Physische Geographie
und Landschaftsökologie

Leibniz
Universität
Hannover

A Machine learning-based approach to assess pollinator habitat suitability

Master thesis at the Institute of Physical Geography and Landscape Ecology at the Gottfried Wilhelm Leibniz University of Hannover in the Master's program Landscape Sciences to obtain the degree of Master of Science (M. Sc.)

Author:	Greta Neumann
Matriculation-No.:	10020488
1st Examiner:	M. Sc. Malte Hinsch
2nd Examiner:	Dr. Marcel Ziems

Hannover, September 2024

Abstract

Pollinators play a crucial role in global food security. Due to the enhanced land-use change through extensive agriculture, natural habitats for bees are increasingly endangered. Because of these past trends in landscape alteration, it is essential to map and understand their species-habitat relationships to protect them from continuing declines in species richness and abundance. To address this, machine learning offers promising tools for habitat suitability modelling, overcoming the limitations of traditional models and enabling more accurate predictions. This thesis explores machine learning-based approaches, focusing on Random Forest, to assess habitat suitability of pollinators in Lower Saxony and Hannover in Germany.

The model was created using data on land use and land cover, geomorphology, vegetation, soil and bioclimate. Bee occurrence data was collected from the GBIF database and a spatial bias reduction approach was developed using the *sampbias* R package. The Random Forest model was developed to assess species density. Different resolutions of the input data, the inclusion of temporally more resolved climate data and fine-tuning of the model were tested. For comparison, a Maxent model was used to predict the species distribution in the same study areas. This thesis demonstrated that the suitability of pollinator habitats can be effectively assessed using Random Forest models. Although the Random Forest model showed a trivial fit given R^2, the better fit of the Maxent model given AUC was a consequence of overfitting in urban areas due to the biased occurrence data. The application of a bias raster for targeted filtering of the GBIF data resulted in enhanced model accuracy for both Random Forest and Maxent. Despite the limitations of comparing modelled species densities with Maxent species distribution, the Random Forest model was able to predict bioclimatic factors, including temperature, precipitation and solar radiation, as being particularly important and in line with documented bee habitat requirements. These findings underscore the potential of Random Forest for habitat suitability modelling and provide a foundation for further refinement.

I Table of Content

II List of Figures

III List of Tables

IV List of Abbreviations

AI	Artificial Intelligence
AUC	Area under the receiver-operator curve
ARIES	ARtificial Intelligence for Environment & Sustainability
BKG	Federal Agency for Cartography and Geodesy
CLC	CORINE Land Cover
CLMS	Copernicus Land Monitoring Service
CRS	Coordinate reference system
DLR	German Aerospace Center
DTM	Digital terrain model
DWD	Germany's national meteorological service
EEA	European Environment Agency
ESTIMAP	Ecosystem Service Mapping Tool
GBIF	Global Biodiversity Information Facility
GIS	Geographic information systems
GLM	Generalised linear model
HSM	Habitat suitability model
InVEST	Integrated Valuation of Ecosystem Services and Trade-offs
LBEG	State Office for Mining, Energy, and Geology
LGLN	State Office for Geoinformation and Land Surveying Lower Saxony
LULC	Land use and land cover
MAE	Mean absolute error
Maxent	Maximum Entropy
ML	Machine Learning
MMU	Minimum mapping unit
MSE	Mean squared error
NDVI	Normalised difference vegetation index
OOB	Out-of-bag
R^2	Coefficient of determination
RF	Random Forest
RMSE	Root mean square error
SD	Standard deviation
SDM	Species distribution modelling
Poll4Pop	Pollinator 4aging and population dynamic model

1 Introduction

As the world's population grows, pollinators play an important role in meeting the increasing demand for food security (Khalifa et al. 2021). The production of about 70% of the world's major crops depends, at least to some extent, on animal pollination. The western honeybee (*Apis mellifera*) is the world's most important crop pollinator (Klein et al. 2007), but wild bees are also essential for crop production. Changes in agriculture, such as the expansion of fields, the loss of semi-natural habitats and the increased use of agrochemicals, are reducing the abundance and diversity of wild bees, thereby reducing pollination services for crops (Häussler et al. 2017). As a benefit to human well-being, wild pollination is an important ecosystem service. Due to declines in species richness and abundance, this service is widely considered to be at risk and may therefore lead to a decline in plant diversity in the long term (Klein et al. 2007).

To address the challenges pollinators are facing, it is necessary to identify and protect suitable habitats. Mapping pollinator habitats is a fundamental step in this process. In the field of ecosystem service assessment, models are used in a variety of ways. Models of the natural environment are applied to assess the structure and function of ecosystems. Land use and land cover (LULC) models are a particularly well-known example of this (Burkhard & Maes 2017). An exemplary tool for the assessment of ecosystem services in Europe is ESTIMAP (Ecosystem Service Mapping Tool). ESTIMAP incorporates a variety of models, including one that addresses crop pollination (Zulian et al. 2014). The primary input data for the ESTIMAP pollination model is LULC data, complemented by expert assessments of the environmental capacity to support insect pollinators (Hinsch et al. 2024). Another common model is the InVEST (Integrated Valuation of Ecosystem Services and Trade-offs) crop pollination model, which generates pollinator abundance and supply indices. The model requires LULC data and various habitat parameters such as estimated nesting site and floral resource availabilities (Wentling et al. 2021). As these models are based on expert judgement, they may be biased towards certain species or groups of species, due to the experience and expertise of the experts involved (Perennes et al. 2021).

Another method used to determine how species respond to changing environmental parameters is species distribution modelling (SDM) (Burkhard & Maes 2017). These models, which include habitat suitability models (HSM), ecological niche models and habitat distribution models, quantify the relationship between species and their environments. HSMs play a vital role in ecological research, linking species data with environmental predictors using response curves derived statistically or theoretically. Advances in science, computing power, geographic information systems (GIS), and remote sensing have greatly improved our ability to model and predict species distributions, which is increasingly important in the face of the ongoing biodiversity crisis. However, when using species data from large web databases, it is crucial to be aware of potential sampling biases and uncertainties in species identification, as these databases often lack controlled sampling designs (Guisan et al. 2017).

Machine learning (ML) is a valuable approach to overcome limitations in mapping and modelling ecosystem services using technological advances. It enables researchers to solve challenges more effectively such as data availability, uncertainty estimates and the integration of socio-ecological aspects. Using ML algorithms, researchers can process large, disparate socio-ecological datasets, leading to more accurate and comprehensive modelling and mapping of ecosystem services. However, despite its potential, the application of ML in ecosystem services research still faces challenges and requires further development (Manley & Egoh 2022).

ML is a rapidly growing field that plays a key role in the broader field of artificial intelligence (AI). AI was originally created with the goal of giving computers the ability to think and reason like humans. ML, a subset of AI, focuses on enabling machines to automatically learn from large datasets and use the learned patterns to make predictions about new, unseen data. Deep learning, in turn, is a specialised subset of ML that uses deep neural networks to enhance the learning process. If AI is thought of as an overarching concept, ML can be considered the cognitive process by which computers learn from data, and deep learning is an advanced, efficient method within this learning process (Yuan 2023).

ML encompasses different approaches to modelling and prediction, with three main types: supervised learning, unsupervised learning and reinforcement learning. Supervised learning uses a training set of labelled data points, where each data point contains features and labels. The goal is to learn a hypothesis that can predict the label of new, unseen data points based on their features alone. This approach focuses on minimising the discrepancy between the predicted and true labels, using a loss function to guide the learning process. Unsupervised learning, on the other hand, does not rely on labelled data. Instead, it seeks to identify underlying patterns or structures in the data itself, often through methods such as clustering, where data points are grouped based on similarity, or feature learning, which extracts important features for efficient processing. Finally, reinforcement learning differs from both by influencing future data points through predictions obtained by a hypothesis (Jung 2022).

When using ML approaches to assess the habitat suitability of species, a distinction can be made between classification and regression problems as part of supervised learning. Classification approaches look at presence and absence of species or suitable and unsuitable habitats, while regressions look at the degree of suitability through density or abundance of species (Džeroski 2009). The most common way in the literature is modelling presence and absence points. Due to the lack of absence data in species occurrence web databases, it is a common approach to create pseudo-absence or background points (Guisan et al. 2017). A popular and well-established ML tool for SDM using the approach of compiling background points is Maxent (Maximum Entropy), which has comparable good predictive performance and is particularly easy to use (Merow et al. 2013). The most widely used ML algorithm in ecosystem service research is Random Forest (RF), which can consider both classification and regression problems (Scowen et al. 2021). Because they are robust to overfitting and produce good prediction models, they are increasingly being used in HSM (Howard et al. 2014).

Although widely used in the analysis of presence-only data, the pseudo-absence approach has several weaknesses. In the case of pseudo-absences, new data are generated to fit a constructed model, rather than constructing a model that fits the data, which can lead to potential uncertainty. Another source of error is the interpretability of the results. These approaches model the probability that a location is a presence rather than a pseudo-absence, so their interpretation is sensitive to the number of pseudo-absences and their location. In addition, implementation is problematic as there are many different approaches to how, where and how many pseudo-absences should be selected (Warton & Shepherd 2010).

By describing suitability in terms of occurrence density or abundance in a regression approach, these weaknesses can be overcome. This method is well established in species distribution research as point process modelling (Guisan et al. 2017; Renner et al. 2015; Warton & Shepherd 2010). Research on the assessment of species density using RF in comparison to the use of absence or background points is available but lacking and has mainly focused on bird species (Howard et al. 2014; Kosicki 2017, 2020; Kosicki & Hromada 2018; Mi et al. 2017; Oppel et al. 2012).

In the assessment of the suitability of habitats for pollinators or pollination services, process-based models such as ESTIMAP, InVEST, Poll4Pop (Pollinator 4aging and population dynamic model) and ARIES (ARtificial Intelligence for Environment & Sustainability) are frequently employed, particularly at local and national scales (Gardner et al. 2020; Hinsch et al. 2024; Łowicki & Fagiewicz 2021; Pashanejad et al. 2023). Statistical models, such as generalised linear models (GLM), are also used in this area of research (Herrera et al. 2014; Radzevičiūtė et al. 2021). The establishment of ML has led to the widespread adoption of Maxent as the most widely used approach for modelling distribution of pollinators, especially at national and continental scales (Giannini et al. 2012; Marshall et al. 2015; Moens et al. 2024; Polce et al. 2013, 2014, 2018). Giménez-García et al. (2023) compared a widely used process-based model with ML models and found that ML methods were particularly good at predicting pollination supply on a global scale. Geue & Thomassen (2020) assessed the habitat preferences of two bumblebee species in Bulgaria and Romania using several different ML approaches, including RF and Maxent. Presence-only data and relative abundance were used with sampling data from 44 sites, which improved the understanding of distribution. Rahimi et al. (2021) used GBIF presence-only data to assess the distribution of wild bees in a region of Iran. Methods used included RF and GLM. However, the research focused on the results of habitat suitability itself, rather than the performance and suitability of the models.

The aim of this work is not only to contribute to a better understanding of pollinator habitat suitability, but more importantly to show how it can be captured using ML methods that have already proven successful in other areas. While many studies have focused on process-based models, there is a clear gap in the use of ML techniques at local and regional scales, particularly in Germany. This research aims to fill this gap by developing and testing new methods that could improve the way pollinator habitat suitability is assessed. Given the global decline in

pollinator populations and their essential role in food security and ecosystem health, improving assessment through ML is both timely and necessary. The results of this thesis could lead to more accurate models and a deeper understanding of how pollinators interact with their environment.

Therefore, the methods of this work include the development of a suitable ML approach for the assessment of pollinator habitat suitability at local and regional scales in Germany. This includes the development of a method based on RF and an approach to reduce spatial bias in occurrence data. As part of the method development, different datasets will be analysed and tested to determine the sensitivity of the approach to different geodata and scales. Furthermore, the suitability of RF for assessing the suitability of pollinator habitats will be evaluated. The secondary objective is to determine which parameters significantly influence pollinator habitat suitability. The following research questions will be addressed and answered in the thesis:

1) How is pollinator habitat suitability assessed in the study areas using a machine learning approach?
2) How can the spatial bias in occurrence data be reduced and how does this affect the modelling?
3) Which input variables are highly relevant in the model and therefore also for habitat suitability modelling?

2 Material

This chapter provides an overview of the selected study areas of Hannover and Lower Saxony, as well as the data to be used for modelling. The data is divided into two categories: predictor data, comprising geodata of the landscape, vegetation and climate, and occurrence data of bees.

2.1 Study area

The study area of Lower Saxony is one of the northern federal states in Germany and is the second largest in terms of area with around 47,630 km². Hannover is the capital of this federal state and is located towards the south of the federal state with an area of 204.2 km² (Fig. 1a). The area measurements were calculated using geodata about the administrative areas in Germany from the Federal Agency for Cartography and Geodesy (BKG).

Lower Saxony forms two geomorphologic zones. To the north is the tidal coast of the North Sea with islands, followed by the North German Lowlands. The southern part of the federal state is characterized by the Lower Saxony Uplands and Hills. Hannover is located at the border between these two landscapes (ZÖLLER ET AL. 2017).

Lower Saxony is largely characterized by deciduous forests, the North Sea coast and the raised and transitional bogs in the North German Lowlands (ELLENBERG & LEUSCHNER 2010). The southern part of the Lowlands is mostly covered with loess, which is known for its particularly fertile soils (ZÖLLER ET AL. 2017).

Lower Saxony is located in the warm-temperate climate zone of the mid-latitudes. The study areas have diverse climates due to the transition from maritime to continental influences and the landscape. The North Germany Lowlands experience milder winters and moderately warm summers due to its proximity to the sea, while the southern uplands and hills are colder due to its higher altitude. Hannover represents average climatic conditions of the region (DWD 2018).

According to geodata from the BKG, *artificial surfaces* make up 9.6% of the area, indicating significant urbanization and infrastructure in Lower Saxony. *Agricultural areas* dominate, making up 64.3% of the land, reflecting the state's extensive farming. *Forests and seminatural areas* account for 24%, showing a substantial number of natural landscapes. Wetlands and water bodies both cover around 1% of the area showing the region's relatively small but important water-related ecosystems (Fig. 1b).

The main LULC types in the city core of Hannover are categorised as follows. *Urban areas* are the most frequent, covering 31.4% of the city. *Industrial areas* account for 19%. *Artificial vegetated areas* make up 14.5%, and *forests* cover 13.6%, showing significant green spaces within the city. *Arable land* represents 9.8%, and pastures cover 5.6%, reflecting some agricultural activity (Fig. 1c).

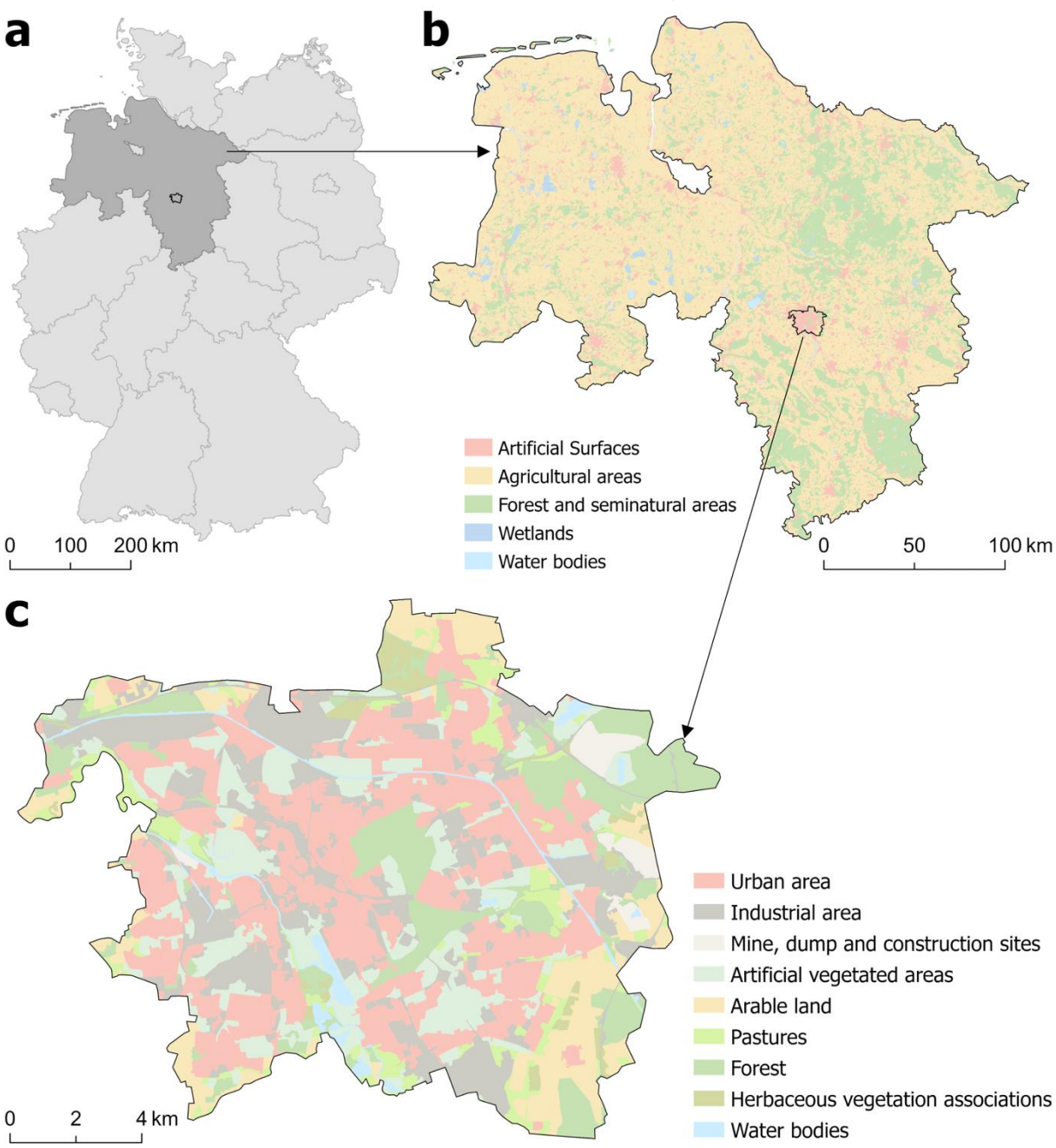


Fig. 1: Study areas within Germany (**a**) (© GeoBasis-DE/BKG 2024), main land cover classes in Lower Saxony (**b**) (© GeoBasis-DE/BKG 2024); land cover and land use of Hannover (**c**) (© GeoBasis-DE/BKG 2024)

2.2 Input data

The input data consists of various variables describing the landscape and climate of the study areas with different spatial resolutions. Efforts were made to select datasets with the highest resolution, which are available as open data and easy to process. The complete list of all geo-data used is provided in Tab. A 1.

2.2.1 Land use and land cover data

Numerous studies indicate that the abundance and diversity of wild bees are strongly influenced by landscape structures and the associated LULC types (MEIER ET AL. 2021). Therefore, the primary input data for assessing pollinator habitat suitability or pollination services in other studies often includes LULC data (GALLANT ET AL. 2014; HINSCH ET AL. 2024; PERENNES ET AL. 2021). For this research, three different datasets have been utilized, each offering a different classification and level of detail.

The CORINE Land Cover (CLC) is the oldest database of the Copernicus Land Monitoring Service (CLMS). The objective is to standardise land data in Europe to support environmental policies. Its first inventory was in 1990 and is currently updated every six years. Managed by the European Environment Agency (EEA) and implemented by national teams, the project now includes 39 European countries, covering nearly 6 million km^2. The CLC uses high-resolution satellite images, topographic maps, orthophotos, and ground survey data for mapping. A standardised nomenclature classifies the data in three levels and 44 classes. The minimum mapping unit (MMU) is 25 ha, and the minimum mapping width is 100 m (BÜTTNER ET AL. 2021).

For the area of Germany, the more detailed dataset CLC5 exists, which provides a vector format description of LULC according to the CLC nomenclature. It is based on Germany's 2018 Land Cover Model (LBM-DE2018). Its MMU of 1 ha is generalised to 5 ha for CLC5. Since 2012, it updates every three years, but 2018 is the most current available dataset (BKG 2022).

In Lower Saxony, there are 30 of the 44 different classes present. The main types are *non-irrigated arable land* with almost 44%, and *pastures* with 20%. *Coniferous forests* cover almost 13%, and *broad-leaved forests* cover 7.5%. *Discontinuous urban fabric* is the fifth largest type with 6.5%. In the city of Hannover, 18 classes are present. *Discontinuous urban fabric* has the largest share in area with 26.8%, followed by *industrial or commercial units* with 15.5%. The classes *broad-leaved forest, sport and leisure facilities*, and *non-irrigated arable land* each have an area share of around 10% (Fig. 2a).

The next used LULC dataset is a new product by CLMS. The CLC+ Backbone (CLCBB) is a high-resolution raster product, providing limited, but robust and consistent thematic detail. The dataset contains 11 different LULC classes and has a spatial resolution of 10 m. It is primarily based on Copernicus Sentinel satellite imagery (EEA 2022; PROBECK ET AL. 2021). The data for the reference year 2018 was used. However, a dataset for 2021 was also published during the process of this work.

In Lower Saxony, there are 8 of the 11 classes present. The two main classes are *periodically herbaceous* with 38% and *permanent herbaceous* with 25%. *Woody broadleaved deciduous trees* have an area share of 15% and *woody needle leaved trees* of 13%. *Sealed* areas account for 6% in Lower Saxony according to the CLCBB of 2018. In the city of Hannover, the same classes are present. With 42% *sealed* areas are the most frequent. The second largest area share is *woody broadleaved deciduous trees* with 29%. In Hannover, there are 14% *permanent herbaceous* and almost 8% *periodically herbaceous* areas (Fig. 2b).

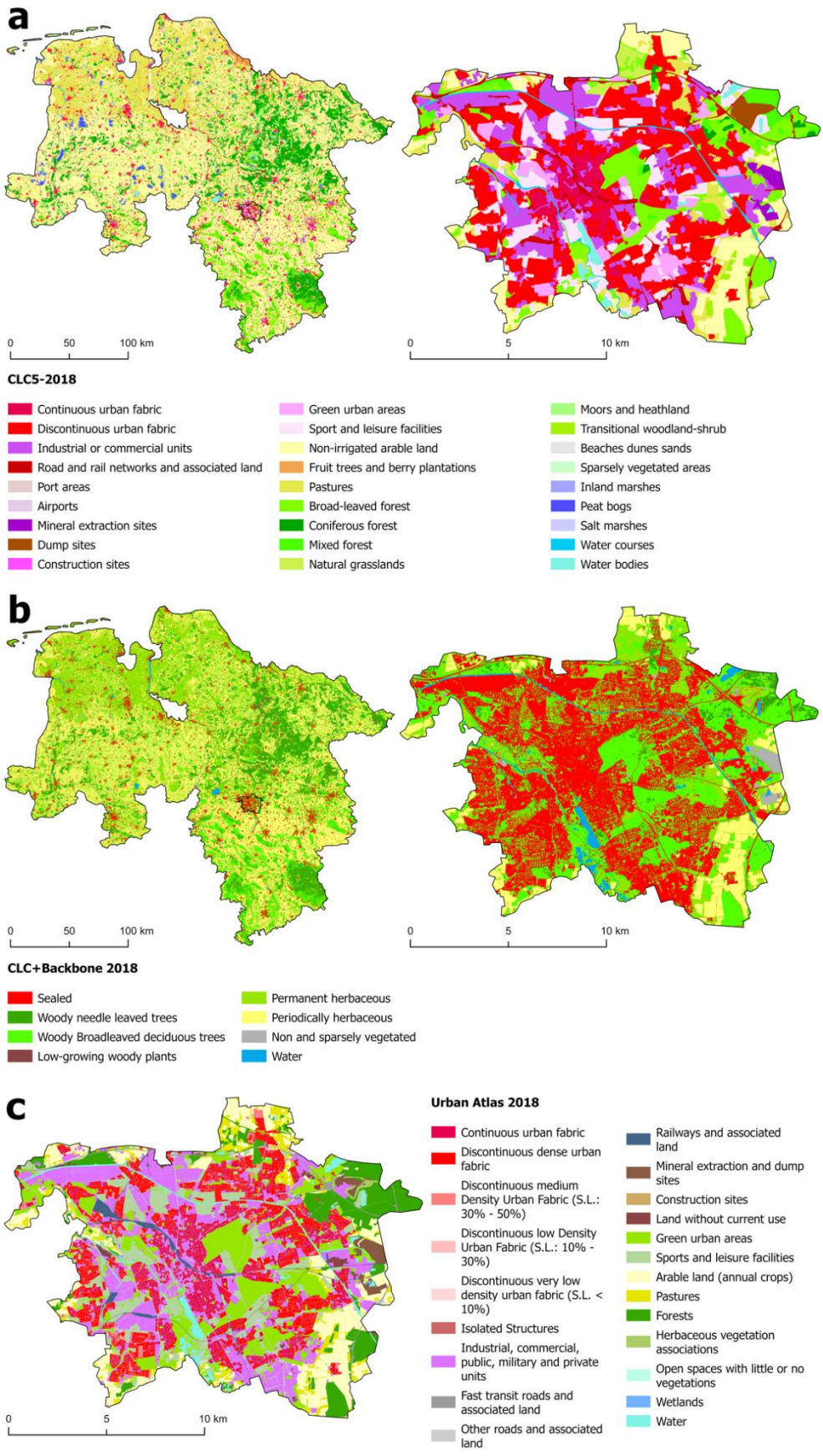


Fig. 2: Land use and land cover data; CLC5-2018 Lower Saxony and Hannover (**a**) (© GeoBasis-DE/BKG 2024); CLCBB Lower Saxony and Hannover (**b**) (© EU, CLMS 2018, EEA); Urban Atlas Hannover (**c**) (© EU, CLMS 2018, EEA)

The last LULC dataset used is the Urban Atlas (UA). The UA provides high-resolution LULC maps of urban areas, covering nearly 800 European cities with more than 50,000 inhabitants by 2012 and 2018. The project started in 2006 with 300 cities with more than 100,000 inhabitants. Each map covers the city and its surroundings. The classification system, derived from CLC, includes 27 classes in 5 thematic groups. The MMU of this dataset is 0.25 ha in urban areas and 1 ha in rural areas. The dataset is derived from satellite image interpretation, topographic, and LULC data (EUROPEAN UNION 2020).
In Hannover, 22 of the 27 classes of the UA are present. The most frequent class is *industrial, commercial, public, military and private units* with 17%. *Green urban areas* and *discontinuous dense urban fabric (S.L.: 50% - 80%)* both have around 13%. S.L. describes the degree of soil sealing. The area share of *arable land (annual crops)* is almost 11%. 9.5% of the area in Hannover is *continuous urban fabric (S.L.: > 80%)* and 8% *forests* according to the UA (Fig. 2c).

2.2.2 Digital terrain model

To provide a clearer description of the landscape, a digital terrain model (DTM) was used. This model describes the surface of the terrain without its natural and artificial objects, such as houses or trees, representing the relief (JÄGER & HEIPKE 2020). The State Office for Geoinformation and Land Surveying Lower Saxony (LGLN) provides the DTM as open data with a precision of 1 m. This was selected for the study area of Hannover (LGLN 2019). Since this resolution would be too detailed for the entire area of Lower Saxony, the DTM200 was chosen for this broader region. This is a nationwide dataset with a spatial resolution of 200 m, derived from the DTMs of the surveying authorities of the federal states (BKG 2021). The DTM 200 covers the area of Lower Saxony with elevation values ranging from -15.86 to 962.83 m (Fig. 3a). The DTM 1 for Hannover ranges from 0 to 121.77 m (Fig. 3b).

2.2.3 Normalised difference vegetation index

GALBRAITH ET AL. (2015) state that variables such as the normalised difference vegetation index (NDVI) can potentially be useful for pollinator studies. The NDVI is a simple ratio of reflectance measurements in the red and near-infrared bands, making it a key tool for monitoring vegetation changes (NIGHTINGALE ET AL. 2008).
The dataset Sentinel-2-Vegetation Index from the German Aerospace Center (DLR) describes the NDVI in Germany. The data was derived from Sentinel-2 images and covers the period from the end of June 2015 to the end of September 2017. Google Earth Engine was used to select all Sentinel-2 images with less than 60% cloud cover. A median mosaic of these images was then created for Germany. The dataset has a spatial resolution of 10 m (DLR 2023). The NDVI values for Lower Saxony range from -0.22 to 0.41 (Fig. 3c) and for Hannover from -0.21 to 0.36 (Fig. 3d).

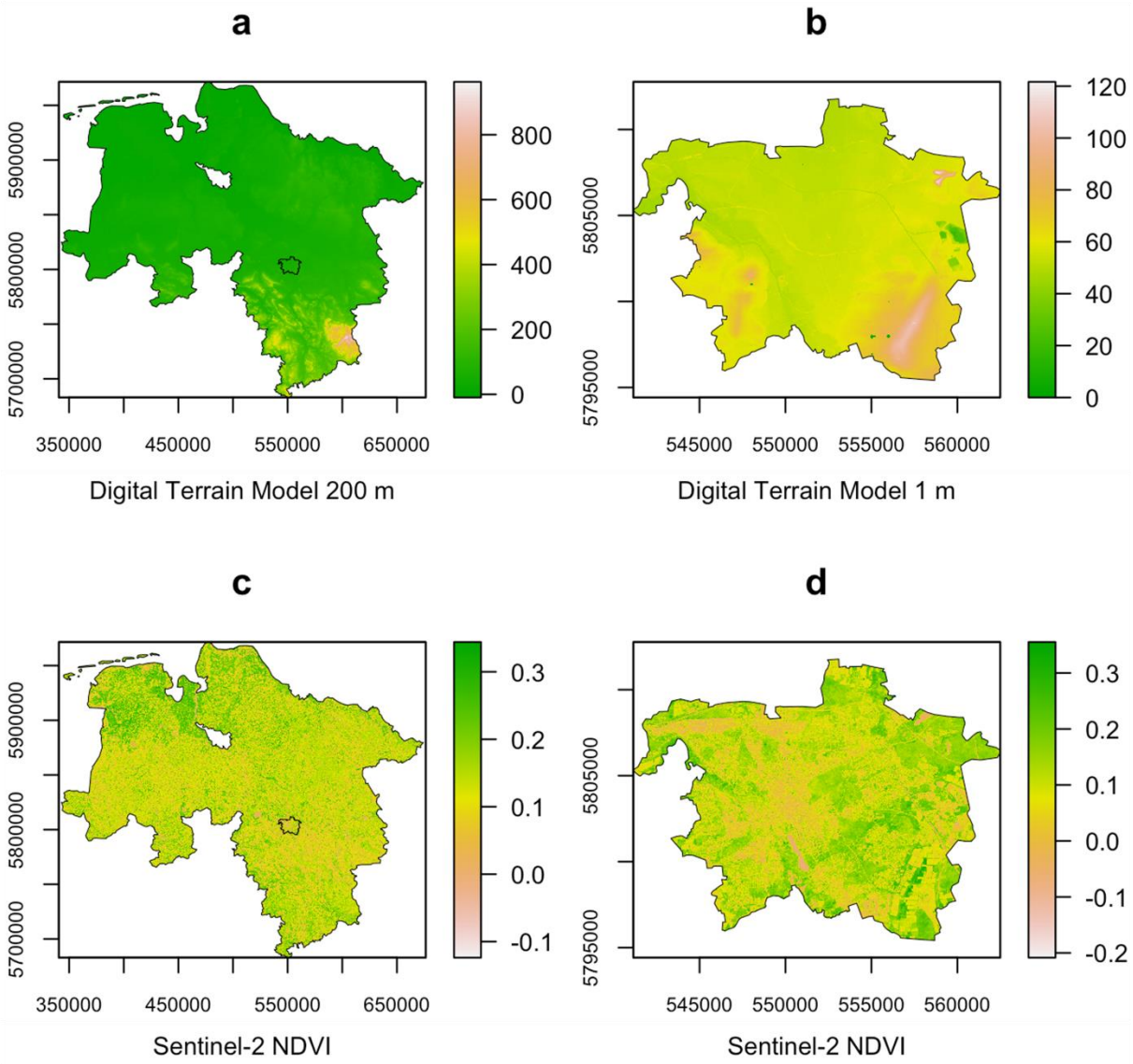


Fig. 3: Digital terrain model 200 m in Lower Saxony (**a**) (© GeoBasis-DE/BKG 2024); Digital terrain model 1 m in Hannover (**b**) (© GeoBasis-DE/LGLN 2024, CC-BY 4.0); Normalized Difference Vegetation Index in Lower Saxony (**c**) and Hannover (**d**) (© DLR Sentinel-2 NDVI 2015 Germany)

2.2.4 Climate data

A combination of temperature and radiation intensity can affect the abundance of bees, as bees become inactive when it falls below a certain threshold (ZULIAN ET AL. 2013). Furthermore, precipitation is also a commonly identified variable that can influence the abundance of bees (KAMMERER ET AL. 2021; MOENS ET AL. 2023).

Germany's national meteorological service (DWD) provides various categories of different climate observations as open data (KASPAR ET AL. 2019). Their database includes historical and current German climate data. It contains various variables from classic meteorological data from near the earth's surface, data from the free atmosphere, the ground and phenological data. The data is gridded and available with a resolution of 1 km x 1 km (KASPAR ET AL. 2013).

In addition to data with very high temporal resolution, such as 5 minutes, hourly, daily and monthly, data with averaged values for seasonal, annual and multi-annual periods are also

provided. Two different temporal resolutions were selected for this study. Multi-annual data covering a period from 1991-2020 and seasonal data to observe how their relevance for the ML model changes with changing temporal resolution.

The four seasons are divided as follows. March, April and May are the first season, June, July and August are the second. September, October and November are the third and December, January and February are the last season. The following six different variables are available as seasonal data: sum of sunshine duration, sum of precipitation, sum of de Martonne drought index, monthly averaged daily minimum, mean and maximum air temperature. All of them were selected for this work. They are shown exemplarily in Fig. 4 for the summer of 2023 in Lower Saxony.

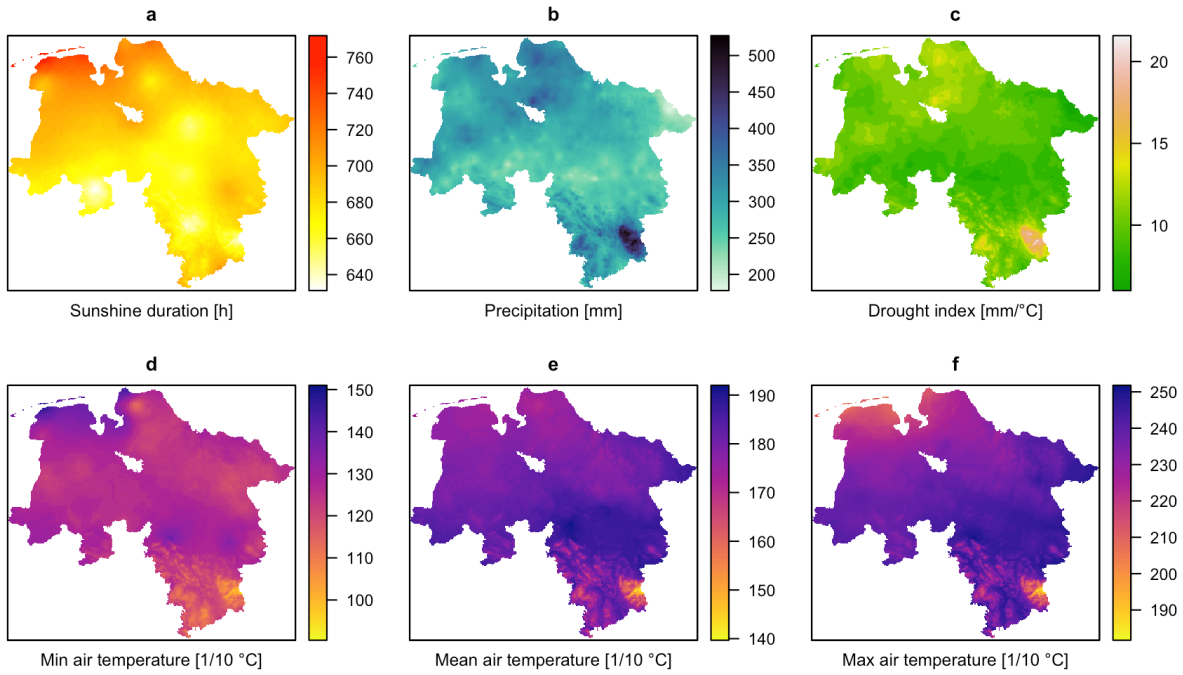


Fig. 4: Exemplary representation of the seasonal climate data in Lower Saxony: Summer 2023; sunshine duration (**a**) (DWD CDC: Seasonal grids of sum of sunshine duration over Germany, version v1.0, 2018); precipitation (**b**) (DWD CDC: Seasonal grids of sum of precipitation over Germany, version v1.0, 2018); drought index (**c**) (Seasonal grids of sum of drought index (de Martonne) over Germany, version v1.0, 2018); min temperature (**d**) (Seasonal grids of monthly averaged daily minimum air temperature (2m) over Germany, version v1.0, 2018); mean temperature (**e**) (Seasonal grids of monthly averaged daily mean air temperature (2m) over Germany, version v1.0, 2018); max temperature (**f**) (Seasonal grids of monthly averaged daily maximum air temperature (2m) over Germany, version v1.0, 2018)

The same variables were selected for the multi-annual period from 1991-2020. However, more than those were available here, which is why the following parameters were included as well: mean sum of global radiation, soil temperature in 5 cm depth under uncovered soil and soil moisture in 5cm depth under grass and sandy loam. The total of 9 parameters are shown in Fig. 5 for Lower Saxony.

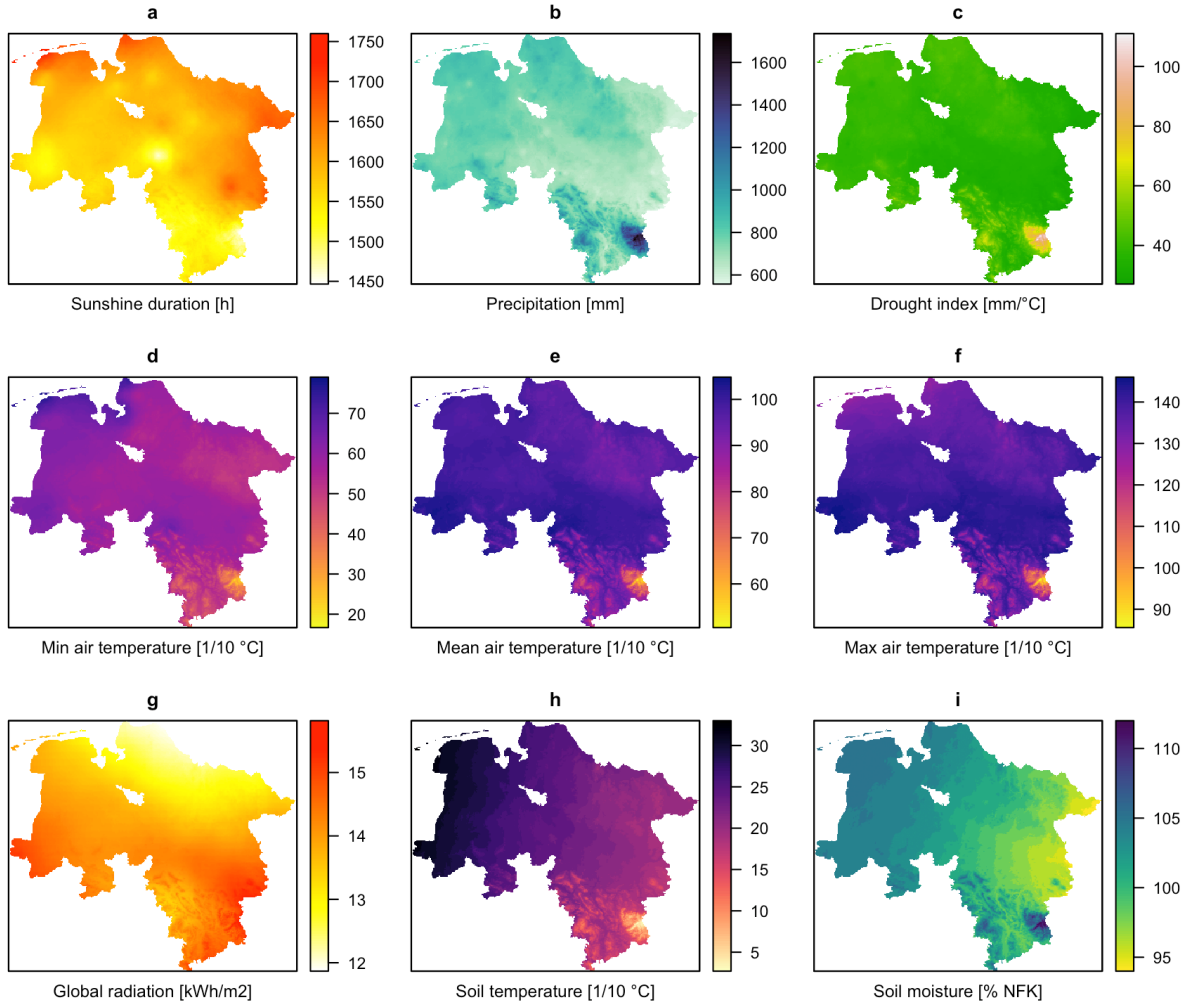


Fig. 5: Multi-annual climate data in Lower Saxony: sunshine duration (**a**) (DWD CDC: Multi-annual grids of annual sunshine duration over Germany 1991-2020, version v1.0, 2018); precipitation (**b**) (DWD CDC: Multi-annual grids of precipitation height over Germany 1991-2020, version v1.0, 2018); drought index (**c**) (DWD CDC: Multi-annual grids of drought index (de Martonne) over Germany 1991-2020, version v1.0, 2018); min temperature (**d**) (DWD CDC: Multi-annual grids of monthly averaged daily minimum air temperature (2m) over Germany 1991-2020, version v1.0, 2018); mean temperature (**e**) (DWD CDC: Multi-annual grids of monthly averaged daily mean air temperature (2m) over Germany 1991-2020, version v1.0, 2018); max temperature (**f**) (DWD CDC: Multi-annual grids of monthly averaged daily maximum air temperature (2m) over Germany 1991-2020, version v1.0, 2018); global radiation (**g**) (DWD CDC: Gridded multi annual monthly mean sums and multi annual yearly mean sum of incoming shortwave radiation (global radiation) on the horizontal plain for Germany based on ground and satellite measurements, Version V003, 2024); soil temperature (**h**) (DWD CDC: Multi-annual grids of soil temperature in 5 cm depth under uncovered soil, version 0.x, 2024); soil moisture (**i**) (DWD CDC: Multi-annual grids of soil moisture in 5cm depth under grass and sandy loam, version 0.x, 2024)

2.2.5 Soil types

Most wild bees are ground nesting species, and the above-mentioned soil temperature can affect the nesting activity (GARDEIN ET AL. 2022). The soil type is another factor that should be considered as soil characteristics are important for many wild bees (MOENS ET AL. 2023).

Information about soil types can be found in the horizon-based data of the attribute tables for soil surface data from the State Office for Mining, Energy, and Geology (LBEG). This table includes details such as horizon number, thickness, and depth. For each horizon, a soil type is

specified. Data points are localized using easting and northing coordinates. The soil texture, defined by its particle size and the relative amounts of sand, silt, and clay is given (EVERTSBUSCH ET AL. 2021).

2.3 Pollinator occurrence data

The most used database for HSMs is the Global Biodiversity Information Facility (GBIF) (GUISAN ET AL. 2017). The international platform provides comprehensive access to digital species occurrence data. Using international data standards enables efficient indexing, searching, and filtering of data (SVENNINGSEN & SCHIGEL 2024). When working with the data, it should be noted that uncertainties in the identification of species may occur, as well as low coordinate accuracy and incomplete or uneven spatial coverage of the actual distribution of a species (GUISAN ET AL. 2017).

The occurrence data from the GBIF database was downloaded in R using the *rgbif* library. In order to obtain as much data as possible, all bee families occurring in Lower Saxony were selected: *Andrenidae, Apidae, Colletidae, Halictidae, Megachilidae,* and *Melittidae.* In total, 46,067 occurrences were found. Duplicate entries and entries with a coordinate uncertainty of over 1,000 m were deleted. Datasets containing NA values were removed. Additionally, since there is a lack of data before the year 2010, these records were also excluded. This resulted in a total of 34,681 occurrences remaining. The spatial distribution and the distribution of families are shown in Fig. 6. It is evident that the family *Apidae* is the most frequently occurring. The occurrences are mainly concentrated in settlement areas. Occurrences have risen sharply over time, especially since 2018. The highest records are in 2022 and 2023.

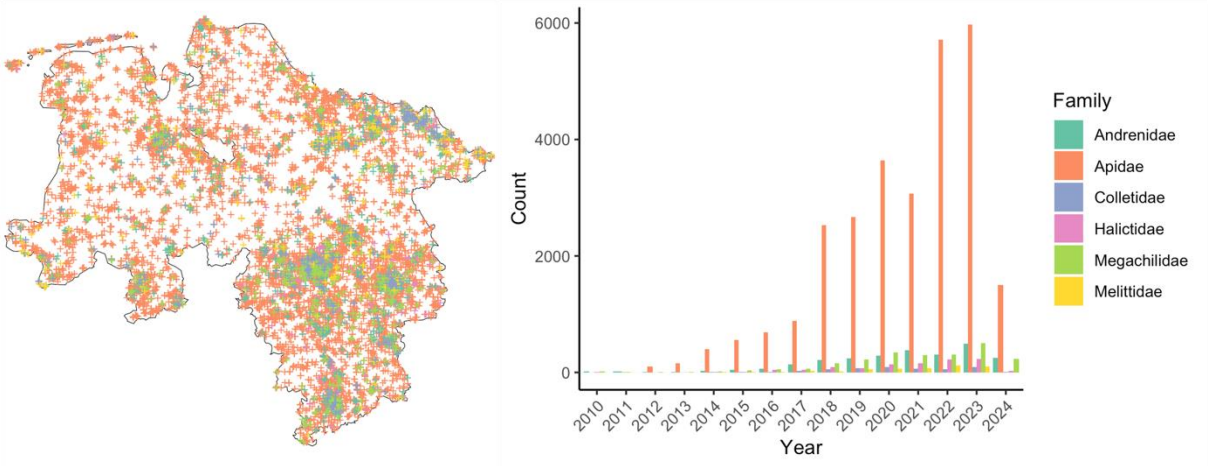


Fig. 6: Bee occurrences in Lower Saxony by family and year (GBIF.org (01 July 2024) GBIF Occurrence Download https://doi.org/10.15468/dl.vnb2vm)

3 Methods

The methods in this thesis are divided into three different parts. The first part comprises the data preparation of the input predictor variables and the occurrence data as training data. Subsequently, two different ML approaches for capturing pollination habitat suitability are introduced. The focus of this work is on the use of the RF approach. Additionally, Maxent is used as a widely known approach to model species distribution using occurrence data and is utilised in this thesis to compare its results with the ones of the RF approach. The methods are implemented in R 3.6.0+ and ArcGIS Pro 3.1.1. In addition, ChatGPT 4o was used to debug and simplify scripts and improve performance.

3.1 Input data preparation

The input data will be prepared in the same way for both ML approaches. For further processing, it needs to be arranged in an identical format, i.e. in the same coordinate reference system (CRS), as raster data with a consistent cell size and the same extent. In addition, further variables are derived and calculated from the input data. The goal is to test as many variables as possible, especially those that are also considered in other research, to find out which of these are rated as particularly important by ML approaches and the stated task. Different cell sizes are tested within the methods to observe how this affects the calculations. In the following, it should be noted that the data preparation steps are repeated with these different cell sizes.

Tab. 1 lists the different variables used for the ML approaches. If they were derived, it is shown which data was used for this. The short name used in the further work is specified. It is also presented for which study area they are employed, as there are some different resolution datasets for Lower Saxony and Hannover. The climate data are not listed here, as no further parameters are derived from them. The processing of this data is described in the end of this chapter.

For the LULC data, one dataset with many classes and one with few classes is selected for each study area to analyse how this affects the calculations. The high-resolution CLCBB is selected for Hannover and Lower Saxony. This dataset is already available as a raster dataset and is therefore only cropped to the respective study areas. As a LULC dataset with several classes, CLC5 is used for Lower Saxony and UA for Hannover, as this offers even more differentiated classes for cities regarding artificial areas. The LULC data is cut to the respective study areas in ArcGIS Pro and saved as a TIFF file using the *Feature to Raster* function.

The ESTIMAP pollination model is a common approach to assess pollination service potential (Hinsch et al. 2024). A central component of the model is the integration of nesting suitability and floral availability, which results from different models that estimate how suitable different landscape types are as sources of food and shelter for insects. The nesting suitability and floral availability parameters refer to different LULC types and are based on the CLC nomenclature. The CLC classes were assigned values from 0 to 1 for both parameters (Zulian et al. 2013).

Tab. 1: Overview of the input variables

Data type	Nr.	Variable	Short name	Source	Study area	Spatial resolution	Data format
LULC	1	CORINE Land Cover 5 ha	CLC5		Lower Saxony	MMU 5 ha	Shape file
		Urban Atlas	UA		Hannover	MMU 0.25 ha	Feature class
	2	CLC+Backbone	CLCBB		Lower Saxony + Hannover	10 m	GeoTIFF
	3	Floral availability	FA	CLC5	Lower Saxony + Hannover		
	4	Nesting suitability	NS	CLC5	Lower Saxony + Hannover		
	5	Distance from forest edges	DisFE	CLC5	Lower Saxony		
				UA	Hannover		
	6	Distance from natural areas	DisNA	CLC5	Lower Saxony		
				UA	Hannover		
	7	Distance from riparian areas	DisRA	CLC5	Lower Saxony		
				UA	Hannover		
	8	Distance from urban areas	DisUA	CLC5	Lower Saxony		
				UA	Hannover		
Geomor-phology	9	Digital terrain model 200 m	DTM200		Lower Saxony	200 m	ASCII
		Digital terrain model 1 m	DTM1		Hannover	1 m	GeoTIFF
	10	Slope 200 m	Slope200	DTM200	Lower Saxony		
		Slope 1 m	Slope1	DTM1	Hannover		
	11	Aspect 200 m	Aspect200	DTM200	Lower Saxony		
		Aspect 1 m	Aspect1	DTM1	Hannover		
Vegetation	12	NDVI	NDVI			10 m	GeoTIFF
Soil	13	Clay amount	Clay	Soil Type	Lower Saxony + Hannover	1 km	ASCII
	14	Silt amount	Silt	Soil Type	Lower Saxony + Hannover	1 km	ASCII
	15	Sand amount	Sand	Soil Type	Lower Saxony + Hannover	1 km	ASCII

The *Join-Field* function in ArcGIS Pro can be used to append the values for floral availability and nesting suitability to the CLC5 classes. The CLC5 vector data for the respective variables is then exported as a TIFF file using the *Feature to Raster* function, specifying the required cell size. The extent of the output file can be determined by specifying a *snap raster*. This is used to limit the file to the needed study area.

Other important parameters for assessing pollinator habitat suitability are forest edges, semi-natural areas and riparian zones since they provide nesting habitats and floral resources (ZULIAN ET AL. 2013). The aim is to calculate the distance per grid cell to the respective areas based on the LULC data. At this point, the distance to urban areas is also calculated to evaluate the spatial bias in the occurrence data due to the higher population density in cities. In consequence, this distance functions as a kind of control value. The corresponding LULC classes are selected for the Lower Saxony study area in the CLC5 and for Hannover in the UA. Processing is carried out in ArcGIS Pro. The size of the forest edges and riparian areas and the selection of the corresponding classes for these two parameters is based on the ESTIMAP pollination model (ZULIAN ET AL. 2013).

To calculate the distance to forest edges, forest areas must first be selected. A negative buffer of 50 m is then created, which represents the forest edges. The *Distance Accumulation* tool is executed for these features to calculate the straight-line distance from the forest edges. When executing this function, the desired cell size is specified again and a mask to cover the required study area. The distance to semi-natural areas is calculated in a similar way: they are selected in the respective LULC dataset and then serve as input data for distance calculation using *Distance Accumulation*. To calculate the riparian areas, a 25 m buffer is first created around wetlands and water bodies. Next, agricultural areas, forests and semi-natural areas are selected and cut to the extent of the riparian areas. The distance can then be calculated again from this output. Finally, the distance to urban surfaces is calculated by selecting artificial surfaces and then calculating the distance from them. The exact CLC5 and UA classes selected for the calculation can be found in Tab. A 2 of the Appendix.

Other important parameters such as slope and aspect can be calculated from elevation models (GUISAN ET AL. 2017). The DTM with a resolution of 1 m, which is used for the study area of Hannover, is provided in tiles of 1 km x 1 km. For further processing, it is merged into a TIFF file using the ArcGIS Pro function *Mosaic to New Raster*. Further processing takes place in R. Using the *raster* library, the DTM files are first cropped to the extent of the respective study area using the *crop* function. The output file is a rectangular raster with the extent of the study area. The *mask* function of the same library is then used to restrict the extent exclusively to the area of the study area. The *terrain* function can then be applied to calculate the slope and aspect using the respective DTMs as input. No further parameters are derived from the NDVI data. This dataset is only cropped to the extent of the study areas using the *crop* and *mask* functions in R.

It is assumed that most ground-nesting bees prefer sandy or sandy-loamy soil (Antoine & Forrest 2021). To further analyse this connection, the respective amounts of sand, clay and silt in the soil are assessed. This data is extracted from the horizon-based data of the attribute tables for soil surface data from the LBEG. The data table was loaded into ArcGIS Pro using *XY Table to Point*. A compilation of nesting data for bees showed that the average minimum nesting depth is 17 cm, while the average maximum depth is 35 cm (Cane & Neff 2011). Therefore, all values with an upper horizon depth of more than 35 cm were removed so that only the relevant horizons were included in the calculation. For each type of soil texture, the mean value of the respective sand, silt and clay content was determined according to AG Boden (2005) (Tab. A 3) and was then attached to the point features using *Join Field*. The points with the proportions of sand, silt and clay were then exported as a shapefile. In R, the respective proportions could then be interpolated as a raster for the study areas using the *Raster* library with the *interpolate* function. The required cell size was used. Since there is no dataset that provides this information, this is one method to get the best possible approximation of the corresponding amounts of sand, clay and silt in the study areas.

The next step is to convert the previously prepared raster files into a standardised format. They are processed in R using the *terra, raster* and *sf* libraries. The outlines of the study areas are read in as shapefiles and defined as the target extent. Using a function, all TIFF files are loaded and, if necessary, transformed into the CRS *UTM ETRS Zone 32 N* (EPSG: 25832) using *project*. The data is cut to the target extent using *crop*. The required cell size is defined. A template raster with the target extent and cell size is created and the TIFF files are resampled to its characteristics using the *resample* function. With *mask*, the data is finally limited to the extent of the study area.

The climate data is prepared in a similar way. However, it is important to note that these are provided in the CRS *DHDN 3-degree Gauss-Krüger Zone 3* (EPSG:31467) and the CRS is not yet linked to the data. Before the data can be transformed into the CRS *UTM ETRS Zone 32 N*, the CRS *Gauss-Krüger Zone 3* first needs to be set. The next steps are the same as the ones above. The climate data and their short names used in the following work are listed in Tab. 2.

Tab. 2: Overview of the climate input variables

Data type	Nr.	Variable	Short name	Study area	Spatial resolution	Data format
Climate	16	Sunshine duration	Sunshine	Lower Saxony + Hannover	1 km	ASCII
	17	Precipitation height	Precipitation			
	18	Drought index	Drought			
	19	Minimum air temperature	TempMin			
	20	Mean air temperature	TempMean			
	21	Maximum air temperature	TempMax			
	22	Global radiation	Radiation			
	23	Soil temperature	SoilTemp			
	24	Soil moisture	SoilMoist			

The steps described in this chapter are repeated several times. One reason for this is that the data is analysed for two different study areas and the other is that different cell sizes are evaluated. The cell sizes 100 m and 500 m are used for the extent of Lower Saxony and the cell sizes 10 m and 50 m for Hannover. The process is therefore repeated a total of four times, which is why the processes were partially automated using scripts. Only the multi-annual climate data is used at first. The seasonal data will be tested at a later stage of this work, but then is processed in the same way. The resulting input data is visualised in Fig. A 1 for 10 m in Hannover and in Fig. A 2 for the cell size of 100 m in Lower Saxony.

3.2 Random Forest

RFs were first introduced by Leo Breiman in 2001 (Breiman 2001). A RF is an ensemble of decision trees built using random samples of data. Instead of using all features to split nodes, each tree uses a random subset of features to find the best split (Fig. 7). This creates many weaker trees, each producing different predictions (Bonaccorso 2017). While individual trees have a high variance, as even a small adjustment changes the tree significantly, RF utilises the combination of many individual trees (Valavi et al. 2021). A key advantage is that the algorithm can be used for both categorical and continuous variables, i.e. for classification and regression problems. Further advantages are that they are relatively fast to train and predict and have a built-in estimation of the generalisation error (Cutler et al. 2012).

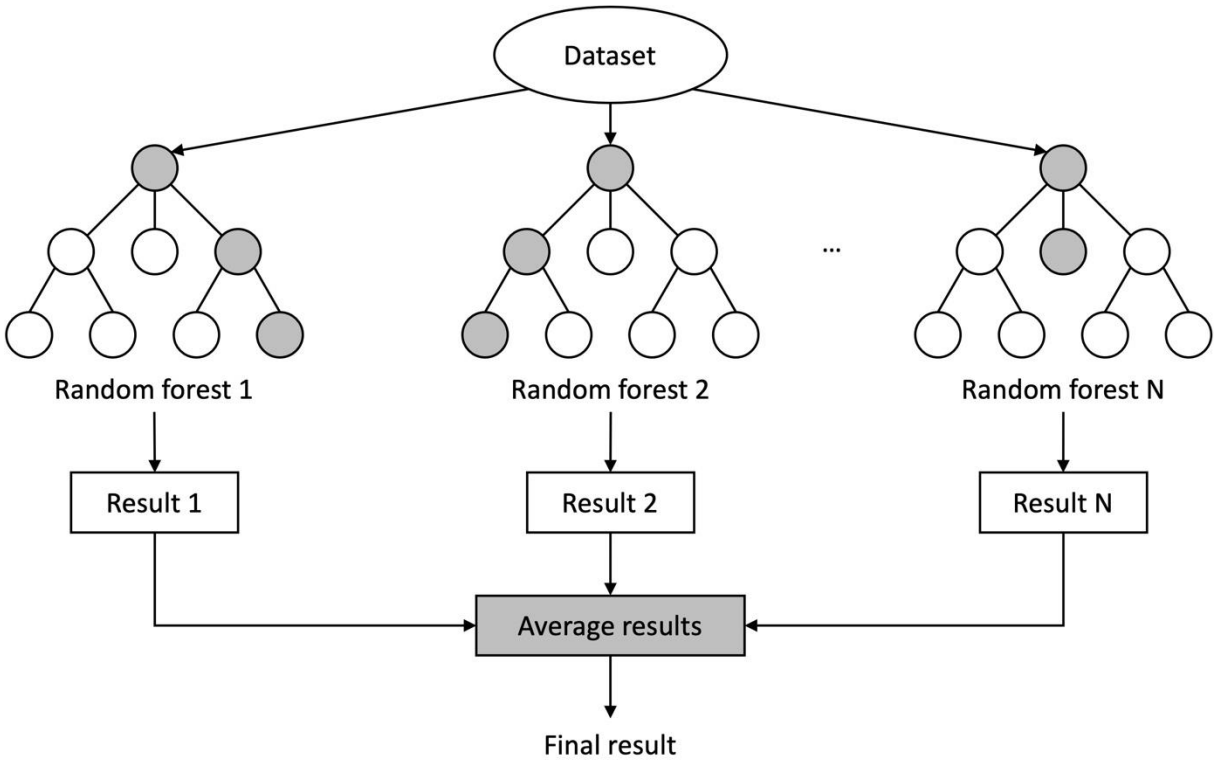


Fig. 7: Flowchart of a random forest algorithm (own figure based on Fu & Qi 2022)

RF is based on the bagging approach (Breiman 2001). Bagging is an ensemble learning technique that improves prediction models by creating multiple new sample datasets from the original training dataset using random sampling with replacement (bootstrap sampling). Each sample is used to train the same model, and the combined predictions reduce variance and overfitting, improving overall performance. RF extends bagging specifically to decision trees. It introduces additional randomness by selecting a random subset of features for each tree, which increases diversity and reduces correlation between trees. This method is advantageous for handling large datasets with many features, requires minimal parameter tuning, and provides estimates of feature importance. RF combines predictions from multiple decision trees to create a more stable, accurate and robust model, making it a popular algorithm in ML (Juwariyem et al. 2024)

3.2.1 Training data preparation

A target variable is required to train the RF model. Since the GBIF data only contains presence and no absence data, this cannot be used directly as a target variable. A common method is to use background samples as a second class. They record the landscape and enable a comparison of the preferred environmental conditions of a species with the existing conditions of the entire landscape observed. Another option is to record background data as locations where the species has not yet been recorded. However, RF models are known to have poorer prediction performance compared to other methods when using background data (Džeroski 2009; Valavi et al. 2021).

In this work, an attempt is made to record habitat suitability without compiling background data. For this approach, the occurrence per grid cell of the input variable is counted. In this way, the abundance and distribution of bees can be approximated. This method changes the problem into one of regression, where the goal is to predict the density or abundance of occurrences within each cell, rather than a binary presence or absence classification. Thus, the target variable is a measure of occurrence records per unit rather than the presence or absence of a particular species (Guisan et al. 2017).

The workflow of preparing the training data is summarised in Fig. 8 and described in detail in the following. The processing is done in R using the libraries *raster*, *sp*, and *dplyr*. The input variables have already been prepared in a standardised format with the exact same spatial extent and resolution. The input variables are stacked to a multi-layer raster object. A template raster is selected from the raster stack to extract cell coordinates. These coordinates are converted into a data frame. The values from the raster stack are extracted and combined with the coordinates of the created data frame. The occurrence data is rasterised using the template grid, while the number of occurrences per cell is counted and saved in the new grid using *rasterize*. The counts of occurrences are extracted from the rasterised data and added as a new column to the data frame. Any rows with missing values or where the count is zero are removed from the data frame to create the training dataset. Columns containing the LULC

values (UA, CLC5, CLCBB) in the training dataset are converted to character type as these values are not numeric.

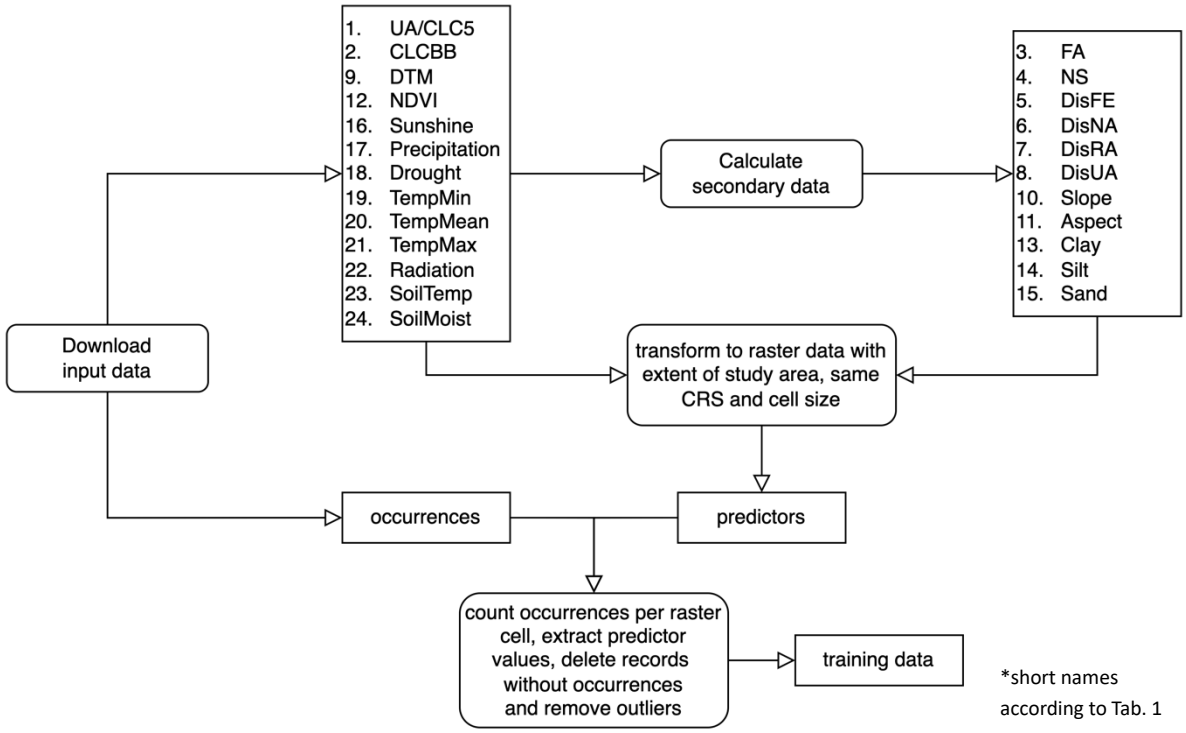


Fig. 8: Training data preparation workflow

The GBIF data are influenced by unevenly distributed sampling efforts (BECK ET AL. 2014). This is also visible in the training data, as there are many very high outliers. To ensure that these do not influence the subsequent calculation, the training data is first sorted by the count of occurrences in descending order. The value at the 5th percentile position is calculated and used to replace all values above it. To allow a comparison between the different models, the relative count value is calculated by dividing the count value by the maximum count value. The training data preparation is repeated several times for the required cell sizes in the respective study areas.

3.2.2 Random forest model

The RF algorithm was implemented in R using the *ranger* (RANdom forest GEneRator) package. The software was first introduced in 2017 as a C++ application and R package and is a comparatively fast and memory efficient implementation of RF particularly suitable for high-dimensional data (WRIGHT & ZIEGLER 2017). The previously prepared data is split into a training dataset and a validation dataset. The most suitable split ratio can vary greatly (JUNG 2022). Initially, the data is split in a ratio of 70% training and 30% test data. The *ranger* function is used to create the RF model. The predictors are set to all columns of the training datasets containing the input variable values. The response variable is the column containing the number of occurrences per raster cell.

The variable importance is measured using the permutation method, which shows how much a predictor variable contributes to the prediction accuracy of the model. This is done by randomly shuffling the values of the variable in question and observing the increase in prediction error or mean squared error (MSE). If the prediction accuracy significantly worsens, the variable is estimated important, as it indicates the model's reliance on that variable for accurate predictions (CUTLER ET AL. 2012). Additionally, a specific random seed is set to ensure reproducibility of the results.

Mtry is defined as the number of variables to potentially split at each node. The default is the square root of the number of predictor variables, rounded down. For 24 predictor variables, this gives an *mtry* of 4 (WRIGHT & ZIEGLER 2017). The default *number of trees* in this package is 500. The *target node size* is defined as the minimum number of observations in a terminal node. Low values result in trees with greater depth, as more splits must be performed to reach the terminal nodes. The default value for regressions is 5 (PROBST ET AL. 2019). The default *splitting rule* for regressions is to use variance, which means that the algorithm chooses the splits that maximise the reduction in variance (WRIGHT & ZIEGLER 2017).

In RFs, each tree is built using a bootstrap sample of the training data. The data that is not used in the construction of a particular tree is referred to as the out-of-bag (OOB) data for that tree. For each observation, its value can be predicted using only the trees for which that observation was OOB. The OOB prediction error is calculated as the MSE between the OOB predictions and the actual observations (CUTLER ET AL. 2012). In addition, the *ranger* function calculates R^2_{OOB}, which measures how well the RF model explains the variance of the target variable using OOB predictions (CHICCO ET AL. 2021).

Next, the *importance* function is used to extract the variable importance of the previously computed RF model. The *predict* function makes predictions on the test dataset using the trained RF model. The results are stored in a column with the test data. Combining the predictions with the test data allows a direct comparison between predicted and actual values, which is essential for further model validation.

The RF models were validated using mean absolute error (MAE), MSE, coefficient of determination (R^2) as well as root mean square error (RMSE) and standard deviation (SD). MAE measures the sum of the absolute errors divided by the sample size and MSE measures the average squared difference between the estimated values and the actual value. The RMSE is calculated by taking the square root of the MSE. Lower values of the three measures indicate better model performance, with the best value being 0. The disadvantage of these measures is that their values go to infinity and therefore as a single value do not say much about the performance of the regression. It is suggested to use R^2, since it tends to be more informative and truthful. R^2 quantifies the proportion of the variance in the dependent variable that is explained by the independent variable. A value of 1 indicates a perfect fit, while 0 corresponds to a trivial fit. Negative values indicate a poor fit. The SD is a measure of the variation around the mean (CHICCO ET AL. 2021).

To best assess pollinator habitat suitability in both study areas, the RF model is applied in several iterations with different settings. The workflow is shown in Fig. 9. Firstly, different cell sizes are tested during data preparation. This means that the validation parameters and permutation importance are considered and compared to select the best cell size for each study area. An attempt is then made to minimise the spatial bias in the occurrence data. Various settings are also tested here. In the third iteration, the periodic climate data is replaced by seasonal data. This involves comparing different time periods. Finally, the model is fine tuned. Input data with low significance is removed and different hyperparameter settings are tested. The methods of these iterations are explained in detail in the following chapters.

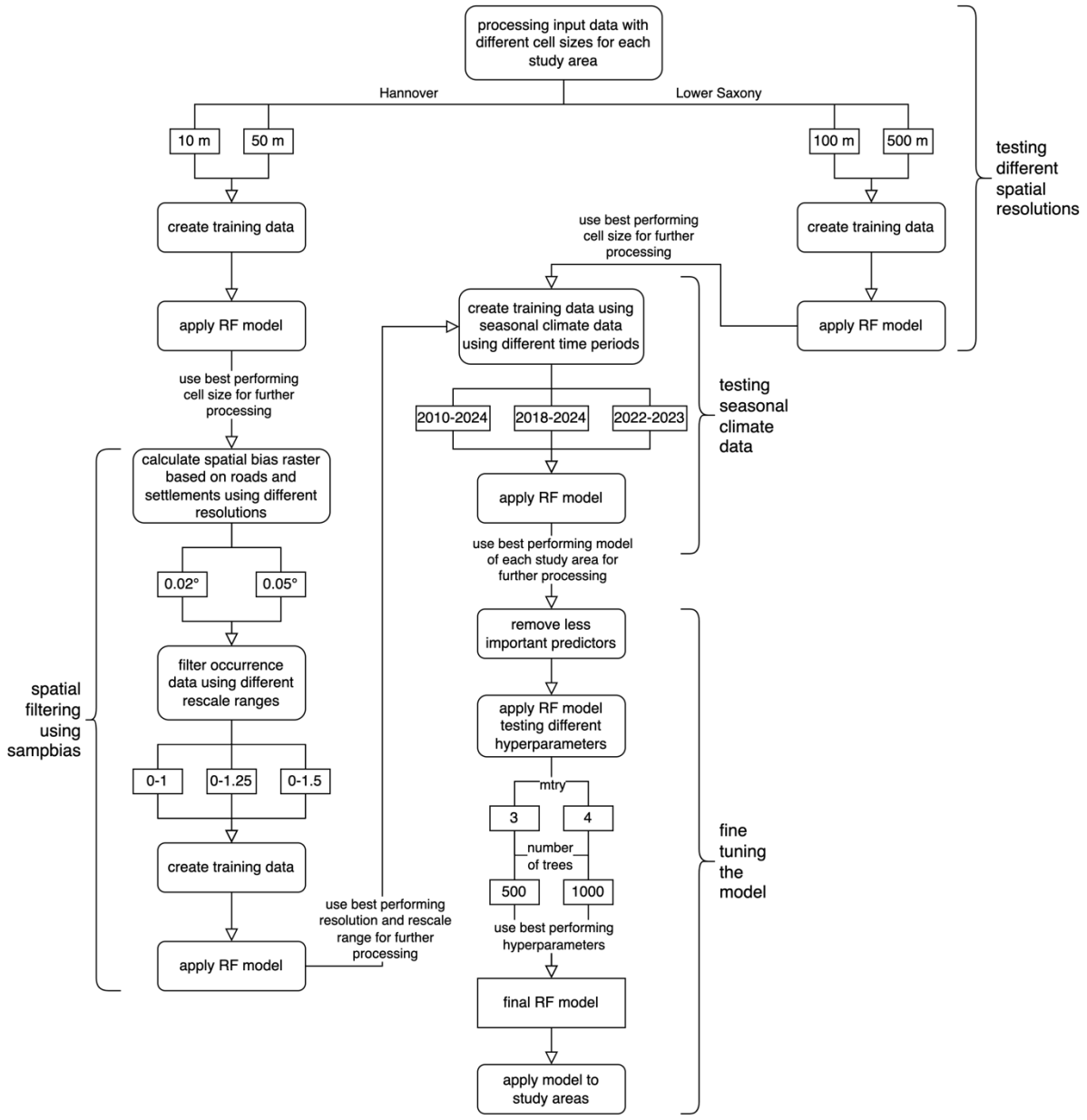


Fig. 9: Workflow of developing the RF model

3.2.3 Spatial filtering using sampbias

Geographic sampling bias occurs when the effort to collect data is unevenly distributed across a study area, often due to physical accessibility factors such as roads, which influences where sampling takes place. This bias is primarily driven by human accessibility, leading to most species observations being made in or near cities and along roads or paths. The biased sampling results can reduce the accuracy of models estimating species distributions. The *sampbias* R package is designed to quantify and address accessibility bias in species occurrence datasets using a Bayesian approach. The key functionalities of *sampbias* include the evaluation of the accessibility bias within a given dataset and the visualisation on how the bias is spatially distributed across the study area. The package can analyse any multi-species occurrence dataset against any geographic gazetteer and allows direct inputs from GBIF. The output of *sampbias* includes measures of sampling rates across space, allowing for comparisons between different bias factors such as roads and rivers (Zizka et al. 2021).

In order to capture the spatial bias of the occurrence data, geodata of settlement areas and roads are extracted from the detailed vector dataset Basis-DLM of the LGLN (© GeoBasis-DE/LGLN 2024, CC-BY 4.0). The object-structured dataset describes the landscape of Lower Saxony as polygons, lines and points, which are described and classified by attributes (LGLN 2024). In ArcGIS Pro, settlement areas were dissolved and those with an area of over 100 ha were selected. To identify the spatial bias, large areas with many inhabitants are particularly relevant. Important roads (motorway, state road and national road) were also selected. Both layers were exported as a shapefile for further processing in R.

The spatial bias is particularly noticeable in Lower Saxony. The Hannover study area serves primarily as a comparative value for a region without this spatial bias. The attempt to filter the bias is therefore only carried out for Lower Saxony.

Using the *terra* library, the shapefiles of the settlements, roads and the region of Lower Saxony are loaded and transformed to CRS *World Geodetic System 1984* (EPSG:4326) since *sampbias* is working with this CRS. A named list of gazetteers is created, containing the roads and settlements shapefiles. This list is used as an input for the bias calculation to consider these features in the sampling bias analysis. The *calculate_bias* function of the *sampbias* library is used to compute the sampling bias, considering the occurrence data, a specified resolution in degrees, the gazetteers list, and the restricted study area of Lower Saxony. The resolutions 0.02° and 0.05° are used to observe how different levels of resolution affect the filtering process.

The *project_bias* function is used to project the sampling bias results, which are then visualized using the *map_bias* function. The type of mapping chosen is *sampling_rate*. A raster representing the sampling bias, derived from the combined effects of roads and settlements, is generated and saved as a TIFF file for later comparison.

A filtering process was applied to remove possible sampling bias in the occurrence data. The process aims to remove points in regions with high sampling bias. The occurrence data is converted to a spatial object using the Longitude and Latitude of the occurrence data. The CRS of

the occurrence data is transformed to match that of the bias raster. For each occurrence point, the corresponding raster cell ID and bias value are extracted. This information is critical for determining the sampling bias associated with each occurrence.
A function is defined to randomly remove points from areas with high sampling bias. The function groups the data by the raster cell IDs. This ensures that the bias adjustment is performed separately for each spatial cell. Within each group, the function performs three steps. If the bias value is missing, it is replaced with 0, to ensure that missing values do not affect the calculation. The bias value is rescaled to a range of 0 to 1 using the *rescale* function. This value represents the proportion of points to remove from each group. The range is later adjusted to 0-1.25 and 0-1.5 to test filtering with varying intensity. The number of points to remove is calculated by multiplying the total number of occurrences in the group by the mean of the rescaled values.
The function then ungroups the data and re-groups it by the cell column and the calculated number of points to remove. This ensures that the filtering operation is correctly applied to each cell. Within each group, points are removed based on the calculated number. The defined function is applied to the occurrence data, resulting in a filtered dataset with reduced sampling bias. The filtered data is converted back to a data frame, including original coordinates, for further analysis. This data can then be processed again with *sampbias* to visualise the reduction of the sampling bias.

3.2.4 Comparison of seasonal and periodic climate data

The next iteration will test the effect of more precise climate data on the performance of the model. In the first runs, averaged climate data for 1991-2020 was used. As the climate varies over the year and there are significant differences, particularly within the seasons, the seasonal climate data will be used for the training data in this step. The seasonal data is available for sunshine duration, precipitation, drought index and minimum, mean and maximum air temperature. The periodic data is still used for the variables global radiation, soil temperature and soil moisture. The seasonal data was also prepared as described above so that it is available in the same extent, resolution and CRS as the other input data.
However, the training data is then prepared somewhat differently, as the seasonal data must be appended to the corresponding years and seasons. The process is carried out in R. The climate data is again analysed for both study areas. The data that previously provided the best result is processed. For Lower Saxony, the occurrence data that produced the best result during spatial filtering is used. For Hannover, the data with the cell size is used, which produced a higher accuracy of the model in the first iteration.
The corresponding occurrence data is loaded and converted to a spatial feature object with coordinates transformed to the CRS of the input raster data. A function is defined to categorise the occurrence months into seasons represented by the code which was also used to identify the season in the file name of the climate data. The result is a new column in the occurrence data frame, which contains the season code for each occurrence based on its month. A raster

template is created representing the input data. The *st_coordinates* function of the *sf* library is used to extract the coordinates from the occurrences data frame. Based on these coordinates and the raster template the function *cellFromXY* of the *raster* library is utilised to assign each occurrence to a cell in the raster template based on its coordinates in form of cell IDs. The data is then grouped by the cell ID, year, and season to count occurrences per year, season and raster cell.

In the filenames of the seasonal climate data information about climate variable, season and year is named. A function is defined to extract this information. The climate data files are loaded, and their values are extracted and combined with the occurrence data based on matching year and season using the *extract* function of the *terra* library. The extracted climate data is merged with the occurrence data based on the cell IDs, years and seasons.

Finally, a multi-layer raster object containing the remaining input variables is created using the *stack* function of the *raster* library. The values from the raster stack are converted into a data frame and combined with the extracted coordinates for each cell in the raster stack. The raster values are merged with the previous results based on the coordinates. Like before any rows containing missing values are removed, the columns containing LULC values are converted to character type, outliers are eliminated by calculating the fifth percentile position of the count values and replacing all higher values by its result. Again, the relative count value is calculated and used.

When using seasonal climate data, a temporal bias is added. Fig. 6 shows that the number of occurrences has increased strongly since 2018. However, this is probably because more people have sampled than more bees being present. To consider this temporal bias, the seasonal data is analysed using three settings. In the first, the sightings of all years are considered. Then the occurrences from 2018 to 2024 are used, as the number has increased strongly here. Finally, a run is made exclusively with the data from 2022 and 2023, as these years had by far the most sightings which is additionally quite similar for both years.

3.2.5 Fine-tuning of the model and application to the study area

By analysing the permutation importance, statements can be made about the relevance of the different predictor variables in capturing pollinator habitat suitability in the study areas. To increase the accuracy of the model, the variables with low importance are removed from the training data. The selection of variables can be different for the two study areas. These are then simply removed from the training data frame for further computation. The aim is to remove data that is of minor importance so that it has no impact on the calculation.

Removing variables can reduce the size of mtry. The hyperparameters are also adjusted in some cases to achieve better model performance. The reduced mtry, as well as the previous value of 4, is utilised. Furthermore, in addition to the default number of trees of 500, 1,000 are tested, as this value is also frequently employed as the default in other implementations (PROBST ET AL. 2019). Other parameters are not tested or changed as this is not the focus of this

work. This means that a further four runs are carried out for both study areas based on the previous best practice runs.
Finally, the hyperparameters that produced the best results are selected. The model is applied to assess the entire study area. In this case, the training dataset includes all occurrence data and the corresponding values of the input variables. A new data frame is created for the test data, which contains all coordinates of the input raster data with the respective values. The model is trained with the selected parameters and then applied to predict the created test data. Subsequently, a predicted count value corresponding to the training data is appended to the test data. The previously extracted coordinates can then be used to create a raster file that maps this value in the study area.

3.3 Maximum Entropy

Using the RF algorithm is relatively time-consuming, not only because of the preparation of the input raster, but also because of the preparation of the training data. To see if this approach is still worthwhile, the results are compared with those of Maxent. Maxent is based on the Maximum Entropy principle, which has become prominent in ecological applications for predicting species distributions. This principle states, from a Bayesian perspective, that the probability distribution with the highest entropy best represents the data within known constraints. Maxent works with presence-only data, defining its probability distribution over all pixels in the study area, using species occurrence pixels as sample points and their environmental characteristics as explanatory variables. First introduced by Phillips et al. (2004), Maxent has evolved into a well-developed stand-alone package for these purposes (Guisan et al. 2017). Maxent utilises a list of species presence locations and a set of environmental predictors across a study area as input. From this landscape, Maxent extracts a sample of background locations where information about presence and absence is unknown and contrasts it against the presence locations (Merow et al. 2013).
For ecological applications, Maxent is presented as a general approach to modelling species distributions using presence-only data. It estimates a target probability distribution by finding the one with maximum entropy. This means that the most spread (or uniform) distribution is chosen without contradicting the known occurrence data and environmental constraints. The model provides information on the probability of the species being present at a given location (Phillips et al. 2006). In ML, the maximum entropy is categorised within the classifications (Bonaccorso 2017).

3.3.1 Maxent model

The *dismo* package was used for the implementation of Maxent in R, as it enables Maxent to be run by calling Java (Sillero et al. 2023). Initially, the occurrence data is loaded and its CRS is transformed to match that of the predictor variables. As was done previously for the RF model, the occurrence data is randomly split into a training set comprising 70% of the data and a test set comprising 30% of the data. Subsequently, the predictor variables are loaded and

combined with the *stack* function of the *raster* package to form a single multi-layered object. The *maxent* function enables the creation of a maxent model with the occurrence train data and predictor grid stack as input data. Within this function, the variables that do not contain continuous values can be specified. Otherwise, the default settings are used. This model is then applied to the study area with the *predict* function to predict the species distribution. The output is a raster file containing the prediction data of the study area. As with RF, it is possible to display variable importance, which is shown as a percentage (HIJMANS ET AL. 2023).
Furthermore, Maxent is executed multiple times for the specific study areas. The initial stage of the process involves working with the processed raster data from the first RF iteration, which resulted in the optimal outcome for the RF model regarding the respective study areas. Subsequently, the filtered occurrence data, which has been subjected to a reduction in spatial bias, is also employed for Lower Saxony. However, it is not as straightforward to incorporate seasonal climate data in the same manner as with RF. Consequently, this step is omitted. Subsequently, a reduction of input variables are also tested for Maxent.

3.3.2 Maxent validation

For model validation, the area under the receiver-operator curve (AUC) is employed as it represents the most prevalent metric in the Maxent literature. AUC is a metric that assesses the predictive accuracy of a model without relying on a specific threshold, focusing solely on the ranking of locations. AUC represents the probability that a randomly selected presence location will be ranked higher than a randomly selected background point. While the AUC is typically employed to assess the efficiency of a model in differentiating between presence and absence locations, in the context of presence-only data, the AUC is used to evaluate the performance of the model in identifying presence locations in relation to background points (MEROW ET AL. 2013).
The initial step involves the generation of a set of random background points. According to the literature, 1,000 background points are generated from the area covered by the predictors using the *randomPoints* function (HIJMANS ET AL. 2023). The background points are employed as a reference point for the evaluation of the model in relation to the occurrence points. Subsequently, the values of the predictors at the locations of the test occurrence points are extracted. Additionally, the predictor values at the locations of the background points are extracted.
Subsequently, the previously trained Maxent model is utilised to predict the probability of species presence at each of the test occurrence points. The resulting values are stored in a vector. The same procedure is then applied to the background points, with the results also stored in a vector. The *evaluate* function is used with the vectors as input, with the objective of evaluating the model's performance based on the predictions made on the test occurrence points and the background points. This function calculates metrics such as the AUC, which can subsequently be plotted.

4 Results

The following chapter presents the results of the RF and Maxent models for assessing pollinator habitat suitability and species distribution. For each iteration, the different settings are explained, and the resulting validation parameters are presented. It is also considered how the assessment of the importance of the predictor variables changes. Finally, it is shown how the respective models predicted the study areas of Hannover and Lower Saxony and where there are differences.

4.1 Random Forest

The RF model was further developed in several iterations. Different spatial resolutions of the predictor data were tested, an approach for filtering the spatial bias was tried, the periodic climate data was replaced by seasonal data and finally the input variables were changed and the hyperparameters adjusted. The following chapters show the results of the respective iterations and finally the result of the prediction of the study areas.

4.1.1 Testing different spatial resolutions

In the first iteration, the training data for both study areas was processed at two different resolutions. The model hyperparameters are listed in Tab. 3, distinguishing between the study area and the cell size used. In addition to the defined and default settings, sample size, prediction error and R^2_{OOB} are shown. The sample size becomes smaller as the cell size increases. In Hannover, the prediction error for a cell size of 10 m is slightly lower at 0.051 than for a cell size of 50 m at 0.056. A clear difference can be seen when looking at R^2_{OOB}. For 10 m this value is 0.057, while for 50 m it is even slightly negative. For Lower Saxony, the prediction errors for 100 and 500 m are hardly different with 0.056 and 0.058. A difference can be seen when looking at R^2_{OOB}. Here the value for 500 m is higher at 0.030 than that for 100 m at 0.016.

Tab. 3: Model hyperparameters of the first RF iteration testing different spatial resolutions

Study area	Hannover		Lower Saxony	
Cell size	10	50	100	500
Type	Regression			
Number of trees	500			
Sample size	550	432	7,713	5,478
Number of independent variables	24			
Mtry	4			
Target node size	5			
Variable importance mode	permutation			
Splitrule	variance			
OOB prediction error (MSE)	0.051	0.056	0.056	0.058
R^2_{OOB}	0.057	-0.006	0.016	0.030

The validation parameters are shown in Tab. 4. MAE, MSE and RMSE show lower values at a resolution of 10 m than at 50 m. R^2 is higher than at 50 m. The SD is lower at 10 m. A cell size of 10 m performs clearly better than 50 m for the Hannover study area. The parameters for Lower Saxony are very similar. The MAE is slightly lower for 500 m and R^2 is slightly higher than for 100 m. In summary, the 500 m cell size performs minimally better.

Tab. 4: Validation parameters of the first RF iteration testing different spatial resolutions

Study area	Hannover		Lower Saxony	
Cell size	10	50	100	500
MAE	0.162	0.175	0.177	0.174
MSE	0.048	0.059	0.057	0.057
R^2	0.119	0.000	0.023	0.027
RMSE	0.219	0.243	0.238	0.239
SD	0.216	0.243	0.238	0.238

Fig. A 3 illustrates the permutation importance of the predictors of the RF model. In general, the predictors show a higher importance in Lower Saxony compared to Hannover, with most showing an increasing trend at 500 m resolution.
For many predictors, Lower Saxony has higher importance values than Hannover. In particular, UA/CLC5, CLCBB, FA, NS, Aspect and NDVI have relatively low importance values in both regions, with slight increases in Lower Saxony. While DisUA has no significance in Hannover, its value is more apparent in Lower Saxony.
Sunshine, Precipitation, Radiation and SoilTemp show different patterns. Sunshine and radiation start with higher importance in Hannover but decrease in Lower Saxony. Precipitation and SoilTemp have a relatively low importance in Hannover but become more important in Lower Saxony.
In summary, the figure shows that most predictors have a higher permutation importance in Lower Saxony compared to Hannover. According to all results the cell size of 10 m works best for Hannover, while the higher cell size of 500 m is preferred for Lower Saxony.

4.1.2 Testing spatial filtering

To reduce the strong spatial bias in Lower Saxony, the occurrence data were filtered with regards to settlement areas and roads. The bias grid was created for two different resolutions of 0.02 and 0.05°. In addition, based on these resolutions, the occurrences were filtered to different degrees with a rescaling range of 0-1, 0-1.25 and 0-1.5.
The spatial bias maps shown in Fig. 10 provide a visual representation of the estimated sampling rate per raster cell at the two different resolutions under different rescaling values compared to the unfiltered bias rasters at the top. These maps help to understand the geographical distribution and intensity of bias across the study area. The unfiltered occurrence data contains

a total of 34,681 records. With a rescale range of 0-1 this number changes to 17,859 at 0.02° and 17,462 at 0.05°, to 12,750 and 12,985 at 0-1.25 and to 8,109 and 8,538 at 0-1.5.

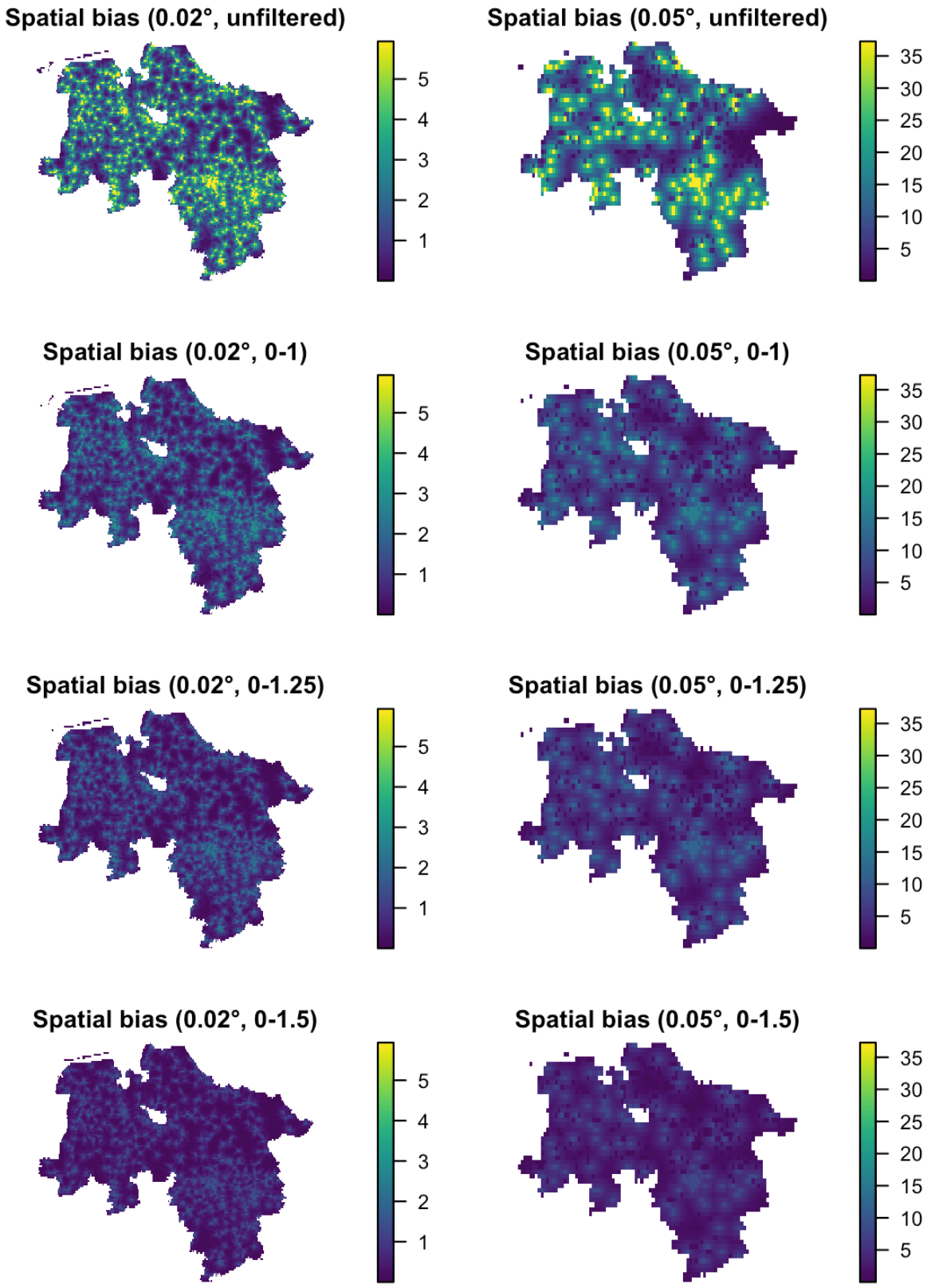


Fig. 10: Spatial bias rasters showing estimated sampling rate for unfiltered and filtered occurrence data

At both resolutions, the unfiltered maps show the highest intensity and most scattered distribution of spatial bias, with noticeable hotspots at larger settlement areas. As the rescale values increase from unfiltered to 0-1.5, there is a clear trend towards decreasing bias intensity and a more uniform spatial distribution. At a resolution of 0.02°, finer bias structures can be seen than at the coarser resolution, suggesting that finer resolution captures more detailed spatial bias patterns. The effect of rescaling is evident, as higher rescaling ranges lead to a significant reduction in both the intensity and number of high bias areas, resulting in a more homogenised bias landscape. Higher rescaling ranges were not tested as the number of records had already fallen to a quarter of the original occurrences. In addition, the 1-1.5 range already produced a very uniform result according to the rasters.

The model hyperparameters of the second iteration are listed in Tab. 5. The results are compared with the previously best performing result from the first iteration, i.e. with a cell size of 500 m. The prediction error decreases slightly with increasing rescale range at a resolution of 0.02°, while it remains approximately the same at a resolution of 0.05°. R^2_{OOB} decreases with increasing rescale for both resolutions. The exception is R^2_{OOB} for 0.02° and a rescale range of 0-1, where the value increases slightly compared to the control value. It is also clear that the sample size decreases with increased filtering. While the occurrence data is reduced to 25% with a rescale range of 0-1.5, the sample size is only reduced to a proportion of 40%.

Tab. 5: Model hyperparameters of the second RF iteration testing different filter parameters with sampbias compared to the result of the first iteration of Lower Saxony with 500 m (control)

Resolution	con-		0.02°			0.05°	
Rescale range	trol	0-1	0-1.25	0-1.5	0-1	0-1.25	0-1.5
Type	Regression						
Number of trees	500						
Sample size	5,478	4,301	3,110	2,284	3,582	2,908	2,223
Number of independent variables,	24						
Mtry	4						
Target node size	5						
Variable importance mode	permutation						
Splitrule	variance						
OOB prediction error (MSE)	0.058	0.055	0.054	0.052	0.058	0.060	0.055
R^2_{OOB}	0.030	0.037	0.007	0.006	0.017	0.009	-0.022

The validation parameters are shown in Tab. 6. At a resolution of 0.05°, MAE, MSE, RMSE and SD increase or remain approximately the same compared to the control value. However, R^2 increases for all variants and is highest for a rescale range of 0-1. At a resolution of 0.02°, the MAE, MSE, RMSE and SD increase at a rescale range of 0-1.25, while they decrease at the other two ranges. The 0-1.5 range performs best. R^2 increases with a range of 0-1 and then decreases

as the range increases. While the 0-1.5 range has the best MAE, MSE, RMSE and SD, R^2 is clearly the worst with these settings.

Tab. 6: Validation parameters of the second RF iteration testing different filter parameters with sampbias compared to the result of the first iteration of Lower Saxony with a cell size of 500 m (control)

Resolution	control	0.02°			0.05°		
Rescale range		0-1	0-1.25	0-1.5	0-1	0-1.25	0-1.5
MAE	0.174	0.168	0.183	0.161	0.178	0.175	0.174
MSE	0.057	0.053	0.063	0.048	0.059	0.056	0.057
R^2	0.027	0.049	0.032	0.018	0.037	0.035	0.034
RMSE	0.239	0.229	0.252	0.219	0.243	0.236	0.239
SD	0.238	0.228	0.252	0.219	0.243	0.235	0.239

Fig. A 4 illustrates the permutation importance of the RF model predictors for the different resolutions and rescale ranges compared to the unfiltered data. The low values for CLCBB, FA, NS, DisFE, DisNA, DisRA, Aspect, NDVI remain low and often decrease with increasing rescale range. CLC5 and DisUA also clearly lose importance with increasing rescale range, indicating that the spatial bias in the filtering process is decreasing. Clay, silt and sand lose slightly in importance with increasing rescale range. For most of the climate data the importance decreases when filtering the data. With higher rescale range the importance for the 0.02° increases again and reaches an importance close to the unfiltered data, while that for the 0.05° resolution continues to decrease.

Based on the results, it was decided to continue working with the values of a rescale range of 0-1.5 at a resolution of 0.02°. Although R^2 shows clearly poor values here, all other measures show the best comparative results. In addition, the losses in importance are lowest for this setting, and in some cases the importance even increases.

4.1.3 Testing seasonal climate data

In the third iteration, the effect of using seasonal climate data for the training data was tested for three different time periods. The model hyperparameters are listed in Tab. 7. For Hannover, the results are compared with those of the first iteration, where the data were modelled with a cell size of 10 m. For Lower Saxony, the data is compared with the second iteration using a resolution of 0.02° and a rescale range of 0-1.5. The results of this third iteration are also based on the previous results, which are presented here as comparative values.

The periods considered are 2010-2024, 2018-2024 and 2022-2023. The sample size increases for both study areas and decreases as the period gets shorter. The sample size for the period 2022-2023 is below the control values. For Hannover, R^2_{OOB} increases and reaches its highest value for the period 2018-2024. It also increases for Lower Saxony but is highest for the period 2010-2024. The prediction error decreases in Lower Saxony with a shorter time span. For Hannover the value partly decreases and is lowest for the period 2010-2024.

Tab. 7: Model hyperparameters of the third RF iteration testing the use of seasonal climate data for different time periods compared to the result of the first iteration of Hannover with a cell size of 10 m and the second iteration of Lower Saxony with a resolution of 0.02° and a rescale range of 0-1.5 (control)

Study area	Hannover				Lower Saxony			
Years	con-trol	2010-2024	2018-2024	2022-2023	con-trol	2010-2024	2018-2024	2022-2023
Type	Regression							
Number of trees	500							
Sample size	550	697	672	310	2,284	3,260	2,816	1,008
Number of independent variables	24							
Mtry	4							
Target node size	5							
Variable importance mode	permutation							
Splitrule	variance							
OOB prediction error (MSE)	0.051	0.046	0.049	0.053	0.052	0.047	0.045	0.042
R^2_{OOB}	0.057	0.136	0.157	0.104	0.006	0.038	0.026	0.007

The validation parameters of the third iteration are shown in Tab. 8. For Hannover, the MAE, MSE, RMSE and SD increase with the use of seasonal climate data. They are highest in 2022-2023. The lowest values are from 2018-2024, which are only slightly different from the control value. For 2010-2024 and 2018-2024 R^2 increases but decreases with shorter time periods. For 2022-2023 R^2 is below the control value. In Lower Saxony the MAE increases in the first two periods. For 2022-2023 it is only slightly below the control value. MSE, RMSE and SD all decrease minimally, and the values for the different time periods are almost identical. R^2 increases for the period of 2010-2024 and then decreases again for the shorter periods.

Tab. 8: Validation parameters of the third RF iteration testing the use of seasonal climate data for different time periods compared to the result of the first iteration of Hannover with a cell size of 10 m and the second iteration of Lower Saxony with a resolution of 0.02° and a rescale range of 0-1.5 (control)

Study area	Hannover				Lower Saxony			
Years	con-trol	2010-2024	2018-2024	2022-2023	con-trol	2010-2024	2018-2024	2022-2023
MAE	0.162	0.165	0.166	0.178	0.161	0.168	0.170	0.158
MSE	0.048	0.056	0.049	0.053	0.048	0.045	0.045	0.046
R^2	0.119	0.195	0.152	0.072	0.018	0.086	0.071	0.029
RMSE	0.219	0.236	0.221	0.229	0.219	0.213	0.213	0.214
SD	0.216	0.235	0.221	0.225	0.219	0.213	0.212	0.214

The permutation importance of the input variables is visualised in Fig. A 5. The study areas are differentiated by colour. The different time periods are shown on the x-axis next to the control value. In several cases, such as DTM, Slope, Precipitation, Drought, Radiation, SoilTemp and SoilMoist, the control values are significantly higher compared to the newer time periods for the Lower Saxony study area. This indicates that these predictors were more influential in the earlier model iterations.
For most predictors, the permutation importance tends to decrease or remain stable over the different time periods compared to the control. However, there are some exceptions: CLCBB shows a significant increase in importance for Hannover in the period 2022-2023. Also, the NDVI importance becomes slightly higher with shorter time periods. For Hannover, the importance tends to decrease with the seasonal climate data, but it is obvious that the period 2018-2024 performs best, as this importance often reaches values close to the control value, sometimes even increasing (UA, DisFE, DisNA, DisRA, Clay, Silt, Sand, Sunshine, TempMin, Radiation). In Lower Saxony, the highest importance values can be seen in the period 2010-2024, but most of them are not close to the control value. TempMean is the only variable with a noticeable increase in importance compared to the control value.
Based on the results, the period from 2018-2024 was selected for Hannover. R^2 increased while MAE, MSE, RMSE and SD remained almost the same. In addition, the importance for this period has often increased or decreased only slightly. This is not so obvious for Lower Saxony. Although the validation parameters were slightly improved using seasonal climate data, the importance has generally decreased significantly. Therefore, it was decided not to work with seasonal data for this study area.

4.1.4 Fine-tuning the model

For Hannover, the work will continue with the data that have been processed with a cell size of 10 m and that contain seasonal climate data for the period from 2018 to 2024. To further improve the model, the variables with low significance are now removed. All variables with an importance of less than 0.002 were dropped from the model (NS, SoilTemp, DisUA, Aspect, Slope, FA, CLCBB, DTM, NDVI, UA). Fewer were also tried, but this gave the best result in terms of validation parameters. As only 14 variables were considered, mtry was automatically set to 3. In addition, the hyperparameters were changed in further runs. An mtry of 4 and a number of trees of 1,000 instead of 500 were tested.
Tab. 9 shows the hyperparameters of this fourth iteration. The prediction error increases slightly as the number of variables is reduced. It is lowest for a number of trees of 1,000 and an mtry of 3, but the others are only slightly higher. R^2_{OOB} drops very much in some cases. It is highest for a number of trees of 1,000 and an mtry of 3, where it is only slightly less than the control value.

Tab. 9: Model hyperparameters of the fourth RF iteration testing less input variables and different mtry and number of trees for Hannover compared to the results of the third iteration with seasonal climate data in the period of 2018-2024

Run	Control	1	2	3	4
Type	Regression				
Number of trees	500	500	500	1,000	1,000
Sample size	672				
Number of independent variables	24	14	14	14	14
Mtry	4	3	4	3	4
Target node size	5				
Variable importance mode	permutation				
Splitrule	variance				
OOB prediction error (MSE)	0.049	0.053	0.059	0.052	0.054
R^2_{OOB}	0.157	0.098	0.051	0.148	0.085

Tab. 10 shows the validation parameters of the fourth iteration. MAE, MSE, RMSE and SD could be reduced in almost all cases by removing some of the variables. However, the differences to the control values are not too strong. R^2 did not increase in all cases, but for the fourth run, i.e. an mtry of 4 and a number of trees of 1,000, it increased significantly.

Tab. 10: Validation parameters of the fourth RF iteration testing less input variables and different mtry and number of trees for Hannover compared to the results of the third iteration with seasonal climate data in the period of 2018-2024

Run	Control	1	2	3	4
MAE	0.166	0.153	0.157	0.159	0.158
MSE	0.049	0.048	0.047	0.051	0.047
R^2	0.152	0.159	0.142	0.080	0.192
RMSE	0.221	0.220	0.216	0.226	0.216
SD	0.221	0.220	0.212	0.226	0.216

Fig. A 6 shows the permutation performance of this iteration for Hannover. The control values from the previous iteration are also shown here. It can be clearly seen how the importance of all the remaining variables increases as the others are removed. While DisFE, DisNA, DisRA, Clay, Silt, Sand, Drought, Radiation and SoilMoist increase only slightly in comparison, Sunshine, Precipitation, TempMin, TempMean, TempMax show a very significant increase. The extent to which the different numbers of trees and mtrys have an effect cannot be seen directly, as they seem to be quite unpredictable.

In terms of validation parameters, run 4 was selected as the final RF model for Hannover. This model was built using input data with a cell size of 10 m. Seasonal climate data for 2018-2024 were used, and an mtry of 4 and a number of trees of 1,000 were selected for the model.

For Lower Saxony, however, it was decided not to continue working with seasonal data. Therefore, the data was used again, with the input data processed at 500 m and then filtered with sampbias at a resolution of 0.02° and a rescale range of 0-1.5. Again, the low importance variables were removed first. The best result in terms of validation parameters was obtained by removing all variables with a permutation importance of less than 0.002 (Aspect, NS, FA, DisNA, NDVI, CLCBB, DisFE, CLC5, DisUA, DisRa, Clay).

Tab. 11 shows the hyperparameters of this fifth iteration. The number of variables was reduced from 24 to 13, resulting in a mtry of 3. Again, an additional mtry of 4 was tested, as well as a number of trees of 1,000 instead of 500. The prediction error has remained almost the same, while R^2_{OOB} has decreased and even become negative for three runs.

Tab. 11: Model hyperparameters of the fifth RF iteration testing less input variables and different mtry and number of trees for Lower Saxony compared to the results of the second iteration using a resolution of 0.02° and a rescale range of 0-1.5

Run	Control	1	2	3	4
Type	Regression				
Number of trees	500	500	500	1,000	1,000
Sample size	2,284				
Number of independent variables	24	13	13	13	13
Mtry	4	3	4	3	4
Target node size	5	5	5	5	5
Variable importance mode	permutation				
Splitrule	Variance				
OOB prediction error (MSE)	0.052	0.050	0.051	0.054	0.052
R^2_{OOB}	0.006	-0.014	0.000	-0.056	-0.030

Tab. 12 shows the validation parameters for this iteration. MAE, MSE, RMSE and SD have almost all increased slightly. However, R^2 could be increased in two cases. It is highest for a number of trees of 1,000 and a mtry of 4, followed by the mtry of 3.

Tab. 12: Validation parameters of the fifth RF iteration testing less input variables and different mtry and number of trees for Lower Saxony compared to the results of the second iteration using a resolution of 0.02° and a rescale range of 0-1.5

Run	Control	1	2	3	4
MAE	0.161	0.171	0.168	0.162	0.158
MSE	0.048	0.056	0.053	0.049	0.051
R^2	0.018	0.016	0.006	0.029	0.030
RMSE	0.219	0.237	0.231	0.222	0.225
SD	0.219	0.237	0.223	0.221	0.225

The permutation importance for this iteration can be found in Fig. A 7. Again, all the remaining variables have increased in importance. DTM, Slope, TempMin and Radiation show a slight but significant increase. Silt, Sand, Sunshine, Precipitation, Drought, TempMean, TempMax, SoilTemp and SoilMoist have in some cases almost doubled compared to the control value. In comparison, the importance is usually slightly higher for a mtry of 4 than for a mtry of 3. In addition, the importance is usually higher for a number of trees of 500 than for one of 1,000. Selecting the optimal setting for Lower Saxony was not a simple process. Based on R^2, the number of trees with 1,000 was selected, and since the importance for mtry of 4 was typically higher, this value was chosen.

4.1.5 Prediction of the study areas

To predict habitat suitability within the study areas, the model which performed best was applied to each respective study area. As seasonal climate data was used in Hannover, the model was applied to spring, summer and autumn of one year. The year 2023 was chosen because it includes the most occurrences. The winter season was not considered, given that bees do not fly during this period. In Lower Saxony, periodic climate data was used, so this step was omitted for this region. In this case, habitat suitability was only predicted once. The results are shown in Fig. 11.

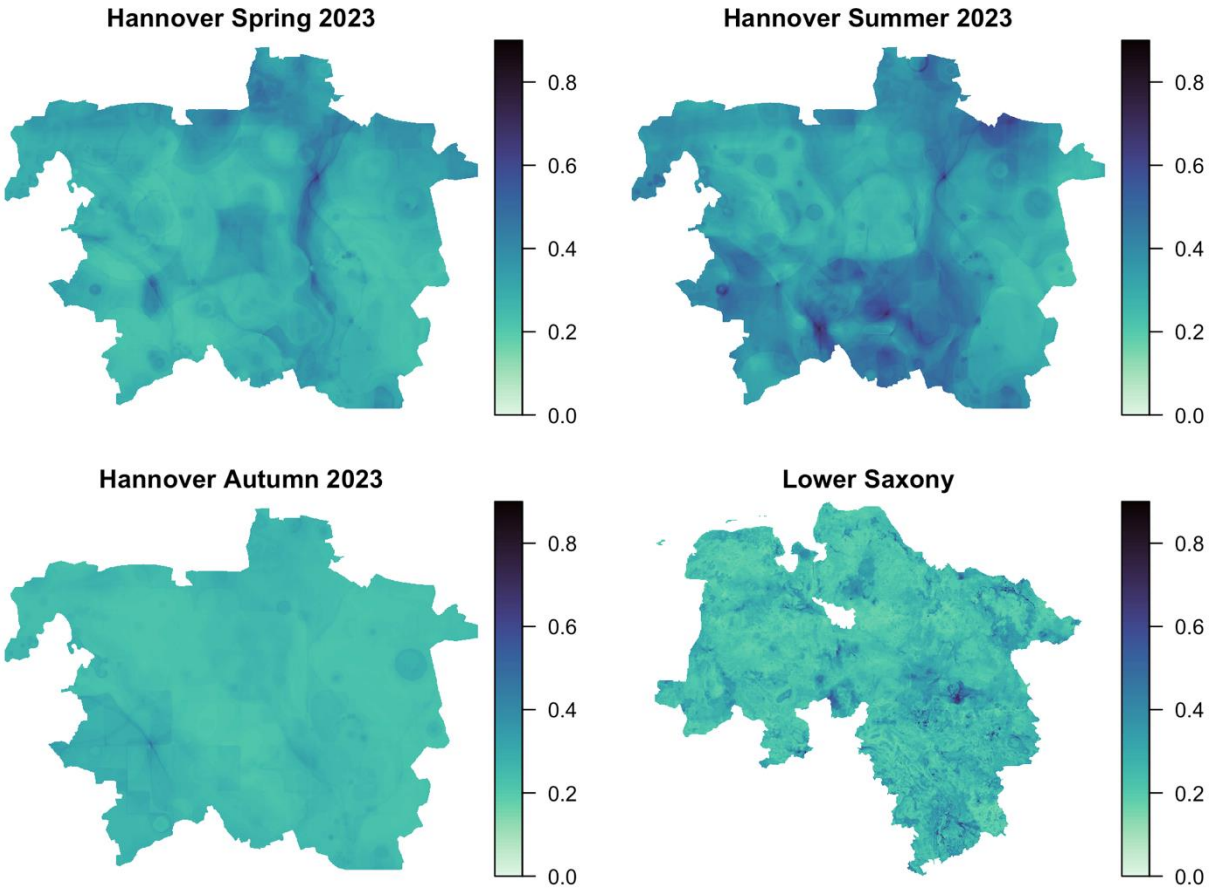


Fig. 11: Results of the habitat suitability prediction using RF for the study areas of Hannover and Lower Saxony

Looking at Hannover, the highest values are reached in summer. While some higher structures can still be seen in spring, in autumn almost only low values are predicted. In summer there is a larger region of higher values, especially in the southern part of the area. High and low values can be seen for Lower Saxony. In order to be able to better evaluate the data, scatter plots have been created showing the predicted value achieved for each pixel and each variable for the four results and the associated level of the variables. The plots can be found in the Appendix.

The LULC data have been excluded from both study regions due to their limited importance. Additionally, information on FA and NS was also deemed to be of minimal value in both regions and is therefore not included as predictor data for the final models. DisUA was also removed. This variable served as a control value for the spatial bias, so its lack of relevance in both study regions is a positive outcome. The initial distinction in the selection of important predictor variables is evident in DisFE, DisNA and DisRA. Although these were deemed relevant for the area of Hannover, they were found to be of small relevance in Lower Saxony and were therefore excluded from the final model. In terms of their importance within the final model of Hannover, the three predictors are of comparatively lower value (Fig. A 6, Fig. A 7).

Fig. A 8 illustrates the scatterplots for the selected predictor variables in Hannover during spring, Fig. A 9 during summer, and Fig. A 10 during autumn. A review of the data reveals that, in particular during the spring season, a higher prediction is more likely to be recognised with a low value of DisFE. The highest distances receive low prediction values in all seasons. However, in summer, the peaks in prediction are not easily assignable, as they are located between 1,000 and 3,000 m. In autumn, the peak is at approximately 2,500 m, and otherwise, only a slight pattern can be observed, which generally shows a lower distance in higher prediction values. Regarding DisNA, the most evident pattern is observed in spring. A distinct peak is evident at a shorter distance. Overall, the prediction declines as the distance increases. In summer, the pattern is less discernible, with multiple peaks across all distances and a greater dispersion of values. In autumn, the trend reverses. The prediction shows a slight increase with increasing distance, with a single peak between 10,000 and 15,000 m. In the spring season, the DisRA value reaches its maximum at a distance between approximately 2,000 to 3,000 m. The prediction exhibits a relatively uniform decrease at distances above and below this point. In contrast, during the summer months, the data reveals a distinct pattern whereby the highest prediction values are observed at shorter distances, with a notable decline at greater distances. In autumn, the pattern is more similar to spring, but the prediction values decline more after the peak with greater distance.

The next stage of the analysis considers the geomorphological predictors DTM, Slope and Aspect. In the Hannover area, all variables were identified as having an insufficient level of importance and were therefore excluded from the final model. In contrast, in Lower Saxony, DTM and Slope were assigned greater importance, although Aspect was not included in the final model as well. The scatterplot for Lower Saxony is shown in Fig. A 11. In examining the data

for DTM and Slope, it is evident that the highest prediction is observed at the lowest values for the two predictors. As the values of the variables increase, the predicted outcome declines.
The NDVI as a variable for vegetation was assessed as having low significance in both study areas and was therefore not included in the final models. In consideration of the soil predictors clay, silt, and sand, it can be stated that all variables for Hannover were incorporated into the final model, despite their comparatively low importance values. In the case of Lower Saxony, only the clay variable was not included. A similar pattern is observed for clay in Hannover across all seasons. The highest predictor values are achieved with a clay content of 0.1 to 0.2. As the quantity in question increases, the values in question decrease. Regarding silt, it can be observed that the highest prediction values are achieved with an amount between 0.5 and 0.7. It is evident that peaks can be identified across all seasons. The peak is most apparent in the spring and autumn months, with a somewhat less distinct appearance in the summer period, due to the higher overall prediction values. The distribution of silt values in Lower Saxony is more dispersed. The highest values are observed between 0.4 and 0.7, though no distinct peaks can be identified. As the predicted value increases and decreases surrounding this range, the predicted value trends towards a decrease. A comparable pattern can be observed regarding sand. In the case of Hannover, clear peaks between 0.2 and 0.3 can be identified throughout the seasons. In the case of Lower Saxony, the values are also more dispersed, lacking clear peaks. However, the highest predicted values fall within the range of 0.2 to 0.5.
The following section presents the results for the predictor variables sunshine, precipitation and drought. All variables were included in the final models for both study areas. In Hannover, the variable of sunshine achieves a very high importance value, while the variable of drought achieves comparatively low importance. In contrast, in Lower Saxony, precipitation and drought are the two highest-rated predictor variables, whereas sunshine tends to be of lesser importance. The seasonal climate data for Hannover was employed for these variables, which is the reason why the y-axes of the scatterplots of these variables differ. It is not possible to make a definitive statement regarding trends in sunshine levels. Although there are some peaks, no discernible pattern can be identified. It is evident that the predictor values for Lower Saxony are particularly elevated within the middle value ranges. It is similarly challenging to draw conclusions regarding precipitation based on the scatterplots for Hannover. However, in Lower Saxony, it is evident that low precipitation values have the highest predictions, which decrease with increasing precipitation. A comparable pattern emerges for drought. Although it is difficult to make a statement for Hannover, in Lower Saxony, the highest predictions are achieved with low drought values, which decrease with increasing drought.
In the following step, the temperature values TempMin, TempMean and TempMax are considered. All variables were employed in both final models. While these predictors are of high importance in Hannover, they are of medium importance in Lower Saxony. Once more, seasonal climate data was employed in the Hannover study area. The peaks for TempMin are evident in Hannover across all seasons, particularly within the medium value ranges. The

structure becomes more apparent in Lower Saxony. The highest predictor values are observed at approximately two-thirds of the TempMin range. The predictor values subsequently decrease around this range. Additionally, higher prediction values for both TempMean and TempMax are observed in Hannover, particularly within the mid-range of the variables. In contrast, the data for Lower Saxony clearly show that both predictor variables have the highest predictive values at the highest variable values, which decreases as the variable value decreases.
Finally, the variables Radiation, SoilTemp and Soil Moist are considered. In the case of Hannover, the predictor SoilTemp was not included in the final model. In contrast, this variable is of comparatively high importance in Lower Saxony. The remaining two variables are employed in both models. As no seasonal climate data was available for these predictors, periodic data was used for both study areas. Regarding the spring season, there is an increase in radiation levels in Hannover as the prediction increases. This is not the case in summer, however, several peaks can be observed across the value range. In autumn, it is evident that the prediction increases slightly as the values decline. A peak is observed at approximately 14.35 kWh/m^2. In Lower Saxony, the values exhibit greater dispersion, with a slight peak at around 14 to 14.5 kWh/m^2. The soil temperature in Lower Saxony reaches its highest values at approximately 20 °C. The prediction for values above this decreases slightly, while for values below it decreases more. SoilMoist in Hannover appears to provide higher predictions with falling values for the seasons spring and summer, although peaks can be identified at the lowest and highest values for SoilMoist. In autumn, this peak is only present at the highest value. For Lower Saxony, the highest prediction values can be identified at a SoilMoist of approximately 95 to 100% NFK, which decreases with increasing SoilMoist.

4.2 Maxent

Subsequently, the outputs of the Maxent model are analysed, initially for Hannover and then for Lower Saxony. In this section, the significance values of the predictors in the various runs and the validation of the models using the AUC are considered. The prediction is also shown. The aim here is to create a comparison to the RF. The focus is on considering the importance of predictor variables and species distribution prediction. It should be noted that the validation cannot be compared in the best possible way, as different methods are used for the different models. The results are based on the optimal results previously achieved by the RF.

4.2.1 Hannover

The Maxent model was initially run with all predictor variables for the Hannover study area. As the data with a resolution of 10 m was found to be the most effective for the RF model, it was selected for use. The occurrence data was not filtered for Hannover, and the integration of seasonal data was not undertaken due to the limitations of the model. As illustrated in Fig. A 12, the initial run reveals that DisFE, DisNA and Sunshine, along with UA and CLCBB, attain remarkably high importance values, ranging from 10 to 15%. The control value DisUA is of minimal significance and is excluded from the subsequent run.

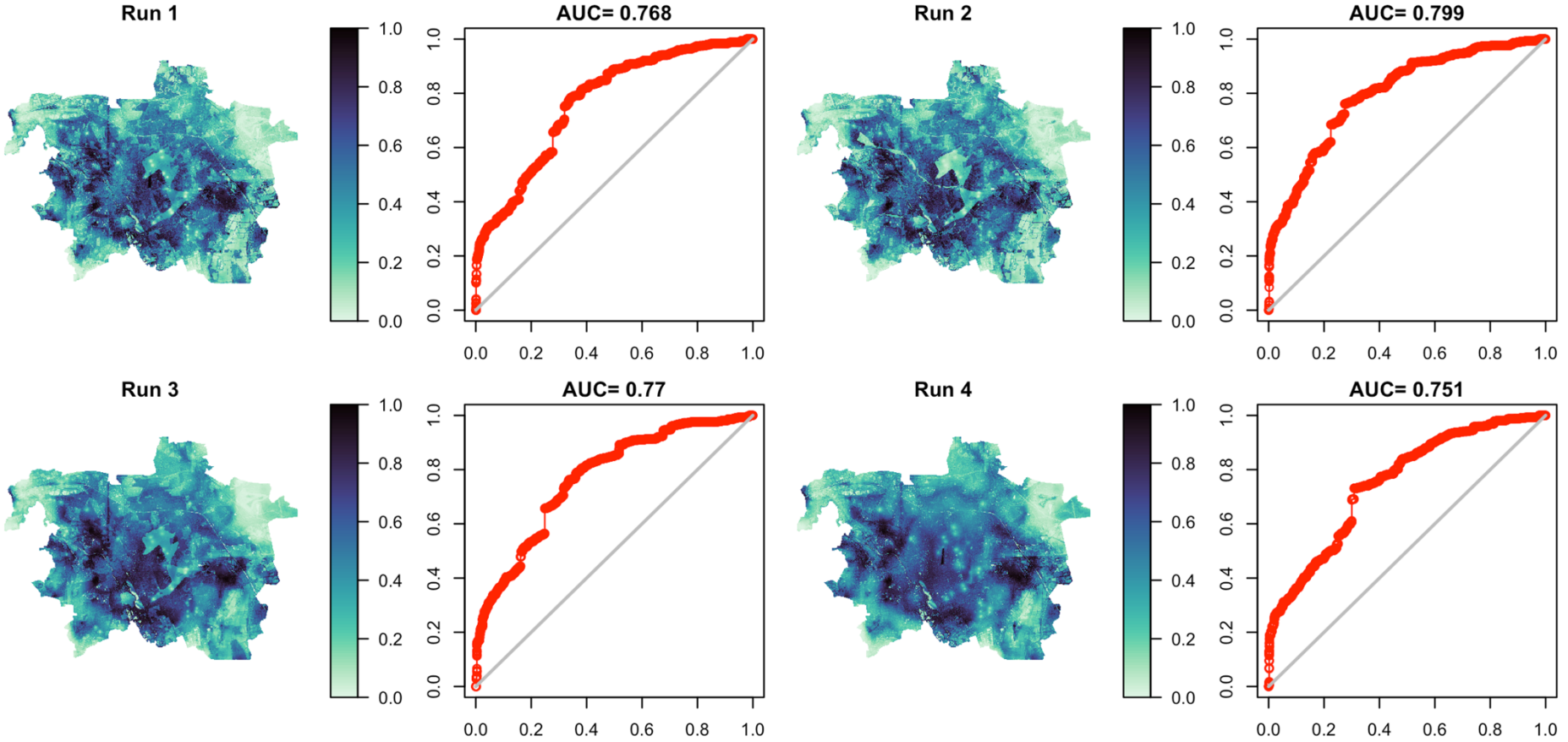


Fig. 12: Prediction and AUC of the Maxent models for Hannover testing different input data in four runs

Consequently, the importance of UA and CLCBB increases slightly in the second run. The greatest increase is observed in the importance of DisFE, which rises above 20%. The importance of DisNA and Sunshine declines slightly. To test the model's performance without the LULC data, which were assessed as irrelevant by RF, they are also removed in the third run. Many values remain similar, but there is an increase in both TempMax and SoilTemp. Finally, a run is performed in which FA and NS are also removed as the were not considered relevant by RF and based on expert opinion. Both are evaluated with approximately 5% in the third run. In the fourth run, some variables increase or decrease slightly, with only Sunshine increasing more strongly.

Fig. 12 illustrates the spatial predictions and AUC of the corresponding runs. It can be observed that the AUC demonstrates an improvement from run 1 to run 2, resulting from the exclusion of DisUA, with a notable increase from 0.768 to 0.799. However, the removal of UA and CLCBB, followed by FA and NS, resulted in a reduction in the AUC to 0.751 in run 4. To obtain a more detailed evaluation of the prediction values, the range of values for the four runs is presented as boxplots in Fig. 13. An examination of the prediction maps for runs 1 and 2 reveals a slight shift in the range towards the lower end, which is corroborated by the boxplots. Furthermore, there is an increase in the value range for runs 3 and 4. Additionally, run 4 exhibits a more homogeneous study area with a reduction in discontinuous areas.

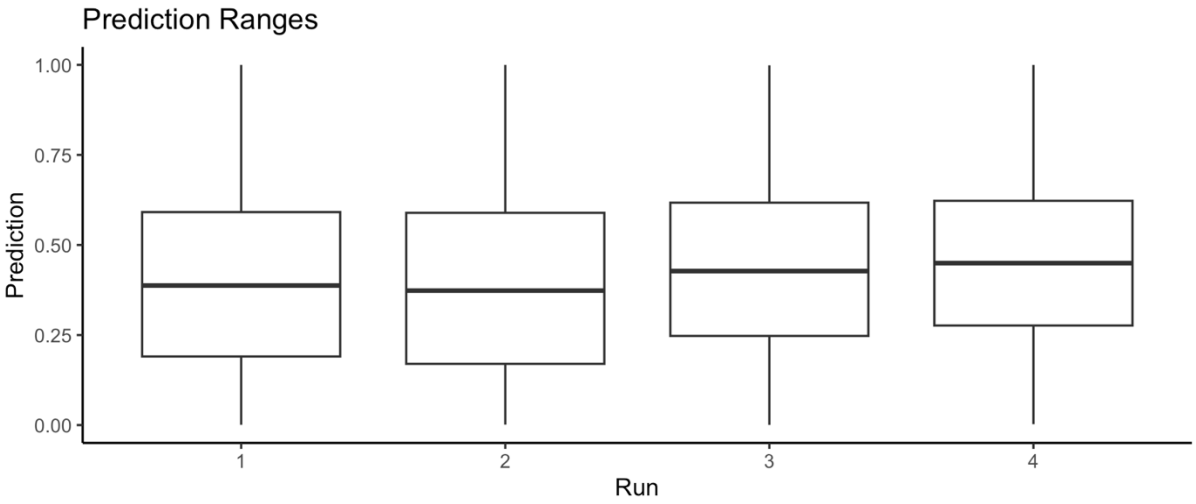


Fig. 13: Prediction ranges of the Maxent models for Hannover testing different input data in four run

4.2.2 Lower Saxony

In the case of the Lower Saxony study area, Maxent was initially applied to the unfiltered occurrence data based on the Predictor raster files with a cell size of 500 m. An examination of the Importance values in Fig. A 13 reveals that DisUA reaches approximately 65%, which indicates a markedly high spatial bias in the data. Subsequently, the filtered occurrence data with a resolution of 0.02° and a rescale range of 0-1.5 was applied in the second run. It can be observed that the importance of DisUA exhibits a slight decline yet remains at a notably high level of 60%. Consequently, this variable was excluded from the model in the third iteration, as it serves only as a control value. It is evident that CLC5 but also CLCBB, become significantly

more important in the third run. The LULC data demonstrated minimal relevance in the RF model. Therefore, in the fourth run, a model was constructed without the two LULC data as predictors. It can be observed that FA and NS become considerably more relevant. These variables are based on expert opinion and are linked to the CLC classes. They are also irrelevant in the RF model. A fifth and final run was conducted in which these variables were also excluded. In this run, the importance values are somewhat more evenly distributed. Drought, SoilMoist and Slope in particular achieve higher values here.

To gain a more detailed insight into the outcomes of the five runs, the prediction results and the AUC are presented in Fig. 15. In run 1, the unfiltered occurrence data was utilised, resulting in an AUC of 0.834. The application of filtering to the data resulted in a slight increase in the AUC, reaching 0.842. It may be assumed that the filtering of the data has a beneficial effect on the model, despite the reduction in the number of occurrences. In the subsequent three runs, predictor variables were removed. As a result, the AUC decreased due to the removal of the control value DisUA. In the fourth run, the two previously very important LULC data, CLC5 and CLCBB, were removed. Consequently, the AUC continued to fall. In the final run, FA and NS were removed, resulting in a further decrease in the AUC to 0.789.

Upon examination of the prediction, it becomes evident that the outcome of runs 1 to 2, with the filtered data, exhibits a more pronounced contrast. This indicates that the number of values within the middle range is reduced, while the number of low prediction values is increased. A visual inspection of the prediction ranges in Fig. 14 provides further confirmation of this hypothesis, with a clear shift in distribution towards the lower range. As variables are removed, the value range increases again. Furthermore, run 5 reaches even higher values than run 1. However, the result of the mapped prediction differs significantly from that of run 1. The result of run 5 appears more uniform and without abrupt transitions, whereas the initial run appears more discontinuous.

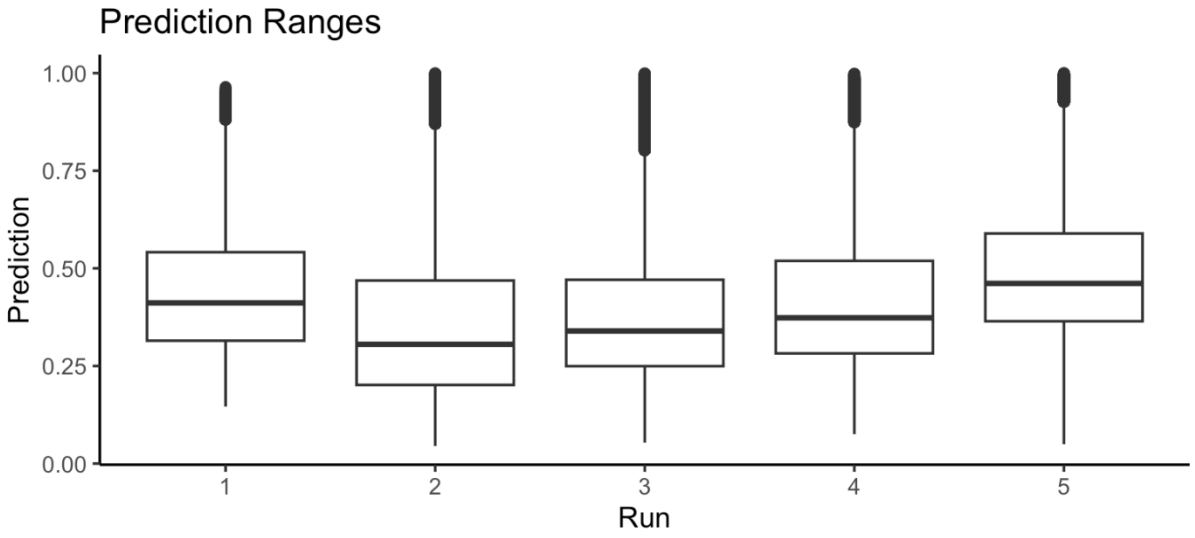


Fig. 14: Prediction ranges of the Maxent models for Lower Saxony testing different input data in five runs

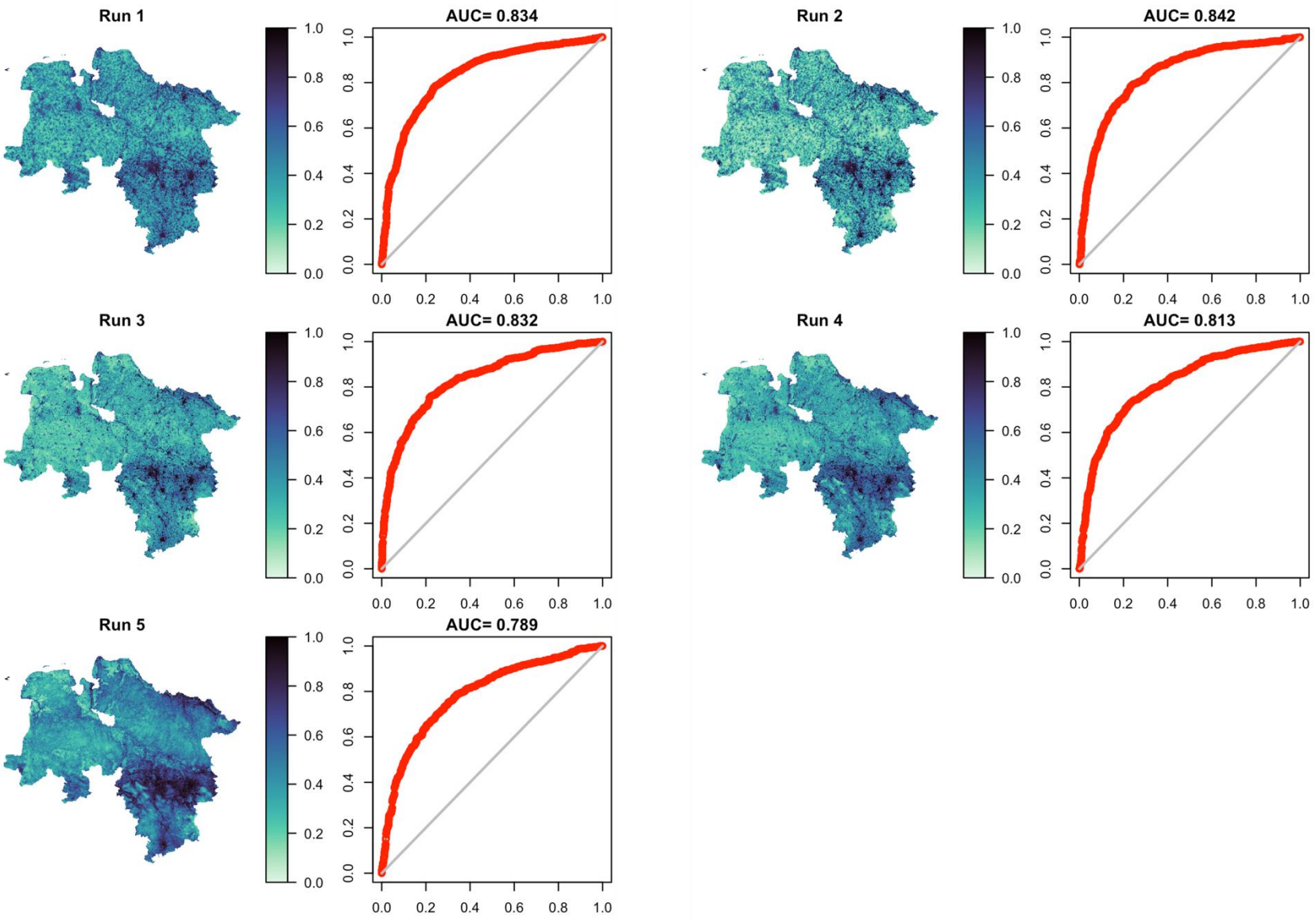


Fig. 15: Prediction and AUC of the Maxent models for Lower Saxony testing different input data in five runs

5 Discussion

This thesis presents the development of a RF modelling approach for the assessment of pollinator habitat suitability in the study areas Hannover and Lower Saxony. While the primary focus was on the RF modelling approach, the study areas were also modelled with Maxent to enable a comparison to be made. Initially, the Maxent modelling approach appears to be more straightforward than the RF method. The model has been implemented in the R software package *dismo* in such a way that it is highly user-friendly and can therefore be rated as very comprehensible. The GBIF data can be incorporated into the function as downloaded, as the model has been designed to work with this kind of presence-only data. The generation of background points is a fully automated process. It should be noted, however, that the default settings have been employed in the majority of cases. A greater number of settings can be made when using Maxent to enhance the quality of the model. In comparison, the creation of the RF model with the *ranger* package was more complex, but still easy to use. RF was not developed to capture habitat suitability, but it remains an effective tool for this purpose. The main challenge was preparing the training data accordingly. A further point for consideration was the preparation of the input predictor variables. The variables were required to be prepared as raster data for both approaches over the same extent, with the same CRS and spatial resolution. The process of searching for and preparing the data proved to be time-consuming, yet it was a uniform procedure for both modelling approaches.

In this thesis, an approach was taken to assess the degree of suitability through the density or abundance of species, due to the lack of absence data and the weaknesses of pseudo-absences (Džeroski 2009; Guisan et al. 2017). A regression approach was developed, where a grid was generated and the number of occurrences in each cell was quantified and relativised, serving as the target variable for the model. When modelling occurrence density, it is important that the probability of detecting an individual when the species is present is the same for the entire study area in order to obtain a reliable abundance estimate (Pearce & Ferrier 2001). It is unlikely that this is the case when using GBIF data. However, the bias inherent in GBIF data is also problematic when modelling presence and absence data and was attempted to be minimised by removing outliers and filtering the data.

As no comparison was made between the recording of occurrence density and the creation of background or absence points for RF, it is difficult to make a statement about the suitability of this method. However, it does have some advantages. This method allows the model to fit the data, rather than the other way around (Warton & Shepherd 2010). The preparation of the training data is easier, as the way in which the background points are created can strongly influence the result and many settings must be tried and tested (Barbet-Massin et al. 2012). While there are evident advantages to modelling relative occurrence density, the applicability of this approach for HSM based on GBIF data remains uncertain due to the spatial bias inherent in the data. Specifically, the occurrences can be underrepresented in areas of high abundance and overrepresented in areas of low abundance (Gomes et al. 2018). In addition to the

clustering of occurrence data in specific locations, the absence of a standardised sampling design further complicates the interpretation of the data (GUISAN ET AL. 2017). For further research, it would be useful to investigate the comparison of modelling presence points and relative density, as it has not yet been considered in HSM with RF using GBIF data.

A comparison of the performance of the RF regression model and the Maxent classification model for the assessment of habitat suitability is inherently challenging due to the fundamental differences in their methods and the validation metrics applied to each. The RF model, which is based on a regression approach, was evaluated using a series of continuous performance metrics, including MAE, MSE, R^2, RMSE and SD. These metrics provide insight into the accuracy of the model in predicting continuous habitat suitability scores (CHICCO ET AL. 2021). In contrast, the Maxent model, which is based on a classification approach, was validated using the AUC, a metric that assesses the model's ability to distinguish between suitable and unsuitable habitats (MEROW ET AL. 2013). This contrast in evaluation techniques demonstrates the difficulty of making a direct comparison between the two models, as each set of metrics focuses on different aspects of model performance. Consequently, while the RF model provides a comprehensive evaluation of prediction accuracy across a range of values, the Maxent model offers a statistical interpretation of habitat suitability that is particularly suited to binary classification tasks. Accordingly, any comparison of these models must be made with consideration of their inherent strengths and limitations.

By fine-tuning the parameters, an R^2 value of just under 0.2 was achieved for the Hannover study area. Although this still indicates that the fit of the model is rather trivial (CHICCO ET AL. 2021), the value is still much higher than that of the Lower Saxony study area, where an R^2 of just over 0 was achieved. The values for MAE, MSE, RMSE and SD are not very different for the two study areas. In contrast, an AUC of 0.75 to 0.80 was obtained using Maxent for Hannover, depending on the setting. Values of 0.79 to 0.84 were achieved for Lower Saxony. The model fit was therefore higher for Lower Saxony than for Hannover, in contrast to RF. The study area of Hannover was selected for analysis due to the relatively low spatial bias observed in the occurrence data. The bias is particularly evident when highly populated areas are compared with their surroundings. Consequently, the higher R^2 value for the RF model in Hannover may indicate improved model fit when the spatial bias is minimal. Nevertheless, the results indicate that Maxent is performing significantly better than RF.

The fit of the Maxent model can be interpreted as *fair* to *good* or *useful* (GUISAN ET AL. 2017). It should be noted, however, that Maxent models are generally prone to overfitting (PHILLIPS ET AL. 2006). While the AUC provides a discrimination measure for all possible ranges of thresholds and corresponds to the probability that a presence has a higher predicted value than an absence (LOBO ET AL. 2008), it does not quantify overfitting (RADOSAVLJEVIC & ANDERSON 2014). It is therefore possible that a poorly fitted model may still demonstrate good discriminatory power as a result of overfitting (LOBO ET AL. 2008). It can be assumed that there is an overfitting effect, given that the importance values demonstrated strong relevance in the LULC data and

FA and NS when modelling with Maxent (BORIA ET AL. 2014). In contrast, this was not the case when modelling with RF.

Another potential explanation for the notable differences between the validation metrics of the two models is the utilisation of relative abundance data for RF and presence point data for Maxent. The results differ significantly in terms of both the validation parameters and the model fit, as well as in terms of prediction. To achieve a more accurate comparison, it would be beneficial to perform the RF modelling additionally with the incorporation of background points. The comparison of the results with Maxent in this thesis, as a commonly used model, is interesting but poses significant interpretation challenges. It should also be noted that the dataset is not optimal despite the filtering. The GBIF data exhibit a high degree of spatial bias, which has limited the scope for creating a model with a good fit. Instead, the aim has been to enable and test new approaches for calculating and evaluating pollinator habitat suitability.

To facilitate a comparison of the results of the HSM, the predictions presented in Fig. 16 for Hannover are displayed in a side-by-side format. The prediction with RF is presented for the summer of 2023, given that this is the period during which the bees are most abundant. Two runs were selected for Maxent. In the second Run, DisUA was excluded from the modelling process, while in the fourth Run, the LULC data, FA and NS were also removed. The top row depicts the results for the entire study area. The second row presents a section of the northern tip of the Hannover study area, predominantly comprising arable land and pastures, as well as urban fabric and forest areas. The third row illustrates a section of the southern part of the area, including the waterbody Maschsee. To the east of the lake is a forest area, and to the west are urban fabrics. Additionally, the right-hand column illustrates the value ranges of the predictions, which correspond to the order in which they are displayed.

The most obvious difference between the predictions is that the structures of the LULC data are still very present in Maxent. These structures can also be seen in run 4, where they are probably based on the NDVI, which also shows a higher relevance in the modelling. The prediction grid appears much more homogeneous from run 2 to run 4, but in contrast the prediction grid of RF is much more uniform, and no hard edges can be seen. It remains questionable how accurate these structures in the Maxent predictions are and whether they really exist in nature. While there can be some impassible barriers that create hard edges for certain species, there tend to be more soft edges with habitat quality that declines continuously (WATTS ET AL. 2024). Looking at the full extent value ranges, there is a clear difference. Maxent's values are much more spread out over the whole range from 0-1. RF, on the other hand, only has values between 0.2 and 0.6. The average of all ranges is around 0.35-0.45. This pattern supports the hypothesis that a different distribution arises between the two models and their approach to describing habitat suitability based on species abundance and occurrence. In the case of RF, there is abundance throughout the area, but it is less dense in some areas and denser in others. There are no locations within the area where the prediction is 0. This is likely because no positions without occurrences were included in the training data of the RF modelling. It would

have been beneficial for the model to include absence points in order to achieve a higher variation of the values. However, this would have involved the creation of pseudo-absences, which could have introduced new uncertainties (Barbet-Massin et al. 2012).
It is also noticeable that the Maxent results show many high values in the inner part of the study area, i.e. where the urban fabric is present. Forest and agricultural areas on the edge tend to have low values. For RF, higher values can be seen in the outer areas. In order to provide a more detailed and comprehensive analysis of the results, the two sections of the study area were subjected to a more in-depth examination. An inspection of run 2 of the first extent in the northern area reveals not only the distinct structural characteristics of the various LULC types, but also the delineation of forest edges. Nevertheless, it is evident that the forest edges are assigned a relatively low rating, whereas the forest core is rated higher. The agricultural areas were assigned a low rating, whereas the urban fabric was rated highly. This effect is diminished by the exclusion of the LULC, FA and NS data, yet the structures remain visible. A comparison of these results with the expert opinions on the availability of flowering plants and nesting sites for pollinators in the ESTIMAP model based on CLC data (Zulian et al. 2013) raises questions about the reliability of the Maxent results, particularly given the high prediction of urban fabric. The low prediction for arable land is in line with the literature, as important arable wild plants have been lost to intensification of arable farming, and with them the livelihoods of many bees (Westrich 2019; Zulian et al. 2013). The natural grassland, which is moderately scored by Maxent, receives one of the highest ratings for flowering plant availability and nesting site availability in ESTIMAP and is generally very important to many bees as there are many pollen sources available. The rating of forest edges having a lower predictive value than forest cores is questionable, as this finding contrasts with the existing literature, which describes forest edges in particular as suitable habitats for bees. (Westrich 2014; Zulian et al. 2013). The low prediction for the small lake in the north-east of extent 1 is reasonable, as the water bodies themselves are not suitable (Hinsch et al. 2024; Zulian et al. 2013).
It is not possible to make any definitive statements regarding the results obtained with RF. The overall rating for this area was slightly higher, at approximately 0.4, in comparison to the rating for Maxent, which was approximately 0.3. Furthermore, the values exhibit minimal dispersion, spanning a range of 0.3 to 0.5. Only slight differences are evident in the area. While the Maxent results may not completely align with existing literature, it remains unclear whether those of RF are more meaningful in this context. However, it appears that RF is less susceptible to the spatial bias of the occurrence data. The occurrences were more frequent in the city core than in the outer parts, which may have led to the urban fabric being rated so highly by Maxent. This phenomenon is not present in RF.

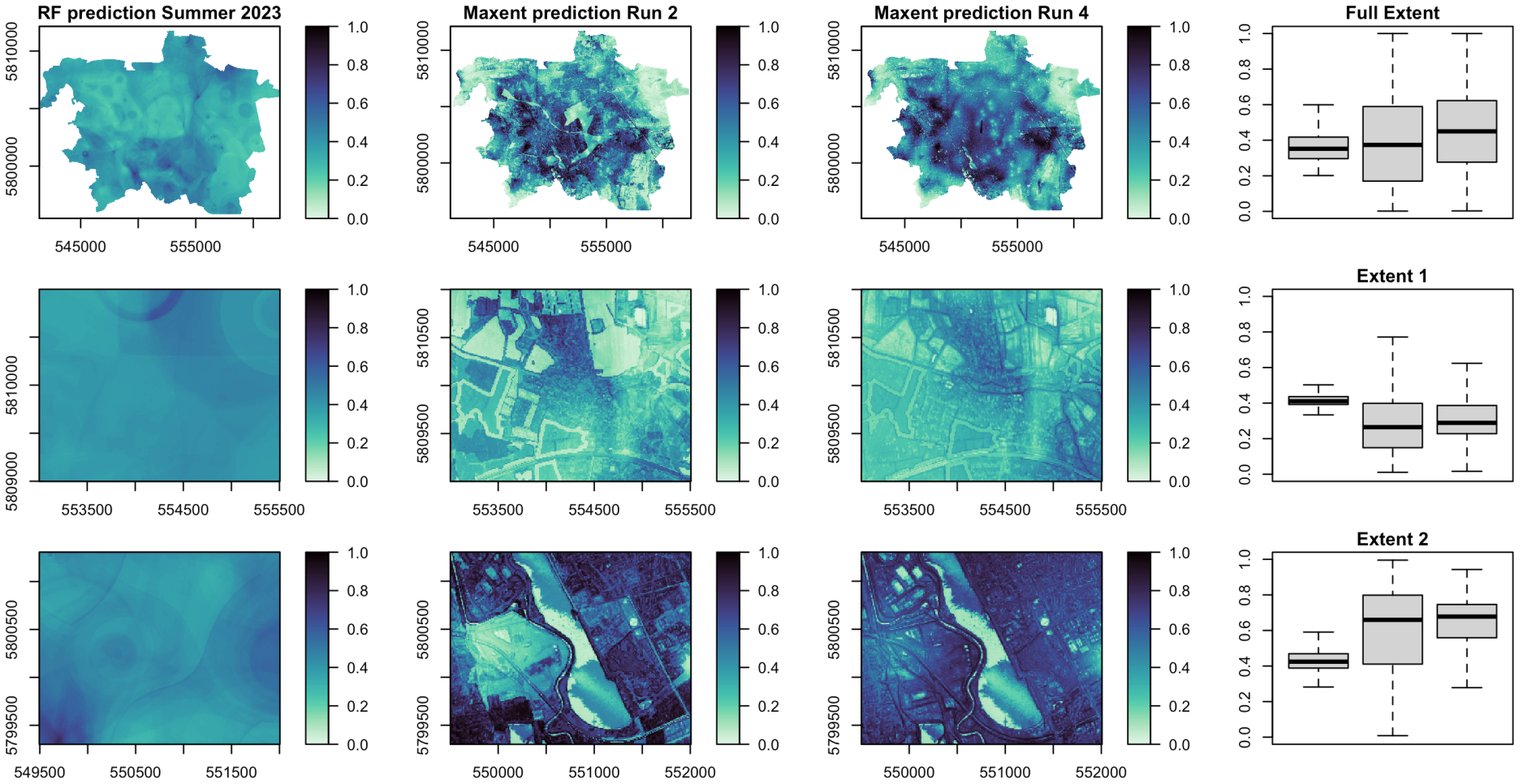


Fig. 16: Comparison of the predictions of the RF and Maxent models for Hannover

A review of the second section of the Maxent results shows that the water area is rated very low in both cases, a finding that is consistent with literature that rates water bodies as unsuitable (Zulian et al. 2013). The forest area and the natural grassland to the west of the lake were assigned low ratings in run 2 but higher ratings in run 4. However, the forest edges are not as clearly delineated as in the first section. Furthermore, it is evident that urban fabric achieves exceptionally high values in this section. In general, the area is distinguished by Maxent with remarkably high prediction values, averaging 0.7. In contrast, RF exhibits an average of 0.4, with instances where values reach up to 0.6. It is therefore slightly more scattered than in the previous section. The literature indicates that urban areas are typically not deemed suitable for nesting but offer certain floral resources (Zulian et al. 2013). Some of the bees native to Germany occur in urban habitats, particularly as natural habitats are being lost in intensively farmed areas. It should be noted, however, that there are certain species that cannot survive in urban environments due to their specific ecological requirements (Westrich 2019). Nevertheless, studies have demonstrated that bees exhibit high species richness and elevated flower visitation rates in urban environments. This underscores the potential of well-managed cities to serve as vital pollination hotspots for urban crops and wildflowers (Theodorou et al. 2020). Still, the high assessment of the urban area by Maxent remains questionable and may be indicative of overfitting and it can be assumed that Maxent is more susceptible to spatial bias. In contrast, the RF analysis did not identify any structures that have emerged from LULC since these variables were not rated as important.

The same analytical approach was employed for the predictions for Lower Saxony, as illustrated in Fig. 17. Run 3 was selected for Maxent, which was executed with the filtered occurrence data and without DisUA as input. Additionally, run 5 was selected, in which LULC, FA and NS data were also removed. Again, two further sections were selected for analysis. The first area of interest is in the north-western region and is characterised by extensive agricultural LULC, including arable land and pastures. The area also encompasses some peat bogs and the town of Oldenburg, which is situated in the north-eastern part of the section. The southern half of the region is characterised by the presence of coniferous forest areas. To the north-east, the river Ems is located. The second section of the study covers the Hannover region. The south-western section includes the city centre and the Deister hill chain. Additionally, the two smaller towns of Celle and Peine are included in this extent, and a part of the city Braunschweig is covered in the east. The northern area is characterised by coniferous forest, whereas the southern part is increasingly defined by the dominance of arable land and pastures.

Upon initial observation, it is evident that the value ranges exhibit a considerable degree of variation. The RF predictions are markedly lower and show minimal dispersion, whereas those of Maxent display a notable increase from run 3 to 5. This discrepancy is less prominent for extent 1 and more evident for extent 2 than for the full extent. This pattern has already been observed in the Hannover study area. It seems reasonable to assume that the difference between the models is caused by the different target variables.

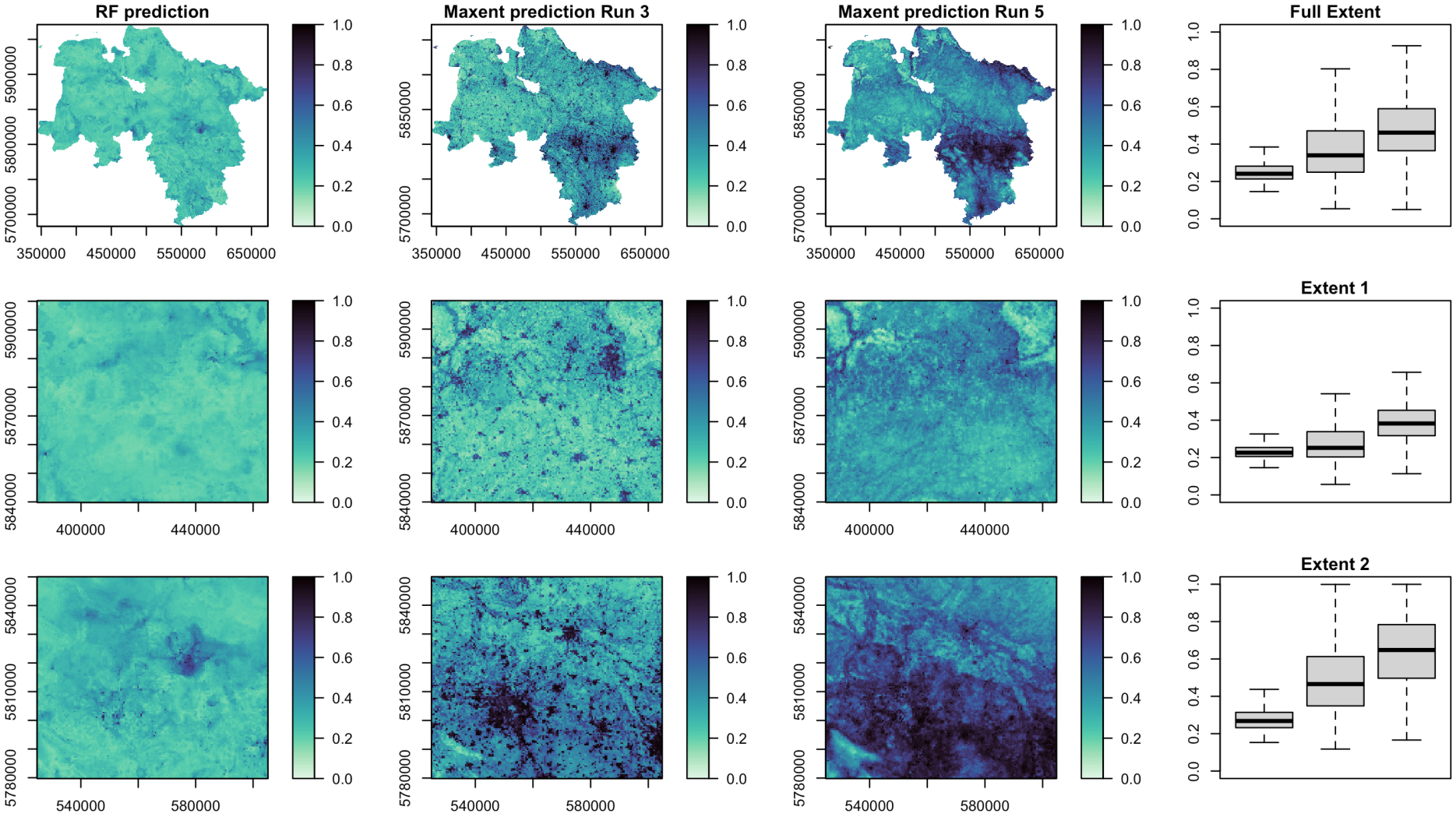


Fig. 17: Comparison of the predictions of the RF and Maxent models for Lower Saxony

The results of run 3 of the Maxent modelling clearly demonstrate that settlement areas were assigned a high rating. This phenomenon is particularly evident in the full extent at the large cities but is also visible in extent 1 and 2. In addition to the large cities, numerous smaller structures with a high prediction are situated on settlement areas. The removal of the LULC data in run 5 serves to diminish the visibility of these structures, yet the regions themselves remain prominent. As the prediction generally increases from run 3 to 5, the urbanised areas are rated even higher than before. As it was the case with Hannover, it is evident that the sharp edges of the prediction are fading and becoming softer. In line with the previous results, pastures and arable land are typically assigned lower ratings. It is notable that forest areas that are well suited to the environment, such as those in the southern part of extent 1, are not identified as such by the model. It is assumed that there is an issue of overfitting due to a spatial bias in the occurrence of the urban fabric, which is reflected in the high prediction values. In contrast to the study area of Hannover, where water areas were assigned a very low prediction, watercourses are given a higher rating in the modelling of Lower Saxony with Maxent. While the riparian areas are important habitats for pollinators, the water areas themselves are not suitable (Westrich 2014, 2019; Zulian et al. 2013). The Maxent model identifies the elevation of the Deister and distinguishes it from its surrounding area, with a lower predicted value. Given the observed decline in temperature and increase in precipitation at this elevation, the prediction is consistent with the existing literature, which indicates that the majority of wild bees require sunlight, warmth, and dry conditions to thrive (Westrich 2019).

An examination of the predictions derived from RF modelling also reveals the presence of varying structures, yet no discernible patterns that can be attributed to LULC. This is because the modelling process did not incorporate data from LULC, FA and NS, nor NDVI. Of particular significance were the bioclimatic variables, the elevation model and soil composition. At extent 1 and 2, a small number of grid cells can be identified that exhibit exceptionally high prediction values yet display notable contrasts to their surrounding areas. It seems reasonable to posit that these cells were included in the training dataset and that the assignment of higher relative abundance data was more straightforward in these locations. This is particularly true given that these grid cells are situated in urban areas, which often exhibited a high abundance in the occurrence data. Consequently, it can be concluded that although RF is generally not susceptible to the spatial bias of the occurrence data and does not typically rank urban areas highly, structures can also be identified here that are associated with the training data.

In general, it can be concluded that Maxent is highly sensitive to spatial bias in occurrence data, resulting in both overfitting of LULC data and overestimation of urban areas, despite the data having gone through a filtering process. This phenomenon is not observed in RF, where only isolated structures of the training data can be identified. It would be beneficial to additionally verify the results with sampling data collected in a standardised sampling design, given that the GBIF data is subject to a significant degree of bias.

The process of filtering GBIF data in order to minimise spatial bias represents a crucial step in the modelling process, with the potential to significantly enhance the quality of HSM. One proven method is to rasterise the study area and randomly select a maximum number of records per raster cell. Another is to select a radius within which a maximum of 1 record may be present. Both methods are similar and result in a reduction of occurrence data, but at the same time have been able to improve the quality of the model, e.g. to better predict less sampled areas (Beck et al. 2014; Boria et al. 2014; Kramer-Schadt et al. 2013).

In general, high outliers were removed when generating the training data for RF modelling. The R package *sampbias* was used to further refine the method. Rasters were created showing the sampling bias in relation to roads and settlement areas, as it is assumed that a higher sampling rate can be expected due to the accessibility and higher population in these locations. It was shown that there is a clear bias in the bee occurrence data in Lower Saxony. The occurrence data were filtered based on these rasters. Grid cells with a high sampling bias were filtered accordingly and conversely those with a low bias were filtered little or not filtered at all. This allowed even more targeted filtering of the data.

In this filtering process, different resolutions of the sampling grid were tried, as well as different rescale ranges to adjust the intensity of the filtering. This step proved to be useful as it resulted in noticeable differences. The highest rescale range was chosen as it produced the most uniform result in relation to the rasters, but at the same time removed the most occurrence records. The quality of the RF model could be improved by the filtering process in terms of the validation parameters. In the case of Maxent, the AUC also showed that filtering brought a slight improvement, although only about a quarter of the original occurrence records were used for further modelling. At the same time, no clear statement can be made about the resolution and rescale ranges used. These should be tested depending on the data basis, study area and number of records in order to select the best option for the model.

Filtering the data not only significantly reduced the number of occurrence records, but also the sample size of the training data. The aim would be to adjust the filter function so that the sample size is not reduced too much despite the removal of records, as a larger sample size is usually better for model quality (Jung 2022). This could be achieved by specifying an even lower resolution when creating the bias raster. The results have shown that the number of records removed is significantly affected by the rescale range. The sample size, on the other hand, is more clearly influenced by the resolution. Further investigation using a lower resolution might also provide clearer results regarding the validation parameters.

The importance values can also be used to make statements about the filtering process. DisUA was included as a control measure to provide information on the spatial bias. The importance of this predictor was significantly reduced by increasing the rescale range when modelling with RF. CLC5 also became less relevant because of filtering. CLCBB, FA and NS were already of low importance but were reduced to almost zero by filtering. The importance values of the Maxent

model show similar results for DisUA. However, this variable is rated much higher and even if it loses relevance through filtering, this value remains the highest rated.
The reduction in the relevance of DisUA in both models through spatial filtering shows that this process worked in terms of reducing spatial bias due to proximity to roads and settlement areas. Spatial filtering can reduce overfitting by addressing spatial biases in occurrence data. These biases often arise from geographic clustering of sampled localities, which can cause models to overfit to specific environmental features associated with these clusters (Boria et al. 2014). Filtering helps to minimise such biases, resulting in a model that is less influenced by localised environmental variables such as those represented by LULC datasets such as CLC5 and CLCBB. It is important to reduce spatial clustering in occurrence data to improve model calibration, especially if there is a strong sampling bias towards certain LULC types (Kramer-Schadt et al. 2013). The reduced importance of LULC variables in the RF model after filtering may be a direct result of such spatial bias reduction, supporting more robust and generalisable model predictions. One explanation for the fact that DisUA and also the LULC variables were rated much higher in the Maxent model than in the RF model is that Maxent is more prone to overfitting than RF (Cutler et al. 2012; Phillips et al. 2006).
It is important to note that when filtering spatially, the presence of spatial clustering may be a factor for certain species with limited distribution ranges. The removal of this ecological signal would result in a weakening of the prediction (Kramer-Schadt et al. 2013). As all bee families were considered in this case, this is an insignificant factor. However, it should be considered when modelling certain species.
While *sampbias* has been established to represent the sampling bias in study areas (Bogoni et al. 2022; Poester-Carvalho et al. 2023), the features of this package have not yet been used to reduce this bias. This thesis has shown that this method can be a successful approach to filter the bias in relation to, e.g. roads and settlement areas, as the quality of the model can be improved further in addition to removing the outliers.
One advantage of RF over Maxent is that, in addition to periodic climate data, more specific data can be incorporated. The results have demonstrated that, while this does not necessarily result in an improved model, it can still be advantageous for certain applications. This includes the analysis of habitat suitability in a changing climate or the estimation of flight times for specific bee species. In the case of Maxent, a model needs to be created for each period of climate data used, with the occurrence data for the corresponding period incorporated into the training process. In contrast, when using RF, a model can be created with all occurrences. The climate data only needs to be attached to the corresponding occurrence. This model can then be projected onto the data for different time periods, as demonstrated in the Hannover case study presented in this thesis. The method of utilising a trained model on additional climate data is frequently employed in the assessment of species distribution in future climate scenarios (Rahimi et al. 2021).

Finally, it was determined which input variables are highly relevant in the model and therefore also for HSM. As the focus of this thesis has been on modelling with RF, the importance of the RF models will be primarily evaluated in the following. The first input parameters were the UA for Hannover and the CLC5 for Lower Saxony as LULC datasets with many classes and additionally the CLCBB for both study areas as LULC dataset with few classes. It is interesting to note that the LULC data were not considered important for RF modelling, as these data form the basis for many process-based models and were also considered most important by Maxent. It has already been concluded that this is a result of overfitting due to the bias of the occurrences. It is therefore even more interesting that RF does not consider either a few or many LULC classes to be important. This result could lead to a completely new approach of modelling pollinator habitat suitability. In any case, this is supported by the fact that LULC itself often does not provide information on whether corresponding nesting and foraging resources are available for bees (Westrich 2014, 2019). Hinsch et al. (2024) addressed this issue by including ecosystem condition in their ESTIMAP-based HSM in the Hannover region. For further research, it would be interesting to consider this ecosystem condition as an input variable for RF. A similar picture emerges from the analysis of FA and NS. The expert-based assessment of food and nesting resources for bees is classified as unimportant in the RF modelling. In Maxent, however, they are more important, especially in Lower Saxony. A similar behavior as with the LULC data is observed, since the expert assessment is based on the same CLC classes.

The next parameters are the distance-based variables DisFE, DisNA and DisRA regarding forest edges, natural areas and riparian areas. In Lower Saxony, these were not considered important with RF, while they are more important in Hannover. The reason for this could be the different resolution of the input data for the two study areas. The forest edges were calculated with a width of 50 m and the riparian areas with only 25 m. The spatial resolution for Lower Saxony was 500 m, so these small structures are not noticeable here, whereas they are much more obvious in Hannover with a resolution of 10 m. The results also showed that smaller distances from riparian areas, forest edges and natural areas tend to have higher prediction values, which is in line with ESTIMAP's expert opinion (Zulian et al. 2013). DisUA as distance to settlements was considered unimportant by RF, again indicating firstly that the filtering process to remove spatial bias was successful and secondly that there was no overfitting regarding settlements.

The following input variables were the geomorphological data, DTM, Slope, and Aspect. In Hannover, all three variables were deemed to have minimal importance. However, in Lower Saxony, both DTM and Slope were identified as significant factors. One potential explanation for the lack of importance attributed to these parameters in Hannover is the relatively limited range of altitudes within the city, in comparison to the more pronounced elevation gradients observed in the south of Lower Saxony. At Maxent, only Slope was of greater importance in both study areas. The decline in temperature and increase in precipitation that occurs with

elevation, coupled with the fact that the majority of wild bees require sunlight, warmth, and dry conditions to thrive, explains the lower predictions in this area (WESTRICH 2019).
The NDVI was classified as unimportant in both study areas by the RF model. The index is used to characterise the productivity of plant communities. HONCHAR (2020) found that higher NDVI was associated with significant flower diversity and high diversity of wild bees. In this case, a mean NDVI over a longer time period was used. It may be beneficial to test more precise data, for example, covering the NDVI during the bees' flight times.
The subsequent variables comprise data characterising the soil texture, specifically the proportions of Clay, Silt and Sand. Except for Clay in Lower Saxony, all proportions were identified as important in both study areas through the use of RF modelling. The importance of soil texture in Maxent modelling was relatively low. Soil texture is a relevant factor for bees, as many wild bee species are soil nesters. Bees are often associated with sandy soils, but there are significant differences in the preferences of different species (ANTOINE & FORREST 2021). However, the scatterplot for Lower Saxony indicates a trend whereby sandy soils tend to have higher prediction values, which would support this assumption. More precise results would likely be obtained if specific bee species with similar nesting behaviours were modelled.
The climate data, including Sunshine, Precipitation, Drought, TempMin, TempMean, TempMax and Radiation, were identified as important for both study areas in the RF modelling and were assigned the highest values. While Precipitation and Drought were identified as particularly important in Lower Saxony, Sunshine and TempMax were found to be more significant in Hannover. This difference may be attributed to the handling of seasonal climate data, as Precipitation and Drought exhibited greater importance before incorporating seasonal climate data. The most evident patterns in the scatterplots were observed in Lower Saxony. Here, the highest temperatures were associated with the highest prediction values, while low precipitation and drought were correlated with high predictions. Additionally, a trend emerged for Radiation and Sunshine, where higher values tended to generate higher predictions. These findings align with the existing literature, which indicates that bees prefer warm temperatures, dry conditions, and sunshine (ANTOINE & FORREST 2021; WESTRICH 2019).
The final two variables, SoilTemp and SoilMoist, describe the temperature and moisture of the soil. In Lower Saxony, both parameters are of significance when modelling with RF. In Hannover, however, only SoilMoist is of importance. One possible explanation for this observation is the relatively small differences in Hannover. Most ground-nesting bees have a preference for soil with a well-drained but not excessively dry composition. This finding is supported by the scatterplots, which indicate that lower SoilMoist values result in higher prediction values. Additionally, the influence of soil temperature on soil-nesting bees is evident, with warmer temperatures positively affecting adult bee activity and the development rate and survival of bees in the larval stage (ANTOINE & FORREST 2021). This is validated by the low SoilTemp values, which correspond to the lowest prediction values.

The results demonstrated that RF, in contrast to Maxent, does not exhibit a pronounced response to the spatial bias inherent in the occurrence data sourced from GBIF. Consequently, the phenomenon of overfitting was not observed. The strategy of filtering the occurrence data using a bias grid was effective and represents a method for filtering the occurrence data with greater specificity, thereby providing the optimal data foundation. It is important to acknowledge that the spatial bias still had a significant impact on the RF modelling, which ultimately prevented the achievement of a satisfactory model fit. Nonetheless, the results of the RF modelling are largely in line with those reported in existing literature. It can be concluded that the use of HSM of pollinators with RF represents a robust approach to avoid the issue of overfitting due to the spatial bias of GBIF data.

It must be acknowledged that the research design of this thesis is subject to limitations. Although the selection of species density instead of modelling background points was a reasonable approach, it resulted in a more challenging comparison with Maxent modelling. In fact, Maxent could only be used as a rough comparative value. A more targeted approach would have been to also include RF modelling using background points. A comparison between the two RF modelling approaches would be as valuable as the comparison with Maxent and should be considered in subsequent research.

6 Conclusion

The application of RF to model the habitat suitability of pollinators has been demonstrated to be an effective approach, employing GBIF occurrence data, which is characterised by a high spatial bias and thus stands out from Maxent, which is prone to overfitting. Whether the modelling of occurrence density is more effective than the use of background points remains to be investigated. Despite the inability to achieve a satisfactory model fit, the prediction results are a promising indication of the potential of this approach.

This thesis presents a methodology for assessing the habitat suitability of pollinators in Lower Saxony and Hannover, based on the use of a RF model. The model was constructed using a variety of data on LULC, geomorphology, vegetation, soil and bioclimate, which were collected and processed in a uniform raster format. Furthermore, the GBIF database was utilised to gather presence-only occurrence data of bees. The occurrence data were employed in the construction of the training dataset for the model, which was represented as relative density of occurrences. The RF model was then further developed in various iterations. Different resolutions for the training data were tested, an approach to filter the occurrence data to remove the spatial bias, the inclusion of more temporally resolved climate data and then a fine-tuning of the model was applied. The final model was used to predict the corresponding study areas in terms of habitat suitability for pollinators. To enable a comparison, the same input data was used to create a prediction using the Maxent model.

The first research question to be answered in this thesis was how pollinator habitat suitability is assessed in the study areas using a ML approach. While Maxent focused intensively on the LULC data due to overfitting regarding urban areas, resulting in very hard edges and an overvaluation of settlement areas, RF was able to produce a more homogeneous result in which these structures were not identified again. However, the use of relative occurrence density results in a narrow range of predicted values. An area-wide abundance was determined, though within the lower range. Including locations with absences in the training data may facilitate the extension of the range and the achievement of a more pronounced variation across the area in question. The model fit of RF was trivial, yet it still provided useful results that could be built upon. Maxent obtained a good fit but was affected by overfitting. The potential of integrating climate data with a higher temporal resolution into the RF model was found to be advantageous in certain cases, but it does not necessarily result in a superior model.

Furthermore, it should be explored how the spatial bias in the occurrence data can be reduced and how this affects the modelling. Although the filtering of heavily clustered occurrence data to reduce spatial bias is a common approach, a method was developed for performing this filtering in a more targeted way. The bias raster, created in relation to roads and settlement areas, enabled the bias to be represented in its distribution and subsequently filtered according to this raster. This approach facilitated the identification of locations with a high bias, which were then filtered accordingly. The filtering had a positive effect on the modelling for both RF

and Maxent, as showed by an improvement in the validation parameters despite the greatly reduced number of occurrences.

Finally, it should be determined which input variables are highly relevant in the model and therefore also for HSM. The bioclimatic variables achieved highest importance when modelling with RF. It became evident that a high prediction was achieved with high temperatures, low precipitation and high solar radiation, which aligns with the documented requirements of bees. Furthermore, the preference for sandy soils was supported. The distances to forest edges, riparian areas and natural areas are only relevant if the resolution of the input data allows for these features to be differentiated. In contrast, the LULC data, assessments of nesting and food resources, and the NDVI were not considered to be important. In contrast to Maxent, these were the variables that were identified as important. Therefore, an approach could be presented for RF that differs significantly in terms of input data from common process-based models, where LULC data often form the basis of modelling.

The limitations of this study can be attributed to the choice of research design, which proved inefficient. The decision to model species density presented a significant challenge when attempting to make a comparison with Maxent. Furthermore, there is a lack of evaluation of the usefulness of species density modelling using GBIF data. It would have been beneficial to incorporate RF modelling with background points, enabling not only a comparison between the two RF approaches but also an enhanced comparison with Maxent.

It would also be beneficial to validate the findings with supplementary occurrence data gathered through a consistent sampling design. Given the low relevance of the LULC data, it would be helpful to include ecosystem condition as a factor, as this may provide a more comprehensive consideration of the diverse ecosystem types. The methods employed in this study was based on modelling all bee families. To obtain more precise results in relation to the various input variables, it would be informative to model specific bee species, as their habitat requirements differ considerably. This approach could improve the model fit.

Although the HSM developed in this study offers valuable insights into the relationship between pollinators and their environment, it is essential to acknowledge that no model can fully capture the complexities of nature. The RF model employed in this study captures significant interactions between important environmental variables and pollinator presence. However, it is a simplified representation of the underlying ecological dynamics. The model's strength lies in its capacity to demonstrate how different variables contribute to the prediction of suitable habitats, thereby offering practical guidance for conservation and habitat management. The results underscore the value of predictive modelling as a tool for understanding ecological systems. However, it is important to recognise that the model should not be viewed as an exact reflection of nature. Instead, it should be considered as a robust basis for understanding the relationship between pollinators and their habitats.

In conclusion, the presented RF model provides a robust foundation for HSM of pollinators at the local and regional scales, based on GBIF data, which can be further developed.

7 References

AG Boden (2005): Bodenkundliche Kartieranleitung: mit 41 Abbildungen, 103 Tabellen und 31 Listen. In: Bundesanstalt für Geowissenschaften und Rohstoffe und den Geologischen Landesämtern in der Bundesrepublik Deutschland Hannover*: Vol. 5.* Stuttgart; 438.

Antoine, C. M., & Forrest, J. R. K. (2021): Nesting habitat of ground-nesting bees: a review. Ecological Entomology, 46(2), 143–159. https://doi.org/10.1111/EEN.12986

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012): Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution, 3(2), 327–338. https://doi.org/10.1111/J.2041-210X.2011.00172.X

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014): Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics, 19, 10–15. https://doi.org/10.1016/J.ECOINF.2013.11.002

BKG (2021): Dokumentation Digitales Geländemodell Gitterweite 200 m. Leipzig; 7. http://www.crs-geo.eu/crs-national.htm

BKG (2022): Dokumentation - CORINE Land Cover 5 ha - CLC5 (2018). Leipzig; 7. https://sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/clc5_2018.pdf

Bogoni, J. A., Peres, C. A., & Ferraz, K. M. P. M. B. (2022): Medium- to large-bodied mammal surveys across the Neotropics are heavily biased against the most faunally intact assemblages. Mammal Review, 52(2), 221–235. https://doi.org/10.1111/MAM.12274

Bonaccorso, G. (2017): Machine learning algorithms: reference guide for popular algorithms for data science and machine learning. Birmingham; 360.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014): Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling, 275, 73–77. https://doi.org/10.1016/J.ECOLMODEL.2013.12.012

Breiman, L. (2001): Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Burkhard, B., & Maes, J. (2017): Mapping Ecosystem Services. In: Advanced Books. Sofia; 374. https://doi.org/10.3897/AB.E12837

Büttner, G., Kosztra, B., Maucha, G., Pataki, R., Kleeschulte, S., Hazeu, G., Vittek, M., & Littkopf, A. (2021): Copernicus Land Monitoring Service CORINE Land Cover User Manual. Copenhagen. https://land.copernicus.eu/en/technical-library/clc-product-user-manual/@@download/file

Cane, J. H., & Neff, J. L. (2011): Predicted fates of ground-nesting bees in soil heated by wildfire: Thermal tolerances of life stages and a survey of nesting depths. Biological Conservation, 144(11), 2631–2636. https://doi.org/10.1016/J.BIOCON.2011.07.019

Chicco, D., Warrens, M. J., & Jurman, G. (2021): The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623/FIG-1

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012): Random Forests. In: C. Zhang & Y. Ma (Eds.): Ensemble Machine Learning: Methods and Applications. New York. https://doi.org/10.1007/978-1-4419-9326-7_5

DLR (2023): *Sentinel-2 - Vegetation Index (NDVI) - Germany, 2015*. https://doi.org/https://doi.org/10.15489/z5rq0pr8wv85

DWD (2018): Klimareport Niedersachsen. Offenbach am Main; 52.

Džeroski, S. (2009): Machine Learning Applications in Habitat Suitability Modeling. In: S. E. Haupt, A. Pasini, & C. Marzban (Eds.): Artificial Intelligence Methods in the Environmental Sciences. Dordrecht. https://doi.org/10.1007/978-1-4020-9119-3_19

EEA. (2022): CLC+ Backbone Product Specification and User Manual: Raster Product. 52. https://land.copernicus.eu/en/technical-library/product-user-manual-for-clc-backbone-raster-only/@@download/file

Ellenberg, H., & Leuschner, C. (2010): Vegetation Mitteleuropas mit den Alpen: in ökologischer, dynamischer und historischer Sicht (6th ed.). Stuttgart; 1333.

European Union. (2020): Mapping Guide v6.3 for a Urban Atlas. 47. https://land.copernicus.eu/en/technical-library/urban_atlas_2012_2018_mapping_guide/@@download/file

Evertsbusch, S., Prause, D., Dieler, S., Sbresny, J., & Gehrt, E. (2021): Erläuterungen zur Bodenkarte von Niedersachsen 1 : 50.000 (BK50). Informationen zu den Bodenflächendaten - Technische Dokumentation -. Hannover; 39.

Fu, H., & Qi, K. (2022): Evaluation Model of Teachers' Teaching Ability Based on Improved Random Forest with Grey Relation Projection. Scientific Programming, 2022. https://doi.org/10.1155/2022/5793459

Galbraith, S. M., Vierling, L. A., & Bosque-Pérez, N. A. (2015): Remote sensing and ecosystem services: Current status and future opportunities for the study of bees and pollination-related services. Current Forestry Reports, 1(4), 261–274. https://doi.org/10.1007/S40725-015-0024-6/FIGURES/4

Gallant, A. L., Euliss, N. H., & Browning, Z. (2014): Mapping Large-Area Landscape Suitability for Honey Bees to Assess the Influence of Land-Use Change on Sustainability of National Pollination Services. PLOS ONE, 9(6), e99268. https://doi.org/10.1371/JOURNAL.PONE.0099268

Gardein, H., Fabian, Y., Westphal, C., Tscharntke, T., & Hass, A. (2022): Ground-nesting bees prefer bare ground areas on calcareous grasslands. Global Ecology and Conservation, 39, e02289. https://doi.org/10.1016/J.GECCO.2022.E02289

Gardner, E., Breeze, T. D., Clough, Y., Smith, H. G., Baldock, K. C. R., Campbell, A., Garratt, M. P. D., Gillespie, M. A. K., Kunin, W. E., McKerchar, M., Memmott, J., Potts, S. G., Senapathi, D., Stone, G. N., Wäckers, F., Westbury, D. B., Wilby, A., & Oliver, T. H. (2020): Reliably predicting pollinator abundance: Challenges of calibrating process-based ecological models. Methods in Ecology and Evolution, 11(12), 1673–1689. https://doi.org/10.1111/2041-210X.13483

Geue, J. C., & Thomassen, H. A. (2020): Unraveling the habitat preferences of two closely related bumble bee species in Eastern Europe. Ecology and Evolution, 10(11), 4773–4790. https://doi.org/10.1002/ECE3.6232

Giannini, T. C., Acosta, A. L., Garófalo, C. A., Saraiva, A. M., Alves-dos-Santos, I., & Imperatriz-Fonseca, V. L. (2012): Pollination services at risk: Bee habitats will decrease owing to climate change in Brazil. Ecological Modelling, 244, 127–131. https://doi.org/10.1016/J.ECOLMODEL.2012.06.035

Giménez-García, A., Allen-Perkins, A., Bartomeus, I., Balbi, S., Knapp, J. L., Hevia, V., Woodcock, B. A., Smagghe, G., Miñarro, M., Eeraerts, M., Colville, J. F., Hipólito, J., Cavigliasso, P., Nates-Parra, G., Herrera, J. M., Cusser, S., Simmons, B. I., Wolters, V., Jha, S., … Magrach, A. (2023): Pollination supply models from a local to global scale. Web Ecology, 23(2), 99–129. https://doi.org/10.5194/we-23-99-2023

Gomes, V. H. F., Ijff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., Coelho, L. D. S., Matos, F. D. D. A., Castilho, C. V., Filho, D. D. A. L., López, D. C., Guevara, J. E., Magnusson, W. E., Phillips, O. L., Wittmann, F., Carim, M. D. J. V., Martins, M. P., Irume, M. V., Sabatier, D., Molino, J. F., … Ter Steege, H. (2018): Species Distribution Modelling: Contrasting presence-only models with plot abundance data. Scientific Reports 2018 8:1, 8(1), 1–12. https://doi.org/10.1038/s41598-017-18927-1

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017): Habitat Suitability and Distribution Models: With Applications in R. In: Habitat Suitability and Distribution Models. Cambridge. https://doi.org/10.1017/9781139028271

Häussler, J., Sahlin, U., Baey, C., Smith, H. G., & Clough, Y. (2017): Pollinator population size and pollination ecosystem service responses to enhancing floral and nesting resources. Ecology and Evolution, 7, 1898–1908. https://doi.org/10.1002/ece3.2765

Herrera, J. M., Ploquin, E. F., Rodríguez-Pérez, J., & Obeso, J. R. (2014): Determining habitat suitability for bumblebees in a mountain system: a baseline approach for testing the impact of climate change on the occurrence and abundance of species. Journal of Biogeography, 41(4), 700–712. https://doi.org/10.1111/JBI.12236

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2023): Package 'dismo'. 68. https://www.gbif.org

Hinsch, M., Zulian, G., Stekker, S., Rega, C., Nabuurs, G. J., Verweij, P., & Burkhard, B. (2024): Assessing pollinator habitat suitability considering ecosystem condition in the Hannover Region, Germany. Landscape Ecology, 39(3). https://doi.org/10.1007/S10980-024-01851-X

HONCHAR, H. (2020): The use of the Normalized Difference Vegetation Index (NDVI) to estimate the diversity of wild bees (Hymenoptera, Apoidea)Honchar H. Ecological Sciences, 1(2), 133–139. https://doi.org/10.32846/2306-9716/2020.ECO.2-29.1.22

HOWARD, C., STEPHENS, P. A., PEARCE-HIGGINS, J. W., GREGORY, R. D., & WILLIS, S. G. (2014): Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution, 5(6), 506–513. https://doi.org/10.1111/2041-210X.12184

JÄGER, E., & HEIPKE, C. (2020): Geotopographie und Photogrammetrie. In: K. Kummer, T. Kötter, H. Kutterer, & S. Ostrau (Eds.): Das deutsche Vermessungs- und Geoinformationswesen. Berlin.

JUNG, A. (2022): Machine Learning. Singapore; 212. https://doi.org/10.1007/978-981-16-8193-6

JUWARIYEM, J., SRIYANTO, S., LESTARI, S., & CHAIRANI, C. (2024): Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms. Sinkron, 8(2), 947–955. https://doi.org/10.33395/SINKRON.V8I2.13448

KAMMERER, M., GOSLEE, S. C., DOUGLAS, M. R., TOOKER, J. F., & GROZINGER, C. M. (2021): Wild bees as winners and losers: Relative impacts of landscape composition, quality, and climate. Global Change Biology, 27(6), 1250–1265. https://doi.org/10.1111/GCB.15485

KASPAR, F., KRATZENSTEIN, F., & KAISER-WEISS, A. K. (2019): Interactive open access to climate observations from Germany. Advances in Science and Research, 16, 75–83. https://doi.org/10.5194/ASR-16-75-2019

KASPAR, F., MÜLLER-WESTERMEIER, G., PENDA, E., MÄCHEL, H., ZIMMERMANN, K., KAISER-WEISS, A., & DEUTSCHLÄNDER, T. (2013): Monitoring of climate change in Germany-data, products and services of Germany's National Climate Data Centre. Adv. Sci. Res, 10, 2012. https://doi.org/10.5194/asr-10-99-2013

KHALIFA, S. A. M., ELSHAFIEY, E. H., SHETAIA, A. A., EL-WAHED, A. A. A., ALGETHAMI, A. F., MUSHARRAF, S. G., ALAJMI, M. F., ZHAO, C., MASRY, S. H. D., ABDEL-DAIM, M. M., HALABI, M. F., KAI, G., AL NAGGAR, Y., BISHR, M., DIAB, M. A. M., & EL-SEEDI, H. R. (2021): Overview of Bee Pollination and Its Economic Value for Crop Production. Insects 2021, Vol. 12, Page 688, 12(8), 688. https://doi.org/10.3390/INSECTS12080688

KLEIN, A. M., VAISSIÈRE, B. E., CANE, J. H., STEFFAN-DEWENTER, I., CUNNINGHAM, S. A., KREMEN, C., & TSCHARNTKE, T. (2007): Importance of pollinators in changing landscapes for world crops. Proceedings of the Royal Society B: Biological Sciences, 274(1608), 303. https://doi.org/10.1098/RSPB.2006.3721

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. https://doi.org/10.1016/J.ECOLMODEL.2017.01.024

KOSICKI, J. Z. (2020): Generalised Additive Models and Random Forest Approach as effective methods for predictive species density and functional species richness. Environmental and Ecological Statistics, 27(2), 273–292. https://doi.org/10.1007/S10651-020-00445-5

Kosicki, J. Z., & Hromada, M. (2018): Cuckoo density as a predictor of functional and phylogenetic species richness in the predictive modelling approach: Extension of Tryjanowski and Morelli (2015) paradigm in the analytical context. Ecological Indicators, 88, 384–392. https://doi.org/10.1016/J.ECOLIND.2018.01.009

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013): The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions, 19(11), 1366–1379. https://doi.org/10.1111/DDI.12096

LGLN (2019): Digitale Geländemodelle (DGM). Hannover; 4. https://www.lgln.niedersachsen.de/download/122447/Produktinformation_DGM.pdf

LGLN (2024): Digitale Landschaftsmodell (DLM). Produktinformation ATKIS®. 4. https://www.lgln.niedersachsen.de/download/126448/Produktinformation_DLM.pdf

Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008): AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17(2), 145–151. https://doi.org/10.1111/J.1466-8238.2007.00358.X

Łowicki, D., & Fagiewicz, K. (2021): A new model of pollination services potential using a landscape approach: A case study of post-mining area in Poland. Ecosystem Services, 52, 101370. https://doi.org/10.1016/J.ECOSER.2021.101370

Manley, K., & Egoh, B. N. (2022): Mapping and modeling the impact of climate change on recreational ecosystem services using machine learning and big data. Environmental Research Letters, 17(5), 054025. https://doi.org/10.1088/1748-9326/AC65A3

Marshall, L., Carvalheiro, L. G., Aguirre-Gutiérrez, J., Bos, M., de Groot, G. A., Kleijn, D., Potts, S. G., Reemer, M., Roberts, S., Scheper, J., & Biesmeijer, J. C. (2015): Testing projected wild bee distributions in agricultural habitats: predictive power depends on species traits and habitat type. Ecology and Evolution, 5(19), 4426–4436. https://doi.org/10.1002/ECE3.1579

Meier, S., Walz, U., Syrbe, R. U., & Grunewald, K. (2021): Das bundesweite Habitatpotenzial für Wildbienen - Ein Indikator für die Bestäubungsleistung. Naturschutz Und Landschaftsplanung, 53(6), 12–19. https://doi.org/10.1399/NUL.2021.06.01

Merow, C., Smith, M. J., & Silander, 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. https://doi.org/10.1111/J.1600-0587.2013.07872.X

Mi, C., Huettmann, F., Sun, R., & Guo, Y. (2017): Combining occurrence and abundance distribution models for the conservation of the Great Bustard. PeerJ, 2017(12). https://doi.org/10.7717/PEERJ.4160/SUPP-3

Moens, M., Biesmeijer, J. C., Huang, E., Vereecken, N. J., & Marshall, L. (2024): The importance of biotic interactions in distribution models of wild bees depends on the type of ecological relations, spatial scale and range. Oikos, e10578. https://doi.org/10.1111/OIK.10578

Moens, M., Biesmeijer, J. C., Klumpers, S. G. T., & Marshall, L. (2023): Are threatened species special? An assessment of Dutch bees in relation to land use and climate. Ecology and Evolution, 13(7). https://doi.org/10.1002/ECE3.10326

Nightingale, J. M., Esaias, W. E., Wolfe, R. E., Nickeson, J. E., & Ma, P. L. A. (2008): Assessing honey bee equilibrium range and forage supply using satelite-derived phenology. International Geoscience and Remote Sensing Symposium (IGARSS), 3(1), 1–4. https://doi.org/10.1109/IGARSS.2008.4779460

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., & Louzao, M. (2012): Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation, 156, 94–104. https://doi.org/10.1016/J.BIOCON.2011.11.013

Pashanejad, E., Thierry, H., Robinson, B. E., & Parrott, L. (2023): The application of semantic modelling to map pollination service provisioning at large landscape scales. Ecological Modelling, 484, 110452. https://doi.org/10.1016/J.ECOLMODEL.2023.110452

Pearce, J., & Ferrier, S. (2001): The practical value of modelling relative abundance of species for regional conservation planning: a case study. Biological Conservation, 98(1), 33–43. https://doi.org/10.1016/S0006-3207(00)00139-7

Perennes, M., Diekötter, T., Groß, J., & Burkhard, B. (2021): A hierarchical framework for mapping pollination ecosystem service potential at the local scale. Ecological Modelling, 444, 109484. https://doi.org/10.1016/J.ECOLMODEL.2021.109484

Phillips, S., Aneja, V. P., Kang, D., & Arya, S. P. (2006): Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3–4), 231–259. https://doi.org/10.1016/J.ECOLMODEL.2005.03.026

Phillips, S., Dudík, M., & Schapire, R. E. (2004): A maximum entropy approach to species distribution modeling. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 655–662.

Poester-Carvalho, J. A., Barão, K. R., da Costa, L. G., & Ferrari, A. (2023): Areas of endemism and sampling bias of Pentatomidae (Heteroptera) in the Americas. Journal of Insect Conservation, 27(5), 781–794. https://doi.org/10.1007/S10841-023-00497-5/TABLES/2

Polce, C., Garratt, M. P., Termansen, M., Ramirez-Villegas, J., Challinor, A. J., Lappage, M. G., Boatman, N. D., Crowe, A., Endalew, A. M., Potts, S. G., Somerwill, K. E., & Biesmeijer, J. C. (2014): Climate-driven spatial mismatches between British orchards and their pollinators: increased risks of pollination deficits. Global Change Biology, 20(9), 2815–2828. https://doi.org/10.1111/GCB.12577

Polce, C., Maes, J., Rotllan-Puig, X., Michez, D., Castro, L., Cederberg, B., Dvorak, L., Fitzpatrick, Ú., Francis, F., Neumayer, J., Manino, A., Paukkunen, J., Pawlikowski, T., Roberts, S. P. M., Straka, J., & Rasmont, P. (2018): Distribution of bumblebees across europe. One Ecosystem, 3, e28143. https://doi.org/10.3897/ONEECO.3.E28143

Polce, C., Termansen, M., Aguirre-Gutié Rrez, J., Boatman, N. D., Budge, G. E., Crowe, A., Garratt, M. P., Phane Pietravalle, S., Potts, S. G., Ramirez, J. A., Somerwill, K. E., & Biesmeijer, J. C. (2013): *Species Distribution Models for Crop Pollination: A Modelling Framework Applied to Great Britain*. https://doi.org/10.1371/journal.pone.0076308

Probeck, M., Ruiz, I., Ramminger, G., Fourie, C., Maier, P., Ickerott, M., Storch, C., Homolka, A., Muller, S. J., Tiwari, H., Stumpf, A., Chun, S., Mattos, C., Lindmayer, A., Jahangir, F., Endara, P., Berndt, F., Dohr, M., Kapferer, W., … Dufourmont, H. (2021): CLC+ BACKBONE: SET THE SCENE IN COPERNICUS FOR THE COMING DECADE. International Geoscience and Remote Sensing Symposium (IGARSS), 2021-July, 2076–2079. https://doi.org/10.1109/IGARSS47720.2021.9553252

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019): Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/WIDM.1301

Radosavljevic, A., & Anderson, R. P. (2014): Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of Biogeography, 41(4), 629–643. https://doi.org/10.1111/JBI.12227

Radzevičiūtė, R., Theodorou, P., Schlegel, M., & Paxton, R. J. (2021): A two-part modelling approach reveals a positive effect of pollinator biodiversity in boosting the pollination of apple flowers. Agriculture, Ecosystems & Environment, 306, 107197. https://doi.org/10.1016/J.AGEE.2020.107197

Rahimi, E., Barghjelveh, S., & Dong, P. (2021): Estimating potential range shift of some wild bees in response to climate change scenarios in northwestern regions of Iran. Journal of Ecology and Environment, 45(1), 1–13. https://doi.org/10.1186/S41610-021-00189-8/TABLES/5

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015): Point process models for presence-only analysis. Methods in Ecology and Evolution, 6(4), 366–379. https://doi.org/10.1111/2041-210X.12352

Scowen, M., Athanasiadis, I. N., Bullock, J. M., Eigenbrod, F., & Willcock, S. (2021): The current and future uses of machine learning in ecosystem service research. Science of The Total Environment, 799, 149263. https://doi.org/10.1016/J.SCITOTENV.2021.149263

Sillero, N., Campos, J. C., Arenas-Castro, S., & Barbosa, A. M. (2023): A curated list of R packages for ecological niche modelling. Ecological Modelling, 476, 110242. https://doi.org/10.1016/J.ECOLMODEL.2022.110242

Svenningsen, C. S., & Schigel, D. (2024): Sharing insect data through GBIF: novel monitoring methods, opportunities and standards. Phil. Trans. R. Soc. B, 379(20230104). https://doi.org/10.1098/rstb.2023.0104

Theodorou, P., Radzevičiūtė, R., Lentendu, G., Kahnt, B., Husemann, M., Bleidorn, C., Settele, J., Schweiger, O., Grosse, I., Wubet, T., Murray, T. E., & Paxton, R. J. (2020): Urban areas as hotspots for bees and pollination but not a panacea for all insects. Nature Communications 2020 11:1, 11(1), 1–13. https://doi.org/10.1038/s41467-020-14496-6

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021): Modelling species presence-only data with random forests. Ecography, 44(12), 1731–1742. https://doi.org/10.1111/ECOG.05615

Warton, D. I., & Shepherd, L. C. (2010): Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. Annals of Applied Statistics, 4(3), 1383–1402. https://doi.org/10.1214/10-AOAS331

Watts, E. F., Waldron, B. P., & Kuchta, S. R. (2024): Hard edges, soft edges, and species range evolution: A genomic analysis of the Cumberland Plateau salamander. Journal of Biogeography, 00, 1–12. https://doi.org/10.1111/JBI.14962

Wentling, C., Campos, F. S., David, J., & Cabral, P. (2021): Pollination potential in Portugal: Leveraging an ecosystem service for sustainable agricultural productivity. Land, 10(4). https://doi.org/10.3390/LAND10040431

Westrich, P. (2014): Wildbienen. Die anderen Bienen (4. Aufl.). München; 168.

Westrich, P. (2019): Die Wildbienen Deutschlands (2nd ed.). Stuttgart (Hohenheim); 821.

Wright, M. N., & Ziegler, A. (2017): Ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(1). https://doi.org/10.18637/JSS.V077.I01

Yuan, H. (2023): Current perspective on artificial intelligence, machine learning and deep learning. Applied and Computational Engineering, 19(1), 116–122. https://doi.org/10.54254/2755-2721/19/20231019

Zizka, A., Antonelli, A., & Silvestro, D. (2021): sampbias, a method for quantifying geographic sampling biases in species distribution data. Ecography, 44(1), 25–32. https://doi.org/10.1111/ECOG.05102

Zöller, L., Beierkuhnlein, C., Faust, D., Eitel, B., & Samimi, C. (2017): Die physische Geographie Deutschlands (L. Zöller, Ed.). Darmstadt; 208.

Zulian, G., Paracchini, M. L., Maes, J., & Liquete, C. (2013): ESTIMAP: Ecosystem services mapping at European scale. Ispra; 54. https://doi.org/10.2788/64369

Zulian, G., Polce, C., & Maes, J. (2014): ESTIMAP: A GIS-based model to map ecosystem services in the European Union. Annali Di Botanica, 4, 1–7. https://doi.org/10.4462/ANNBOTRM-11807

Appendix

Tab. A 1: Overview of the geodata used in this thesis

Publisher	Dataset	Reference
BKG	CORINE Land Cover 5 ha 2018	© GeoBasis-DE/BKG 2024
BKG	Digital terrain model 200 m	© GeoBasis-DE/BKG 2024
CLMS	Urban Atlas Hannover 2018	© EU, CLMS 2018, EEA
CLMS	CLC+Backbone 2018	© EU, CLMS 2018, EEA
DLR	Sentinel-2 NDVI Germany 2015	© DLR Sentinel-2 NDVI 2015 Germany
DWD	Multi-annual grids of annual sunshine duration over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seasonal grids of sum of sunshine duration over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Multi-annual grids of precipitation height over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seasonal grids of sum of precipitation over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Multi-annual grids of drought index (de Martonne) over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seaonal grids of sum of drought index (de Martonne) over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Multi-annual grids of monthly averaged daily minimum air temperature (2m) over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seasonal grids of monthly averaged daily minimum air temperature (2m) over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Multi-annual grids of monthly averaged daily mean air temperature (2m) over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seasonal grids of monthly averaged daily mean air temperature (2m) over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Multi-annual grids of monthly averaged daily maximum air temperature (2m) over Germany 1991-2020	© DWD CDC, version v1.0, 2018
DWD	Seasonal grids of monthly averaged daily maximum air temperature (2m) over Germany 2010-2024	© DWD CDC, version v1.0, 2018
DWD	Gridded multi annual monthly mean sums and multi annual yearly mean sum of incoming shortwave radiation (global radiation) on the horizontal plain for Germany based on ground and satellite measurements 1991-2020	© DWD CDC, version V003, 2024
DWD	Multi-annual grids of soil temperature in 5 cm depth under uncovered soil 1991-2020	© DWD CDC, version 0.x, 2024
DWD	Multi-annual grids of soil moisture in 5cm depth under grass and sandy loam 1991-2020	© DWD CDC, version 0.x, 2024
LBEG	BK50 Sachdaten	© LBEG, Germany, 2024
LGLN	Digital terrain model 1 m	© GeoBasis-DE/LGLN 2024, CC-BY 4.0
LGLN	Basis-DLM	© GeoBasis-DE/LGLN 2024, CC-BY 4.0

Tab. A 2: Selected CLC5 classes in Lower Saxony and UA classes in Hannover to create the distance raster

	CLC5 Lower Saxony		Urban Atlas Hannover	
Artificial areas	111	Continuous urban fabric	11100	Continuous urban fabric
			11210	Discontinuous dense urban fabric
			11220	Discontinuous medium urban fabric
	112	Discontinuous urban fabric	11230	Discontinuous low density urban fabric
			11240	Discontinuous very low density urban fabric
			11300	Isolated structures
	121	Industrial or commercial units	12100	Industrial, commercial, public, military and private units
	122	Road and rail networks and associated land	12210	Fast transit roads and associated land
			12220	Other roads and associated land
			12230	Railways and associated land
	123	Port areas		
	124	Airports		
	131	Mineral extraction site:	13100	Mineral extraction and dump sites
	132	Dump sites		
	133	Construction sites	13300	Construction sites
			13400	Land without current use
	141	Green urban areas	14100	Green urban areas
	142	Sport and leisure facilities	14200	Sports and leisure facilities
Agricultural areas	211	Non-irrigated arable land	21000	Arable land (annual crops)
	222	Fruit trees and berry plantations		
	231	Pastures	23000	Pastures
Forest	311	Broad-leaved forest	31000	Forests
	312	Coniferous forest		
	313	Mixed forest		
Semi-natural areas	321	Natural grasslands	32000	Herbaceous vegetation associations
	322	Moors and heathland		
	324	Transitional woodland-shrub		
	331	Beaches, dunes, sands	33000	Open spaces with little or no vegetations
	333	Sparsely vegetated areas		
Wetlands	411	Inland marshes	40000	Wetlands
	412	Peat bogs		
	421	Salt marshes		
Waterbodies	511	Water courses	50000	Water
	512	Water bodies		

Tab. A 3: Soil texture types and their mean amounts of clay, silt and sand according to AG BODEN (2005)

Soil texture type	Clay	Silt	Sand
Ls2	0.21	0.45	0.34
Ls3	0.21	0.35	0.44
Ls4	0.21	0.225	0.565
Lt2	0.3	0.4	0.3
Lt3	0.4	0.4	0.2
Lts	0.35	0.225	0.425
Lu	0.235	0.575	0.19
Sl2	0.065	0.175	0.76
Sl3	0.1	0.25	0.65
Sl4	0.145	0.25	0.605
Slu	0.125	0.45	0.425
Ss	0.025	0.05	0.925
St2	0.11	0.05	0.84
St3	0.21	0.075	0.715
Su2	0.025	0.175	0.8
Su3	0.04	0.325	0.635
Su4	0.04	0.45	0.51
Tl	0.55	0.225	0.225
Ts2	0.55	0.075	0.375
Ts3	0.4	0.75	0.525
Ts4	0.3	0.075	0.625
Tt	0.825	0.175	0.175
Tu2	0.55	0.425	0.125
Tu3	0.375	0.575	0.1
Tu4	0.3	0.7	0.05
Uls	0.125	0.575	0.3
Us	0.04	0.65	0.31
Ut2	0.1	0.785	0.135
Ut3	0.15	0.765	0.115
Ut4	0.21	0.74	0.09
Uu	0.04	0.9	0.1

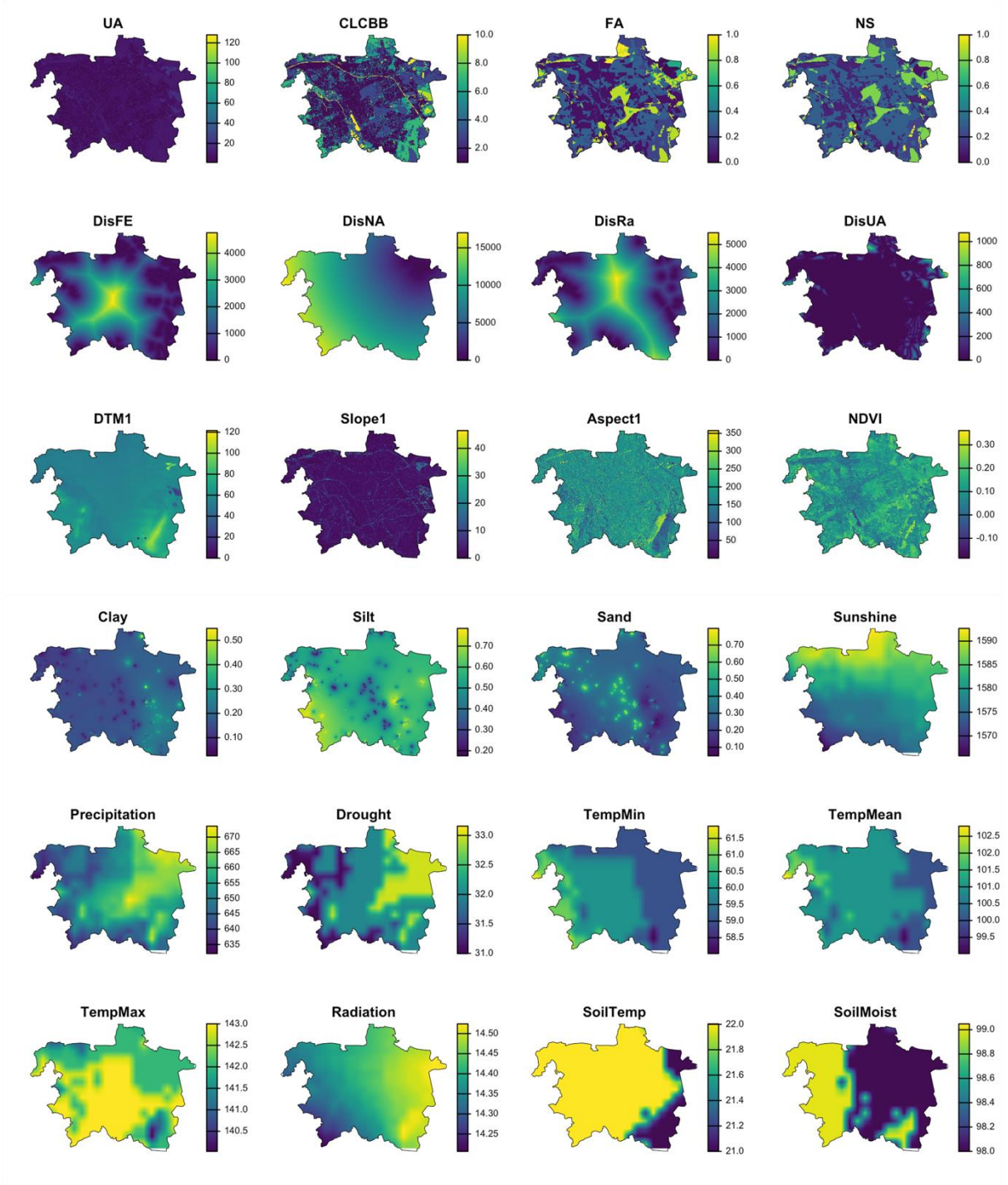


Fig. A 1: Input data containing multi-annual climate data for the Hannover study area, resampled to a cell size of 10 m

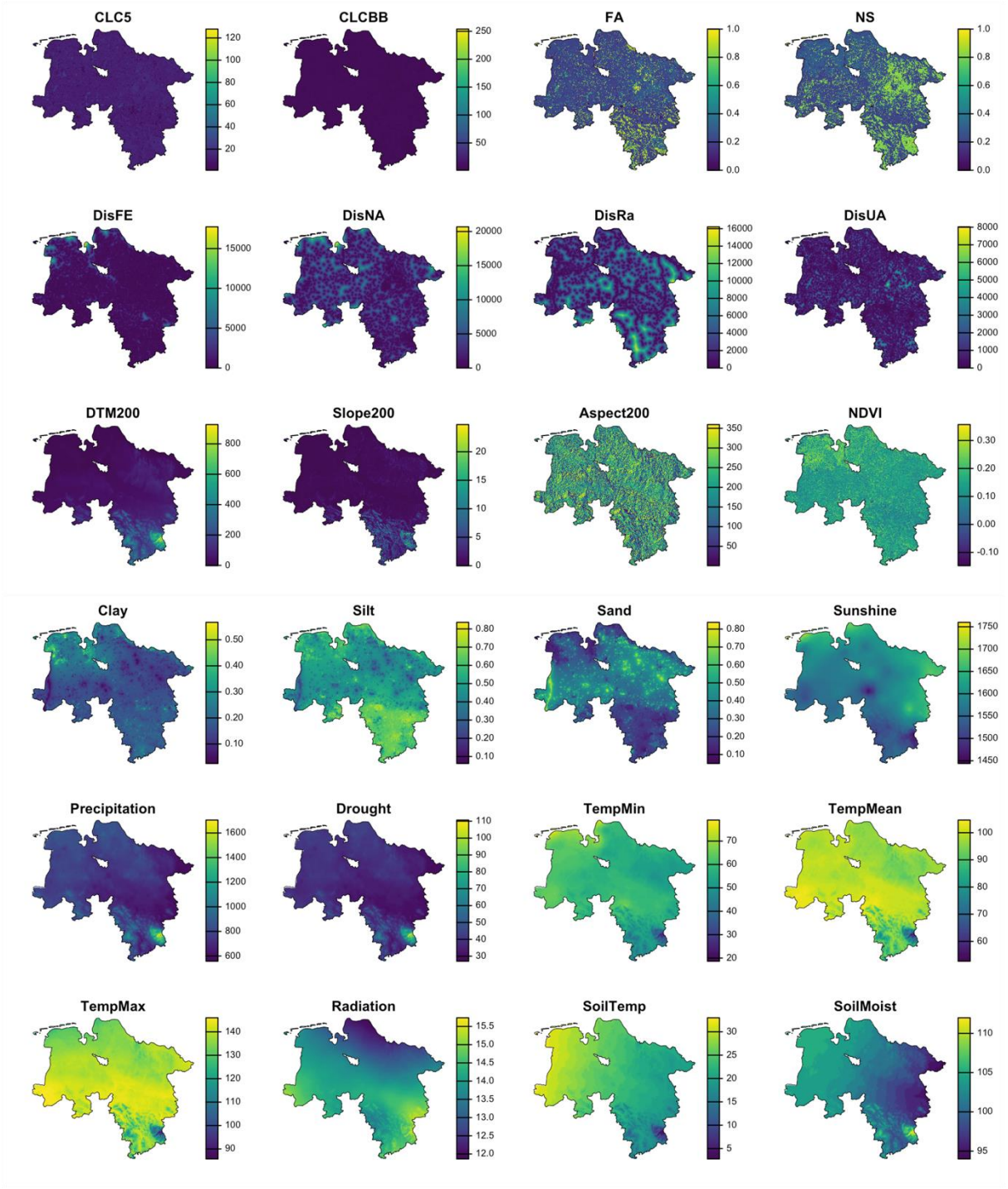


Fig. A 2: Input data containing multi-annual climate data for the Lower Saxony study area, resampled to a cell size of 100 m

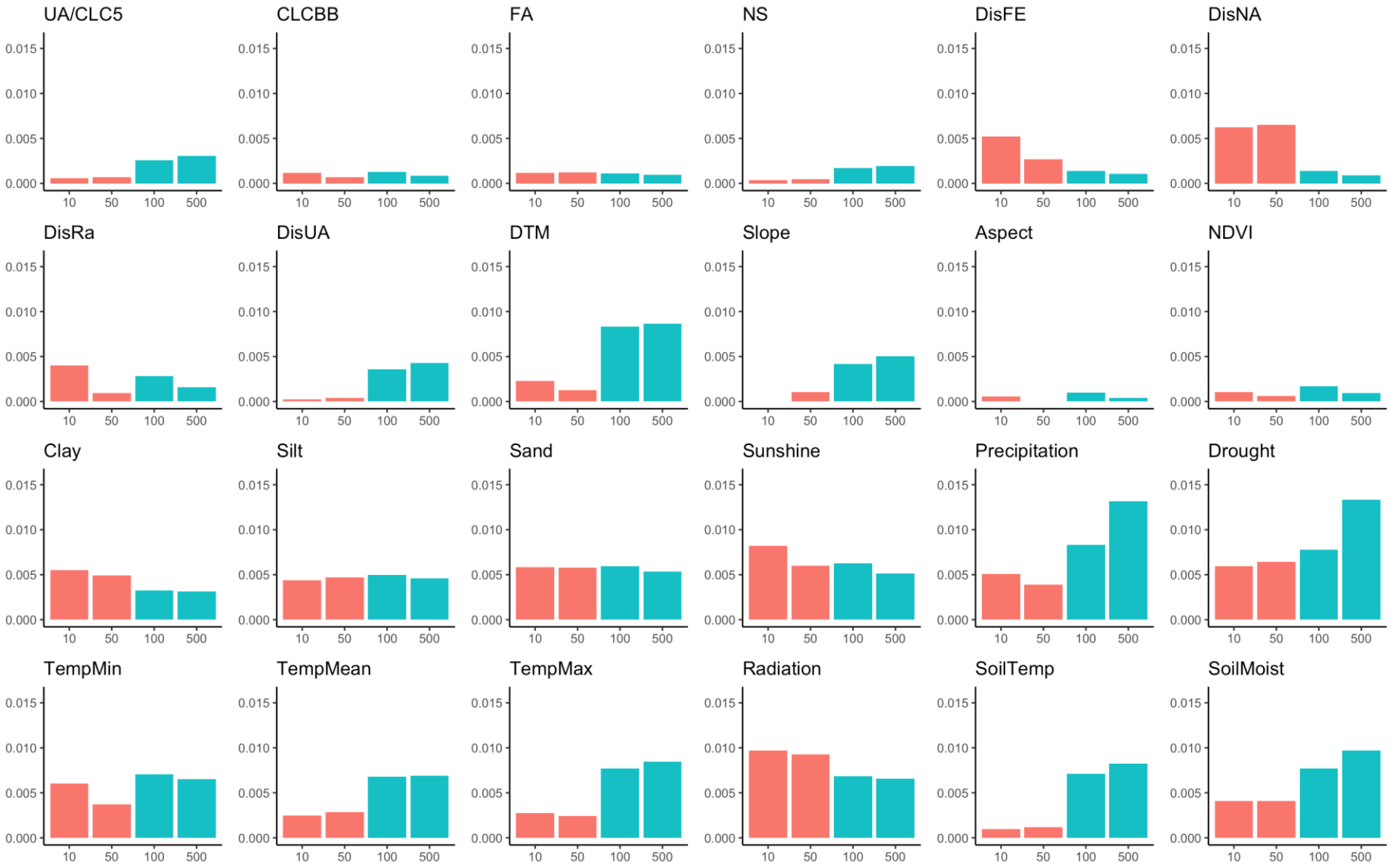


Fig. A 3: Permutation importance (y) of the first RF iteration testing different spatial resolutions (x) for the study areas Hannover (red) and Lower Saxony (blue)

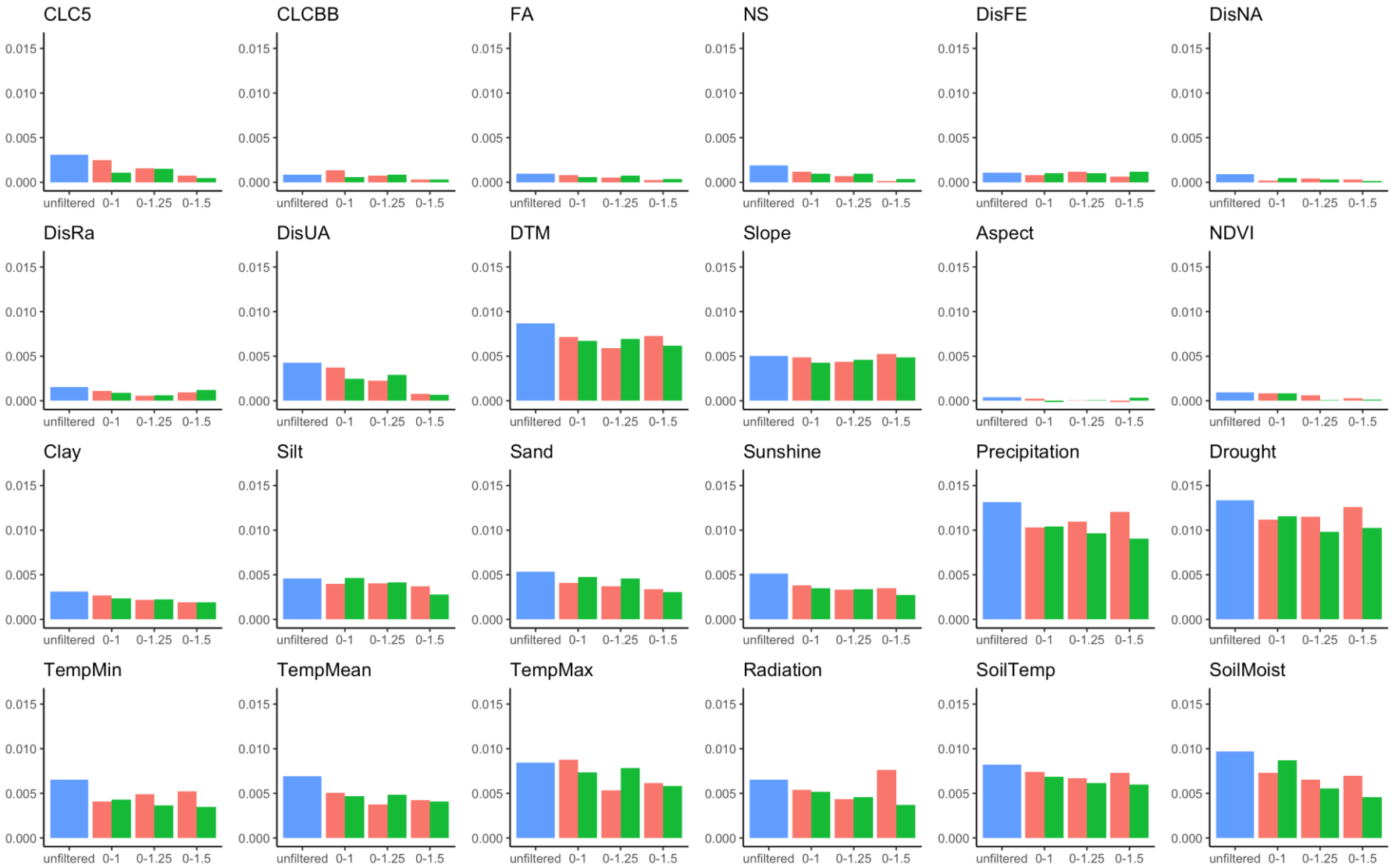


Fig. A 4: Permutation importance (y) of the second RF iteration testing different rescale ranges (x) and resolutions (red: 0.02°, green: 0.05°) with sampbias compared to the result of the first iteration of Lower Saxony with 500 m (blue)

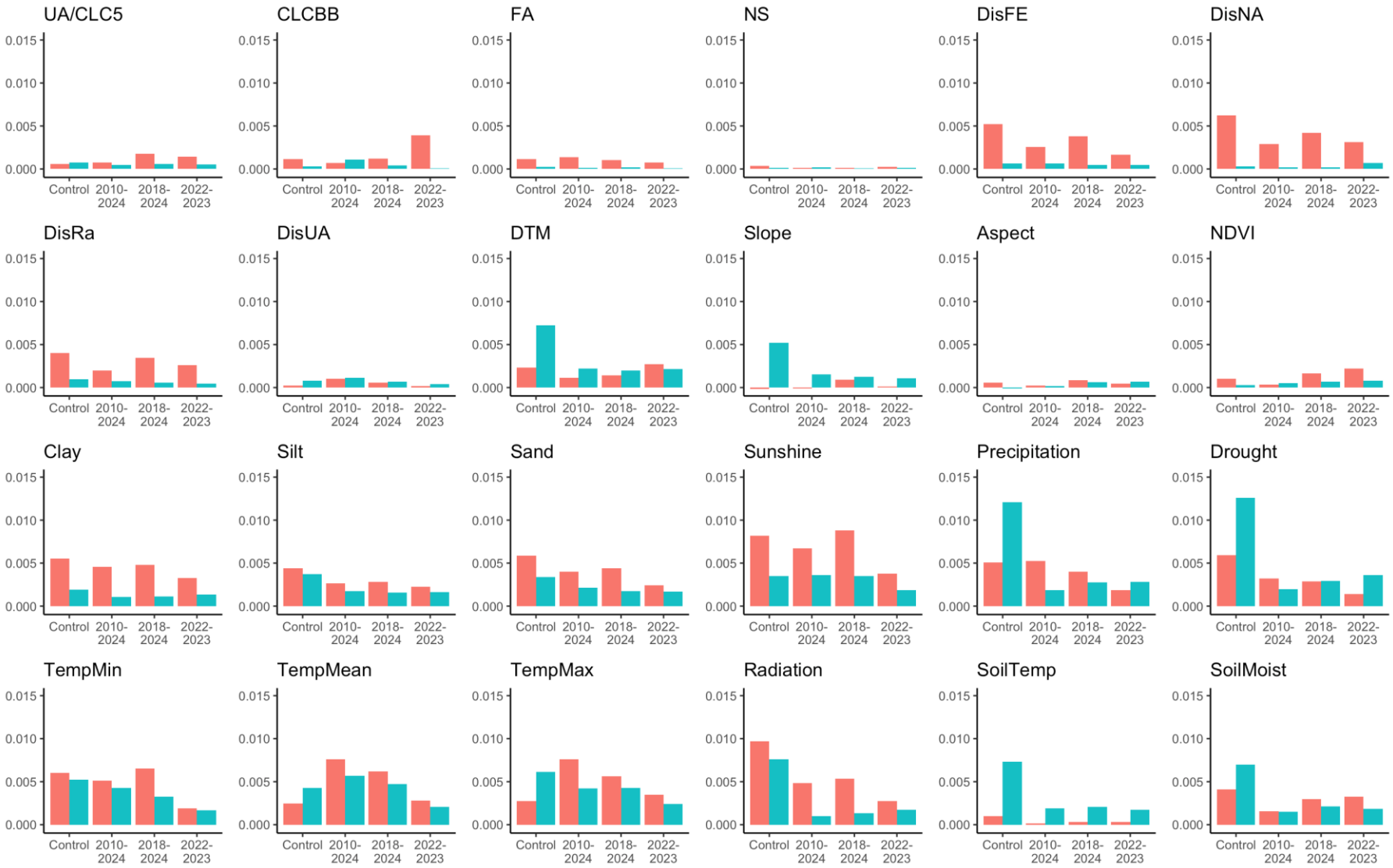


Fig. A 5: Permutation importance (y) of the third RF iteration testing seasonal climate data for different periods (x) for the study areas Hannover (red) and Lower Saxony (blue) compared to the result of the first iteration of Hannover with 10 m cell size and the second iteration of Lower Saxony with 0.02° resolution and a rescale range of 0-1.5

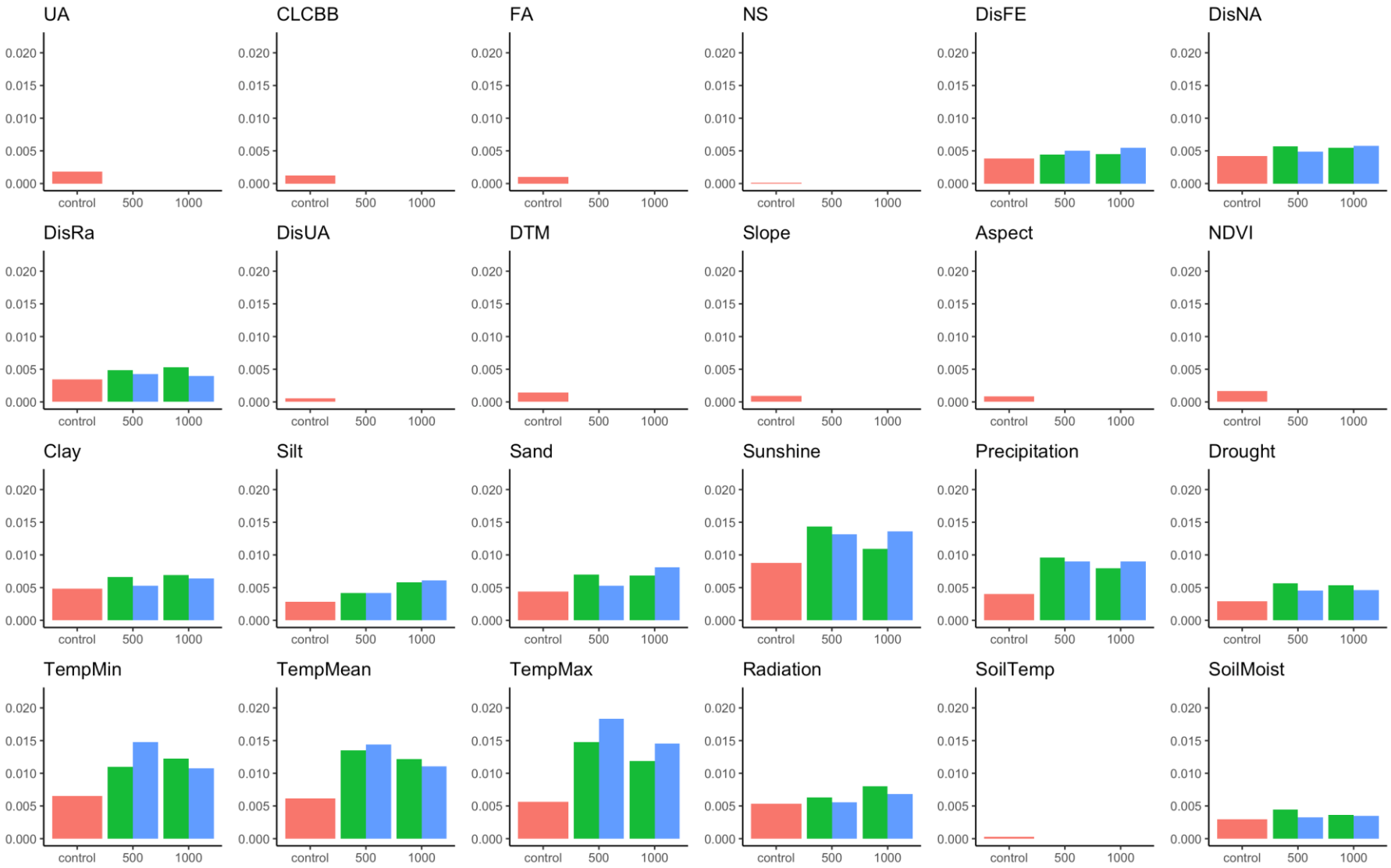


Fig. A 6: Permutation importance (y) of the fourth RF iteration testing less input variables and different mtry (3: green, 4: blue) and number of trees (x) in Hannover compared to the results of the third iteration with seasonal climate data in the period of 2018-2024

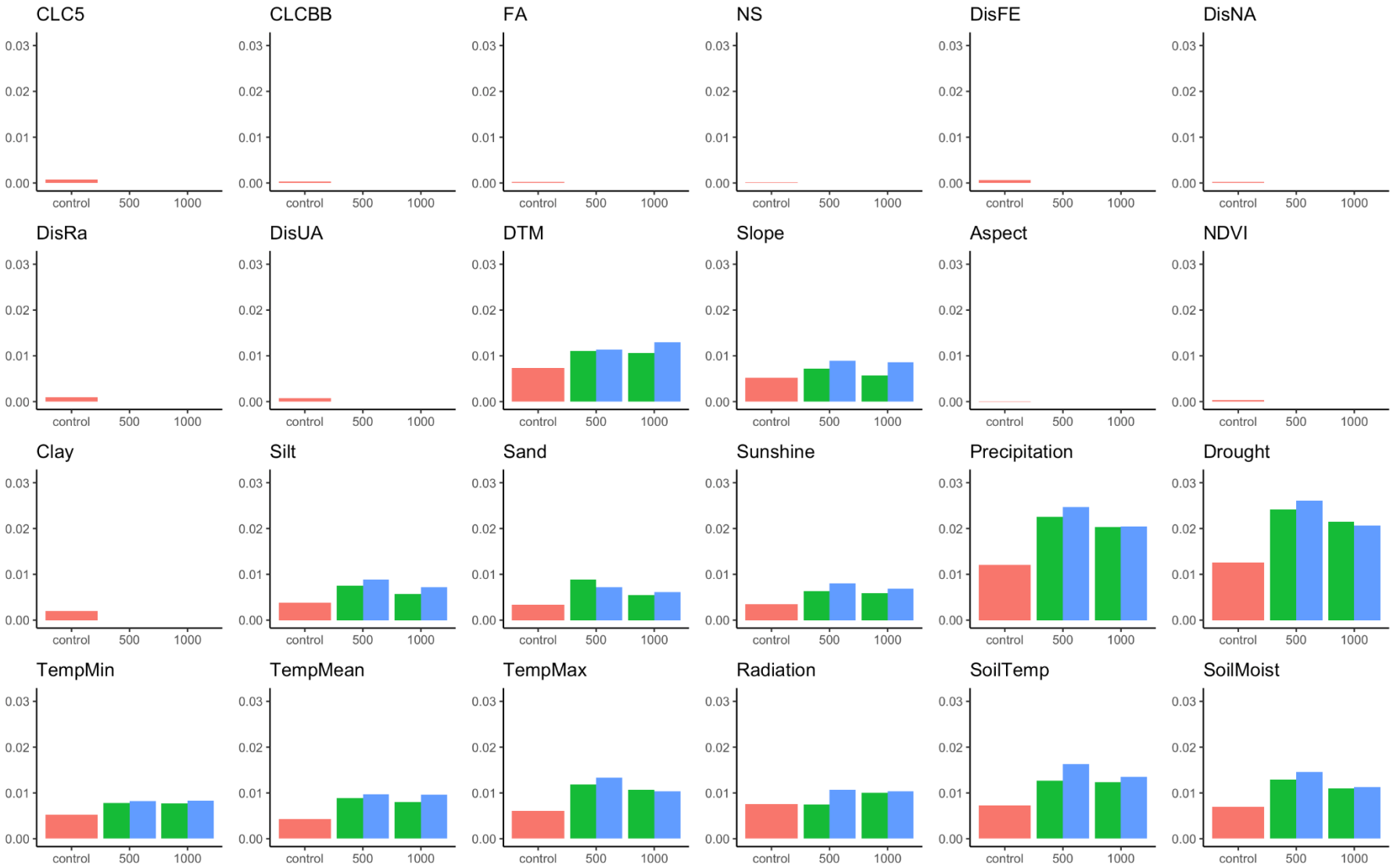


Fig. A 7: Permutation importance (y) of the fifth RF iteration testing less input variables and different mtry (3: green, 4: blue) and number of trees (x) in Lower Saxony compared to the results of the second iteration using a resolution of 0.02° and a rescale range of 0-1.5

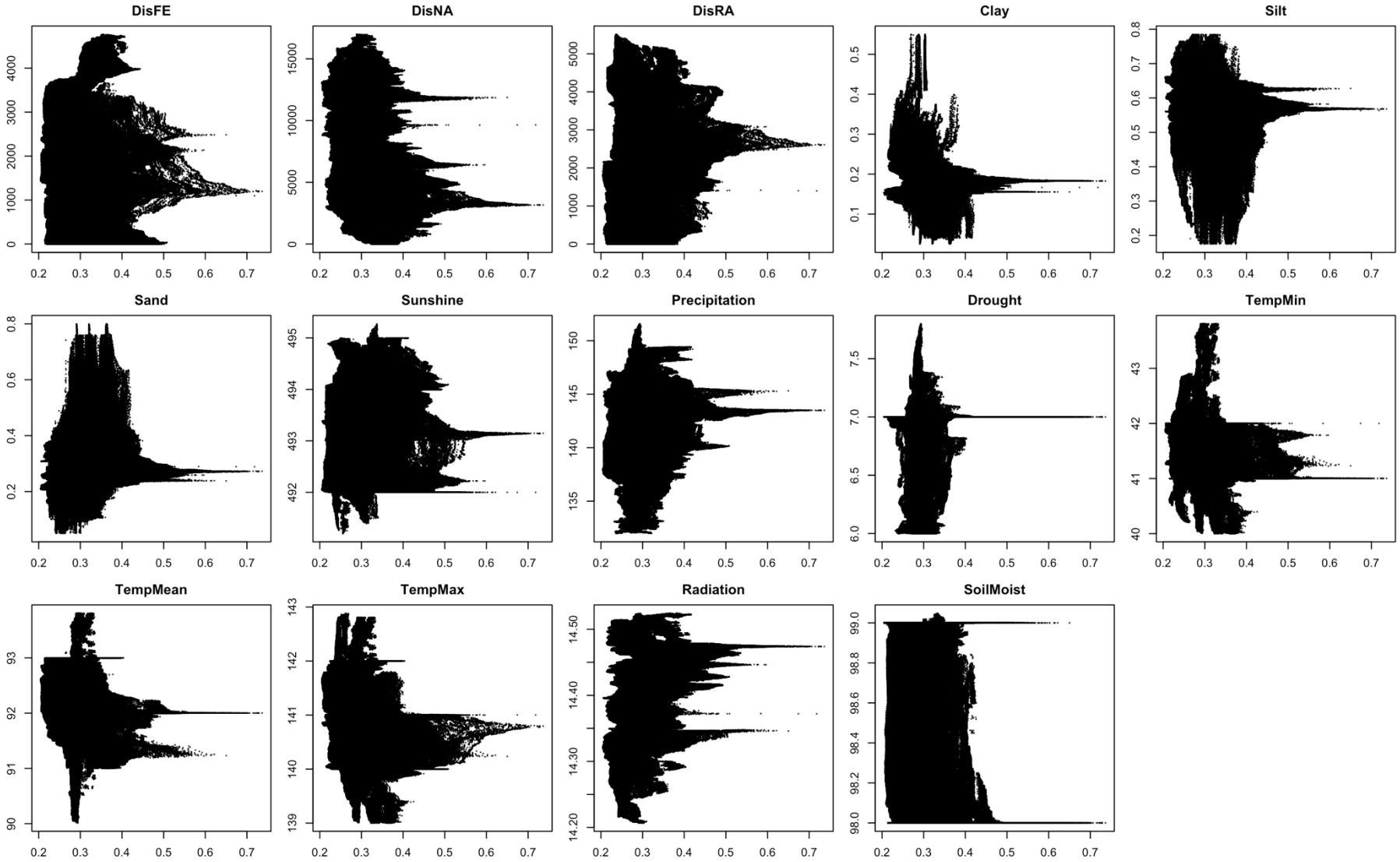


Fig. A 8: Scatter plots with values of the input variables (y) and RF prediction (x) in Hannover in spring 2023

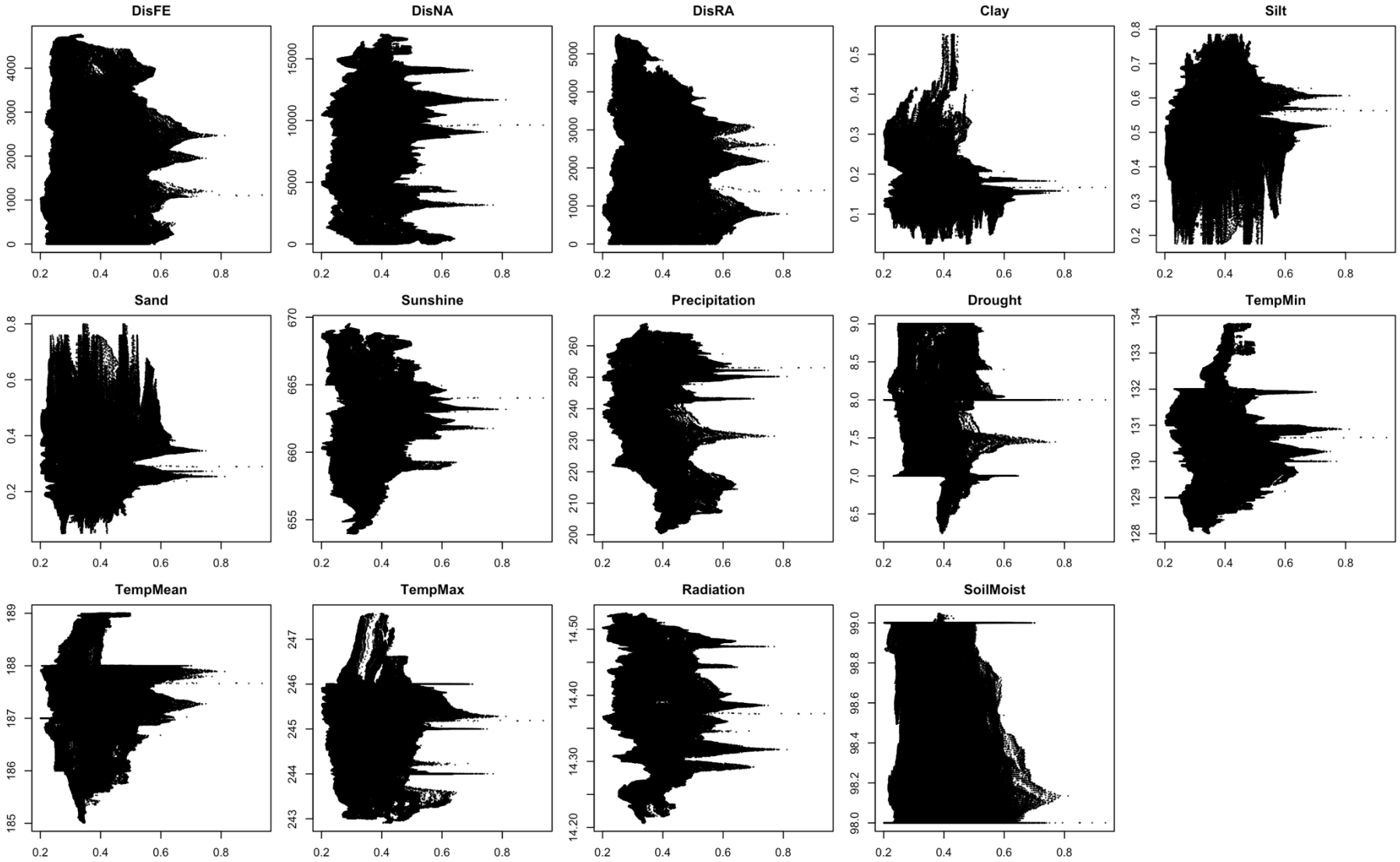


Fig. A 9: Scatter plots with values of the input variables (y) and RF prediction (x) in Hannover in summer 2023

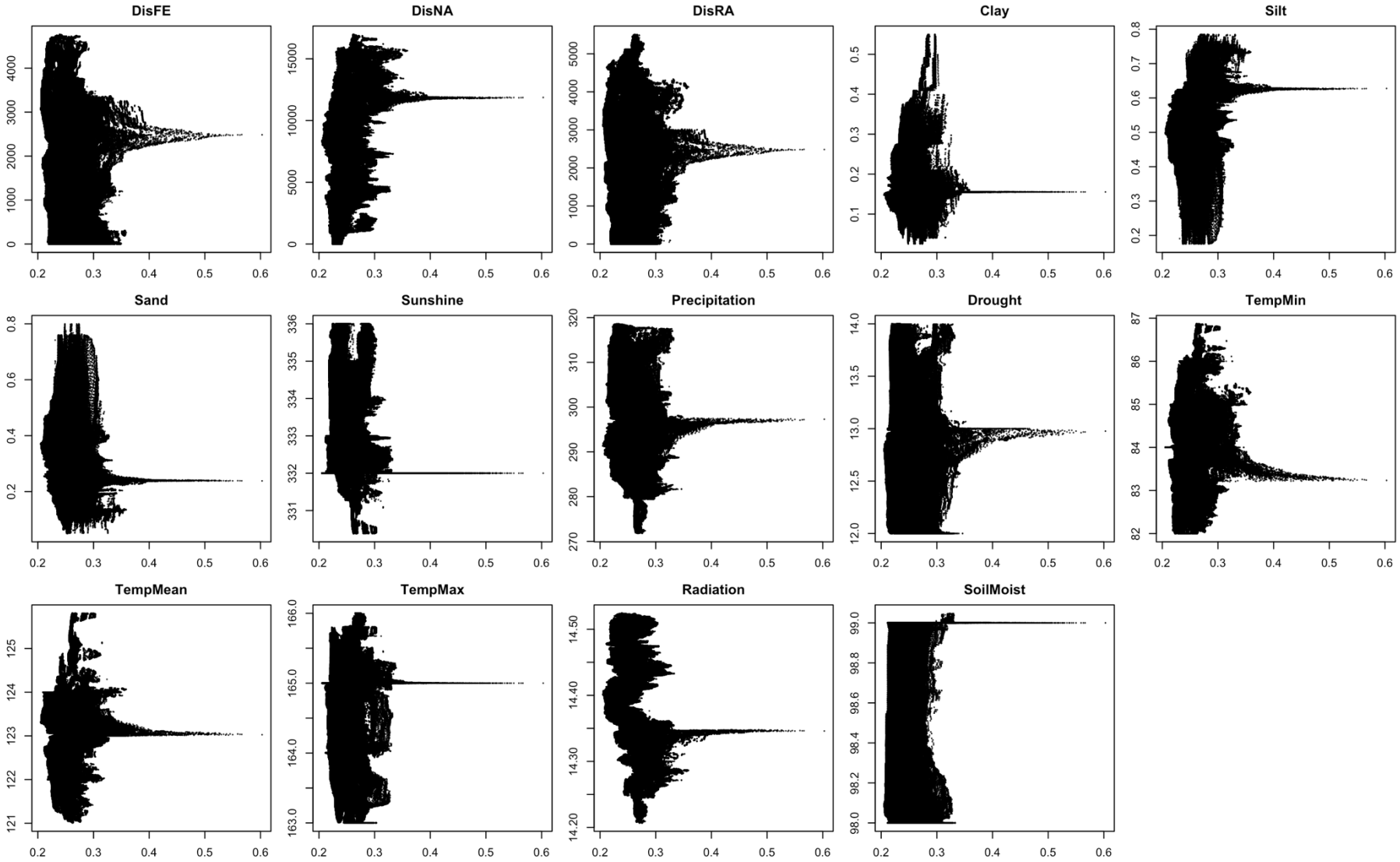


Fig. A 10: Scatter plots with values of the input variables (y) and RF prediction (x) in Hannover in autumn 2023

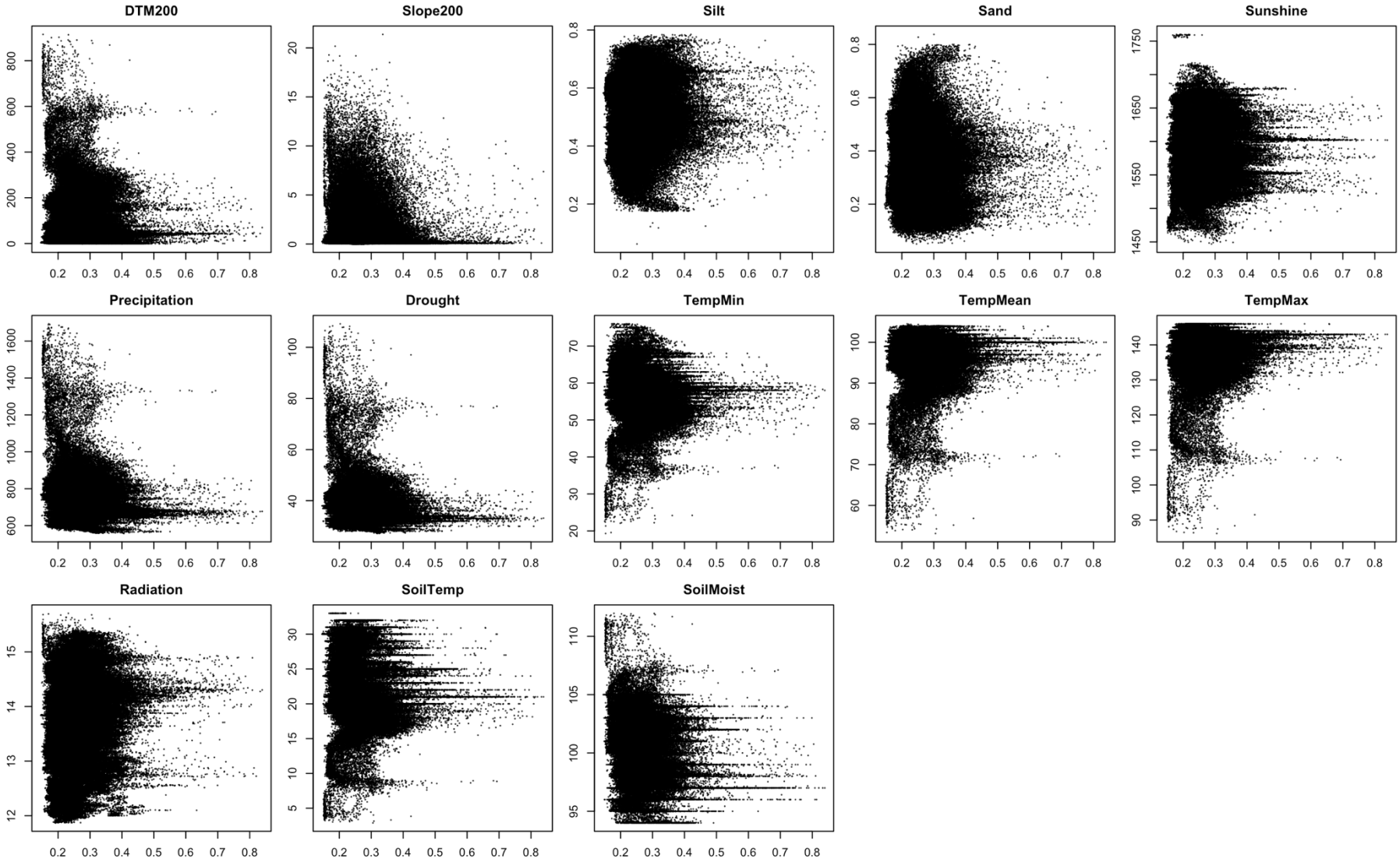


Fig. A 11: Scatter plots with values of the input variables (y) and RF prediction (x) in Lower Saxony

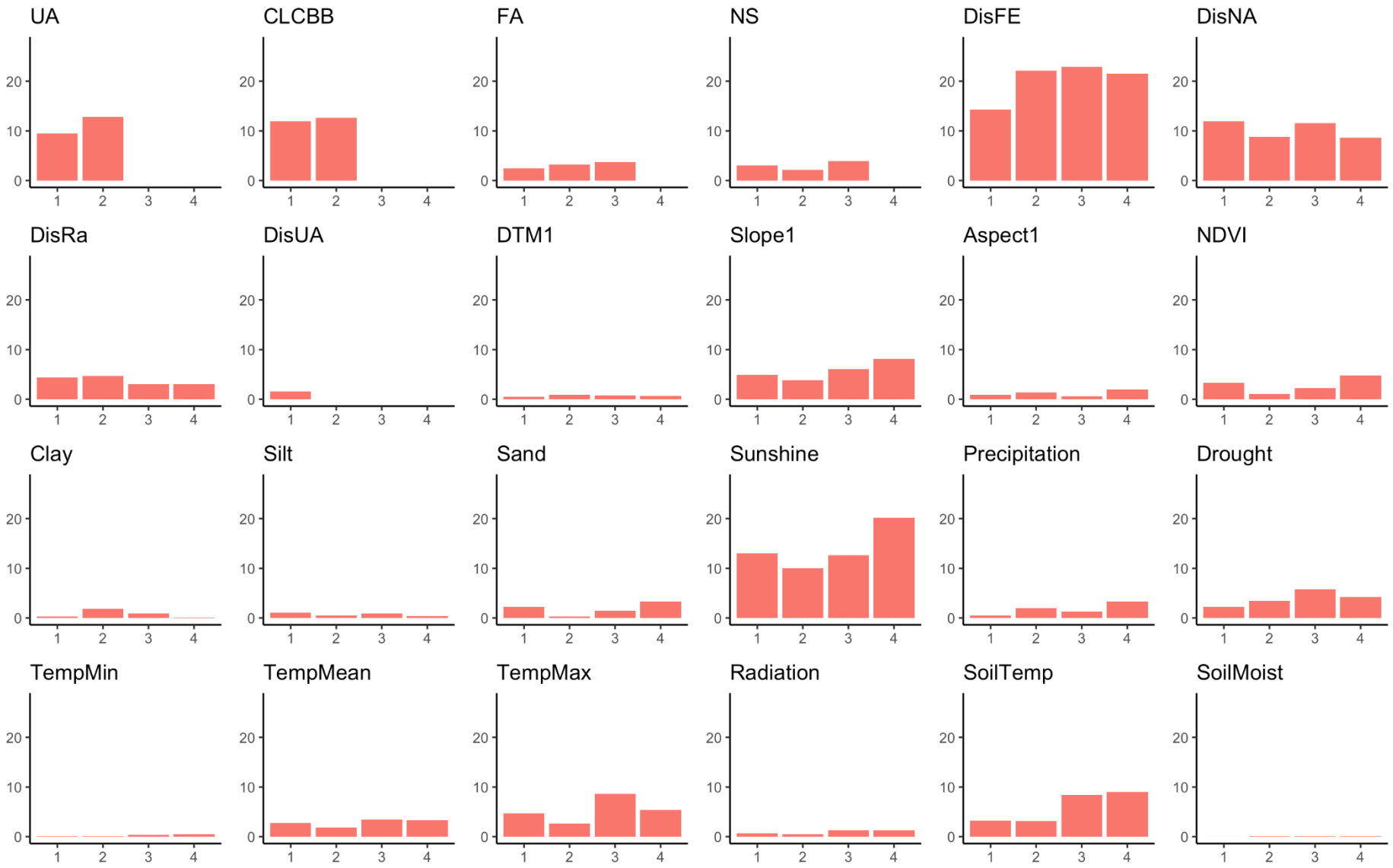


Fig. A 12: Importance (y) of the Maxent models in % for Hannover testing different input data in four runs (x)

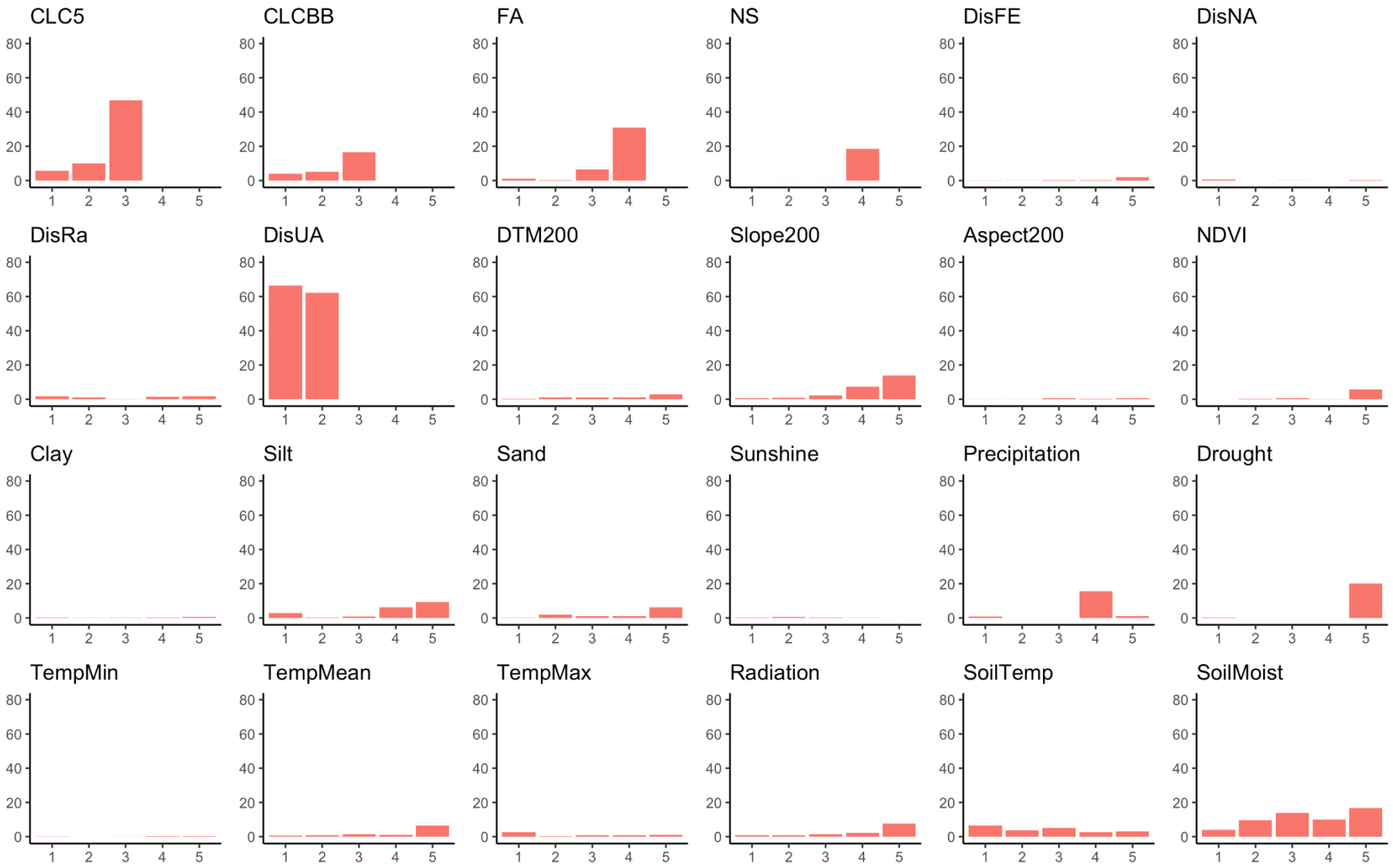


Fig. A 13: Importance (y) of the Maxent models in % for Lower Saxony testing different input data in five runs (x)

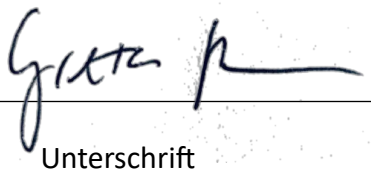
Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich meine Masterarbeit mit dem Titel

A Machine learning-based approach to assess pollinator habitat suitability

selbständig verfasst und die benutzten Hilfsmittel und Quellen sowie gegebenenfalls die zu Hilfeleistungen herangezogenen Institutionen vollständig angegeben habe. Alle Stellen der Arbeit, die anderen Quellen dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht.

Mit der Übermittlung meiner Arbeit auch an externe Dienste zur Plagiatsprüfung erkläre ich mich einverstanden.

Hannover, 12.09.2024 

Ort, Datum Unterschrift