MASTERARBEIT | MASTER'S THESIS

Titel | Title

Predicting the spatio-temporal presence of Aedes albopictus in Graz, Austria, using citizen science network data

verfasst von | submitted by

Dominik Knabe BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 856
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Kartographie und Geoinformation
Betreut von | Supervisor:	Assoz. Prof. Dr. Ourania Kounadi BSc MSc

Acknowledgements

I would like to say thank you to all my family, friends, professors, and colleagues who supported and motivated me throughout my academic journey. First, I want to thank my supervisor Dr. Ourania Kounadi from the University of Vienna for the valuable feedback and the help she provided whenever it was necessary. The regular meetings with her and other master's students were a nice opportunity to exchange ideas and to calm the nerves whenever some parts of the work did not go as planned.

Furthermore, I am very grateful for the cooperation with Dr. Raphael Spiekermann and Dr. Stefan Steger from GeoSphere Austria, who helped me greatly with their expertise in statistical modeling and geospatial risk assessments. Their professional advice and constructive feedback guided me in applying appropriate methods and strengthened the overall quality of my research. I would also like to thank them for including me in the ClimateDiseaseRisk project, which gave me valuable insights into the wider scientific context of my thesis and allowed me to benefit from an inspiring professional network.

I would also like to thank the *Magistrat Graz, Stadtvermessungsamt* for providing access to the relevant spatial predictor data. Without their support in making these high-quality data sets available, this research would not have been feasible.

Lastly, I am particularly grateful to my parents, Elke and Helmut, who always supported me throughout the years and for their trust in me. Without you and your support, I would not be where I am today, THANK YOU!

Abstract

The Asian tiger mosquito (*Aedes albopictus*) is one of the most invasive species in the world and a vector for numerous arboviruses. Since its arrival in Austria in 2012, populations have been growing, raising public health concerns. In this work, the spatio-temporal presence of *Aedes albopictus* was modeled for Graz, Austria, using citizen science data from Mosquito Alert. A framework of generalized additive (mixed) models (GAM/GAMM) was employed to predict spatial habitat suitability and the daily number of mosquito counts using a set of static (land cover, land use, etc.) and dynamic (weather parameters) predictors. Two approaches were tested to account for observer bias in Mosquito Alert data due to heterogeneous sampling effort. The spatial model identified residential areas, allotment gardens and cemeteries as areas at risk, while industrial and rural areas showed low suitability. The presence of green spaces in residential areas was found to have a positive effect on habitat suitability. The temporal model highlighted the seasonal activity of *Aedes albopictus,* which was primarily driven by temperature with peak mosquito season in late summer. Precipitation was not identified as a relevant driver for temporal mosquito activity. The results showed that citizen science data can be used exclusively, when bias-corrected, to yield biologically plausible results to gain insight into invasive mosquito distributions. The results provide a basis for enhanced mosquito surveillance and targeted strategies for vector management in Austrian cities.

Kurzfassung

Die Asiatische Tigermücke (*Aedes albopictus*) ist eine der invasivsten Arten weltweit und Überträger zahlreicher Arboviren. Seit ihrer ersten Sichtung in Österreich im Jahr 2012 wuchs ihre Population stetig an, was zu Bedenken hinsichtlich der öffentlichen Gesundheit führte. In dieser Arbeit wird die raumzeitliche Präsenz von *Aedes albopictus* für Graz, Österreich, anhand von Citizen-Science-Daten von Mosquito Alert modelliert. Ein Set aus generalized additive (mixed) models (GAM/GAMM) wird verwendet, um die räumliche Lebensraumeignung und die tägliche Anzahl der Mücken anhand einer Reihe von statischen (Landbedeckung, Landnutzung, usw.) und dynamischen (Wetterparameter) Prädiktoren vorherzusagen. Es werden zwei Ansätze getestet, um den Beobachterbias in den Mosquito Alert-Daten aufgrund heterogene Stichprobenaufwand zu berücksichtigen. Das räumliche Modell identifiziert Wohngebiete, Kleingärten und Friedhöfe als Risikogebiete, während Industrie- und ländliche Gebiete eine geringe Eignung aufweisen. Das Vorhandensein von Grünflächen in Wohngebieten hat sich als positiv für die Lebensraumeignung erwiesen. Das zeitliche Modell hob die saisonale Aktivität von *Aedes albopictus* hervor, die in erster Linie von der Temperatur bestimmt wird, mit einer Hochsaison für Mücken im Spätsommer. Niederschläge wurden nicht als relevanter Faktor für die zeitliche Aktivität der Mücken identifiziert. Die Ergebnisse zeigen, dass Citizen-Science-Daten, wenn sie um Verzerrungen bereinigt werden, verwendet werden können, um biologisch plausible Ergebnisse zu erzielen und Einblicke in die Verbreitung invasiver Mücken zu gewinnen. Die Ergebnisse bilden eine Grundlage für eine verbesserte Mückenüberwachung und gezielte Strategien zur Verhinderung der weiteren Ausbreitung der Spezies in österreichischen Städten.

Table of contents

Acknowledgements ii

Abstract iii

Kurzfassung iv

Table of contents v

List of figures vii

List of tables viii

List of abbreviations ix

Chapter 1 Introduction 1

1.1 Previous Work 1

1.2 Research Gap 2

1.3 Research Aim 3

Chapter 2 Fundamentals 4

2.1 Species distributions 4

2.1.1 Condition of dispersal 5

2.1.2 Abiotic environment and the fundamental niche 5

2.1.3 Biotic environment and the realized niche 7

2.2 Habitat Suitability Modeling 8

2.2.1 Framework of habitat suitability models 9

2.2.2 From observations to niche quantification 11

2.2.3 Predicting spatial distributions from the quantified niche 12

2.2.4 Which model to use? 12

2.2.5 Predictor variables and geographic scale 14

2.3 Aedes albopictus 14

2.3.1 Spread and geographical distribution 15

2.3.2 Future distribution in Europe 15

2.3.3 Characteristics of *Aedes albopictus* 16

2.3.4 A vector for arboviruses 17

2.4 GAM Model 18

Chapter 3 Data and Methods 20

3.1 Data 20

3.2 Study area 23

3.3 Software 23

3.4 Methodology 24

3.4.1 Filtering of presences and sampling of pseudo-absences 27

3.4.2 Selection and preprocessing of predictor data 31

3.4.3 Spatial and temporal modelling using GAM/GAMM 41

3.4.4 Model validation.. 46

3.4.5 Visualization and classification 49

Chapter 4 Results.. 50

4.1 Spatial model 50

4.1.1 Performance of MS1and MS2 50

4.1.2 Feature importance and partial effects.. 55

4.1.3 Classification of risk levels and model stability 58

4.2 Temporal model.. 61

4.2.1 Model accuracy and performance.. 61

4.2.2 Feature importance and partial effects.. 63

4.2.3 Classification of risk levels 65

Chapter 5 Discussion... 66

5.1 Comparison of bias correction approaches.. 67

5.2 Drivers for mosquito presence... 68

5.3 Applicability of citizen science data.. 70

Chapter 6 Conclusion and potential for future work 71

References 73

Appendix 95

List of figures

Figure 1: Regression of mosquito sightings on SE ... 22

Figure 2: Workflow for spatio-temporal modeling of *Aedes albopictus* ... 26

Figure 3: Gridded SE values and mosquito presences ... 27

Figure 4: Spatial dispersion of *Aedes albopictus* in Graz ... 29

Figure 5: Pseudo-absence sampling using SE weights ... 30

Figure 6: Temporal distribution of mosquito sightings and SE in Graz ... 31

Figure 7: Concept of spatio-temporal modeling of *Aedes albopictus* ... 32

Figure 8: Land use classes in Graz ... 34

Figure 9: Land cover roof sealing ... 37

Figure 10: Population density in Graz ... 37

Figure 11: Kernel density of storm drains in Graz. ... 38

Figure 12: Correlation matrix of meteorological predictors ... 41

Figure 13: Areas for spatial cross-validation. ... 48

Figure 14: Classification thresholds for habitat suitability and activity ... 50

Figure 15: Land use class distributions for MS1 & MS2. ... 51

Figure 16: AUROC after spatial cross-vlaidation ... 52

Figure 17: Predicted habitat suitability from MS1 and MS2 ... 54

Figure 18: Spatial permutation-based feature importance for MS1 and MS2 ... 56

Figure 19: Spatial partial effect plots ... 57

Figure 20: Interaction effects ... 58

Figure 21: Spatial classification of habitat suitability ... 59

Figure 22: Uncertainty of modeled results ... 60

Figure 23: Bivariate choropleth map of habitat suitability and prediction uncertainty. ... 61

Figure 24: Temporal prediction of daily mosquito counts ... 63

Figure 25: Temporal permutation-based feature importance ... 64

Figure 26: Temporal partial effect plots ... 65

Figure 27: Temporal classification of daily mosquito activity ... 66

List of tables

Table 1: Data overview 22

Table 2: Reclassification of land use data 34

Table 3: Reclassification of land cover data. 35

Table 4: Saptial predictors 43

Table 5: Temporal predictors. 45

List of abbreviations

AUROC	Area Under the Receiver Operating Characteristic Curve
BRT	Boosted Regression Trees
CV	Coefficient of Variation
EDA	Exploratory Data Analysis
GAM	Generalized Additive Model
GAMM	Generalized Additive Mixed Model
GIS	Geographic Information System
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HSM	Habitat Suitability Model
MCDA	Multi-Criteria Decision Analysis
PPS	Probability Proportional To Size
ROC	Receiver Operating Characteristic
RS	RS
SE	Sampling Effort
TWI	Topographic Wetness Index

Chapter 1 Introduction

Aedes albopictus (Culicidae) (Skuse, 1895), also known as the Asian tiger mosquito or forest day mosquito, is an aggressive mosquito that acts as a vector for up to 26 different arboviruses. Prominent examples include dengue, yellow fever or chikungunya (Medlock et al., 2012; Paupy et al., 2009). The mosquito is endemic to the temperate and tropical regions of Southeast Asia. During the last four decades, it spread globally to every continent excluding Antarctica (M. U. Kraemer et al., 2015). Climate change and associated rise in average temperature lead to the further expansion of *Aedes albopictus* into cooler climates and higher latitudes, thereby increasing public health risks (Cunze et al., 2016). Today, the arthropod is found in the top 100 list of the worst invasive species by the Invasive Species Specialist Group due to its high adaptability to diverse environments, strong competitive ability, the effects of globalization, insufficient monitoring, and the absence of effective control measures (*Global Invasive Species Database*, 2025; Medlock et al., 2015). To identify the best options for preventing and managing further spread of *Aedes albopictus*, it is necessary to determine and predict suitable environments for its establishment.

Considering the species introduction to Austria in 2012, this work targets the gap in research related to small-scale habitat suitability of *Aedes albopictus* specifically within Austria by focusing on the city of Graz (Seidel et al., 2012). As discussed in the following sections, while global- and continental-scale models primarily consider climatic factors, achieving substantial consensus, it is important to account for variables like land use, vegetation, and other localized factors at smaller scales, such as a country or city. Despite the advancements in modeling at these larger scales, there remains a lack of targeted research on the Austrian environment. Therefore, this analysis seeks to identify and understand the relevant factors affecting the habitat suitability of *Aedes albopictus* in Austria, using Graz as the study region. While traditionally the spread and surveillance of mosquitoes were usually observed using ovitraps, the citizen science network Mosquito Alert offers a novel, cheap and reliable tool to obtain a broad spatial coverage of expert-verified mosquito data (Palmer et al., 2017). By employing a spatial habitat suitability model and a temporal model to predict daily mosquito activity, this research helps to provide further knowledge that can guide effective mosquito management strategies in the region and potentially identify future areas of risk.

1.1 Previous Work

The increasing threat of climate change-induced spread of *Aedes albopictus* in Europe is already a well-established field of research. Several studies have modeled the current habitat distribution and future habitat suitability based on various climate change scenarios. In the past decades, suitable areas for *Aedes albopictus* were identified using a Random Forest classifier, expert-based multi-criteria decision analysis (MCDA), and high-resolution gridded climate data (Avenell et al., 2009; Caminade et al., 2012). According to these papers favorable conditions are found along the Mediterranean coast of southern Europe. Fischer et al. (2011) compared four different modeling approaches, distinguishing between statistical climate variable selection and expert-based selection, as well as native and global ranges, to predict future vector distribution. The paper predicted a shift in habitat suitability towards northwestern Europe, with a temporal delay in eastern Europe and a decrease in suitability in southern Europe between 2011 and 2040. Georgiades et al. (2023) used climate data as well as global population and land use data to perform a combination of Random Forest and XGBoost binary classifiers.

While many studies exist on present and future habitat suitability across Europe, there is a lack of studies at smaller scales, such as countries or cities, especially in Austria. In Switzerland, habitat suitability modeling (HSM) was performed using land surface temperature satellite data with a spatial resolution of 1000 m and a temporal resolution of four data points per day (Neteler et al., 2013). Similar simulations for Germany used only climate data as model input, with spatial resolutions of 5 km and 10 km (Koch et al., 2016; Thomas et al., 2018).

On an even smaller scale, the relationship between egg abundance (from ovitraps) and land cover data was explored in the city of Rome using generalized linear mixed models (GLMM) with high-resolution land cover data, solar radiation, and capture month as predictor variables (Cianci et al., 2015). It was found that vegetation and solar radiation are positively correlated with egg numbers. Torina et al. (2023) modeled the time series of *Aedes albopictus* abundance in Palermo, Italy, analyzing how meteorological parameters like humidity, air pressure, and wind speeds influence it. Additionally, three supervised learning models successfully predicted micro-scale mosquito abundance in Charlotte, NC, USA, using socioeconomic and landscape factors as predictor variables (Chen et al., 2019). Li et al. (2014) showed that ongoing urbanization in China accelerates mosquito survivorship and development. Land cover attributes like canopy type, water clearance, or depth were important factors associated with mosquito abundance. On a smaller scale, *Aedes albopictus* habitats are expected to be most favorable in urbanized small green areas, such as recreation spots or private gardens due to the animal's preference of small water containers like pot plants for breeding compared to larger open areas (Manica et al., 2016).

While the majority of these studies analyze abundance data from traditional ovitraps, citizen science network data provides a reliable and cost-efficient tool to track and monitor mosquitoes globally (Carney et al., 2022; Palmer et al., 2017). Consequently, distribution patterns of *Aedes albopictus* can be analyzed in areas without the need for costly and widely distributed insect traps that need to be operated by experts.

1.2 Research Gap

As previously discussed, individual research has been undertaken to identify current and future distributions of *Aedes albopictus* as well as relevant predictors. These models are performed on various spatial scales, ranging from global and national-scale models to micro-scale models. While most models analyze the bioclimatic range of habitat suitability of *Aedes albopictus* (Caminade et al., 2012; D. Fischer et al., 2011; M. U. Kraemer et al., 2015), some studies also analyze the impact of predictors like land cover, vegetation, socioeconomic data or meteorological parameters on the spatial and temporal distribution of the mosquito (Chen et al., 2019; Cianci et al., 2015; Gardner et al., 2013; Torina et al., 2023; Unlu et al., 2011; Westby et al., 2021). None of these studies were performed in an Austrian environment, especially not on a city level. Similar environmental characteristics as they can be found in Germany were only analyzed for climate data (Koch et al., 2016). In Rome, additional parameters like land cover data were used but only within the university campus as study area, and therefore only reflecting a limited range of environmental conditions (Cianci et al., 2015).

Given the fact that environmental predictors like land use, land cover or vegetation have their very own individual characteristics depending on the study area, it is difficult to assume that findings from previous research conducted in China or the USA can easily be transferred to Austria. For example, vegetation in China contains completely different plant species compared to Austria and therefore provides different habitats and ecological interactions. Additionally, urban environments have distinctive characteristics like unique types of land use categories (e.g. allotment gardens), different levels of infrastructure (e.g. waste management), or socioeconomic background, thus yielding different conditions, habitats, or breeding sites (e.g. neglected plastic containers or overgrown spaces in low-income areas). To describe the heterogeneous characteristics of a given study area, most previous studies used coarser indicators like the amount of impervious surface or broad land use classes (e.g. urban, rural, forest) (Baldacchino et al., 2017; Shragai & Harrington, 2019). Consequently, there is potential to explore the effect of more detailed predictor data to analyze fine-scale differences in mosquito distribution across an urban area.

Most studies use ovitrap egg abundance data exclusively, without considering citizen science data. Mosquito Alert data offer a broader and more detailed spatiotemporal coverage of mosquito sightings, potentially providing deeper insights into how different factors influence *Aedes albopictus* habitat suitability. Garamszegi et al. (2024) utilized a combination of Mosquito Alert and ovitrap data to analyze the distribution of *Aedes albopictus* across Hungary. Little to no work has been done to predict habitat suitability of *Aedes albopictus* on a city scale using citizen data exclusively and to discuss the plausibility and applicability of such an approach.

1.3 Research Aim

Given the gap in research, this study aims to utilize the extensive spatiotemporal coverage of mosquito sighting data from Mosquito Alert to train generalized additive (mixed) models (GAM/GAMM) to further assess the individual effects of predictors like land use, land cover or vegetation within a single study area, especially in the environmental setting of an Austrian city which has not been subject of research so far. Interaction effects are utilized to analyze if the presence of vegetation in an urban context influences mosquito distribution. Additionally, there has been limited research examining the temporal aspect of mosquito distribution based on meteorological parameters. By examining the influence of meteorological conditions on the temporal distribution of mosquito occurrences using high-temporal-resolution weather data, the aim is to assess the weather-driven impact on mosquito activity while also predicting the risk of mosquito abundance.

Through cooperation with Geosphere Austria, access was granted to a set of land use and land cover data for Graz. Given that Graz has the highest number of mosquito sightings in Austria, reported by the citizen science network Mosquito Alert and the ovitrap network operated by the Austrian agency for health and food security (AGES), it serves as an ideal location for a case study in Austria.

Since citizen science data are utilized exclusively, this work aims to explore the applicability and plausibility of such an approach by comparing modeled results with field observations from Graz and similar regions. Two approaches are tested to correct for observer-bias due to heterogeneous sampling effort (SE) inherent to such data inventories. With the help of this analysis, high-risk areas and dominant drivers can be detected to guide further prevention measures in Graz and possibly other Austrian cities.

Main objective

To test the use of citizen science data for predicting the presence of *Aedes albopictus* in space and time at a city scale in Graz – Austria.

Research Questions:

1. Two approaches were tested to correct spatial bias inherent in citizen science data. Which is the better performing one in terms of biological plausibility?
2. Which factors are most influential on modeling the presence of *Aedes albopictus*?
3. Which areas are most prone to the presence of *Aedes albopictus*?
4. How do results compare to findings from similar studies in other regions?
5. Given the bias inherent in citizen science data, can the model demonstrate accurate and plausible predictions? What are the advantages and disadvantages of using these data?

Chapter 2 Fundamentals

2.1 Species distributions

The fundamental questions in biogeography include: what are we observing (i.e., type of species, communities, or ecosystems)? Where do these observations occur in terms of space and time? What influences the distribution of organisms in their respective locations (Cox et al., 2020)? The search for answers to these questions is a longstanding attempt (Guisan & Thuiller, 2005). Interest in estimating the range of ecological niches for species to model their geographic distributions can be found in many fields like systematics, ecology, public health, conservation, etc. With the increase in ecological research during the 20th century, the understanding and explanation of biodiversity and its distribution across different temporal and spatial scales remain an important domain in macroecology (Cox et al., 2020; McGill, 2010). An abundance of new or enhanced theories arose such as the theory of metapopulation, the neutral theory, the ecological niche theory, or the metabolic theory that explicitly incorporated geographic space (J. H. Brown et al., 2004; Chase, 2011; Hanski, 1998; Hubbell, 2001). In the following chapter, the most important aspects of ecological theories and species distributions, most relevant for modeling habitat suitability, are explained.

Before diving into the fundamentals of species distributions, the keywords habitat, environment and niche need to be defined according to Kearney (2006). A **habitat** is a representation of a physical location, at a specific scale of space and time, where an organism either exists or has the potential to exist. Compared to an environment or niche, a habitat can be described without a specific relation to an organism. An **environment** contains biotic and abiotic factors that surround and may interact with a specific species. Lastly, a **niche** can be described as a specific subset of environmental factors that influence a certain organism, in which the mean absolute fitness of individuals within a population is at least one or more.

When modeling habitat suitability, the fundamental conditions that describe how and why species are distributed across space and time as they are, need to be considered. Guisan et al. (2017) define these conditions as follows:

I. the species has to reach the site, i.e. to access the region (Barve et al., 2011) and disperse there;
II. the abiotic environmental conditions must be ecophysiologically suitable for the species;
III. the biotic environment (interactions) must be suitable for the species.

The interaction of these factors happens dynamically with a varying strength depending on the observed scale to form the complex phenomenon called the geographic distribution of a species (Soberón &

Peterson, 2005). The first consideration describes the conditions that must be fulfilled for a species to occupy and enter new areas from its origin where it can **disperse**. It respects the biogeographic background of the species, along with all factors that restrict its distribution from its original location, including physical barriers to migration (e.g., oceans, mountains, deserts), as well as biotic and abiotic dispersal mechanisms. The second requirement relates to the **abiotic habitat suitability** for the target species, indicating that the interaction of abiotic environmental factors at the location, commonly known as environmental suitability, is part of the conditions necessary for a species to grow and sustain populations. This second condition defines the basis of habitat suitability modeling (as described in section 2.2.2). The third criterion accounts for **biotic interactions**, which refer to the relationships with other organisms that can be either beneficial (commensalism, mutualism) or harmful (competition, predation). The upcoming sections describe these three conditions in detail.

2.1.1 Condition of dispersal

In the process of speciation at some point in space and time, every single species on earth emerged from evolutionary processes. One cause of speciation, the allopatric speciation, is a result of a species not being able to overcome geographic barriers to reach new areas. One famous example is the group of Darwin's finches endemic to the Galapagos Islands (Sato et al., 2001). By not fulfilling the first condition, Darwin's finches, originating from one common ancestor, were isolated on their respective islands divided by the ocean and therefore were not able to disperse. As a result, 15 different species evolved through the process of allopatric speciation. The very large-scale distribution of many species today can largely be described by the biogeographic history and dispersal limitations, thus being distributed near areas from which they originate (Cox et al., 2020). In the scope of HSM, it is important to know that these processes only happen at a very slow pace and will also occur in the future. As stated by Guisan et al. (2017), it is essential to examine the extent to which species retain the ecological characteristics of their ancestral species after divergence (niche conservatism (J. J. Wiens & Graham, 2005)). The effect of niche changes between ancestral and invaded ranges can have important implications on how to forecast and model invasions (Guisan et al., 2014).

Nonetheless, numerous invasive species (so called "exotic" or "alien species") state their capability of overcoming these dispersal limitations in the form of geographic barriers by making use of abiotic or biotic dispersal mechanisms. Due to rapid globalization in the 20th century, human activity helps invasive species to overcome dispersal limitations, therefore replacing them with limitations to the abiotic environment (II) (Capinha et al., 2015). *Aedes albopictus* provides a clear example of this process by overcoming the physical barrier in the form of the Indian and Pacific Ocean by the traffic via plane and ship during World War II (Gratz, 2004).

2.1.2 Abiotic environment and the fundamental niche

Not only do physical barriers constrain the distribution of species but also abiotic environmental characteristics. Individuals of the same species can be found in regions with very different conditions, with each combination of environmental variables being one distinct habitat (Crosby et al., 2019). Therefore, each species can occupy areas with a range of different habitats. In the scope of this knowledge, several questions arise. How does the non-living environment affect the spread of living organisms? What are the various kinds of environmental factors that impact the distribution of species?

In what ways do multiple factors collectively shape a species' geographical range? These key questions are fundamental as they form the basis for all further HSM.

Abiotic factors describe all non-living parameters surrounding an organism. These include climatic conditions (like temperature, humidity, radiation or wind), soil parameters (like soil temperature, soil humidity or minerals) as well as water parameters (like the type of water bodies, temperature, pH-value or salinity). Hutchinson (1978) referred to these factors as scenopoetic, deriving from the Greek origins meaning “setting the scene”. Soberón (2007) proposes that it may be meaningful that scenopoetic variables such as topography or macroclimate are assessed for a large area with a coarse spatial resolution (10^{-1}-10^{2} km^{2}) and consequently show high spatial autocorrelation. Nonetheless, these variables depend on the analyzed species and thus, high resolution scenopoetic variables may also exist. In the case of modeling *Aedes albopictus,* land use is also a scenopoetic variable that is only useful when observed with a high resolution, especially when analyzing species distribution on a city scale.

For each organism a subset of scenopoetic attributes exists that favors the fitness and reproduction of a species. Consequently, no species can reach its maximum fitness in every environment, thus only occupying limited geographic regions. The explanation to that can be found in each organism’s physiology (Woodward & Kelly, 2003). During evolutionary processes, organisms develop their own physiological adaptations to their surrounding environments. Therefore, based on its unique physiological expression, each organism reacts differently to its surrounding abiotic factors, also called environmental gradients. Each organism’s response along an environmental gradient can be displayed using physiological response curves derived from gradient analysis (Whittaker, 1967). These curves display a species’ abundance in relation to an environmental gradient. The peaks of response curves are often linked to an organism's physiological optimum along the gradient, while fitness tends to decrease as one moves further away from that point (Ellenberg, 1952). Depending on the species and the type of gradient, response curves can vary strongly in shape and distribution. Taking solar radiation as an example for an environmental gradient, a lizard as an ectotherm relies heavily on external heat and therefore shows a positive response to the presence of solar radiation. A mole on the other hand spends most of its life underground and consequently shows a minimal response to solar radiation. While one species might thrive in a specific habitat another would suffer and consequently become extinct.

When examining the influence of each environmental variable in isolation, one might overlook how interacting effects of multiple variables affect the species' physiology, with one potentially reducing or enhancing the impact of another. Consequently, all relevant variables must be analyzed together to accurately describe what is known as the environmental niche of species (Chase & Leibold, 2003). The combination of all environmental attributes that are relevant to the species’ fitness is called the ecological niche (Kearney, 2006). Following this concept, the same environment can have totally different effects on the fitness of different organisms, based on their respective physiology, morphology and behavior. For example, two organisms that are identical in every trait except for their reflectivity to solar radiation might have significantly different body temperatures when exposed to the same environmental factors such as wind, humidity, air temperature, and solar radiation. Likewise, two distinct species may consume the same food, yet the nutritional results can differ if their digestive systems are not the same. The influence of organisms on their surroundings leads to the phenomenon of organisms “creating” their own ecological niches (Odling-Smee et al., 2003).

Based on previous work by Grinnell (1917), who introduced the term ‘niche’ to an ecological context, Hutchinson (1957) quantitatively formalized the fundamental niche concept as a n-dimensional hypervolume with environmental variables in which for every point the intrinsic growth rate of a species is positive. At that time, Hutchinson did not differentiate between biotic and abiotic variables, which he later did in (1978), leading to two fundamental niches that can be defined as the fundamental and the realized niche. The *fundamental Grinellian niche* $\mathbf{N}_\mathrm{F}$, or just *fundamental niche*, can be described as

$$\mathbf{N}_F = \{\vec{e_j} | r(\vec{e_j}) > 1\} \tag{1}$$

where $r(\vec{e_j})$ is defined as the density-independent intrinsic growth rate of a population as a function of all abiotic environmental variables in a cell j (Soberón, 2007). If we now assume the relation $\gamma : \mathbf{E} \rightarrow \mathbf{G}$, a geographic information system (GIS) can be used to map $\mathbf{N}_\mathrm{F}$ to the geographic space **G** to get a set of locations where the species can potentially live (Guisan & Zimmermann, 2000). Soberón (2007) calls this the *fundamental area of distribution of a species* **A** that can be described as

$$\mathbf{A} = \gamma(\boldsymbol{N}_F) = \{j \in \mathbf{G} | r(\vec{e_j}) > 1\} \tag{2}$$

In nature, **A** is almost impossible to calculate given the fact that almost every organism interacts with the biotic environment to either profit or suffer from its consequences as discussed in section 2.1.3. Consequently, the fundamental niche is often only applied in mechanistic models which include the connection between functional traits of organisms and their environments into spatial distribution models (e.g. Kearney and Porter (2009)). Further useful applications could be based on *ex situ* data like from aquariums or botanical gardens. Another reason that makes the realization of this niche so hard is the fact that prior to modeling, the effect of each environmental variable on every organism needs to be tested in the laboratory (Guisan et al., 2017).

2.1.3 Biotic environment and the realized niche

In nature, interactions between organisms, on the same or different trophic levels, are everywhere while either promoting or reducing a species' fitness. Some species even rely completely on the existence of others. As an example, the yucca plant can't exist without the co-existence of the yucca moth and vice versa (Pellmyr & Leebens-Mack, 1999). While the plant can only be pollinated by the moth, the animal's larvae depend on the plant's seeds for food. Historically, competition between species was considered as the first biotic interaction in the definition of the niche concept by Hutchinson. Nonetheless, research shows that also interactions between different trophic levels, as for example plant-herbivore or predator-prey, as well as interactions in functional groups like symbiosis or host-parasite can be important (Broitman et al., 2009; Lira-Noriega & Peterson, 2014).

As described by the fundamental niche, an organism can occupy every location with suitable environmental/abiotic conditions. However, in nature, when multiple species compete for the same resources or environmental conditions the one with the higher fitness will succeed and suppress the other as predicted by the *competitive exclusion principle* (Hardin, 1960). The fundamental niche concept can be extended to the *realized niche* concept, by adding biotic effects, to obtain a more realistic depiction of the real world. In geographic space this can be seen as the intersecting space between the area with suitable abiotic conditions $\mathbf{A}$ and the area with suitable biotic conditions $\mathbf{B}$. According to Soberón (2007), this area is called the *realized area of distribution* $\mathbf{J_R}$ and can be defined using the overall growth rate of the population dx_j/d_t:

$$\mathbf{J_R} = \left\{ j \middle| \frac{dx_j}{dt} |_{x_j \approx 0} > 0 \right\} = \mathbf{A} \cap \mathbf{B} \tag{3}$$

Following this definition, $\mathbf{J_R}$ also contains potential areas of distribution where the respective species possibly don’t have access to. Consequently, in nature, the area in which a species population can grow and persist is the intersection of areas where all three conditions are fulfilled (dispersal, abiotic environment and biotic environment). This area can be called the *occupied area of distribution* (Gaston, 2003).

2.2 Habitat Suitability Modeling

The twenty-first century is witnessing an outstanding biodiversity crisis, with our planet at risk of a sixth major species extinction, driven by human activities (Ceballos et al., 2010). Today, major threats include pollution, habitat destruction, alteration, and fragmentation. Climate change and anthropogenic impact have led to an alteration of ecosystems, changes of habitats, biodiversity loss and biological invasions (Githeko et al., 2000; Jaureguiberry et al., 2022; Van Der Wal et al., 2008). Consequently, conservation of endangered species, protection of ecosystems and biodiversity, or the management of vector-carrying species become relevant fields of research. Therefore, effective strategies and techniques are necessary to predict the future distribution of life on our planet. Simultaneously, scientific progress is accelerating rapidly. Due to the rapid developments in statistical analysis and computing power, the capacity to model and forecast ecological patterns, with applications in evolutionary biology, biogeography, and conservation biology, has increased strongly (Nordhaus, 2007).

According to Guisan et al. (2017), there has been a large increase in interest in predictive HSMs during the twenty-first century to identify areas that are suitable for a particular species based on environmental conditions. This rise in interest can be attributed to two main factors. First, the last twenty years have seen an exponential rise in computing capabilities, with the emergence of geographic information systems and remote-sensing technologies leading to the creation of high-resolution environmental data sets, which together provide enhanced opportunities to analyze, model, and predict a wide range of species across various spatial scales (ranging from local to global). Second, driven by these technological advancements and in the absence of alternative methods suitable for assessing numerous species across diverse regions or large geographic areas, HSMs have become a necessary tool in applied ecological and environmental sciences. They have proven especially valuable for assessing the potential impacts of global human-induced environmental changes on biodiversity and ecosystems (C. J. Johnson & Gillingham, 2005; Pearson & Dawson, 2003; Schwartz, 2012).

Furthermore, in the context of this work, HSMs were previously successfully used to model the ecological niche of various disease-carrying vectors (Ayala et al., 2009; Eisen et al., 2006; Koch et al., 2016). The following sections will include the framework of HSMs (section 2.2.1), the process of quantifying the niches (section 2.2.2), how to use them to make predictions about a species’ distribution (section 2.2.3), the models that can be used (section 2.2.4), and how the choice of predictors relates to the geographic scale (section 2.2.5).

2.2.1 Framework of habitat suitability models

Before diving into HSMs, it needs to be clarified that in literature the terms habitat suitability models, species distribution models, ecological niche models or habitat distribution models belong to the same class of models.

HSMs are empirical models that connect observation data from the field (target, response or dependent variable) with environmental predictor variables (independent variables), utilizing response surfaces that are derived either statistically or theoretically (Guisan & Zimmermann, 2000). A response surface can be viewed as the outcome resulting from the interaction of multiple predictor variables, effectively describing the combined effect of physiological response curves (see section 2.2.2). Observation data can be characterized differently, like simple presence, presence-absence or opportunistically acquired data like from natural history collections (Graham et al., 2004). Environmental predictors can have direct or indirect effects on the spatial distribution of species as described in chapter 2.1. These predictors often follow a hierarchical order and are relevant on different spatial scales (McGill, 2010). The prediction of species distribution for a large area with a coarse resolution often relies on abiotic factors, primarily on climate variables which are obtained from remote sensing applications. This potential distribution can be referred to as the bioclimatic range of a species (Guisan & Thuiller, 2005). The potential distribution can further be constrained by adding dispersal and disturbance factors to the model. For instance, extreme events like natural disasters or the impact of human activity can alter this distribution on smaller scales, which can lead to the extinction or creation of new habitats. For instance, Fischer et al. (1990) recognized land use as the dominant predictor for plant distributions in a human-disturbed environment. Further predictors on smaller scales are influences of the biotic environment and resource factors like food or nutrients (Guisan & Thuiller, 2005). Including all these elements will yield the realized distribution as it builds on the realized niche concept as described in chapter 2.1.3.

According to Guisan & Thuiller (2005), the conceptual procedure in building an HSM involves the following six steps:

a. Conceptualization
b. Data preparation
c. Model fitting
d. Model evaluation
e. Spatial predictions
f. Assessment of model applicability

In the **conceptualization** phase, a suitable framework for the given problem will be developed. This involves determining the correct spatial scale for which predictions will be made, choosing appropriate predictors based on profound knowledge of the given species as well as determining a suitable sampling design. Given the problem statement and the available data, a suitable modeling technique needs to be chosen as further described in section 2.2.4.

The process of **data preparation** includes collection, cleaning, manipulation, alignment and integration of predictor and target data. In data collection, data on species distribution like presence or presence-absence data, as well as corresponding predictor data are gathered. Next, data sets need to be cleaned to only contain information relevant to the given problem and to remove errors, duplicates or inconsistencies. As data can be characterized by numerous formats like raster, vector or numerical data, it needs to be transformed into a suitable format for the analysis. This may involve spatial operations (e.g. clustering, kernel density, aggregation, etc.), dealing with missing values, normalizing or transforming data for better model integration. Finally, post-processed data needs to be integrated into a coherent data set used for modeling.

For most species and regions, there aren't enough resources to enable comprehensive surveys that record a species' presence and absence. Most data only contain presence observations as they are derived from the commonly applied general atlas distribution framework or from large museum collections (Peterson, 2003; Underhill & Gibbons, 2002). Citizen science network data have also been proven to be reliable and cost efficient in monitoring the presence of mosquito species (Palmer et al., 2017). Data from large community databases like the GBIF or Mosquito Alert must be treated carefully due to (i) uneven spatial coverage of the actual distribution of a species due to heterogeneous SE, (ii) uncertainties in the identification of a species, (iii) insufficient sampling design, (iv) spatial autocorrelation in sample locations and (v) low or unknown accuracy of sample locations (Guisan et al., 2017). During data preparation, a well-structured sampling design for pseudo-absences may help to reduce bias inherent to large community data inventories by applying the same bias to the sampling of pseudo-absences (Phillips et al., 2009).

In the next step during **model fitting**, post-processed data will be split into sets of training and test data. Training data are used to fit a suitable machine learning algorithm. During this process, the model estimates its parameters based on the given training data to capture the relationship between target and predictor data in the best way possible (Molnar, 2022). To evaluate how well the model describes the data, techniques like goodness-of-fit metrics, such as *Akaike Information Criterion* (AIC) or *Bayesian Information Criterion* (BIC) and other metrics like R^2, *Root Mean Square Error* (RMSE), or in case of binary target data the *Area Under the Receiver Operating Characteristic* (AUROC) can be used (Akaike, 1998; Bradley, 1997; Gao, 2024; Hodson, 2022; Schwarz, 1978).

During **evaluation,** test data are used to assess how well the fitted model performs on unseen data. Cross-validation techniques are used to split data into multiple sets of training and test data to subsequently apply evaluation metrics described previously. Random cross-validation may fail to capture spatial, temporal or hierarchical (random effects) structures, often inherent to ecological data. Block cross-validation can be used to address this issue by splitting data based on space, time or random effects (Roberts et al., 2017). Therefore, model performance can be evaluated for different regions or time ranges (e.g. neighborhoods, years).

The fitted and evaluated model can now be used to **predict** species distribution or habitat suitability in unsampled regions. This phase is described in detail in section 2.2.3.

Lastly, the **assessment of model applicability** evaluates if a model was suitable for the given problem and context. In this phase, researchers consider the model's domain of validity and the environmental predictors the model was trained on, and whether model predictions are trustworthy. Potential drawbacks or future improvements, as well as applicable regions will be analyzed and discussed (Guisan & Zimmermann, 2000).

2.2.2 From observations to niche quantification

As discussed in chapter 2.1, every region on Earth can be characterized with a set of environmental variables, consequently defining it as a habitat that may possibly be inhabited by a certain species (Kearney, 2006). If one wants to model the habitat suitability of such a region for a given species, the ecological niche of that species determines if that region is a suitable candidate by defining the environmental limits for which a population can grow and persist. Therefore, if the realized ecological niche of an organism is quantified and all relevant environmental parameters are at hand, habitat suitability can be determined for every geographic region. This insight is the basic principle of HSMs. HSMs therefore provide an estimate of the realized environmental niche of a species in a specific geographic area during a particular timeframe, determined by the dates of sampling and the environmental variables applied (Guisan et al., 2017). But how can we use field observations to quantify the ecological niche of a species?

In statistical modeling, the realized niche of a species is quantified by analyzing known locations and their associated environmental attributes (e.g. remote sensing data) to model the species-environment relationship (Franklin, 1995). Instead of only presence data, using presence-absence or even abundance data can enhance discrimination between suitable and unsuitable habitats (Brotons et al., 2004; Howard et al., 2014). Using a variety of statistical techniques the link between species presence (and absence) data and the environmental gradients can be derived by quantifying its response curves. To be able to capture the full range of the realized environmental niche of a species, the **assumption of the species being in equilibrium** with its environment needs to be made (Elith & Leathwick, 2009; Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000). This suggests that the species' population is stable and well-adapted to the current environmental conditions, implying that the reproductive success, survival rates, and overall population dynamics are balanced with the availability of resources, habitat characteristics, and ecological interactions within that environment.

Given the complex and dynamic nature of our environment, it is challenging to assume that an organism is in equilibrium with its environment. Species and environmental data are usually collected for a certain range of time and space and therefore only reflect a snapshot of the environment-species relationship in a permanently changing system (climate change, natural disasters, biological invasion, change of land use, etc.). Therefore, in practical HSM research, the assumption of a pseudo-equilibrium is made (Franklin, 1995; Guisan & Theurillat, 2000). When modeling invasive species like the Asian tiger mosquito, this assumption does not hold given the fact that the animal just arrived in a new environment and thus not being able to fully establish in the given study area. In this case, literature suggests fitting of HSMs using data from its native range (Peterson, 2003). This is especially useful when analyzing potential large-scale species distributions based on the climatic niche (Broennimann et al., 2007).

Unfortunately, using data from the native range of *Aedes albopictus* is not meaningful when analyzing its local distribution within the Austrian environment since climatic variations are negligible on that scale. The environmental factors assumed to influence mosquito distribution in Graz are unique and cannot be precisely represented by data from its native region. Austria's landscape, characterized by specific land use, vegetation, and urbanization patterns, may result in different mosquito distributions compared to those in its native Southeast Asian range. Furthermore, as this work aims to statistically derive drivers for the species distribution in Graz, using data from its native range would not be meaningful. Consequently, despite the growing population and the ongoing establishment of the Asian tiger mosquito in Graz, the only possibility is to assume pseudo-equilibrium (Bakran-Lebl et al., 2025). Understanding the context of the assumption of pseudo-equilibrium is important for the interpretation of model predictions and for potential improvements during model conceptualization. Simply predicting habitat suitability of the Asian tiger mosquito in Graz might underestimate the future distribution as the potential realized niche cannot be fully quantified.

2.2.3 Predicting spatial distributions from the quantified niche

After the realized environmental niche or a part of it (depending on the study area) has been fitted, the next step is to determine how this quantified niche can be projected into geographic space to determine the potential distribution of a species. If the environmental values for a specific set of species presence–absence locations are incorporated into a statistical model as the input and then directly projected onto a series of environmental layers, without any visualization in environmental space, it can be classified as a "blind approach". A more effective method for comprehending how a species' niche response has been integrated into a model is to first examine the prediction process in environmental space rather than solely in geographic space. This can be accomplished by projecting the habitat suitability into the environmental space where the niche was originally established (Guisan et al., 2017).

To predict the potential distribution of a species the statistical model representing the species' environmental responses must be applied to the geographic space where the species was originally observed. This involves using environmental conditions for specific locations to apply the statistical functions developed in environmental space. This process allows for the spatially explicit expression of habitat suitability for the species across the study area. For spatial predictions, it is essential to calculate a habitat suitability value for each geographic cell within the region by implementing the model on the environmental layers that contain the niche variables utilized in the model. By creating a stack of layers representing environmental variables used to fit the HSM, a habitat suitability score can be calculated for any cell where predictor variables are available (Soberón, 2007).

Consequently, potential species distributions can be analyzed in unsampled regions or in different temporal ranges by projecting the model in space and time. Therefore, we can analyze potential areas at risk for invasive species or potential future habitat suitability distributions due to climate change (Bonizzoni et al., 2013; Schwartz, 2012). Projecting a model into space and time needs to be treated with caution due to reasons like environmental comparability, niche stability or incomplete knowledge of the full niche (Guisan et al., 2017).

2.2.4 Which model to use?

When conceptualizing a modeling framework, researchers should carefully consider the intentions of the research and the available data to analyze. A large variety of methods are applied in modeling habitat suitability (Franklin, 2010). Models in this field can broadly be categorized into three groups: **Descriptive, explanatory** and **predictive models**. Descriptive models try to capture the link between predictor variables (e.g. environmental gradients) and the response variable (e.g. species distribution) without explaining underlying mechanisms. This kind of analysis is often made prior to explanatory or predictive models to select the right amount and type of predictors (Guisan et al., 2017). Multivariate statistical analyses are prominent methods to achieve these goals (Dray et al., 2003).

When using predictive models (Côté & Reynolds, 2002), accurate spatial predictions are prioritized over detailed explanations of underlying mechanisms. The goal is to optimize prediction accuracy by choosing the optimal set of predictors, often using parsimonious models with few, uncorrelated variables to minimize for instance the sum of errors (J. B. Johnson & Omland, 2004). Even if a model's underlying mechanisms are not fully understood (e.g. "black box" models), its predictive power might still be valuable. However, simply maximizing fit on training data can lead to biased parameter estimates and overfitting. The best predictive model is the one that generalizes well and yields the most accurate predictions on independent or cross-validated data (Merow et al., 2014).

Lastly, explanatory models aim to explain the ecological processes that lead to patterns in the response variable (e.g. plant species) (Austin et al., 1990). This approach is also called the hypothetico-deductive approach as it utilizes prior knowledge of the system to formulate a set of hypotheses which can be confirmed or disproven by estimating model parameters (Guisan et al., 2017). Deductive approaches are generally biophysical models that represent known relationships by parameterizing the models using (field) measurements. Because machine-learning determines the effect of the various predictors in the response variable, it’s important that the model is interpretable, otherwise the knowledge gain is limited. Therefore, the chosen algorithms must not only provide reliable predictive power but also remain easily interpretable. Depending on the type of response variable, as well as the characteristics of observation data, a variety of approaches are available.

While in the best-case scenario, an explanatory model with maximum predictive performance would be ideal, there is a trade-off between these two in reality (Guisan & Zimmermann, 2000). Another major difference when selecting the method arises from the quality of data needed. The first category of models includes classification and regression tree analyses, generalized linear models (GLM), generalized additive models (GAM) or artificial neural networks. These methods rely on quality presence-absence data to be able to derive statistical functions to correctly classify habitat suitability scores based on the distribution of presence and absence of a species (Brotons et al., 2004).

Generalized regression models like GLMs and GAMs are very flexible models offering high interpretability and are therefore commonly used in HSMs (Elith & Leathwick, 2009; Guisan et al., 2017). While the classical least squares (LS) regression is only valid when the response variable is normally distributed with constant variance (homoscedasticity), GLMs offer a more adaptable type of regression that can handle various distributions and non-constant variance (Molnar, 2022). In GLMs, predictors are linked to the response variable through a link function, enabling transformations to linearity and keeping predictions within logical limits. This flexibility allows GLMs to work with distributions like Gaussian, Poisson, Binomial, or Gamma (Guisan & Zimmermann, 2000). In ecology, this is important as species observation data are often count data (Poisson distribution) or presence-absence data (Binomial distribution). GAMs build on GLMs by allowing non-linear relationships between predictors and the response variable using several non-linear smooth functions instead of polynomial functions. Like GLMs, they can accommodate various distributions and non-constant variance, providing greater flexibility for modeling complex data patterns (Guisan et al., 2002; Wood, 2017). In applied ecological research, numerous examples illustrate the use of generalized regression models in studies related to habitat suitability in fields of biodiversity loss (Fragnière et al., 2022), marine biology (E. J. Brown et al., 2019), conservation planning (Hunt et al., 2020), or climate change ecology (Descombes et al., 2020).

The second category comprises all methods which can work with presence-only data. Common approaches are Bioclim or Ecological Niche Factor Analysis (ENFA) (Brotons et al., 2004). Without the need for absence locations, these models analyze the environmental envelope (all relevant environmental conditions) under which a species is present, which can then be projected to other areas

(Hirzel et al., 2002). ENFA was applied in numerous ecological fields of research including marine biology (Bryan & Metaxas, 2007), conservation planning (Santos et al., 2006) or climate change impact (Estrada-Peña & Venzal, 2007). Nonetheless, by generating pseudo-absences, presence-only data can also be analyzed by GLMs and GAMs as they are able to handle binary data.

2.2.5 Predictor variables and geographic scale

When creating HSMs, the choice of suitable environmental variables used as predictors is crucial depending on the physiological characteristics of the species and the geographic and temporal scale for which the prediction is done (McGill, 2010; J. A. Wiens, 1989). To give a few examples, when analyzing the distribution of a species that lives underground one might not focus on solar radiation as a predictive variable. Regarding geographic scale, performing a global prediction of habitat suitability for a given plant species, one might fail to use soil type as a predictor, which has a strong impact on local or regional scales, because water availability and particle-size distributions vary strongly over short distances and is therefore almost impossible to integrate in a global-scale model. The connections between climate and vegetation that are obvious at larger scales might decrease in importance at smaller scales, overshadowed by the influences of various biological processes or competition (Greig-Smith, 1979). In a human-disturbed landscape, land use was identified as the factor with the highest predictive power for modeling plant community distributions (H. S. Fischer, 1990).

Consequently, during model conceptualization, an appropriate spatial and temporal scale depending on the goal of research should be determined allowing for the correct choice of predictor variables. Additionally, ecological and physiological knowledge about the species helps to identify relevant predictors, since every organism reacts differently to certain environmental gradients (Woodward & Kelly, 2003). The better we understand an organism's characteristics, the more accurately we can predict its distribution.

The predictive power of HSMs using multiple predictors relies heavily on the sample size of observation data. Sufficient data are necessary to fit qualitative response curves. The effect of sample size on the quality of HSMs is analyzed in several papers (Guisan et al., 2007; Hernandez et al., 2006; Thibaud et al., 2014). According to Guisan et al. (2017), the predictive power of HSMs decreases severely with sample sizes smaller than 30 while the effect of sample size becomes less relevant with more than 50 presences. Too little sample size can lead to overfitted models as predictor variables cannot be fitted in a probabilistic way. At least ten observations per predictor can be seen as a rule of thumb, since more complex models need more training data to be fitted.

2.3 Aedes albopictus

Since explanatory models rely on prior knowledge of the species and its environmental system to select suitable predictor variables, the upcoming sections will explain fundamental characteristics of *Aedes albopictus* to understand its role and interactions within the ecosystem. First, it will be discussed how the species could spread globally, its geographical distribution (section 2.3.1) as well as its bioclimatic envelope and future dimensions within Europe (section 2.3.2). Next, general characteristics like life cycle, feeding, and breeding behavior will be discussed (section 2.3.3) and lastly, the animal's role as vector for arboviruses will be explained (section 2.3.4).

2.3.1 Spread and geographical distribution

Originating from the temperate and tropical forests of Southeast Asia, the mosquito spread rapidly to other regions during World War II. Several dengue outbreaks were recorded in Japan between 1942 and 1944. It is assumed that *Aedes albopictus* was introduced as a vector to Nagasaki through warships coming from Southeast Asia. In 1943, pilots coming from Fiji were responsible for the introduction of the mosquito to Hawaii (Gratz, 2004).

After World War II, due to the rapidly increasing need for global trade, intercontinental ship traffic is assumed to be the dominant driver of the global spread of *Aedes albopictus.* The extensive global trade of 'lucky bamboo' and used tires, acting as water containers containing eggs and larvae, is the major pathway for the establishment of the tiger mosquito in continents like Europe or North America. The latter led to the high-impact introduction of the mosquito to Texas (USA), São Paulo (Brazil) and Padua (Italy) (Scholte & Schaffner, 2007). As of now, the insect arrived in other parts of Asia, Africa, the Americas, Australia, and Europe and was therefore found on every continent excluding Antarctica (M. U. Kraemer et al., 2015).

In Europe, first sightings of *Aedes albopictus* were reported in Albania in 1979 without any further reports until 1990 in Italy (Sabatini et al., 1990). Due to its mild winters, the largest established populations of the species in Europe are located along the Mediterranean and the Adriatic coast in countries like Spain, France, Italy, Greece or Albania. Through the passive transportation by cars and trucks, the species could further establish populations in more northern countries (Scholte & Schaffner, 2007). The mosquito can now be found in numerous European countries of higher latitudes, including Germany, Austria, Switzerland and Sweden (*Aedes Albopictus - Current Known Distribution*, 2024).

Since *Aedes albopictus* was first documented in Austria in 2012 (Seidel et al., 2012), it has been observed nationwide, reaching all federal provinces by 2022, with populations in Vienna and Graz capable of surviving winters ("Asian tiger mosquito throughout Austria," 2023; Bakran-Lebl et al., 2022). Since then, annual mosquito counts are rising continuously and are likely to increase in the future due to climate-change-related rise of average temperatures as discussed in the next section (Bakran-Lebl, 2025b).

2.3.2 Future distribution in Europe

The sole fact that *Aedes albopictus* is ranked as the highest-listed insect species on the top 100 list of the world's most invasive species is reason enough to comprehend and analyze the future spread due to its strong adaptability (www.iucngisd.org). Global- or continental-scale models for predicting habitat suitability are mainly based on climatic variables as they analyze the potential distribution based on the bioclimatic envelope of the species (Guisan & Thuiller, 2005). Climatic thresholds for the establishment of a population and for the ability of overwintering are determined as follows by Medlock et al. (2015):

○ Minimum mean annual rainfall:	500 mm
○ Minimum mean winter temperature:	0°C
○ Mean annual temperature:	11°C

Additionally, the period with temperatures exceeding 11°C should consistently last for more than 186 days annually (Scholte & Schaffner, 2007).

Climate change significantly impacts the distribution of *Aedes albopictus* by altering the climatic conditions that determine habitat suitability. As global temperatures rise and precipitation patterns shift, regions previously unsuitable for the mosquito may become viable habitats, while areas that have traditionally supported their populations may face changes in suitability.

Warmer temperatures can extend the mosquitoes' active season and expand their range into higher latitudes and altitudes, where milder winters and sufficient rainfall provide favorable conditions.

Based on future climatic conditions, Fischer et al. (2011) predict a further expansion of *Aedes albopictus* habitat into western and central Europe between 2011 and 2040. Eastern Europe will experience the same trend with a temporal delay. Caminade et al. (2012) support those findings, highlighting a future increase in suitable conditions over Northwestern Europe and a slight future decrease over Spain, due to drier and warmer summers. In the future, low minimum temperatures will be the limiting factor over Eastern Europe for the establishment of the mosquito, whereas over Central Europe temperature thresholds will be replaced by limits for precipitation (Cunze et al., 2016).

2.3.3 Characteristics of *Aedes albopictus*

In 1895*, Aedes albopictus* was described for the first time as 'the banded mosquito of Bengal' in India by the English entomologist Frederick A. A. Skuse (Huang, 1968). The species is endemic to the forests of Southeast Asia, where it feeds on wildlife as a zoophilic mosquito. Like all other flies, in its lifecycle, the animal goes through four different stages, namely egg, larva, pupa and adult (Marini et al., 2020). As the animal breeds in aquatic environments, its favored breeding sites were water-filled tree holes or bamboo stumps located at the forest edges, which led to its classification as a rural vector (Bonizzoni et al., 2013).

The species' high ecological plasticity enabled it to quickly adjust to the impacts of human interaction (domestication), as it is described by Tabachnick (1991) for their close relatives *Aedes aegypti*. The mosquito quickly turned on domestic animals and humans and started to use alternative man-made containers like plastic boxes, plant pots or used tires as breeding sites. As an example, during World War II an important factor for the dengue epidemic in Japan was the installation of water tanks across cities to prevent fires from bombardment, in which the insects could breed (Gratz, 2004). Due to this domestication, *Aedes albopictus* quickly became a significant and often sole vector for arboviruses in rural areas across the world like in France, Italy or southern China (Gould et al., 2010; Rezza et al., 2007; Wu et al., 2010). Li et al. (2014) showed that urbanization in Guangzhou, China, increased larval habitats with two-fold higher densities in pupae and three-fold higher densities in adult-stage mosquitoes compared to suburban and rural settings. In the temperate climate of St. Louis, Missouri, mosquito counts in an urban environment were an order of magnitude higher than ones in a rural setting (Westby et al., 2021). Additionally, the survival time of adults was also increased in urban areas. These findings agree with a large collection of immature *Aedes* mosquitoes in southeastern Côte d'Ivoire across different landscapes with the highest proportion of positive breeding sites being in the urban area (Zahouli et al., 2017). Nonetheless, while most studies agree with the fact that the mosquito thrives in urban environments, it needs to be mentioned that findings are not consistent throughout the global range. For example, in Lambaréné, Gabon, *Aedes albopictus* abundance was higher in the rural environment (Bikangui et al., 2023).

The tiger mosquito, as an exophilic and exophagic animal, is primarily active during the day and tends to feed outdoors, with peak biting times in the early morning and late afternoon. However, exceptions to this behavior have been noted, influenced by factors such as season, geographic region, availability of hosts, and the characteristics of the human environment. The mosquito shows a preference for biting mammals with inconsistent findings on preference between humans and animals (Paupy et al., 2009). A study by Richard et al. (2006) classifies *Aedes albopictus* as a poor vector for human pathogens like arboviruses, due to its zoophilic and opportunistic feeding behaviors. However, other studies show that the preference of *Aedes albopictus* for their host choice highly depends on the mosquito population's geographical origin. By observing blood meals from wild mosquitoes or by conducting host choice experiments on wild populations an anthropophilic behavior of *Aedes albopictus* could be proven (Delatte et al., 2010; Niebylski et al., 1994).

Moreover, *Aedes albopictus'* tendency to feed on various animal species not only boosts its biological traits, such as reproductivity and survival, but also increases the risk of spreading zoonotic pathogens between animals and from animals to humans. The wide range of hosts it utilizes significantly contributes to its ability to invade and establish itself in diverse environments, from forests to urban areas (Paupy et al., 2009).

Compared to the mosquito's close relative *Aedes aegypti, Aedes albopictus'* ability to adapt to changing climatic conditions is unmatched. While tropical populations did not show overwintering, temperate ones were able to survive winters by producing eggs that are able to undergo a winter diapause (Hawley, 1988). While European diapausing eggs could survive a cold spell of -10°C, tropical eggs were only able to survive -2°C (Thomas et al., 2012). This characteristic is what makes it possible for *Aedes albopictus* to establish itself in Europe and other temperate regions in the first place.

2.3.4 A vector for arboviruses

Aedes albopictus acts as a vector for up to 26 different arboviruses, including dengue, chikungunya, or yellow fever, therefore being able to transmit diseases between humans via bites (Paupy et al., 2009). Following the geographical expansion of *Aedes albopictus* and other *Aedes* species, it is estimated that almost half of the world's population is at risk of dengue virus infections (Brady et al., 2012). The primary vector in the Pacific responsible for several epidemics is the closely related *Aedes aegypti* (Calvez et al., 2016). Nonetheless, *Aedes albopictus* poses a significant threat to regions outside of the tropics and subtropics like Europe, North America or Japan, due to the insect's strong climatic adaptation as discussed previously.

Between 1942 and 1944, the tiger mosquito was responsible for the first ever widespread dengue outbreak in Japan reoccurring every summer with at least 200,000 infections in total (Gratz, 2004). In 1978, dengue reemerged in China after 32 years, leading to large epidemics affecting hundreds of thousands in Guangdong, Guangxi, and Hainan Island. All four dengue virus types were involved with *Aedes albopictus* being the vector in the inland (Fan et al., 1989). Between 2004 and 2007, the largest Chikungunya virus (CHIKV) outbreak ever recorded occurred in the Indian Ocean Islands and India with an estimated 5000 cases for the Union of the Comoros, over 200,000 for the island of La Réunion and up to 1.3 million for India with many further infections on Madagascar, Mauritius and the Seychelles (Njenga et al., 2008).

Following its introduction to Europe in 1979 in Albania, several arboviruses were able to be locally transmitted across the continent. In 2007, the first ever recorded Chikungunya outbreak in Europe was reported with 197 cases (Angelini et al., 2007). *Aedes albopictus* was identified as the vector for isolated cases of dengue infections in France in 2007 (Gould et al., 2010). Similar observations occurred in Catalonia, Spain, where 65 patients tested positive for dengue between April and December 2015 (Aranda et al., 2018). According to the European Centre for Disease Prevention and Control (ECDC 2024; www.ecdc.europa.eu), since 2010, when the first recent dengue outbreak was documented, there have been 48 vector-borne dengue outbreaks. From 2010 to 2017, there were up to 3 outbreaks annually, which increased to 5 outbreaks in 2018, 7 in 2020, and 10 in 2022. The highest number of outbreaks occurred last year, in 2023, with 8 in France, 4 in Italy, and 2 in Spain. This trend, combined with the recent increase in observations of *Aedes albopictus* in Austria, highlights the relevance of understanding their small-scale distribution in the Austrian context (Bakran-Lebl et al., 2025).

2.4 GAM Model

In ecological research, two of the most popular regression modeling techniques are GLMs and GAMs, potentially extended to GLMMs and GAMMs by including mixed effects to describe intragroup variations (see Guisan et al. (2002) for an overview of generalized regression models in habitat suitability modeling). One fundamental reason for this trend is their ability to handle the wide range of distributions found in ecological data (Poisson, Gamma, binomial, negative binomial, Gaussian) (Bolker et al., 2009; Guisan et al., 2002; Wood, 2017).

GLMs are extensions of simple linear models allowing for non-constant variance and non-linearity typically found in nature. They use so-called link functions to describe an assumed relationship between the linear combination of predictors and the mean of the response variable. These link functions depend on the distribution of ecological data and thus allow for a more flexible modeling of ecological relationships, often not properly represented by Gaussian distributions assumed in simple linear regression models. Mathematically, GLMs can be described as

$$g\big(E_Y(y|x)\big) = \beta_0 + \beta_1 x_1 +_{\dots} \beta_p x_p = X^T\beta \tag{4}$$

with the link function g, the weighted sums $X^T\beta$ (often called the linear predictor) describing the weighted sum of predictors with coefficients β, and the assumed probability distribution E_Y of the response *y* from the exponential family. GAMs are extensions of GLMs using a linear combination of smooth functions to describe the predictor-response relationship. Consequently, the shift from weighted linear predictors to smooth functions allows for the modeling of more complex relationships. The previous mathematical definition can therefore be altered to

$$g\big(E_Y(y|x)\big) = \beta_0 + f_1(x_1) + f_2(x_2) +_{\dots} + f_p(x_p) \tag{5}$$

with the linear term $\beta_p x_p$ being replaced by a more flexible function $f_p(x_p)$. Since individual smooth functions can be analyzed visually, GAMs allow for high interpretability of the modeled relationships (Molnar, 2022).

Each smooth function is built from simpler, fixed, weighted *basis-functions*. These can be viewed as building blocks that, when combined, create the smooth function of a predictor. Consequently, a smooth function can be defined by the sum of *K* basis-functions $b_{j,k}$ multiplied by their respective coefficients $\beta_{j,k}$:

$$f_j(x_j) = \sum_{k=1}^{K} \beta_{j,k} b_{j,k}(x_j) \tag{6}$$

While low values for *K* might lead to an underrepresentation of variation in the data, excessive values might lead to overfitting and increased computational effort. To counteract overfitting, a *smoothing penalty* is applied to coefficients of basis-functions, thus avoiding wiggliness and noise and ensuring appropriate complexity. A *smoothing parameter* λ controls the balance between overfitting and smoothing. When set to zero, no smoothing penalty will be applied, resulting in overfitting while large values ($\lambda \rightarrow \infty$) will yield a straight line. Using the R package *mgcv,* λ can be selected automatically using the method *REstricted Maximum Likelihood* (REML, (Wood, 2017)). The *effective degrees of freedom* (edf) parameter can be used to measure the complexity of a smoother as it represents the number of parameters used to fit the model. Consequently, *edf* cannot be larger than *K* since *K* represents the maximum number of basis-functions. Values for *edf* well below *K* indicate that only a small proportion of basis splines contribute to the predictor's explanation of data variation. Values close to *K* might indicate that the current number of basis-functions might not be enough to sufficiently describe the data (Pedersen et al., 2019). Consequently, by successively comparing *K* and *edf,* the correct number of basis-functions can be estimated for each predictor. In general, *K* is set too high rather than too low since the main drawback is high computational costs compared to underrepresentation of the data. Since smoothers are penalized by λ, high values of *K* do not represent a major problem with respect to overfitting (Wood, 2017).

When it comes to the type of smoothers, there is a variety to choose from depending on the needs and the associated penalty matrix. The most used spline type is the *Thin Plate Regression Spline* (TPRS) since they are general purpose splines. They use a penalty matrix which is based on the squared derivatives of basis-functions. Since a derivative describes the slope of a function, the applied penalty is determined by the wiggliness of the underlying basis-functions. Other smoother types relevant in this work are *cyclic cubic regression splines* (for periodical predictors, like *day of the year*) and *random effects* (for mixed models) (Wood, 2003).

In GLMs and GAMs, smooth functions represent fixed effects since they are applied globally within the model, thus influencing each observation equally in the data set. Since data can be spatially, temporally or categorically nested, intragroup variations might lead to varying relationships between predictors and the response due to underlying processes within a group. To address this issue, GAMs can be extended to GAMMs by including random effects for variables representing the nested structure (e.g., years, neighborhoods, species types and other categorical features). Random effects are a certain type of smoother used to model intragroup variations. They can be applied globally, affecting all data (random intercept) or to specific predictors only (random slope). Consequently, random slopes need to be defined using interaction terms between the grouping factor and the predictor variables, allowing the effect of those predictors to vary across groups (Pedersen et al., 2019).

As an example, one wants to analyze the relationship between the height of plants and soil nitrogen levels for a given study area. A variety of plant species can be found within the study area with each species having its own response curve to the level of nitrogen, but all of them are subject to the same overall nitrogen levels. Consequently, each species will grow to different heights influenced by their evolutionary adaptations and physiological traits. However, since all of them coexist in the same environmental range, there are shared environmental interactions and influences that shape their growth. While the level of nitrogen can be used globally as a fixed effect covering all species, a random intercept can be applied to account for variations in the baseline growth (height) among different spatial areas, to account for nested unobserved reasons. Each area might have unique characteristics, such as differing amounts of other soil nutrients or microclimate conditions, which affect plant growth but are not directly measured. This spatially varying random intercept captures the inherent differences in average plant height due to these unobserved, location-specific influences. Additionally, a random slope could be added to specify that the effect of nitrogen on plant height might vary between species. This means that while one species might experience a steep increase in height with increasing nitrogen, another species might only exhibit a small response. Without this random slope effect, the model would fail to predict a species-specific response since only the average relationship was known. Contrarily, groups could be modeled separately resulting in several individual models with potentially lacking training data due to data sparsity within groups (i.e., a very rare plant species).

Additionally, GAMs allow for the interaction of predictors using interaction terms. These are applied when the effects of predictors are assumed to change depending on the values of other predictors (Molnar, 2022). To give an easy example, when analyzing the growth rate of a plant species, the influence of sunlight might change depending on the availability of water. Under very dry conditions, the duration of sunshine might be irrelevant since water for plant growth is not abundant. On the other hand, when water is available, the amount of sunshine might heavily influence the plant's growth rate.

Chapter 3 Data and Methods

After setting the fundamental background of this work, this chapter covers data and methods applied in this work. Sources and characteristics of the data used in this research are described in section 3.1. Subsequently, the study area and climatology of Graz will be described (section 3.2) followed by the software and packages used for data preprocessing and modeling (section 3.3). Data preparation and methodology applied in this modeling framework are described in section 3.4.

3.1 Data

This section covers a short description and an overview of Mosquito Alert data and other data layers used in the course of this work. A more detailed description, as well as descriptive statistics are covered in the chapter data preparation (section 3.4.1 for Mosquito Alert data and section 3.4.2 for predictor data).

Mosquito Alert data

Mosquito Alert report data are updated on a daily basis and is publicly accessible (Mosquito Alert, 2024). The citizen science system Mosquito Alert is an initiative by public research institutions. This project aims to bring scientists, citizens and further stakeholders together to monitor, analyze and control the worldwide spread of *Aedes albopictus* and other *Aedes* species without facing geographical and financial constraints of traditional observation networks (Bartumeus et al., 2018; Eritja et al., 2019, 2021; Palmer et al., 2017). Users can submit pictures of mosquito sightings via the Mosquito Alert application[1] which is available for iOS and Android. Observations are marked with a location and a timestamp and are classified and verified by experts, with a score indicating the confidence with which a report can be assigned to a specific taxonomy (e.g. *Aedes albopictus*). The original data frame covers the years 2021-2024 and contains 218,277 observations from across the globe.

To correct bias inherent to Mosquito Alert data, SE data are provided. SE describes the modeled probability of at least one report being sent on that day within a given cell. Data are collected daily with a sampling grid size of 0.025 degrees (~2.8 km × 1.9 km at the study site). The modeled probability is calculated based on the number of participants in that cell, the duration since they started participating in that project, and the intrinsic participant motivation (Palmer et al., 2017).

Given the nature of citizen science data, the number of observations obviously correlates with the effort of citizen scientists to collect data, as no observations can be done when no one is watching. To show this characteristic in the data, the number of mosquito sightings per cell and day was aggregated and plotted against the average SE for a respective cell and month (see **Figure 1**). A *least-squares fit* was calculated with $R^2 = 0.79$. This means that 79% of the variance in the dependent variable can be explained, indicating a strong relationship between these two variables. It is crucial to consider this data characteristic during model conceptualization to avoid bias propagation from response data to mapped predictions. On the other hand, one can argue that SE is higher in certain areas due to a higher abundance of mosquitoes. Someone experiencing a stronger presence of mosquitoes is probably more likely to use the app compared to one who has never encountered them. When comparing mosquito sightings from Mosquito Alert with those from the nationwide monitoring project using ovitraps run by the AGES (Bakran-Lebl, 2025b), we can see a large agreement for the nationwide distribution patterns of *Aedes albopictus* in Austria, with the largest populations being in the urban area, especially Wien, Graz, and Linz, and barely any sightings outside of cities. The biggest discrepancy can be found at service stops near highways outside the cities where ovitraps show a high abundance of *Aedes albopictus* due to the passive transport of the animal by cars and trucks (Scholte & Schaffner, 2007). These populations cannot be properly monitored using citizen science since users barely spend enough time at these locations. Since this work focuses on the area of Graz, this should not be a significant problem.

[1] www.mosquitoalert.com

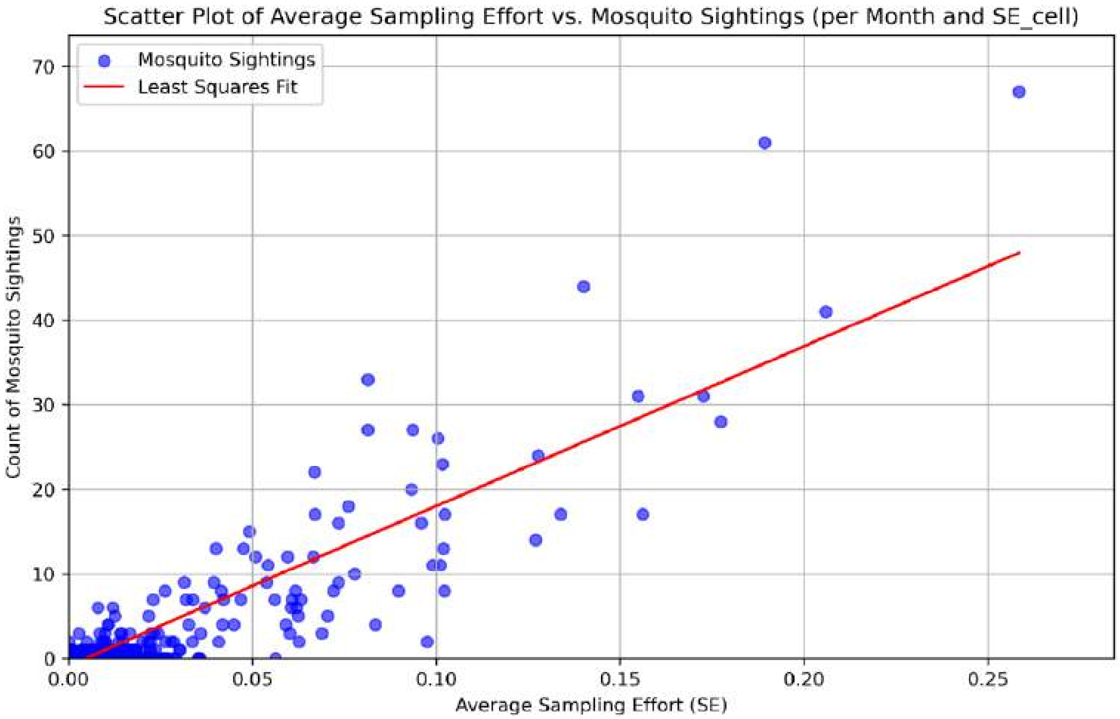

Figure 1: Number of mosquito sightings per month and SE-cell plotted against the average SE value for the respective month and cell (**blue**). A least-squares-fit with $R^2 = 0.79$ indicating a strong relationship between the number of sightings and SE (**red**).

Other data

All data sets used in this work are summarized in **Table 1**. Further details are covered in section 3.4.1 and 3.4.2.

Table 1: Overview of all data sets used in the workflow of this research. Data from the Magistrat Graz, Stadtvermessungsamt is not available publicly and was provided for research purposes. All other data sources are openly accessible.

Data	Type	Date of creation	Description, statistics, additional information	Source
Mosquito Alert observations	Point	Updated daily	Years: 2021-2024 218,277 samples Parameters: Taxonomy, Type of observation, validation score, date, coordinates	https://github.com/MosquitoAlert/Data
Weather data (daily)	Point	Updated daily	**Parameters:** Station, date, T_max (mean = 17.6 °C), T_min (mean = 7.2 °C), T_mean (mean = 12.5 °C), RH_mean (mean = 67.8%), WS_mean (mean = 1.6 m/s), P (mean = 2.3 mm)	https://doi.org/10.60669/gs6w-jd70
Weather data (hourly)	Point	Updated daily	**Parameters:** Station, date, WS_mean (mean = 1.31 m/s)	https://doi.org/10.60669/9bdm-yq93

Sampling effort	CSV	Updated daily	Probability of at least one report being sent from the 0.025° cell during the day. Global data from 2021-2024 with 417,308 samples **Parameters:** Date, cell coordinates, sampling effort	https://github.com/Mosquito-Alert/sampling_effort_data
Land use	Vector	2022	6 classes, 29 categories	Magistrat Graz, Stadtvermessungsamt
Land cover	Raster	2022	1 m resolution, 28 classes	Magistrat Graz, Stadtvermessungsamt
City boundary	Vector	-	City boundary of Graz	Magistrat Graz, Stadtvermessungsamt
Population	Vector	2024	Population data on district level	Magistrat Graz, Stadtvermessungsamt
Sorm drains	Point	1997-2024	171,032 samples	Magistrat Graz, Stadtvermessungsamt
Digital terrain model	Raster	Between 2008-2014	1 m resolution	https://www.landesentwicklung.steiermark.at
OSM Districts	Vector	-	Accessed from overpass-turbo.eu	https://www.openstreetmap.org

3.2 Study area

The study area of this work was limited to the administrative city boundaries of Graz, Austria, given its large population of *Aedes albopictus* (for Austrian standards) within this area (Bakran-Lebl et al., 2025). Furthermore, land cover and population data made available by the city of Graz is constrained to this range.

The area covers 127.6 km^2 and is approximately located between the latitudes 47.00 - 47.15 N and the longitudes 15.35 – 15.55 E. Graz is the capital of the federal state of Styria and the country's second-largest city with a total population of 343,461[2]. The city is situated on the Mur River, surrounded by hills, forests and agricultural farmland. The city's urban landscape is diverse, representing its rich history and development. In the city center of Graz, to the south of the central "Schlossberg", lies the old town characterized by medieval multi-story buildings and pedestrian zones. Surrounding the old town and the "Schlossberg", typical multi-story buildings from the 20th century with their green inner courtyards shape the cityscape, showing the highest population density in Graz. The further one moves away from the city center, the more one- and two-family homes, often featuring private gardens, can be found. Several parks, cemeteries, allotment gardens and street greening across the city and next to the Mur River add vegetation and green spaces to the cityscape.

Graz experiences a temperate climate and a clear annual trend of temperatures and precipitation with a mean of 9.2 °C and 810 mm of precipitation[3]. The highest mean monthly temperatures are in July, averaging 25 °C during the day and 15 °C at night, while minimum temperatures are in January of 1 °C and -5 °C, respectively. While the summer months experience the most precipitation with a monthly average of 95 mm in June, winter represents the dry season with a minimum monthly precipitation of 33 mm in January.

3.3 Software

For data wrangling and preparation of Mosquito Alert data, Python was the primary programming, language while QGIS was used to modify predictors in the form of raster and vector data. Preprocessing

[2] As of 01.01.2015, www.graz.at

[3] www.climate-data.org

of raster data and modeling of mosquito presence, including model fitting, validation, and visualization were performed in R. All software and libraries used in this work are open source.

Python

Python was used to manipulate, clean, and filter Mosquito Alert data, including mosquito sightings and SE data using the libraries *pandas* and *geopandas*. *pandas* is a commonly used library to analyze and manipulate two-dimensional data frames. Similarly, *GeoPandas* extends the functions of *pandas* for handling spatial data and was used to perform spatial analysis and visualization techniques. The library *sklearn* was used during the exploratory data analysis (EDA) to fit a simple linear regression model.

QGIS

The geographical information system (GIS) application QGIS was used for raster calculations and manipulations. Given the broad range of spatial analysis tools and plugins, QGIS was the optimal choice for manipulating, calculating and transforming raster and vector data. Furthermore, the on-the-fly visualization of changes made to data layers allows for instant feedback and adjustments. The *GRASS* plugin was used to perform raster resampling, allowing for extended aggregation methods when resampling raster data to different cell sizes. The *GDAL* plugin was used for terrain analysis to calculate the topographic wetness index (TWI).

R

Data sampling, model building, and model validation were performed in R. The widely used package *dplyr* was used for data wrangling and manipulation, comparable to *pandas* in Python. The *mgcv* package is a powerful tool for fitting and analyzing GAMs. During model validation, the packages *sperrorest* and *pROC* were used for data partitioning for subsequent (spatial) cross-validation and calculations of ROC curves and AUROC, respectively (Brenning, 2012).

AI

OpenAI’s ChatGPT 4o was used to support the writing process of this thesis. It assisted in structuring text sections, reformulated sentences for improved readability. Furthermore, it was used as an assistant in coding by generating individual snippets of codes in R and Python. All outputs were critically revised, adapted and integrated into the entire framework by the author of this work.

DeepL was used to translate the abstract with the results being adapted by the author.

The use of AI tools served as an aid in language improvement and efficiency, without replacing the authors analytical and scientific work.

3.4 Methodology

The general workflow of this work is described in **Figure 2**, with individual steps being explained in detail in sections 3.3.1-3.3.5. Since spatial predictors (land use, land cover, etc.) were assumed to remain constant across the temporal range and, vice versa, dynamic predictors were assumed to be constant spatially, the presence of *Aedes albopictus* in space and time was analyzed by two individual models. Mosquito Alert data are presence-only data from citizen science. Presences were filtered, and pseudo-absences were sampled to compile a binary data set required for the spatial HSM, while daily aggregated mosquito counts were used for the temporal model (section 3.4.1). Two different approaches were tested to reduce biases inherent to presence-only data with the better-performing one being selected. The selection and preparation of predictor data for the respective models are described in section 3.4.2. Spatial habitat suitability was carried out on a grid of 100 m × 100 m using a binomial GAM. Daily mosquito counts were predicted using a negative binomial GAMM (GAM with random effects). As the predicted spatial suitability scores reflect proportional differences, they were directly used to distribute predicted mosquito counts spatially. Consequently, by combining spatial habitat suitability and daily predicted mosquito counts, the presence of *Aedes albopictus* was visualized across space and time (section 3.4.3). Model performance and feature importance were evaluated individually using metrics like AUROC (for binary data), R^2, and RMSE (for count data) (section 3.4.4). Thresholds for habitat suitability and predicted mosquito counts were used to group results into several risk classes. Visualized results as well as an animated time series of spatially distributed predicted mosquito counts from 01.05.2023 to 30.11.2024 were uploaded to a GitHub repository[4].

[4] https://github.com/Digital-Geography/Master_Thesis_Knabe

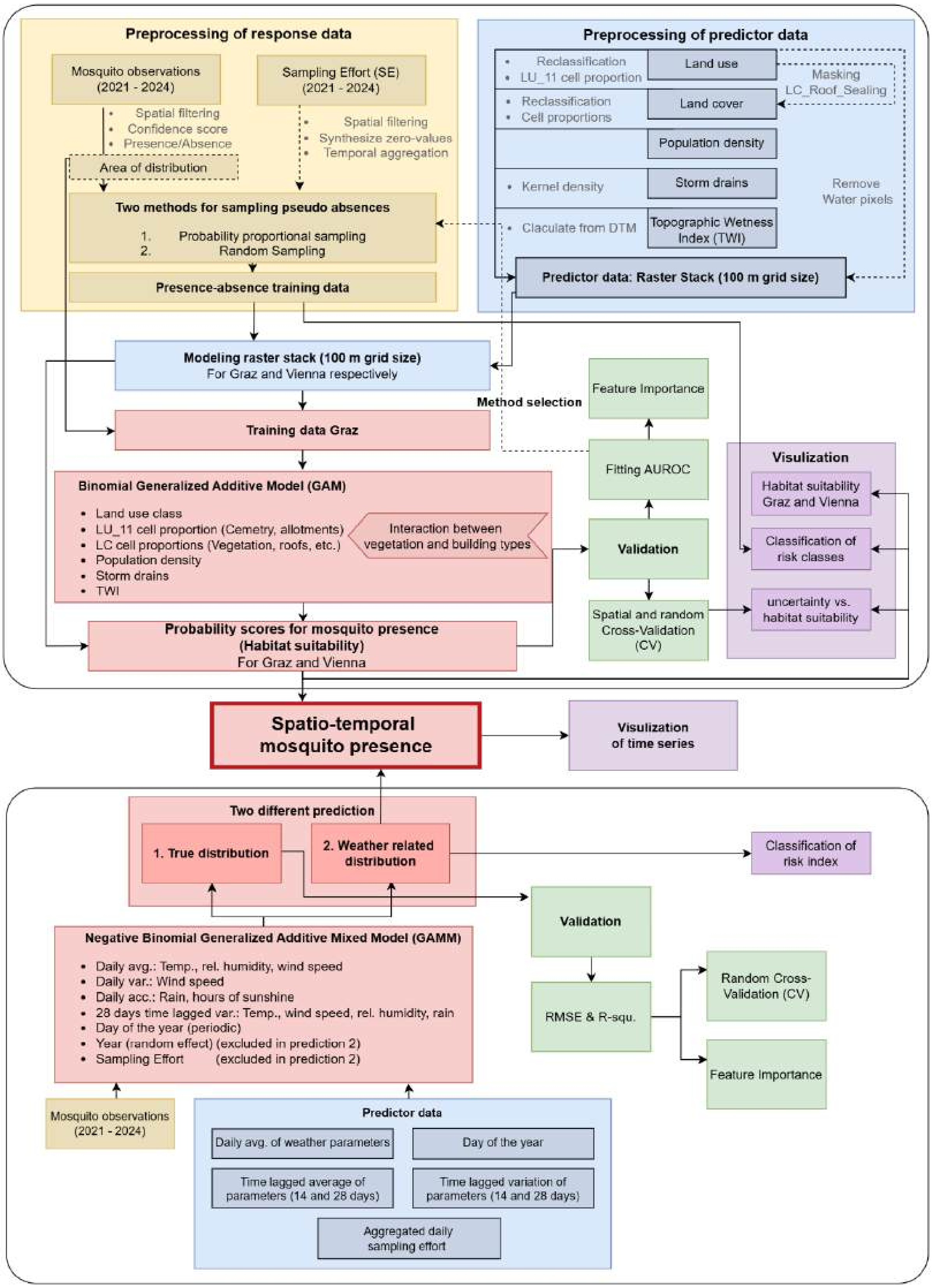
Preprocessing of response data
Mosquito observations (2021 - 2024)
Sampling Effort (SE) (2021 - 2024)
Spatial filtering
Confidence score
Presence/Absence
Spatial filtering
Synthesize zero-values
Temporal aggregation
Area of distribution
Two methods for sampling pseudo absences
1. Probability proportional sampling
2. Random Sampling
Presence-absence training data
Preprocessing of predictor data
Reclassification
LU_11 cell proportion
Land use
Masking LC_Roof_Sealing
Reclassification
Cell proportions
Land cover
Population density
Kernel density
Storm drains
Remove Water pixels
Claculate from DTM
Topographic Wetness Index (TWI)
Predictor data: Raster Stack (100 m grid size)
Modeling raster stack (100 m grid size)
For Graz and Vienna respectively
Method selection
Feature Importance
Training data Graz
Fitting AUROC
Visulization
Habitat suitability Graz and Vienna
Binomial Generalized Additive Model (GAM)
Land use class
LU_11 cell proportion (Cemetry, allotments)
LC cell proportions (Vegetation, roofs, etc.)
Population density
Storm drains
TWI
Interaction between vegetation and building types
Validation
Classification of risk classes
Spatial and random Cross-Validation (CV)
uncertainty vs. habitat suitability
Probability scores for mosquito presence (Habitat suitability)
For Graz and Vienna
Spatio-temporal mosquito presence
Visulization of time series
Two different prediction
1. True distribution
2. Weather related distribution
Classification of risk index
Validation
Negative Binomial Generalized Additive Mixed Model (GAMM)
Daily avg.: Temp., rel. humidity, wind speed
Daily var.: Wind speed
Daily acc.: Rain, hours of sunshine
28 days time lagged var.: Temp., wind speed, rel. humidity, rain
Day of the year (periodic)
Year (random effect) (excluded in prediction 2)
Sampling Effort (excluded in prediction 2)
RMSE & R-squ.
Random Cross-Validation (CV)
Feature Importance
Mosquito observations (2021 - 2024)
Predictor data
Daily avg. of weather parameters
Day of the year
Time lagged average of parameters (14 and 28 days)
Time lagged variation of parameters (14 and 28 days)
Aggregated daily sampling effort

Figure 2: General workflow for the spatio-temporal modeling of the presences of *Aedes albopictus*. Individual steps described in sections 3.4.1 to 3.4.5 include the preprocessing and sampling of response data (**yellow**), preprocessing of predictor data (**blue**), model fitting and prediction (**red**), model validation (**green**) and visualization (**purple**).

3.4.1 Filtering of presences and sampling of pseudo-absences

First, Mosquito Alert presence data were cropped to only contain mosquito sightings between 2021 and 2024 within the city boundaries of Graz. Next, only entries related to the species *Aedes albopictus* were filtered by using the attribute *class_id = 4*. Furthermore, bites and breeding sites were excluded from the data set by selecting *type = adult*. Each report submitted by a user was rated by an expert with a score ranging from -3 to 2, indicating the confidence of species classification. In other words, a -3 means "definitely not a mosquito" while a 2 means "definitely a mosquito". The data set used in this work contained entries with a confidence score of 1 or 2. After preprocessing, the resulting data set contained 1115 entries within the study area (see **Figure 3**), resulting in approx. 0.088 presence records per cell.

SE data were cropped to only contain entries between 2021 and 2024. Data was cropped to a geographical extent of 46.95°-47.15° N and 15.325°-15.55° E (coordinates of lower left cell corners). Given the fact that entries without SE for a given cell and day (e.g. no user in this cell on that day) were not represented in the data set, null entries were generated for each missing entry to create a more comprehensive data set that covers the total spatial and temporal range. For the final spatial model, SE data was aggregated for each cell over the entire time span to obtain the total SE between 2021 and 2024 across the study area. Finally, SE data were transformed into gridded raster data, thereby assigning a respective SE value to each presence/absence cell. Since some cells cover areas from two different SE cells, the average SE value was calculated in proportion to the area covered by each SE cell. In the final step, SE values were normalized to cover a range from 0 to 1, resulting in an average SE value of 0.22 per cell with a standard deviation of 0.27. As expected, SE is low in rural regions with increased values of SE in the central, densely populated regions (see **Figure 3**). The highest values of SE are located to the east of the city center where the university buildings are.

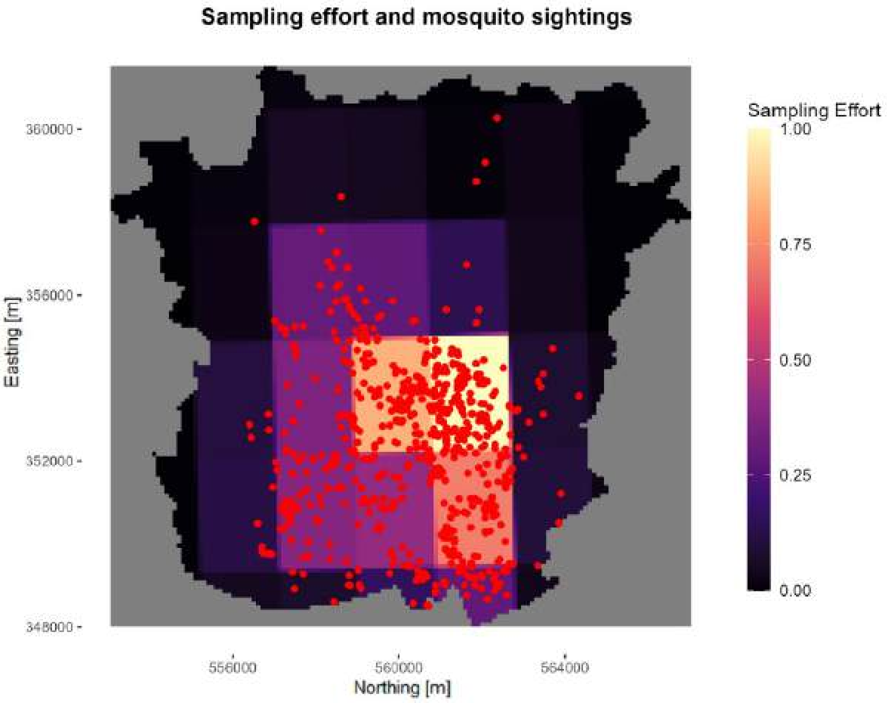

Figure 3: Postprocessed gridded SE data with dark areas indicating low SE and light areas high SE, as well as mosquito presences (**red points**).

Logistic models like GAMs require binary target data for model fitting (Guisan et al., 2002). Therefore, filtered observation data were transformed into a binary grid representing the species presence (at least one observation). In total, the resulting data frame contains 520 presence samples. Another advantage of transforming raw count data to binary presence data is the reduction of spatial autocorrelation. As Mosquito Alert data are obtained by citizen scientists, some locations are over proportionally sampled since some users might be more active than others. As users are the most active in the first days after downloading the app, only a few use the app for a longer period of time (Palmer et al., 2017). When using mosquito counts instead of presence/absence data to train the model, locations with very active users would be overrepresented.

Mosquito Alert data used to fit the statistical model are presence-only data as no reliable absence observations are made. Consequently, artificial absences (known as pseudo-absences or background data) need to be generated to yield further environmental background information. As a result, the model was fitted using data reflecting the full range of environmental parameters instead of only describing areas where the species is present. Determining a suitable sampling design for pseudo-absences requires various considerations based on the given task and data. First, *Aedes albopictus* is a very mobile species, traveling several hundred meters during its lifetime, which exceeds the spatial range of a modeling unit (100 m) (Vavassori et al., 2019). Consequently, there is a substantial chance for the animal to be present in a neighboring cell. Therefore, cells being adjacent to presence cells were removed from the set of potential pseudo-absences as proposed by Barbet-Massin et al. (2012). An adjacency matrix was created for the modeling data frame using the R package *prioritizr* (Hanson et al., 2025). This binary matrix describes neighborhood relations and was used to filter all absence cells adjacent to presence cells from the potential set of pseudo-absences.

As described in section 2.2.2, a fundamental assumption in HSMs is that the modeled species is in equilibrium with its environment and consequently occupies all suitable areas within the observed range. In the scope of this work, this assumption does not hold since the animal just established itself in Graz in the last year with annually growing populations (Bakran-Lebl et al., 2025). When plotting presences based on the year of observation, the animal's spatial dispersion in Graz can be visualized (**Figure 4**). In 2021, the species was first detected by Mosquito Alert with sightings being mainly clustered in an allotment garden in the district "Jakomini". In the next year, the animal spread in the lower half of the study area with a few individual sightings further north. In 2023 and 2024, further dispersal occurred, resulting in a denser distribution across the city center and further spread to all directions. Consequently, sampling pseudo-absences in areas where the species isn't established yet should be avoided since presence data for newly established species can misrepresent the potential distribution and might lead to inaccurate model predictions. Since presence data are missing beyond the range of the species' expansion, sampling of pseudo-absences would be arbitrary and thus lead to erroneous assumptions about the species' absence. Instead, it is crucial to consider areas where the species has had the opportunity to occupy suitable environments, ensuring that the pseudo-absences reflect the current dynamics of the species' range. This careful consideration will improve the reliability of the model. To address this consideration, a convex hull was created around mosquito presences and used to clip the presence/absence data frame, representing the actual area of distribution (see **Figure 4**).

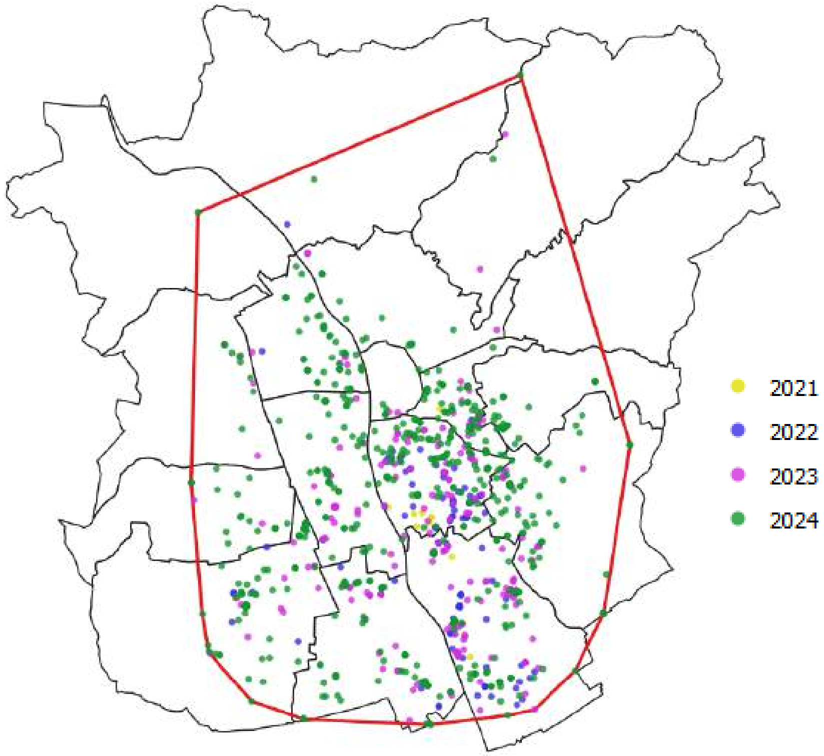

Figure 4: Spatial dispersion of *Aedes albopictus* in Graz. Colors of points indicate the year of mosquito observation. A convex hull is formed around all observations to indicate the current area of distribution used for model training (**red**).

Since Mosquito Alert data are biased due to varying SE, bias can be reduced by applying the same bias to the sampling of pseudo-absences (Phillips et al., 2009). Following this idea, areas with high SE should be overrepresented by pseudo-absences compared to areas with low values. Consequently, Probability Proportional to Size (PPS) sampling was performed on the clipped data frame using default R functions. Aggregated SE values for each absence cell were used as weights during the PPS sampling process. Garamszegi et al. (2024) followed a similar approach when analyzing the spread of *Aedes* species across Hungary using citizen science and field observation data. Instead of weighting during sampling, weights were assigned to each pseudo-absence during model fitting (using Boosted Regression Trees (BRT)) to enhance the impact of pseudo-absences where SE was high. Since the predictive accuracy of GAMs tends to be higher using equal weights for presences and pseudo-absences, the choice was made to reduce bias using PPS sampling instead of applying weights to pseudo-absences during model fitting (Barbet-Massin et al., 2012). A second set of pseudo-absence data was created using Random Sampling (RS) instead of PPS sampling to test a second approach of reducing sampling bias inherent to Mosquito Alert data. This second approach of reducing data inventory bias respected SE during model fitting instead of applying weights during the sampling process (chapter 3.4.3.1).

Barbet-Massin et al. (2012) analyzed the effect of the ratio between presences and absences (known as prevalence) on the accuracy of a model. GAMs were the only method for which prevalence did not have an effect. Hengl et al. (2022) state the rule of thumb that the number of pseudo-absences shouldn't exceed that of presences by 10-20%. Nonetheless, during preliminary testing of this work, two sets of pseudo-absences were tested with one set containing 572 (factor 1.1) and the other 2,600 (factor 5.0) samples. Fitting AUROC was substantially higher for the first set, confirming the rule of thumb by Hengl et al. (2022). Consequently, the following modeling workflow was carried out with the set of 572 pseudo-absences. A comparison of the two sampling methods using 572 pseudo-absences (PPS sampling and random sampling) can be seen in **Figure 5**.

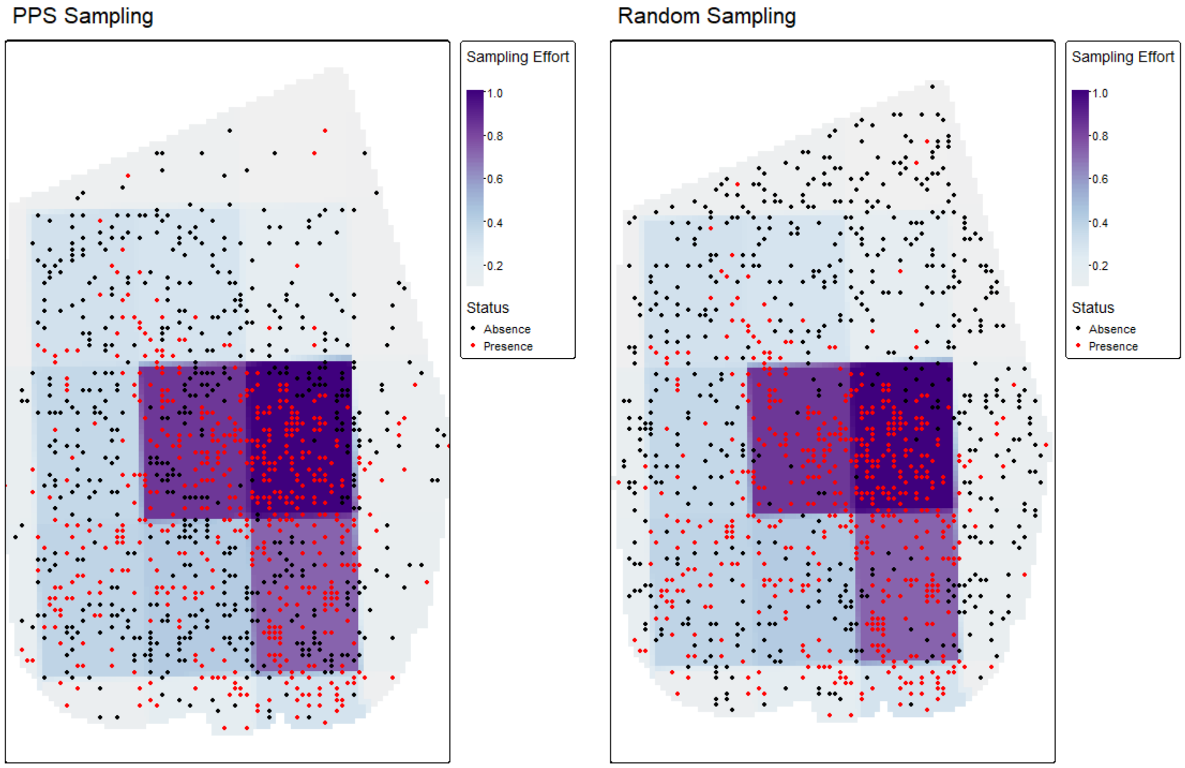

Figure 5: Sampling of pseudo-absences (**black**, n = 572) within the area of distribution of confirmed presences cells (**red**, n = 520) in Graz. PPS sampling (**left**) using aggregated SE values as weights and random sampling (**right**). Light colors indicate low values of SE aggregated over the total time range and dark colors indicate high values.

Since the temporal model used aggregated mosquito count data instead of binary presence/absence data, no pseudo-absence samples were required. For each day between 2022 and 2024, raw mosquito observations and SE values were aggregated for the entire study area (**Figure 6**). The reduction of bias due to SE is covered during model prediction (section 3.4.3.1).

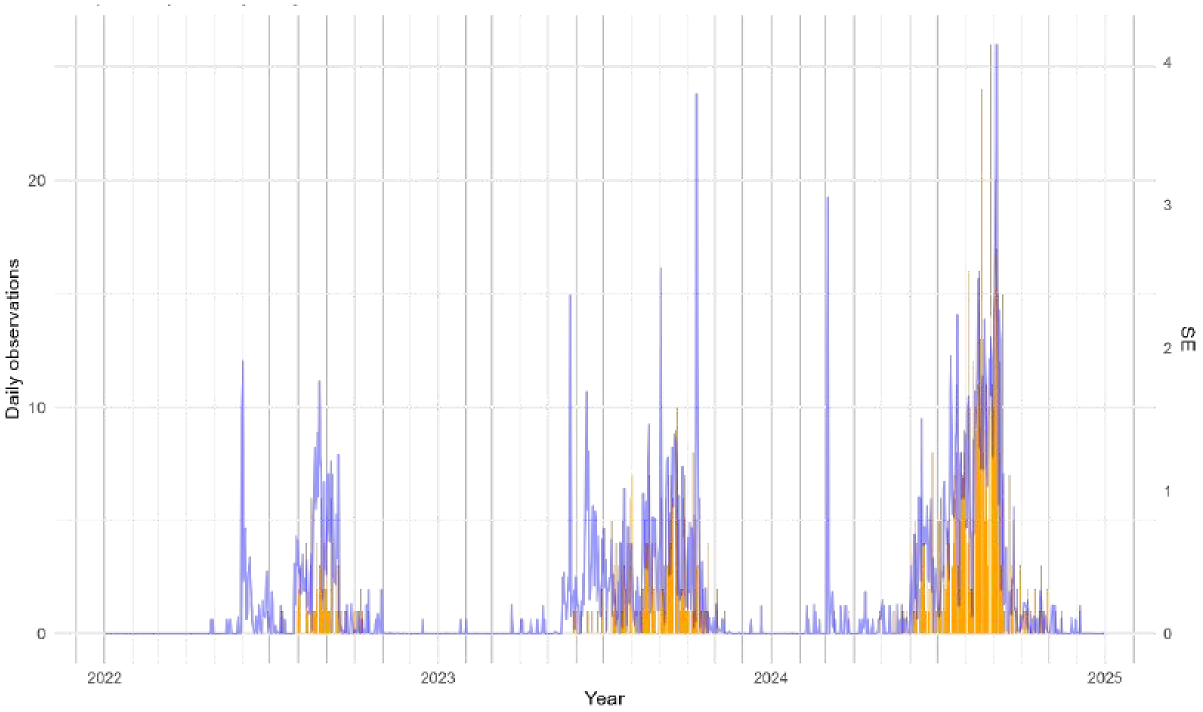

Figure 6: Daily mosquito sightings in Graz (**yellow**) and daily SE (**blue**, secondary y-axis) between 2022 and 2024. Peak mosquito season in late summer (August, September).

3.4.2 Selection and preprocessing of predictor data

A conceptual overview of processes influencing the spatio-temporal presence of *Aedes albopictus* is displayed in **Figure 7**. Spatial predictors (e.g. land cover or land use) might indirectly shape spatial variations in microclimate, host, and breeding site abundance, or the presence of shelter or nutrients (e.g. by vegetation) and therefore influence mosquitoes' ability to reproduce. Since mosquitoes breed in aquatic conditions, time-lagged precipitation prior to sampling dates leads to filling of breeding sites and potentially increasing mosquito presence afterwards while excessive rain might flush eggs and consequently reduce subsequent mosquito counts. Time-lagged temperatures might shape mosquito populations since they influence the growth and survival rates of mosquitoes and their development stages. Daily meteorological conditions are assumed to influence mosquito activity (e.g. reduced activity under windy conditions). In a final step, the habitat suitability model is used to distribute predicted daily mosquito counts spatially to visualize the risk of mosquito presence in space and time. The selection of predictors is based on findings from previous studies and will be explained separately for the spatial model (section 3.4.2.1) and the temporal model (section 3.4.2.2) together with preprocessing of data layers.

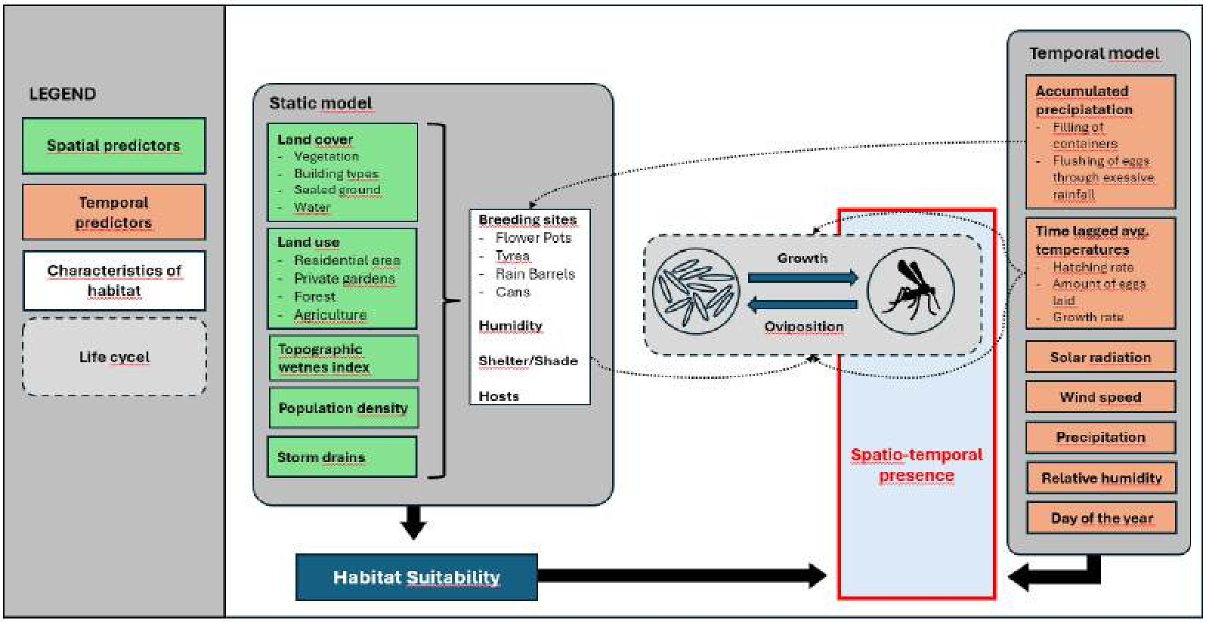

Figure 7: General concept behind the modeling of *Aedes albopictus* in space and time. A static model was trained to predict habitat suitability while a dynamic model predicted seasonal mosquito activity based on meteorological conditions.

3.4.2.1 Spatial predictors

In section 2.1, main drivers influencing the spatial distribution of a species were discussed. Climatic variables are either relevant on a very broad spatial scale like continents or on a very small scale due to microclimate (e.g. shade or sun) (McGill, 2010). Besides numerous observation campaigns, large-scale bioclimatic envelope models have already demonstrated the potential current and future spread across Europe, including Austria (see section 2.3.2). Microclimate variables can be used as predictors to capture spatial variations in mosquito populations (Murdock et al., 2017). Since mosquito presences are not obtained from single point locations (e.g. insect traps) but comprehensively across the whole study area due to the location-independent sampling of Mosquito Alert, variability in microclimate at locations of sampling cannot be reflected with a sufficient spatial resolution. In contrast when sampling mosquito presences with insect traps, microclimate variations could be captured using automatic weather stations located at the respective trap locations. Consequently, in this work, climate variables are not implemented as spatial predictors since variability of microclimate is not quantified for the study area and since differences in remote sensing climate data are neglectable. A set of available predictor variables is chosen to indirectly reflect the heterogeneous characteristics of the study area.

Mosquito distribution patterns are influenced by abiotic factors like shade where the animals can rest (Service, 1971), the availability of larval habitats (Westby et al., 2021) and the microclimate (Murdock et al., 2017). Murdock et al. (2017) showed that land use types, determined by the proportion of impervious surfaces, successfully capture underlying variations in microclimate to explain differences in mosquito population dynamics. Microclimates in an urban environment tend to be warmer and drier due to the *urban heat island* effect and, vice versa, rural environments being characterized by cooler and more humid conditions. Based on the local climate of the study area and the season (see section 3.2 for description of Graz), this can lead to positive or negative effects on the animal's larval survival, development rates and body sizes.

Also, local biotic factors like the availability of blood hosts (Richards et al., 2006), or different vegetation types that offer shade, humidity, protection and fast-decomposing detritus inputs (Murrell et al., 2011), can shape the local habitat suitability (Gardner et al., 2013). Urban areas provide a large number of potential breeding sites as the animal lays its eggs in small stagnant pools of water like in planter dishes, rain gutters, rain barrels, tires or plastic containers from litter (Bartlett-Healy et al., 2012; Dowling et al., 2013). Private gardens, allotment gardens or graveyards are likely to yield suitable conditions for the mosquito as they offer a variety of breeding sites with the additional presence of vegetation (Kuhlisch et al., 2018; Reichl et al., 2023). In general, average temperatures are higher in the city compared to the rural area, potentially increasing the number of mosquito generations and growth rates (Westby et al., 2021).

Data preparation

The upcoming subsections explain the preprocessing of spatial predictor data. Individual resulting raster layers were merged with the raster describing presence/absence and SE (see section 3.4.1) to form one coherent raster stack (100 m grid size) used for modeling (section 3.4.3.1). This multi-band raster was created using the tool *Merge* from the QGIS plugin *GDAL*.

Land use

The original land use data from 2020 in vector format were grouped into six major classes and 29 categories. These six classes were regrouped into nine classes, according to **Table 2**, as some of the land use categories are believed to yield favorable conditions for the mosquito (see **Figure 8**). Allotment gardens and cemeteries (Class 11) are expected to favor the presence of the Asian tiger as there are many potential breeding sites, artificial watering and a lot of vegetation (Kuhlisch et al., 2018; Reichl et al., 2023). Class 50 that previously described all built up areas was divided into three individual classes. Class 51 described homes for one or two families, as well as buildings in open land, as these are mostly accompanied by either a private garden or surrounding meadows. Class 52 described residential multi-story buildings, including the oldtown of Graz. The new class 50 described all other buildings like office buildings or industrial facilities. These classes were split, as they offer a unique set of conditions like the amount of breeding sites, vegetation and availability of blood hosts.

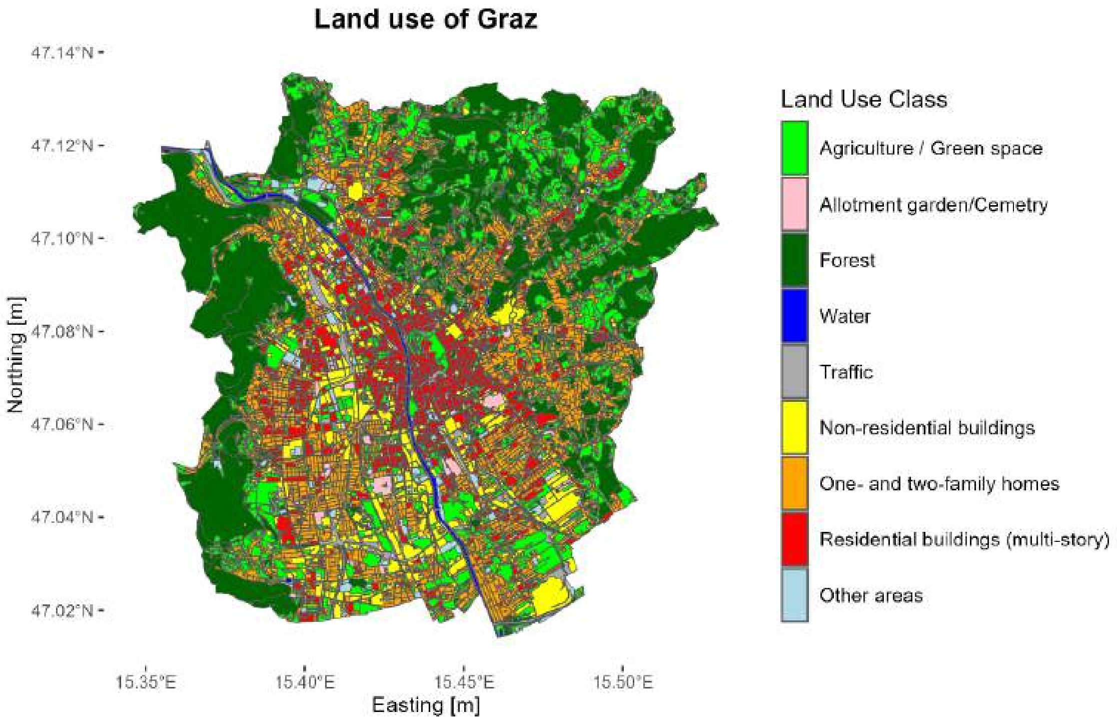

Figure 8: Land use classes in the city of Graz after reclassification.

Using the *Rasterize* tool in QGIS, land use vector data were rasterized with the most dominant land use class being assigned to each grid cell. Pixels representing water areas (class 30) were removed as mosquitoes can't be observed in these areas. After the previous steps, 19.7% of the study area is covered by class 10, 1.0% (class 11), 29.3% (class 20), 0% (class 30), 9.4% (class 40), 9.0% (class 50), 20.1% (class 51), 9.9% (class 52) and 2.3% (class 60). As previously discussed, class 11 was believed to be critical in evaluating the presence of *Aedes albopictus*. However, this class is notably sparse within the study area. When vector data were rasterized, there was a risk of losing detailed spatial information critical for precise analysis. To address this, the proportion of Class 11 was specifically calculated within each pixel. This approach aimed to more accurately capture the influence of this land use class, ensuring that its impact is thoroughly understood in the context of mosquito habitat suitability. This was not done for class 51 and 52 as these land use types will be precisely captured by the land cover data described in the following section.

Table 2: Classes and categories of original land use data used as predictor. Categories in red were regrouped into new classes as they play an important role in predicting the presence of Aedes albopictus.

Class Description	Reclassified	Category
Agriculture / Green space	**10**	Green area
		Arable land
		Garden center
		Fruit growing/plantation
		Orchard meadow

		Vineyard
		Wasteland
		Parc
Allotment garden/Cemetery	**11**	Allotment garden
		Cemetery
Forest	**20**	Forest
Water	**30**	Fluent water
		Still water
Traffic	**40**	Storage area
		Road system
		Rail system
Non-residential buildings	**50**	Business and industry
		Other buildings
One- and two-family homes	**51**	Building area in open land
		One- and two-family homes
Residential buildings (multi-story)	**52**	Multi-story residential buildings
		Buildings from founders' era
		Medieval buildings
Other areas	**60**	Mining area
		Recreation area
		Technical supply and disposal
		Hedge - Alley
		Sports facility
		Other areas

Land cover

While land use data broadly describes the use case of an area, high-resolution (1 m) land cover data (28 classes) can better describe the composition of a space within a land use class. E.g., while land use class 51 coarsely describes an area used by a one- or two-family home, potentially including a green space or a pool, land cover data can precisely describe the amount of green space, vegetation, water, or built-up area within a given pixel. To reduce the number of variables, classes were reclassified according to **Table 3**.

Table 3: Overview of land cover classes after reclassification. Secondary land cover beneath vegetation is written in brackets. Low vegetation (sealed ground), low vegetation (sealed roof), high vegetation (sealed ground) and high vegetation (sealed roof) were assigned to both the vegetation (high and low) and the sealed ground/roof classes. By combining land use and land cover data, *LC_Sealed_Roof* is further split into three classes representing the different types of buildings.

Description	**Reclassified**	**Variable name**
Green Space	Green space	LC_Green_Space
Agriculture	Soil	LC_Soil
Open Soil		
Construction site		
Water	Water	LC_Water
Pool		
Sealed ground	Sealed ground	LC_Sealed_Ground
Sealed ground (low vegetation)		
Sealed ground (high vegetation)		
Sealed roof	Sealed roof	LC_Sealed_Roof (LC_Roof_50, LC_Roof_51, LC_Roof_52)
Sealed roof (low vegetation)		
Sealed roof (high vegetation)		
Low vegetation (green space)	Low vegetation	LC_Vegetation_Low
Low vegetation (agriculture)		
Low vegetation (open soil)		
Low vegetation (water)		
Low vegetation (pool)		

Low vegetation (construction)		
Low vegetation (sealed ground)		
Low vegetation (sealed roof)		
High vegetation (green space)	High vegetation (> 4 m height)	LC_Vegetation_High
High vegetation (agriculture)		
High vegetation (open soil)		
High vegetation (water)		
High vegetation (pool)		
High vegetation (construction)		
High vegetation (sealed ground)		
High vegetation (sealed roof)		

Reclassified classes were transformed into individual binary raster layers (1 m grid size) using the QGIS raster calculator with pixel values representing class membership (0/1). Binary raster layers were resampled to a 100 m grid, with values representing the sum of 1-meter pixels. The resulting land cover raster layers contained values between 0 and 10,000. For clarity, raster values were divided by 100 to represent the percentage of a grid cell that is covered by the respective land cover types (compare **Figure 9**).

Further refinement was applied to the *LC_Sealed_Roof* layer, which includes various building types such as houses, multi-story buildings, allotments, and industrial structures. Before resampling to 100 m grid cells, land use data were employed to create specific subclasses by masking land cover roof sealing data. To avoid any errors during masking, invalid polygons (self-intersecting ring) were identified and repaired. This enabled the creation of three variables:

- **LC_Roof_50:** Created by masking with land use class 50, it represents industrial and other non-residential buildings.
- **LC_Roof_51:** Formed by masking with both land use classes 51 and 11, as these classes represent smaller individual buildings, typically accompanied by gardens.
- **LC_Roof_52:** Derived from land use class 52, this subclass encompasses multi-story buildings. These are either used for residential or commercial purposes.

Lower values of *LC_Roof_51* compared to *LC_Roof_50* and *LC_Roof_52* support the statement that buildings from these classes are often surrounded by green areas or gardens (see **Figure 9**). A very high proportion of roof coverage doesn’t allow for the presence of vegetation as no space is available.

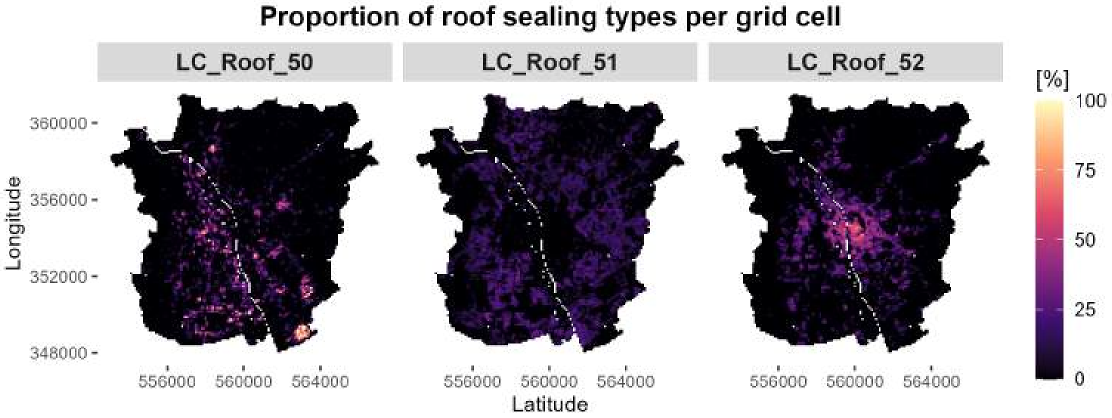

Figure 9: Land cover roof sealing layers representing the proportion per grid cell covered by the three building types. Light colors indicate a dense coverage by buildings. While multi-story buildings are mainly found towards the city center (*LC_Roof_52*), single homes and allotment gardens are further away from the center (*LC_Roof_51*). Other buildings like industrial structures can be found randomly across the city (*LC_Roof_50*).

Population density

Population density for each of the 259 census districts was calculated in *QGIS* by dividing raw population counts by district area. Next, the vector layer is rasterized to match the extent and cell size of other predictor layers. Most densely populated areas of Graz were in the city center around the "Schlossberg" with up to 0.033 residents per square meter (see **Figure 10**).

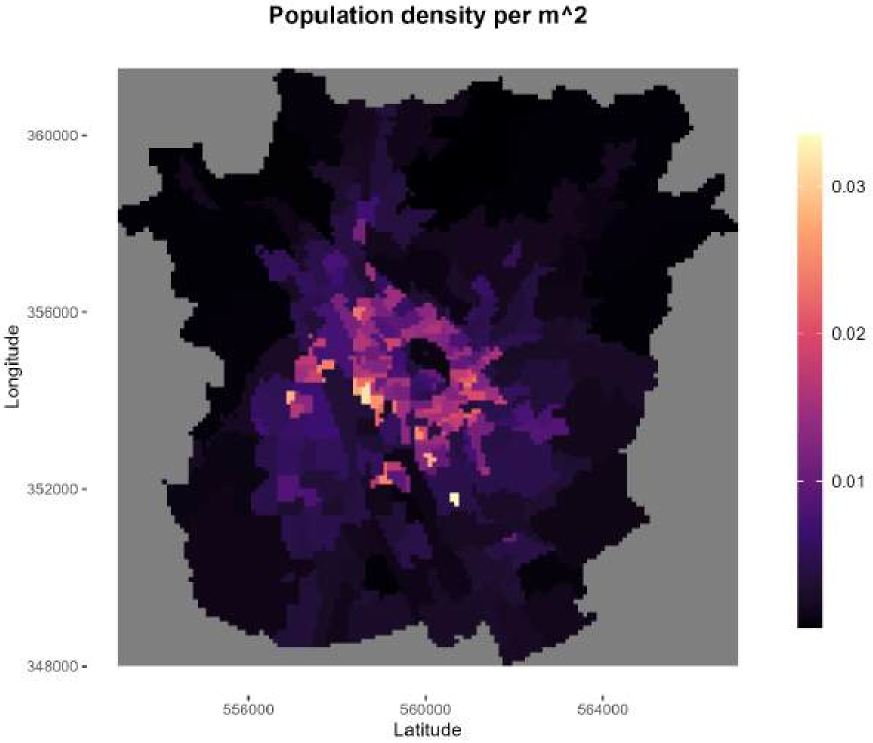

Figure 10: Population density of Graz on census district level, with lighter colors indicating higher values.

Storm drains

Data obtained from the Magistrat Graz, Stadtvermessungsamt, contained a total of 171,032 points. Since the hypothesis is tested that storm drains might act as breeding sites for the mosquitoes due to warm and wet conditions in the canalization, data were filtered to only contain this type of shaft covers (*OBJEKTTYP = "Regeneinlauf"*) resulting in 39,875 features. Other shaft covers or feature types were removed from the data as many of them do not have openings to the underground sewage system.

In the next step, using the QGIS tool *Heatmap*, the point layer was transformed into a raster layer that describes the kernel density of storm drains (see **Figure 11**). In this method, a circular curved surface is fitted over each point with a given kernel radius and a kernel shape. At the point location, the value of the curved surface is highest, diminishing with increased distance and reaching zero at the kernel radius. For each output raster cell, the values of all kernel surfaces where they overlay the pixel center are added. Therefore, the resulting values are unitless but rather represent a relative density.

The kernel radius parameter describes the radial neighborhood around each input point that can be affected by that point. As the Tiger mosquito flies several hundred meters in its life (Vavassori et al., 2019), a kernel radius of 300 m was selected as the presence of a storm drain could potentially influence the presence of a mosquito within that distance. The kernel shape was set to *quartic.* No data values were replaced by 0.

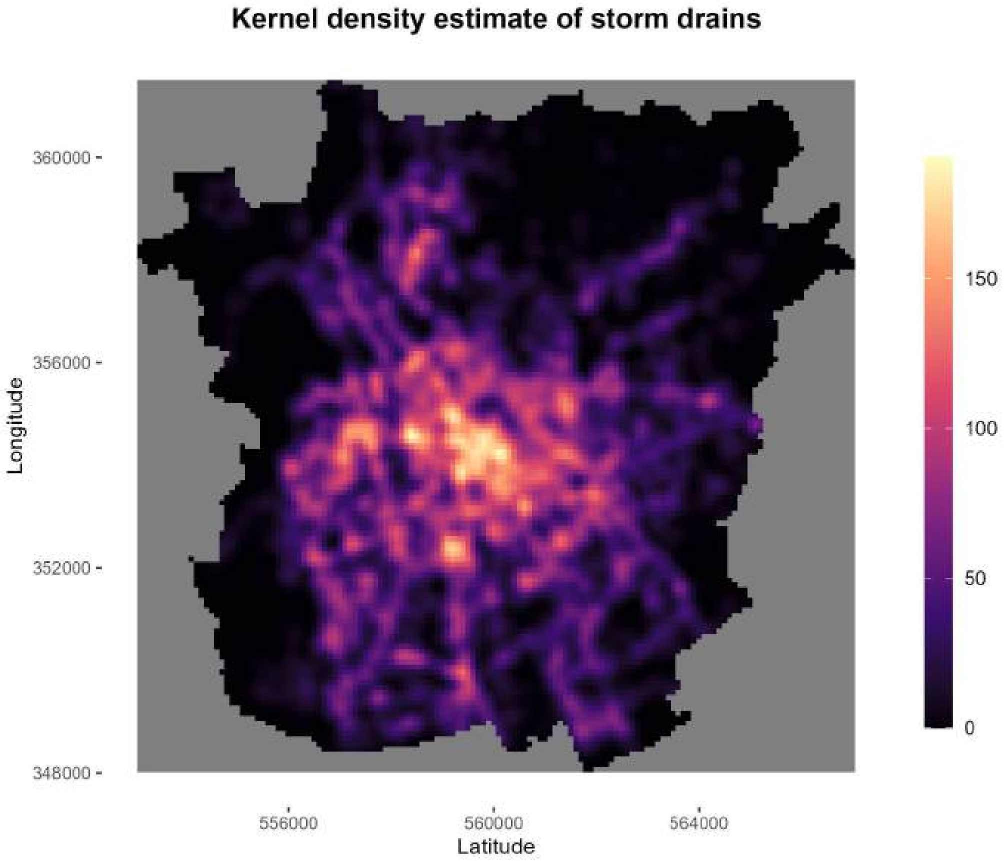

Figure 11: Heatmap of storm drains after kernel density estimation. Values were proportional to the number of storm drains per spatial unit with light colors indicating a higher density of storm drains.

Topographic Wetness Index

The topographic wetness index (TWI) describes spatial variation in soil moisture and the potential for saturation by integrating information about the slope and the contributing upslope area of a landscape. Since moist and saturated areas can increase the relative humidity locally, TWI might help to predict the presence of *Aedes albopictus* as their lifetime increases under high humidity (Cai et al., 2023). A digital terrain model (DTM) with 1 m grid size for Graz was used to calculate the TWI according to

$$TWI = \ln\left(\frac{a}{tan(\beta)}\right) \tag{7}$$

where a describes the local contributing upslope area for each pixel and β the angle of the slope in radians (Sørensen et al., 2006). The calculation of TWI was done in *QGIS*. First, the slope was calculated and values smaller or equal to zero were replaced by one as the natural logarithm is applied later. When calculating the slope, the Zevenberg-Thorne formula was applied instead of Horn's formula due to diverse and complex terrain features like hills and rivers and due to the high spatial resolution of the DTM (Lee & Clarke, 2005). Next, values were transformed to radians. The contributing upslope area was calculated using the tool *Flow accumulation (qm of esp)* from the *SAGA* toolbox. In a last step, TWI was calculated as previously stated. When using grid sizes other than 1 m, the contributing upslope area needs to be multiplied by the area of one pixel as the tool calculates the area in pixels. Finally, the raster was resampled to 100 m to match other predictor layers.

3.4.2.2 Temporal predictors

Daily observed mosquito counts were assumed to be shaped by daily average and time-lagged weather correlates due to several reasons. Higher temperatures accelerate transition rates between various stages of the mosquito's life cycle such as from larva to pupa and from pupa to adult, as well as decreased mortality rates (Tran et al., 2013). Precipitation acts as a trigger for egg hatching but too much precipitation leads to the overflow of containers and the flushing of eggs (Dieng et al., 2012; Soti et al., 2012). Solar radiation was found to be positively correlated with egg counts when temperature wasn't available as predictor (Cianci et al., 2015). Since the temperature variable is given here, solar radiation or hours of sunshine are neglected as predictors given their positive correlation with air temperature (**Figure 12**). Cai et al. (2023) demonstrated that high values of relative humidity are likely to increase the mosquitos lifetime.

Since this model aims to predict the number of observed mosquitoes for each day, daily means of meteorological variables alone would not successfully explain the data. Daily means of wind speeds, precipitation or temperature might explain the daily activity of the animal (e.g. reduced flight activity due to wind, (Service, 1980)), but will fail to explain the meteorological effect on the animals preliminary life cycle. Consequently, time-lagged variables help to predict the abundance of *Aedes albopictus* (Torina et al., 2023). As suggested by Schmalhausen's law on drivers of biological systems, variations of weather corelates might influence mosquito populations stronger than the average of parameters (Poh et al., 2019). Thus, in addition to the mean of time-lagged variables, the coefficient of variation (CV) was calculated for time-lagged variables.

Data preparation

Raw weather data were obtained from GeoSphere Austria collected by two weather stations in Graz (station 16412: Lat. 47.0777, 15.4489; station 16413: Lat. 47.0462, Lon. 15.4102) to analyze the time series of mosquito abundance between 2022 and 2024 (sample size for 2021 was too low). Since no rainfall for a given day is indicated with -1, negative precipitation values were replaced by zeroes. For each day, the average of both weather stations was calculated to represent the average meteorological condition across the study area. Weather data processed in this work contained daily records of *minimum/maximum/average air temperature* (T; mean = 7.2 °C/17.6 °C/12.5 °C), *average relative humidity* (RH; mean = 67.8%), *average wind speed* (WS; mean = 1.6 m/s) and *accumulated precipitation* (P; mean = 2.3 mm). Since wind speeds can vary heavily during the day, the mean alone might fail to properly describe the influence of wind speed on the presence of mosquitoes. For example, two days could have the same average wind speed, but different coefficients of variation might indicate more variable or gusty conditions on one day. Consequently, hourly records of wind speeds were used to calculate the *coefficient of variation of wind speed* (WS_cv) for each day. Since the size of mosquito populations is heavily dependent on seasonality (with peaks in summer), a *day of the year* (doy) variable was created using the R package *lubridate.* Similarly, a *year* attribute, describing the year of observation, was created to account for the nested structure of the data.

To test the effect of time-lagged variables, the 14-day and 28-day averages of all daily means were calculated for the period preceding each day (including the current day) reflecting the mosquito's typical 3 to 4-week life cycle. To determine which of the two time-lags is used for model fitting, the correlation between the observed daily mosquito sightings and the time-lagged variables was analyzed using the R package *Hmisc.* Only days between 01.05 and 31.10 were included in this correlation analysis since this time range describes peak mosquito season. With the exception of relative humidity, every other time-lagged average of 28 days showed a higher correlation with observed mosquito counts compared to the ones for 14 days. Consequently, for reasons of parsimony, 14-day time-lags were not included during model fitting. The same reduction of predictors has been applied to time-lagged and daily means of the three temperature variables *T_min*, *T_max* and *T_mean*. For the time-lagged variable, *T_mean* was found to have the strongest correlation. In contrast, for daily averages, *T_max* was identified as the most highly correlated variable.

The variation of 28 days time-lagged variables was calculated using the coefficient of variation

$$CV = \frac{\sigma}{\mu} \tag{8}$$

with σ being the standard deviation and μ the mean of the respective variable within the given time range (Cook et al., 2014). Therefore, *CV* reflects the variation of a variable with small values of *CV* indicating a narrow distribution and high values of *CV* indicate a wide distribution.

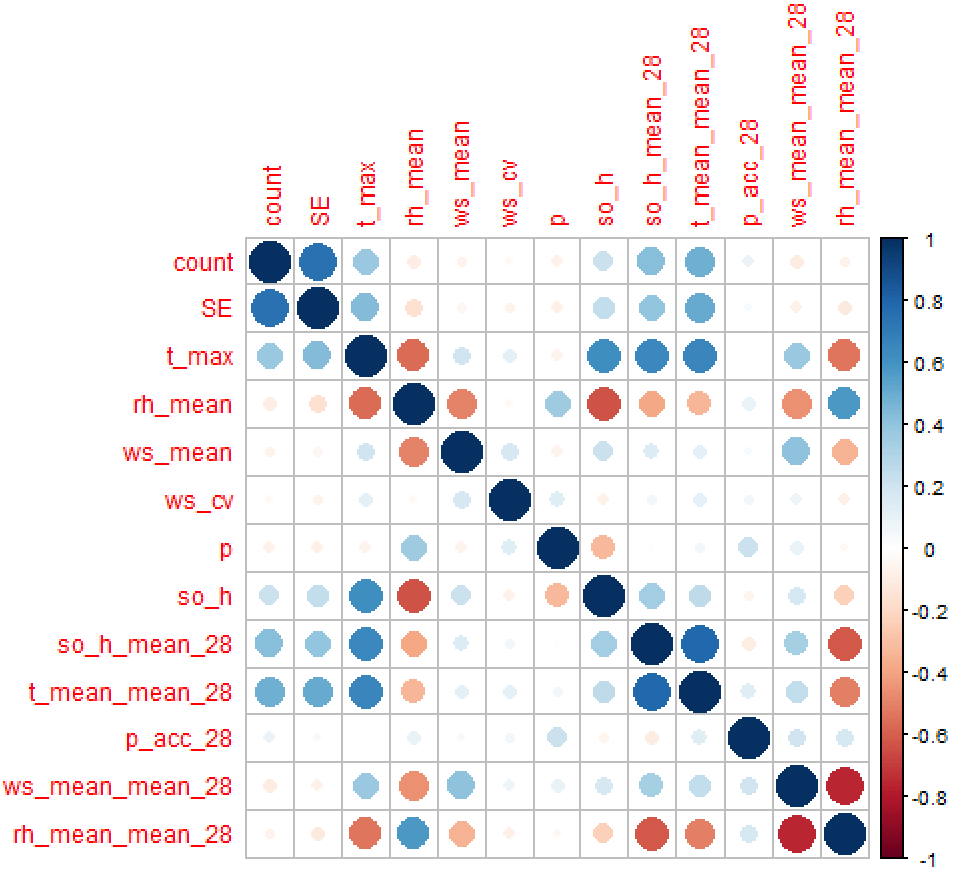

Figure 12: Correlation matrix including daily mosquito counts, aggregated SE, and weather correlates with colors indicating positive (**blue**), negative (**red**) and no correlation (**white**). Correlations were calculated using days of mosquito season (01.05 to 31.10).

3.4.3 Spatial and temporal modelling using GAM/GAMM

3.4.3.1 Spatial model

Spatial habitat suitability prediction was carried out using a GAM for multiple reasons. Various methods allow for the processing of presence-only data including BIOCLIM or ENFA (Brotons et al., 2004). Elith et al. (2006) showed that novel presence-absence methods including GAMs outperform presence-only models. This application of a binary classifier is particularly appropriate because mosquito presence is not random but determined by environmental factors that define certain areas as more suitable than others. By modeling the relationship between habitat variables and presence/absence, one can identify key habitat features that influence mosquito distribution. Consequently, to better interpret relationships between predictors and the response, an approach from the family of explainable models is chosen. GLMs and GAMs have proven to be particularly useful in predicting species distributions while also explaining ecological processes that lead to spatial patterns (Guisan et al., 2002). The analysis and visualization of underlying smoothing functions allow for the assessment of effects of individual predictors on the response variable. In contrast to GLMs, GAMs can handle highly non-linear and non-monotonic relationships between predictor variables and the target variable (Molnar, 2022).

Additionally, GAMs allow to deal with spatially or temporally nested data by including random effects during model fitting (Zuur et al., 2009).

Therefore, a GAM was fitted using species presence-only data from Mosquito Alert to predict habitat suitability with a spatial resolution of 100 m x 100 m using the R package *mgcv*. The relatively small amount of mosquito sightings (1,115 observations) would not allow for a higher resolution as there isn't enough training data to successfully represent suitable habitats at such a small scale. Consequently, further reduction of spatial resolution would lead to an increased fragmentation of suitable habitats into smaller cells. Without sufficient presence data, there is an increased likelihood of potential suitable habitats being erroneously represented by absence locations. To test this hypothesis, during preliminary testing, habitat suitability modeling was carried out for a spatial resolution of 20 m but was not further used in the scope of this work due to very poor modeling results. Additionally, as the mosquito is a very mobile animal (Vavassori et al., 2019), modeling with an even higher resolution than 100 m would be pointless.

Spatial data frames (PPS and RS) of sampled presence/pseudo-absence and SE data (section 3.4.1), including postprocessed predictor data (section 3.4.2.1), were used for model fitting. Each row of the data frame describes one individual cell with a species being absent/present (0/1) as well as associated predictor data. A group of 14 predictor variables was utilized for model fitting, incorporating six additional *tensor product interaction terms* (Pedersen et al., 2019) used to consider potential interactions between the main effects representing class proportions of non-residential building types (LC_Roof_51 and LC_Roof_52) and vegetation types (LC_Vegetation_High, LC_Vegetation_Low, and LC_Green_Space), respectively. Vegetation in residential areas was hypothesized to play a decisive role in predicting habitat suitability of *Aedes albopictus* under the presence of artificial containers, used as breeding sites. Since residential areas offer a variety of these containers (e.g. flowerpots, rain gutters, tires, waste, etc.) as well as large numbers of blood hosts, the assumption is made that the effect of vegetation varies depending on the proportion of developed residential areas. To give an example, a few trees or hedges in the proximity of an apartment complex with a variety of breeding sites and blood host might have a stronger influence on the presence of mosquitoes compared to the same number of trees scattered across a parking lot.

A binomial GAM with logistic link function was trained using the *gam()* function of the package *mgcv* with associated smoothing functions being automatically fitted using internal cross-validation (Wood, 2017) according to

fit = mgcv::gam(fo, data=train, family=binomial, method = "REML", select = TRUE).

An alternative function is the *bam()* function, increasing computational efficiency for large data sets (Pedersen et al., 2019). The formula *fo* defines the effects (fixed and interaction) used to fit the model (see **Table 4**). Default spline types are used (TPRS) to construct smoothers for all predictors except for the categorical predictor *LU_Majority_Class*. Since the latter is a categorical variable (called *factor* in R) it was defined as a factor fixed effect. Each unique category has its own estimated effect (coefficient) on the response variable. Thus, instead of representing smooth response curves like continuous predictors, the factor effect is piecewise constant without smooth transitions in between. The smoothing parameter λ, inherent to each penalty matrix, is estimated using REML (see section 2.4). Automatic variable selection was applied by defining the argument *select = TRUE.* As explained in section 2.4, penalty matrices are used to reduce overfitting by reducing the complexity and wiggliness of basis-functions. These matrices only penalize the range space of basis-functions with the null space (e.g. linear trend in cubic splines) being unaffected. Consequently, if null spaces of smoothers can't be penalized, they can't be zeroed out during variable selection. The automatic variable selection applied uses a double penalty approach by adding a second term to the penalty matrix, effectively zeroing out the null spaces of smoothers. Therefore, unnecessary smoothers can be removed completely, making the resulting model easier to interpret (Marra & Wood, 2011).

Two different approaches were tested to reduce bias inherent to presence data. The first spatial approach (MS1) accounts for data bias during PPS sampling of pseudo-absences (see section 3.4.1). In the second approach (MS2), the relationship between SE and the response variable was determined during model training by including SE as a numerical predictor (using random sampling of pseudo-absences) (see section 3.4.1 for sampling designs). By setting SE values to zero during predictions, the model can separate the true ecological signal (ecological predictors) from the observer signal (SE). The resulting predictions reflect conditions under which SE is equally distributed in space. Using the *gam()* function, this can be done using the *exclude* argument. This approach was applied in fields of species distribution modeling or landslide modeling. For example, Steger et al. (2021) used this approach for landslide susceptibility models where sampling bias was indirectly described by a categorical variable and thus being fitted as a random effect, and finally excluded during prediction. O'Neill et al. (2023) used proxies of SE (i.e. population density, distance to nearest roads or distance to major population center) as covariates during model fitting. For the final predictions, values for respective predictors were set constant across the study area to account for the effect of SE. Since in this study, SE is represented directly by a continuous variable it was included directly using a smooth fixed effect. Including SE during predictions would potentially yield better model performance since data bias would be described by the predictors. Nonetheless, since the aim is to model the actual probability of mosquito presence rather than the probability of a mosquito being observed by a Mosquito Alert user, the smooth effect related to SE should be excluded during predictions.

To predict habitat suitability scores, the fitted GAM was applied to the prediction data frame representing the whole study area (not cropped to area of distribution) to represent potential areas at risk for current and future mosquito presence. The predicted values from the binomial GAM represent the estimated probability that a given cell belongs to the "positive" class (i.e. presence).

Table 4: Set of predictor data used for predicting spatial habitat suitability. Interaction effects will be applied to land cover data of vegetation and roof area. Land use (LU) and land cover data (LC) are used to calculate cell proportions.

Predictor	*Description*	*Type*	*Levels of freedom (k)*
LU_Majority_Class	Majority land use	Fixed effect (Factor)	-
Population_Density	Population density (per m^2)	Fixed effect (Smooth)	3
LU_Allot_Grave	Cell proportion of allotment garden and graveyards	Fixed effect (Smooth)	3
Rain_Inlet	Kernel-density of Storm drains	Fixed effect (Smooth)	3
TWI	Topographic wetness index	Fixed effect (Smooth)	3
LC_Soil	Cell proportion of open soil	Fixed effect (Smooth)	3
LC_Vegetation_Low	Cell proportion of low vegetation	Fixed effect (Smooth)	3
LC_Vegetation_High	Cell proportion of high vegetation	Fixed effect (Smooth)	3
LC_Green_Space	Cell proportion of green space	Fixed effect (Smooth)	3
LC_Water	Cell proportion of water	Fixed effect (Smooth)	3

LC_Sealed_Ground	Cell proportion of sealed ground	Fixed effect (Smooth)	3
LC_Roof_50	Cell proportion of roof area (industrial or non-residential)	Fixed effect (Smooth)	3
LC_Roof_51	Cell proportion of roof area (multistory buildings)	Fixed effect (Smooth)	3
LC_Roof_52	Cell proportion of roof area (one/two-family homes, allotments)	Fixed effect (Smooth)	3
6 Interaction Terms: LC_Roof_51/LC_Vegetation_High LC_Roof_51/LC_Vegetation_Low LC_Roof_51/LC_Green_Space LC_Roof_51/LC_Vegetation_High LC_Roof_52/LC_Vegetation_Low LC_Roof_52/LC_Green_Space	Interaction effect between different vegetation types and residential building types	Interaction effect: Tensor product interaction (ti())	(4, 4)
Sampling_Effort	Temporally aggregated Sampling Effort (only for MS2)	Fixed effect (Smooth) (excluded in prediction)	3

3.4.3.2 Temporal model

To predict daily mosquito counts in Graz, a negative binomial GAMM (GAM including random effects) was fitted using the *gam()* function from the *mgcv* package according to

fit = mgcv::gam(fo, data = train, family = nb(), select = TRUE)

Associated smooth functions were fitted automatically using internal cross-validation (Wood, 2017). When modeling non-negative count data, one can choose between the Poisson family and the negative binomial family. One typically chooses the *Poisson* family when the variance is close to the mean (equidispersion). However, if the variance exceeds the mean one typically chooses the *negative binomial family* (overdispersion). The negative binomial family was applied, as the variance of daily mosquito counts exceeds the mean by a factor of 7 since data are zero-inflated (no mosquito sightings in winter and spring). Three different formulas (**Table 5**) were tested to fit the GAMM with the best performing one being integrated in the spatio-temporal analysis. The first temporal model was trained with all predictors (MT1), the second model ignores variations of time-lagged variables (MT2), and the third one ignores time-lagged averages (MT3). Automatic variable selection was applied using *select = TRUE* (see section 3.4.3.1 for details). Default spline types (TPRS) were used to construct smoothers except for the predictor *doy*. Latter was constructed using *cyclic cubic regression splines* (cc) to account for periodic nature of this variable (start and end point are equal). After comparing parameters *k* and *edf* in the model summary statistics, *k* was set to three for all predictors except for *doy*. Latter was modeled using *k = 12* since there are 12 months per year.

The true distribution of daily mosquito sightings from Mosquito Alert depends not only on the effect of weather or seasonal climate. Since mosquito populations in Graz are growing annually (**Figure 6**, Bakran-Lebl et al. (2025)), weather predictors alone might fail to successfully predict observed mosquito counts. Promotion campaigns for Mosquito Alert like in radio shows or articles in newspapers might lead to peaks of SE (see **Figure 6**). Consequently, the observed daily number of mosquitoes might be heavily biased due to the effort made by citizen scientists to collect the data. To address this bias, the variables *year* (random effect) and *SE* (smooth function) were added to the sets of predictors. Including both in predictions would yield the most accurate model since not only the influence of weather and seasonality is accounted for but also underlying variations due to SE and the total growth of mosquito populations. To assess the raw effect of weather and seasonality, to examine conditions under which peak mosquito activity is expected, the effects *SE* and *year* need to be zeroed out during prediction. This approach follows the same principle as the second approach of reducing sampling bias in the spatial model (see section 3.4.3.1).

Table 5: Three different sets were used to predict daily counts of mosquito sightings in Graz. Model set 1 includes all parameters (MT1), set 2 contains daily parameters and time-lagged (Tl) average/accumulation (MT2) and set 3 contains daily parameters and the variation of time-lagged parameters (MT3). 28-day time-lagged variables are indicated (_28).

Predictor	Type	MT1	MT2	MT3
Doy (bs = cc)	Individual day	✓	✓	✓
T_max	Average	✓	✓	✓
RH_mean	Average	✓	✓	✓
WS_mean	Average	✓	✓	✓
WS_cv	Daily variation	✓	✓	✓
P	Daily accumulation	✓	✓	✓
T_mean_mean_28	Tl-average	✓	✓	
T_mean_cv_28	Tl-variation	✓		✓
RH_mean_28	Tl-average	✓	✓	
RH_cv_28	Tl-variation	✓		✓
WS_mean_28	Tl- average	✓	✓	
WS_cv_28	Tl- variation	✓		✓
P_acc_28	Tl-accumulation	✓	✓	
P_cv_28	Tl-variation	✓		✓
Year	Random effect	✓ / X	✓ / X	✓ / X
Sampling_Effort	Daily sum	✓ / X	✓ / X	✓ / X

3.4.3.3 Spatio-temporal model

While the binomial spatial GAM predicts the probability of the mosquito being present for each cell (values between 0 and 1), the temporal negative binomial GAMM predicts daily mosquito counts within Graz. Predictions from presence/(pseudo-)absence data are frequently used to indicate variations in species abundance (Greaves et al., 2006; Sarà, 2008). Since the aim is not to exactly predict the count of mosquitoes within each cell but rather predict the trend or risk of mosquito presence in space and time, a linear relationship between their probability of spatial presence and their total abundance within each cell was assumed. Therefore, spatio-temporal mosquito presence is determined by: (1) the spatial habitat suitability and (2) the dynamic weather and seasonality effects. Consequently, since predicted habitat suitability scores reflect proportional differences, these scores were used to spatially distribute daily predicted mosquito counts (i.e. a location with a score of 0.8 is expected to host twice as many mosquitoes as one with 0.4) to indicate the risk of mosquito presence. The predicted abundance within each cell and day was calculated according to

$$A_{i,t} = C_t * \frac{P_i}{\sum_j P_j} \qquad (9)$$

The predicted daily count from the temporal model C_t was multiplied by the normalized spatial habitat suitability indices that summarize to 1. Subscripts indicate the current day t, the cell i, and the total number of cells j. To avoid any bias regarding SE and growth of mosquito populations due to non-equilibrium, predictions of C_t were used, where the effects *Year* and *Sampling_Effort* were zeroed out. Therefore, the temporal dynamics of weather and seasonality, expressed by C_t, were projected into space by distributing the total value using spatial scores of habitat suitability. Since habitat suitability scores are normalized, the sum of all cells yielded the total count of mosquitoes within Graz

$$\sum_{i=1}^{j} A_{i,t} = C_t \tag{10}$$

as predicted by the temporal model.

3.4.4 Model validation

3.4.4.1 Spatial model

Model performance for binary response data is typically analyzed using ROC curves and AUROC (or AUC). Calculation and visualization of these metrics were done using the R package *pROC.* The ROC curve is used to show the diagnostics of a binary classifier by plotting the true-positive rate (sensitivity) on the y-axis against the false-positive rate (1 - specificity) on the x-axis. To generate the ROC curve, each unique predicted probability value is used as threshold for classification of presences. Each point of the ROC curve describes the associated true-positive and true-negative rate for the respective threshold (Bradley, 1997).

A diagonal line from the lower left corner to the upper right corner of the diagram indicates random guessing and reflects no discriminative power. The further the ROC curve is above the diagonal line, the higher the predictive capability of the associated model. The overall discriminative capacity of the model can be quantified by calculating the AUROC that measures the area under the curve. The AUROC describes the probability that a randomly selected positive instance is ranked higher than a randomly selected negative one (Hosmer et al., 2013).

Calculating AUROC over the whole study, including areas outside the current area of distribution, would lead to erroneous results. In this situation of non-equilibrium, the classification of presences does not reflect the model's ability to distinguish between presences and absences. Instead, it compares true presences from the current distribution with predicted habitat suitability scores based on the ecological potential for the whole area, leading to inflated false positives or false negatives. Therefore, fitting AUROC was calculated using the data frame used for model fitting (clipped to area of distribution), to evaluate the overall fitting performance of each model. The *sperrorest* package is utilized to partition (five folds; 80% training and 20% testing) fitting data (cropped to area of distribution) into test and training sets for the subsequent spatial cross-validation (Brenning, 2012). Factor partitioning was performed based on five subregions within the study area (see **Figure 13**) to analyze how well the model performs on: unseen data (1) and how model accuracy varies for different spatial characteristics (e.g.

inner city vs. rural) (2). Cross-validation procedure is based on the principle of fitting the model using a large part of the data (i.e. data from subregion 1 to 4) and testing its performance with the remaining data (i.e. data from subregion 5) using the AUROC.

To examine the uncertainty of model predictions within each cell, random cross-validation was performed to partition data into five folds with each fold being used for testing once. This is only done for the better performing model. Since partitioning of data was random (in contrast to factor-based) this process can be repeated multiple times with each repetition yielding different sets of train and test data. Consequently, this process results in up to $N\ (N = n_{folds} * n_{reps})$ predictions for each instance. Using the range of all predictions per cell the standard deviation was calculated to indicate model uncertainty. Since partitioning was random the number of predictions per cell might vary from one to another (max. *N*). To minimize this variation, the process was repeated ten times ($N = 50$).

Modeled relationships between predictors and response were evaluated on two different levels. Permutation-based feature importance was performed using the R package *vip* to assess importance on the inter-variable level (Greenwell & Boehmke, 2020). Partial effect plots were analyzed to examine how changing values of relevant effects influence habitat suitability scores. The core principle of permutation-based feature importance is to compare baseline model performance using all predictors with the performance after a predictor being randomly permuted. In the first step, the fitted model is evaluated on the original data to obtain the baseline AUROC value. Next, model fitting is repeated with values of one predictor being permuted (randomly shuffled) across the data set. By doing this, the relationship between predictor and response is lost. By comparing the AUROC value of the permuted model with the baseline model, the overall importance (difference in AUROC) of the permuted variable can be quantified. This process is repeated until each variable is permuted once. Since permutation of variables was random, each permutation was repeated 30 times to obtain stable importance estimates and associated uncertainty scores.

The statistical significance of predictors (both main and interaction effects) was assessed by approximated p-values, automatically derived from each penalized regression spline. These p-values represent the probability of an effect being observed, assuming that there is no true relationship between the predictor and the response (i.e. under the null hypothesis). For interaction terms, the null hypothesis describes no interaction between associated main effects (combined effect is sum of individual main effects). Consequently, the lower a p-value for a given predictor is, the higher its statistical significance. A common threshold for p-values of 0.05 was chosen to distinguish between statistically significant and non-significant predictors (Wood, 2017).

Partial effect plots were used to visualize how changing predictor values influence the response allowing for the interpretation of underlying relationships. Interaction between the different roof cover and vegetation predictors were visualized using 2D and 3D contour plots (Steger et al., 2023). The change of the response at a given predictor value is represented in log-odds (for binomial GAMs). Visualization of partial effects and interaction effects were performed using the package *mgcViz.* For plotting interaction effects, the *too.far* argument in the vis.gam() function was set to 0.2 to remove parts of the plotted surface that are extrapolated far beyond the range of observed data. Specifically, it excludes grid nodes that are more than 0.2 units away (in scaled space) from any observed data point in the two-dimensional predictor space.

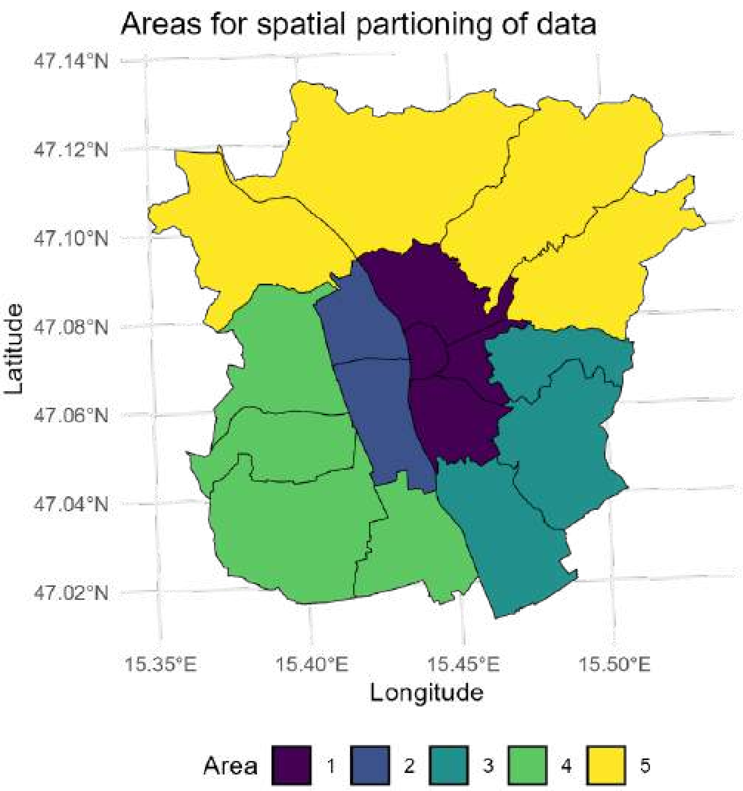

Figure 13: Areas used for spatial portioning of training data to perform spatial cross-validation. Administrative districts of Graz compromising each area: I, II, III, VI (**1**); IV, V (**2**), VII, VIII, IX (**3**); XIV, XV, XVI, XVII (**4**), X, XI, XII, XIII (**5**).

3.4.4.2 Temporal model

Fitting performance of the three models was evaluated using the coefficient of determination R^2, effectively describing the proportion of the variance in the outcome being explained by the model (Gao, 2024). Additionally, in analogy to Torina et al. (2023), the fitted negative binomial models were analyzed using RMSE. It describes the square root of the MSE and is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

describing the average magnitude of errors (Hodson, 2022). Random cross-validation was performed to test the model's capability to handle unseen data using five folds and five repetitions, with the RMSE being calculated for each combination of test and training data. Since the overall accuracy of the three models was compared, the predictors *Year* and *Sampling_Effort* were included during predictions in the cross-validation process. The best set of predictors (MT1, MT2, MT3) was selected by comparing the average R^2 and RMSE after cross-validation. Since a random effect for the *year* variable was included during model fitting, a *leave one year out* partitioning (temporal cross-validation) could not be applied since each factor level (i.e. *year*) must be represented by test and training data.

Permutation-based feature importance was performed using the *vip* package by analyzing the difference in RMSE (Greenwell & Boehmke, 2020). Since the aim is to assess the effect of weather correlates on the response, model prediction during feature importance estimates was performed by zeroing out *Year* and *Sampling_Effort*. Since shuffling of permuted variables is random, individual importance estimates were analyzed using 30 simulations.

Like for the spatial model, partial effect plots were used to visualize how changing predictor values influence the response allowing for the interpretation of underlying relationships. The change in the response at a given predictor value is expressed in terms of log-expected counts, as the model uses a log link function with a negative binomial distribution. Visualization of partial effects was performed using the package *mgcViz.*

3.4.5 Visualization and classification

3.4.5.1 Spatial model

To convert continuous model outputs into interpretable habitat suitability classes, a classification scheme was applied based on the distribution of predicted probabilities at known mosquito presence locations. Specifically, predicted probabilities were extracted for all pixels with observed *Aedes albopictus* presence. These values were then sorted in descending order and used to calculate percentiles. Three thresholds were derived: the 0th percentile (maximum suitability at a presence point), the 25th percentile (medium suitability), and the 95th percentile (least 5% of suitability values among observed presences) (see **Figure 14a**). These cut-off points were then used to classify the prediction into *low*, *moderate*, and *high* suitability classes. A similar percentile-based classification method has been employed by Kraemer et al. (2019) in their global mapping of *Aedes* vectors. In their study, thresholds were derived from predicted environmental suitability values corresponding to specific quantiles of occurrence data, allowing for the definition of ecologically realistic risk zones while accounting for potential observational uncertainty.

Model uncertainty values for each cell, described by the standard deviation of predicted probabilities after random cross-validation, were transformed into three quantile-based bins (low, medium and high). One common way to visualize uncertainty in spatial data is the use of bivariate choropleth maps (Lucchesi & Wikle, 2017). Therefore, a bivariate choropleth map with 3x3 classes was created to visualize the risk of mosquito presence with associated model uncertainties.

3.4.5.2 Temporal model

The influence of recent meteorological conditions on the daily presence of *Aedes albopictus* was reflected by the predicted number of mosquitoes per day, with *Sampling_Effort* and *Year* being zeroed out during prediction. These daily predictions reflect the relative temporal risk based solely on weather-driven variation and the day of the year.

In analogy to the classification scheme used for spatial habitat suitability, three temporal risk categories were defined. The predicted values of daily mosquito counts associated with presence records were ranked in descending order. From this distribution, three quantile-based thresholds were extracted: the

0th (minimum), 25th, and 95th percentiles (see **Figure 14b**). These thresholds were then used to classify the temporal risk of mosquito occurrence into three ordinal categories. Consequently, a given day is categorized as *high risk* if the predicted number of mosquitoes is equal to or greater than the 25th percentile of predictions associated with true presences. Days with predictions falling between the 25th and the 95th percentile were assigned to the class of *medium risk and* days with predictions lower than the 95th percentile were defined as *low risk* days.

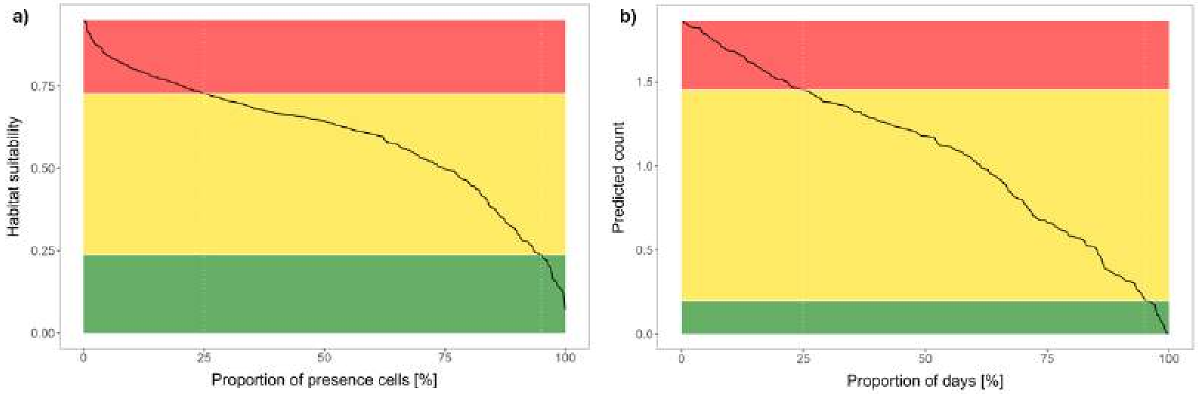

Figure 14: Definition of cut-off points for classification of risk classes for spatial habitat suitability (**a**): The black line indicates the percentage of true presence cells with an associated predicted habitat suitability score smaller than the associated score on the y-axis. The 25th and the 95th percentiles were used as cut-off points to define the classes *low risk* (**green**), *medium risk* (**orange**) and *high risk* (**red**). Classification thresholds for the temporal predictions are based on recent weather conditions and the day of the year (**b**): The black line indicates the percentage of true presence days with an associated predicted daily mosquito count smaller than the associated value on the y-axis.

Chapter 4 Results

4.1 Spatial model

4.1.1 Performance of MS1and MS2

Two different models were fitted using two different sets of presence/pseudo-absence data. MS1 accounted for sampling bias by applying PPS sampling of pseudo-absences based on SE values. MS2 employed a fixed effect for SE on a set of randomly sampled pseudo-absences. Fitting AUROC of 0.82 for MS2 was slightly higher than the one for MS1 (AUROC = 0.80). Generally, an AUROC of 0.5 indicates that model predictions are equally effective as random guessing. According to Swets (1988), values of 0.5 < AUROC < 0.6 are interpreted as "fail", 0.6 < AUROC < 0.7 as "poor", 0.7 < AUROC < 0.8 as "fair", 0.8 < AUROC < 0.9 as "good" and AUROC > 0.9 as "excellent". Since AUROC calculations of MS1 and MS2 are based on different data sets, it is difficult to compare model performance. This is especially the case since pseudo-absences of MS1 were more dominant in urban areas due to high values of SE while pseudo-absences of MS2 were randomly distributed (**Figure 16**). Consequently, pseudo-absences of MS2 showed an increased distribution across land use class 10 (green land, agriculture) and 20 (forest), while the ones from MS1 showed higher numbers in urban land use class 11 (allotment, graveyard), 40 (traffic), 50 (non-residential), 52 (residential multi-story buildings) and 60 (other areas) (**Figure 15**). Since mosquito presences were predominantly concentrated in urban areas, pseudo-absences of MS2 covered more areas where mosquito presence is unlikely. Therefore, due to comparably low numbers of presences in forests, meadows and acres (class 10 and 20), rural areas are assumed to be easily classifiable as unsuitable areas, typically known as *trivial areas*. Steger & Glade (2017) showed that overrepresentation of trivial areas in landslide modeling, characterized by flat areas, led to increased AUROC values. Similarly, the ability of MS2 to distinguish between presences and absences was enhanced by such distributions, resulting in higher AUROC scores compared to MS1. To further visualize this characteristic, spatial cross-validation was performed to evaluate the model's capability of classifying unseen data and to analyze spatial differences in model accuracy. **Figure 16** shows individual AUROC values for each group of neighborhoods used for testing. Again, both models showed similar AUROCs with overall values of 0.72 (MS1) and 0.73 (MS2). Both models showed the best accuracy in cluster 3 (purple). MS2 slightly outperformed MS1 in cluster 1 (green), 2 (red) and 5 (blue), while MS1 showed better results in cluster 4 (yellow). This overperformance was expected especially for cluster 5 due to increased numbers of pseudo-absences in trivial terrain.

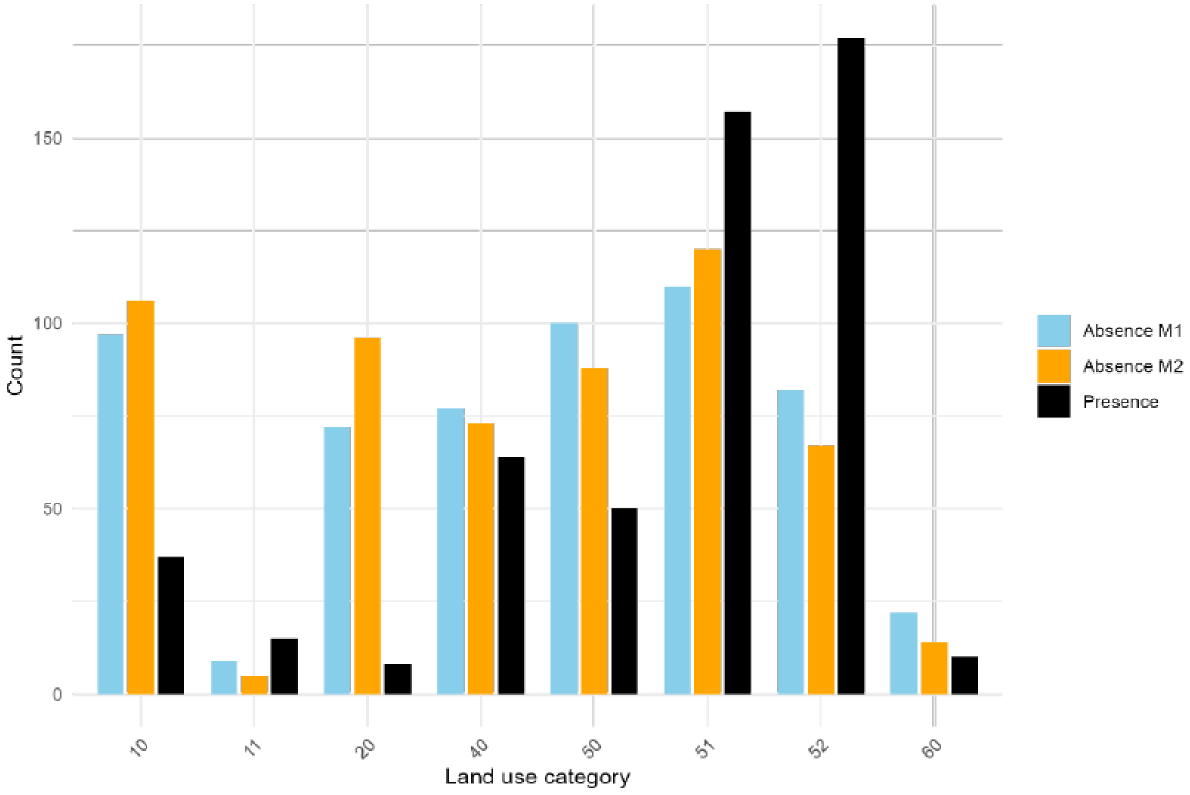

Figure 15: Land use class distributions for presences (**black**), pseudo-absences (PPS sampling, **blue**) and pseudo-absences (random sampling, **orange**).

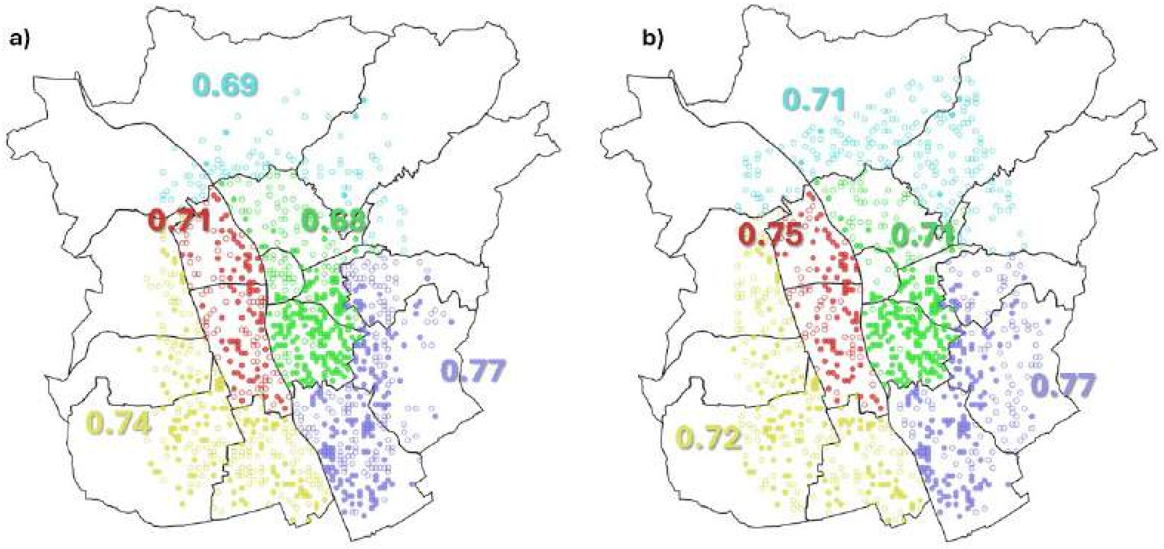

Figure 16: Mean AUROC results based on spatial partitioning of data of MS1 (**a**) and MS2 (**b**) into test and training sets. Filled circles depict presence and empty circles absence locations of *Aedes albopictus.* Black lines indicate district boarders within Graz.

To compare the two approaches, fitting AUROC was calculated on the complete unsampled data frame (including all potential absences within the whole study area). AUROC of 0.85 (MS1) and 0.84 (MS2) indicated "good" model performance with non-significant differences in model accuracy. Interpretation of AUROC values must be treated carefully given the geographical extent of the area to which the model is applied. As previously discussed, including larger proportions of trivial terrain might increase AUROC values (Lobo et al., 2008; Steger & Glade, 2017). Since large parts of the study area, surrounding the urban area of Graz, are covered by forests, fields and meadows, model accuracy might be inflated. To reduce the effect of trivial terrain, AUROC was additionally calculated only for the area of distribution (convex hull) resulting in scores of 0.77 (MS1) and 0.76 (MS2) indicating "fair" predictive accuracy.

According to Lobo et al. (2008), AUROC primarily measures a model's ability to discriminate between presences and absences but does not necessarily reflect the quality of predicted probabilities and how well these correspond to associated habitats and ecological relevance. Steger et al. (2016) support that statement by analyzing how quantitative validation relates to geomorphic plausibility of predicted landslide suitability maps. He demonstrated that models with similar AUROC scores might yield discrepancies in how well they align with the actual distribution of landslide events. Therefore, the AUROC does not always reveal deeper issues like propagation of biases inherent to incomplete data inventories. Consequently, MS1 cannot be prioritized over MS2 solely due to the marginal difference in AUROC. Despite similar AUROC values, there was a substantial difference in prediction patterns across the study area (see **Figure 17**). MS1 generally tended to predict the highest probabilities of mosquito presence within the inner city, characterized by multi-story residential buildings. The medieval old town in the center depicted an exception, potentially due to missing vegetation. Slightly lower (but still high) probabilities were predicted in residential areas further outside of the city center, typically characterized by one- and two-family homes with garden areas. Cemeteries and allotment gardens were not identified as high-risk areas. In contrast, MS2 rated the suburban areas with one- and two-family homes with garden areas (southern part of study area) higher than the inner city of Graz. Furthermore, the highest probabilities were predicted for allotment gardens and cemeteries. Another major difference to MS1 depicted the medieval old town in the city center where street greening is missing. While MS2 identified the old town as an area at risk for mosquito presence, MS1 did not. Both models predicted low probability for commercial, industrial and traffic areas (e.g., southeast of Graz or in the center to the West of the river Mur). The lowest probabilities were predicted for rural areas characterized by grassland, agriculture and forests.

To select the model that better reflects the true distribution of the response, one might relate to the concept of *biological plausibility* typically referred to in fields of epidemiology and medicine (Hoffer, 2003). Therefore, the model that better reflects biological knowledge about the relationship between *Aedes albopictus* and its environment was selected. Since *Aedes albopictus* is classified as a *container breeding species* it is expected that their habitat preference strongly correlates with the abundance of natural or artificial small containers. According to the Austrian institute for health and nutrition safety (AGES), private properties with gardens are potential areas at risk since there are numerous breeding sites like plant dishes, rain barrels, rain gauges or tree holes (Bakran-Lebl, 2025a). Additionally, in residential areas with green spaces, artificial watering might lead to filling of containers during absence of precipitation. Furthermore, vegetation plays an important role in offering protection, shade, nutrients and humidity (Murrell et al., 2011). These habitat characteristics are especially suitable for cemeteries and allotment gardens, typically found across Austrian cities. This statement was proven by ovitrap data obtained from AGES in nine different locations across Graz (Bakran-Lebl et al., 2025). Most *Aedes albopictus* eggs were caught at trap location GFP, located next to the cemetery in the neighborhood "St. Peter", with 1.73 eggs per day. Second most observations were done at location GFL (1.39 eggs per day) next to the allotment garden "Blumenfreunde" in "Lend", followed by trap GLB (1.25 eggs per day). Latter is in the neighborhood "Liebenau" next to several single-family homes with large garden areas. In contrast, trap location AGG on the campus of University of Graz did not indicate much mosquito activity with only 0.11 eggs per day. This location reflects the typical urban characteristics found in the inner city with larger multi-story buildings and green areas in between. Also, in "Baden-Württemberg" in southwestern Germany, allotment gardens and urban residential areas with garden areas were identified as areas at risk for the presence of the Asian tiger mosquito[5].

Since MS2 identified allotment gardens, cemeteries and residential areas with garden areas as most probable habitats, the model is assumed to be more plausible in terms of ecological and biological knowledge about the species' habitat characteristics, even though it yielded slightly worse AUROC values. High probabilities predicted in the inner city by MS1 are not supported by ovitrap data and expert knowledge since potential breeding sites are expected to be less compared to previously mentioned sites. Therefore, only MS2 was further analyzed and implemented in spatio-temporal predictions.

[5] https://www.gesundheitsamt-bw.de/lga/de/kompetenzzentren-netzwerke/arbo-baden-wuerttemberg/verbreitung-von-tigermuecken/

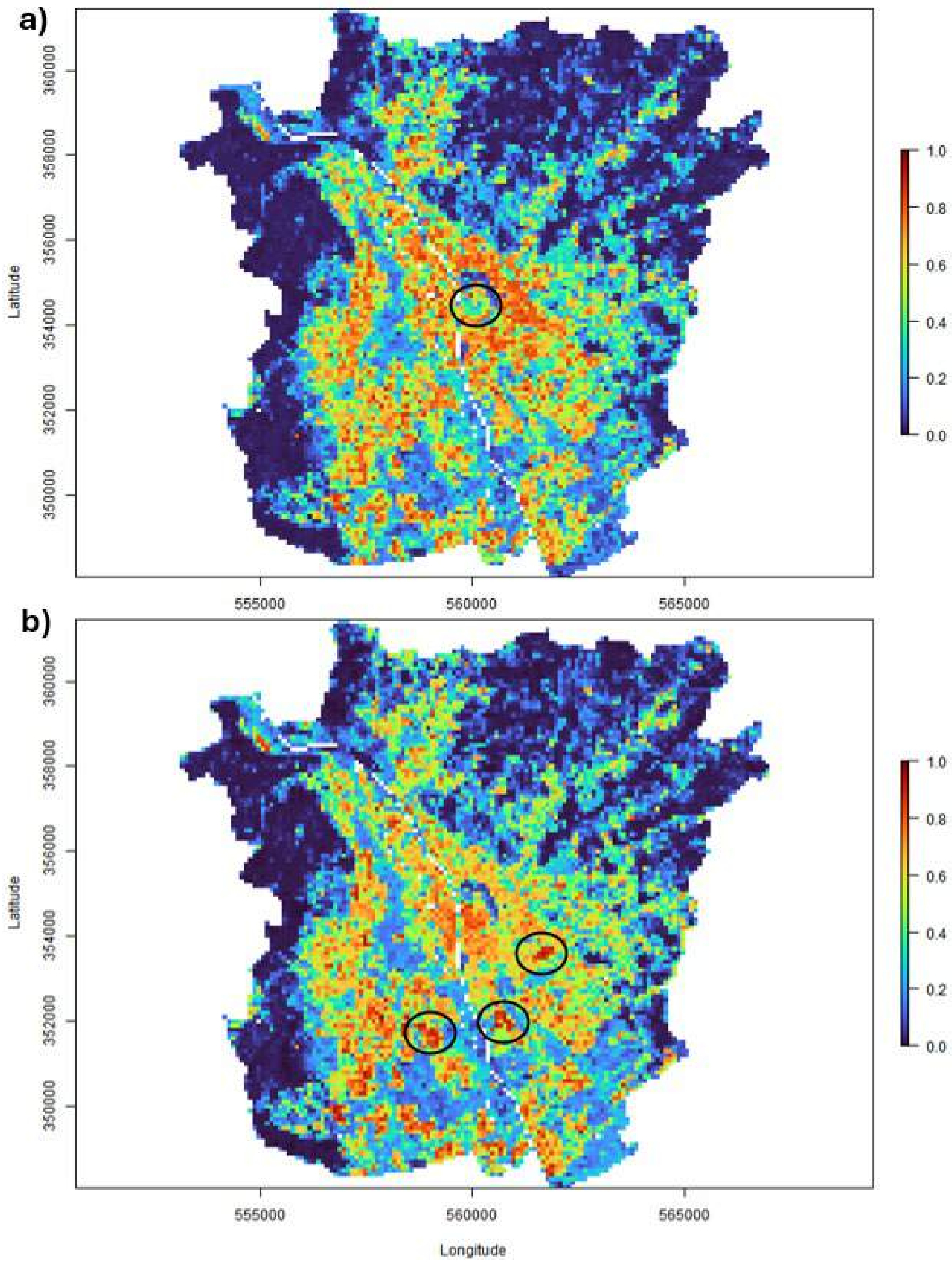

Figure 17: Predicted habitat suitability from MS1 (**a**) and MS2 (**b**) ranging from 0 (**blue**) to 1 (**red**). Major differences between the two approaches are in the old town (**circle, a**) and the surrounding inner city, as well as allotment gardens and cemeteries (**circles, b**). In contrast to MS1, MS2 tended to predict higher probabilities for one- and two-family homes with garden area (southern part of Graz) compared to the inner city.

4.1.2 Feature importance and partial effects

To better understand predicted patterns of habitat suitability, one needs to examine the global feature importance, as well as the individual partial effect plots of smooth functions and interaction terms. For MS1, *LC_Roof_52* was most influential for predicting the presence of *Aedes albopictus*, followed by *LC_Roof_51* and *LC_Vegetation_High* (see **Figure 18**). Automatic variable selection identified and removed *LC_Sealed_Ground, LC_Soil, Population_Density, LC_Water,* and *Rain_Inlet* since they did not explain variability in the response, leading to a feature importance score of zero for those predictors.

For MS2, the integrated fixed effect for *Sampling_Effort* resulted in the highest feature importance highlighting the strong relationship between observed mosquito presence (response) and SE, confirming the bias inherent to training data. With SE being integrated into model fitting, the importance of *LC_Roof_52* decreased substantially by more than 50%, even below LC_Roof_51 followed by *LC_Vegetation_High* and *LU_Allot_Grave*. This difference in feature importance explains why MS1 predicted higher probabilities for the inner city while MS2 predicted higher probabilities for residential areas in the suburbs (described by *LC_Roof_51*). *TWI, LC_Vegetation_Low, LU_Majority_Class* and *LC_Green_Space* only yielded a marginal contribution to predicting the probability of mosquito presence for MS1 and MS2. For MS2, *LC_Roof_50, LC_Sealed_Ground, LC_Soil* did not contribute to the explanation of response data.

Only predictors of MS2 were further analyzed, since MS2 was selected as the better performing model due to biological plausibility. *Sampling_Effort* (p-value < 0.001)*, LC_Roof_51* (p-value < 0.001)*, LC_Roof_52* (p-value < 0.001)*, LU_Allot_Grave* (p-value = 0.001)*, TWI* (p-value = 0.004) and *LC_Vegetation_High* (p-value = 0.028) were identified as significant main effects (p-value < 0.05). Non-significant main effects were *Rain_Inlet* (p-value = 0.060)*, Population_Density* (p-value = 0.101)*, LC_Water* (p-value = 0.109)*, LC_Vegetation_Low* (p-value = 0.300)*, LC_Soil* (p-value = 0.302)*, LC_Roof_50* (p-value = 0.480)*, LC_Sealed_Ground* (p-value = 0.547) and *LC_Green_Space* (p-value = 0.737). Regarding interaction terms, *ti(LC_Roof_51, LC_Vegetation_High)* (p-value = 0.015)*, ti(LC_Roof_51, LC_Green_Space)* (p-value = 0.048) and *ti(LC_Roof_52, LC_Green_Space)* (p-value = 0.043) were identified as statistically significant, indicating a significant interaction between associated main effects. *ti(LC_Roof_51, LC_Vegetation_Low)* (p-value = 0.109), *ti(LC_Roof_52, LC_Vegetation_High)* (p-value = 0.077) and *ti(LC_Roof_52, LC_Vegetation_Low)* (p-value = 0.855) were classified as non-significant.

Partial effect plots describe the estimated relationship between individual predictors and the response after model fitting. Since the model used the logit link function to predict binomial data, values on the y-axis describe log-odds which can be transformed into probabilities using the inverse link function. The predictor with the strongest relationship to mosquito presence was *LU_Allot_Grave* with a log-odds value close to 6 (probability > 99%) at 100% cell coverage. The associated smooth function revealed a linear trend indicating that the probability of mosquito presence increases as the coverage by allotment gardens or cemeteries rises within a cell. The smooth function for *LC_Roof_51* showed a similar trend, featuring a more gradual slope that resulted in smaller effects at higher values. In contrast, *LC_Roof_52* followed a non-linear trend with a maximum around 38% coverage followed by decreasing log-odds for higher predictor values. The associated smooth function for the fixed effect *LC_Vegetation_High* revealed a non-linear negative trend with high proportions of coverage being associated with low probabilities for mosquito presence. Bias inherent in the presence data inventory was further visualized by the smooth function of *Sampling_Effort* indicating higher probabilities for high values of SE. The smooth function of *TWI* followed a linear negative trend.

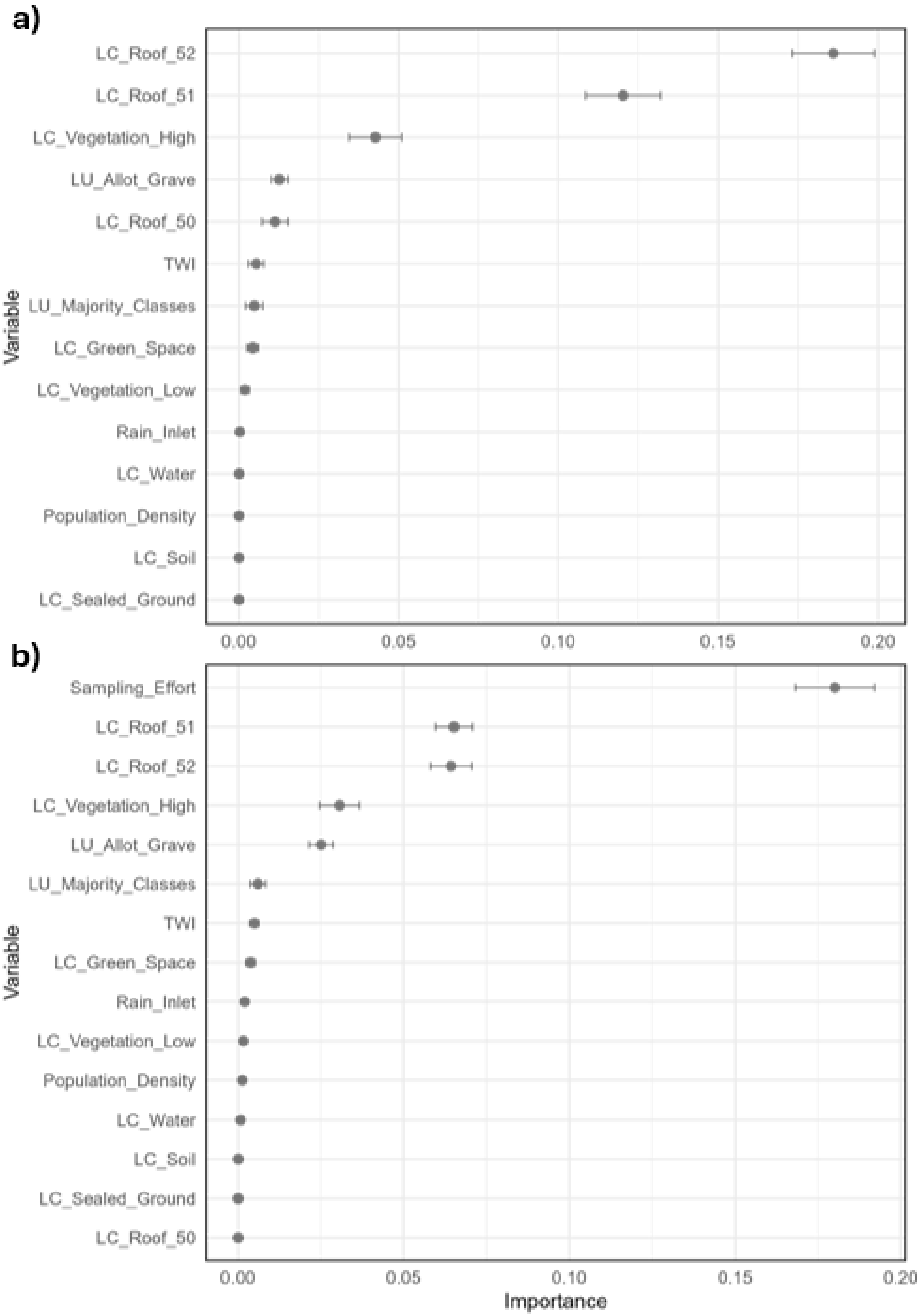

Figure 18: Permutation-based feature importance for MS1 (**a**) and MS2 (**b**). The importance-score describes the difference in AUROC between the baseline model and a model where the respective variables were randomly permuted. Standard error of feature importance was calculated based on 30 simulations.

Four out of six interaction effects integrated during model fitting are visualized in **Figure 20**. Interaction effects of *LC_Vegetation_Low* were not further analyzed due to low significance scores. In general, relevant interactions of *LC_Roof_52* resulting in log-odds of up to 1 were lower than those for

LC_Roof_51 reaching log-odds larger than 2. There was a clear interaction between *LC_Roof_51* and *LC_Green_Space*. The effect of *LC_Roof_51* on the presence of the Asian tiger became increasingly positive as the proportion of *LC_Green_Space* increased. This interaction explains the high predicted values of mosquito presence in the southern outskirts of Graz, represented by one- and two-family homes with garden areas. Contrarily, the interaction between *LC_Roof_51* and *LC_Vegetation_High* showed a strong negative interaction where the highest log-odds are associated with high values of *LC_Roof_51* but low values for *LC_Vegetation_High*. Regarding *LC_Roof_52*, a weaker but positive interaction was observed with green spaces indicating that the presence of *Aedes albopictus* is likely to increase in the city center when green areas are present. The highest values were in the center of the plot, indicating that mosquito presence is most likely when the area of a cell is equally covered by residential multi-story buildings and green spaces. The interaction of *LC_Roof_52* with *LC_Vegetation_High* revealed a similar pattern with a peak in the center of the plot, highlighting the importance of high vegetation in densely populated areas. Interestingly, a second peak in positive log-odds was observed at the lower end of the vegetation gradient, where high vegetation is nearly absent, but multi-story residential buildings cover 30%-60% of the respective cells. All four interaction effects showed negative log-odds when the percentage of residential areas is close to 0, with three of them highlighting the importance of vegetation in built-up residential areas. These findings support the statement that the presence of *Aedes albopictus* is more likely in residential areas compared to rural settings. Furthermore, the presence of green spaces in these areas benefited mosquito presence.

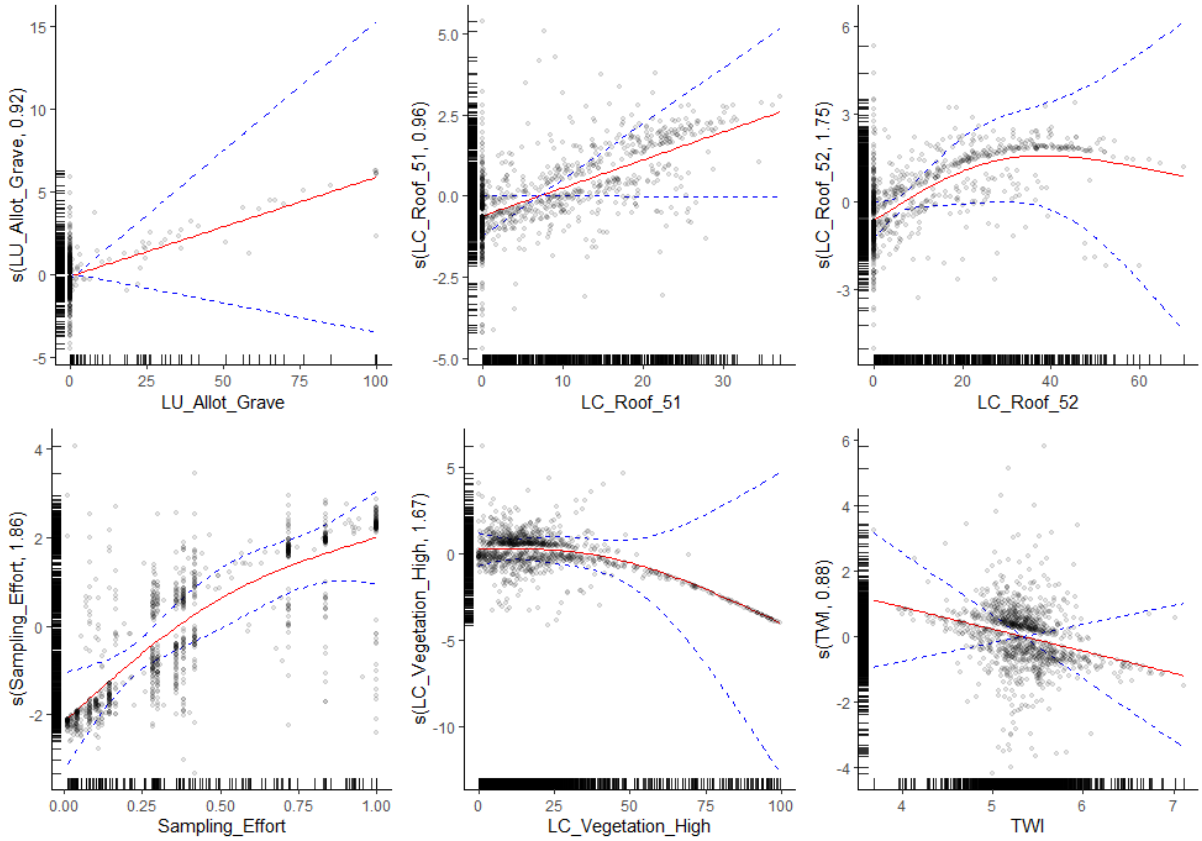

Figure 19: Visualization of partial effect plots illustrating the relationships between various fixed effects and the response variable, with fitted smooth effects (**red**), confidence lines at 5 standard deviations (**blue**), and data residuals (**points**) to enable comparison of predictor effects within the model. Rug marks along the axes show the distribution of predictor values and associated log-odds.

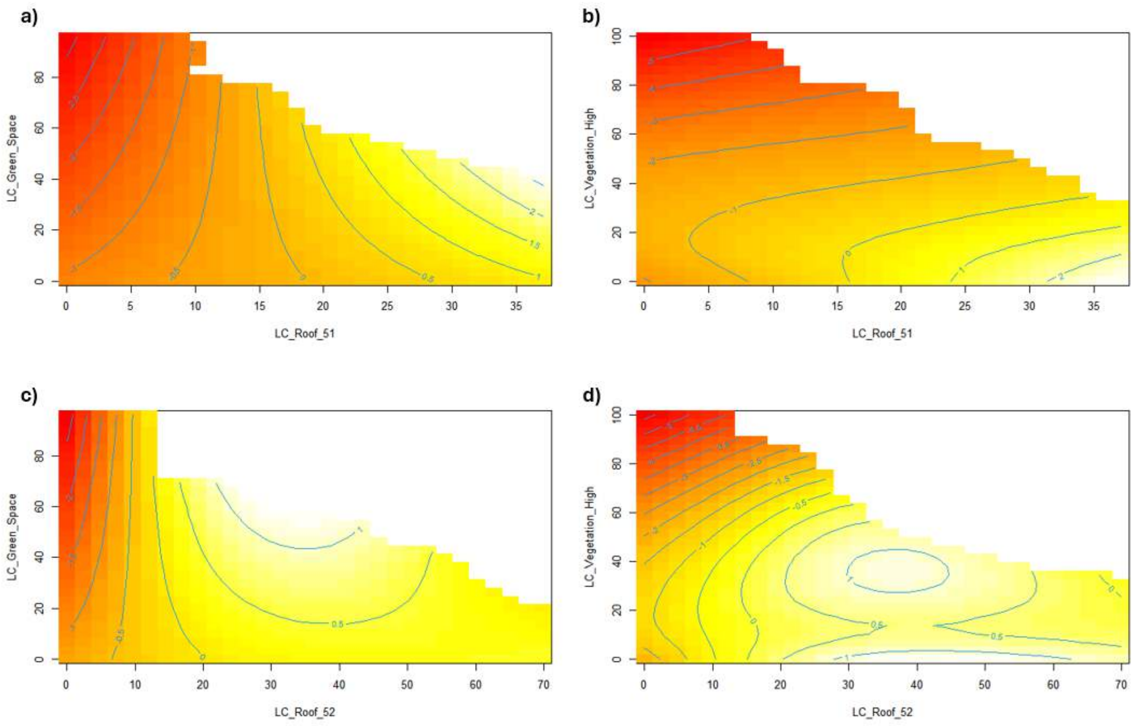

Figure 20: Illustration of interaction effects between residential structures and vegetation type predictors: *LC_Roof_51* & *LC_Green_Space* (**a**), *LC_Roof_51* & *LC_Vegetation_High* (**b**), LC_Roof_52 & LC_Green_Space (**c**) and *LC_Roof_52* & *LC_Vegetation_High* (**d**). The color gradient represents the partial effect on the log-odds of *Aedes albopictus* presence, with lighter (**yellow-white**) shades indicating higher predicted log-odds and **darker red** indicating lower predicted log-odds. Blue isolines represent contours of equal log-odds.

4.1.3 Classification of risk levels and model stability

The GAM (MS2) yielded continuous habitat suitability scores for *Aedes albopictus* across the study area. For easier interpretation, these continuous results were transformed into discrete risk classes, based on a percentile-based approach, informed by the distribution of predicted probabilities at true presence cells. The classification resulted in three distinct risk levels (low, medium and high) with cut-off points being the 0th, 25th, and 95th percentile of observed probabilities associated with true presence locations (see **Figure 14**). The spatial distribution of the three classes is shown in **Figure 21** with *low suitability* covering 53.2%, *medium suitability* 40.6% and *high suitability* 6.2% of the study area. Generally, areas with low suitability are unpopulated, rural regions characterized by forests, agriculture and meadows. In the urban setting, the "Schlossberg" and large infrastructural areas like the railways or construction sites to the West of the old town were classified as low risk. The same accounts for industrial complexes like those in the southeastern part of Graz with large factories and storage buildings. The medium level comprises residential areas in general, since *LC_Roof_51* and *LC_Roof_52* were identified as the most important predictors by the permutation-based feature importance. High risk areas were scattered around the urban area and could roughly be grouped into three categories or a combination of them: cemeteries and allotment gardens (see **Figure 17b**) **(1)**, densely built-up residential areas associated with high values of *LC_Roof_52* (city center) **(2)** and large coverage of one- and two-family homes in combination with green spaces and small amounts of high vegetation (especially southern part of Graz) **(3)**. The combination of predictor values incorporated in interaction effects were decisive if a residential zone was classified as a medium or high-risk area (see **Figure 20**). The interaction between *LC_Roof_51* and *LC_Vegetation_High* depicts a clear example. Both neighborhoods, "Waltendorf" to the East of the city center and "Straßgang" in the southwestern corner of the study area show similar characteristics with large parts being covered by one- and two-family homes with garden areas. Nonetheless, high risk areas were predominantly located in "Straßgang" while "Waltendorf" was primarily classified as a medium-risk area. The reason for this was the difference in the amount of high vegetation with increased coverage in "Waltendorf". Log-odds associated with the interaction of the predictors decreased with increasing coverage of high vegetation (see **Figure 20b**).

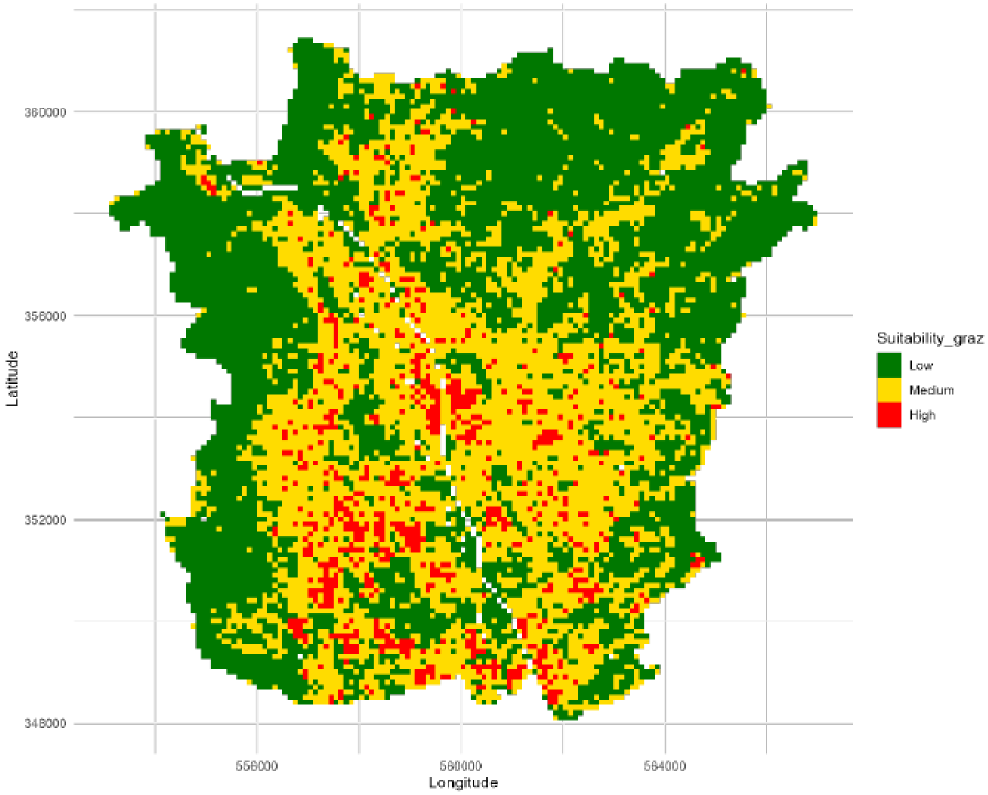

Figure 21: Spatial classification of habitat suitability for *Aedes albopictus* in Graz. The map shows predicted risk levels for mosquito presence based on classified habitat suitability scores. The classification was based on percentiles of predicted probabilities at true presence locations. **Green** indicates low suitability (≤ 25th percentile), **yellow** moderate suitability (25th–95th percentile), and **red** high suitability (> 95th percentile).

Spatial uncertainty in model predictions of MS2 was assessed using repeated random cross-validation. For each cell, the standard deviation σ of predicted mosquito presence probabilities across 50 model iterations was calculated to quantify the variability in model outputs. The mean standard deviation across all cells is 0.047. The distribution of model uncertainty is shown in **Figure 22a**. Most values were below 0.07 with a strong right-skewed distribution, indicating that most cells revealed a low to moderate uncertainty. A peak was observed between 0.02 and 0.05, suggesting high model stability in these areas. The number of cells decreased as uncertainty increased, with only a few cells showing values up to 0.24. The results showed a clear spatial pattern in uncertainty. Clusters of increased model instability were found in the medieval city center to the southwest of the central "Schlossberg", as well as in areas where the dominant land use features were cemeteries or allotment gardens. The observed uncertainty may result from limited data representing the environmental characteristics of these areas, which likely led to their underrepresentation in the training sets produced by random partitioning during cross-validation. Regarding cemeteries and allotment gardens (land use class 11), this becomes clear when comparing the number of samples in the training data (used for random partitioning) with those from other important classes associated with high predicted probabilities (e.g., class 51; see **Figure 15**). This assumption does not directly hold regarding the high uncertainties in the old town (land use class 52), since this class was well represented in the training data set. Increased uncertainties in the old town were potentially due to the implemented interaction effects between *LC_Roof_52* and *LC_Vegetation_High.* While most cells with increased values for *LC_Roof_52* were also associated with at least some high vegetation due to street greening or green courtyards, the amount of high vegetation was close to zero in the central old town. Consequently, as described previously, this specific combination of predictor values (represented by the interaction effect) might be underrepresented in some sets of training data after partitioning during cross-validation. Since this exact combination of predictor values, representing high coverage of residential multi-story buildings without the presence of high vegetation, was associated with increased log-odds (see **Figure 20d**), uncertainty in predictions is expected due to the sparsity of training data. Further isolated clusters of increased uncertainty were likely to be due to rare but important combinations of predictor values expressed by interaction effects.

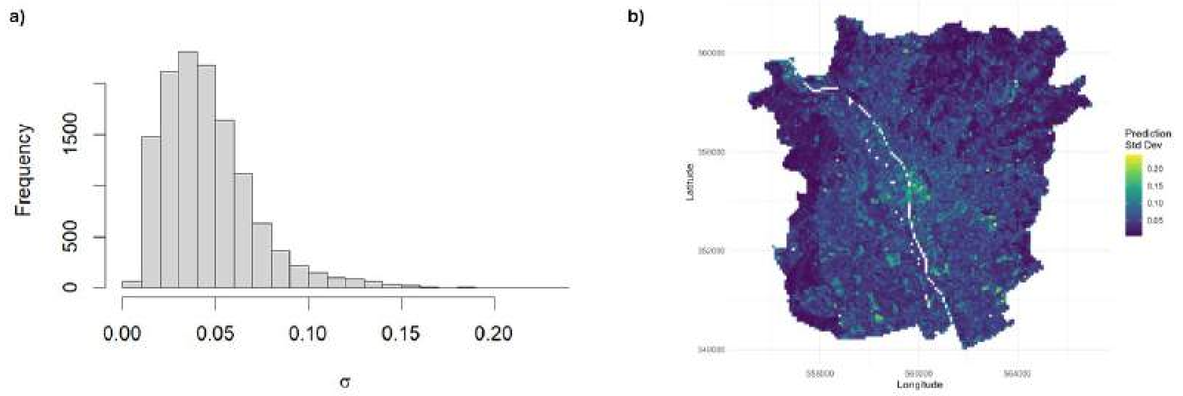

Figure 22: Training data used for model fitting was randomly partitioned to create 50 different data sets (80% of original data set). For each set, habitat suitability scores were predicted using the fitted MS2. For each cell, the standard deviation *σ* of all predictions was calculated. Values ranged from 0 to 0.24 with a right-skewed distribution and a peak between 0.02 and 0.05 **(a)**. Spatial pattern of prediction uncertainty across the city of Graz **(b)**.

To provide a spatial representation of habitat suitability in combination with prediction uncertainty, a bivariate choropleth map was created (**Figure 23**). The map visualizes a combination of the previously discussed risk classes (derived from habitat suitability scores) and model uncertainty (quantified by the standard deviation of predictions) using a 3x3 matrix. The horizontal axis corresponds to the increasing uncertainty and the vertical axis to the increasing risk level. Consequently, each color on the map represents a distinct combination of uncertainty and risk level. Dark blue areas, for example, represent cells with high risk of mosquito presence associated with high model uncertainty.

Overall, the map shows that areas with low suitability generally correspond with low uncertainties, especially in rural zones. In contrast, urban areas with high suitability scores show a broad spectrum of uncertainty levels. Especially areas like the inner city, cemeteries and allotment gardens show high uncertainties (dark blue). On the other hand, the outskirts of Graz in the south and southwest yielded suitable habitats associated with lower uncertainty scores (magenta and purple).

This indicates that the model's predictions in certain urban areas were more sensitive to the composition of the training data. The higher uncertainty in these locations suggests that the partial effects of predictors and their interactions relevant to these environments, such as *LU_Allot_Grave* or the combination of *LC_Roof_52* with predictors related to vegetation coverage, were less stable across different cross-validation folds. This could be due to complex or rare combinations of predictor values, which were not consistently represented in the training sets after random partitioning. In other words, small changes in the training data can lead to noticeable shifts in the predicted effects of these variables, especially when interaction terms were involved. This highlights a lower robustness of the model in these areas. In contrast, the more stable predictions in the southern and southwestern outskirts of Graz indicate that the model performs more robustly where predictor combinations were well represented.

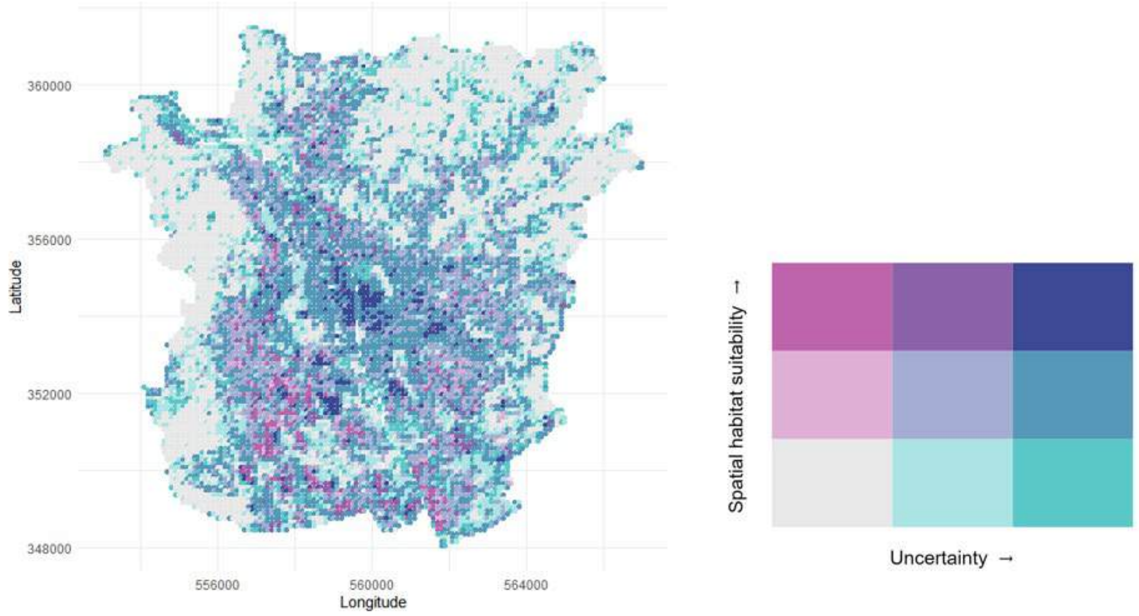

Figure 23: Bivariate choropleth map showing the combination of habitat suitability and prediction uncertainty.

4.2 Temporal model

4.2.1 Model accuracy and performance

A GAMM was trained using three different sets of predictors, with the first set including daily, time-lagged means and time-lagged variations of meteorological parameters (MT1). The second incorporated daily and time-lagged means (MT2) and the third included daily and time-lagged variations (MT3). To evaluate fitting performance, R^2 was derived for all three models. MT1 showed the highest R^2 with a value of 0.78, followed by MT3 with 0.77 and MT2 with 0.75. All models were able to explain a high proportion of mosquito counts with the total difference in performance being minor ($\Delta R^2 = 0.023$). Repeated cross-validation resulted in values of R^2 of 0.72 (RMSE = 1.39) for MT3, followed by 0.71

(RMSE = 1.40) for MT2 and 0.71 (RMSE = 1.41) for MT1. Again, differences in model accuracy and variance explained were negligible ($\Delta R^2 = 0.01;\ \Delta RMSE = 0.02$).

Given the minor differences in model performance and accuracy, selecting one of the three, solely based on the statistics given, would not be meaningful. Model MT1, despite showing the highest value of R^2, was not selected for further analysis. The small increase in explained variation did not justify the increased complexity due to the incorporation of all available predictors. Following the principle of parsimony, also known as Occam's razor, the simpler model should be chosen if performance metrics are similar. This principle suggests that when predictive performance is comparable, the simpler model with less assumptions generalizes better and provides clearer and more interpretable insights (Bargagli Stoffi et al., 2022). Furthermore, since MT1 yielded the highest R^2 associated with the lowest R^2 after cross-validation, one can argue that this model was prone to overfitting since it performed worse on unseen data compared to the other models.

Based on the previously described results, one should prioritize MT3 over MT2. Nonetheless, since time-lagged averages are easier to interpret and more intuitive than time-lagged variations, MT2 was further analyzed and implemented in the spatio-temporal prediction, despite the minimal differences in accuracy. This was especially reasonable since the goal of this model was to not only maximize predictive performance but also to explain ecological relationships. Variations of time-lagged variables can be more difficult to link to biological processes.

MT2 was able to successfully predict the active season of *Aedes albopictus* in Graz in the summer and autumn months with peak mosquito activity in August and September (see **Figure 24a**). The model tended to underpredict peaks of mosquito activity but slightly overestimated the total number of observations between 2022 and 2024 by 21 sightings. Furthermore, by including the main effect *Sampling_Effort* and the random effect *Year*, the annual increase of mosquito sightings, due to growing populations and increased SE, was accurately predicted by the model. This approach is appropriate if one wants to predict the number of mosquito sightings from Mosquito Alert made between 2022 and 2024.

Nonetheless, in this research, the goal was not to accurately predict the number of mosquito reports but rather to explain how mosquito activity relates to a given meteorological condition. To give an easy example, given two identical days in terms of meteorological conditions, the number of observations could vary substantially due to a peak in SE in one day (e.g., after a promotional campaign on the local news). Therefore, to isolate the effect of meteorological predictors on mosquito activity, the variables *Year* and *Sampling_Effort* were zeroed out during prediction (see **Figure 24b**). This is meaningful, since both predictors either reflect human reporting behavior or annual population growth instead of the actual climatic suitability. By zeroing out respective predictors, the model allowed for a clearer assessment of how past and current weather and seasonality shapes the risk of mosquito occurrence.

Additionally, if one wants to predict future risk levels of mosquito presence, zeroing out the *Year* variable is essential. Since this predictor is a factor variable with its levels only being fitted to past data (e.g. 2022 to 2024), future predictions would not be possible. This makes the full model unsuitable for forecasting applications (e.g. mosquito risk assessment based on weather forecast). In contrast, the reduced prediction approach allows for future predictions under different scenarios, independent from sampling behavior or population growth.

The reduced model clearly underestimated the observed number of mosquitoes, especially during peak mosquito season. It is important to mention that this underestimation was no drawback in the context of this study. The goal of the reduced prediction was to estimate the relative risk driven by meteorological conditions alone. Moreover, the observed daily counts from Mosquito Alert itself are very likely to underestimate the true number of mosquitoes in Graz.

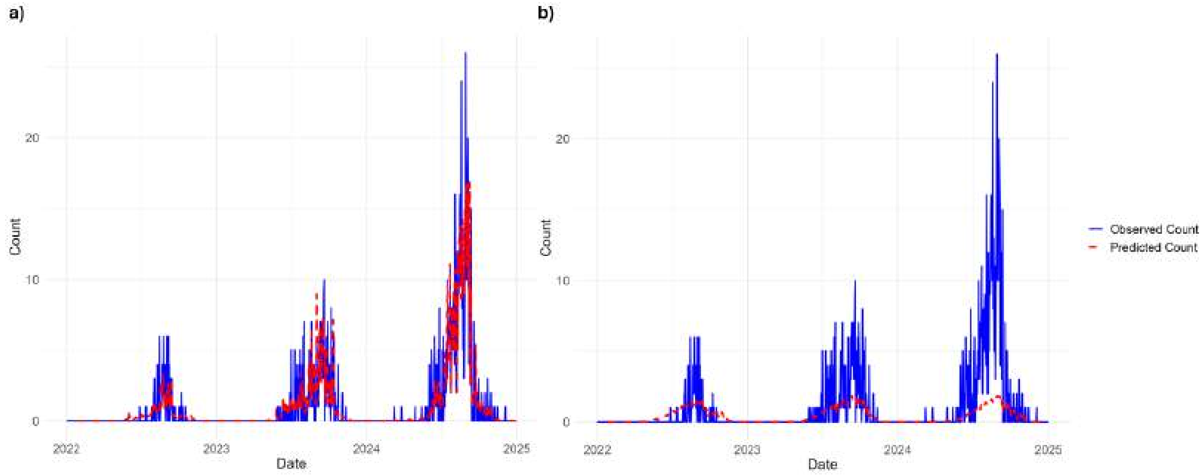

Figure 24: Comparison of observed and predicted daily mosquito counts in Graz using MT2.
a) Predictions included the smooth effect of *Sampling_Effort* and the random effect for *Year*. This allowed the model to reflect annual growth in mosquito reports and temporal variation in observer activity.
b) Predictions excluded these two effects, isolating the influence of meteorological predictors. While predicted counts were lower, this approach enabled unbiased assessment of weather-driven mosquito activity and supports risk classification independent of human reporting behavior.

4.2.2 Feature importance and partial effects

In the following, global feature importance, partial effects and significance of predictors are only described for MT2 due to previously discussed reasons. After model fitting, *Sampling_Effort* (p-value < 0.001)*, Year* (p-value < 0.001)*, Doy* (p-value < 0.001)*, T_max* (p-value = 0.001)*, T_mean_mean_28* (p-value = 0.005) were identified as significant effects (p-value < 0.05). Non-significant effects were *RH_mean* (p-value = 0.061), *P_acc_28* (p-value = 0.200)*, Rh_mean_mean_28* (p-value = 0.246)*, WS_cv* (p-value = 0.256)*, WS_mean_mean_28* (p-value = 0.467)*, WS_mean* (p-value = 0.469) and *P* (p-value = 0.731).

Permutation-based feature importance was performed to assess the difference in RMSE for each predictor with the effects *Sampling_Effort* and *Year* being zeroed out (see **Figure 25**). Highest feature importance was identified for the time-lagged average of the daily mean temperature (*T_mean_mean_28*) with an associated difference in RMSE of 0.241. The day of the year (*Doy*) variable was second most influential for explaining the variability of daily mosquito counts with a difference in RMSE of 0.215. Further relevant predictors were the daily maximum temperature (*T_max*, ΔRMSE = 0.161), daily mean relative humidity (*RH_mean*, ΔRMSE = 0.025), and time-lagged accumulated precipitation (P*_acc_28*, ΔRMSE = 0.003). Automatic variable selection identified and removed WS*_cv, WS_mean, P, WS_mean_mean_28* and *RH_mean_mean_28* since they did not explain variability in the response, leading to a feature importance score of zero for those predictors. Negative values for ΔRMSE occur when randomly shuffled values of predictors better explain variability in the response than the actual data. This implies that associated predictors are not useful in predicting daily mosquito counts.

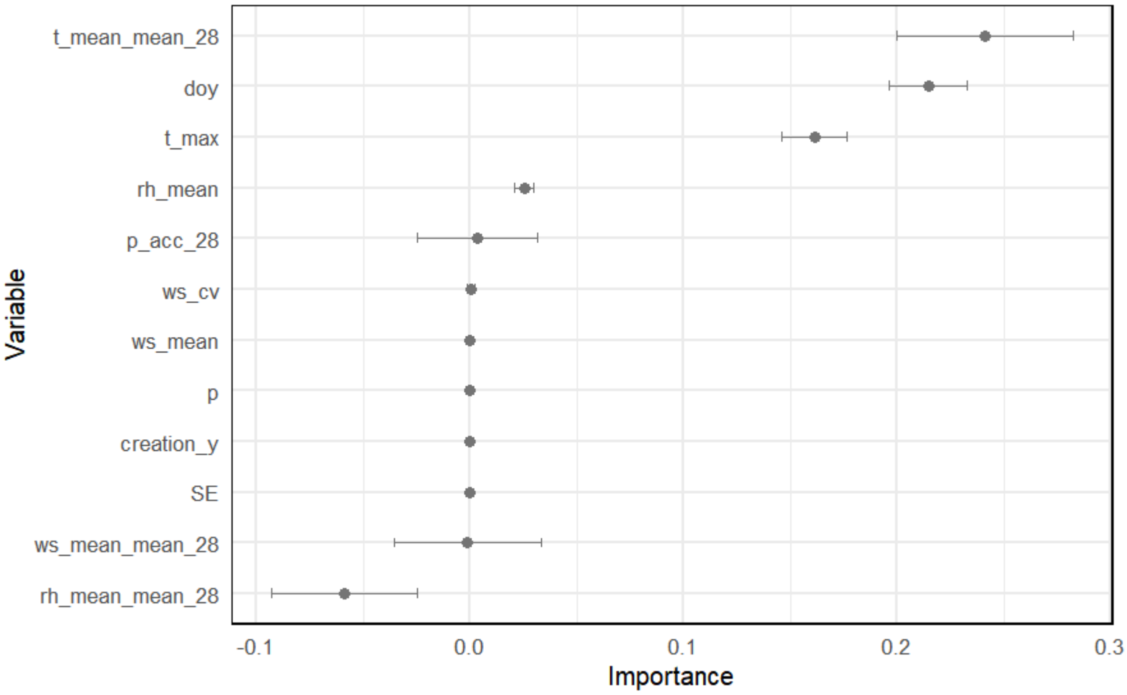

Figure 25: Permutation-based feature importance for MT2. The importance-score describes the difference in RMSE between the baseline model and a model where the respective variable was randomly permuted. Standard error of feature importance was calculated based on 30 simulations.

Partial effect plots of relevant main effects are shown in **Figure 26**. Since the GAMM modeled a negative binomial distribution of the response variable using a log link function, the y-axis of each plot represents the effect of the predictor on the log of the expected count. Consequently, a negative value on the y-axis corresponds to a decrease and positive values lead to an increase in expected counts.

Similar to the spatial model, a positive relationship between *Sampling_Effort* and the response was observed, underpinning the bias in citizen science data and supporting the decision to zero out *Sampling_Effort* during prediction. The time-lagged average temperature (*T_mean_mean_28*) showed a non-linear effect on the log of the expected mosquito counts. At temperatures below approximately 10 °C the partial effect was negative, resulting in a reduction in expected mosquito counts. Above this threshold, increasing values were associated with increasingly positive effects on the log-expected count. Consequently, increasing temperatures in the preceeding 4 weeks were linked to elevated predicted mosquito activity. The partial effect plot for *Doy* revealed a clear seasonal pattern with a positive relationship during summer and autumn and a negative one in winter and early spring. A maximum was observed around the 260th day of the year (mid-September). The partial effect for the daily maximum temperature (*T_max*) revealed a similar trend compared to the time-lagged temperature, with associated values on the y-axis being slightly lower. The zero-crossing of the graph was located between 13 °C and 14 °C. With increasing values of *T_max*, the slope of the smooth effect decreases. Mean relative humidity (*RH_mean*) had a weak and nearly flat effect. The function remained close to zero across the range, with minor fluctuations, suggesting that this variable had little influence on mosquito counts in the model.

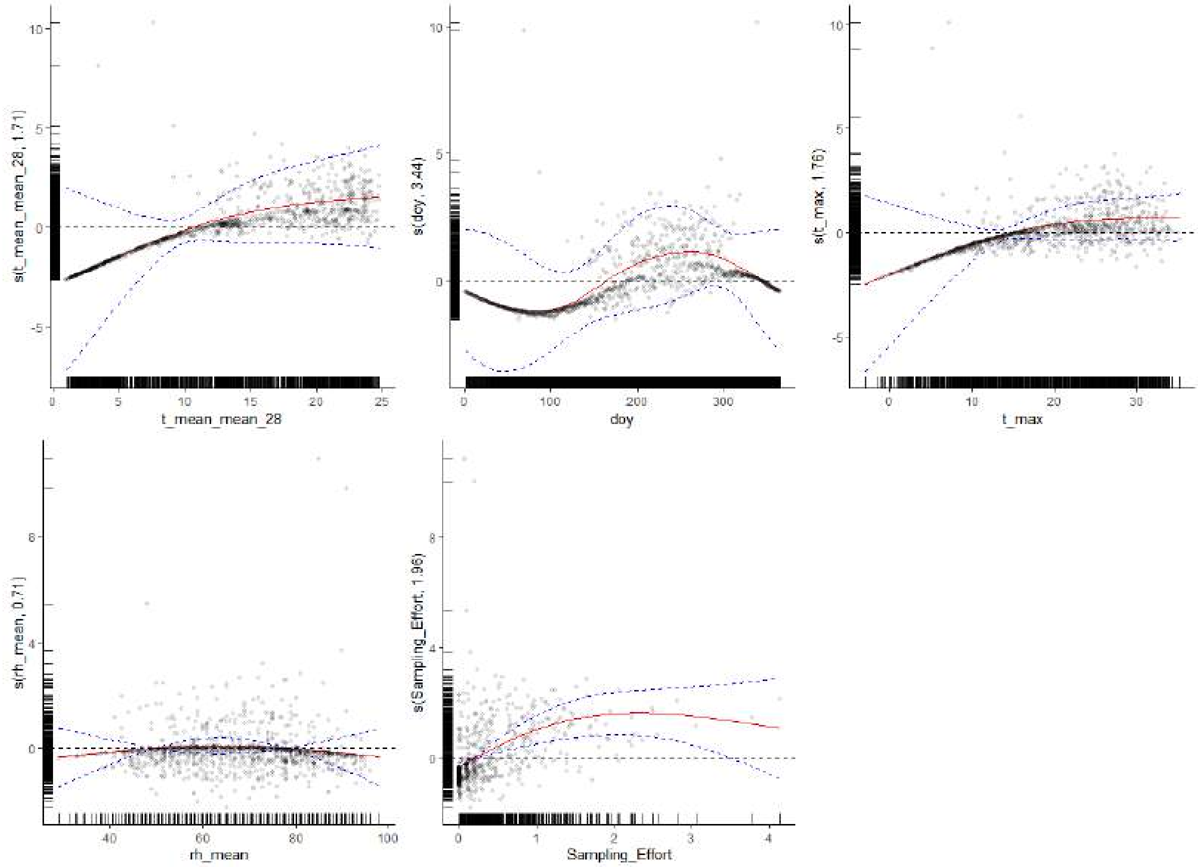

Figure 26: Visualization of partial effect plots illustrating the relationships between various fixed effects and daily mosquito counts, with fitted smooth effects (**red**), confidence lines at 5 standard deviations (**blue**), and data residuals (**points**) to enable comparison of predictor effects within the model. Rug marks along the axes show the distribution of predictor values and associated log-odds.

4.2.3 Classification of risk levels

Continuous daily mosquito counts predicted by the reduced MT2 were transformed into discrete risk classes, based on a percentile-based classification scheme informed by the distribution of predicted counts on days of true mosquito observations. This approach resulted in three distinct risk levels (low, medium, and high) with thresholds being the 0th, 25th, and 95th percentile of predicted mosquito counts associated with true presence days (see **Figure 14**). The resulting temporal risk index over three full years (2022-2024) is shown in **Figure 27**. The figure illustrates the predicted mosquito counts per day based on meteorological conditions and the day of the year, classified into three risk classes. The seasonal pattern of mosquito activity became evident with high-risk periods occurring in late summer between mid-August and early September. Medium-risk periods exceeded the summer months, ranging from the beginning of June to approximately mid-October. Low-risk periods spanned from late autumn to the end of May with the risk of mosquito presence being close to zero in winter. Over the entire time span (2022-2024), *low-risk* days accounted for 52.7% of all days, while *medium-* and *high-risk* days represented 38.0% and 6.7%, respectively. An additional 2.6% of all days could not be classified due to missing predictor data.

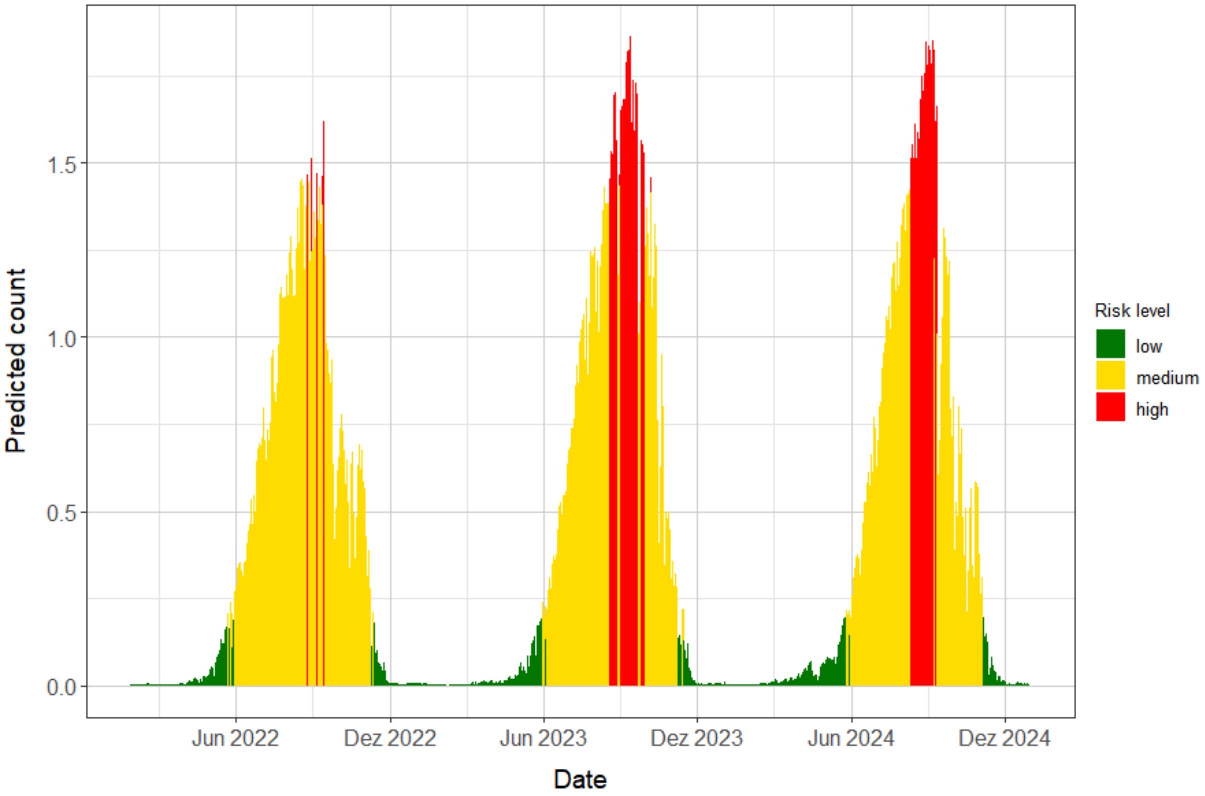

Figure 27: Temporal classification of daily mosquito activity into low **(green)**, medium **(yellow)**, and high **(red)** risk levels based on weather-driven predictions from MT2 (2022–2024). The classification highlights seasonal patterns, with peak risk occurring in late summer and early autumn.

Chapter 5 Discussion

The aim of this work was to predict the presence and activity of *Aedes albopictus* in space and time on a city-scale in Graz, Austria, based on citizen science data by employing a generalized additive (mixed) modeling framework. For this purpose, two separate models were developed: a spatial GAM to predict habitat suitability across the city of Graz and a temporal GAMM to explain the daily activity of *Aedes albopictus* based on meteorological conditions. By combining both approaches, daily predicted mosquito counts were spatially distributed using the predicted habitat suitability scores to visualize the presence of the Asian tiger mosquito in space and time. Both the spatial and the temporal model showed biologically plausible predictions and were able to capture the main drivers of mosquito activity.

The use of citizen science data offers clear advantages, like the broad spatial and temporal coverage, as well as overcoming financial constraints that are typically associated with traditional monitoring networks. At the same time, given the nature of citizen science data, one needs to deal with the bias due to varying SE across the study area (Palmer et al., 2017). Therefore, this work explicitly used only citizen science data (and due to the insufficient spatial coverage by insect traps) to analyze the plausibility of the results and the applicability of this approach.

This chapter discusses the main findings in relation to the following five research questions:

1. Two approaches were tested to correct spatial bias inherent in citizen science data. Which is the better performing one in terms of biological plausibility?
2. Which factors are most influential on modeling the presence of *Aedes albopictus*?
3. Which areas are most prone to the presence of *Aedes albopictus*?
4. How do results compare to findings from similar studies in other regions?
5. Given the bias inherent in citizen science data, can the model demonstrate accurate and plausible predictions? What are the advantages and disadvantages of using these data?

Each subsection is structured around one to three research questions, starting with the comparison of the two approaches tested to account for spatial bias in observation data (section 5.1). In section 5.2, the most influential predictors are discussed and compared with other studies, leading to the identification of areas most at risk of mosquito presence (RQ 2, 3, and 4). Section 5.3 answers research question 5 by discussing the overall plausibility of these results, including the advantages and disadvantages of using citizen science data.

5.1 Comparison of bias correction approaches

Correcting for sampling bias inherent to Mosquito Alert data is essential since otherwise bias in citizen science data would directly propagate into the modeled results. To address this problem, two different approaches were tested and compared in terms of accuracy and biological plausibility. The first approach (MS1) employed a probability proportional to size sampling of pseudo-absences based on associated values of SE. This observer-oriented approach is commonly applied in habitat suitability modeling when dealing with biased presence-only data (Moua et al., 2020; Phillips et al., 2009). Instead of using SE as a weight for sampling pseudo-absences, the second approach (MS2) directly incorporated SE as a covariate during model fitting, while excluding them from predictions.

According to Swets (1988), both approaches resulted in a "good" model fit with differences in accuracy being marginal with AUROC scores of 0.82 (MS1) and 0.80 (MS2). By only comparing both AUROC values, one might fail to identify the "better" model, since this value only represents a model's ability to differentiate between presences and absences. Consequently, it does not quantify the quality of predicted probabilities and their associated habitats and ecological relevance (Lobo et al., 2008). Despite the similarities in accuracy, the distribution of predicted habitat suitability scores showed very different spatial patterns. MS1 identified the central residential area of Graz (excluding the old town), while MS2 highlighted cemeteries, allotment gardens, and the suburban areas in the southwest (where trees are absent) as areas with highest probabilities for the presence of *Aedes albopictus.*

The results show that MS2 yields a biologically more plausible prediction of the presence of *Aedes albopictus* in Graz. Habitats and hotspots identified by this approach, like cemeteries, allotment gardens and residential areas with private gardens are known to yield suitable conditions for this species. These areas provide numerous natural or artificial water containers, shaded areas for rest and regular human activity, with all these aspects being important requirements for the species' survival and reproduction. In contrast, the inner city, which is characterized by dense construction, less breeding sites and increased values of SE, was highlighted by MS1 as area being most prone to mosquito occurrence. This indicates that the approach of sampling pseudo-absences based on SE is more prone to bias propagation, due to increased habitat suitability scores in densely populated areas (Carney et al., 2022). In contrast, the approach of including SE as a main effect during model fitting better reflected biologically expected patterns of habitat suitability. Field observations in Austria and Germany support the general findings from MS2. Ovitraps operated at three positions in Graz with one being in an allotment garden (217 eggs), a second one in the city center (only a few eggs) and a third in an industrial area (no eggs) (Reichl et al., 2023). Similar observations were done by Bakran-Lebl et al. (2025), with highest egg counts being found near cemeteries and allotment gardens. In southwestern Germany, allotment gardens and residential areas with garden areas were identified as areas at risk[6].

However, the high habitat suitability values predicted for the medieval old town were unexpected, as these areas lack street greenery and provide little opportunity for artificial container habitats such as gardens or balconies (based on Google Street View inspection). One explanation might be the very small extent of the old town within the study area, which resulted in only a single pseudo-absence being randomly sampled (n = 1, see **Figure 5**). Consequently, the absence characteristics of this area were insufficiently represented during model fitting, which led to artificially increased log-odds for these areas (see **Figure 21d**).

5.2 Drivers for mosquito presence

The spatial GAM and the temporal GAMM identified several predictors as relevant to explain the presence and activity of *Aedes albopictus* in Graz. Permutation-based feature importance revealed that predictors related to the land cover of residential buildings were most important to explain the distribution of mosquito presences. *LC_Roof_51*, related to the coverage of one- and two-family homes was identified as most important predictor, followed by *LC_Roof_52,* related to residential multi-story buildings. In contrast, *LC_Roof_50* (other buildings like industrial or storage buildings) did not contribute to explaining variability in the response. Further relevant predictors were *LC_Vegetation_High* and *LU_Allot_Grave*, while the contribution of the remaining predictors was negligible or close to zero. However, the global permutation-based feature importance only informs about how much each predictor contributed to the predictive performance of the model, expressed by the change in AUROC. It does not yield any information about whether a predictor increases or reduces the probability of mosquito presence. To assess if the effect of a predictor is beneficial or disadvantageous for the occurrence of the Asian tiger mosquito, partial effect plots for smooth effects and interaction effects were analyzed.

Increasing values for *LC_Roof_51* and *LC_Roof_52* were associated with increasing log-odds. Higher log-odds of *LC_Roof_51* indicate that land coverage of one- and two-family homes had a stronger positive effect on the occurrence of the Asian tiger mosquito than the land coverage of residential multi-story buildings. Furthermore, the partial effect plot for the coverage of cemeteries and allotment gardens revealed the highest log-odds across all predictors, indicating the important role of these areas in future risk management. The predictor was only ranked 4th in global feature importance due to the zero-inflated values of this predictor. Compared to the other three predictors ranked above, *LU_Allot_Grave* was only relevant in a few small areas within the study area while other land use types covered large parts of the study area (e.g., forests represented by *LC_Vegetation_High*). These findings are consistent with citizen science and ovitrap surveys in Graz (Bakran-Lebl et al., 2025; Reichl et al., 2023), and they align well with results from other European and international studies. In northeastern Italy, Baldacchino et al. (2017) showed that the abundance of *Aedes albopictus* strongly correlated with the proportion of urban area, while agricultural and forested area did not play a role. Li et al. (2014) support these findings, showing that urbanization in China increased the density of larval habitats and accelerated their

[6] https://www.gesundheitsamt-bw.de/lga/de/kompetenzzentren-netzwerke/arbo-baden-wuerttemberg/verbreitung-von-tigermuecken/

development and survival rates. In an urban environment in Guangzhou, China, the highest number of habitats were found in residential areas, followed by parks, construction sites and schools (Guo et al., 2023). Also in southern New York, sites with greater impervious surface had significantly higher abundance of *Aedes albopictus* (Shragai & Harrington, 2019). Taken together, these studies agree that urbanization and peri-domestic environments fulfill habitat preferences of *Aedes albopictus.* However, to date there has been little work on differentiating between different land use classes within the urban environment. The present results provide a more detailed picture by further separating built-up areas and vegetation types.

Increasing values for high vegetation were associated with decreasing log-odds, due to the absence of reports in forested areas. Therefore, the isolated smooth effect *LC_Vegetation_High* had a negative relationship with the probability of mosquito presence, agreeing with findings from Oklahoma, USA (McMahon et al., 2022). However, partial effect plots of interaction effects revealed that the presence of high vegetation in densely populated residential areas (interaction with *LC_Roof_52*) was beneficial for mosquito presence. The same characteristic was found for *LC_Green_Space.* The isolated smooth effect showed a negative effect on mosquito presences due to the absence of reports in large meadows while green spaces in residential areas (*Roof_51* and *Roof_52*) benefited mosquito occurrence. Surprisingly, while green spaces in loosely built-up residential areas had a positive effect, the presence of trees resulted in negative log-odds and consequently reduced the probability of mosquito presence in these areas. Generally, large parts of these findings agree with previous work, since variations in microclimate due to vegetation in an urban environment counteracts urban heat and dryness, thus favoring habitat conditions (Murdock et al., 2017). Furthermore, vegetation in an urban setting provides shades and organic matter, beneficial for mosquito activity (Shragai & Harrington, 2019). In Rome, Italy, trees were positively correlating with egg numbers but grass cover showed a negative effect (Cianci et al., 2015).

In the reduced temporal model, time-lagged average temperature (*T_mean_mean_28*) contributed the most in explaining variability in daily mosquito counts, followed by the day of the year parameter (*Doy*), the daily maximum temperature (*T_max*) and the daily average of relative humidity (*RH_mean*). A time-lagged average temperature of at least 10 °C was needed to favor mosquito occurrence, which is consistent with experimental studies showing a developmental threshold of around 10-13 °C for *Aedes albopictus* eggs and larvae (Cai et al., 2023; Marini et al., 2020; Thomas et al., 2012). Surprisingly, permutation-based feature importance of daily precipitation and time-lagged accumulated precipitation was zero or very close to zero, despite the necessity of breeding sites being filled with water. One reason might be due to artificial watering of vegetation and green spaces in garden areas or on balconies, etc. Therefore, potential breeding sites might be filled with water, despite ongoing dry periods. Comparable results were obtained in southern Europe, where precipitation showed only a weak association with mosquito abundance, since artificial breeding sites maintain water availability despite the absence of precipitation. Furthermore, when modeling invasive species, the assumption being made that the species is in equilibrium with its local environment is typically hurt (Guisan et al., 2017). Since mosquito populations in Graz are growing annually, there is further potential for reproduction and larger populations. At this early stage of invasion (first *Aedes albopictus* sighting in Graz was in 2021), even low mosquito numbers could be sustained by a limited number of breeding sites despite the absence of precipitation events (Roques et al., 2023). Another reason might be due to the nature of citizen science data. The Asian tiger mosquito, being an exophilic and exophagic species, primarily encounters humans (Mosquito Alert users) outdoors. Consequently, the number of reported sightings may be biased on rainy days, as people are less likely to spend time outside and submit observations. Since the SE variable provided by Mosquito Alert only reflects the number of registered users within a cell, but not their actual activity on a given day, it cannot sufficiently account for this reduction in reporting during rainfall events (Palmer et al., 2017).

Overall, the results suggest that peri-domestic residential environments, especially one- and two-family homes with green spaces (but no trees), allotment gardens, and cemeteries are the most suitable areas of *Aedes albopictus* presence in Graz. Vegetation effects are context-dependent, but in dense urban settings, shaded green spaces enhance mosquito occurrence. From a temporal perspective, temperature (especially time-lagged average above 10 °C) emerged as the strongest predictor while precipitation did not explain temporal mosquito activity.

5.3 Applicability of citizen science data

The present study relied solely on citizen science data from Mosquito Alert to model the spatio-temporal presence of *Aedes albopictus* in Graz. This decision was made due to the insufficient coverage of ovitraps in Graz, but also to explore the applicability of citizen science data in this field.

A major advantage of citizen science data is the extensive spatial and temporal coverage without facing financial constraints in comparison to traditional observation networks with ovitraps and maintenance personnel (Carney et al., 2022; Palmer et al., 2017). This is especially the case for invasive species, where the spatial expansion of a species outpaces the expansion of observation networks (Bartumeus et al., 2018). Therefore, the Mosquito Alert network allowed for early-warning of newly invaded areas by several *Aedes* species as well as first detections across Europe (Carney et al., 2022).

However, the use of citizen science data also comes with challenges. The quality of modeled results depends on the possibility of correcting for heterogeneous SE. Otherwise, the biased distribution of presence samples towards areas with high population would translate into the spatial distribution of modeled habitat suitability scores. In this study, the strong relationship between SE and presence observations was demonstrated by SE being the most important predictor to explain variability in the response data (in MS2). Additionally, heterogeneous observer bias resulted in fluctuations of presence observations over time, influenced by Mosquito Alert promotion campaigns, administrative activities or maybe even due to human behavior based on the weather.

Despite the bias, modeled probabilities showed that biologically plausible results can be obtained when explicitly accounting for SE. The second spatial approach (MS2), in which SE was employed as a main effect during model fitting, yielded results that mostly agree with previous works from around the globe (see section 5.2). The broad pattern of habitat suitability revealed a gradient from rural to urban environments. Within the urban context, cemeteries and allotment gardens were successfully detected as areas being most prone to mosquito presence followed by residential areas. While the broad predictions were biologically plausible, some patterns need to be interpreted with caution. In the inner city (around the "Schlossberg"), modeled results are likely to overestimate habitat suitability since field observations only show very low egg counts in areas of this characteristic (Bakran-Lebl et al., 2025; Reichl et al., 2023). The same counts for the old town as explained in section 5.1. In general, while the interaction between vegetation and residential built-up area revealed a positive effect, the negative interaction between high vegetation and loosely built-up residential areas was unexpected. As a result, predicted probabilities in the eastern suburbs are lower than predicted probabilities in the southern suburbs. These findings should be treated with caution, as the low number of presences in the eastern suburbs may simply reflect that the species has not yet fully established itself in this area (**Figure 4** shows an ongoing spread towards the east over the past four years), instead of indicating that the habitat is unsuitable itself. To reduce the risk of false pseudo-absences in areas where the species is still expanding, a convex hull was formed around all presence observations to constrain the sampling area of pseudo-absences. Nonetheless, this problem related to the modeling of invasive species cannot be fully corrected, since the assumption of equilibrium with the environment is typically violated. Consequently, instead of capturing the potential environmental niche, the model only quantifies the realized environmental niche of the current invasion state (Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000).

The reduced temporal model successfully predicted the seasonality of daily mosquito counts with increasing mosquito activity from spring to early autumn, mainly based on parameters related to seasonality (*daily maximum temperature*, *time-lagged temperature,* and *day of the year*). However, predictors that fluctuate on short time scales, like wind speed or precipitation, did not contribute to predictions. Consequently, the predictive power of the temporal model for the daily activity is limited, with predicted values representing rather a seasonal risk assessment instead of a daily one. This might be due to the temporal fluctuation of SE related to humans' motivation to go outdoors (e.g. rainy vs. sunny weather), which is not captured by Mosquito Alert's SE parameter (Palmer et al., 2017), thus possibly distorting modeled results.

Overall, this research showed that the use of citizen science data allows for biologically plausible spatio-temporal modeling of *Aedes albopictus* in Graz. This cost-effective approach offered broad spatial and temporal coverage which would not have been feasible with traditional observation methods. At the same time, the uneven SE and the non-equilibrium state of invasive species limited fine-scale plausibility.

Chapter 6 Conclusion and potential for future work

This work demonstrated the applicability of citizen science data to predict the spatio-temporal presence of *Aedes albopictus* in Graz, Austria. By employing generalized additive (mixed) models, the spatial habitat suitability, as well as seasonal mosquito activity were modeled with biologically plausible results. Peri-domestic environments, such as loosely built-up neighborhoods with garden areas (and no trees), allotment gardens and cemeteries were identified as "high-risk" areas. Predicted habitat suitability in the inner city (especially the old town) is expected to be inflated due to the insufficient coverage by pseudo-absences or the influence of sampling bias which could not be eliminated completely. On the temporal scale, 28-day time-lagged temperature was the main driver for seasonal mosquito activity, while precipitation had no significant impact.

A major achievement of this work was to show that the use of citizen science data can yield valuable insights, despite their associated challenges regarding observation-bias. This work suggests that including SE as a predictor (if available as numerical value) during model fitting yields biologically more plausible results compared to the weighted sampling of pseudo-absences. Nonetheless, results should be interpreted with caution since bias due to heterogeneous SE can only be corrected for but not fully eliminated. Given the fact that the species population is growing annually, one can assume that it has not reached its full potential yet. Consequently, it is important to consider that the results presented here only reflect the current state of invasion of *Aedes albopictus,* as only the realized niche could be modeled instead of the full potential niche. For the same reason, some environmental drivers may have not been identified as influential in the present models, as the ongoing non-equilibrium state of invasion can mask their ecological relevance. Another important contribution of this work lies in the detailed usage of spatial predictors. While most works only consider coarse indicators of urbanization like broad land use types or the amount of impervious surface (Baldacchino et al., 2017; Shragai & Harrington, 2019), this study differentiated between different types of residential areas and vegetation, cemeteries or allotment gardens. This more detailed approach allows for a more precise picture of heterogeneous habitat suitability across an urban area. According to my knowledge, this is the first attempt in modeling habitat suitability of *Aedes albopictus* on a city-scale in Austria using statistical models, thus providing an entry point for further research.

Looking forward, several directions for future research emerge. First, the uncertainty of modeled results could be improved by running multiple simulations based on different sets of pseudo-absence data. Consequently, fitted interaction effects between residential building types and vegetation types would be more robust, leading to more reliable predictions in sparse areas that are represented by certain rare characteristics (e.g., the old town). Another approach to improve uncertainty and credibility is the combination of field data from ovitraps and citizen science data to further minimize the effect of observation bias (Eritja et al., 2025). Another promising next step would be to integrate the spatial and temporal model within one single spatio-temporal modeling framework. This approach would allow for space-time interactions (e.g. the effect of urban heat might differ depending on the month). Da Re et al. (2025) have demonstrated the feasibility of such an approach by using a set of spatio-temporal weather correlates to predict egg numbers in southern Europe using ensemble modelling techniques. However, this approach requires spatially resolved microclimatic data that describes spatial gradients due to e.g. the urban heat effect. Another promising extension would be the integration of the species mobility patterns into modeled results. Since this model predicts the probability of adult mosquito presence within each cell, the mosquitoes are not confined to the location but may also move to neighboring cells. Vavassori et al. (2019) showed that most individuals of *Aedes albopictus* move more than 250 m and up to 1 km. Combining the presence model with mobility patterns would yield more realistic and smoother results, better reflecting actual risk levels.

References

Aedes albopictus - current known distribution: July 2024. (2024, July 8). https://www.ecdc.europa.eu/en/publications-data/aedes-albopictus-current-known-distribution-july-2024

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer New York. https://doi.org/10.1007/978-1-4612-1694-0_15

Angelini, R., Finarelli, A. C., Angelini, P., Po, C., Petropulacos, K., Macini, P., Fiorentini, C., Fortuna, C., Venturi, G., Romi, R., Majori, G., Nicoletti, L., Rezza, G., & Cassone, A. (2007). An outbreak of chikungunya fever in the province of Ravenna, Italy. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *12*(9), E070906.1. https://doi.org/10.2807/esw.12.36.03260-en

Aranda, C., Martínez, M. J., Montalvo, T., Eritja, R., Navero-Castillejos, J., Herreros, E., Marqués, E., Escosa, R., Corbella, I., Bigas, E., Picart, L., Jané, M., Barrabeig, I., Torner, N., Talavera, S., Vázquez, A., Sánchez-Seco, M. P., & Busquets, N. (2018). Arbovirus surveillance: First dengue virus detection in local Aedes albopictus mosquitoes in Europe, Catalonia, Spain, 2015. *Eurosurveillance*, *23*(47). https://doi.org/10.2807/1560-7917.ES.2018.23.47.1700837

Austin, M. P., Nicholls, A. O., & Margules, C. R. (1990). Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species. *Ecological Monographs*, *60*(2), 161–177. https://doi.org/10.2307/1943043

Avenell, D., Medlock, J., Ducheyne, E., Scholte, E.-J., Hendrickx, G., & Schaffner, F. (2009). *Development of Aedes albopictus risk maps*.

Ayala, D., Costantini, C., Ose, K., Kamdem, G. C., Antonio-Nkondjio, C., Agbor, J.-P., Awono-Ambene, P., Fontenille, D., & Simard, F. (2009). Habitat suitability and ecological niche profile of major malaria vectors in Cameroon. *Malaria Journal*, *8*(1), 307. https://doi.org/10.1186/1475-2875-8-307

Bakran-Lebl, K. (2025a). *Asiatische Tigermücke (Aedes albopictus), Informationen und Empfehlungen für betroffene Regionen*. https://www.ages.at/en/news/detail/asiatische-tigermuecke-in-ganz-oesterreich-gefunden

Bakran-Lebl, K. (2025b). *Ovitrap-Monitoring gebietsfremder Gelsenarten in Österreich*.

Bakran-Lebl, K., Seebacher, B., & Reichl, J. (2025). *Ovitrap-Monitoring gebietsfremder Gelsenarten in Österreich—Jahresbericht 2024*.

Baldacchino, F., Marcantonio, M., Manica, M., Marini, G., Zorer, R., Delucchi, L., Arnoldi, D., Montarsi, F., Capelli, G., Rizzoli, A., & Rosà, R. (2017). Mapping of Aedes albopictus Abundance at a Local Scale in Italy. *Remote Sensing*, *9*(7), 749. https://doi.org/10.3390/rs9070749

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

Bargagli Stoffi, F. J., Cevolani, G., & Gnecco, G. (2022). Simple Models in Complex Worlds: Occam's Razor and Statistical Learning Theory. *Minds and Machines*, *32*(1), 13–42. https://doi.org/10.1007/s11023-022-09592-z

Bartlett-Healy, K., Unlu, I., Obenauer, P., Hughes, T., Healy, S., Crepeau, T., Farajollahi, A., Kesavaraju, B., Fonseca, D., Schoeler, G., Gaugler, R., & Strickman, D. (2012). Larval Mosquito Habitat Utilization and Community Dynamics of Aedes albopictus and Aedes japonicus (Diptera: Culicidae). *Journal of Medical Entomology*, *49*(4), 813–824. https://doi.org/10.1603/ME11031

Bartumeus, F., Oltra, A., & Palmer, J. R. B. (2018). Citizen Science: A Gateway for Innovation in Disease-Carrying Mosquito Management? *Trends in Parasitology*, *34*(9), 727–729. https://doi.org/10.1016/j.pt.2018.04.010

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*(11), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Bikangui, R., Boussougou-Sambe, S. T., Saidou, M., Ngossanga, B., Doumba Ndalembouly, A. G., Djida, Y., Ayong More, Beh Mba, R., Abe, H., Ushijima, Y., Borrmann, S., Lell, B., Yasuda, J., & Adegnika, A. A. (2023). Distribution of Aedes mosquito species along the rural–urban gradient in Lambaréné and its surroundings. *Parasites & Vectors*, *16*(1), 360. https://doi.org/10.1186/s13071-023-05901-2

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bonizzoni, M., Gasperi, G., Chen, X., & James, A. A. (2013). The invasive mosquito species Aedes albopictus: Current knowledge and future perspectives. *Trends in Parasitology*, *29*(9), 460–468. https://doi.org/10.1016/j.pt.2013.07.003

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W., & Hay, S. I. (2012). Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. *PLoS Neglected Tropical Diseases*, *6*(8), e1760. https://doi.org/10.1371/journal.pntd.0001760

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *2012 IEEE International Geoscience and Remote Sensing Symposium*, 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393

Broennimann, O., Treier, U. A., Müller-Schärer, H., Thuiller, W., Peterson, A. T., & Guisan, A. (2007). Evidence of climatic niche shift during biological invasion. *Ecology Letters*, *10*(8), 701–709. https://doi.org/10.1111/j.1461-0248.2007.01060.x

Broitman, B., Szathmary, P., Mislan, K., Blanchette, C., & Helmuth, B. (2009). Predator-prey interactions under climate change: The importance of habitat vs body temperature. *Oikos*, *118*, 219–224. https://doi.org/10.1111/j.1600-0706.2008.17075.x

Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, *27*(4), 437–448. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Brown, E. J., Kokkalis, A., & Støttrup, J. G. (2019). Juvenile fish habitat across the inner Danish waters: Habitat association models and habitat growth models for European plaice, flounder and common sole informed by a targeted survey. *Journal of Sea Research*, *155*, 101795. https://doi.org/10.1016/j.seares.2019.101795

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a Metabolic Theory of Ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/10.1890/03-9000

Bryan, T., & Metaxas, A. (2007). Predicting suitable habitat for deep-water gorgonian corals on the Atlantic and Pacific Continental Margins of North America. *Marine Ecology Progress Series*, *330*, 113–126. https://doi.org/10.3354/meps330113

Cai, X., Zhao, J., Deng, H., Xiao, J., Liu, T., Zeng, W., Li, X., Hu, J., Huang, C., Zhu, G., & Ma, W. (2023). Effects of temperature, relative humidity, and illumination on the entomological parameters of Aedes albopictus: An experimental study. *International Journal of Biometeorology*, *67*(4), 687–694. https://doi.org/10.1007/s00484-023-02446-y

Calvez, E., Guillaumot, L., Millet, L., Marie, J., Bossin, H., Rama, V., Faamoe, A., Kilama, S., Teurlai, M., Mathieu-Daudé, F., & Dupont-Rouzeyrol, M. (2016). Genetic Diversity and Phylogeny of Aedes aegypti, the Main Arbovirus Vector in the Pacific. *PLOS Neglected Tropical Diseases*, *10*(1), e0004374. https://doi.org/10.1371/journal.pntd.0004374

Caminade, C., Medlock, J. M., Ducheyne, E., McIntyre, K. M., Leach, S., Baylis, M., & Morse, A. P. (2012). Suitability of European climate for the Asian tiger mosquito Aedes albopictus: Recent trends and future scenarios. *Journal of The Royal Society Interface*, *9*(75), 2708–2717. https://doi.org/10.1098/rsif.2012.0138

Capinha, C., Essl, F., Seebens, H., Moser, D., & Pereira, H. M. (2015). The dispersal of alien species redefines biogeography in the Anthropocene. *Science*, *348*(6240), 1248–1251. https://doi.org/10.1126/science.aaa8913

Carney, R. M., Mapes, C., Low, R. D., Long, A., Bowser, A., Durieux, D., Rivera, K., Dekramanjian, B., Bartumeus, F., Guerrero, D., Seltzer, C. E., Azam, F., Chellappan, S., & Palmer, J. R. B. (2022). Integrating Global Citizen Science Platforms to Enable Next-Generation Surveillance of Invasive and Vector Mosquitoes. *Insects*, *13*(8), Article 8. https://doi.org/10.3390/insects13080675

Ceballos, G., Garcia, A., & Ehrlich, P. (2010). The sixth extinction crisis. *J. Cosmol.*, *8*, 180–185.

Chase, J. M. (2011). 5. Ecological Niche Theory. In S. M. Scheiner & M. R. Willig (Eds.), *The Theory of Ecology* (pp. 93–108). University of Chicago Press. https://doi.org/10.7208/9780226736877-006

Chase, J. M., & Leibold, M. A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo3638660.html

Chen, S., Whiteman, A., Li, A., Rapp, T., Delmelle, E., Chen, G., Brown, C. L., Robinson, P., Coffman, M. J., Janies, D., & Dulin, M. (2019). An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. *Landscape Ecology*, *34*(6), 1295–1311. https://doi.org/10.1007/s10980-019-00839-2

Cianci, D., Hartemink, N., Zeimes, C. B., Vanwambeke, S. O., Ienco, A., & Caputo, B. (2015). High Resolution Spatial Analysis of Habitat Preference of Aedes Albopictus (Diptera: Culicidae) in an Urban Environment. *Journal of Medical Entomology*, *52*(3), 329–335. https://doi.org/10.1093/jme/tjv026

Cook, D., Julias, M., & Nauman, E. (2014). Biological variability in biomechanical engineering research: Significance and meta-analysis of current modeling practices. *Journal of Biomechanics*, *47*(6), 1241–1250. https://doi.org/10.1016/j.jbiomech.2014.01.040

Côté, I. M., & Reynolds, J. D. (2002). Predictive Ecology to the Rescue? *Science*, *298*(5596), 1181–1182. https://doi.org/10.1126/science.1079074

Cox, C. B., Ladle, R. J., & Moore, P. D. (2020). *Biogeography: An ecological and evolutionary approach* (Tenth edition). Wiley.

Crosby, A. D., Bayne, E. M., Cumming, S. G., Schmiegelow, F. K. A., Dénes, F. V., & Tremblay, J. A. (2019). Differential habitat selection in boreal songbirds influences estimates of population size and distribution. *Diversity and Distributions*, *25*(12), 1941–1953. https://doi.org/10.1111/ddi.12991

Cunze, S., Kochmann, J., Koch, L. K., & Klimpel, S. (2016). Aedes albopictus and Its Environmental Limits in Europe. *PLOS ONE*, *11*(9), e0162116. https://doi.org/10.1371/journal.pone.0162116

Da Re, D., Marini, G., Bonannella, C., Laurini, F., Manica, M., Anicic, N., Albieri, A., Angelini, P., Arnoldi, D., Bertola, F., Caputo, B., De Liberato, C., Della Torre, A., Flacio, E., Franceschini, A., Gradoni, F., Kadriaj, P., Lencioni, V., Del Lesto, I., … Rosà, R. (2025). Modelling the seasonal dynamics of Aedes albopictus populations using a spatio-temporal stacked machine learning model. *Scientific Reports*, *15*(1), 3750. https://doi.org/10.1038/s41598-025-87554-y

Delatte, H., Desvars, A., Bouétard, A., Bord, S., Gimonneau, G., Vourc'h, G., & Fontenille, D. (2010). Blood-feeding behavior of Aedes albopictus, a vector of Chikungunya on La Réunion. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *10*(3), 249–258. https://doi.org/10.1089/vbz.2009.0026

Descombes, P., Pitteloud, C., Glauser, G., Defossez, E., Kergunteuil, A., Allard, P.-M., Rasmann, S., & Pellissier, L. (2020). Novel trophic interactions under climate change promote alpine plant coexistence. *Science*, *370*(6523), 1469–1473. https://doi.org/10.1126/science.abd7015

Dieng, H., Rahman, G. M. S., Abu Hassan, A., Che Salmah, M. R., Satho, T., Miake, F., Boots, M., & Sazaly, A. (2012). The effects of simulated rainfall on immature population dynamics of Aedes albopictus and female oviposition. *International Journal of Biometeorology*, *56*(1), 113–120. https://doi.org/10.1007/s00484-011-0402-0

Dowling, Z., Ladeau, S. L., Armbruster, P., Biehler, D., & Leisnham, P. T. (2013). Socioeconomic Status Affects Mosquito (Diptera: Culicidae) Larval Habitat Type Availability and Infestation Level. *Journal of Medical Entomology*, *50*(4), 764–772. https://doi.org/10.1603/ME12250

Dray, S., Chessel, D., & Thioulouse, J. (2003). Co-Inertia Analysis and the Linking of Ecological Data Tables. *Ecology*, *84*(11), 3078–3089. https://doi.org/10.1890/03-0178

Eisen, L., Eisen, R. J., & Lane, R. S. (2006). Geographical Distribution Patterns and Habitat Suitability Models for Presence of Host-Seeking Ixodid Ticks in Dense Woodlands of Mendocino County, California. *Journal of Medical Entomology*, *43*(2), 415–427. https://doi.org/10.1093/jmedent/43.2.415

Elith, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(Volume 40, 2009), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ellenberg, H. (1952). Physiologisches und ökologisches Verhalten derselben Pflanzenarten. *Berichte Der Deutschen Botanischen Gesellschaft*, *65*(10), 350–361. https://doi.org/10.1111/j.1438-8677.1953.tb00671.x

Eritja, R., Delacour-Estrella, S., Ruiz-Arrondo, I., González, M. A., Barceló, C., García-Pérez, A. L., Lucientes, J., Miranda, M. Á., & Bartumeus, F. (2021). At the tip of an iceberg: Citizen science and active surveillance collaborating to broaden the known distribution of Aedes japonicus in Spain. *Parasites & Vectors*, *14*(1), 375. https://doi.org/10.1186/s13071-021-04874-4

Eritja, R., Ruiz-Arrondo, I., Delacour-Estrella, S., Schaffner, F., Álvarez-Chachero, J., Bengoa, M., Puig, M.-Á., Melero-Alcíbar, R., Oltra, A., & Bartumeus, F. (2019). First detection of Aedes japonicus in Spain: An unexpected finding triggered by citizen science. *Parasites & Vectors*, *12*(1), 53. https://doi.org/10.1186/s13071-019-3317-y

Eritja, R., Sanpera-Calbet, I., Delacour-Estrella, S., Ruiz-Arrondo, I., Puig, M. À., Bengoa-Paulís, M., Alarcón-Elbal, P. M., Barceló, C., Mariani, S., Martínez-Barciela, Y., Bravo-Barriga, D., Polina, A., Pereira-Martínez, J. M., González, M. A., Escartin, S., Melero-Alcíbar, R., Blanco-Sierra, L., Magallanes, S., Collantes, F., … Bartumeus, F. (2025). Integrating Citizen Science and Field Sampling into Next-Generation Early-Warning Systems for Vector Surveillance: Twenty Years of Municipal Detections of Aedes Invasive Mosquito Species in Spain. *Insects*, *16*(9), 904. https://doi.org/10.3390/insects16090904

Estrada-Peña, A., & Venzal, J. M. (2007). Climate Niches of Tick Species in the Mediterranean Region: Modeling of Occurrence Data, Distributional Constraints, and Impact of Climate Change. *Journal of Medical Entomology*, *44*(6), 1130–1138. https://doi.org/10.1603/0022-2585(2007)44%255B1130:cnotsi%255D2.0.co;2

Fan, W., Yu, S., & Cosgriff, T. M. (1989). The Reemergence of Dengue in China. *Reviews of Infectious Diseases*, *11*(Supplement_4), S847–S853. https://doi.org/10.1093/clinids/11.Supplement_4.S847

Fischer, D., Thomas, S. M., Niemitz, F., Reineking, B., & Beierkuhnlein, C. (2011). Projection of climatic suitability for Aedes albopictus Skuse (Culicidae) in Europe under climate change conditions. *Global and Planetary Change*, *78*(1–2), 54–64. https://doi.org/10.1016/j.gloplacha.2011.05.008

Fischer, H. S. (1990). Simulating the Distribution of Plant Communities in an Alpine Landscape. *Coenoses*, *5*(1), 37–43.

Fragnière, Y., Gremaud, J., Pesenti, E., Bétrisey, S., Petitpierre, B., Guisan, A., & Kozlowski, G. (2022). Mapping habitats sensitive to overgrazing in the Swiss Northern Alps using habitat suitability modeling. *Biological Conservation*, *274*, 109742. https://doi.org/10.1016/j.biocon.2022.109742

Franklin, J. (1995). Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography: Earth and Environment*, *19*(4), 474–499. https://doi.org/10.1177/030913339501900403

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511810602

Gao, J. (2024). R-Squared (R^2) – How much variation is explained? *Research Methods in Medicine & Health Sciences*, *5*(4), 104–109. https://doi.org/10.1177/26320843231186398

Garamszegi, L. Z., Soltész, Z., Szentiványi, T., Kurucz, K., Nagy, G., & Bede-Fazekas, Á. (2024). Identifying ecological factors mediating the spread of three invasive mosquito species: Citizen science informed prediction. *Journal of Pest Science*. https://doi.org/10.1007/s10340-024-01841-7

Gardner, A. M., Anderson, T. K., Hamer, G. L., Johnson, D. E., Varela, K. E., Walker, E. D., & Ruiz, M. O. (2013). Terrestrial vegetation and aquatic chemistry influence larval mosquito abundance in catch basins, Chicago, USA. *Parasites & Vectors*, *6*(1), 9. https://doi.org/10.1186/1756-3305-6-9

Gaston, K. J. (2003). *The Structure and Dynamics of Geographic Ranges*. Oxford University Press.

Georgiades, P., Proestos, Y., Lelieveld, J., & Erguler, K. (2023). Machine Learning Modeling of Aedes albopictus Habitat Suitability in the 21st Century. *Insects*, *14*(5), 447. https://doi.org/10.3390/insects14050447

Githeko, A. K., Lindsay, S. W., Confalonieri, U. E., & Patz, J. A. (2000). *Climate change and vector-borne diseases: A regional analysis*.

Global invasive species database. (2025). https://www.iucngisd.org

Gould, E. A., Gallian, P., De Lamballerie, X., & Charrel, R. N. (2010). First cases of autochthonous dengue fever and chikungunya fever in France: From bad dream to reality! *Clinical Microbiology and Infection*, *16*(12), 1702–1704. https://doi.org/10.1111/j.1469-0691.2010.03386.x

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*(9), 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Gratz, N. G. (2004). Critical review of the vector status of Aedes albopictus. *Medical and Veterinary Entomology*, *18*(3), 215–227. https://doi.org/10.1111/j.0269-283X.2004.00513.x

Greaves, R. K., Sanderson, R. A., & Rushton, S. P. (2006). Predicting species occurrence using information-theoretic approaches and significance testing: An example of dormouse distribution in Cumbria, UK. *Biological Conservation*, *130*(2), 239–250. https://doi.org/10.1016/j.biocon.2005.12.017

Greenwell, B., M., & Boehmke, B., C. (2020). Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, *12*(1), 343. https://doi.org/10.32614/RJ-2020-013

Greig-Smith, P. (1979). Pattern in Vegetation. *The Journal of Ecology*, *67*(3), 755. https://doi.org/10.2307/2259213

Grinnell, J. (1917). The Niche-Relationships of the California Thrasher. *The Auk*, *34*(4), 427–433. https://doi.org/10.2307/4072271

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, *157*(2–3), 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., & Kueffer, C. (2014). Unifying niche shift studies: Insights from biological invasions. *Trends in Ecology & Evolution*, *29*(5), 260–269. https://doi.org/10.1016/j.tree.2014.02.009

Guisan, A., & Theurillat, J.-P. (2000). Equilibrium modeling of alpine plant distribution: How far can we go? *Phytocoenologia*, *30*(3–4), 353–384. https://doi.org/10.1127/phyto/30/2000/353

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139028271

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). WHAT MATTERS FOR PREDICTING THE OCCURRENCES OF TREES: TECHNIQUES, DATA, OR SPECIES' CHARACTERISTICS? *Ecological Monographs*, *77*(4), 615–630. https://doi.org/10.1890/06-1060.1

Guo, X., Luo, L., Long, Y., Teng, P., Wei, Y., Xie, T., Li, L., Yin, Q., Li, Z., Wang, Y., He, J., Ji, X., Zhou, H., Zhang, X., Chen, S., Zhou, Y., Xu, K., Liang, G., Kuang, H., … Zhou, X. (2023). Field investigation combined with modeling uncovers the ecological heterogeneity of Aedes albopictus habitats for strategically improving systematic management during urbanization. *Parasites & Vectors*, *16*(1), 382. https://doi.org/10.1186/s13071-023-05926-7

Hanski, I. (1998). Metapopulation dynamics. *Nature*, *396*(6706), 41–49. https://doi.org/10.1038/23876

Hanson, J. O., Schuster, R., Strimas-Mackey, M., Morrell, N., Edwards, B. P. M., Arcese, P., Bennett, J. R., & Possingham, H. P. (2025). Systematic conservation prioritization with the prioritizr R package. *Conservation Biology*, *39*(1), e14376. https://doi.org/10.1111/cobi.14376

Hardin, G. (1960). The Competitive Exclusion Principle. *Science*, *131*(3409), 1292–1297. https://doi.org/10.1126/science.131.3409.1292

Hawley, W. A. (1988). The biology of Aedes albopictus. *Journal of the American Mosquito Control Association Supplement*, *1*, 1–39.

Hengl, T., Parente, L., & Bonannella, C. (2022). *Spatial and Spatiotemporal Interpolation / Prediction using Ensemble Machine Learning* (Version v0.1). Zenodo. https://doi.org/10.5281/ZENODO.5894878

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-Niche Factor Analysis: How to Compute Habitat-Suitability Maps Without Absence Data? *Ecology*, *83*(7), 2027–2036. https://doi.org/10.1890/0012-9658(2002)083%255B2027:ENFAHT%255D2.0.CO;2

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hoffer, L. J. (2003). Complementary or alternative medicine: The need for plausibility. *CMAJ*, *168*(2), 180–182.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (1st ed.). Wiley. https://doi.org/10.1002/9781118548387

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

Huang, Y.-Min. (1968). Neotype designation for Aedes (Stegomyid) albopictus (Skuse) (Diptera: Culicidae). *Proceedings of the Entomological Society of Washington*, *70*, 297--302.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.

Hunt, T. N., Allen, S. J., Bejder, L., & Parra, G. J. (2020). Identifying priority habitat for conservation and management of Australian humpback dolphins within a marine protected area. *Scientific Reports*, *10*(1), 14366. https://doi.org/10.1038/s41598-020-69863-6

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*(0), 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Hutchinson, G. E. (1978). *An Introduction to Population Ecology*. Yale University Press.

Hutchinson, G. E. (George E. (with Internet Archive). (1978). *An introduction to population ecology*. New Haven : Yale University Press. http://archive.org/details/introductiontopo0000hutc

Jaureguiberry, P., Titeux, N., Wiemers, M., Bowler, D. E., Coscieme, L., Golden, A. S., Guerra, C. A., Jacob, U., Takahashi, Y., Settele, J., Díaz, S., Molnár, Z., & Purvis, A. (2022). The direct drivers of recent global anthropogenic biodiversity loss. *Science Advances*, *8*(45), eabm9982. https://doi.org/10.1126/sciadv.abm9982

Johnson, C. J., & Gillingham, M. P. (2005). An evaluation of mapped species distribution models used for conservation planning. *Environmental Conservation*, *32*(2), 117–128. https://doi.org/10.1017/S0376892905002171

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*, *19*(2), 101–108. https://doi.org/10.1016/j.tree.2003.10.013

Kearney, M. (2006). Habitat, environment and niche: What are we modelling? *Oikos*, *115*(1), 186–191. https://doi.org/10.1111/j.2006.0030-1299.14908.x

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*(4), 334–350.

Koch, L. K., Cunze, S., Werblow, A., Kochmann, J., Dörge, D. D., Mehlhorn, H., & Klimpel, S. (2016). Modeling the habitat suitability for the arbovirus vector Aedes albopictus (Diptera: Culicidae) in Germany. *Parasitology Research*, *115*(3), 957–964. https://doi.org/10.1007/s00436-015-4822-3

Kraemer, M. U. G., Reiner, R. C., Brady, O. J., Messina, J. P., Gilbert, M., Pigott, D. M., Yi, D., Johnson, K., Earl, L., Marczak, L. B., Shirude, S., Davis Weaver, N., Bisanzio, D., Perkins, T. A., Lai, S., Lu, X., Jones, P., Coelho, G. E., Carvalho, R. G., … Golding, N. (2019). Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nature Microbiology*, *4*(5), 854–863. https://doi.org/10.1038/s41564-019-0376-y

Kraemer, M. U., Sinka, M. E., Duda, K. A., Mylne, A. Q., Shearer, F. M., Barker, C. M., Moore, C. G., Carvalho, R. G., Coelho, G. E., Van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I. R., Teng, H.-J., Brady, O. J., Messina, J. P., Pigott, D. M., Scott, T. W., Smith, D. L., … Hay, S. I. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus. *eLife*, *4*, e08347. https://doi.org/10.7554/eLife.08347

Kuhlisch, C., Kampen, H., & Walther, D. (2018). The Asian tiger mosquito Aedes albopictus (Diptera: Culicidae) in Central Germany: Surveillance in its northernmost distribution area. *Acta Tropica*, *188*, 78–85. https://doi.org/10.1016/j.actatropica.2018.08.019

Lee, S.-J., & Clarke, K. (2005). *An Assessment of Differences in Algorithms For Computing Fundamental Topographic Parameters*.

Li, Y., Kamara, F., Zhou, G., Puthiyakunnon, S., Li, C., Liu, Y., Zhou, Y., Yao, L., Yan, G., & Chen, X.-G. (2014). Urbanization Increases Aedes albopictus Larval Habitats and Accelerates Mosquito Development and Survivorship. *PLOS Neglected Tropical Diseases*, *8*(11), e3301. https://doi.org/10.1371/journal.pntd.0003301

Lira-Noriega, A., & Peterson, A. T. (2014). Range-wide ecological niche comparisons of parasite, hosts and dispersers in a vector-borne plant parasite system. *Journal of Biogeography*, *41*(9), 1664–1673. https://doi.org/10.1111/jbi.12302

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

Lucchesi, L. R., & Wikle, C. K. (2017). Visualizing uncertainty in areal data with bivariate choropleth maps, map pixelation and glyph rotation. *Stat*, *6*(1), 292–302. https://doi.org/10.1002/sta4.150

Manica, M., Filipponi, F., D'Alessandro, A., Screti, A., Neteler, M., Rosà, R., Solimini, A., Della Torre, A., & Caputo, B. (2016). Spatial and Temporal Hot Spots of Aedes albopictus Abundance inside and outside a South European Metropolitan Area. *PLoS Neglected Tropical Diseases*, *10*(6), e0004758. https://doi.org/10.1371/journal.pntd.0004758

Marini, G., Manica, M., Arnoldi, D., Inama, E., Rosà, R., & Rizzoli, A. (2020). Influence of Temperature on the Life-Cycle Dynamics of Aedes albopictus Population Established at Temperate Latitudes: A Laboratory Experiment. *Insects*, *11*(11), Article 11. https://doi.org/10.3390/insects11110808

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*(7), 2372–2387. https://doi.org/10.1016/j.csda.2011.02.004

McGill, B. J. (2010). Matters of Scale. *Science*, *328*(5978), 575–576. https://doi.org/10.1126/science.1188528

McMahon, A., França, C. M. B., & Wimberly, M. C. (2022). Comparing Satellite and Ground-Based Measurements of Environmental Suitability for Vector Mosquitoes in an Urban Landscape. *Journal of Medical Entomology*, *59*(6), 1936–1946. https://doi.org/10.1093/jme/tjac145

Medlock, J. M., Hansford, K. M., Schaffner, F., Versteirt, V., Hendrickx, G., Zeller, H., & Bortel, W. V. (2012). A Review of the Invasive Mosquitoes in Europe: Ecology, Public Health Risks, and Control Options. *Vector-Borne and Zoonotic Diseases*, *12*(6), 435–447. https://doi.org/10.1089/vbz.2011.0814

Medlock, J. M., Hansford, K. M., Versteirt, V., Cull, B., Kampen, H., Fontenille, D., Hendrickx, G., Zeller, H., Van Bortel, W., & Schaffner, F. (2015). An entomological review of invasive mosquitoes in Europe. *Bulletin of Entomological Research*, *105*(6), 637–663. https://doi.org/10.1017/S0007485315000103

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (Second edition). Christoph Molnar.

Mosquito Alert. (2024). *Mosquito Alert Reports* [Dataset]. https://github.com/MosquitoAlert/Data

Moua, Y., Roux, E., Seyler, F., & Briolant, S. (2020). Correcting the effect of sampling bias in species distribution modeling – A new method in the case of a low number of presence data. *Ecological Informatics*, *57*, 101086. https://doi.org/10.1016/j.ecoinf.2020.101086

Murdock, C. C., Evans, M. V., McClanahan, T. D., Miazgowicz, K. L., & Tesla, B. (2017). Fine-scale variation in microclimate across an urban landscape shapes variation in mosquito population dynamics and the potential of Aedes albopictus to transmit arboviral disease. *PLOS Neglected Tropical Diseases*, *11*(5), e0005640. https://doi.org/10.1371/journal.pntd.0005640

Murrell, E. G., Damal, K., Lounibos, L. P., & Juliano, S. A. (2011). Distributions of Competing Container Mosquitoes Depend on Detritus Types, Nutrient Ratios, and Food Availability. *Annals of the Entomological Society of America*, *104*(4), 688–698. https://doi.org/10.1603/AN10158

Neteler, M., Metz, M., Rocchini, D., Rizzoli, A., Flacio, E., Engeler, L., Guidi, V., Lüthy, P., & Tonolla, M. (2013). Is Switzerland Suitable for the Invasion of Aedes albopictus? *PLoS ONE*, *8*(12), e82090. https://doi.org/10.1371/journal.pone.0082090

Niebylski, M. L., Savage, H. M., Nasci, R. S., & Craig, G. B. (1994). Blood hosts of Aedes albopictus in the United States. *Journal of the American Mosquito Control Association*, *10*(3), 447–450.

Njenga, M. K., Nderitu, L., Ledermann, J. P., Ndirangu, A., Logue, C. H., Kelly, C. H. L., Sang, R., Sergon, K., Breiman, R., & Powers, A. M. (2008). Tracking epidemic Chikungunya virus into the Indian Ocean from East Africa. *The Journal of General Virology*, *89*(Pt 11), 2754. https://doi.org/10.1099/vir.0.2008/005413-0

Nordhaus, W. D. (2007). Two Centuries of Productivity Growth in Computing. *The Journal of Economic History*, *67*(1), 128–159. https://doi.org/10.1017/S0022050707000058

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). *Niche Construction*. Princeton University Press; JSTOR. http://www.jstor.org/stable/j.ctt24hqpd

O'Neill, D., Häkkinen, H., Neumann, J., Shaffrey, L., Cheffings, C., Norris, K., & Pettorelli, N. (2023). Investigating the potential of social media and citizen science data to track changes in species' distributions. *Ecology and Evolution*, *13*(5), e10063. https://doi.org/10.1002/ece3.10063

Palmer, J. R. B., Oltra, A., Collantes, F., Delgado, J. A., Lucientes, J., Delacour, S., Bengoa, M., Eritja, R., & Bartumeus, F. (2017). Citizen science provides a reliable and scalable tool to track disease-carrying mosquites. *Nature Communications*, *8*(1), 916. https://doi.org/10.1038/s41467-017-00914-9

Paupy, C., Delatte, H., Bagny, L., Corbel, V., & Fontenille, D. (2009). Aedes albopictus, an arbovirus vector: From the darkness to the light. *Microbes and Infection*, *11*(14–15), 1177–1185. https://doi.org/10.1016/j.micinf.2009.05.005

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*(5), 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, *7*, e6876. https://doi.org/10.7717/peerj.6876

Pellmyr, O., & Leebens-Mack, J. (1999). Forty million years of mutualism: Evidence for eocene origin of the yucca-yucca moth association. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(16), 9178–9183. https://doi.org/10.1073/pnas.96.16.9178

Peterson, A. T. (2003). Predicting the Geography of Species' Invasions via Ecological Niche Modeling. *The Quarterly Review of Biology*, *78*(4), 419–433. https://doi.org/10.1086/378926

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Poh, K. C., Chaves, L. F., Reyna-Nava, M., Roberts, C. M., Fredregill, C., Bueno, R., Debboun, M., & Hamer, G. L. (2019). The influence of weather and weather variability on mosquito abundance and infection with West Nile virus in Harris County, Texas, USA. *Science of The Total Environment*, *675*, 260–272. https://doi.org/10.1016/j.scitotenv.2019.04.109

Reichl, J., Prossegger, C., Eichholzer, B., Plauder, P., Unterköfler, M. S., Bakran-Lebl, K., Indra, A., & Fuehrer, H.-P. (2023). A citizen science report—Tiger mosquitoes (Aedes albopictus) in allotment gardens in Graz, Styria, Austria. *Parasitology Research*, *123*(1), 79. https://doi.org/10.1007/s00436-023-08106-9

Rezza, G., Nicoletti, L., Angelini, R., Romi, R., Finarelli, A. C., Panning, M., Cordioli, P., Fortuna, C., Boros, S., Magurano, F., Silvi, G., Angelini, P., Dottori, M., Ciufolini, M. G., Majori, G. C., & Cassone, A. (2007). Infection with chikungunya virus in Italy: An outbreak in a temperate region. *The Lancet*, *370*(9602), 1840–1846. https://doi.org/10.1016/S0140-6736(07)61779-6

Richards, S. L., Ponnusamy, L., Unnasch, T. R., Hassan, H. K., & Apperson, C. S. (2006). Host-feeding patterns of Aedes albopictus (Diptera: Culicidae) in relation to availability of human and domestic animals in suburban landscapes of central North Carolina. *Journal of Medical Entomology*, *43*(3), 543–551. https://doi.org/10.1603/0022-2585(2006)43%255B543:hpoaad%255D2.0.co;2

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Roques, L., Boivin, T., Papaïx, J., Soubeyrand, S., & Bonnefon, O. (2023). Dynamics of Aedes albopictus invasion insights from a spatio-temporal model. *Biological Invasions*, *25*(8), 2679–2695. https://doi.org/10.1007/s10530-023-03062-y

Sabatini, A., Raineri, V., Trovato, G., & Coluzzi, M. (1990). [Aedes albopictus in Italy and possible diffusion of the species into the Mediterranean area]. *Parassitologia*, *32*(3), 301–304.

Santos, X., Brito, J., Sillero, N., Pleguezuelos, J., Llorente, G., FAHD, S., & Parellada, X. (2006). Inferring Habitat-suitability Areas with Ecological Modelling Techniques and GIS: A Contribution to Assess the Conservation Status of Vipera latastei. *Biological Conservation*, *130*, 416–425. https://doi.org/10.1016/j.biocon.2006.01.003

Sarà, M. (2008). Breeding abundance of threatened raptors as estimated from occurrence data. *Ibis*, *150*(4), 766–778. https://doi.org/10.1111/j.1474-919X.2008.00856.x

Sato, A., Tichy, H., O'hUigin, C., Grant, P. R., Grant, B. R., & Klein, J. (2001). *On the Origin of Darwin's Finches*.

Scholte, E.-J., & Schaffner, F. (2007). Waiting for the tiger: Establishment and spread of the Aedes albopictus mosquito in Europe. In W. Takken & B. G. J. Knols (Eds.), *Emerging pests and vector-borne diseases in Europe* (pp. 241–260). Brill | Wageningen Academic. https://doi.org/10.3920/9789086866267_016

Schwartz, M. W. (2012). Using niche models with climate projections to inform conservation management decisions. *Biological Conservation*, *155*, 149–156. https://doi.org/10.1016/j.biocon.2012.06.011

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

Seidel, B., Duh, D., Nowotny, N., & Allerberger, F. (2012). Erstnachweis der Stechmücken Aedes (Ochlerotatus) japonicus japonicus (Theobald, 1901) in Österreich und Slowenien in 2011 und für Aedes (Stegomyia) albopictus (Skuse, 1895) in Österreich 2012 (Diptera: Culicidae). *Entomol Z*, *122*, 223–226.

Service, M. W. (1971). The Daytime Distribution of Mosquitoes Resting in Vegetation. *Journal of Medical Entomology*, *8*(3), 271–278. https://doi.org/10.1093/jmedent/8.3.271

Service, M. W. (1980). Effects of wind on the behaviour and distribution of mosquitoes and blackflies. *International Journal of Biometeorology*, *24*(4), 347–353. https://doi.org/10.1007/BF02250577

Shragai, T., & Harrington, L. C. (2019). Aedes albopictus (Diptera: Culicidae) on an Invasive Edge: Abundance, Spatial Distribution, and Habitat Usage of Larvae and Pupae Across Urban and Socioeconomic Environmental Gradients. *Journal of Medical Entomology*, *56*(2), 472–482. https://doi.org/10.1093/jme/tjy209

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, *10*(12), 1115–1123. https://doi.org/10.1111/j.1461-0248.2007.01107.x

Soberón, J., & Peterson, A. T. (2005). Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, *2*. https://doi.org/10.17161/bi.v2i0.4

Sørensen, R., Zinko, U., & Seibert, J. (2006). On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrology and Earth System Sciences*, *10*(1), 101–112. https://doi.org/10.5194/hess-10-101-2006

Soti, V., Tran, A., Degenne, P., Chevalier, V., Lo Seen, D., Thiongane, Y., Diallo, M., Guégan, J.-F., & Fontenille, D. (2012). Combining Hydrology and Mosquito Population Models to Identify the Drivers of Rift Valley Fever Emergence in Semi-Arid Regions of West Africa. *PLoS Neglected Tropical Diseases*, *6*(8), e1795. https://doi.org/10.1371/journal.pntd.0001795

Steger, S., Brenning, A., Bell, R., Petschko, H., & Glade, T. (2016). Exploring discrepancies between quantitative validation results and the geomorphic plausibility of statistical landslide susceptibility maps. *Geomorphology*, *262*, 8–23. https://doi.org/10.1016/j.geomorph.2016.03.015

Steger, S., & Glade, T. (2017). The Challenge of "Trivial Areas" in Statistical Landslide Susceptibility Modelling. In M. Mikos, B. Tiwari, Y. Yin, & K. Sassa (Eds.), *Advancing Culture of Living with Landslides* (pp. 803–808). Springer International Publishing. https://doi.org/10.1007/978-3-319-53498-5_92

Steger, S., Mair, V., Kofler, C., Pittore, M., Zebisch, M., & Schneiderbauer, S. (2021). Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling – Benefits of exploring landslide data collection effects. *Science of The Total Environment*, *776*, 145935. https://doi.org/10.1016/j.scitotenv.2021.145935

Steger, S., Moreno, M., Crespi, A., Zellner, P. J., Gariano, S. L., Brunetti, M. T., Melillo, M., Peruccacci, S., Marra, F., Kohrs, R., Goetz, J., Mair, V., & Pittore, M. (2023). Deciphering seasonal effects of triggering and preparatory precipitation for improved shallow landslide prediction using generalized additive mixed models. *Natural Hazards and Earth System Sciences*, *23*(4), 1483–1506. https://doi.org/10.5194/nhess-23-1483-2023

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tabachnick, W. (1991). Evolutionary Genetics and Arthropod-borne Disease: The Yellow Fever Mosquito. *American Entomologist*, *37*, 14–26. https://doi.org/10.1093/ae/37.1.14

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C., & Guisan, A. (2014). Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, *5*(9), 947–955. https://doi.org/10.1111/2041-210X.12203

Thomas, S. M., Obermayr, U., Fischer, D., Kreyling, J., & Beierkuhnlein, C. (2012). Low-temperature threshold for egg survival of a post-diapause and non-diapause European aedine strain, Aedes albopictus (Diptera: Culicidae). *Parasites & Vectors*, *5*(1), 100. https://doi.org/10.1186/1756-3305-5-100

Thomas, S. M., Tjaden, N. B., Frank, C., Jaeschke, A., Zipfel, L., Wagner-Wiening, C., Faber, M., Beierkuhnlein, C., & Stark, K. (2018). Areas with High Hazard Potential for Autochthonous Transmission of Aedes albopictus-Associated Arboviruses in Germany. *International Journal of Environmental Research and Public Health*, *15*(6), Article 6. https://doi.org/10.3390/ijerph15061270

Torina, A., La Russa, F., Blanda, V., Peralbo-Moreno, A., Casades-Martí, L., Di Pasquale, L., Bongiorno, C., Vitale Badaco, V., Toma, L., & Ruiz-Fons, F. (2023). Modelling time-series *Aedes albopictus* abundance as a forecasting tool in urban environments. *Ecological Indicators*, *150*, 110232. https://doi.org/10.1016/j.ecolind.2023.110232

Tran, A., L'Ambert, G., Lacour, G., Benoît, R., Demarchi, M., Cros, M., Cailly, P., Aubry-Kientz, M., Balenghien, T., & Ezanno, P. (2013). A Rainfall- and Temperature-Driven Abundance Model for Aedes albopictus Populations. *International Journal of Environmental Research and Public Health*, *10*(5), Article 5. https://doi.org/10.3390/ijerph10051698

Underhill, L., & Gibbons, D. (2002). Mapping and monitoring bird populations: Their conservation uses. In D. J. Pain & K. Norris (Eds.), *Conserving Bird Biodiversity: General Principles and their Application* (pp. 34–60). Cambridge University Press. https://doi.org/10.1017/CBO9780511606304.004

Unlu, I., Farajollahi, A., Healy, S. P., Crepeau, T., Bartlett-Healy, K., Williges, E., Strickman, D., Clark, G. G., Gaugler, R., & Fonseca, D. M. (2011). Area-wide management of Aedes albopictus: Choice of study sites based on geospatial characteristics, socioeconomic factors and mosquito populations. *Pest Management Science*, *67*(8), 965–974. https://doi.org/10.1002/ps.2140

Van Der Wal, R., Truscott, A., Pearce, I. S. K., Cole, L., Harris, M. P., & Wanless, S. (2008). Multiple anthropogenic changes cause biodiversity loss through plant invasion. *Global Change Biology*, *14*(6), 1428–1436. https://doi.org/10.1111/j.1365-2486.2008.01576.x

Vavassori, L., Saddler, A., & Müller, P. (2019). Active dispersal of Aedes albopictus: A mark-release-recapture study using self-marking units. *Parasites & Vectors*, *12*(1), 583. https://doi.org/10.1186/s13071-019-3837-5

Westby, K. M., Adalsteinsson, S. A., Biro, E. G., Beckermann, A. J., & Medley, K. A. (2021). Aedes albopictus Populations and Larval Habitat Characteristics across the Landscape: Significant Differences Exist between Urban and Rural Land Use Types. *Insects*, *12*(3), 196. https://doi.org/10.3390/insects12030196

Whittaker, R. H. (1967). GRADIENT ANALYSIS OF VEGETATION*. *Biological Reviews*, *42*(2), 207–264. https://doi.org/10.1111/j.1469-185X.1967.tb01419.x

Wiens, J. A. (1989). Spatial Scaling in Ecology. *Functional Ecology*, *3*(4), 385–397. https://doi.org/10.2307/2389612

Wiens, J. J., & Graham, C. H. (2005). Niche Conservatism: Integrating Evolution, Ecology, and Conservation Biology. *Annual Review of Ecology, Evolution, and Systematics*, *36*(Volume 36, 2005), 519–539. https://doi.org/10.1146/annurev.ecolsys.36.102803.095431

Wood, S. N. (2003). Thin Plate Regression Splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Woodward, F. I., & Kelly, C. K. (2003). Why are species not more widely distributed? Physiological and environmental limits. In *T. Blackburn / K. Gaston (publ.): Macroecology: Concepts and Consequences* (pp. 239–255). Oxford.

Wu, J.-Y., Lun, Z.-R., James, A. A., & Chen, X.-G. (2010). *Dengue Fever in Mainland China*. https://doi.org/10.4269/ajtmh.2010.09-0755

Zahouli, J. B. Z., Koudou, B. G., Müller, P., Malone, D., Tano, Y., & Utzinger, J. (2017). Urbanization is a main driver for the larval ecology of Aedes mosquitoes in arbovirus-endemic settings in south-eastern Côte d'Ivoire. *PLOS Neglected Tropical Diseases*, *11*(7), e0005751. https://doi.org/10.1371/journal.pntd.0005751

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer New York. https://doi.org/10.1007/978-0-387-87458-6

Appendix

This appendix contains the R code used for data preparation of spatial and temporal predictor data, the sampling of pseudo-absences, the fitting and validation of the spatial and temporal models, and finally the visualization and classification of results. Codes are accessible at https://github.com/Digital-Geography/Spatiotemporal_presence_of_Aedes_albopictus.

Spatial data preparation

```
rm(list = ls(all.names = TRUE))

gc()

library(sf) #simple features

library(dplyr) #For mutating dfs

library(terra) #Raster analysis

library(readxl) #Read excel files

library(ggplot2) # For plotting


#Load Convex Hull (area of distribution (from QGIS) as Sampling Area) and boundaries

convex_hull_graz <- st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Area_Of_Equilibrium/Convex_Hull_31287.shp")

city_bound_graz <- st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Stadtgrenze/Stadtgrenze_31287.shp")

districts_graz <- st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Stadtgrenze/Bezirke.geojson")


#Clean up districts graz

districts_graz <- districts_graz[1:17, ] %>% dplyr::select(name, geometry)


#Order df according to numbering of districts and create attribute district_id

districts_graz <- districts_graz[c(1, 3, 4, 5, 6, 7, 2, 8:17), ]
districts_graz <- districts_graz %>% mutate(
  district_id = row_number()
)
#Aggregate districts to clusters for spatial CV
districts_graz <- districts_graz %>% mutate(
  cluster = case_when(
    district_id %in% c(1, 2, 3, 6) ~ 1,
    district_id %in% c(4, 5) ~ 2,
    district_id %in% c(7, 8, 9) ~ 3,
    district_id %in% c(14, 15, 16, 17) ~ 4,
    district_id %in% c(10, 11, 12, 13) ~ 5,
  )
)

#Reproject layer
districts_graz <- st_transform(districts_graz, crs = 31287)

#Plot clusters
ggplot(data = districts_graz) +
  geom_sf(aes(fill = factor(cluster)), color = "black") +
  scale_fill_viridis_d(name = "Area") +
  theme_minimal() +
  labs(title = "Areas for spatial partioning of data",
       x = "Longitude",
       y = "Latitude") +
  theme(legend.position = "bottom")

ggsave(plot = last_plot(), "D:/Masterarbeit/Figures/Clusters_CV.png")

# Load Raster, that was created in QGIS, and rename bands
    # For Graz
    modeling_raster_graz<-
rast("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Modeling_
Raster_100m.tif")

    new_band_names <- c("LC_Agriculture", "LC_Construction",
"LC_Green_Space","LC_OpenSoil","LC_RoofSealing",

                        "LC_Roof_50", "LC_Roof_51",
"LC_Roof_52","LC_Sealed_Ground","LC_Vegetation_High","LC_Vegetation_
Low","LC_Water", "LU_Class50", "LU_Class51",
"LU_Class52","LU_Majority_Class", "LU_Majority_Use",


"LU_Allot_Grave","Population_Density","Presence","Rain_Inlet","TWI",
"Sampling_Effort")

    names(modeling_raster_graz) <- new_band_names

    writeRaster(modeling_raster_graz,

                filename =
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/modeling_raste
r_graz_pp.tif",

                overwrite = TRUE)


    plot(modeling_raster_graz)




#Raster to dataframe

    df_pred_graz <- as.data.frame(modeling_raster_graz, xy = TRUE,
na.rm = TRUE) #resulting point corrdinates represent pixel centroids


### Regrouping of Land Use categories ###

    #Graz: define subsets to be grouped

    LU_10 <- c(14, 3, 10, 18, 19, 5, 29, 20)

    LU_11 <- c(21, 9)

    LU_20 <- c(28)

    LU_30 <- c(8, 25)

    LU_40 <- c(2, 4, 26)

    LU_50 <- c(12, 23)
    LU_51 <- c(6, 11)
    LU_52 <- c(16, 17, 13)
    LU_60 <- c(1, 7, 15, 22, 24, 27)

    #Reclassify land use classes
    df_pred_graz <- df_pred_graz %>%
      mutate(LU_Majority_Class = case_when(
        LU_Majority_Use %in% LU_10 ~ 10,
        LU_Majority_Use %in% LU_11 ~ 11,
        LU_Majority_Use %in% LU_20 ~ 20,
        LU_Majority_Use %in% LU_30 ~ 30,
        LU_Majority_Use %in% LU_40 ~ 40,
        LU_Majority_Use %in% LU_50 ~ 50,
        LU_Majority_Use %in% LU_51 ~ 51,
        LU_Majority_Use %in% LU_52 ~ 52,
        LU_Majority_Use %in% LU_60 ~ 60,
      ))

    #Proportion of each land use class
    prop.table(table(df_pred_graz$LU_Majority_Class))

### Regrouping of Land Cover Attributes ###
df_pred_graz <- df_pred_graz %>%
  mutate(LC_Soil = LC_OpenSoil + LC_Construction + LC_Agriculture) %>%
    dplyr::select(-LC_OpenSoil, -LC_Construction, -LC_Agriculture)

#Transform presence variable and LU_Majority_Use to factor
df_pred_graz$Presence <- as.factor(df_pred_graz$Presence)
df_pred_graz$LU_Majority_Class                                    <-
as.factor(df_pred_graz$LU_Majority_Class)
```

```
#Remove unwanted attributes
df_pred_graz$LU_Majority_Use <- NULL
df_pred_graz$LU_Class51 <- NULL
df_pred_graz$LU_Class52 <- NULL
df_pred_graz$LU_Class50 <- NULL
df_pred_graz$LC_RoofSealing <- NULL

#Clip data to city boundaries
sf_pred_graz <- st_as_sf(df_pred_graz, coords = c("x", "y"), crs =
31287)
sf_pred_graz <- st_intersection(sf_pred_graz, city_bound_graz)
sf_pred_graz <- sf_pred_graz %>% dplyr::select(-GEMEINDE, -area)

# Remove water pixels
sf_pred_graz <- sf_pred_graz[sf_pred_graz$LU_Majority_Class != 30, ]

#Add spatial cluster to sf_pred_graz. For spatial cross-validation,
therefore only for Graz
sf_pred_graz <- st_join(sf_pred_graz, districts_graz[, c("geometry",
"cluster")])

#Divide by 100 so LC and LU variables represent the percentage of
pixel being covered
    sf_pred_graz$LC_Green_Space <- sf_pred_graz$LC_Green_Space/100
    sf_pred_graz$LC_Roof_50 <- sf_pred_graz$LC_Roof_50/100
    sf_pred_graz$LC_Roof_51 <- sf_pred_graz$LC_Roof_51/100
    sf_pred_graz$LC_Roof_52 <- sf_pred_graz$LC_Roof_52/100
    sf_pred_graz$LC_Sealed_Ground                              <-
sf_pred_graz$LC_Sealed_Ground/100
    sf_pred_graz$LC_Vegetation_High                            <-
sf_pred_graz$LC_Vegetation_High/100
    sf_pred_graz$LC_Vegetation_Low                             <-
sf_pred_graz$LC_Vegetation_Low/100
```

```
    sf_pred_graz$LC_Water <- sf_pred_graz$LC_Water/100

    sf_pred_graz$LU_Allot_Grave <- sf_pred_graz$LU_Allot_Grave/100

    sf_pred_graz$LC_Soil <- sf_pred_graz$LC_Soil/100


#Create training set by clipping with area of mosquito spread

    sf_train_graz <- st_intersection(sf_pred_graz, convex_hull_graz)

    sf_train_graz  <-  sf_train_graz  %>%  dplyr::select(-id,  -area,  -
perimeter)


# Plot data frames

    plot(sf_pred_graz)

    plot(sf_train_graz)


#Save dfs

    st_write(sf_train_graz,
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postprocessed/
sf_train_graz.gpkg",  layer = "layer_name", append = FALSE)

    st_write(sf_pred_graz,
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postprocessed/
sf_pred_graz.gpkg",  layer = "layer_name", append = FALSE)


#Statsitics SE

    mean(sf_pred_graz$Sampling_Effort, na.rm = TRUE)

    sd(sf_pred_graz$Sampling_Effort, na.rm = TRUE)
```

Temporal data preparation

```
rm(list = ls(all.names = TRUE))

gc()

library(sf) #simple features

library(dplyr) #For mutating dfs

library(ggplot2)  #Plotting histograms

library(gridExtra) #For subplotting
library(lubridate) #Date transformations

library(tidyr) #data wrangling

library(raster) #Claculating CV Score

library(zoo) #For rollapply()

library(corrplot) #Plotting Correlation Matrix

library(Hmisc) #Calculating correlation matrix

library(MASS) #For negative binomial distribution

library(mgcv) # For gam()

library(lme4) # For glmm

library(patchwork) #stacking subplots


#Load Weather data

#(t = Lufttemp [°C], rh = rel. Feuchte [%], ffx = max.
Windgeschwindigkeit [m/s], ws_mean = mittere Windspeed [m/s],

                    #p = NS 24h [mm],  so_h = Sonnenscheindauer [h],
cglo_j = Globalstrahlung [J/cm2])

  #daily

  Meteo_d <-
read.csv("D:/Masterarbeit/Jupyter/Data/Meteo/Messstationen
Tagesdaten v2 Datensatz_20210101_20241231.csv")

  #hourly (wind only)

  Meteo_h <-
read.csv("D:/Masterarbeit/Jupyter/Data/Meteo/Messstationen
Stundendaten v2 Datensatz_20210101T0000_20241231T0000.csv")


#Load validated Mosquito Sightings from Mosquito Alert for Graz (after
preprocessing in MosAl_explo_prepro.ipynb)

Mos <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Mosquito/Mosquitos_val1_3
1287.shp")


#Load daily Sampling_Effort (after MosAl_explo_prepro.ipynb)

Sampling_Effort <- read.csv("D:/Masterarbeit/Jupyter/Data/Sampling
Effort/Postprocessed/SE_pp.csv")
```

```
Sampling_Effort$date <- as.Date(Sampling_Effort$date)

Sampling_Effort$X <- NULL



#Manipulate Meteo data

  #Remove station 16401 as there is no data

  Meteo_d <- Meteo_d[Meteo_d$station != 16401,]

  #Remove _flag attributes

  Meteo_d <- Meteo_d %>% dplyr::select(-ends_with("_flag"))

  #Transform time to date

  Meteo_d <- Meteo_d %>%

    mutate(date = as.Date(sub("T.*","",time))) %>%

    dplyr::select(-time)

  #Remove 2021

  Meteo_d$year <- year(Meteo_d$date)

  Meteo_d   <-   Meteo_d   %>%   filter(Meteo_d$year   !=   2021)   %>%
dplyr::select(-"year")

  #Drop unwanted columns

  Meteo_d <- Meteo_d %>%

     dplyr::select(-"ffx", -"cglo_j")

  #Rename columns

  col_names  <-  c("station","t_max",  "t_min",  "t_mean",  "rh_mean",
"ws_mean", "p", "so_h", "date")

  colnames(Meteo_d) <- col_names


  #Replace negative rain values (No precip) with 0

  Meteo_d <- Meteo_d %>%

    mutate(

      p = ifelse(p < 0, 0, p)

    )

    #Calcualte daily means from both stations
```

```
 Meteo_d <- Meteo_d %>%
       group_by(date)%>%
        summarise(
         t_max = mean(t_max, na.rm = TRUE),
         t_min = mean(t_min, na.rm = TRUE),
         t_mean = mean(t_mean, na.rm = TRUE),
         rh_mean = mean(rh_mean, na.rm = TRUE),
         ws_mean = mean(ws_mean, na.rm = TRUE),
         p = mean(p, na.rm = TRUE),
         so_h = mean(so_h, na.rm = TRUE),
        )

#Manipulation of Mosquito data
 Mos <- Mos %>%
  mutate(creation_d = ymd(creation_d))

 #Clean up df
  Mos <- Mos %>% dplyr::select("creation_d","creation_y",
"creation_m")
 #Remove 2021 due to sparsity of data
  Mos_22_24 <- Mos %>%
   filter(creation_y != 2021)

 # Prepare daily counts of mosquitoes and create month and year
attribute
  #Create sequence of days
   date_range <- seq(min(Meteo_d$date), max(Meteo_d$date), by =
"day")
  #Df from sequence
  full_dates <- data.frame(creation_d = date_range)
```

```
daily_data <- Mos_22_24 %>%
  count(creation_d) %>% #aggregate daily counts
  right_join(full_dates, by = "creation_d") %>% # Right join to keep all days
  mutate(
    creation_m = month(creation_d),    # create month and year attribute
    creation_y = year(creation_d))%>%
  rename(count = n, date = creation_d)%>% #rename columns
  replace_na(list(count = 0)) #replace NAN with 0


#drop geometry column
daily_data <- st_drop_geometry(daily_data)
daily_data <- daily_data %>% filter(daily_data$creation_y != 2021)


# transfrom factor for second y-axis
transform_factor <- max(daily_data$count)/max(Sampling_Effort$SE)


#Plot SE and daily counts for 2022-2024
ggplot(daily_data, aes(x = date, y = count)) +
  geom_col(fill = "orange") +
  geom_line(data = Sampling_Effort, aes(x = date, y = SE * transform_factor), color = "blue", alpha = 0.4) +
  labs(title = "Daily Mosquito Sightings between 2022 and 2024", x = "Year", y = "Daily observations") +
  theme_minimal() +
  scale_y_continuous(
    sec.axis = sec_axis(~ . / transform_factor, name = "Sampling_Effort")
     ) +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y", date_minor_breaks = "1 month")
```

```
    #Save plot
    ggsave(plot = last_plot(), filename = "D:/Masterarbeit/Figures/Daily_sightings_22_24.png")

#Calculate mean and variation of hourly windspeed
    #Hourly means between stations
    daily_ws_mean <- Meteo_h %>%
      group_by(time) %>% # Group by hour
      summarise(
        ws_mean = mean(ff, na.rm = TRUE)
      ) %>%
      mutate(
        date = date(time)
      )

    #CV Score per day
    daily_ws_cv <- daily_ws_mean %>%
      group_by(date) %>%
      summarise(
        ws_cv = raster::cv(ws_mean)
      )
   #Join daily df with ws_CV
   Meteo_d <- left_join(Meteo_d, daily_ws_cv, by = "date")

   # Claculate Meteo parameters across 14 and 28 days in advance
    # Compute rolling mean, standard deviation and coefficient of variation for each parameter
   Meteo_d <- Meteo_d %>%
     arrange(date) %>%
     mutate(
       # 14-days time-lagged
       t_max_mean_14 = rollapply(t_max, width = 14, FUN = mean, fill
= NA, align = "right"),

       t_max_sd_14 = rollapply(t_max, width = 14, FUN = sd, fill = NA,
align = "right"),

       t_max_cv_14 = ifelse(t_max_mean_14 != 0, (t_max_sd_14 /
t_max_mean_14) * 100, NA),


       t_min_mean_14 = rollapply(t_min, width = 14, FUN = mean, fill
= NA, align = "right"),

       t_min_sd_14 = rollapply(t_min, width = 14, FUN = sd, fill = NA,
align = "right"),

       t_min_cv_14 = ifelse(t_min_mean_14 != 0, (t_min_sd_14 /
t_min_mean_14) * 100, NA),


       t_mean_mean_14 = rollapply(t_mean, width = 14, FUN = mean, fill
= NA, align = "right"),

       t_mean_sd_14 = rollapply(t_mean, width = 14, FUN = sd, fill =
NA, align = "right"),

       t_mean_cv_14 = ifelse(t_mean_mean_14 != 0, (t_mean_sd_14 /
t_mean_mean_14) * 100, NA),


       rh_mean_mean_14 = rollapply(rh_mean, width = 14, FUN = mean,
fill = NA, align = "right"),

       rh_mean_sd_14 = rollapply(rh_mean, width = 14, FUN = sd, fill
= NA, align = "right"),

       rh_mean_cv_14 = ifelse(rh_mean_mean_14 != 0, (rh_mean_sd_14 /
rh_mean_mean_14) * 100, NA),


       ws_mean_mean_14 = rollapply(ws_mean, width = 14, FUN = mean,
fill = NA, align = "right"),

       ws_mean_sd_14 = rollapply(ws_mean, width = 14, FUN = sd, fill
= NA, align = "right"),

       ws_mean_cv_14 = ifelse(ws_mean_mean_14 != 0, (ws_mean_sd_14 /
ws_mean_mean_14) * 100, NA),
      p_mean_14 = rollapply(p, width = 14, FUN = mean, fill = NA,
align = "right"),
      p_sd_14 = rollapply(p, width = 14, FUN = sd, fill = NA, align
= "right"),
      p_cv_14 = ifelse(p_mean_14 != 0, (p_sd_14 / p_mean_14) * 100,
NA),
      p_acc_14 = rollapply(p, width = 14, FUN = sum, fill = NA, align
= "right"),

      so_h_mean_14 = rollapply(so_h, width = 14, FUN = mean, fill =
NA, align = "right"),
      so_h_sd_14 = rollapply(so_h, width = 14, FUN = sd, fill = NA,
align = "right"),
      so_h_cv_14 = ifelse(so_h_mean_14 != 0, (so_h_sd_14 /
so_h_mean_14) * 100, NA),
      )

  # 28-days time-lagged
  Meteo_d <- Meteo_d %>%
    arrange(date) %>%
    mutate(
      t_max_mean_28 = rollapply(t_max, width = 28, FUN = mean, fill
= NA, align = "right"),
      t_max_sd_28 = rollapply(t_max, width = 28, FUN = sd, fill = NA,
align = "right"),
      t_max_cv_28 = ifelse(t_max_mean_28 != 0, (t_max_sd_28 /
t_max_mean_28) * 100, NA),

      t_min_mean_28 = rollapply(t_min, width = 28, FUN = mean, fill
= NA, align = "right"),
      t_min_sd_28 = rollapply(t_min, width = 28, FUN = sd, fill = NA,
align = "right"),
      t_min_cv_28 = ifelse(t_min_mean_28 != 0, (t_min_sd_28 /
t_min_mean_28) * 100, NA),
       t_mean_mean_28 = rollapply(t_mean, width = 28, FUN = mean, fill
= NA, align = "right"),

       t_mean_sd_28 = rollapply(t_mean, width = 28, FUN = sd, fill =
NA, align = "right"),

       t_mean_cv_28 = ifelse(t_mean_mean_28 != 0, (t_mean_sd_28 /
t_mean_mean_28) * 100, NA),


       rh_mean_mean_28 = rollapply(rh_mean, width = 28, FUN = mean,
fill = NA, align = "right"),

       rh_mean_sd_28 = rollapply(rh_mean, width = 28, FUN = sd, fill
= NA, align = "right"),

       rh_mean_cv_28 = ifelse(rh_mean_mean_28 != 0, (rh_mean_sd_28 /
rh_mean_mean_28) * 100, NA),


       ws_mean_mean_28 = rollapply(ws_mean, width = 28, FUN = mean,
fill = NA, align = "right"),

       ws_mean_sd_28 = rollapply(ws_mean, width = 28, FUN = sd, fill
= NA, align = "right"),

       ws_mean_cv_28 = ifelse(ws_mean_mean_28 != 0, (ws_mean_sd_28 /
ws_mean_mean_28) * 100, NA),


       p_mean_28 = rollapply(p, width = 28, FUN = mean, fill = NA,
align = "right"),

       p_sd_28 = rollapply(p, width = 28, FUN = sd, fill = NA, align
= "right"),

       p_cv_28 = ifelse(p_mean_28 != 0, (p_sd_28 / p_mean_28) * 100,
NA),

       p_acc_28 = rollapply(p, width = 28, FUN = sum, fill = NA, align
= "right"),


       so_h_mean_28 = rollapply(so_h, width = 28, FUN = mean, fill =
NA, align = "right"),

       so_h_sd_28 = rollapply(so_h, width = 28, FUN = sd, fill = NA,
align = "right"),

       so_h_cv_28 = ifelse(so_h_mean_28 != 0, (so_h_sd_28 /
so_h_mean_28) * 100, NA),
      )
 #Clean df
   Meteo_d <- Meteo_d %>%
     dplyr::select(-contains("_sd"), -"p_mean_28", -"p_mean_14")

 #Join Mos_Count with Meteo_d and sampling effort
   df_train <- left_join(daily_data, Meteo_d, by = "date")
   df_train <- left_join(df_train, Sampling_Effort, by = "date")
   #Create day of the year attribute
   df_train$doy <- yday(df_train$date)



#Save Meteo_data
   write.csv(Meteo_d,
"D:/Masterarbeit/Jupyter/Data/Meteo/Postprocessed/Meteo_d.csv",
row.names = FALSE)
#Save df_train_meteo_MA
   write.csv(df_train,
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/Meteo/df_train_mete
o_MA.csv",
             row.names = FALSE)

###########################
### Correlation matrix ####
###########################

   vars_eda <- c(
     "SE",
     "t_mean",
     "t_max",
     "t_min",
     "rh_mean",
     "ws_mean",
     "ws_cv",
     "p",
     "so_h",
     "so_h_mean_28",
     "so_h_mean_14",
     "t_mean_mean_28",
     "t_mean_mean_14",
     "t_max_mean_28",
     "t_max_mean_14",
     "t_min_mean_28",
     "t_min_mean_14",
     "p_acc_28",
     "p_acc_14",
     "ws_mean_mean_28",
     "ws_mean_mean_14",
     "rh_mean_mean_28",
     "rh_mean_mean_14"
   )

    ## Poisson or negative binomial distribution?
    count_mean <- mean(df_train$count, na.rm = TRUE)
    count_variance <- var(df_train$count, na.rm = TRUE)

    print(count_mean)
    print(count_variance)

    ## Variance exceeds mean with factor 7 --> Negative binomial family

    ###########################
    ### Correlation matrix ####
```

```
###########################

vars_eda <- c(
  "SE",
  "t_mean",
  "t_max",
  "t_min",
  "rh_mean",
  "ws_mean",
  "ws_cv",
  "p",
  "so_h",
  "so_h_mean_28",
  "so_h_mean_14",
  "t_mean_mean_28",
  "t_mean_mean_14",
  "t_max_mean_28",
  "t_max_mean_14",
  "t_min_mean_28",
  "t_min_mean_14",
  "p_acc_28",
  "p_acc_14",
  "ws_mean_mean_28",
  "ws_mean_mean_14",
  "rh_mean_mean_28",
  "rh_mean_mean_14"
)
#Correlation of weather parameters only during active season
eda_cor <- df_train %>%
  dplyr::filter(doy >= 121 & doy <= 304) %>%
  dplyr::select(count, all_of(vars_eda))
```

```
    #Pearson-correlations of each attribute in vars_eda with count

    cor_values <- sapply(eda_cor, function(x) cor(x, eda_cor$count,
use = "complete.obs")) #ignores missing values

    #Create df for readability

    cor_df <- data.frame(Predictor = names(cor_values), Correlation =
cor_values)


    # Sorting in descending order

    sorted_cor_df <- cor_df %>%

      arrange(desc(Correlation))


    print(sorted_cor_df)


    ### Correlation Matrix ###


    vars_mat <- c(

      "count",

      "SE",

      "t_max",

      "rh_mean",

      "ws_mean",

      "ws_cv",

      "p",

      "so_h",

      "so_h_mean_28",

      "t_mean_mean_28",

      "p_acc_28",

      "ws_mean_mean_28",

      "rh_mean_mean_28"

    )
    #Select data for matrix
    eda_mat <- eda_cor[, vars_mat]

    #Create matrix
    cormat <- rcorr(as.matrix(eda_mat))
    corrplot(cormat$r)

    #Plotting
    png("D:/Masterarbeit/Figures/CorMat_Meteo.png")
    corrplot(cormat$r)
    dev.off()
```

Sampling of pseudo-absences

```
rm(list = ls(all.names = TRUE))
gc()
library(sf) #simple features
library(dplyr) #For mutating dfs
library(tmap) #Plotting geometries
library(mgcv) #For GAM modeling
library(pROC) #For AUROC validation
library(ggplot2)  #Plotting histograms
library(gridExtra) #For subbplotting
library(corrplot) #Plotting Correlation Matrix
library(Hmisc) #Calculating correlation matrix
library(prioritizr) #Adjancency matrix
library(lubridate) #Date transformation
library(tidyr) #For unnesting lists containing columns

#Load data frame clipped to area of distribution for sampling of pseudo
absences
```

```
    sf_train_graz                                                    <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postpr
ocessed/sf_train_graz.gpkg",  layer = "layer_name")

    sf_train_graz$LU_Majority_Class                                  <-
as.factor(sf_train_graz$LU_Majority_Class)

    sf_train_graz$Presence <- as.factor(sf_train_graz$Presence)



#Load Mosquito Sightings and create day of the year attribute

    Mos                                                              <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Mosquito/Mosquitos_val1_3
1287.shp")

    Mos$doy <- yday(Mos$creation_d)


## define color ramp # HEX codes from https://colorbrewer2.org/

    custom_palette   <-   colorRampPalette(c("#f0f0f0",   "#e0ecf4",
"#9ebcda", "#8856a7", "#3f007d"))

    custom_palette2 <- colorRampPalette(c("black", "red"))


##########################################

###### Sampling of pseudo absences #######

##########################################


# Seed for reproducibility of random sampling

    set.seed(123)


#### Remove absences adjacent to presences from sampling process ####

    # Create an adjacency matrix for the points

    am_points <- proximity_matrix(sf_train_graz, distance = 145)  #145
m to include all eight surrounding pixels


    # Identify rows of presence points

    presence_indices <- which(sf_train_graz$Presence == 1)
    #Create empty df
    remove_indices_0 <- numeric()

    for (i in presence_indices) {
      # Get the row indices of absences that are adjacent to this
presence point
      adjacent_absences <- which(am_points[i, ] == 1)

      # Add those indices to the removal list
      remove_indices_0         <-         unique(c(remove_indices_0,
adjacent_absences))
    }

    # Now, filter out the absence points from sf_0 based on the removal
indices
    sf_train_graz_filtered <- sf_train_graz[-remove_indices_0, ]

## Separate presence and absence
    sf_1 <- sf_train_graz[sf_train_graz$Presence == 1, ]
    sf_0 <- sf_train_graz_filtered[sf_train_graz_filtered$Presence ==
0, ]

### PPS and RS sampling of absences
    #Sample size absences
    num_samples    <-    sum(sf_train_graz$Presence    ==    1)*1.1
#https://opengeohub.github.io/spatial-prediction-eml/spatiotemporal-
machine-learning-for-species-distribution-modeling.html

    # PPS-Sampling
    sf_0_pps <- sf_0[sample(1:nrow(sf_0), size = num_samples,
                            replace = FALSE,
                            prob = sf_0$Sampling_Effort), ]
    # Random sampling

    sf_0_rs <- sf_0[sample(1:nrow(sf_0), size = num_samples, replace
= FALSE), ]


    # Combine the two

    sf_resamp_pps = rbind(sf_1, sf_0_pps)

    sf_resamp_rs = rbind(sf_1, sf_0_rs)


    sf_resamp_pps <- st_sf(sf_resamp_pps)

    sf_resamp_rs <- st_sf(sf_resamp_rs)




### Create map with all Presence/Absence and Sampling Effort

    #Transform Presence to factor

    sf_resamp_pps$Presence <- factor(sf_resamp_pps$Presence, levels =
c("0", "1"))  # "0" for Absence, "1" for Presence

    sf_resamp_rs$Presence  <-  factor(sf_resamp_rs$Presence,  levels  =
c("0", "1"))


    # Create the map with pps-sampled Presence/Absence and Sampling
Effort

    map_sf_resamp_pps <- tm_shape(sf_train_graz) +

      tm_squares(

        fill = "Sampling_Effort",

        fill.scale        =        tm_scale_continuous(values        =
custom_palette(10)),

        size = 1,

        fill_alpha = 1,

        col_alpha = 0,

        fill.legend = tm_legend(title = "Sampling Effort", reverse =
TRUE)

      ) +

      tm_title("PPS Sampling") +
   tm_shape(sf_resamp_pps) +
   tm_dots(
    fill = "Presence",
    fill.scale = tm_scale_categorical(
     values = custom_palette2(2),
     labels = c("Absence", "Presence")
    ),
    size = 0.4,
    shape = 20,
    fill_alpha = 1,
    col_alpha = 1,
    fill.legend = tm_legend(title = "Status")
   ) +
   tm_layout(
    outer.margins = c(0.0, 0.02, 0.02, 0.02),
    inner.margins = c(0, 0, 0, 0)
   )


  # Create the map with rs-sampled Presence/Absence and Sampling
Effort
  map_sf_resamp_rs <- tm_shape(sf_train_graz) +
   tm_squares(
    fill = "Sampling_Effort",
    fill.scale = tm_scale_continuous(values =
custom_palette(10)),
    size = 1,
    fill_alpha = 1,
    col_alpha = 0,
    fill.legend = tm_legend(title = "Sampling Effort", reverse =
TRUE)
   ) +
   tm_title("Random Sampling") +
      tm_shape(sf_resamp_rs) +
      tm_dots(
        fill = "Presence",
        fill.scale = tm_scale_categorical(
          values = custom_palette2(2),
          labels = c("Absence", "Presence")
        ),
        size = 0.4,
        shape = 20,
        fill_alpha = 1,
        col_alpha = 1,
        fill.legend = tm_legend(title = "Status")
      )

      # Combine both maps
      tmap_arrange(map_sf_resamp_pps, map_sf_resamp_rs, ncol=2)



### CD plots - Exploratory data analysis (EDA) ###

    # Numerical predictors to analyze
    vars_eda <- c("LC_Green_Space","LC_Roof_50","LC_Roof_51",
                  "LC_Roof_52",                   "LC_Sealed_Ground",
"LC_Sealed_Ground","LC_Vegetation_High",
                  "LC_Vegetation_Low","LC_Water", "LU_Allot_Grave",
                  "Population_Density","Rain_Inlet","TWI")

    # Drop geometry to work with data as a regular dataframe
    eda_pps <- st_drop_geometry(sf_resamp_pps)
    eda_rs <- st_drop_geometry(sf_resamp_rs)
    #Check correlation between predictors
    eda_cor    <-    eda_pps    %>%    dplyr::select(-c("Presence",
"LU_Majority_Class", "cluster"))
    cormat <- rcorr(as.matrix(eda_cor))
    corrplot(cormat$r)

    # Compute threshold for binary response
    threshold <- sum(eda_pps$Presence == 1) / nrow(eda_pps)



## Create single plots for each variable

# corresponding adjust values, and bins for histograms
adjust_values <- 5
bins_values <- 50

# Initialize an empty list to store plots
plots_list <- list()

# Vector to store AUROC values
auroc_values <- numeric(length(vars_eda))

# Loop through each variable and create a separate plot
for (i in seq_along(vars_eda)) {
  var_name <- vars_eda[i]

  # Fit univariate GAM model (numerical)
  fit <- gam(as.formula(paste("Presence ~ s(", var_name, ", k = 3)")),
data = eda_pps, family = binomial)
  probs <- predict(fit, type = "response", newdata = eda_pps)
  # Compute AUC (Univariate Analysis)
  auroc <- roc(response = eda_pps$Presence, predictor = probs)
  auroc_value <- round(auc(auroc), 3)
  # Store AUROC value
  auroc_values[i] <- auroc_value

  # Create the Conditional Density Plot with Histogram
  p <- ggplot(eda_pps, aes_string(x = var_name)) +
    # Density plot (Conditional Density)
    geom_density(aes(y = ..count../max(..count..), fill = Presence),
position = "fill", adjust = adjust_values) +
    scale_fill_manual(values = c("0" = "grey", "1" = "#e34a33")) +
    # Histogram on secondary y-axis with custom bin size
    geom_histogram(aes(y = ..density../max(..density..)), bins =
bins_values, fill = "grey", color = "black", alpha = 0.3) +
    # Set primary y-axis between 0 and 1
    scale_y_continuous(limits = c(0, 1), sec.axis = sec_axis(~ . *
max(density(eda_pps[[var_name]], na.rm = TRUE)$y), name = "Density
(Histogram)")) +
    # Labels and Title
    labs(x = var_name, y = "Conditional Density", title = paste("GAM-
based AUROC:", auroc_value)) +
    theme(plot.title = element_text(size = 10, face = "bold", hjust =
0.5)) +
    # Axis limits and threshold line
    xlim(quantile(eda_pps[[var_name]], 0.01, na.rm = TRUE),
quantile(eda_pps[[var_name]], 0.99, na.rm = TRUE)) +
    geom_hline(yintercept = threshold, lwd = 2, linetype = "dashed",
color = "black")
  # Store the plot in the list
  plots_list[[var_name]] <- p
}
```

```
# Print AUROC values
names(auroc_values) <- vars_eda
print(auroc_values)

grid.arrange(grobs = plots_list, nrow = 3, ncol = 6)

# Histograms of Land Use classes
#Sort in ascending order
eda_pps$LU_Majority_Class <- factor(eda_pps$LU_Majority_Class,
                                  levels                             =
sort(as.numeric(levels(eda_pps$LU_Majority_Class))))
# Distribution of LU Classes across Presence and Absence
plot1 <- ggplot(eda_pps, aes(x = LU_Majority_Class)) +
  geom_bar(fill = "skyblue") +
  labs(title = "LU Classes across Presences/Absences", x = "Land Use
Category", y = "Count") +
  theme_minimal()

# Distribution of LU Classes only across Presence
plot2   <-   ggplot(subset(eda_pps,   Presence   ==   1),   aes(x   =
LU_Majority_Class)) +
  geom_bar(fill = "lightcoral") +
  labs(title = "LU Classes across Presences", x = "Land Use Category",
y = "Count") +
  theme_minimal()


grid.arrange(plot1, plot2, ncol = 1)

#Distribution of LU Classes across pseudo absences and presences
```

```
plot3 <- ggplot(subset(sf_resamp_pps, Presence == 0), aes(x = LU_Majority_Class)) +
  geom_bar(fill = "skyblue") +
  labs(title = "LU Classes across Pseudo absences (M1)", x = "Land Use Category", y = "Count") +
  theme_minimal() +
  coord_cartesian(ylim = c(0, 130))

# Distribution of LU Classes only across Presence
plot4 <- ggplot(subset(sf_resamp_rs, Presence == 0), aes(x = LU_Majority_Class)) +
  geom_bar(fill = "skyblue") +
  labs(title = "LU Classes across Pseudo absences (M2)", x = "Land Use Category", y = "Count") +
  theme_minimal() +
  coord_cartesian(ylim = c(0, 130))



grid.arrange(plot2, plot3, plot4,  ncol = 1)

##################
# Count land use across presences and absences of the two approaches
plot_data <- bind_rows(
  eda_pps %>% filter(Presence == 1) %>% count(LU_Majority_Class) %>% mutate(Dataset = "Presence"),
  sf_resamp_pps %>% filter(Presence == 0) %>% count(LU_Majority_Class) %>% mutate(Dataset = "Absence M1"),
  sf_resamp_rs %>% filter(Presence == 0) %>% count(LU_Majority_Class) %>% mutate(Dataset = "Absence M2")
)

# Make sure all classes are present in each dataset
plot_data <- plot_data %>%
  complete(LU_Majority_Class, Dataset, fill = list(n=0))


# Plotting

ggplot(plot_data, aes(x=LU_Majority_Class, y=n, fill=Dataset)) +

  geom_bar(stat="identity",      position=position_dodge(width=0.8),
width=0.7) +

  theme_minimal() +

  labs(title="Land use class distribution by presences and pseudo
absences",

       x="Land use category",

       y="Count") +

  scale_fill_manual(name   =   NULL,   values=c("Presence"="black",
"Absence M1"="skyblue", "Absence M2"="orange")) +

  theme(axis.text.x=element_text(angle=45, hjust=1))


ggsave(plot                     =                    last_plot(),
"D:/Masterarbeit/Figures/Distribution_pseudoabsences.png")


#Save resampled data frames

    st_write(sf_resamp_pps,
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postprocessed/
sf_resamp_pps_no_doy.gpkg",  layer = "layer_name", append = FALSE)

    st_write(sf_resamp_rs,
"D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postprocessed/
sf_resamp_rs_no_doy.gpkg",  layer = "layer_name", append = FALSE)
```

Spatial model fitting and validation

```
rm(list = ls(all.names = TRUE))

gc()

library(sperrorest) #Data partitioning for cross-validation

library(magick) # image processing and editing

library(tmap) #thematic mapping

library(dplyr) #For mutating df
library(sf) #Simple features

library(raster) #For raster data

library(ggplot2) #For plotting

library(vip) #visualizing variable importance

library(terra) #Raster analysis

library(tibble) #for data wrangling

library(mgcv) #For model fitting

library(viridisLite) #color palettes

library(pROC) # AUROC calculation

library(mgcViz) # Visualization of partial effects and interaction
effects

library(patchwork) #For plotting


## define color ramp # HEX codes from https://colorbrewer2.org/

custom_palette <- colorRampPalette(c('#006837', '#1a9850', '#66bd63',
'#a6d96a', '#d9ef8b', '#fee08b', '#fdae61', '#f46d43', '#d73027',
'#a50026'))

custom_palette2 <- colorRampPalette(c("black", "red"))


##Read data and manipulations


    # PPS and RS are resampled data from sf_train (cropped to current
species distribution) to fit the model

    sf_train_pps <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postpr
ocessed/sf_resamp_pps_no_doy.gpkg", layer = "layer_name")

    sf_train_rs <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postpr
ocessed/sf_resamp_rs_no_doy.gpkg", layer = "layer_name")


    sf_train_graz <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postpr
ocessed/sf_train_graz.gpkg", layer = "layer_name")
```

```
    # Data frames that cover the whole study areas (not cropped to
species current distribution)

    sf_pred_graz                                                    <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/100m/Postpr
ocessed/sf_pred_graz.gpkg",  layer = "layer_name")


    ##Create factors

      sf_train_pps$Presence <- as.factor(sf_train_pps$Presence)

      sf_train_pps$LU_Majority_Class                                <-
as.factor(sf_train_pps$LU_Majority_Class)

      sf_train_pps$cluster <- as.factor(sf_train_pps$cluster)


      sf_train_rs$Presence <- as.factor(sf_train_rs$Presence)

      sf_train_rs$LU_Majority_Class                                 <-
as.factor(sf_train_rs$LU_Majority_Class)


      sf_pred_graz$Presence <- as.factor(sf_pred_graz$Presence)

      sf_pred_graz$LU_Majority_Class                                <-
as.factor(sf_pred_graz$LU_Majority_Class)


#########################

### Fitting GAM Model ###

#########################


  ## define formulas including interaction terms

  #pps

  fo_all_pps <- Presence ~

        factor(LU_Majority_Class) +

        s(Population_Density, k = 3) +

        s(LU_Allot_Grave, k = 3) +

        s(Rain_Inlet, k = 3) +

        s(LC_Soil, k = 3) +

        s(TWI, k = 3) +
        s(LC_Roof_51, k = 3) +
        s(LC_Vegetation_Low, k = 3) +
        s(LC_Vegetation_High, k = 3) +
        s(LC_Water, k=3) +
        s(LC_Sealed_Ground, k=3) +
        s(LC_Green_Space, k = 3) +
        s(LC_Roof_52, k = 3) +
        s(LC_Roof_50, k = 3) +
        ti(LC_Roof_51, LC_Vegetation_Low, k=c(4,4)) +
        ti(LC_Roof_51, LC_Vegetation_High, k=c(4,4)) +
        ti(LC_Roof_51, LC_Green_Space, k=c(4,4)) +
        ti(LC_Roof_52, LC_Vegetation_High, k=c(4,4)) +
        ti(LC_Roof_52, LC_Vegetation_Low, k=c(4,4)) +
        ti(LC_Roof_52, LC_Green_Space, k=c(4,4))


    #rs
    fo_all_rs <- Presence ~
      factor(LU_Majority_Class) +
      s(Population_Density, k = 3) +
      s(LU_Allot_Grave, k = 3) +
      s(Rain_Inlet, k = 3) +
      s(LC_Soil, k = 3) +
      s(TWI, k = 3) +
      s(LC_Roof_51, k = 3) +
      s(LC_Vegetation_Low, k = 3) +
      s(LC_Vegetation_High, k = 3) +
      s(LC_Water, k=3) +
      s(LC_Sealed_Ground, k=3) +
      s(LC_Green_Space, k = 3) +
      s(LC_Roof_52, k = 3) +
      s(LC_Roof_50, k = 3) +

      s(Sampling_Effort, k=3) +

      ti(LC_Roof_51, LC_Vegetation_Low, k=c(4,4)) +

      ti(LC_Roof_51, LC_Vegetation_High, k=c(4,4)) +

      ti(LC_Roof_51, LC_Green_Space, k=c(4,4)) +

      ti(LC_Roof_52, LC_Vegetation_High, k=c(4,4)) +

      ti(LC_Roof_52, LC_Vegetation_Low, k=c(4,4)) +

      ti(LC_Roof_52, LC_Green_Space, k=c(4,4))


    ##fit models

    #pps

    myfit_all_pps = mgcv::gam(fo_all_pps, data=sf_train_pps,
family=binomial, method = "REML", select = TRUE)

    summary(myfit_all_pps)


    #rs

    myfit_all_rs = mgcv::gam(fo_all_rs, data=sf_train_rs,
family=binomial, method = "REML", select = TRUE)

    summary(myfit_all_rs)


    # Viszualize interaction effects

        #PPS

        vis.gam(myfit_all_pps, plot.type="contour",
view=c("LC_Roof_51","LC_Vegetation_Low"), too.far = 0.05)

        vis.gam(myfit_all_pps, plot.type="contour",
view=c("LC_Roof_51","LC_Vegetation_High"), too.far = 0.05)

        vis.gam(myfit_all_pps, plot.type="contour",
view=c("LC_Roof_52","LC_Green_Space"), too.far = 0.05)

        vis.gam(myfit_all_pps, plot.type="contour",
view=c("LC_Roof_52","LC_Vegetation_High"), too.far = 0.05)

        vis.gam(myfit_all_pps, plot.type="contour",
view=c("LC_Roof_52","LC_Vegetation_Low"), too.far = 0.05)
```

```
        #RS

        par(mfrow=c(2,2))

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_51","LC_Green_Space"), too.far = 0.2)

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_51","LC_Vegetation_Low"), too.far = 0.2)

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_51","LC_Vegetation_High"), too.far = 0.2)

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_52","LC_Green_Space"), too.far = 0.2)

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_52","LC_Vegetation_High"), too.far = 0.2)

        vis.gam(myfit_all_rs,                   plot.type="contour",
view=c("LC_Roof_52","LC_Vegetation_Low"), too.far = 0.2)


    par(mfrow=c(3,4))


    #Create visGAM objects for partial effect plots

    vis_gam_pps <- getViz(myfit_all_pps)

    vis_gam_rs <- getViz(myfit_all_rs)


    print(plot(vis_gam_pps, allTerms = T), pages = 1)

    print(plot(vis_gam_rs, allTerms = T), pages = 1)


    #Partial effects RS

    P_LU_Allot <- plot( sm(vis_gam_rs, 2) )

    P_LU_Allot   <-   P_LU_Allot   +   l_fitLine(colour   =   "red")   +
l_rug(mapping = aes(x=x, y=y), alpha = 0.8) +

      l_ciLine(mul = 5, colour = "blue", linetype = 2) +

      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()


    P_TWI <- plot( sm(vis_gam_rs, 5) )
    P_TWI <- P_TWI + l_fitLine(colour = "red") + l_rug(mapping =
aes(x=x, y=y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()

    P_LC_Roof_51 <- plot( sm(vis_gam_rs, 6) )
    P_LC_Roof_51 <- P_LC_Roof_51 + l_fitLine(colour = "red") +
l_rug(mapping = aes(x=x, y=y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()

    P_LC_Roof_52 <- plot( sm(vis_gam_rs, 12) )
    P_LC_Roof_52 <- P_LC_Roof_52 + l_fitLine(colour = "red") +
l_rug(mapping = aes(x=x, y=y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()

    P_Sampling_Effort <- plot( sm(vis_gam_rs, 14) )
    P_Sampling_Effort <- P_Sampling_Effort + l_fitLine(colour = "red")
+ l_rug(mapping = aes(x=x, y=y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()

    P_LC_Vegetation_High <- plot( sm(vis_gam_rs, 8) )
    P_LC_Vegetation_High <- P_LC_Vegetation_High + l_fitLine(colour =
"red") + l_rug(mapping = aes(x=x, y=y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) + theme_classic()

    gridPrint(P_LU_Allot, P_LC_Roof_51, P_LC_Roof_52,
P_Sampling_Effort, P_LC_Vegetation_High, P_TWI, ncol=3)

###################################
```

```
##### Graz AUROC calculations ######
####################################

    #Predict habitat suitability values for sf_train (complete df for graz cropped to area of species distribution): for model comparison
    sf_train_graz$prob_pps <- predict(myfit_all_pps, type = "response", newdata = sf_train_graz)
    sf_train_graz$prob_rs <- predict(myfit_all_rs, exclude = c("s(Sampling_Effort)"), type = "response", newdata = sf_train_graz)

    #Prediction on resampled dfs for fitting AUROC calculation
    sf_train_pps$prob_pps <- predict(myfit_all_pps, type = "response", newdata = sf_train_pps)
    sf_train_rs$prob_rs <- predict(myfit_all_rs, exclude = c("s(Sampling_Effort)"), type = "response", newdata = sf_train_rs)

    #Calculate fitting ROC
    roc_graz_pps <- roc(sf_train_graz$Presence, sf_train_graz$prob_pps)
    roc_graz_rs <- roc(sf_train_graz$Presence, sf_train_graz$prob_rs)
    fitting_roc_pps <- roc(sf_train_pps$Presence, sf_train_pps$prob_pps)
    fitting_roc_rs <- roc(sf_train_rs$Presence, sf_train_rs$prob_rs)

    #Calculate fitting AUROC
    auroc_graz_pps <- auc(roc_graz_pps)
    auroc_graz_rs <- auc(roc_graz_rs)
    fitting_auroc_pps <- auc(fitting_roc_pps)
    fitting_auroc_rs <- auc(fitting_roc_rs)

    print(auroc_graz_pps)
    print(auroc_graz_rs)
    print(fitting_auroc_pps)
    print(fitting_auroc_rs)


###################################
### Prediction of Probabilities ###
##################################
    #Graz
    sf_pred_graz$prob_pps <- predict(myfit_all_pps, type =
"response", newdata = sf_pred_graz)
    sf_pred_graz$prob_rs <- predict(myfit_all_rs, type = "response",
exclude = c("s(Sampling_Effort)"), newdata = sf_pred_graz)


    #Export predicitons
    st_write(sf_pred_graz,
"D:/Masterarbeit/Jupyter/Data/Probabilities/sf_predicted_graz.gpkg",
layer = "layer_name", append = FALSE)


    #AUROC after prediction
    roc_graz_pred_pps <- roc(sf_pred_graz$Presence,
sf_pred_graz$prob_pps)
    roc_graz_pred_rs <- roc(sf_pred_graz$Presence,
sf_pred_graz$prob_rs)


    auroc_graz_pred_pps <- auc(roc_graz_pred_pps)
    auroc_graz_pred_rs <- auc(roc_graz_pred_rs)


    print(auroc_graz_pred_pps)
    print(auroc_graz_pred_rs)


######################################
### Plot predictions #################
######################################


    #Pixel size
    target_resolution <- 100

    extent_data_graz <- st_bbox(sf_pred_graz)  # extent of Raster

    #Color palette

    custom_palette <- viridisLite::turbo(100)


    coords_graz <- st_coordinates(sf_pred_graz)


    probs_graz_pps <- sf_pred_graz$prob_pps

    probs_graz_rs <- sf_pred_graz$prob_rs


    raster_i_graz <- raster(extent(extent_data_graz),

                       res = c(target_resolution, target_resolution),

                       crs = CRS("+init=epsg:31287"))


    # Transform df to spatial point data

    spdf_graz_pps <- SpatialPointsDataFrame(coords_graz, data =
data.frame(probs_graz_pps), proj4string =
CRS(st_crs(sf_pred_graz)$proj4string))

    spdf_graz_rs <- SpatialPointsDataFrame(coords_graz, data =
data.frame(probs_graz_rs), proj4string =
CRS(st_crs(sf_pred_graz)$proj4string))


    #Rasterization

    raster_i_graz_pps <- rasterize(spdf_graz_pps, raster_i_graz,
field = "probs_graz_pps", na.rm = TRUE)

    raster_i_graz_rs <- rasterize(spdf_graz_rs, raster_i_graz, field
= "probs_graz_rs", na.rm = TRUE)


    #Saving

    writeRaster(raster_i_graz_pps, filename =
"D:/Masterarbeit/Figures/Probabilities_pred_graz_pps.tif", overwrite
= TRUE)
```

```
    writeRaster(raster_i_graz_rs, filename = "D:/Masterarbeit/Figures/Probabilities_pred_graz_rs.tif", overwrite = TRUE)


    #Plotting

        #PPS

png("D:/Masterarbeit/Figures/Habitat_suitability_Graz_pps.png", width = 800, height = 600)

        plot(raster_i_graz_pps, col = custom_palette, zlim = c(0, 1),

             main = "Habitat suitability in Graz (PPS)",

             xlab = "Longitude", ylab = "Latitude")


        #RS

png("D:/Masterarbeit/Figures/Habitat_suitability_Graz_rs.png",  width = 800, height = 600)

        plot(raster_i_graz_rs, col = custom_palette, zlim = c(0, 1),

             main = "Habitat suitability in Graz (RS)",

             xlab = "Longitude", ylab = "Latitude")


        dev.off()


########################################

### FEATURE IMPORTANCE based on auroc###

########################################


    #For PPS

        ds <- as_tibble(sf_train_pps)

        mynsim = 30 # number of simulations


        set.seed(666)
        # Select the relevant columns
        cn_s <- colnames(myfit_all_pps$model)
        cn_s[cn_s == "factor(LU_Majority_Class)"] =
"LU_Majority_Class"  # Correcting column name
        train_sel_s <- ds %>%
          dplyr::select(any_of(cn_s)) %>%
          dplyr::select(-Presence)

        # Define the target variable
        target_s <- ds$Presence

        # Calculate variable importance
        result_pps <- vi_permute(
          nsim = mynsim,
          object = myfit_all_pps,
          train = train_sel_s,
          target = target_s,
          event_level = "second",               # Because predicted
instance is 1 and not 0
          metric = "roc_auc",               # auroc based...
          pred_wrapper = predict.gam
        ) %>%
          dplyr::arrange(-Importance) %>%
          dplyr::mutate(process = "Mosquito Presence")
        ###
        ##
        #

        result_pps

        # Ensure the Variable column is ordered by Importance in
descending order
```

```
    result_pps <- result_pps %>%
     arrange(desc(Importance)) %>%  # Sort data frame descending
     mutate(Variable = factor(Variable, levels = rev(Variable))) # Reverse factor levels
    result_pps


    #Plot feature importance
    Importance_Mosquito_pps <- ggplot(result_pps, aes(x = Importance, y = Variable)) +
     geom_point(color = "#747474", size = 3) +
     geom_errorbarh(aes(xmin = Importance - StDev, xmax = Importance + StDev),
                height = 0.2, color = "#747474", linewidth = 0.5) +  # Increase line width
     theme_minimal(base_size = 16) +  # Increase font size
     labs(title = "Feature Importance M1", x = "Importance", y = "Variable") +
     theme(axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(size = 18, face = "bold"),
        panel.border = element_rect(color = "black", fill = NA, size = 1))
    Importance_Mosquito_pps


    ggsave(plot=Importance_Mosquito_pps, "D:/Masterarbeit/Figures/Feature_Importance_PPS.png")


  #For RS
    ds <- as_tibble(sf_train_rs)
    mynsim = 30 # number of simulations


    set.seed(666)
```

```
        # Select the relevant columns
        cn_s <- colnames(myfit_all_rs$model)
        cn_s[cn_s == "factor(LU_Majority_Class)"] =
"LU_Majority_Class"  # Correcting column name
        train_sel_s <- ds %>%
          dplyr::select(any_of(cn_s)) %>%
          dplyr::select(-Presence)

        # Define the target variable
        target_s <- ds$Presence

        # Calculate variable importance
        result_rs <- vi_permute(
          nsim = mynsim,
          object = myfit_all_rs,
          train = train_sel_s,
          target = target_s,
          event_level = "second",             # Because my predicted
instance is 1 and not 0
          metric = "roc_auc",                 # auroc based...
          pred_wrapper = predict.gam
        ) %>%
          dplyr::arrange(-Importance) %>%
          dplyr::mutate(process = "Mosquito Presence")
        ###
        ##
        #

        result_rs

        # Ensure the Variable column is ordered by Importance in
descending order
```

```
        result_rs <- result_rs %>%
          arrange(desc(Importance)) %>%  # Sort data frame
          mutate(Variable = factor(Variable, levels = rev(Variable)))
# Reverse factor levels
        result_rs


        # Plotting
        Importance_Mosquito_rs <- ggplot(result_rs, aes(x =
Importance, y = Variable)) +
          geom_point(color = "#747474", size = 3) +
          geom_errorbarh(aes(xmin = Importance - StDev, xmax =
Importance + StDev),
                         height = 0.2, color = "#747474", linewidth =
0.5) +  # Increase line width
          theme_minimal(base_size = 16) +  # Increase font size
          labs(title = "Feature Importance M2", x = "Importance", y =
"Variable") +
          theme(axis.title.x = element_text(size = 15),
                axis.title.y = element_text(size = 15),
                plot.title = element_text(size = 18, face = "bold"),
                panel.border = element_rect(color = "black", fill =
NA, size = 1))
        Importance_Mosquito_rs


        ggsave(plot=Importance_Mosquito_rs,
"D:/Masterarbeit/Figures/Feature_Importance_RS.png")


################################################
### Spatial cross validation using sperrorest (Brenning, 2012)###
################################################
    #Transform geometry to coords
    coords_pps <- st_coordinates(sf_train_pps)
    coords_rs <- st_coordinates(sf_train_rs)
    # Bind coordinates back as separate columns
    sf_train_pps <- sf_train_pps %>%
      st_drop_geometry() %>%
      cbind(coords_pps)

    sf_train_rs <- sf_train_rs %>%
      st_drop_geometry() %>%
      cbind(coords_rs)

    nfolds = 5 #Five fold since there are 5 clusters
    nreps = 1 #one repetition is enough since partitioning is factor based

    #Factor partioning of data into train and test
    parti_pps <- partition_factor_cv(sf_train_pps, coords = c("X", "Y"), nfold = nfolds, repetition = nreps, seed1 = 123, fac = "cluster")
    parti_rs <- partition_factor_cv(sf_train_rs, coords = c("X", "Y"), nfold = nfolds, repetition = nreps, seed1 = 123, fac = "cluster")

    #Initzialize data frame for results
    results_cv_mos_pps <- data.frame(Repetition = integer(), Fold = integer(), AUROC = numeric(), stringsAsFactors = FALSE)
    results_cv_mos_rs <- data.frame(Repetition = integer(), Fold = integer(), AUROC = numeric(), stringsAsFactors = FALSE)

    #Loop through folds and reps to fit model, predict and calculate AUROC
    #pps
    for (j in 1:nreps) {
      partiloop <- parti_pps[[j]]  # Access the jth element
      for (i in 1:nfolds) {
        first <- partiloop[[i]][[2]]  # Access the second element of the ith fold
        test <- sf_train_pps[first,]; ntesti <- nrow(test); train <- sf_train_pps[-first,]  # Exclude the train set
        # Fit and calculate AUROC
        myfit <- mgcv::gam(fo_all_pps, data = train, family = binomial, method = "REML", select = TRUE)
        test$prob <- predict.gam(myfit, type = "response", newdata = test)
        myroc <- pROC::roc(response = test$Presence, predictor = test$prob, auc = TRUE)
        auroc <- round(myroc$auc, 5)
        print(auroc)
        # Store results
        results_cv_mos_pps <- rbind(results_cv_mos_pps, data.frame(Repetition = j, Fold = i, AUROC = auroc))}}


    #rs
    for (j in 1:nreps) {
      partiloop <- parti_rs[[j]]  # Access the jth element
      for (i in 1:nfolds) {
        first <- partiloop[[i]][[2]]  # Access the second element of the ith fold
        test <- sf_train_rs[first,]; ntesti <- nrow(test); train <- sf_train_rs[-first,]  # Exclude the test set
        # Fit and calculate AUROC
        myfit <- mgcv::gam(fo_all_rs, data = train, family = binomial, method = "REML", select = TRUE)
        test$prob <- predict.gam(myfit, type = "response", exclude = "s(Sampling_Effort)", newdata = test)
        myroc <- pROC::roc(response = test$Presence, predictor = test$prob, auc = TRUE)
        auroc <- round(myroc$auc, 5)
        print(auroc)
        # Store results
        results_cv_mos_rs <- rbind(results_cv_mos_rs,
data.frame(Repetition = j, Fold = i, AUROC = auroc))}}

    #average results
    average_results_pps <- aggregate(AUROC ~ Repetition, data =
results_cv_mos_pps, FUN = mean)
    average_results_rs <- aggregate(AUROC ~ Repetition, data =
results_cv_mos_rs, FUN = mean)
    print(average_results_pps)
    print(average_results_rs)

    ######################################################
    ##### Random CV for uncertainty calculation ##########
    ######################################################

    coords_2 <- st_coordinates(sf_pred_graz)

    sf_pred_graz <- sf_pred_graz %>%
      st_drop_geometry() %>%
      cbind(coords_2)

    nfolds = 5 #folds
    nreps = 10 #repetitions

    # Initzialize list to store predicitons of each fold
    predictions_list <- vector("list", length = nfolds)

    #Random partioning of data into train and test
    parti <- partition_cv(sf_train_rs, nfold = nfolds, repetition =
nreps, seed1 = 123)
```

```
    #initzialize empty df
    results_cv_mos <- data.frame(Repetition = integer(), Fold =
integer(), AUROC = numeric(), stringsAsFactors = FALSE)


    #Loop through folds and reps to fit model and predict habitat
suitability scores
    for (j in 1:nreps) {
      partiloop <- parti[[j]]  # Access the jth element
      for (i in 1:nfolds) {
        first <- partiloop[[i]][[2]]  # Access the second element of
the ith fold
        train <- sf_train_rs[-first,]  # Exclude the test set
        pred <- sf_pred_graz
        # Fitting
        myfit <- mgcv::gam(fo_all_rs, data = train, family = binomial,
method = "REML", select = TRUE)
        #Prediciton
        pred$prob <- predict.gam(myfit, type = "response", exclude =
c("s(Sampling_Effort)"), newdata = pred)
                #Store predictions of each fold
        predictions_list[[i]] <- rbind(predictions_list[[i]], pred[,
c("X", "Y", "prob")])
      }
      #Print repetition number
      print(j)
    }
    #Combine lists
    all_predictions <- do.call(rbind, predictions_list)


    # Calculate standard deviation of predictions by location
    uncertainty <- all_predictions %>%
      group_by(X, Y) %>%
      summarise(sd_pred = sd(prob, na.rm = TRUE))
    #Plot uncertainties
    ggplot(uncertainty, aes(x = X, y = Y, fill = sd_pred)) +
      geom_tile() +
      scale_fill_viridis_c(option = "viridis", name = "Prediction\nStd Dev") +
      coord_fixed() +
      labs(title = "Spatial Uncertainty in Predictions",
           x = "Longitude",
           y = "Latitude") +
      theme_minimal()

    ggsave(plot = last_plot(), "D:/Masterarbeit/Figures/Model_uncertainties.png")
    #Merge uncertainties and sf_pred_graz
    sf_pred_graz <- sf_pred_graz %>%
      left_join(uncertainty %>% dplyr::select(X, Y, sd_pred), by = c("X", "Y"))

    st_write(sf_pred_graz, "D:/Masterarbeit/Jupyter/Data/Probabilities/sf_predicted_graz.gpkg", layer = "layer_name", append = FALSE)
```

Temporal model fitting and validation

```
rm(list = ls(all.names = TRUE))
gc()
library(dplyr) #For mutating dfs
library(mgcv) #For model fitting
library(MASS) #For negative binomial distribution
library(sperrorest) #Partitioning in cross validation
library(vip) #Feature Importance
library(ggplot2)  #Plotting histograms
library(mgcViz) #visualization of partial effects


#load postprocessed data

df_train                                                   <-
read.csv("D:/Masterarbeit/Jupyter/Data/QGIS/Sampling_data/Meteo/df_t
rain_meteo_MA.csv")

#Rename columns

df_train <- df_train %>%

  dplyr::rename(Sampling_Effort = SE) %>%

  dplyr::rename(year = creation_y)


df_train$year <- as.factor(df_train$year)


########################

######### GAMM ##########

########################


# 1. Set: Daily values and 28-days time-lagged averages/accumulations

fo <- count ~

  s(t_max, k = 3) +

  s(rh_mean, k = 3) +

  s(ws_mean, k = 3) +

  s(ws_cv, k = 3) +

  s(p, k = 3) +

  s(t_mean_mean_28, k = 3) +

  s(rh_mean_mean_28, k = 3) +

  s(ws_mean_mean_28, k = 3) +

  s(p_acc_28, k = 3) +

  s(doy, k = 12, bs = "cc") +  s(year, bs = "re") +  s(Sampling_Effort,
k = 3)


# 2. Set: Daily values and 28-days time-lagged variations
```

```
fo_cv <- count ~
  s(t_max, k = 3) +
  s(rh_mean, k = 3) +
  s(ws_mean, k = 3) +
  s(ws_cv, k = 3) +
  s(p, k = 3) +
  s(t_mean_cv_28, k = 3) +
  s(ws_mean_cv_28, k = 3) +
  s(rh_mean_cv_28, k = 3) +
  s(p_cv_28, k = 3) +
  s(doy, k = 12, bs = "cc") +  s(year, bs = "re") +  s(Sampling_Effort,
k = 3)


# # 1. Set: Daily values, 28-days time-lagged averages/accumulations
and 28-days time-lagged variations
fo_all <- count ~
  s(t_max, k = 3) +
  s(rh_mean, k = 3) +
  s(ws_mean, k = 3) +
  s(ws_cv, k = 3) +
  s(p, k = 3) +
  s(t_mean_mean_28, k = 3) +
  s(t_mean_cv_28, k = 3) +
  s(p_acc_28, k = 3) +
  s(p_cv_28, k = 3) +
  s(rh_mean_mean_28, k = 3) +
  s(rh_mean_cv_28, k = 3) +
  s(ws_mean_mean_28, k = 3) +
  s(ws_mean_cv_28, k = 3) +
  s(doy, k = 12, bs = "cc") + s(year, bs = "re") +  s(Sampling_Effort,
k = 3)
```

```
#### Model Fitting using three sets of predictors ####
myfit = mgcv::gam(fo, data=df_train, family= nb(), select = TRUE)
summary(myfit)

plot(myfit, page=1, scale=-1)

myfit_cv = mgcv::gam(fo_cv, data=df_train, family= nb(), select =
TRUE)
summary(myfit_cv)

plot(myfit_cv, page=1, scale=0)

myfit_all = mgcv::gam(fo_all, data=df_train, family= nb(), select =
TRUE)
summary(myfit_all)

plot(myfit_all, page=1, scale=0)

#R-squared
summary(myfit)$r.sq
summary(myfit_cv)$r.sq
summary(myfit_all)$r.sq

##########################
#### Partial effects #####
##########################

#cretae visGAM object
    vis_gam <- getViz(myfit)
    print(plot(vis_gam, allTerms = T), pages = 1)
```

```
#PLot partial effects
P_t_mean_28 <- plot(sm(vis_gam, 6))
P_t_mean_28 <- P_t_mean_28 +
  l_fitLine(colour = "red") +
  l_rug(mapping = aes(x = x, y = y), alpha = 0.8) +
  l_ciLine(mul = 5, colour = "blue", linetype = 2) +
  l_points(shape = 19, size = 1, alpha = 0.1) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_classic()

P_doy <- plot(sm(vis_gam, 10))
P_doy <- P_doy +
  l_fitLine(colour = "red") +
  l_rug(mapping = aes(x = x, y = y), alpha = 0.8) +
  l_ciLine(mul = 5, colour = "blue", linetype = 2) +
  l_points(shape = 19, size = 1, alpha = 0.1) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_classic()

P_t_max <- plot(sm(vis_gam, 1))
P_t_max <- P_t_max +
  l_fitLine(colour = "red") +
  l_rug(mapping = aes(x = x, y = y), alpha = 0.8) +
  l_ciLine(mul = 5, colour = "blue", linetype = 2) +
  l_points(shape = 19, size = 1, alpha = 0.1) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_classic()

P_rh_mean <- plot(sm(vis_gam, 2))
P_rh_mean <- P_rh_mean +
  l_fitLine(colour = "red") +
      l_rug(mapping = aes(x = x, y = y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) +
      geom_hline(yintercept = 0, linetype = "dashed") +
      theme_classic()

    P_SE <- plot(sm(vis_gam, 12))
    P_SE <- P_SE +
      l_fitLine(colour = "red") +
      l_rug(mapping = aes(x = x, y = y), alpha = 0.8) +
      l_ciLine(mul = 5, colour = "blue", linetype = 2) +
      l_points(shape = 19, size = 1, alpha = 0.1) +
      geom_hline(yintercept = 0, linetype = "dashed") +
      theme_classic()

    gridPrint(P_t_mean_28, P_doy, P_t_max, P_rh_mean, P_SE, ncol = 3)

##########################
#### Cross validation ####
##########################

nfolds = 5 #folds
nreps = 5 #repetitions

# Random partioninig
parti_factor <- partition_cv(df_train, nfold = nfolds, repetition =
nreps, seed1 = 123)

# Initzialize empty df for results
results_cv_meteo <- data.frame(
  Repetition = integer(),
  Fold = integer(),
  R2 = numeric(),
  RMSE = numeric(),
  stringsAsFactors = FALSE
)
results_cv_meteo_cv <- data.frame(
  Repetition = integer(),
  Fold = integer(),
  R2 = numeric(),
  RMSE = numeric(),
  stringsAsFactors = FALSE
)
results_cv_meteo_all <- data.frame(
  Repetition = integer(),
  Fold = integer(),
  R2 = numeric(),
  RMSE = numeric(),
  stringsAsFactors = FALSE
)

# CV loop for fo
for (j in 1:nreps) {
  partiloop <- parti_factor[[j]]  # Access the j-th element
  for (i in 1:nfolds) {
    first <- partiloop[[i]]$test  # Access the test set of the i-th
fold
    test <- df_train[first,]; ntesti <- nrow(test);
    train <- df_train[-first,]  # Exclude the test set
    # Fit and predict
    fit_cv <- mgcv::gam(fo, data = train, family = nb(), select =
TRUE)
```

```
    test$pred <- predict.gam(fit_cv, type = "response", newdata = test)

    #RMSE
    mse <- mean((test$count - test$pred)^2, na.rm = TRUE)
    rmse <- sqrt(mse)

    #Calculate R2
    ss_res <- sum((test$count - test$pred)^2, na.rm = TRUE)
    ss_tot <- sum((test$count - mean(test$count, na.rm = TRUE))^2, na.rm = TRUE)
    r2 <- 1 - ss_res/ss_tot

    # Store results
    results_cv_meteo <- rbind(results_cv_meteo, data.frame(Repetition = j, Fold = i, R2 = r2, RMSE = rmse))
  }
}

# CV loop  for fo_cv
for (j in 1:nreps) {
  partiloop <- parti_factor[[j]]  # Access the j-th element
  for (i in 1:nfolds) {
    first <- partiloop[[i]]$test  # Access the test set of the i-th fold
    test <- df_train[first,]; ntesti <- nrow(test);
    train <- df_train[-first,]  # Exclude the test set
    # Fit and predict
    fit_cv <- mgcv::gam(fo_cv, data = train, family = nb(), select = TRUE)
    test$pred <- predict.gam(fit_cv, type = "response", newdata = test)
    #RMSE
    mse <- mean((test$count - test$pred)^2, na.rm = TRUE)
    rmse <- sqrt(mse)

    #Calculate R2
    ss_res <- sum((test$count - test$pred)^2, na.rm = TRUE)
    ss_tot <- sum((test$count - mean(test$count, na.rm = TRUE))^2, na.rm = TRUE)
    r2 <- 1 - ss_res/ss_tot

    # Store results
    results_cv_meteo_cv <- rbind(results_cv_meteo_cv, data.frame(Repetition = j, Fold = i, R2 = r2, RMSE = rmse))
  }
}

# CV loop  for fo_all
for (j in 1:nreps) {
  partiloop <- parti_factor[[j]]  # Access the j-th element
  for (i in 1:nfolds) {
    first <- partiloop[[i]]$test  # Access the test set of the i-th fold
    test <- df_train[first,]; ntesti <- nrow(test);
    train <- df_train[-first,]  # Exclude the test set
    # Fit and predict
    fit_cv <- mgcv::gam(fo_all, data = train, family = nb(), select = TRUE)
    test$pred <- predict.gam(fit_cv, type = "response", newdata = test)

    #RMSE
    mse <- mean((test$count - test$pred)^2, na.rm = TRUE)
    rmse <- sqrt(mse)

    #Calculate R2
    ss_res <- sum((test$count - test$pred)^2, na.rm = TRUE)
    ss_tot <- sum((test$count - mean(test$count, na.rm = TRUE))^2,
na.rm = TRUE)
    r2 <- 1 - ss_res/ss_tot

    # Store results
    results_cv_meteo_all <- rbind(results_cv_meteo_all,
data.frame(Repetition = j, Fold = i, R2 = r2, RMSE = rmse))
  }
}

# Print results
average_results_r2 <- aggregate(R2 ~ Repetition, data =
results_cv_meteo, FUN = mean)
average_results_r2_cv <- aggregate(R2 ~ Repetition, data =
results_cv_meteo_cv, FUN = mean)
average_results_r2_all <- aggregate(R2 ~ Repetition, data =
results_cv_meteo_all, FUN = mean)
print(mean(average_results_r2$R2))
print(mean(average_results_r2_cv$R2))
print(mean(average_results_r2_all$R2))

average_results_rmse <- aggregate(RMSE ~ Repetition, data =
results_cv_meteo, FUN = mean)
average_results_rmse_cv <- aggregate(RMSE ~ Repetition, data =
results_cv_meteo_cv, FUN = mean)
average_results_rmse_all <- aggregate(RMSE ~ Repetition, data =
results_cv_meteo_all, FUN = mean)
print(mean(average_results_rmse$RMSE))
print(mean(average_results_rmse_cv$RMSE))
print(mean(average_results_rmse_all$RMSE))



############################################
################# Prediction ###############
############################################
#Predictors to be zeroed out during prediction
my_exclude <- c("s(year)" , "s(Sampling_Effort)")

#predict
df_train$prediction <- predict(myfit, type = "response", newdata =
df_train)
df_train$prediction_exc <- predict(myfit, type = "response", exclude
= my_exclude, newdata = df_train)

#difference between prediction and true count
print(sum(df_train$count)-sum(df_train$prediction, na.rm = TRUE))

#Plot without my_exclude
plot <- ggplot(df_train, aes(x = as.Date(date))) +
  geom_line(aes(y = count, color = "Observed Count"), linewidth = 0.7)
+
  geom_line(aes(y = prediction, color = "Predicted Count"), linewidth
= 0.9, linetype = "dashed") +
  labs(
    x = "Date",
    y = "Count"
  ) +
  scale_color_manual(values = c("Observed Count" = "blue", "Predicted
Count" = "red")) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),

    axis.title.y = element_text(size = 14),

    axis.text = element_text(size = 12),

    legend.title = element_blank(),

    legend.text = element_text(size = 12),

    plot.title = element_text(size = 16, face = "bold", hjust = 0.5)

  )

    print(plot)


    ggsave(plot                    =                    last_plot(),
"D:/Masterarbeit/Figures/Meteo_pred.png")

#Plot with my_exclude

    plot_exc <- ggplot(df_train, aes(x = as.Date(date))) +

      geom_line(aes(y = count, color = "Observed Count"), linewidth =
0.7) +

      geom_line(aes(y = prediction_exc, color = "Predicted Count"),
linewidth = 0.9, linetype = "dashed") +

      labs(

        x = "Date",

        y = "Count"

      ) +

      scale_color_manual(values  =  c("Observed  Count"  =  "blue",
"Predicted Count" = "red")) +

      theme_minimal() +

      theme(

        axis.title.x = element_text(size = 14),

        axis.title.y = element_text(size = 14),

        axis.text = element_text(size = 12),

        legend.title = element_blank(),

        legend.text = element_text(size = 12),

        plot.title = element_text(size = 16, face = "bold", hjust =
0.5)
```

```
      )


    print(plot_exc)

    ggsave(plot                    =                    last_plot(),
"D:/Masterarbeit/Figures/Meteo_pred_exc.png")

#Root Mean Square Error

    mse_gam <- mean((df_train$count - df_train$prediction)^2, na.rm =
TRUE)

    rmse_gam <- sqrt(mse_gam)

    rmse_gam


write.csv(df_train,
"D:/Masterarbeit/Jupyter/Data/Meteo/Predicted/Meteo_pred.csv")


############################################

########## Feature Importance ##############

############################################

#transfrom to tibble

ds <- as_tibble(df_train)

mynsim = 30

#Seed for reproducability

set.seed(666)


# Select the relevant columns

cn_s <- colnames(myfit$model)

train_sel_s <- ds %>%

  dplyr::select(any_of(cn_s)) %>%

  dplyr::select(-count)


# Define the target variable

target_s <- ds$count
```

```
#define pred_wrapper since we want to exclude year and Sampling_Effort

pfun_prob <- function(object, newdata) {

  # prediction wrapper

  predict.gam(object, newdata = newdata, type = "response", exclude =
my_exclude, select = TRUE)}


# Calculate variable importance

result_s <- vi_permute(

  nsim = mynsim,

  object = myfit,

  train = train_sel_s,

  target = target_s,

  metric = "RMSE",

  pred_wrapper = pfun_prob

) %>%

  dplyr::arrange(-Importance) %>%

  dplyr::mutate(process = "Mosquito Presence")

###


# Ensure the Variable column is ordered by Importance in descending
order

result_s <- result_s %>%

  arrange(desc(Importance)) %>%  # Sort data frame

  mutate(Variable = factor(Variable, levels = rev(Variable)))   #
Reverse factor levels

result_s


# Sort the data by Importance in descending order

Importance_Mosquito <- ggplot(result_s, aes(x = Importance, y =
Variable)) +

  geom_point(color = "#747474", size = 3) +
  geom_errorbarh(aes(xmin = Importance - StDev, xmax = Importance +
StDev),
                 height = 0.2, color = "#747474", linewidth = 0.5) +
  theme_minimal(base_size = 16) +
  labs(title = "Feature Importance in temporal model", x =
"Importance", y = "Variable") +
  theme(axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(size = 18, face = "bold"),
        panel.border = element_rect(color = "black", fill = NA, size
= 1))
Importance_Mosquito


#When we see negative values for the feature importance, it can happen
that the predictions on the shuffled
#data are more accurate than the real data. This occurs when the
feature doesn’t matter,
#but random chance causes the predictions on the shuffled data to be
more accurate. (https://someshfengde.medium.com/machine-learning-
explainability-permutation-importance-7a9a69bf5943)
```

Visualization and classification

```
rm(list = ls(all.names = TRUE))
gc()


library(terra)
library(raster)
library(dplyr)
library(ggplot2)
library(sf)
library(here)
library(stars)
library(magick)
library(sf)

library(tibble)

library(patchwork)

library(purrr)

library("biscale")

library(grid)

library(scales)


colors <- c("#007800", "#FFDC00", "#FF0000")

custom_palette <- colorRampPalette(c('#006837', '#1a9850', '#66bd63',
'#a6d96a', '#d9ef8b', '#fee08b', '#fdae61', '#f46d43', '#d73027',
'#a50026'))


# Load objects

    Suitability_graz                                               <-
st_read("D:/Masterarbeit/Jupyter/Data/Probabilities/sf_predicted_gra
z.gpkg",  layer = "layer_name")

    Meteo_pred                                                     <-
read.csv("D:/Masterarbeit/Jupyter/Data/Meteo/Predicted/Meteo_pred.cs
v")

    Mos_graz                                                       <-
st_read("D:/Masterarbeit/Jupyter/Data/QGIS/Mosquito/Mosquitos_val1_3
1287.shp")


#Cleaning and date manipulation

    Mos_graz <- Mos_graz %>% rename(date = creation_d)

    Mos_graz$date <- as.character(Mos_graz$date)

    Meteo_pred   <-   Meteo_pred   %>%   dplyr::select(date,   count,
creation_y, prediction, prediction_exc)


### Plot frequency of index values and presence observations###

    #Breaks for histograms

    breaks <- seq(0, 1, by = 0.025)
```

```
  # Calculate presence observations per bin
     presence_counts_graz <- Suitability_graz %>%
      # Cut into bins based on prob_rs
      mutate(bin = cut(prob_pps, breaks = breaks, include.lowest
= TRUE)) %>%
      # Group by bin
      group_by(bin) %>%
      # Count the number of presences
      summarise(presence_count = sum(Presence == "1", na.rm =
TRUE))

      #Add midpoint value for plotting
      # For graz, last bin is missing
      presence_counts_graz <- presence_counts_graz %>%
       mutate(midpoint = 0.0125 + (row_number() - 1) * 0.025)

  # Calculate scale factor between the axes
      max_freq_graz <- max(hist(Suitability_graz$prob_rs,
breaks = breaks, plot = FALSE)$counts)
      max_presence_graz <-
max(presence_counts_graz$presence_count)

      #Determine overall maximum for y-limit
      max_presence_overall <- max(max_presence_graz)
      scale_factor_graz <- max_freq_graz / max_presence_graz

   ### Plotting ###
   ## Graz
   a <- ggplot(Suitability_graz, aes(x = prob_rs)) +
    # Histogram for frequencies
    geom_histogram(breaks = breaks, fill = "steelblue", color =
"white", alpha = 0.7) +
        # Line for presence counts scaled to primary axis
        geom_line(
          data = presence_counts_graz,
          aes(
            x = midpoint,
            y = presence_count * scale_factor_graz
          ),
          color = "red",
          size = 1
        ) +
        # Points for presence counts scaled
        geom_point(
          data = presence_counts_graz,
          aes(
            x = midpoint,
            y = presence_count * scale_factor_graz
          ),
          color = "red",
          size = 2
        ) +
        # Add secondary axis
        scale_y_continuous(
          name = "Frequency",
          sec.axis = sec_axis(~./scale_factor_graz, name =
"Presences")
        ) +
        labs(
          x = "Habitat suitability [%]",
          title = "Frequency of habitat suitability indices and number
of observed presences within in each bin"
        ) +
        theme_minimal() +
        theme(
          axis.title.y.right = element_text(angle = 90)
        )
       plot(a)

ggsave("D:/Masterarbeit/Figures/Distribution_HS_Index_with_presences
.png",
                plot = a,
                height = 4, width = 6, units = "in", dpi = 300)


##### Calculate classification thresholds based on percentiles #####
    #### For Habitat suitability #####
    #Identify presence pixels
    obs_graz <- Suitability_graz[Suitability_graz$Presence == "1", ]


    # New df with probabilites of presence cells. sort descending and
calculate the percentage of observations
      ind_val <- data.frame(Probability = obs_graz$prob_rs) %>%
        arrange(desc(Probability)) %>%  #sort descending
        dplyr::mutate(
          observations = (row_number() / n()) * 100 #percentage of
observations
          )


      #define threshholds
      p25   <-   ind_val   %>%   filter(abs(observations   -   25)   ==
min(abs(observations - 25))) %>% slice(1) %>% pull(Probability)
      p95   <-   ind_val   %>%   filter(abs(observations   -   95)   ==
min(abs(observations - 95))) %>% slice(1) %>% pull(Probability)
      p0    <-    ind_val    %>%       filter(abs(observations)    ==
min(abs(observations))) %>%  slice(1) %>%  pull(Probability)
```

```
      ### For Meteo ###
      #Identify presence pixels
      obs_meteo <- Meteo_pred %>%
        semi_join(Mos_graz, by = "date")

      #New df with probabilites of presence cells. sort descending and
calculate the percentage of observations
      ind_val_meteo <- data.frame(Predicted =
obs_meteo$prediction_exc) %>%
        arrange(desc(Predicted)) %>%
        dplyr::mutate(
          observations = (row_number() / n()) * 100
        )

      #define threshholds
      p25_meteo <- ind_val_meteo %>% filter(abs(observations - 25) ==
min(abs(observations - 25))) %>% slice(1) %>% pull(Predicted)
      p95_meteo <- ind_val_meteo %>% filter(abs(observations - 95) ==
min(abs(observations - 95))) %>% slice(1) %>% pull(Predicted)
      p0_meteo  <-  ind_val_meteo  %>%   filter(abs(observations)  ==
min(abs(observations))) %>%  slice(1) %>%  pull(Predicted)

### Plotting cumulative distribution plots ###
###### For Habiat suitability ###
   theme_set(
     theme_test() +
      theme(
        legend.position.inside = c(0.8, 0.8),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b
= 0, l = 0), size = 8),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b
= 0, l = 0), size = 8),
        axis.text = element_text(size = 8)
        )
    )


    prob_reclass <- ggplot(ind_val, aes(x = observations, y =
Probability)) +
      annotate("rect", xmin = 0, xmax = 100, ymin = 0, ymax = p95,
alpha = 0.6, fill = "#007800") +
      annotate("rect", xmin = 0, xmax = 100, ymin = p95, ymax = p25,
alpha = 0.6, fill = "#FFDC00") +
      annotate("rect", xmin = 0, xmax = 100, ymin = p25, ymax = p0,
alpha = 0.6, fill = "#FF0000") +
      geom_line(linewidth = 0.6) +
      geom_vline(xintercept = c(25, 95), linetype = "dotted", alpha =
0.6, color = "white", linewidth = 0.6) +
      geom_hline(yintercept = c(p95, p25), color = "white", alpha =
0.6, linewidth = 0.6) +
      xlab("Proportion of presence cells [%]") + ylab("Habitat
suitability") +
      theme(
        plot.title = element_text(face = "bold",  hjust = 0.5),
        axis.title.x = element_text(size = 18),
        axis.title.y = element_text(size = 18),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14))
    prob_reclass


    # Save plot
    ggsave("D:/Masterarbeit/Figures/Classification_thresholds.png",
           plot = prob_reclass)
    #Define classifiaction matrix for bivariate mapping

    reclass_matrix <- matrix(c(0, p95, 1, p95, p25, 2, p25, Inf, 3),
ncol = 3, byrow = TRUE)


  ### Now the same for temporal model ###

    pred_reclass <- ggplot(ind_val_meteo, aes(x = observations, y =
Predicted)) +

      annotate("rect", xmin = 0, xmax = 100, ymin = 0, ymax =
p95_meteo, alpha = 0.6, fill = "#007800") +

      annotate("rect", xmin = 0, xmax = 100, ymin = p95_meteo, ymax =
p25_meteo, alpha = 0.6, fill = "#FFDC00") +

      annotate("rect", xmin = 0, xmax = 100, ymin = p25_meteo, ymax =
p0_meteo, alpha = 0.6, fill = "#FF0000") +

      geom_line(linewidth = 0.6) +

      geom_vline(xintercept = c(25, 95), linetype = "dotted", alpha =
0.6, color = "white", linewidth = 0.6) +

      geom_hline(yintercept = c(p95_meteo, p25_meteo), color =
"white", alpha = 0.6, linewidth = 0.6) +

      xlab("Proportion of days [%]") + ylab("Predicted count") +

      theme(

        plot.title = element_text(face = "bold",  hjust = 0.5),

        axis.title.x = element_text(size = 18),

        axis.title.y = element_text(size = 18),

    axis.text.x = element_text(size = 14),

    axis.text.y = element_text(size = 14))

    pred_reclass


  # Save plot

ggsave("D:/Masterarbeit/Figures/Classification_thresholds_meteo.png"
,

         plot = pred_reclass)
  ### For all days including days without observations (e.g., winter)
###

  ind_val_meteo_2 <- data.frame(Predicted =
Meteo_pred$prediction_exc) %>%

    arrange(Predicted) %>%

    dplyr::mutate(

      observations = (row_number() / n()) * 100

    )


  # Calculate proportion of all days within each class

  prob_low_risk <- mean(Meteo_pred$prediction_exc < p95_meteo, na.rm
= TRUE) * 100

  prob_medium_risk <- mean(Meteo_pred$prediction_exc >= p95_meteo &
Meteo_pred$prediction_exc < p25_meteo, na.rm = TRUE) * 100 +
prob_low_risk

  prob_high_risk <- mean(Meteo_pred$prediction_exc >= p25_meteo &
Meteo_pred$prediction_exc <= p0_meteo, na.rm = TRUE) * 100 +
prob_medium_risk


  #Dtermine y-lim

  y_max <- max(ind_val_meteo_2$Predicted, na.rm = TRUE)


  #Plotting

  pred_reclass_2 <- ggplot(ind_val_meteo_2, aes(x = observations, y =
Predicted)) +

    annotate("rect", xmin = 0, xmax = prob_low_risk, ymin = 0, ymax =
y_max, alpha = 0.6, fill = "#007800") +

    annotate("rect", xmin = prob_low_risk, xmax = prob_medium_risk,
ymin = 0, ymax = y_max, alpha = 0.6, fill = "#FFDC00") +

    annotate("rect", xmin = prob_medium_risk, xmax = 100, ymin = 0,
ymax = y_max, alpha = 0.6, fill = "#FF0000") +

    geom_line(linewidth = 0.6) +

    geom_vline(xintercept = c(prob_low_risk, prob_medium_risk),
linetype = "dotted", alpha = 0.6, color = "white", linewidth = 0.6) +

    xlab("Proportion of day [%]") + ylab("Predcited count") +
    ggtitle("Cumulative distribution of days between 2022 and 2024")
+
    theme(
      plot.title = element_text(face = "bold",  hjust = 0.5),
      axis.title.x = element_text(size = 12),
      axis.title.y = element_text(size = 12))

  pred_reclass_2

  # Save plot
  ggsave("D:/Masterarbeit/Figures/Distribution_risk_all_days.png",
         plot = pred_reclass_2,
         height = 4, width = 6, units = "in", dpi = 300)

# Apply classification threshholds to Meteo_pred and Suitability
  Meteo_pred <- Meteo_pred %>% mutate(
    classification = case_when(
      prediction_exc < p95_meteo ~0, # low risk
      prediction_exc >= p95_meteo & prediction_exc < p25_meteo ~ 1, #
Medium risk
      prediction_exc >= p25_meteo ~ 2  # High risk
    )
  )

  Suitability_graz <- Suitability_graz %>% mutate(
    classification = case_when(
      prob_rs < p95  ~ 1, # Low risk
      prob_rs >= p95 & prob_rs < p25 ~ 2, # Medium risk
      prob_rs >= p25  ~ 3  # High risk
    )
  )
  ### Plot seasonality of weather risk ###

  Meteo_pred$date <- as.Date(Meteo_pred$date)

  ggplot(Meteo_pred,  aes(x  =  date,  y  =  prediction_exc,  fill  =
factor(classification))) +

    geom_col(color = NA) +

    labs(

      x = "Date",

      y = "Predicted count"

    ) +

    theme(

      legend.position = "right",

      panel.grid.major = element_line(color = "gray80"),

      panel.grid.minor = element_line(color = "gray90"),

      axis.title.x = element_text(size = 14),

      axis.title.y = element_text(size = 14),

      axis.text.x = element_text(size = 12),

      axis.text.y = element_text(size = 12),

      legend.text = element_text(size = 10)

    ) +

    scale_fill_manual(

      name = "Risk level",

      values = c("#007800", "#FFDC00", "#FF0000"),

      breaks = c("0", "1", "2"),

      labels = c("low", "medium", "high")

    ) +

    scale_x_date(

      breaks = date_breaks(width = "6 months"),

      labels = date_format("%b %Y")

    )
```

```
###### Apply classification to suitability map #####

    map                                                          <-
rast("D:/Masterarbeit/Figures/Probabilities_pred_graz_rs.tif")


    #reclassify raster using classification matrix

    classified <- classify(map, reclass_matrix)

    classified[classified == 0] <- NA

    plot(classified)

    # Save classified raster

    writeRaster(classified,
"D:/Masterarbeit/Figures/Habitat_classified_graz.tif",
overwrite=TRUE)


    # Now as PNG plot

    df_plot <- as.data.frame(classified, xy = TRUE)

    df_plot$Class  <-  factor(df_plot$Probabilities_pred,  labels  =
c("Low", "Medium", "High"))


    ggplot(df_plot, aes(x = x, y = y, fill = Class)) +

      geom_raster() +

      scale_fill_manual(values = colors, na.translate = FALSE) +

      coord_equal() +

      theme_minimal() +

      theme(legend.position = "right",

            plot.title = element_text(face = "bold", hjust = 0.5)) +

      labs(title = paste("Habitat suitability of Aedes albopictus in
Graz"),

           fill = "Suitability_graz",

           x = "Longitude",

           y = "Latitude")


    ggsave("D:/Masterarbeit/Figures/Habitat_classified_gg.png",
         plot = last_plot(), width = 10, height = 8, dpi = 300)

    #### Bivariate mapping: Suitability vs Uncertainty ####

    #Check distribution of uncertainty

png("D:/Masterarbeit/Figures/Prediction_Uncertainty_Histogram.png",
width = 800, height = 600, res = 150)

    hist(Suitability_graz$sd_pred,    breaks    =    20,    xlab    =
expression(sigma))
    dev.off()

      # Create three classes of uncertainties using quantiles
      Suitability_graz <- Suitability_graz %>% mutate(
        uncertainty_lvl      =      cut(sd_pred,      breaks      =
classInt::classIntervals(
          var = sd_pred, n = 3, style = "quantile"
        )$brks, include.lowest = TRUE, dig.lab = 3, labels = c("1",
"2", "3"))
      )

    #Create factor for combined classification
    Suitability_graz <- Suitability_graz %>% mutate(
      facet_cat   =   as.factor(paste0(classification,   "   -   ",
uncertainty_lvl))
    )

    color_palette <- c(
      "1 - 1" = "#e8e8e8",
      "1 - 2" = "#ace4e4",
      "1 - 3" = "#5ac8c8",
      "2 - 1" = "#dfb0d6",
      "2 - 2" = "#a5add3",
      "2 - 3" = "#5698b9",
      "3 - 1" = "#be64ac",
      "3 - 2" = "#8c62aa",
      "3 - 3" = "#3b4994"
    )
    #Plotting
    ggplot(Suitability_graz, aes(x = X, y = Y, color = facet_cat)) +
      geom_point() +
      scale_color_manual(values = color_palette) +
      coord_equal() +
      theme_minimal() +
      theme(
        strip.text = element_text(face = "bold"),
        legend.position = "none"
      ) +
      labs(
        x = "Longitude",
        y = "Latitude"
      )

    ggsave("D:/Masterarbeit/Figures/Facet_plot.png",
           plot = last_plot(), dpi = 600)


#########################################
#### Combine spatial and temporal #######
#########################################

    #Extract habitat suitability values and transform to sf
    suitability_prob <- Suitability_graz[, c("X", "Y", "prob_rs")]
    suitability_prob <- suitability_prob %>%
      st_as_sf(coords = c("X", "Y"), crs = 31287)

    #Normalize suitability values
    suitability_prob$prob_norm                                    <-
suitability_prob$prob_rs/sum(suitability_prob$prob_rs)

    #Select 2023 for time series
    Meteo_pred_GIF <- Meteo_pred[Meteo_pred$creation_y==2023,]

    #Sort according to date
    Meteo_pred_GIF <- Meteo_pred_GIF %>%
      arrange(date)

    rownames(Meteo_pred_GIF) <- NULL

    # Select only mosquito season
    Meteo_pred_GIF <- Meteo_pred_GIF[121:334,]
    Meteo_pred_GIF <- na.omit(Meteo_pred_GIF)

    #Transform normalized df to raster
    target_resolution <- 100
    extent_data <- st_bbox(suitability_prob)  # extent of Raster

    coords <- st_coordinates(suitability_prob)
    probs <- suitability_prob$prob_norm

    # Create empty raster
    raster <- raster(extent(extent_data),
                       res = c(target_resolution, target_resolution),
                       crs = st_crs(suitability_prob)$proj4string)
    # Transform df to spatial point data

    spdf <- SpatialPointsDataFrame(coords, data = data.frame(probs),
proj4string = CRS(st_crs(suitability_prob)$proj4string))


    # Rasterize the spdf using the probability field

    raster_norm <- rasterize(spdf, raster, field = "probs", na.rm =
TRUE)

    #save raster

    png_filename <-
"D:/Masterarbeit/Figures/suitability_raster_normalized.png"

    png(png_filename, width = 800, height = 600)

    plot(raster_norm, main = "suitability_raster_normalized", col =
custom_palette(100), zlim = c(0, max(suitability_prob$prob_norm)))

    dev.off()


    # Create one abundance raster per day

      #Initilize stack

      abundance_stack <- stack()


      for (i in seq_along(Meteo_pred_GIF$prediction_exc)) {

        #distribute daily predicted mosquito counts using the
normalized spatial habitat suitability scores

        abundance_raster <- raster_norm *
Meteo_pred_GIF$prediction_exc[i]

        abundance_stack <- addLayer(abundance_stack,
abundance_raster)

        }


      # Calculate highest possible Value for upper zlim

      max_value <- max(values(abundance_stack), na.rm = TRUE)

      #transform values to consisten scale (0 to 1)
      abundance_stack_norm <- calc(abundance_stack, fun = function(x)
x / max_value)
      #New max value for zlim
      max_value_2 <- max(values(abundance_stack_norm), na.rm = TRUE)


    #Create GIF
    #Temporarily write images from raster stack
    images <- list()
    for (i in 1:nlayers(abundance_stack_norm)) {
      plot_file <- paste0("D:/Masterarbeit/Figures/Temp/plot_day_",
i, ".png")
      png(plot_file, width = 800, height = 600)
      plot(abundance_stack_norm[[i]], main =
as.character(Meteo_pred_GIF$date[i]),
           col = custom_palette(100),
           zlim = c(0, max_value_2),
           legend.args = list(text = "Risk", side = 3, line = 1.2, cex
= 0.95, adj = 0.5, font = 2))
      dev.off()
      images[[i]] <- plot_file
    }


    #Read images
    img_list <- lapply(images, image_read)


    #Combine to GIF
    gif <- image_animate(image_join(img_list), fps = 25)
    image_write(gif, "D:/Masterarbeit/Figures/space-
time_animation.gif")


    print(gif)
```