Conservation, 241, 108374.

Sanguet, A., Wyler, N., Petitpierre, B., Honeck, E., Poussin, C., Martin, P., and Lehmann, A. (2022). Beyond topo-climatic predictors: Does habitats distribution and remote sensing information improve predictions of species distribution models?. Global Ecology and Conservation, e02286.

Santamarina, S., Alfaro-Saiz, E., Llamas, F., and Acedo, C. (2019). Different approaches to assess the local invasion risk on a threatened species: Opportunities of using high-resolution species distribution models by selecting the optimal model complexity. Global Ecology and Conservation, 20, e00767.

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., and Huijbregts, M. A. (2021). Assessing the reliability of species distribution projections in climate change research. Diversity and Distributions, 27(6), 1035-1050.

Sauer, J., Niven, D., Hines, J., Ziolkowski Jr, D., Pardieck, K. L., Fallon, J. E., and Link, W. (2017). The North American breeding bird survey, results and analysis 1966-2015.

Schleuning, M., Neuschulz, E. L., Albrecht, J., Bender, I. M., Bowler, D. E., Dehling, D. M., ... and Kissling, W. D. (2020). Trait-based assessments of climate-change impacts on interacting species. Trends in Ecology and Evolution, 35(4), 319-328.

Schmeller, D. S., Weatherdon, L. V., Loyau, A., Bondeau, A., Brotons, L., Brummitt, N., ... and Regan, E. C. (2018). A suite of essential biodiversity variables for detecting critical biodiversity change. Biological Reviews, 93(1), 55-71.

Schneider, D. C. (2001). The rise of the concept of scale in ecology: The concept of scale is evolving from verbal expression to quantitative expression. BioScience, 51(7), 545-553.

Schroeder, B. (2008). Challenges of species distribution modeling belowground. Journal of Plant Nutrition and Soil Science, 171(3), 325-337.

Segal, R. D., Massaro, M., Carlile, N., and Whitsed, R. (2021). Small-scale species distribution model identifies restricted breeding habitat for an endemic island bird. Animal Conservation.

Seo, C., Thorne, J. H., Hannah, L., and Thuiller, W. (2009). Scale effects in species distribution models: implications for conservation planning under climate change. Biology letters, 5(1), 39-43.

Shipley, B. R., Bach, R., Do, Y., Strathearn, H., McGuire, J. L., and Dilkina, B. (2022). megaSDM: integrating dispersal and time-step analyses into species distribution models. Ecography, 2022(1).

Sillero, N., and Goncalves-Seco, L. (2014). Spatial structure analysis of a reptile community with airborne LiDAR data. International Journal of Geographical Information Science, 28(8), 1709-1722.

Sillero, N., and Barbosa, A. M. (2021). Common mistakes in ecological niche models. International Journal of Geographical Information Science, 35(2), 213-226.

Silveira, E. M. O., Pidgeon, A. M., Farwell, L. S., Hobi, M. L., Razenkova, E., Zuckerberg, B., ... and Radeloff, V. C. (2023). Multi-grain habitat models that combine satellite sensors with different resolutions explain bird species richness patterns best. Remote Sensing of Environment, 295, 113661.

Simoes, M., Romero-Alvarez, D., Nuñez-Penichet, C., Jiménez, L., and Cobos, M. E. (2020). General theory and good practices in ecological niche modeling: a basic guide. Biodiversity Informatics, 15(2), 67-68.

Simonson, W. D., Allen, H. D., and Coomes, D. A. (2014). Applications of airborne lidar for the assessment of animal species diversity. Methods in Ecology and Evolution, 5(8), 719-729.

Šímová, P., Moudrý, V., Komárek, J., Hrach, K., and Fortin, M. J. (2019). Fine scale waterbody data improve prediction of waterbird occurrence despite coarse species data. Ecography, 42(3), 511-520.

Smith, T. J., and McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. Multiple Linear Regression Viewpoints, 39(2), 17-26.

Smith, A. B., and Santos, M. J. (2020). Testing the ability of species distribution models to infer variable importance. Ecography, 43(12), 1801-1813.

Smith, A. B., Murphy, S. J., Henderson, D., and Erickson, K. D. (2021). Imprecisely georeferenced specimen data provide unique information on species' distributions and environmental tolerances: Don't let the perfect be the enemy of the good. bioRxiv.

Somveille, M., Wikelski, M., Beyer, R. M., Rodrigues, A. S., Manica, A., and Jetz, W. (2020). Simulation-based reconstruction of global bird migration over the past 50,000 years. Nature communications, 11(1), 1-9.

Soultan, A., and Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. PloS one, 12(11), e0187906.

Šťastný, K., Hudec, K., and Bejček, V. (2006). Atlas of breeding bird distribution in the Czech Republic 2001-2003. Aventinum.

Stanton, J. C., Pearson, R. G., Horning, N., Ersts, P., and Resit Akcakaya, H. (2012). Combining static and dynamic variables in species distribution models under climate change. Methods in Ecology and Evolution, 3(2), 349-357.

Stark, J. R., and Fridley, J. D. (2022). Microclimate-based species distribution models in complex forested terrain indicate widespread cryptic refugia under climate change. Global Ecology and Biogeography, 31(3), 562–575.

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

Stuber, E. F., and Fontaine, J. J. (2019). How characteristic is the species characteristic selection scale?. Global Ecology and Biogeography, 28(12), 1839-1854.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... and Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. Biological conservation, 169, 31-40.

Sutton, L. J., Ibañez, J. C., Salvador, D. I., Taraya, R. L., Opiso, G. S., Senarillos, T. L. P., and McClure, C. J. (2023). Priority conservation areas and a global population estimate for the critically endangered Philippine Eagle. Animal Conservation.

Syfert, M. M., Smith, M. J., and Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PloS one, 8(2), e55158.

Synes, N. W., and Osborne, P. E. (2011). Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. Global Ecology and Biogeography, 20(6), 904-914.

Tan, C. O., Özesmi, U., Beklioglu, M., Per, E., and Kurt, B. (2006). Predictive models in ecology: comparison of performances and assessment of applicability. Ecological Informatics, 1(2), 195-211.

Termansen, M., McClean, C. J., and Preston, C. D. (2006). The use of genetic algorithms and Bayesian classification to model species distributions. Ecological modelling, 192(3-4), 410-424.

Tessarolo, G., Ladle, R. J., Lobo, J. M., Rangel, T. F., and Hortal, J. (2021). Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. Ecography, 44(12), 1743-1755.

Title, P. O., and Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography, 41(2), 291-307.

Trainor, A. M., Schmitz, O. J., Ivan, J. S., and Shenk, T. M. (2014). Enhancing species distribution modeling by characterizing predator–prey interactions. Ecological Applications, 24(1), 204-216.

Tuanmu, M. N., and Jetz, W. (2015). A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. Global Ecology and Biogeography, 24(11), 1329-1339.

Tulowiecki, S. J., Larsen, C. P., and Wang, Y. C. (2015). Effects of positional error on modeling species distributions: a perspective using presettlement land survey records. Plant Ecology, 216, 67-85.

Tulowiecki, S. J. (2020). Modeling the historical distribution of American chestnut (Castanea dentata) for potential restoration in western New York State, US. Forest Ecology and Management, 462, 118003.

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J. B., Pe'er, G., Singer, A., ... and Travis, J. M. (2016). Improving the forecast for biodiversity under climate change. Science, 353(6304), aad8466.

Václavík, T., Kupfer, J. A., and Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). Journal of Biogeography, 39(1), 42-55.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs, 92(1), e01486.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2023). Flexible species distribution modelling methods perform well on spatially separated testing data. Global Ecology and Biogeography, 32(3), 369-383.

Van Moorter, B., Kivimäki, I., Panzacchi, M., Saura, S., Brandão Niebuhr, B., Strand, O., and Saerens, M. (2023). Habitat functionality: Integrating environmental and geographic space in niche modeling for conservation planning. Ecology, e4105.

van Proosdij, A. S., Sosef, M. S., Wieringa, J. J., and Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. Ecography, 39(6), 542-552.

VanDerWal, J., Shoo, L. P., Graham, C., and Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?. Ecological modelling, 220(4), 589-594.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography, 37(11), 1084-1091.

Velásquez-Tibatá, J., Graham, C. H., and Munch, S. B. (2016). Using measurement error models to account for georeferencing error in species distribution models. Ecography, 39(3), 305-316.

Velazco, S. J. E., Ribeiro, B. R., Laureto, L. M. O., and Júnior, P. D. M. (2020). Overprediction of species distribution models in conservation planning: A still neglected issue with strong effects. Biological Conservation, 252, 108822.

Venne, S., and Currie, D. J. (2021). Can habitat suitability estimated from MaxEnt predict colonizations and extinctions?. Diversity and Distributions, 27(5), 873-886.

Venter, Z. S., Roos, R. E., Nowell, M. S., Rusch, G. M., Kvifte, G. M., and Sydenham, M. A. (2023). Comparing global Sentinel-2 land cover maps for regional species distribution modeling. Remote Sensing, 15(7), 1749.

Vierling, K. T., Bässler, C., Brandl, R., Vierling, L. A., Weiß, I., and Müller, J. (2011). Spinning a laser web: predicting spider distributions using LiDAR. Ecological Applications, 21(2), 577-588.

Virkkala, R., Luoto, M., Heikkinen, R. K., and Leikola, N. (2005). Distribution patterns of boreal marshland birds: modelling the relationships to land cover and climate. Journal of Biogeography, 32(11), 1957-1970.

Visscher, R. D. (2006). GPS measurement error and resource selection functions in a fragmented landscape. Ecography, 29(3), 458-464.

Vogeler, J. C., Hudak, A. T., Vierling, L. A., Evans, J., Green, P., and Vierling, K. T. (2014). Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. Remote Sensing of Environment, 147, 13-22.

Volis, S., and Tojibaev, K. (2021). Defining critical habitat for plant species with poor occurrence knowledge and identification of critical habitat networks. Biodiversity and Conservation, 30, 3603-3611.

Vollering, J., Halvorsen, R., Auestad, I., and Rydgren, K. (2019). Bunching up the background betters bias in species distribution models. Ecography, 42(10), 1717-1727.

Vollering, J., Schuiteman, A., de Vogel, E., van Vugt, R., and Raes, N. (2016). Phytogeography of New Guinean orchids: patterns of species richness and turnover. Journal of Biogeography, 43(1), 204-214.

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W., Edgar, G. J., Mouillot, D., ... and Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. Ecography, 2022(1).

Walthert, L., and Meier, E. S. (2017). Tree species distribution in temperate forests is more influenced by soil than by climate. Ecology and evolution, 7(22), 9473-9484.

Wang, L., and Jackson, D. A. (2023). Effects of sample size, data quality, and species response in environmental space on modeling species distributions. Landscape Ecology, 1-23.

Watcharamongkol, T., Christin, P. A., and Osborne, C. P. (2018). C4 photosynthesis evolved in warm climates but promoted migration to cooler ones. Ecology letters, 21(3), 376-383.

Wheater, C. P., Bell, J. R., and Cook, P. A. (2020). Practical field ecology: a project guide. John Wiley and Sons.

Wieczorek, J., Guo, Q., and Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International journal of geographical information science, 18(8), 745-767.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ... and Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. PloS one, 7(1), e29715.

Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L., and Ferrier, S. (2012). Which environmental variables should I use in my biodiversity model?. International Journal of Geographical Information Science, 26(11), 2009-2047.

Wilson, A. M., and Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLoS biology, 14(3), e1002415.

Windsor, F. M., van den Hoogen, J., Crowther, T. W., and Evans, D. M. (2023). Using ecological networks to answer questions in global biogeography and ecology. Journal of Biogeography, 50(1), 57-69.

Winner, K., Noonan, M. J., Fleming, C. H., Olson, K. A., Mueller, T., Sheldon, D., and Calabrese, J. M. (2018). Statistical inference for home range overlap. Methods in Ecology and Evolution, 9(7), 1679-1691.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., and NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. Diversity and distributions, 14(5), 763-773.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., ... and Svenning, J. C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological reviews, 88(1), 15-30.

Work, T. T., Onge, B. S., and Jacobs, J. M. (2011). Response of female beetles to LIDAR derived topographic variables in Eastern boreal mixedwood forests (Coleoptera, Carabidae). ZooKeys, (147), 623.

Wright, M.N., Wager, S., Probst, P., Wright, M.M.N., (2018). Package 'ranger.'

uncertainty analysis in ecology (pp. 3-15). Springer, Dordrecht.

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., ... and Karger, D. N. (2020). Macroecology in the age of Big Data–Where to go from here?. Journal of Biogeography, 47(1), 1-12.

Wunder, J., Reineking, B., Bigler, C., and Bugmann, H. (2008). Predicting tree mortality from growth data: how virtual ecologists can help real ecologists. Journal of Ecology, 96(1), 174-187.

Wunderlich, R. F., Mukhtar, H., and Lin, Y. P. (2022). Comprehensively evaluating the performance of species distribution models across clades and resolutions: choosing the right tool for the job. Landscape Ecology, 37(8), 2045-2063.

Yanco, S. W., McDevitt, A., Trueman, C. N., Hartley, L., and Wunder, M. B. (2020). A modern method of multiple working hypotheses to improve inference in ecology. Royal Society open science, 7(6), 200231.

Yang, L., Chen, T., Shi, K. C., Zhang, L., Lwin, N., and Fan, P. F. (2023). Effects of climate and land-cover change on the conservation status of gibbons. Conservation Biology, 37(1), e14045.

Yee, T. W., and Mitchell, N. D. (1991). Generalized additive models in plant ecology. Journal of vegetation science, 2(5), 587-602.

Young, M. A., Critchell, K., Miller, A. D., Treml, E. A., Sams, M., Carvalho, R., and Ierodiaconou, D. (2023). Mapping the impacts of multiple stressors on the decline in kelps along the coast of Victoria, Australia. Diversity and Distributions, 29(1), 199-220.

Yu, H., Cooper, A. R., and Infante, D. M. (2020). Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. Ecological Modelling, 432, 109202.

Zarzo-Arias, A., Penteriani, V., Gábor, L., Šímová, P., Grattarola, F., and Moudrý, V. (2022). Importance of data selection and filtering in species distribution models: A case study on the Cantabrian brown bear. Ecosphere, 13(12), e4284.

Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D., and Coomes, D. (2019). Advances in microclimate ecology arising from remote sensing. Trends in Ecology and Evolution, 34(4), 327-341.

Zhang, G., Zhu, A. X., Huang, Z. P., and Xiao, W. (2018). A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. Transactions in GIS, 22(1), 202-216.

Zhang, X., and Vincent, A. C. (2018). Predicting distributions, habitat preferences and associated conservation implications for a genus of rare fishes, seahorses (Hippocampus spp.). Diversity and Distributions, 24(7), 1005-1017.

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., and Mi, J. (2021). GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery. Earth System Science Data, 13(6), 2753-2776.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... and Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods in Ecology and Evolution, 10(5), 744-751.

Zizka, A., Carvalho, F. A., Calvente, A., Baez-Lizarazo, M. R., Cabral, A., Coelho, J. F. R., ... and Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. PeerJ, 8, e9916.

Zuckerberg, B., Fink, D., La Sorte, F. A., Hochachka, W. M., and Kelling, S. (2016). Novel seasonal land cover associations for eastern North American forest birds identified through dynamic species distribution modelling. Diversity and Distributions, 22(6), 717-730.

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., ... and Grimm, V. (2010). The virtual ecologist approach: simulating data and observers. Oikos, 119(4), 622-635.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., ... and Merow, C. (2020). A standard protocol for reporting species distribution models. Ecography, 43(9), 1261-1277.

Zuur, A. F., Ieno, E. N., and Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. Methods in ecology and evolution, 1(1), 3-14.

Author's Academic CV

Personal

Name: Lukáš Gábor
Date of birth: January 1^{st}, 1991, Czech Republic
E-mail: gabor.lucas@gmail.com
ORCID: https://orcid.org/0000-0001-6137-0994
Google Scholar: https://scholar.google.cz/citations?user=pLQXY5wAAAAJ&hl=cs

Affiliations

2021 – 2023
Yale University
Jetz Lab
Department of Ecology and Evolutionary Biology
Yale University

2017 – 2024
Department of Spatial Sciences,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague

Education

2020 – present
Department of Spatial Sciences,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
PhD studies in Spatial Sciences
Thesis topic: Beyond Assumptions: Unraveling Data Limitations in Predicting Species Distribution

2016 – 2020
Department of Applied Geoinformatics and Spatial Planning,
Faculty of Environmental Sciences,
Czech University of Life Sciences Prague
PhD studies in Applied and Landscape Ecology
Dissertation 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?

Research Grants

2022 – 2023
Could be species distribution models affected by positional uncertainty in species occurrences still ecologically interpretable?
Grant Agency of the Czech University of Life Sciences Prague
Project leader

2021 – 2023
DivLand - Center for landscape and biodiversity
Czech Republic Technological Agency
Team member

2021 – 2022
Could species distribution models be successfully fitted with simple presence or absence of essential/rare habitats?
Grant Agency of the Czech University of Life Sciences Prague
Project leader

2020 – 2021
Modelling species distributions: Addressing the trade-off between spatial scale and positional accuracy of species occurrences
Grant Agency of the Czech University of Life Sciences Prague
Project leader

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

8/2021 – 12/2021
Jetz Lab
Department of Ecology and Evolutionary Biology
Yale University, USA
Fulbright Scholar

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

2022
Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable
Gábor L.
10th Biennial conference of the International Biogeography Society.
Vancouver, Canada
Talk

2021
Habitat as predictors in species distribution models: Shall we use continuous or binary data?
Gábor L., Šímová P. and Moudrý V.
Early Career Biogeographers Conference, International Biogeography Society.
Virtual
Talk

2021
Positional error in species distribution modeling are not overcome by the coarser grain of analysis
Gábor L. and Moudrý V.
Early Career Biogeographers Conference, International Biogeography Society.
Virtual
Poster

2019
Does positional error affect fine-scale species distribution models?
Gábor L., Moudrý V., Lecours V., Malavasi M., Barták V. and Václavík T.
International Society for Ecological Modelling Global Conference.
Salzburg, Austria
Talk

2019
Contrasting fine-scale environmental preferences of cold-water coral species in the northwest Atlantic
Lecours V., **Gábor L.**, Edinger E. and Devillers R.
18th International Symposium GeoHab.
St. Petersburg, Russia
Talk

2017
The quality of spatial data and its effect on species distribution models
Gábor L. and Moudrý V.
Doctoral Consortium on 3rd International Conference on Geographical Information Systems Theory, Applications and Management, INSTICC.
Porto, Portugal
Talk

Teaching Experience

2016 – 2023 Czech University of Life Sciences Prague

- Seminars of GIS I, II; Computer Technology Utilization; Cartography
- Final thesis supervisor or consultant; over 8 successfully defended Bachelor's or Master's Thesis

List of publications

Publications in Journals with Impact Factor

Gábor L., Cohen, J., Moudrý, V., & Jetz, W. (2024). Assessing the applicability of binary land-cover variables to species distribution models across multiple grains. Landscape Ecology, 39(3), 66.

Rocchini, D., Nowosad, J., D'Introno, R., Chieffallo, L., Bacaro, G., Gatti, R. C., ... **Gábor L.** ... & Thouverai, E. (2023). Scientific maps should reach everyone: The cblindplot R package to let colour blind people visualise spatial patterns. *Ecological Informatics*, 76, 102045.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A. M., Bazzichetto, M., ... **Gábor L.** ... & Malavasi, M. (2023). A quixotic view of spatial bias in modelling the distribution of species and their diversity. *npj Biodiversity*, 2(1), 10.

Gábor L., Jetz, W., Zarzo-Arias, A., Winner, K., Yanco, S., Pinkert, S., ... & Moudrý, V. (2023). Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable. *Ecography*, e06358.

Moudrý, V., Keil, P., **Gábor L.**, Lecours, V., Zarzo-Arias, A., Barták, V., ... & Šímová, P. (2023). Scale mismatches between predictor and response variables in species distribution modelling: A review of practices for appropriate grain selection. *Progress in Physical Geography: Earth and Environment*, 47(3), 467-482.

Moudrý, V., Cord, A. F., **Gábor L.**, Laurin, G. V., Barták, V., Gdulová, K., ... & Wild, J. (2023). Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. *Diversity and Distributions*, 29(1), 39-50.

Zarzo-Arias, A., Penteriani, V., **Gábor L.**, Šímová, P., Grattarola, F., & Moudrý, V. (2022). Importance of data selection and filtering in species distribution models: A case study on the Cantabrian brown bear. *Ecosphere*, 13(12), e4284.

Gábor L., Jetz, W., Lu, M., Rocchini, D., Cord, A., Malavasi, M., ... & Moudrý, V. (2022). Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. *Methods in Ecology and Evolution*, 13(10), 2289-2302.

Moudrý, V., Gdulová, K., **Gábor L.**, Šárovcová, E., Barták, V., Leroy, F., ... & Prošek, J. (2022). Effects of environmental conditions on ICESat-2 terrain and canopy heights retrievals in Central European mountains. *Remote Sensing of Environment*, 279, 113112.

Gábor L., Šímová, P., Keil, P., Zarzo-Arias, A., Marsh, C. J., Rocchini, D., ... & Moudrý, V. (2022). Habitats as predictors in species distribution models: Shall we use continuous or binary data?. *Ecography*, 2022(7), e06022.

Marešová, J., Gdulová, K., Pracná, P., Moravec, D., **Gábor L.**, Prošek, J., ... & Moudrý, V. (2021). Applicability of data acquisition characteristics to the identification of local artefacts in global digital elevation models: Comparison of the copernicus and tandem-x dems. *Remote Sensing*, 13(19), 3931.

Gábor L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M., … & Václavík, T. (2020). The effect of positional error on fine scale species distribution models increases for specialist species. *Ecography*, 43(2), 256-269.

Gábor L., Moudrý, V., Barták, V., & Lecours, V. (2020). How do species and data characteristics affect species distribution models and when to use environmental filtering?. *International Journal of Geographical Information Science*, 34(8), 1567-1584.

Moudrý, V., Lecours, V., Malavasi, M., Misiuk, B., **Gábor L.**, Gdulová, K., … & Wild, J. (2019). Potential pitfalls in rescaling digital terrain model-derived attributes for ecological studies. *Ecological Infomatics*, 54, 100987.

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

Books

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

Other Publications

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

Rubín, L., Mitáš, J., Dygrýn, J., Šmída, J., **Gábor L.**, & Pátek, A. (2015). Active commuting of the inhabitants of Liberec city in low and high walkability areas. *Acta Gymnica*a, 45(4), 195-202.

Recognizing ChatGPT's Contribution

I would like to acknowledge that the AI tool ChatGPT was utilized as an assistance in organizing my thoughts and refining the text for the introduction and discussion sections of this dissertation. Throughout the writing process, I was mindful of the limitations and potential biases of the AI tool. I approached it critically, carefully evaluating its suggestions and ensuring they aligned with academic standards. I was responsible for critically reviewing and refining the content, ensuring its accuracy, coherence, and adherence to scholarly conventions. Therefore, the intellectual integrity of the work lies with me as the author, with the AI tool serving as a support in the writing process.

Presented insight, ideas, and thoughts are my own and result from extensive research, critical thinking, and scholarly engagement in ecological predictive modeling. More importantly, the references used in this text were independently sourced through reputable academic platforms such as Google Scholar or Web of Science. The proper citation of these references throughout the dissertation demonstrates that they were not generated by the AI tool. While writing and completing my dissertation (summer 2024), there were no explicit regulations or guidelines at the Czech University of Life Sciences or at the Faculty of Environmental Sciences that prohibited using AI tools like ChatGPT for dissertation writing.

Meet the Mind Behind the Words

The author (see Figure 9.1), an data science and ESG expert, has always been fascinated by the profound influence of data on our lives and the world around us. This early fascination sparked a keen interest in exploring various types of data and their potential applications. Consequently, when it came time to choose a topic for Ph.D. studies, he made a deliberate decision to delve into the realm of spatial data quality and types within the field of predictive ecology.

It soon became clear that while numerous studies focused on modeling methodologies and theoretical frameworks, there was a need for deeper exploration into the critical aspects of spatial data quality and the wide range of spatial data types applicable in predicting species distribution.

The presented work represents the culmination of years of academic pursuit and a profound passion for unraveling the intricacies of ecological predictive modeling. The research within these pages strives to shed light on the influence of spatial data quality, challenge commonly held assumptions, and offer insights that can enhance the accuracy, reliability, and robustness of ecological predictive models. May the ideas presented within these pages ignite curiosity, foster collaboration, and deepen appreciation for the marvelous intricacies of predictive modeling in ecology.

Figure 9.1: Author, Fall 2023

„The dilemma is, as it is often said, correlation does not imply causation. The discovery of a predictive relationship between A and B does not mean one causes the other, not even indirectly. No way, nohow.“

— Eric Siegel

„You can have all of the fancy tools, but if data quality is not good, you‘re nowhere.“

— Veda Bawo

„Torture the data, and it will confess to anything.“

— Ronald Coase

„How do we start to regulate the mathematical models that run more and more of our lives? I would suggest that the process begin with the modelers themselves. Like doctors, data scientists should pledge a Hippocratic Oath, one that focuses on the possible misuses and misinterpretations of their models.“

—Cathy O‘Neil

„With too little data, you won‘t be able to make any conclusions that you trust. With loads of data you will find relationships that aren‘t real. Big data isn‘t about bits, it‘s about talent.“

— Douglas Merrill

„Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.“

— Dan Ariely

Czech University
of Life Sciences Prague

Faculty of Environmental Sciences
Department of Applied Geoinformatics and Spatial Planning
2024
Dissertation Thessis