Modelling the spatial abundance distribution of *Melanargia galathea* in a changing environment

Master Thesis Gaetano Paganini

April 2016

WSL, Swiss Federal Research Institute for Forest, Snow and Landscape Research
ETH Zurich, Department of Environmental Systems Science

Supervisors
PD Dr. Janine Bolliger
Dr. Rafael Wüest

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Modelling the spatial abundance distribution of Melanargia galathea in a changing environment

Authored by (in block letters):

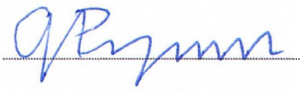
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Paganini	Gaetano

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Gossau SG, 3.4.2016	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

1 Abstract

This *Master thesis* assesses and predicts the abundance and spatial distribution of the butterfly *Melanargia galathea* in Switzerland and Liechtenstein for present and future environmental conditions. The main objective was to explain the spatial pattern of butterfly monitoring observations of *M. galathea* at 474 locations as function of landscape, distance, relief and climate variables across Switzerland and Liechtenstein. For this, different circular areas around each butterfly observation location were used as surrogates for the movement potential of the species, within which the explanatory variables were calculated. Four statistical models and two different ensemble model approaches were used to project the butterfly abundances spatially explicitly and to analyse the influence of the different models on the predictions.

Major results showed that temperature and slope have a strong positive effect on the abundance of *M. galathea*, while the proportion of buildings should not be over 25% and the annual precipitation total is optimal around 800 mm. Currently, the hotspot of *M. galathea* abundances are located in the Swiss Alps (Valais, Ticino, Graubünden) and in the Jura Mountains. I used two climate warming scenarios (RCP4.5 and RCP8.5). The low warming scenario RCP4.5 (+4°C; -20% precipitation in 2100) has a rather negative effect on the abundances of *M. galathea* in lower elevations and a rather positive effect in higher elevations. The same is the case with the strong warming scenario RCP8.5 (+6°C; -40% precipitation in 2100), but the positive effect on higher elevations is more distinct. Under warming an expansion of the distribution ranges are observed in the Valais, the northern edge of the Swiss Alps, partly in the Jura Mountains and in the continental High Alps. A range contraction is to expected in Ticino, Graubünden, Swiss Plateau and partly in the Jura Mountains.

The comparison of the predicted abundances of *M. galathea* for the present and under climate change with three land-use scenarios, revealed no influence of these land-use scenarios on the spatial abundance distribution of *M. galathea*.

Keywords: *M. galathea* · Landscape · Geographical distance · Relief · Climate · Ensemble modelling · Spatial projection · Climate change · Land-use change

2 Introduction

The butterfly *M. galathea* is a widespread species in Europe. Baguette et al. (2000) analysed the dispersal capacity and behaviour and have discovered that an increasing distance between patches decreases the probability of movements between local populations and the number of moving butterflies. Vandewoestijne et al. (2004) analysed the landscape occupancy and genetic population structure in Belgium and Hungary. They discovered that the genetic polymorphism within the sampled populations was high and the genetic differentiation between population was low. This is characteristic for species with high dispersal capacity and/or high density. They also showed, that the dispersal was influenced by the spatial distribution of the habitat patches.

A first focus of my thesis was therefore to assess the spatial distribution of the abundance

of *M. galathea*. My analysis revealed the influence of range of environmental characteristics and elements of the anthroposphere on *M. galathea* to better understand the spatial abundances of the species which allows to come up with scientific and evidence-based conservation measure for the species. An additional interesting question to investigate referred to the landscape occupancy of the species under environmental change. Both, climate and land-use change are important drivers which will shape future landscapes and the availability and spatial arrangement of suitable habitat for biodiversity (Maggini et al. (2014); Bolliger et al. (2011)). A second aim of my thesis was therefore to assess the influence of climate and land-use change scenarios on the spatial abundance distribution of *M. galathea*. Hence assessment of their influence could be crucial for *M. galathea* conservation and management methods. I used three land-use scenarios from Price et al. (2015).

To accomplish these goals the observed abundance at 474 locations in Switzerland was explained by landscape, distance, relief and climate variables at these locations with four statistical models. The spatial abundance information was provided from the Swiss biodiversity monitoring (BDM), which is a program of the Federal Office for the Environment (FOEN). The explanatory variables were calculated with the geographic information system software ArcGIS, with data from FOEN, Federal Office of Topography (Swisstopo) and Swiss Federal Research Institute for Forest, Snow and Landscape Research (WSL). The explanatory variables were calculated for different buffer zones at theses 474 locations to fit the models and to predict the abundance of *M. galathea* in a 1x1 km raster grid in Switzerland and Liechtenstein for spatial projection.

I used two ensemble model approaches for the predictions and to analyse their properties (differences in performance of prediction). The first ensemble model approach combines one model from each model type (GLM, GAM, RPART, RF) and the second ensemble model approach combines all models from the complete variable combination set (Fig. 4). For the abundance distribution of *M. galathea* the ensemble models were used and for the range shifts the difference between ensemble models from the present and the future were taken. The range shift is useful to note and assess the range expansion and range contraction. For the comparison of different ensemble models a two-sided and paired t-test was used. The best best performing ensemble model was used for the prediction in the future with climate change and compared with the land-use change.

3 Materials and Methods

3.1 Study area and species

The study area encompassed the whole of Switzerland and Liechtenstein (ca. 42'000 km^2). The climate of Switzerland is heavily influenced by the Atlantic Ocean. The prevailing currents from westerly directions push, humid air to Switzerland, causing a cooling effect in summer and a warming effect in winter. The Alps act as a prominent climatic barrier between northern and southern Switzerland. The southern part is characterized by an insubric climate, with mild winters and warm summers. The inner-alpine valleys are characterised by a distinct continental climate, because they are sheltered against precipitation from both the north as well as the south. As a consequence, dry conditions prevail in these regions. Starting at an altitude of 1200-1500 m above sea level, precipitation during winter usually occurs as snowfall, such that the area is often covered by a solid layer of snow for months. The temperatures in Switzerland are primarily dependent on the

level of altitude (MeteoSwiss, 2016). Approximately 35.9% of Switzerland is covered by agriculture, 7.5% by settlement and 35.9% by wooded area (SFSO, 2013). The butterfly *M. galathea* is common in Europe, North Africa, Turkey and in Transcaucasia. Its habitat varies from grass and flower dominated overgrown sites. The imagos feed nectar from flowers of Centaurea, Scabiosa, Cirsium and Carduus and fly from June to August, seldom from May to September (Tolman et al., 1998).

3.2 Concept of analysis

The goal of this analysis was twofold. First, the study aimed to identify the drivers that determine the spatial distribution of *M. galathea* abundance. Second, the identified drivers are used to project the abundance of *M. galathea* for current and future climatic conditions across Switzerland using an ensemble modelling approach. Also land-use change scenarios for 2035 were used to estimate their influence on the abundance distribution of *M. galathea* qualitatively. The workflow highlighting the steps conducted in this thesis is illustrated in Figure 1. First, the datasets are described that were used in the individual and ensemble models. One data set contained the abundance data of *M. galathea*, derived from the Swiss Biodiversity Monitoring (BDM; section 3.3). The other two datasets contain the explanatory variables used to identify major environmental drivers of the spatial butterfly distribution (section 3.5), these two datasets were calculated for different buffer zones (section 3.4). Here, I generated two data sets because one dataset was needed to calibrate the predictive models, relying on 474 *M. galathea* observation locations. The second dataset was used to spatially project *M. galathea* abundances, relying on 41'442 locations on a 1x1 km raster grid across Switzerland and Liechtenstein. Second, the modelling framework is explained (section 3.6). And finally, the spatial projections for M. galathea for current (section 3.7) and future environmental conditions are described (section 3.8).

3.3 Abundance data

In 2001, the Federal Office for the Environment (FOEN) in Switzerland started a biodiversity monitoring programme (BDM), with the task to monitor the biological diversity in Switzerland (FOEN, 2014). The Figure 3 shows hotspots in the Jura Mountains, Valais and Ticino, low abundance in the Swiss Plateau and huge gaps of *M. galathea* across Switzerland. The Swiss Plateau has many observation locations with less than 10 bufferflies, especially in the center and west of Switzerland. In total the BDM contained 474 locations with *M. galathea* observations. The monitoring is designed such that each site is visited every five years, so roughly 20% (ca. 90 locations) of the overall BDM sites are visited every year (Fig. 2(a)). Each year from May to August (rarely also at the end of April and at the beginning of September), surveyors walked at distinct dates and a distinct time frame through a defined 2'500 m long transect and identify and counted butterfly individuals (FOEN, 2012a). The midpoint of these transects was used as the butterfly observation location. These surveys were repeated maximally seven times a year. For analysis, an equal amount of measurements per location is very important, hence seven measurements per location and year would be optimal. However measurements were missing in May, June, at the end of August (also a the end of April and at the beginning of September), when the observations of *M. galathea* were mostly low or zero. Also, many locations had years with no observations. Between 2003 and 2014, the time frame accounted for in this thesis, every location had maximally three years with a complete

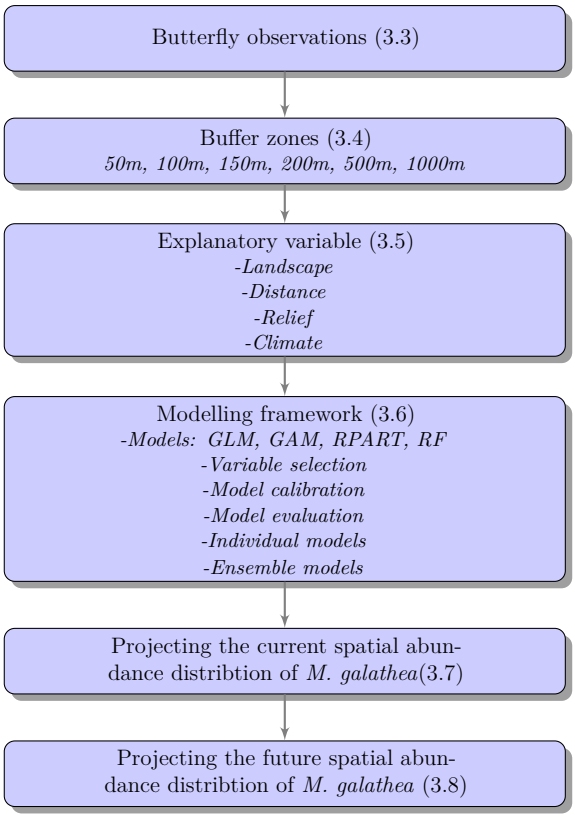
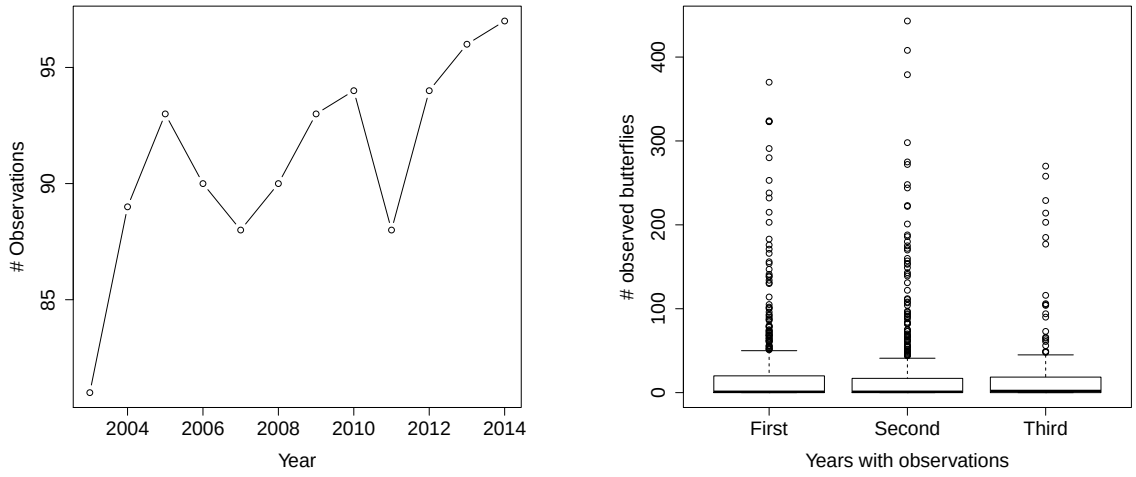


Figure 1: The workflow in this thesis, the number in the brackets refers to the sections.

set of measurements, but every location had at least one year with a complete set of measurements. To make the abundance data comparable across years, I first took the mean across the maximal seven measurements per location and year. Second, the mean across the maximally three years with measurements was calculated. Fig. 2(b) shows all observations with measurements in one, two and three years. This figures reveals that using the mean number of observations across maximal three years is appropriate, because the number of observations are in a similar order of magnitude for all three years (Fig. 2(b)). Finally, the mean number of observations was rounded to obtain integers of the average abundance of *M. galathea* per year. Integers were required, because the applied statistical model assumed a Poisson-distribution (section 3.6). The mean observations of a location represents the number of butterflies a person is likely to observe if he/she goes to this location during a summer-day.

3.4 Buffer zones

The abundance of butterflies is dependent on the environment. The environment may affect the actual dispersal range and determine the feeding and habitat availability. To assess to which degree the environment determines the abundance of *M. galathea*, various circular radii (buffer zones) were considered around each butterfly observation location (radii of 50, 100, 150, 200, 500 and 1000 meters). Within these buffer zones, all environmental variables were calculated (section 3.5). These buffer zones were also used to spatially project *M. galathea* abundance in a 1x1 kilometer raster grid across Switzerland and Liechtenstein (sections 3.7 and 3.8).



(a) Observation locations per year

(b) Number of observed bufferflies across observation years

Figure 2: Number of observation locations per year of the analysis (a) and the distribution of the observations for the first, second and third year (b).

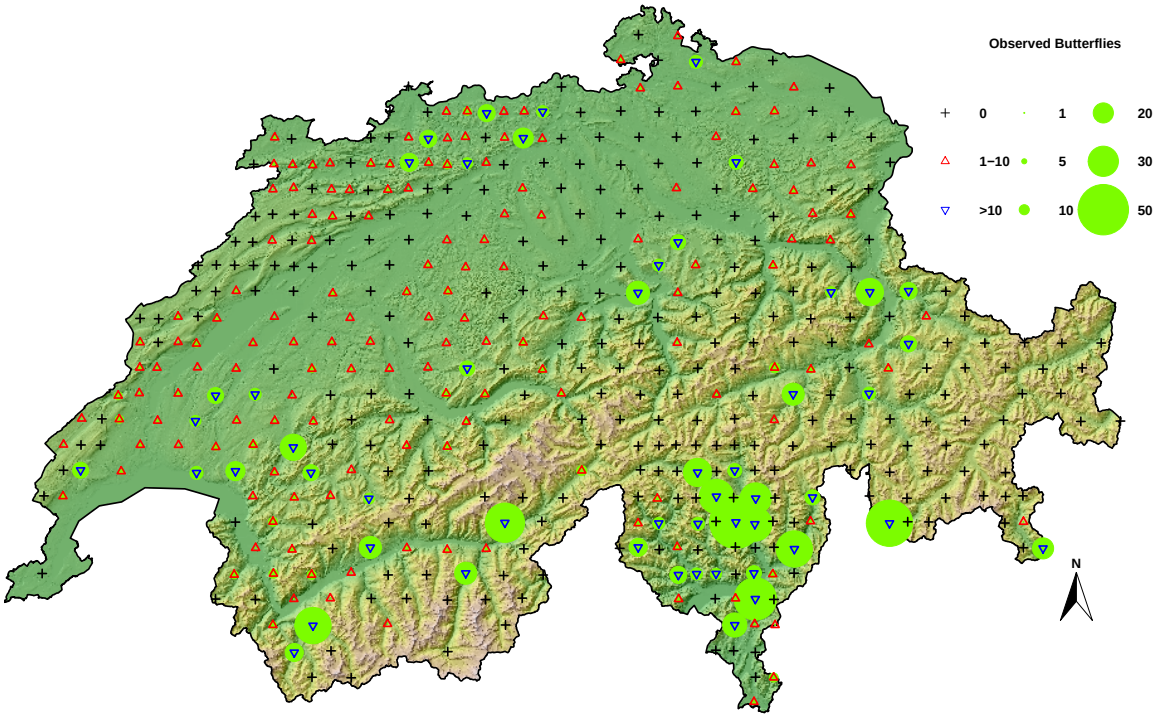


Figure 3: The 474 butterfly observation locations with the mean of observed *M. galathea* per year between 2003 and 2014 across Switzerland.

3.5 Explanatory variables

For the analysis 19 variables were chosen to explain the spatial abundance distribution of *M. galathea* (Table 1). Nine of them related to the landscape, four to geographical

distance, four to relief and two to climatic properties. The variables were generated in ArcGIS version 10.3.1 (ESRI, 2015). The topographical landscape model TLM from the Federal Office of Topography (Swisstopo) covers the whole of Switzerland and Liechtenstein as nationwide basic landscape model. From the TLM, the landscape and geographical distance variables were derived (Table 1). The geographical distance was measured between the butterfly observation locations and the edge of the considered variable. The data for the dry meadows were from FOEN and consist of two geodata raster files (FOEN (2012c) and FOEN (2012d)). From a digital elevation model with a 25x25 m resolution (DEM25; Swisstopo (2005)) the variables elevation, slope (degree) and aspect were calculated (Table 1). The climate variables included annual precipitation total and mean annual temperature and were provided by the Landscape Dynamics unit at WSL, relying on data from Federal Office of Meteorology and Climatology (MeteoSwiss) and spatial interpolation software from Thornton et al. (1997).
The landscape variables represent important elements of the environment, which could be crucial for *M. galathea*. Open land and dry meadow represent landscape elements that include basic food resources of *M. galathea*. Single trees, lines of trees and hedgerows and forests provide hiding places. Roads, railway tracks and buildings represent the influence of modern society on the environment. To which degree *M. galathea* depends on water is not known, hence rivers and lakes were included in the analysis.
The distance variables are surrogates for connectivity of favourable habitats across the landscape, also representing the dispersal capacity of *M. galathea*. Dry meadows are rare in Switzerland and are habitats with many species. The term "dry meadow" is equal to "dry meadows and pastures" in this thesis. Both habitats are highly endangered, hence the Federal Council is monitoring them in an inventory (FOEN, 2010). The great biodiversity and the rarity of dry meadows are the reason to include them into this analysis.

Landscape variables

The landscape variables were all derived from the Swiss TLM and supplemented with the dry meadows inventory (FOEN; Table 1). To account for the area of linear and point landscape elements, buffers were applied. Here the term "buffer size" means a radius, which is applied as radius on points and a buffer on each side of a line. There was no information about the width of lines of trees and hedgerows and single trees, hence the buffer size was estimated by myself, without any background literature. Lines of trees were buffered with a radius of two meters and hedgerows with a radius of one meter. Point features were also attributed an area: single trees were buffered with a radius of two meters. Buildings, represented as footprints in the TLM, were merged into spatially aggregated settlement units using buffers of five meters for each building. The TLM provides information about road widths, sometimes given as intervals. So the chosen buffer size was always the half of the mean of these intervals (Appendix, Table 14). Tunnels, underpasses and underground roads were not considered in the analysis. Also, minor roads such as exit ramps, slip roads, road links, roadhouses and access roads were excluded. Similarly to roads, the TLM provided information on the width of the railways. The buffer width for the respective railway category was always half the mean of the indicated intervals. Also here, tunnels, underpasses, underground rails and no longer used rails were not considered in the analysis (Appendix, Table 15). Surface water bodies were represented by area (lakes, large rivers) and lines (small rivers). Small rivers were buffered with one meter on each side. The spatially explicit information on the area of dry meadows as provided by FOEN (FOEN (2012c), FOEN (2012d)) was merged together. Forest includes closed and open forests and

Table 1: Variables calculated and mostly used to explain *M. galathea* abundance. (*) indicates line or point features, buffered as described in section 3.5

Variable	Data object (ArcGIS)	Data source	Buffer zones
Landscape:			
Line of tree and hedgerow (*)	TLM_BAUM_GEBUESCHREIHE_2015	(Swisstopo, 2015)	50, 100, 150
Single tree (*)	TLM_EINZELBAUM_GEBUESCH_2015	(Swisstopo, 2015)	1000
Forest	TLM_BODENBEDECKUNG_2015	(Swisstopo, 2015)	none
Building (*)	TLM_GEBAEUDE_FOOTPRINT_2015	(Swisstopo, 2015)	50, 100, 150, 200, 500
Road (*) (Table 14)	TLM_STRASSE_2015	(Swisstopo, 2015)	50, 100, 150
Dry meadow	tww and TWW_A2	(FOEN, 2012b)	50, 100, 150, 200, 500
River (*)	TLM_FLIESSGEWAESSER_2015	(Swisstopo, 2015)	all
River, Lake	TLM_BODENBEDECKUNG_2015	(Swisstopo, 2015)	all
Railway (*) (Table 15)	TLM_EISENBAHN_2015	(Swisstopo, 2015)	all
Open land			none
Distance to:			
Line of tree and hedgerow	TLM_BAUM_GEBUESCHREIHE_2015	(Swisstopo, 2015)	all
Single tree	TLM_EINZELBAUM_GEBUESCH_2015	(Swisstopo, 2015)	all
Forest	TLM_BODENBEDECKUNG_2015	(Swisstopo, 2015)	all
Dry meadow	tww and TWW_A2	(FOEN, 2012b)	all
Relief:			
Elevation	DEM25	(Swisstopo, 2005)	none
Slope	DEM25	(Swisstopo, 2005)	50, 100, 150, 200, 500
Aspect NO	DEM25	(Swisstopo, 2005)	all
Aspect SW	DEM25	(Swisstopo, 2005)	none
Climate:			
Temperature	te2003avgy-te2006avgy	(section 3.5)	all
Precipitation	pp2003sumy-pp2006sumy	(section 3.5)	all

are already represented as areas, thus did not require buffering. The remaining landscape not covered by data of the Swiss TLM or dry meadows was assigned open land, which mostly represents intensively managed agricultural land.

Distance variables

Distance between landscape elements determine the permeability of the landscape as they may refer to dispersal and movement distances covered by the species. Here, I accounted for four distance variables which likely facilitate movement of *M. galathea* across the landscape (Table 1). The four distance variables measure the distance from the butterfly observation locations to the next single trees, lines of trees and hedgerows, dry meadows and forests. In the case of dry meadows and forests the distance was measured from the butterfly observation locations to the edge of dry meadows or forests. For single trees, lines of trees and hedgerows the distance was calculated between the butterfly observation locations and the non-buffered lines of trees and hedgerows and of single tree points. The data set for the predictions and spatial projection measured the distance from a 1x1 km point grid in Switzerland to the four distance variables.

Relief variables

Elevation, slope and aspect were chosen to represent the relief. Elevation is most strongly correlated with temperature and hence has probably the same relevance for *M. galathea* as temperature. The aspect can have a big influence on the climate of an area, because

wind strength and direction, sunshine duration, and precipitation are influenced by aspect. So aspect represents different climatic conditions. Slope represents special wind conditions in an area and interacts with the aspect. The range of aspects in an area ranges from non-uniform to uniform aspect. So the slope of an area with a non-uniform aspects implies an area with a high surface roughness, which leads to lower wind speed near the surface and to more wind vortex. The opposite is the case in an area with a nearly uniform aspect distribution.
Elevation, slope and aspect were calculated from the digital elevation model (DEM25) with a resolution of 25x25 m. For elevation and slope, the mean elevation or degree slope was taken across all raster cells within a buffer zone (section 3.4). The aspect was divided into two hemispheric angular sections: one from 315 to 135 degree with the northeast direction as midpoint (alias NO, "cold") and the other the opposite half (alias SW, "warm"). For the analysis the percentage of NO and SW in each buffer size was calculated. This implies, that an area with a NO proportion of 100% is particularly exposed to the morning sun, whereas areas with a SW proportion of 100% are particularly exposed to the midday and the evening sun. This division of the aspect into cold and warm could help to analyse the activity period of *M. galathea*.

Climate variables

The climate variables mean annual temperature and the annual precipitation total were provided by the research unit Landscape Dynamics, WSL. They are available as raster files with a resolution of 100 m and were calculated by interpolating daily measurements of MeteoSwiss weather stations from 1930 to 2013 using the Daymet algorithm of Thornton et al. (1997). The daily values were aggregated to monthly, annual, decadal or 30 year averages. For this analysis the annual averages were used, because they include the influence of all seasons on *M. galathea*, including the influence on the larvae. So a long winter and/or cold autumn and/or cold spring reduces the number of *M. galathea* in the summer. To represent the climate for the available *M. galathea* observation between 2003 and 2014, I averaged the annual mean temperature and the annual precipitation totals between 2003 and 2006. For the annual precipitation total the log was taken, because a certain increment of precipitation at high precipitation amounts has not the same importance than the same increment at low precipitation amounts. This is partly because soil can just store a distinct amount of water, dependent on the physical and chemical conditions and the past development of the soil. The rest of the precipitation will just drain off and will not be available for plants and animals.
The calculation tool used in ArcGIS (namely "Zonal Statistics as Table") could sometimes not calculate the mean annual temperature and the annual precipitation total for a 50 m buffer zone, because these raster pixel size is 25x25 m, which is a too small pixel size. Hence the mean annual temperature and the annual precipitation total from the 100 m buffer zone were also used for the 50 m buffer zone.

Missing data in the explanatory variables

The rasters DEM25, annual precipitation total and mean annual temperature did not cover all butterfly observation locations. To exclude the respective locations was not an option, because all 474 observations were required for the statistical analysis. The rasters annual precipitation total and mean annual temperature are congruent and in both lie one location outside and in the DEM25 lies also one locations outside, but a different one. So

these two locations needed approximated values for the relief and climate variables. But these two locations have one thing in common, namely the buffer zones of 500 and 1000 m at these locations have an intersection with all three rasters. So for these two locations the 50, 100, 150 and 200 m buffer zones have no values for the climate and relief variables and the means of these variables in the 500 and 1000 m buffer zones were used for the smaller buffer zones as an approximation.

3.6 Modelling framework

The modelling framework relies on individual models and on two ensemble model approaches (Fig. 4), which is inspired by Engler et al. (2013). Engler et al. (2013) created an ensemble prediction of tree species occurrence using six different individual modelling techniques and remote sensing predictors. The ensemble model approach for species distribution was also used by Araújo and New (2007) and is widely used in climate sciences. In my thesis, I used four statistical models to model *M. galathea* abundance as a function of environmental predictors. Generalized linear models (GLM; McCullagh and Nelder (1989)) and generalized additive models (GAM; Hastie and Tibshirani (1990); Wood (2009)) represent classical regression analysis and recursive partitioning and regression trees (RPART; Breiman et al. (1984)) and random forests (RF; Breiman (2001)) belong to classification and regression trees. This mix of different methods from different concepts allows to compare the models and also to do an ensemble model. A GLM as implemented in the 'base' R package (Davies, 2016) fits a linear model to explain the dependent variable Y through the independent variables X. The 'mixed GAM computation vehicle' (mgcv) package (Wood, 2016), is similar to a GLM, but it replaces the linear model with a smoothing model and is also called non-parametric regression. The smoothing function was fit with five dimensions, so four degrees of freedom were used from GAM to fit the smoothing function, if it was necessary to fit the explanatory variable properly. As the response variable (abundance of *M. galathea*) constitutes a count variable, we used a Poisson error distribution with a log-link-function for GLM and GAM. This method should convert the data to get a Gaussian distribution of the errors, which was verified with Q-Q plots. Also the component smooth function from GAM was plotted, to analyse the influence of the single variables on the abundance variable. The classification and regression tree methods have no assumptions on the distribution of the errors and no assumption that the response has a linear or smooth relationship with the predictors. RPART generates a binary tree where each node represents a threshold for an explanatory variable. This threshold value decides what direction (left or right) an observation takes. This procedure is continued to reach the terminal nodes, where a prediction for the response variable is assigned to an observation. The RPART tree was estimated with the recursive partitioning method using the 'rpart' package in R (Therneau et al., 2015). A problem of RPART is the overfitting, which can be prevented using random forest (RF), that effectively generate a large number of trees in the form of RPART's, which are based on random samples of observations. RF is an ensemble learning method, because it splits the data into training and testing data set and fits different trees. The final prediction is a combination of these trees, where the average over the prediction of these trees is taken. In R the RF was implemented with the 'randomForest' package (Breiman et al., 2015).

3.6.1 Variable selection

All explanatory variables from the datasets for the model calibration were tested for multicollinearity. To do so, a Spearman correlation matrix was calculated and for each pair of variables with a correlation higher than 0.7, one was excluded from the further analysis. The Spearman correlation was suitable because it also tests for non-linear relationships. The Spearman correlation matrix was calculated for each buffer zone individually, resulting in six variable sets for the corresponding buffer zone (50, 100, 150, 200, 500, 1000m). These six variable sets were also use for the spatial projection in the present and future, without testing for multicollinearity in the datasets for the projections. I didn't exclude variables, which influence on the abundance of *M. galathea* is biological not explainable. For example the relief variable slope were kept in the models. So the aim was a good prediction of the abundance distribution and not a good biological explanation for the used variables. Also the explainable variables railway and dry meadow were converted into factors in all buffer zones, because they are so rare.

3.6.2 Measure of model accuracy

The calculation and assessment of the model accuracy is a crucial part in model evaluation. I used four statistical models for predicting *M. galathea* abundance, which limits the option of model evaluation methods. Here, the accuracy of model predictions was assessed using the root mean square error (RMSE, Eq. 1), a measure for model accuracy which is applicable to all four model types (GLM, GAM, RPAR, RF). RMSE indicates at the scale of the abundance observations how far, on average, predictions are from abundance observations. RMSE (Eq. 1) takes the square root of the mean of the squared deviations of the predicted ($\hat{y}_t$) and observed (y_t) abundances, divided by the sample size n (Stanford, 2016). RMSE was also used as weight for the ensemble models, the weights have two properties. First they add up to one and second the lowest RMSE results in the highest weight and vica versa (Eq. 2). Willmott and Matsuura (2005) claim that the mean absolute error (MAE) is better in assessing average model performance than the RMSE. However, Chai and Draxler (2014) investigated MAE and RMSE performance again, with the conclusion that the RMSE is suitable as long the error distribution is expected to be Gaussian. Here the Gaussian distributed errors have just to be fulfilled from GLM and GAM.

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

$$Weight_i = \frac{RMSE_i^{-1}}{\sum_{i=1}^{n} RMSE_i^{-1}} \tag{2}$$

3.6.3 Model calibration

The individual and ensemble models were calibrated using the observations at the 474 butterfly observation locations and the uncorrelated explanatory variables (section 3.6.1). I used two main approaches to produce ensemble models (Fig. 4), one that combines the four model types and the other combines all possible variable combinations for each of the

four model types. So the first approach combined four models in an ensemble model and the second approach combined 4095 models for each model type, if the 500 m buffer zone with 12 explanatory variables is used, hence the second approach combines 16'380 models in total. Both approaches used the RMSE as weights (section 3.6.2), but the ensemble model with the whole variable combination set was also calculated without weights to analyse the influence of the weights.

3.6.4 Model evaluation

Model evaluation is an important part in each research, which is based on models and spatial and/or temporal projections. It assesses the prediction error of a model, if it is used to unknown data or to data other than the training data set. I evaluated all models using Monte Carlo Cross Validation (MCCV), also called the Repeated Random Sub-Sampling Validation (Xu and Liang, 2001). MCCV randomly splits the dataset into a training and validation data set. I used here 80% of the data as a training data set. After the model was fitted to the training data set, the validation data set was used for the prediction. The prediction accuracy was assessed using the validation (or out-of-bag) data set based on RMSE (OOB-RMSE). This procedure was repeated 100 times resulting in 100 OOB-RMSEs. I assessed OOB-RMSE across the four models and across the six buffer zones.

3.6.5 Individual models

The individual models were fit to meet the following objectives. The first objective was to identify the buffer zone with the best model performance, thus lowest OOB-RMSE distribution. I obtained the OOB-RMSE distribution using MCCV with 100 repetitions for each buffer zone. The resulting distribution of 100 OOB-RMSE values were compared across buffer zone using boxplots. I combined the OOB-RMSE from all four model types to identify the buffer zone with best performance over all four model types. The second objective was i) to identify the significant/important variables in the four models with all variables and ii) the best model from the complete variable combination set. For comparison of i) the full model, ii) the best and iii) worst model from the complete variable combination set, and iv) the ensemble model with the weighted models from the complete variable combination set, the OOB-RMSE from the four model types were used to calculate the two-sided, paired t-test. The influence of a variable on the response variable in the model is measured differently in the GLM/GAM and RPART/RF models. GLM/GAM used the p-value of a variable to assess its significance, while the RPART/RF calculate the importance of a variable in the model.

3.6.6 Ensemble models

The ensemble model approach combines uncertainties from all involved models and transfers them to the spatial projection, hence two ensemble model approaches were used to compare them (Fig. 4). The ensemble models were used to spatial project *M. galathea* abundance for the present and the future conditions. Firstly two ensemble models combined all four model types with the complete variable set and with the variables of the best performing model from the complete variable combination set. The OOB-RMSE of the four model types were used as weights (section 3.6.2).

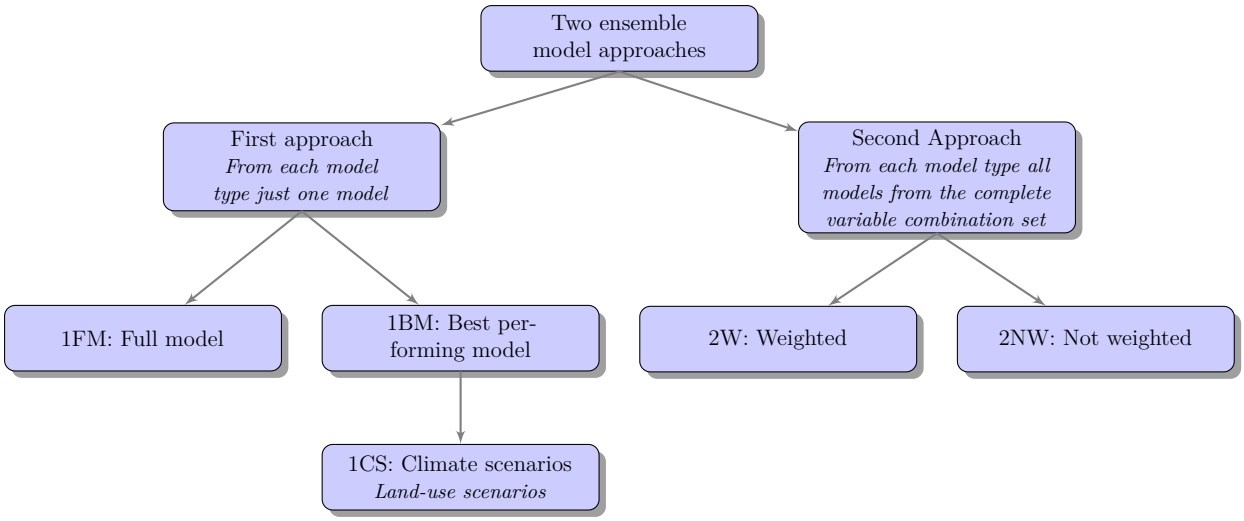


Figure 4: The overview of all ensemble models with the four statistical model types GLM, GAM, RPART and RF.

Secondly, for each of the four model types, all models of the complete variable combination set were combined. I produced two average projections, one non-weighted average, and one weighted mean over all models with the OOB-RMSEs as weights. For both ensemble models I then further combined the four ensemble models using a weighted mean with the OOB-RMSEs as weights. To complete this ensemble model the variance and quantiles were calculated for each of the four model types. The quantiles are for both ensemble models the same, but the weighted mean (Eq. 3) and the unbiased weighted estimator of the sample variance (Eq. 4) had to be calculated differently. The ensemble model without weights from the second ensemble model approach is completely in the Appendix. Also the variance and quantiles of the two ensemble model from the second ensemble model approach are in the Appendix (Fig. 25-32).

Thirdly for the spatial projection of *M. galathea* in the future the best model from the complete variable combination set was taken for each of the four model types and combined using a weighted mean with their OOB-RMSE as weights. All variables were retained, except the climate variables mean annual temperature and annual precipitation total were replaced with values from the future climate scenarios.

These two ensemble model approaches were only used for one buffer zone, because an ensemble model with 12 variables resulted in 4095 models and an ensemble model with 14 variables led to 16'383 individual models. The calculations were computationally intense, hence just the buffer zone with the best performance in the complete model evaluation (section 3.6.4 and 4.3) was chosen to generate the ensemble models.

I used a two-sided paired t-test for the comparison of the ensemble models, because I have repeated prediction for each pixel. The t-test verifies, if the mean of two spatial projections differs significantly. For the first ensemble model approach the difference between the full and the best model were tested. For the second ensemble model approach the difference between the ensemble model with weighted mean and the ensemble model with normal mean was tested. Also for the first ensemble model approach the difference between the ensemble models with the climate scenarios and the ensemble model of the present were tested.

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i \cdot x_i}{\sum_{i=1}^{N} w_i} \tag{3}$$

$$s^2 = \frac{\sum_{i=1}^{N} w_i}{\left(\sum_{i=1}^{N} w_i\right)^2 - \sum_{i=1}^{N} w_i^2} \cdot \sum_{i=1}^{N} w_i \cdot (x_i - \bar{x})^2 \tag{4}$$

3.7 Projecting the current spatial abundance distribution of *M. galathea*

The spatial projections of the abundance of *M. galathea* in Switzerland and Liechtenstein were calculated in a 1x1 km raster for the present. This created 41'442 locations and all needed the explanatory variables for the best performing buffer zone. These variables were calculated analogously to the explanatory variables in the 474 observation locations. I used both ensemble model approaches to project the current spatial abundance distribution of *M. galathea* in Switzerland.
The explanatory variables for the spatial projections (41'442 locations) had some locations, which were not included in the DEM25, precipitation and temperature rasters. So the DEM25 did not cover six locations for the 50, 100, 150 and 200 m buffer zone and hence the mean of the surrounding locations of the 500 m buffer zone was taken as an approximation for these missing values. The 1000 m buffer zone was excluded from the analysis, because of long calculation time and huge data files, which could not be handled by ArcGIS. The temperature and precipitation files did not cover 140, 152, 155, 157 and 175 locations for the 500, 200, 150, 100 and 50 m buffer zones. These locations were located in Liechtenstein. For all buffer zones, the mean of the surroundings locations was taken as an estimation for all missing locations in Liechtenstein.

3.8 Projecting the future spatial abundance distribution of *M. galathea*

3.8.1 Climate scenarios

I assessed the influence of climate change on the abundance of *M. galathea* using the ensemble model 1BM (section 3.6.6), because it uses the best model from the complete variable combination set. WSL provided downscaled regional climate model outputs from the "Coordinated Regional Climate Downscaling experiment" (CORDEX, www.cordex.org), which was used for the fifth IPCC report (Pachauri et al., 2014). The fifth IPCC report used "general circulation models" (GCMs) also called "earth system model" (ESM) and model chains. The GCM was renamed "earth system model" (ESM), due to growing model complexity. GCMs/ESMs model the climate for the whole world in a coarse resolution (e.g.

2.5°). Many GCMs/ESMs are available, developed and used by many institutes around the globe. The "regional climate model" (RCM) models a section (e.g. Europe) with a higher resolution (e.g. 12x12 km) and requires output of a GCM/ESM as input. This model set-up is also called a model chain. Different institutes calculated different RCMs for CORDEX, e.g. the Swiss Federal Institute of Technology in Zurich. The project "forest cover changes in mountainous regions" (FORECOM; IGSM JU (2016)) used the output of five model chains of the CORDEX initiative. These were downscaled to 100x100 m for Switzerland and also used in this thesis. A more precise description of the used climate scenarios follows in the next paragraph.

First, the GCM/ESM from CORDEX were downscaled to a 12x12 km raster grid with RCM by the "Consortium for Small Scale Modelling - Climate Limited-area Model" (COSMO-CLM; CLM-C (2016)). Second the 12x12 km raster were further downscaled to a 100x100 m raster by Dirk Schmatz at WSL from the Landscape Dynamics research unit, resulting in monthly gridded data sets that cover Switzerland (CH1903-LV03 projection) for the years 2006 to 2100. They are calculated for the 4.5 and 8.5 "representative concentration pathways" (RCP). The RCPs represent the possible range of radiative forcing in the year 2100 relative to pre-industrial values. To estimate a range of possible effects of climate change, the "earth system model" (ESM-LR) from the "Max Planck Institute for Meteorology" (MPI-M) was chosen, because of its strong temperature and precipitation anomalies (MPI, 2016). So two climate change scenarios were used for assessing the influence of the climate change to the spatial abundance distribution of *M. galathea*. This two climate scenarios have the abbreviation RCP4.5 and RCP8.5. I created future spatial projections of *M. galathea* abundances for two time periods; 2021-2050 and 2071-2100. For the sake of simplicity the mean over the period from 2021-2050 and 2071-2100 for annual precipitation total and for the annual mean temperature was calculated.

The temperature and precipitation anomalies of the two climate scenarios are presented in Fig. 5 and in Fig. 6. The time line is represented as color and thickness changing line from 2010 until 2098, so the line gets thicker and changes the color from yellow to blue with increasing time. It is evident that the temperature and precipitation anomalies are higher for the RCP8.5 scenario compared to the RCP4.5 scenario. For the integration of the two climate variables from this two climate scenarios in the explanatory variable set, the rasters had to be interpolated from 100x100 m to a 1x1 km raster, because this is the raster pixel size for all predictions with the ensemble models. For the interpolation the nearest neighbour method was used.

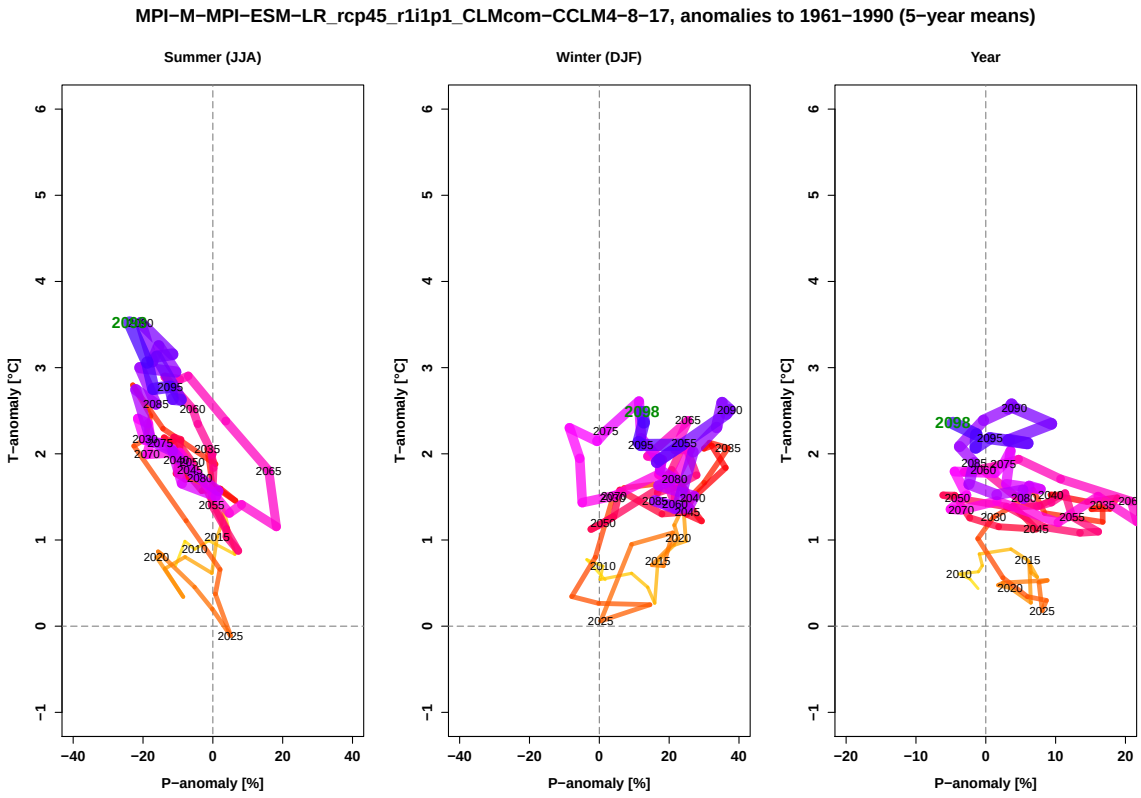


Figure 5: The temperature and precipitation anomalies for the RCP4.5 climate scenarios based on the period 1961 to 1990. The time line is represented as color and thickness changing line from 2010 until 2098, so the line gets thicker and changes the color from yellow to blue with increasing time. The description of the abbreviations in the title is in Table 2 (Schmatz, 2015a).

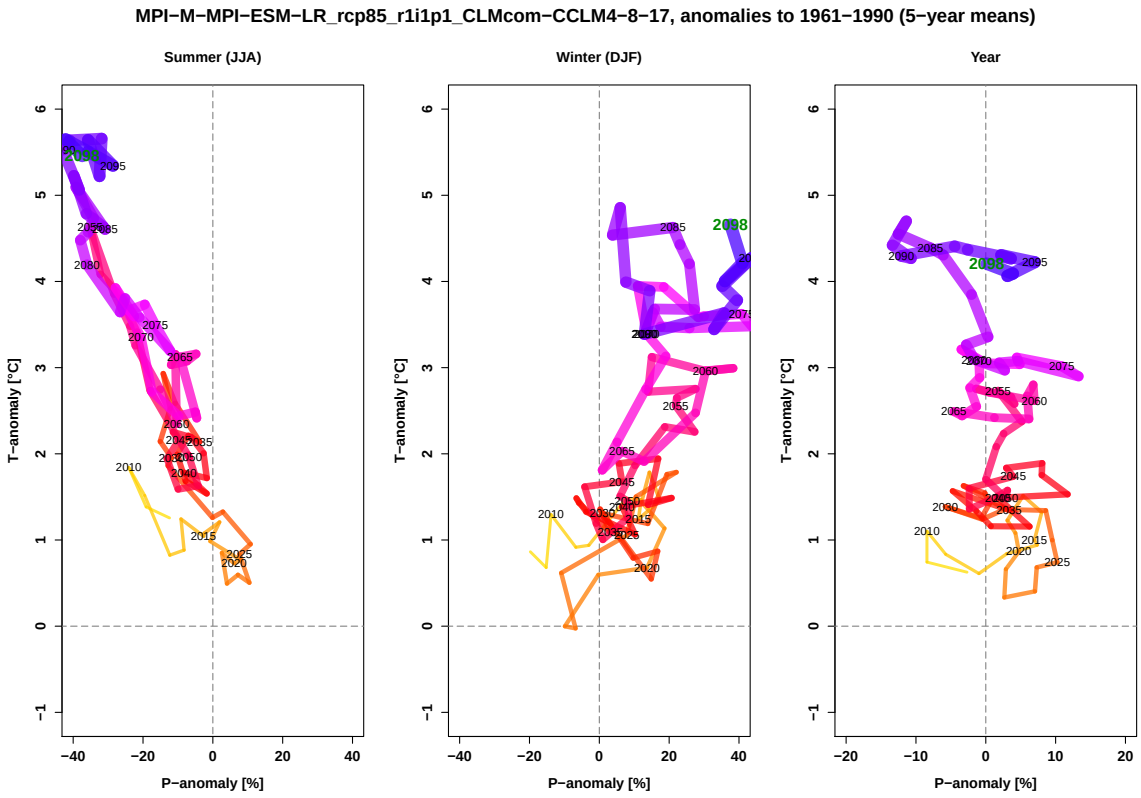


Figure 6: The temperature and precipitation anomalies for the RCP8.5 climate scenarios based on the period 1961 to 1990. The time line is represented as color and thickness changing line from 2010 until 2098, so the line gets thicker and changes the color from yellow to blue with increasing time. The description of the abbreviations in the title is in Table 2 (Schmatz, 2015b).

Table 2: Description of the abbreviation of the titles from the Figures 5 and 6.

Abbreviation	Meaning
MPI-M	The institute which had ran GCM/ESM (Max Planck Institute for Meteorology)
MPI-ESM-LR	GCM/ESM
rcp45/rcp85	The modelled representative concentration pathway (RCP4.5 or RCP8.5)
r1i1p1	Rip nomenclature of the GCM runs (realization 1, initialisation 1, physics 1)
CLMcom	The community which had run the RCM (http://www.clm-community.eu)
CCLM4-8-17	The used RCM: COSMO-CLM or CCLM version 4-8-17

3.8.2 Land-use change scenarios

The land-use scenarios relied on projections of expected land-use and land-cover changes for the year 2035 (Price et al., 2015). The scenarios were calculated for six land classes: closed forest, open forest, overgrown, agriculture pasture, agriculture arable and urban/sealed surface, all derived from Areal Statistics (SFSO (2013); Price et al. (2015)). Three scenarios were used to qualitatively estimate the influence of land-use change on the spatial abundance distribution of *M. galathea* (section 3.8.3). Scenario A1 is also referred to as the

globalisation scenario with a high weighting of the globalisation and less intervention in environmental issues. Concern for environmental issues and the support for conservation and agricultural subsidies are all low. B2 is the opposite of A1, as it focuses on regions and provides more sustainable intervention regarding environmental issues. A Trend scenario is based on past land-use transitions (1985, 1997, 2009; SFSO (2016)) whose trends were linearly extrapolated for the year 2035 (Price et al., 2015).

3.8.3 Assessing *M. galathea* abundance and spatial distributions in a changing environment

For the assessing of the abundance and spatial distribution of *M. galathea* in a changing environment, the ensemble models 1BM and 1CS (Fig. 4) were overlay with the three used land-use change scenarios. Here only the three land classes overgrown, agriculture pasture and agriculture arable were used for the assessment, because *M. galathea* exist especially in open areas. With the combined raster the areas of the three land-use classes were calculated with an abundance higher than 20 predicted butterflies. For these combinations, the rasters of the ensemble model predictions had to be downscaled from a 1x1 km to a 100x100 m pixel size raster. This was done with the nearest neighbour method, because the bilinear interpolation calculated negative values for the abundance. From the two climate scenarios only the first period from 2021 until 2050 was taken, because it has a time intersection with the land-use scenarios, which are predicted for 2035. I used three different combination of the ensemble models 1BM and 1CS with the land-use change scenarios (Trend, A1, B2). First the ensemble model 1BM was combined with the three land-use scenarios. Second the ensemble model 1CS for both climate scenarios was combined with the land-use in 2009, which is also the reference for the land-use scenarios. Third the ensemble model 1CS for both climate scenarios were combined with all three land-use scenarios. For these three combinations the predicted abundance in each pixel for the five land classes were calculated and boxplots were created (Appendix Fig. 33-36). But around 9% of the abundances are outlier, dependent on the land class it can be a few ten thousands outliers, hence they were not included in these boxplots.

4 Results

4.1 Variable selection

Strongly correlated variables (Spearman correlation greater than 0.7) were eliminated (section 3.6.1). The Spearman correlation for the explanatory variables in each buffer zone (50,100, 150, 200, 500, 1000m) at the butterfly observation locations revealed three different variable sets; see Table 1 and Figures 7, 17-21 in the Appendix for detail on what variables were retained for which buffer zone. The 50, 100 and the 150 m buffer zones have the same variable set with 14 variables each, with single tree, forest, open land, SW (aspect) and elevation as excluded variables (Appendix Figures 17-19). The 200 and 500 m buffer zones have the same variable set with 12 variables each, with line of tree and hedgerow, single tree, forest, road, open land, SW (aspect) and elevation as excluded variables (Fig. 7; Appendix Fig. 20). The third variable set with 10 variables is for the 1000 m buffer zone, with line of tree and hedgerow, forest, building, dry meadow, road, open land, SW (aspect), slope and elevation as excluded variables (Fig. 21).
The Spearman correlation matrix is shown in Fig. 7 for the 500 m buffer zone as a

representative example, the correlation matrices of the remaining buffer zones are shown in Appendix Figures 17-21.

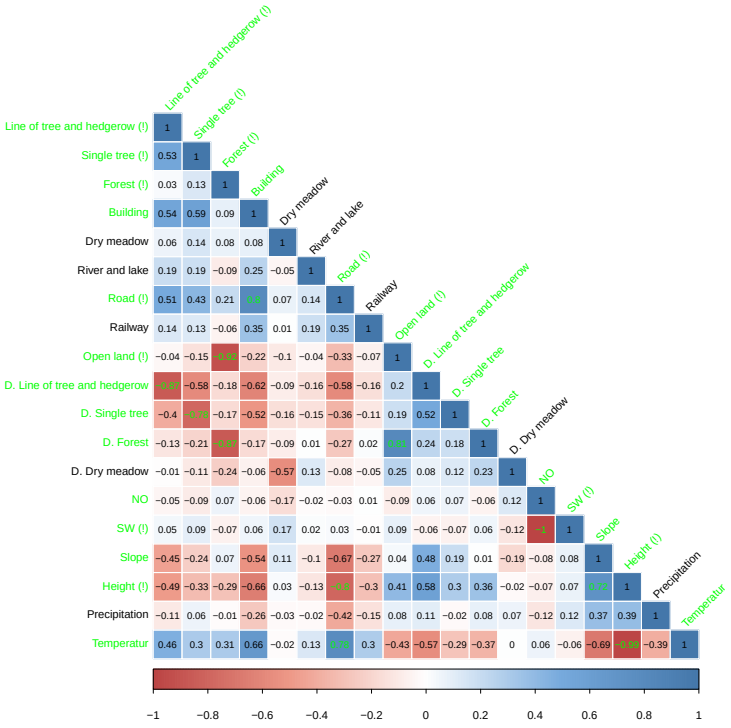
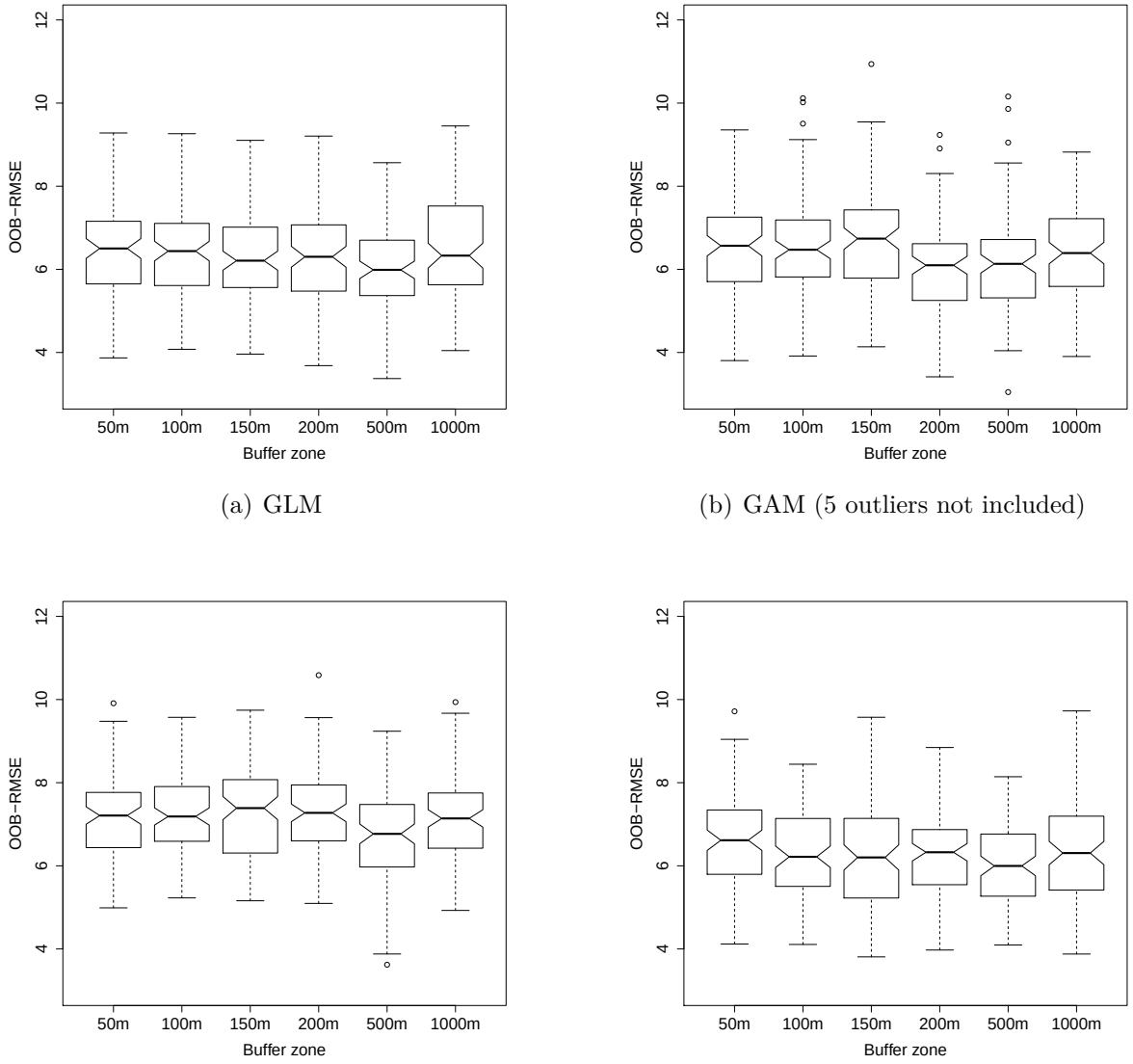


Figure 7: The correlation between all variables for the 500 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

4.2 Individual models

The individual models identified the buffer zone with the lowest OOB-RMSE distribution of the MCCV across all models (Fig. 8). If the notches in the boxplots do not overlap, there is strong evidence that the two medians significantly differ (Chambers, 1983). However, Fig. 8 shows that most boxplots exhibit considerable overlap between notches, indicating that overall, the buffer zones have comparable accuracy. Buffer zone 500 m, however, was consistently lower across all model types, except for GAM. Therefore, I conducted all further analyses with the 500 m buffer zone as a representative. The boxplots also reveal model specific properties of the models. For example, GAM has huge outliers, as the model was fit with a smoothing function that allows for a maximum of four degrees of freedom. Hence the outliers can be very high if the number of observations in the calibration data is reduced (here to 80% of the full data set). Removing the testing data in the MCCV, can drastically change the smoothing functions. OOB-RMSE for GAM has five outliers, which are not included in Fig. 8(b). Four outlier are from the 50 m buffer zone (21'287, 310'624, 380'671, 8'930'159'172) and one from the 1000 m buffer zone (1'310'651). Such performance is typcial for GAM. The plots of the component smooth functions for GAM are in the Appendix, one is from the full model Fig. 22, the other is from the best model of the complete variable combination set Fig. 23. They show for example that butterfly abundance is positively affected by slope steepness across the entire range of zero to 50°.



(a) GLM

(b) GAM (5 outliers not included)

(c) RPART

(d) RF

Figure 8: Boxplots of the MCCV for the four model types and all buffer zones. The MCCV had 100 repetitions.

Results for the four model types using the 500 m buffer zone data based on 12 variables are shown in Table 3. For all four model types slope, building, temperature and precipitation are among the most important variables explaining the abundance of *M. galathea*. The distance variables have negative parameter estimation, so big distances are worse for *M. galathea* than low distances. So according to GLM the abundance of *M. galathea* increases in locations where landscape elements such as line of tree and hedgerow, single tree, dry meadow and forest are close.

Table 3: The variables of the full models, the "∗" are the significance levels, on which the variables are still significant. "∗ ∗ ∗" is a 0.001 significance level, "∗∗" is a 0.01 significance level and "∗" is a 0.05 significance level. They are marked blue, if the parameter estimation is positive and red if it is negative and they are marked black, if they have a smoothing function. For RPART and RF the variables were sorted dependent of their importance. Variables with the "-" sign are not important or significant.

Variable	GLM	GAM	RPART	RF
Building	∗	∗ ∗ ∗	2	2
Dry meadow (factor)	–	∗	11	8
River and lake	∗ ∗ ∗	∗ ∗ ∗	10	10
Railway (factor)	–	–	–	12
Distance line of tree and hedgerow	∗ ∗ ∗	∗ ∗ ∗	7	5
Distance single tree	∗ ∗ ∗	∗ ∗ ∗	9	7
Distance forest	∗ ∗ ∗	–	8	9
Distance dry meadow	∗ ∗ ∗	∗ ∗ ∗	6	6
NO	∗	∗ ∗ ∗	5	11
Slope	∗ ∗ ∗	∗ ∗ ∗	1	1
Precipitation	∗ ∗ ∗	∗ ∗ ∗	4	4
Temperature	∗ ∗ ∗	∗ ∗ ∗	3	3

Table 4 shows the variables of the best-performing models from the complete variable combination set. GLM and GAM have the same variable combination in the best model, but river and lake was not statistically significant in the GAM. RPART needs only slope and temperature for the best model, but another model that additionally includes railway has the same OOB-RMSE. However, in this model railway is much less important than slope and temperature. Because a model should be parsimonious in terms of the number of included variables, the model with two variables slope and temperature was chosen for the representation of the best model of RPART (Table 4).

Table 4: The variables of the best models from the complete variable combination set, the "∗" are the significance levels, on which the variables are still significant. "∗ ∗ ∗" is a 0.001 significance level, "∗∗" is a 0.01 significance level and "∗" is a 0.05 significance level. They are marked blue, if the parameter estimation is positive and red if it is negative and they are marked black, if they have a smoothing function. For RPART and RF the variables were sorted dependent of their importance. Variables with the "-" sign are not important or significant. Note that GLM/GAM, RPART and RF have different variable sets.

Variable	GLM	GAM	RPART	RF
River and lake	∗ ∗ ∗	–	Slope (1)	Slope (1)
Distance line of tree and hedgerow	∗ ∗ ∗	∗ ∗ ∗	Temperature (2)	Building (2)
Distance single tree	∗ ∗ ∗	∗ ∗ ∗		Temperature (3)
Distance forest	∗ ∗ ∗	∗		Precipitation (4)
Slope	∗ ∗ ∗	∗ ∗ ∗		Distance line of tree and hedgerow (5)
Precipitation	∗ ∗ ∗	∗ ∗ ∗		River and lake (6)
Temperature	∗ ∗ ∗	∗ ∗ ∗		Railway (7)

Both the full model and the best model from the complete variable combination set include temperature, slope and precipitation as significant/important variables. The

two-sided, paired t-tests for the OOB-RMSE from the full models, the worst and best models from the complete variable combination set, and the ensemble model 2W (Fig. 4) are in Table 6 and the OOB-RMSE are in Table 5. Note that GAM has the lowest OOB-RMSE in the best model and the largest OOB-RMSE in the worst model. Also not that all variables in the worst GLM and GAM models are significant at a 0.01 significance level. And the t-tests show no significant difference of the OOB-RMSE (Table 6), except between the full model and the best model from the complete variable combination set. Also note that the degrees of freedom is only 3, which is not optimal for a t-test.

Table 5: The OOB-RMSE from the full models, the worst and best models from the complete variable combination set, and the ensemble model 2W (Fig. 4).

Model	GLM	GAM	RPART	RF
Full model	6.09	6.22	6.85	6.11
Best model	5.96	5.72	6.51	5.77
Worst model	7.05	93'623	7.75	8.28
Ensemble model 2W	6.62	130	7.14	6.49

Table 6: The two-sided, paired t-tests for the OOB-RMSEs of the full models, the worst and best models from the complete variable combination set, and the ensemble model 2W (Fig. 4). The OOB-RMSE are in Table 5. The subscript i in t_i are the degrees of freedom.

	Full models	Best models	Worst models
Best models	$t_3 = -4.3152$ $p = 0.02292$		
Worst models	$t_3 = -1.0001$ $p = 0.391$	$t_3 = -1.0001$ $p = 0.391$	
Ensemble models 2W	$t_3 = 1.0131$ $p = 0.3856$	$t_3 = 1.0218$ $p = 0.3821$	$t_3 = -1$ $p = 0.391$

4.3 Projecting the current spatial abundance distribution of *M. galathea*

The spatial projection for the present was done with the first and second ensemble model approach, resulting in seven different spatial projections; one ensemble across the four models using all variable, one ensemble using only the best-performing variable-combination for each model type, four ensemble across all variable combinations (separately for each model type), and one final ensemble combining the latter four ensembles.

First ensemble model approach with the full models 1FM and the best models from the complete variable combination set 1BM

The spatial projection of the ensemble model 1FM with their OOB-RMSE as weights is displayed in Fig. 9, while the spatial projection for the ensemble model 1BM is visible in Fig. 10. Both ensemble models identify the same areas as abundance hotspots for *M. galathea*. The species has its greatest abundance primarily in Valais, Ticino, Graubünden, the Jura Mountains and the northern edge of the Swiss Alps. Both differ significantly according to the t-test ($t_{41'441} = -1$, $p = 8.55e - 10$). Figure 11 represents the difference between the ensemble model 1BM and the ensemble model 1FM. The negative values

(red) show that the ensemble model 1FM has mostly higher abundance prediction in the hotspots zones in Valais, Ticino and Graubünden and lower abundance predictions in the Jura Mountains and the northern edge of the Swiss Alps. These differences come from the different variable sets used for the ensemble models.

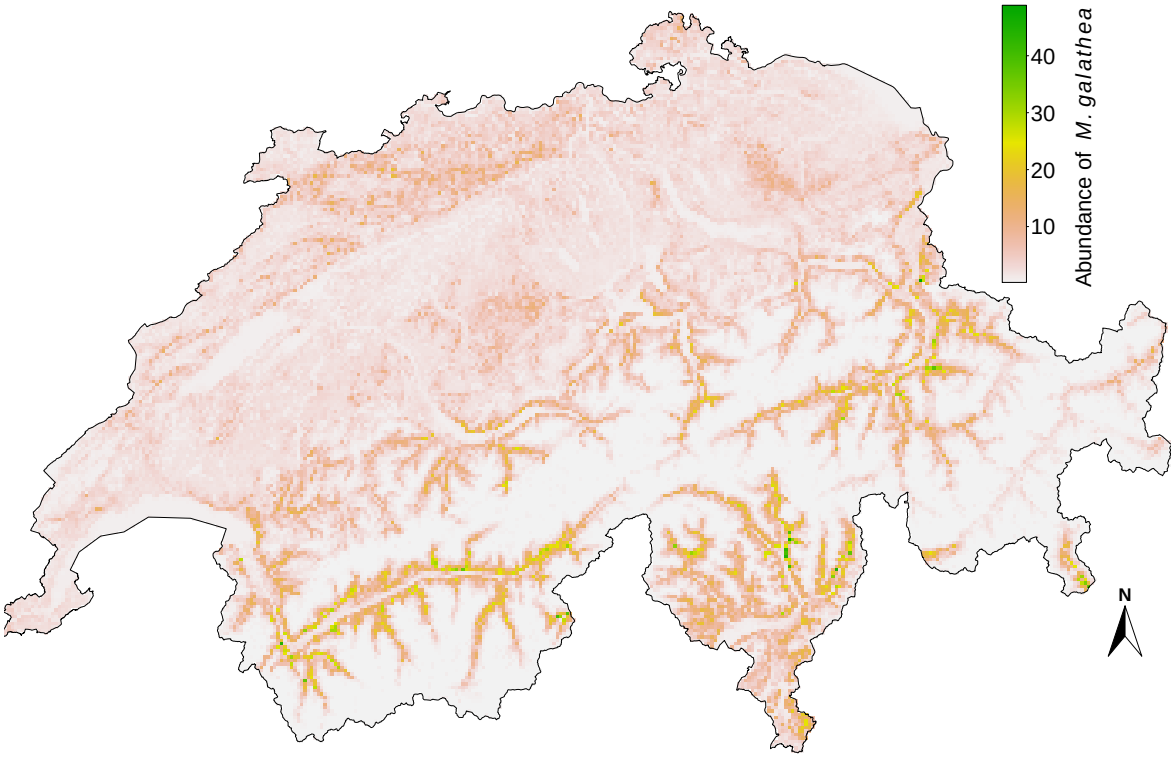


Figure 9: The spatial projection of the weighted mean of all four model types including all variables.

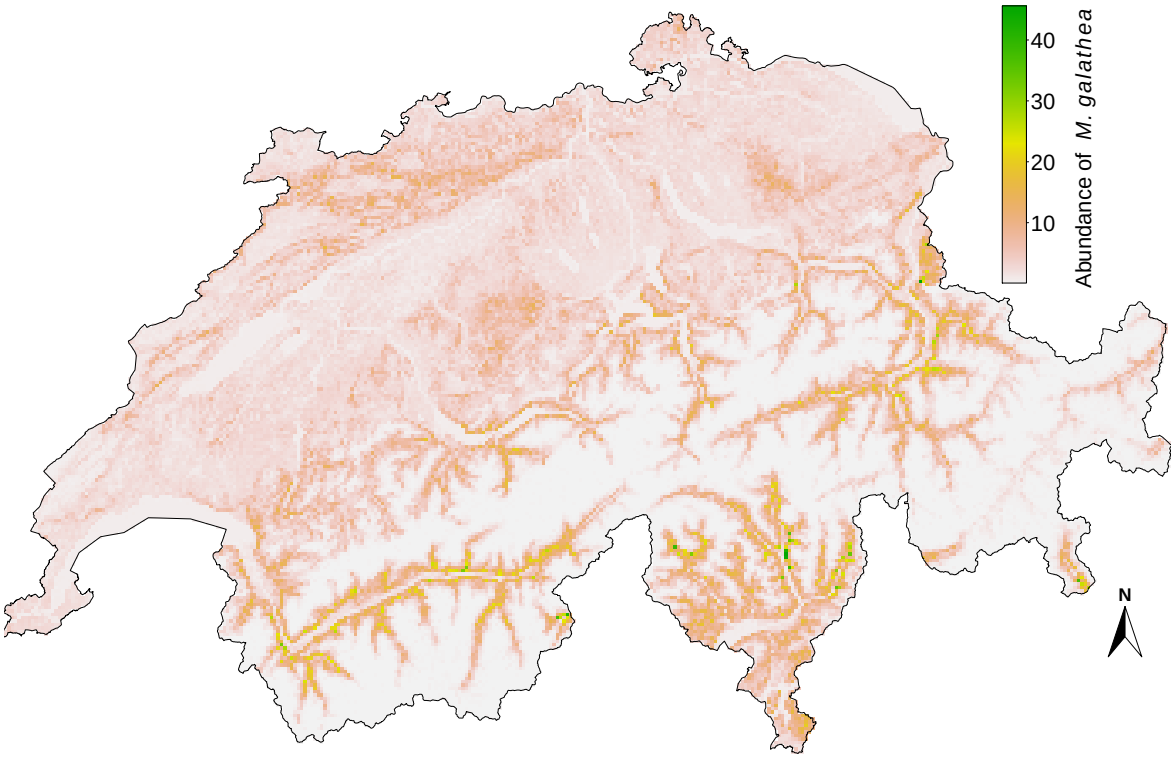


Figure 10: The spatial projection of the weighted mean of all four model types with the best model from the complete variable combination set.

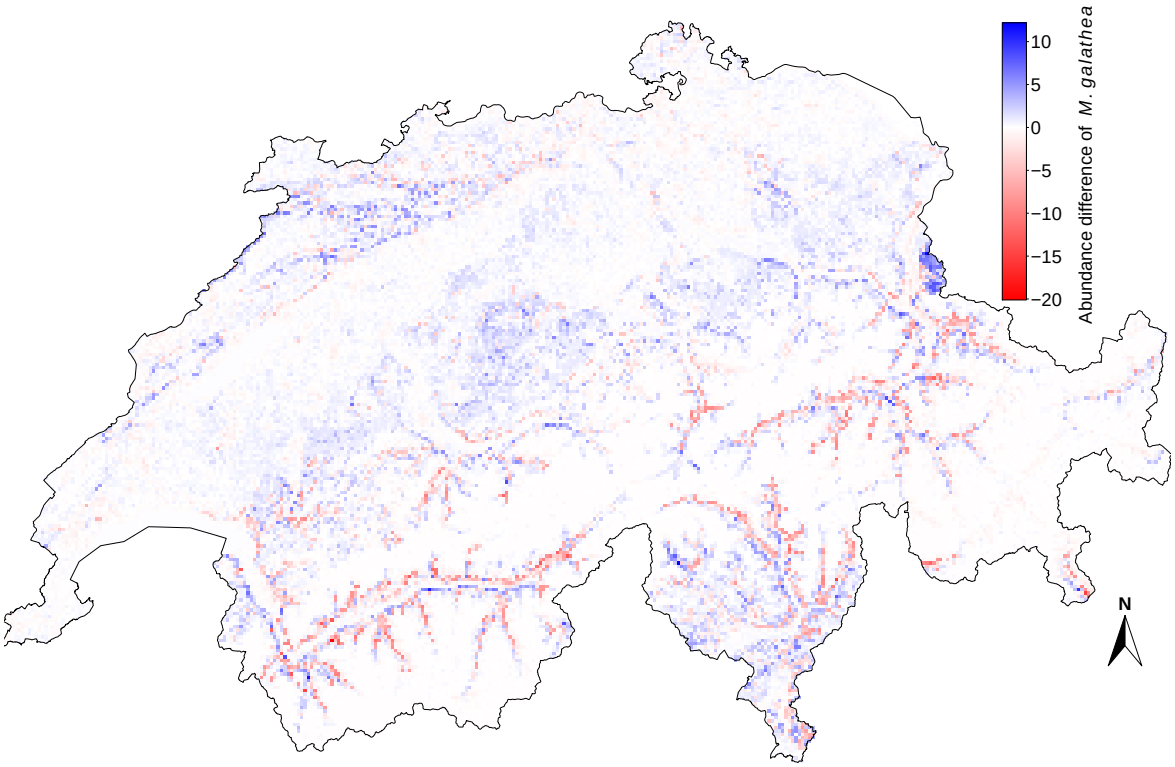
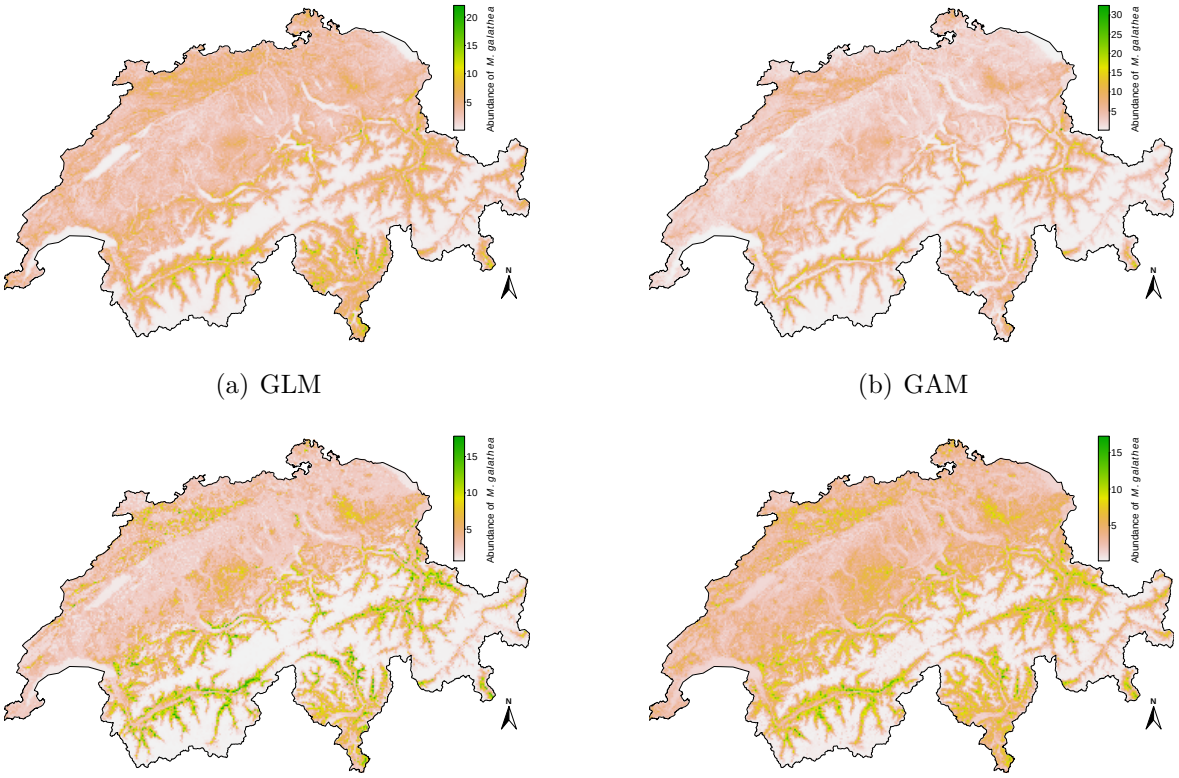


Figure 11: The difference in the spatial projection of the ensemble model with the best model from the complete variable combination set and the ensemble model with the full models. Positive (negative) values are given in blue (red) and indicate where the ensemble of the best models predicts more (less) *M. galathea* individuals.

Second ensemble model approach with the complete variable combination set for each of the four model types (2W; 2NW)

For all four model types an ensemble model was created, each with all models from the complete variable combination set. Fig. 12 shows the weighted average with the OOB-RMSE as weights. The variance of these ensemble models is displayed in Appendix Fig. 25. These four ensemble models were further combined using a weighted mean; each ensemble prediction was weighted by the average OOB-RMSE across the complete variable combination set (Fig. 13). The same ensemble projections were also generated without weighting, these figures are in the Appendix (Fig. 26; 27; 28). Differences between the ensemble model with the weighted mean and the ensemble model without weighting range from 4 to 1 abundances and are shown in Fig. 14. The two-sided, paired t-test is significant ($t_{41'441} = -17.872$, $p - value = 2.2 * 10^{-16}$) and the weighted ensemble projections are rather lower than projections from the ensemble model without weights. This is expected to happen, because the weights minimize the influence the models with over- and underestimation of *M. galathea* abundances. The quantiles from the single ensemble models are for the two ensemble models the same (Fig. 29-32). The second ensemble model approach has a similar abundance distribution as the first ensemble model approach, with the difference that the maxima are around 15 butterflies and not 40. Also, here the influence of the 4095 models from the complete variable combination set is

recognizable. These models have a huge variability with huge over- and underestimations, but the normal or weighted mean gives a lower maxima, than the ensemble models with 4 models. The four ensemble models from the two approaches have the same abundance distribution pattern. The valleys from Graubünden, Ticino, Valais and the northern edge of the Swiss Alps have higher abundance. And in the Swiss Plateau are low abundance, the hotspot are especially in Graubünden, Ticino and Valais. The four different model types have also nearly the same abundance distribution pattern, but the abundance range is for RF smaller than for GLM, GAM, RPART (max 15 vs 40). So RF fits 500 trees and the mean probably decreases the abundance range. RPART is not so smooth than GLM, GAM and RF, because the abundance prediction depends only on one trees, this can cause abrupt abundance changes for adjacent pixels. Also RF and RPART have more hotspots than GLM and GAM, which is probably also a result of the binary tree decision of the abundance prediction. GLM has rather higher abundance than GAM, because it fits linear functions. This does not have to be case, but here it is.



(a) GLM (b) GAM

(c) RPART (d) RF

Figure 12: The spatial projections with four ensemble models for the four model types, each has the complete variable combination set with the OOB-RMSE as weights.

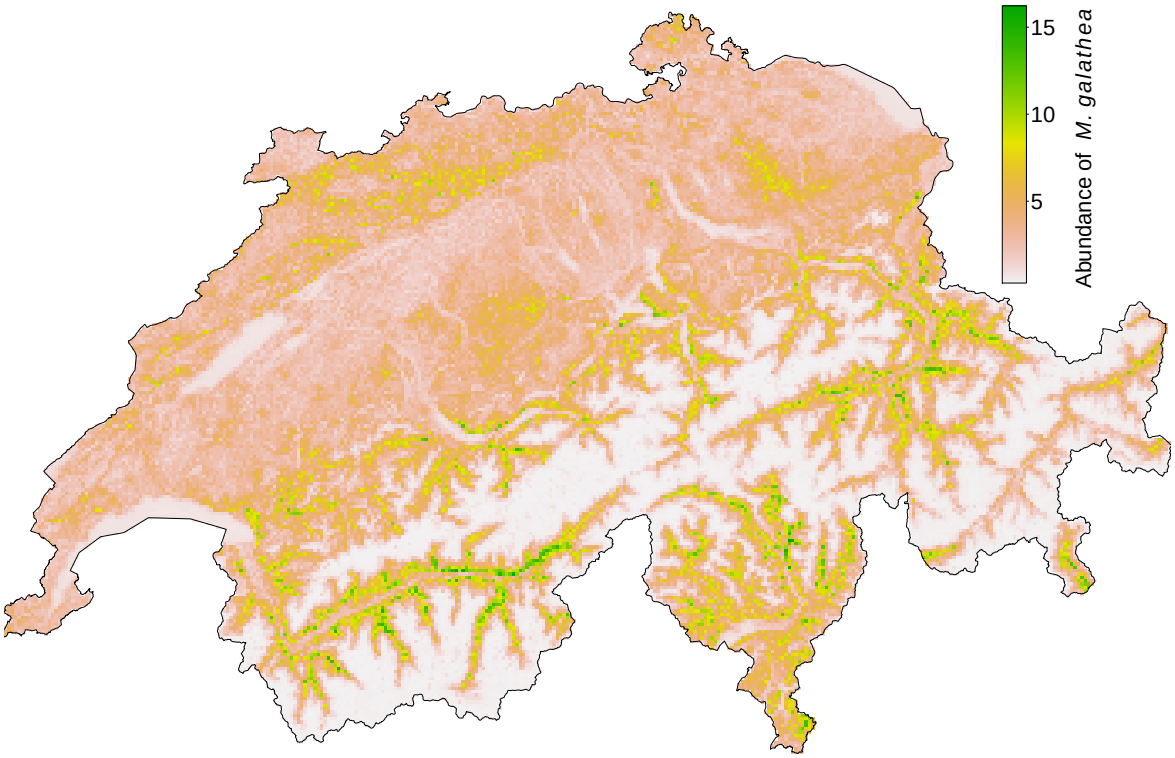


Figure 13: The mean of the four ensemble models from Figure 12, each with the complete variable combination set and their average OOB-RMSE as weights.

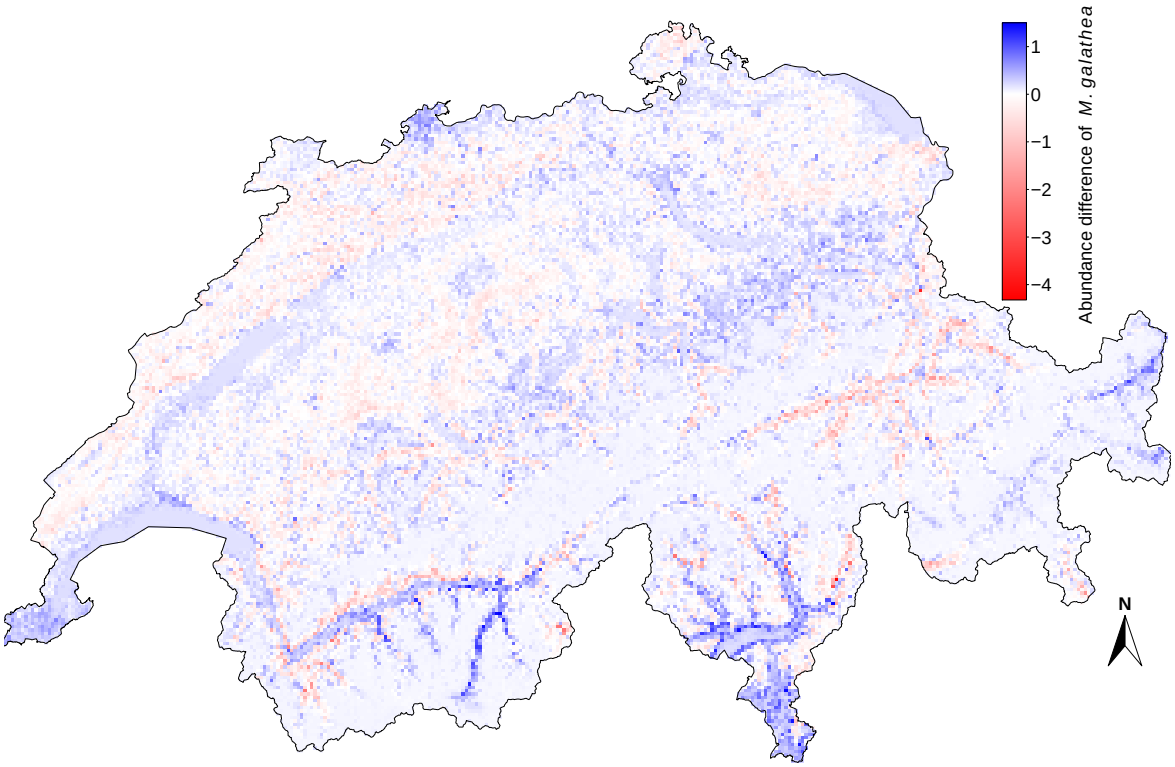


Figure 14: Differences between the ensemble models with the complete variable combination set: model with the weighted mean (2W) versus the model with the normal mean (2NW). Positive (negative) values in the blue (red) indicate where the ensemble of the best models predicts more (less) *M. galathea* individuals.

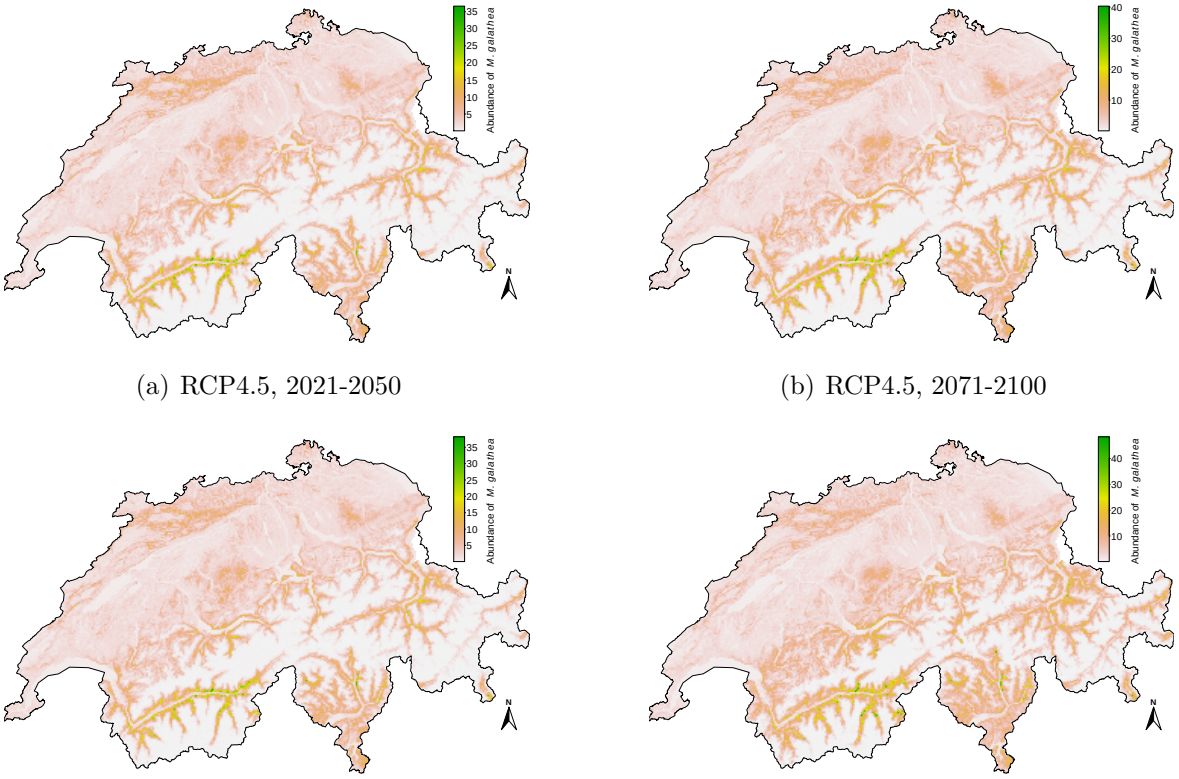
4.4 Assessing *M. galathea* abundance and spatial distributions in a changing environment

First ensemble model approach with the best models from the complete variable combination set combined with the climate scenarios

The spatial projections of the abundance of *M. galathea* with the climate scenarios show a positive and negative influence on the spatial abundance distribution of *M. galathea* (Fig. 15 and Fig. 16). Also all t-tests showed significant difference between these two climate scenarios and the present (Table 7). The RCP8.5 scenario in the period from 2071 to 2100 has the most positive effect on *M. galathea* abundance, this scenario has a high radiative forcing and the temperature/precipitation anomalies are also highest in this period. In contrast, the RCP4.5 climate scenario has the most negative effect on *M. galathea* abundance. Overall the positive effects are especially well visible in Valais, the northern edge of the Swiss Alps and the Jura Mountains, while the negative effects are most prominent in the Swiss Plateau, Ticino and Graubünden.

The differences between the future predictions of the climate scenarios and the two periods with the present prediction show, that RCP8.5 has rather higher abundances than RCP4.5 and the second period (2071-2100) has rather higher abundances than the first period (2021-2050) (Fig. 16). These difference plots were used to analyse the range shifts, e.g. range expansion and range contraction. A range expansion is recognizable to higher

elevations, mostly at the valley sides. A range contraction occur in the flatlands, especially in the Swiss Plateau. In the first period from the climate scenarios is a rather decrease of the abundance in lower elevation areas visible and in the second period is a rather increase in higher elevation areas visible. Also Valais is always a good habitat, because there are nearly no range shifts recognizable and the abundances increase. The range expansion is at its highest with the climate scenario RCP8.5 in the second period.



(a) RCP4.5, 2021-2050

(b) RCP4.5, 2071-2100

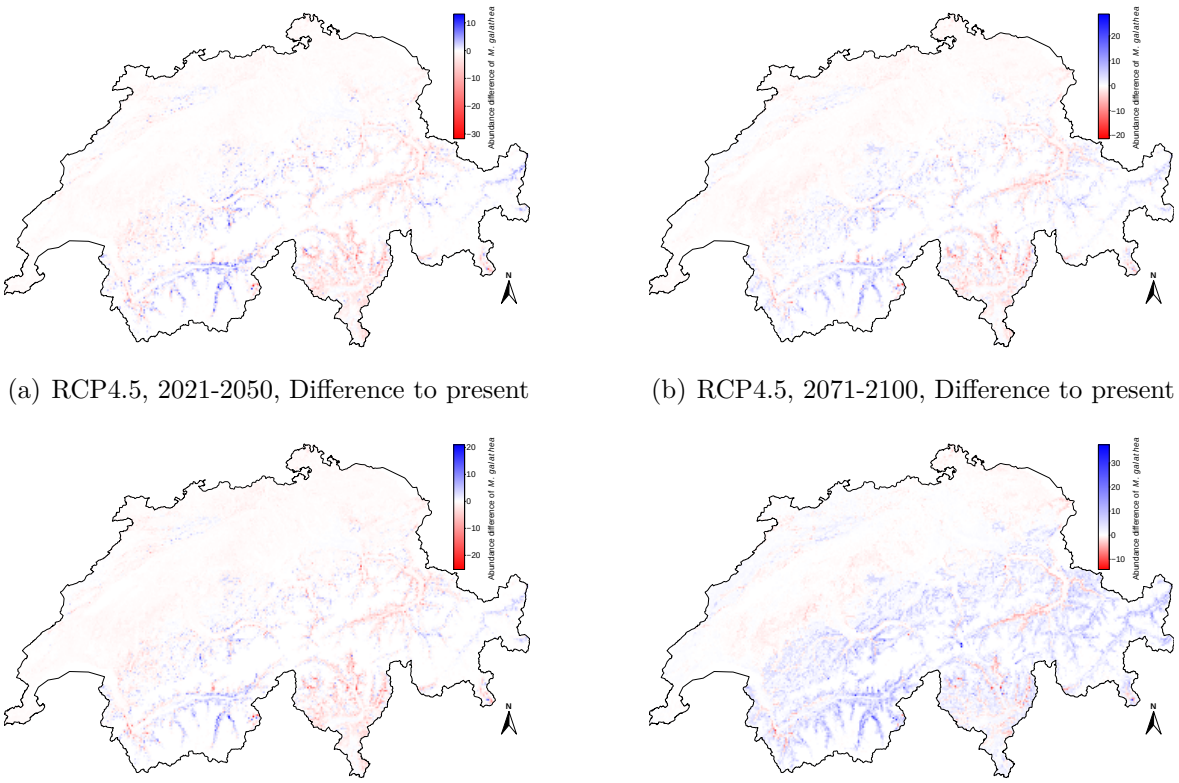
(c) RCP8.5, 2021-2050

(d) RCP8.5, 2071-2100

Figure 15: Spatial projections of M. galathea abundance using two climate scenarios (RCP4.5 and RCP 8.5) and two time periods. The ensemble model 1BM was combined with the temperature and precipitation of the climate scenarios to make theses spatial projections.

Table 7: The two sided, paired t-tests between the spatial projections for the present and future under climate change. The degrees of freedom are always 41'307 and the p values are always $2.2 * 10^{-16}$.

	Present	RCP4.5; 21-50	RCP4.5; 71-00	RCP8.5; 21-50
RCP4.5; 21-50	$t = -58.909$			
RCP4.5; 71-00	$t = 10.261$	$t = -73.076$		
RCP8.5; 21-50	$t = -39.829$	$t = -27.531$	$t = 57.045$	
RCP8.5; 71-00	$t = 87.122$	$t = -120.01$	$t = -103.29$	$t = -115.85$



(a) RCP4.5, 2021-2050, Difference to present

(b) RCP4.5, 2071-2100, Difference to present

(c) RCP8.5, 2021-2050, Difference to present

(d) RCP8.5, 2071-2100, Difference to present

Figure 16: Differences between the climate scenarios and the present. Positive (negative) values in the blue (red) indicate where the ensemble of the best models predicts more (less) *M. galathea* individuals.

Land-use change scenarios

Here the land-use scenarios were combined with the abundance distributions from the present and future under climate change. In Table 8 are the areas of the five land classes visible for the land-use in 2009, which is used as reference, and the three land-use scenarios Trend, A1, B2. The three land-use scenarios seems to have no different influence on the predicted abundances for the present (Table 9), overgrown and agriculture arable have an decrease and agriculture pasture has an increase. The predicted abundances for the future combined with the land-use reference in 2009 reveals higher gains for RCP8.5 (Table 10), it has a higher increase in all three land classes, RCP4.5 has in overgrown a decrease. If the climate scenarios and the land-use scenarios were combined together is also no big difference between the three land-use scenarios identifiable. In all three land-use scenarios have RCP4.5 and RCP8.5 similar decrease in overgrown and agriculture arable, but the increase in agriculture pasture is for RCP8.5 two times greater than for RCP4.5 (Table 11-13).

Table 8: The areas of the five land classes from the land-use change scenarios Trend, A1, B2 and the land-use in 2009 as reference. The percentage of a land class refers to the change compared with the reference land-use in 2009.

Land class	Reference (2009) [km^2]	Trend [km^2]	A1 [km^2]	B2 [km^2]
Closed forest	11'524	11'681 (1%)	11'748 (+2%)	12'059 (+5%)
Open forest	1'858	1'849 (-1%)	1'871 (+1%)	1'346 (-28%)
Overgrown	501	450 (-10%)	405 (-19%)	116 (-77%)
Agriculture pasture	9'959	9'687 (-3%)	9'648 (-3%)	9'924 (0%)
Agriculture arable	4'578	4'322 (-6%)	4'576 (0%)	4'580 (0%)
Sealed surfaces	3'075	3'507 (+14%)	3'246 (+6%)	3'470 (+13%)

Table 9: The overlapping areas of the land-use scenarios with the ensemble model 1BM for abundances of *M. galathea* higher than 20. The percentages are the changes of the land classes respective to the land-use in 2009 with the ensemble model 1BM.

Land class	Reference (2009) [km^2]	Trend [km^2]	A1 [km^2]	B2 [km^2]
Overgrown	0.71	0.13	0.11	0.11
Agriculture pasture	17.34	19.59	19.72	19.61
Agriculture arable	2.06	0.19	0.19	0.21

	Change Trend [km^2]	Change A1 [km^2]	Change B2 [km^2]
Overgrown	-0.58 (-81.69%)	-0.6 (-84.51%)	-0.6 (-84.51%)
Agriculture pasture	2.25 (12.98%)	2.38 (13.73%)	2.27 (13.09%)
Agriculture arable	-1.87 (-90.78%)	-1.87 (-90.78%)	-1.85 (-89.81%)

Table 10: The overlapping areas of the land-use 2009 with the ensemble model 1CS for abundances of *M. galathea* higher than 20. The percentages are the changes of the land classes respective to the land-use in 2009 with the ensemble model 1BM.

Land class	Reference (2009) [km^2]	RCP4.5 [km^2]	RCP8.5 [km^2]
Overgrown	0.71	0.6	0.84
Agriculture pasture	17.34	18.13	20.38
Agriculture arable	2.06	2.08	2.35

	Change RCP4.5 [km^2]	RCP8.5 [km^2]
Overgrown	-0.11 (-15.49%)	0.13 (18.31%)
Agriculture pasture	0.79 (4.56%)	3.04 (17.53%)
Agriculture arable	0.02 (0.97%)	0.29 (14.08%)

Table 11: The overlapping areas of the land-use scenario Trend with the ensemble model 1CS for abundances of *M. galathea* higher than 20. The percentages are the changes of the land classes respective to the land-use in 2009 with the ensemble model 1BM.

Land class	Reference (2009) [km^2]	Trend RCP4.5 [km^2]	Trend RCP8.5 [km^2]
Overgrown	0.71	0.21	0.21
Agriculture pasture	17.34	20.18	22.92
Agriculture arable	2.06	0.32	0.33
		Change Trend RCP4.5 [km^2]	Change Trend RCP8.5 [km^2]
Overgrown		-0.5 (-70.42%)	-0.5 (-70.42%)
Agriculture pasture		2.84 (16.38%)	5.58 (32.18%)
Agriculture arable		-1.74 (-84.47%)	-1.73 (-83.98%)

Table 12: The overlapping areas of the land-use scenario A1 with the ensemble model 1CS for abundances of *M. galathea* higher than 20. The percentages are the changes of the land classes respective to the land-use in 2009 with the ensemble model 1BM.

Land class	Reference (2009) [km^2]	A1 RCP4.5 [km^2]	A1 RCP8.5 [km^2]
Overgrown	0.71	0.18	0.19
Agriculture pasture	17.34	20.24	22.97
Agriculture arable	2.06	0.32	0.33
		Change A1 RCP4.5 [km^2]	Change A1 RCP8.5 [km^2]
Overgrown		-0.53 (-74.65%)	-0.52 (-73.24%)
Agriculture pasture		2.9 (16.72%)	5.63 (32.47%)
Agriculture arable		-1.74 (-84.47%)	-1.73 (-83.98%)

Table 13: The overlapping areas of the land-use scenario B2 with the ensemble model 1CS for abundances of *M. galathea* higher than 20. The percentages are the changes of the land classes respective to the land-use in 2009 with the ensemble model 1BM.

Land class	Reference (2009) [km^2]	B2 RCP4.5 [km^2]	B2 RCP8.5 [km^2]
Overgrown	0.71	0.18	0.19
Agriculture pasture	17.34	20.17	22.9
Agriculture arable	2.06	0.34	0.35
		Change B2 RCP4.5 [km^2]	Change B2 RCP8.5 [km^2]
Overgrown		-0.53 (-74.65%)	-0.52 (-73.24%)
Agriculture pasture		2.83 (16.32%)	5.56 (32.06%)
Agriculture arable		-1.72 (-83.5%)	-1.71 (-83.01%)

5 Discussion

This thesis explains the spatial distribution of the abundance of *M. galathea* in Switzerland and Liechtenstein using ensemble modelling. I generated 19 explanatory variables within various sets of buffers (50-1000 m) at 474 butterfly observation locations across Switzerland and used the abundance observations at these 474 locations to fit four statistical models (GLM, GAM, RPART, RF). To account for multicollinearity, only 12 largely uncorrelated variables entered the models for the 500 m buffer zone. In a 1x1 km raster grid over Switzerland and Liechtenstein the explanatory variables were calculated and used for the spatial predictions with the ensemble models. For the ensemble modelling, I used just the 500 m buffer zone, because it had the best performance in all four model types. The ensemble model relied on four statistical model types (GLM, GAM, RPART, RF) and two ensemble model approaches were used to compare their performance and use the better one for future predictions with climate change scenarios. The prediction for the present and future were also used to analyse the land-use change scenarios.

The fact that climate (temperature and precipitation) together with slope are the main drivers for the distribution of the abundance of *M. galathea* lead to the conclusion that *M. galathea* has a suitable environment along the valley sides in the Swiss Alps and the Jura Mountains, as long as the share of buildings is not too high (according to GAM not over 25%; Fig. 23). The best models of the complete variable combination set used additional variables than temperature and slope, but they are the only retained variables across each of the four model types when selecting for the best-performing variable combination.

The ensemble models for the present show that the valley sides of the Swiss Alps and the Jura mountain as suitable habitats for *M. galathea*.

The climate scenarios seems to show a contradictory on *M. galathea*, a low increase of the temperature causes tendentially a lower increase of the abundance of *M. galathea* than a higher increase of the temperature. So the climate scenario RCP8.5 has tendentially higher abundance prediction than RCP4.5 and the first period from 2021 to 2050 has tendentially lower abundance prediction than the second period form 2071 to 2100. The annual precipitation total has an optimum around 800 mm for *M. galathea* (Fig. 23; log of precipitation). So *M. galathea* seems to prefer arid areas and with decreasing precipitation in the climate scenarios are such areas suitable as habitats for *M. galathea*. But also the valleys and the valley sides at distinct elevation get too arid and hence *M. galathea* shows a range shift to higher elevations. The annual mean temperature has an optimum around 8 °C for *M. galathea* (Fig. 23) and higher temperature seems to be okay for the abundance of *M. galathea* but not lower temperature. Also here is the preference for valley sides of *M. galathea* derivable and also visible in all figures of the ensemble models.

Table 8 depicts potential changes in the areas of certain landscape elements as derived form the three land-use change scenarios Trend, A1, B2 and the land-use in 2009 as reference. First of all, closed forest, agriculture arable and agriculture pasture have small changes. Open forest has especially in B2 a decrease of approximately one third. The variable forest of my variable set includes open and closed forest and is negatively correlated with the variable distance to forest (Fig. 7). The best models of GAM and GLM have distance to forest as significant variable. While GLM has a negative estimation of the coefficient and hence close-by forests increase the abundance of *M. galathea* (Fig. 4), GAM has a smoothing function with a positive gradient (Fig. 23), so close-by forest decrease the abundance of *M. galathea*. Furthermore, the distance to forests is not included in the best RPART and RF models. This contradiction in the models makes it hard to estimate

the influence of the strong increase of open forest in B2. Overgrown areas potentially strongly decrease in the B2 (80%) and the A1 (20%) scenarios. This variable is probably a part of my variable open land, which is positively correlated with distance to forests and negatively correlated with forest cover (Fig. 7). The negative correlation of open land with the forests could again suggest that the changes in overgrown area are negligible, because just distance to forest is in the best models from the complete variable combination set and just for GLM and GAM. Urban/sealed surfaces is comparable with the variable building of my variable set. The component smooth function of the best GAM (Fig. 22) along the building variable shows a decrease of the abundance starting form ca. 25%, but only the best RF includes building as an important variable. The GAM smooth functions suggest that, if the percentage of buildings or urban/sealed surface is too high, the abundance of *M. galathea* decreases. In all land-use change scenarios the cities and urban agglomeration will increase especially in the Swiss Plateau, hence a decrease of *M. galathea* abundance in these areas is possible. Given that the Swiss Plateau currently low abundance of *M. galathea*, hence a permanent colonisation of the Swiss Plateau is less likely in the future. Also has agriculture pasture always an increase of the abundances and overgrown and agriculture arable mostly a decrease of the abundances. For the land-use scenarios combined with the abundance distribution of the present and future under climate change the focus was on the three land classes overgrown, agriculture pasture and agriculture arable. The three different combinations revealed that the land-use scenarios have no different influence on the abundance distribution of *M. galathea* in the present and the future.

The final conclusions for the spatial abundance distribution of *M. galathea* suggest that a optimal habitat has a slope of 20°-40°, the urbanisation should no cover more than 25% of an area, the climate should be arid with an annual precipitation total of ca. 800 mm and a annual mean temperature of ca. 8 °C (Fig. 23). The line of trees and hedgerows seems to have a little influence on the abundance of *M. galathea*, according to the models. In the 500 m buffer zone for the 474 butterfly observation locations have ca. 50% of the locations a line of trees and hedgerows, hence the models have probably a problem to use this variable to distinguish the abundance of *M. galathea*. According to the component smooth function of GAM from the best model (Fig. 23), the abundance will decrease if the distance to line of trees and hedgerows is greater than 2 km. The climate change show a range shift to higher elevations especially at the valley sides and the three used land-use scenarios show a minimal or even negligible influence on the abundance of *M. galathea*.

The ensemble model approach has found broad application in ecological research (Araújo and New (2007); Engler et al. (2013)). I used this approach for the prediction of the abundance of *M. galathea* in the present and future under a changing environment, as it was done from Lütolf et al. (2009) for some butterfly species.

Also the influence of protected areas like dry meadows on *M. galathea* was investigated. Dry meadow areas decrease, because of growing agriculture and forest areas. Their conservation requires high investments, but they differ in the number of species, hence a conservation priority for dry meadows is advisable to save money. This was done from Bolliger et al. (2011), they used the number of plant species for the prioritization. So with this thesis I investigated, how important dry meadows for *M. galathea* are and if the conservation effort for dry meadows also pays out for *M. galathea*. The models show no big influence of dry meadows on the spatial abundance distribution of *M. galathea*. This does not mean, that *M. galathea* avoid dry meadows, but it uses other open areas, for example the agriculture pasture.

This thesis generated some problems and issues for further analysis, which are addressed

in the following paragraph.
The error distribution of GLM and GAM do not fulfil the Gaussian distribution (Appendix Fig. 24), here could a negative binomial error distribution help to obtain a Gaussian distributions of the errors. GAM shows often huge outliers, the ensemble model with weights should weaken this effect of GAM. So GAM has the lowest OOB-RMSE in the best model and the largest OOB-RMSE in the worst model, which can be justified with the smoothing function. Another point of discussion is the fact that in the worst models of the complete variable combination set all variable were significant for GLM and GAM. This was not the case in the best models of the complete variable combination set. So a variable selection according to their significance seems not always lead to the best model, but the OOB-RMSE of the single models showed mostly also no significant difference with the two-sided, paired t-test. The ensemble model approach could be expanded with further models like artificial neural networks, generalized boosted regression etc. Further Liechtenstein in the 1x1 km raster grid had over 100 missing points for temperature and precipitation, these missing values were treated superficial. Because these 100 missing observations locations were interpolated with the surrounding locations, which is a coarse approximation. A further investigation could extend the variable slope, because the same slope value can originate form very different terrain features. For example, rough terrain may lower wind speed and vertex possibly affecting *M. galathea* abundance. Furthermore steep areas probably have higher wind speed and hence a different climate, which again could influence the abundance of *M. galathea*. Also the values of the variables, which describe a optimal habitat or a threshold, come just from the component smooth function of GAM of the full model and the best model from the complete variable combination set (Appendix Fig. 22 and 23).

6 Acknowledgement

I thank Janine Bolliger, Rafael Wüest and Dirk Schmatz for their supervision. This work was offered from Janine Bolliger, her ideas were a driving force and Rafael Wüest was the supervisor for the statistics. Dirk Schmatz provided the climate data and climate scenarios. All were scientists with a lot of experience and knowledge, on which this work is based on.

References

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

Baguette, M., Petit, S., and Quéva, F. (2000). Population spatial structure and migration of three butterfly species within the same habitat network: consequences for conservation. *Journal of Applied Ecology*, 37(1):100–108.

Bolliger, J., Eggenberg, S., IsmailL, S., Seidl, I., Kienast, F., et al. (2011). Balancing forest-regeneration probabilities and maintenance costs in dry grasslands of high conservation priority. *Conservation Biology*, 25(3):567–576.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2015). randomforest: Breiman and cutler's random forests for classification and regression. https://cran.r-project.org/web/packages/randomForest/.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chambers, J. M. (1983). *Graphical methods for data analysis*.

CLM-C, Climate Limited-area Model - Community. (2016). Climate limited-area modelling community. http://www.clm-community.eu/.

Davies, S. (2016). Fitting generalized linear models. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html.

IGSM JU, Department of GIS, Cartography and Remote Sensing, Institute of Geography and Spatial Management, Jagiellonian University. (2016). Forecom forest cover changes in mountainous regions - drivers, trajectories and implications. http://www.gis.geo.uj.edu.pl/forecom/project.html.

Engler, R., Waser, L. T., Zimmermann, N. E., Schaub, M., Berdos, S., Ginzler, C., and Psomas, A. (2013). Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. *Forest Ecology and Management*, 310:64–73.

ESRI, Environmental Systems Research Institute, Redlands, California. (2015). Arcgis for desktop version 10.3.1.4959. http://www.esri.com/software/arcgis/arcgis-for-desktop.

FOEN, Federal Office for the Environment. (2010). Trockenwiesen und -weiden von nationaler Bedeutung: Vollzugshilfe zur Trockenwiesenverordnung. http://www.bafu.admin.ch/publikationen/publikation/01553/index.html?lang=de.

FOEN, Federal Office for the Environment. (2012a). Anleitung für die Feldarbeit zum Indikator «Z7-Tagfalter». http://www.biodiversitymonitoring.ch/fileadmin/user_upload/documents/daten/anleitungen/1010_Anleitung_Z7-Tagf_v15.pdf.

FOEN, Federal Office for the Environment. (2012b). Bundesinventar der Trockenwiesen und -weiden von nationaler Bedeutung. http://webintra.wsl.ch/land/inventory/specinvs/gis/geolib/buwal_tww/DB%20TWW%20-%20J031-0639.pdf.

FOEN, Federal Office for the Environment. (2012c). Metadaten Datei: Bundesinventar der Trockenwiesen und -weiden von nationaler Bedeutung. http://webintra.wsl.ch/land/inventory/specinvs/gis/geolib/buwal_tww/GM03_TWW.pdf.

FOEN, Federal Office for the Environment. (2012d). Metadaten Datei: Bundesinventar der Trockenwiesen und -weiden von nationaler Bedeutung - Anhang 2. http://webintra.wsl.ch/land/inventory/specinvs/gis/geolib/buwal_tww/GM03_TWW_A2.pdf.

FOEN, Federal Office for the Environment. (2014). Biodiversitätsmonitoring Schweiz BDM. http://www.bafu.admin.ch/publikationen/publikation/01766/index.html?lang=en.

MeteoSwiss, Federal Office of Meteorology and Climatology. (2016). Climate of switzerland. http://www.meteoswiss.admin.ch/home/climate/past/climate-of-switzerland.html.

SFSO, Federal Office of Statistic. (2013). Land use in Switzerland: Results of the Swiss land use statistics . CH-Neuchatel.

SFSO, Federal Office of Statistic. (2016). Arealstatistik nach Nomenklatur 2004 . http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/arealstatistik_2004.html.

Swisstopo, Federal Office of Topography. (2005). DHM Das digitale Höhenmodell der Schweiz. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/height/dhm25.html.

Swisstopo, Federal Office of Topography. (2015). Objektkatalog swissTLM3D 1.3. http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/swissTLM3D.parsysrelated1.47641.downloadList.94330.DownloadFile.tmp/201503swisstlm3d13dbarrierefrei.pdf.

FEDRO, Federal Roads Office. (2002). Normalprofile, Rastplätze und Raststätten der Nationalstrassen. http://www.astra.admin.ch/dienstleistungen/00129/00183/00515/index.html?lang=de.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models, volume 43 of monographs on statistics and applied probability.

Lütolf, M., Bolliger, J., Kienast, F., and Guisan, A. (2009). Scenario-based assessment of future land use change on butterfly species distributions. *Biodiversity and Conservation*, 18(5):1329–1347.

Maggini, R., Lehmann, A., Zbinden, N., Zimmermann, N. E., Bolliger, J., Schröder, B., Foppen, R., Schmid, H., Beniston, M., and Jenni, L. (2014). Assessing species vulnerability to climate and land use change: the case of the swiss breeding birds. *Diversity and distributions*, 20(6):708–719.

MPI, Max Planck Institute. (2016). New earth system model of max planck institute for meteorology. http://www.mpimet.mpg.de/en/science/models/mpi-esm.html.

McCullagh, P. and Nelder, J. (1989). Generalised linear models. chapman and hall. *London, UK*.

Pachauri, R. K., Allen, M., Barros, V., Broome, J., Cramer, W., Christ, R., Church, J., Clarke, L., Dahe, Q., Dasgupta, P., et al. (2014). Climate change 2014: Synthesis report. contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change.

Price, B., Kienast, F., Seidl, I., Ginzler, C., Verburg, P. H., and Bolliger, J. (2015). Future landscapes of switzerland: Risk areas for urbanisation and land abandonment. *Applied Geography*, 57:32–41.

Schmatz, D. (2015a). Mpi-m-mpi-esm-lr_rcp45_r1i1p1_clmcom-cclm4-8-17_ano_1961-1990. Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf.

Schmatz, D. (2015b). Mpi-m-mpi-esm-lr_rcp85_r1i1p1_clmcom-cclm4-8-17_ano_1961-19900. Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf.

Stanford, Stanford University. (2016). Rms error. `http://statweb.stanford.edu/~susan/courses/s60/split/node60.html`.

Therneau, T., Atkinson, B., and Ripley, B. (2015). rpart: Recursive partitioning and regression trees. `https://cran.r-project.org/web/packages/rpart//`.

Thornton, P. E., Running, S. W., and White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, 190(3):214–251.

Tolman, T., Lewington, R., and Nuß, M. (1998). *Die Tagfalter Europas und Nordwestafrikas*. Franckh-Kosmos.

Vandewoestijne, S., Martin, T., Liégeois, S., and Baguette, M. (2004). Dispersal, landscape occupancy and population structure in the butterfly melanargia galathea. *Basic and Applied Ecology*, 5(6):581–591.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79.

Wood, S. (2009). Generalized additive models with integrated smoothness estimation. *R help for package mgcv*.

Wood, S. (2016). mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation. `https://cran.r-project.org/web/packages/mgcv/`.

Xu, Q.-S. and Liang, Y.-Z. (2001). Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11.

7 Appendix supplementary material

Table 14: The used road types for the variable road and the width/width range from the literature. From a width range the mean was used for the buffer width of the lines.

Road type	Width	Used width	Source
Motorway	27.5m	27.5m	(FEDRO, 2002)
Limited-access road	25m	25m	(FEDRO, 2002)
10m Road	>10.2 m	10.2m	(Swisstopo, 2015)
8m Road	8.21-10.2 m	9.205m	(Swisstopo, 2015)
6m Road	6.21-8.2 m	7.205m	(Swisstopo, 2015)
4m v	4.21-6.2 m	5.205m	(Swisstopo, 2015)
3m Road	2.81-4.2 m	3.505m	(Swisstopo, 2015)
2m Walkway	1.81-2.8 m	2.305m	(Swisstopo, 2015)
1m Walkway	<1.8 m	1.4m	(Swisstopo, 2015)
1m partially walkway	<1.8 m	1.4m	(Swisstopo, 2015)
2m partially walkway	1.81-2.8 m	2.305m	(Swisstopo, 2015)

Table 15: The used railway types for the variable railway and the width/width range from the literature. From a width range the mean was used for the buffer width of the lines.

Railway type	Width (Swisstopo, 2015)	Used width
Normalspur	1435 mm	1435 mm
Schmalspur	750 – 1435 mm	1093 mm
Schmalspur mit NS		1435 mm
Kleinbahn	500 – 750 mm	625 mm

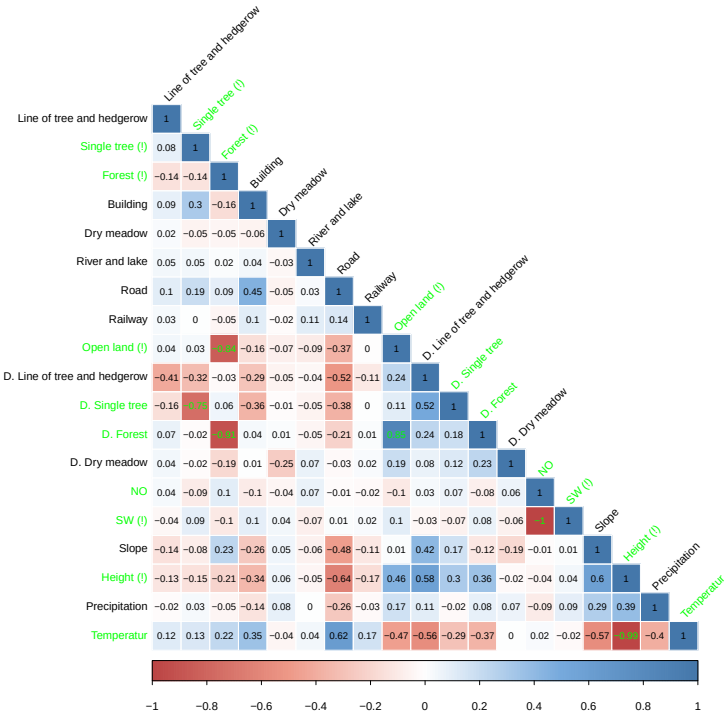


Figure 17: The correlation between all variables for the 50 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

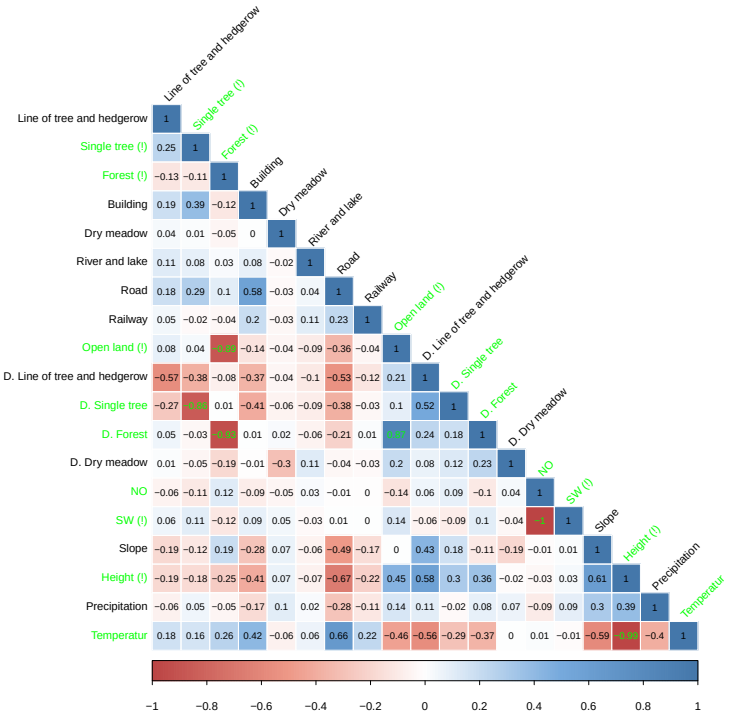


Figure 18: The correlation between all variables for the 100 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

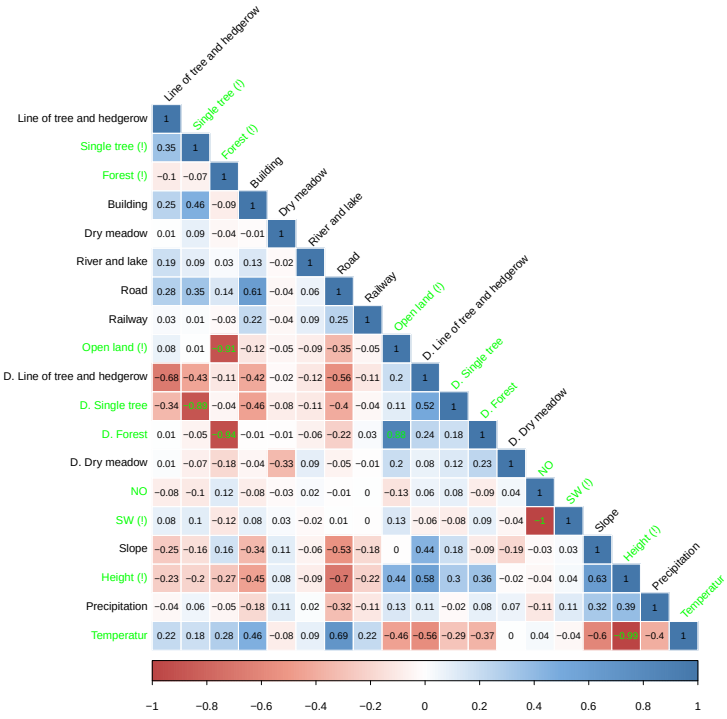


Figure 19: The correlation between all variables for the 150 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

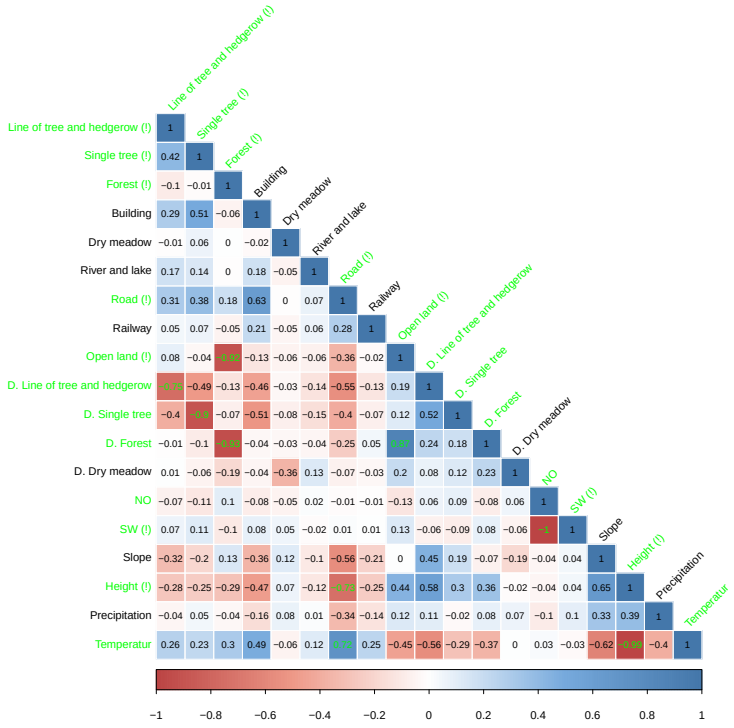


Figure 20: The correlation between all variables for the 200 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

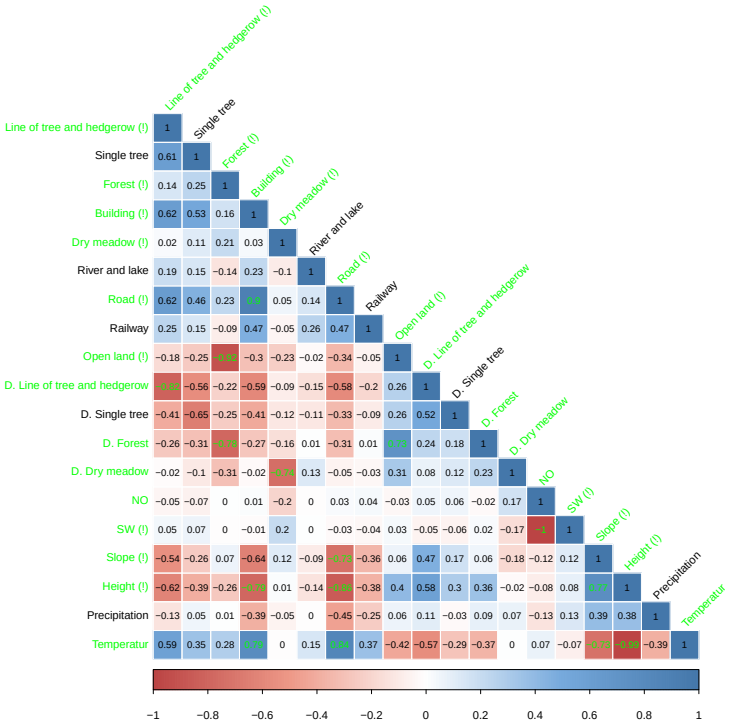


Figure 21: The correlation between all variables for the 1000 m buffer zone. The variables correlated > 0.7 are shown in green, (!) indicates variables removed from the analysis.

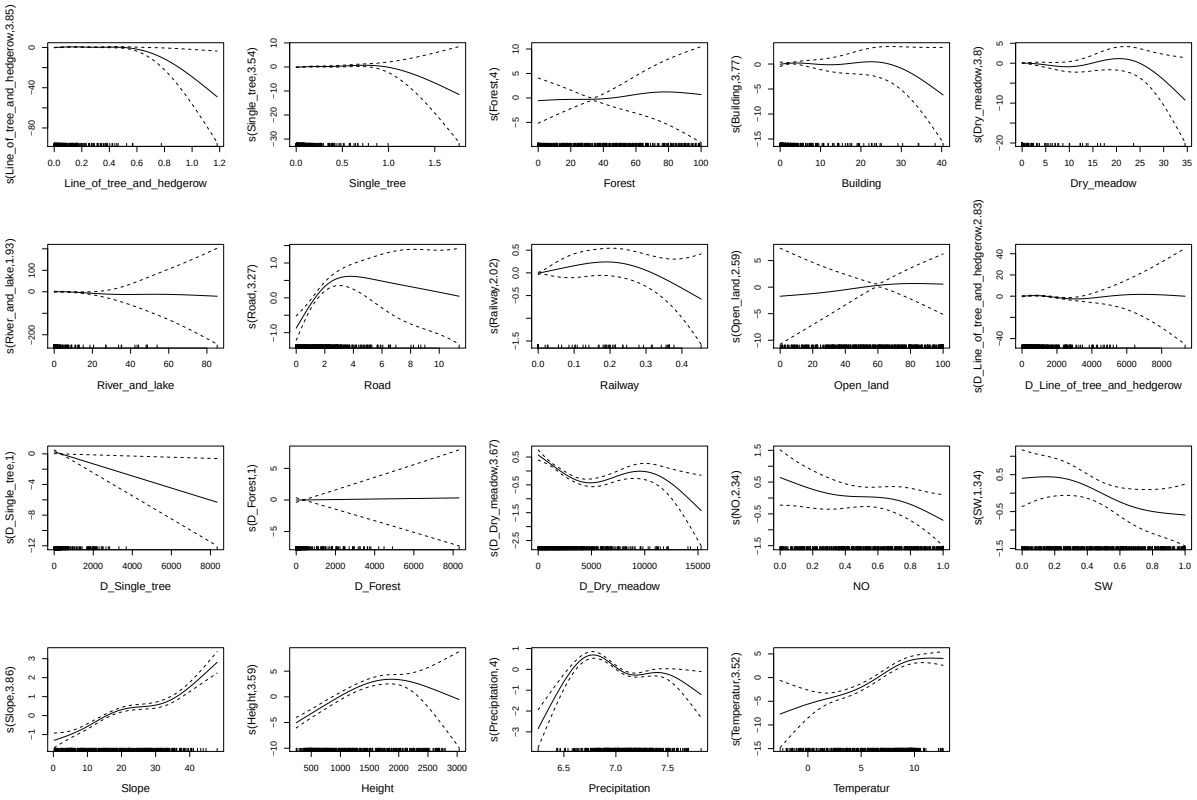


Figure 22: The component smooth function of GAM from the full model from the 500 m buffer zone, which are 12 variables.

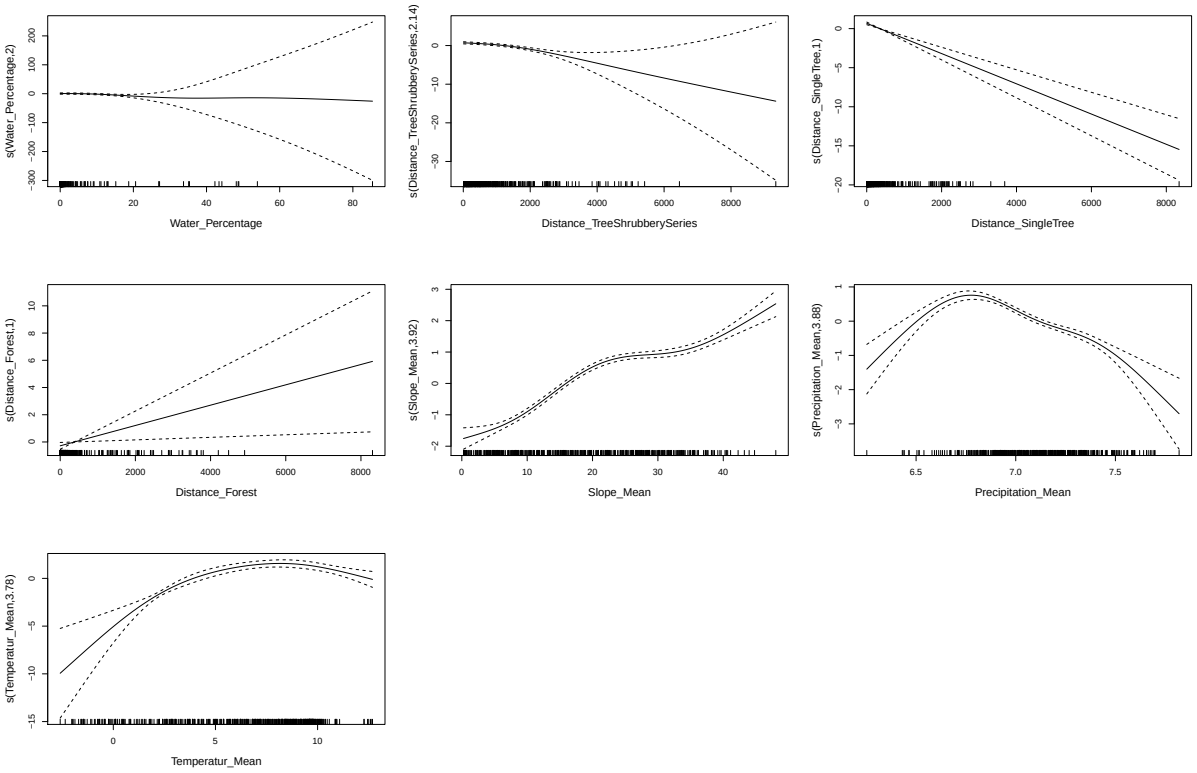
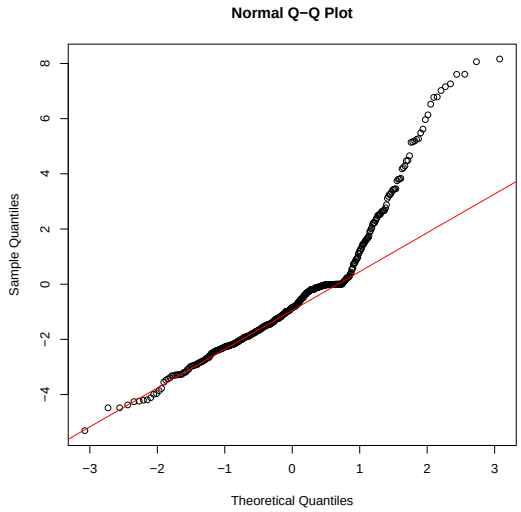
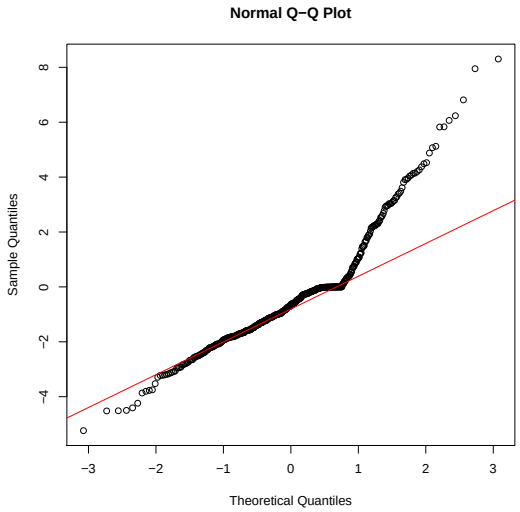


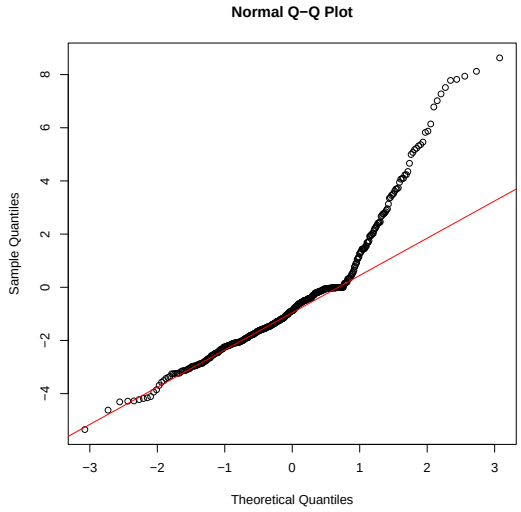
Figure 23: The component smooth function of GAM from the best model of the complete variable combination set from the 500 m buffer zone.



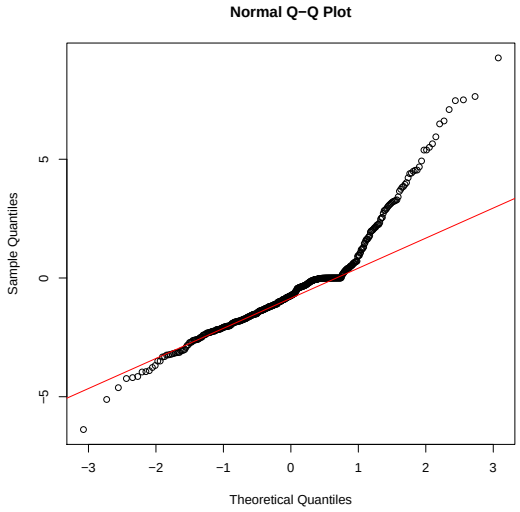
(a) GLM complete model



(b) GAM complete model

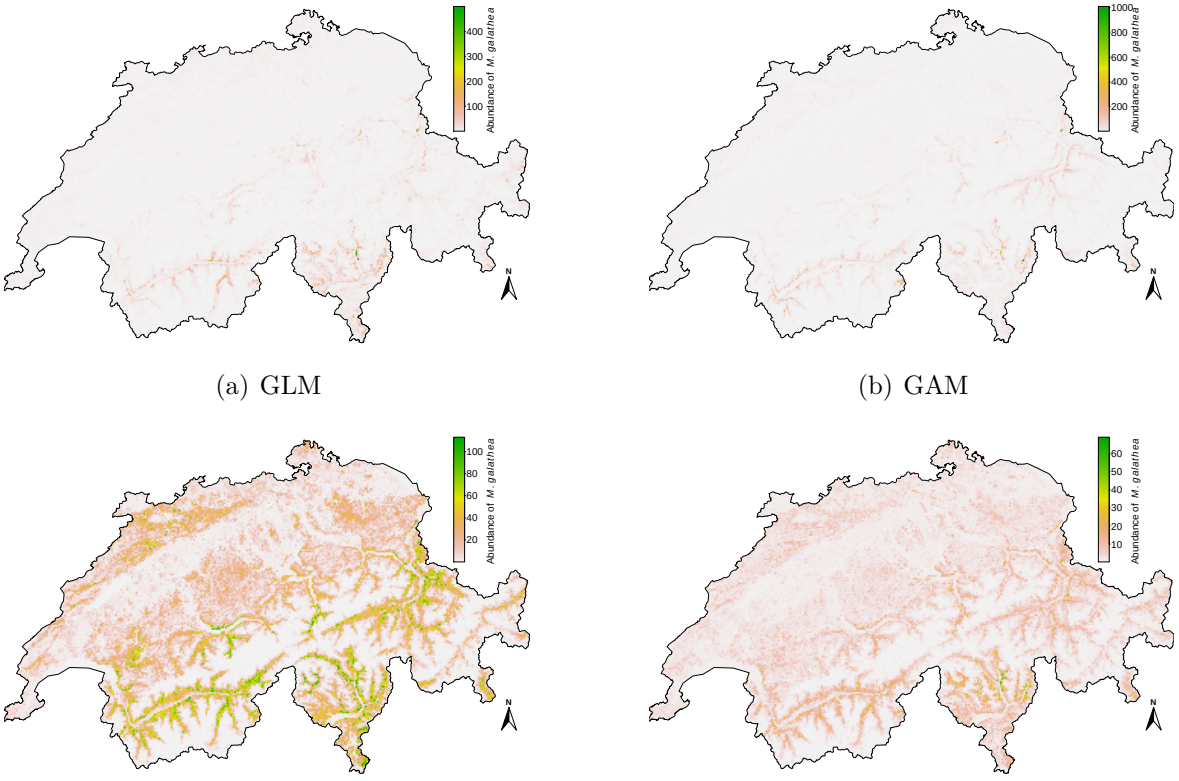


(c) GLM best model



(d) GAM best model

Figure 24: The Q-Q plots for GLM and GAM, one with the complete model (12 variable) and the other with the best model from the complete variable combination set from the 500 m buffer zone.



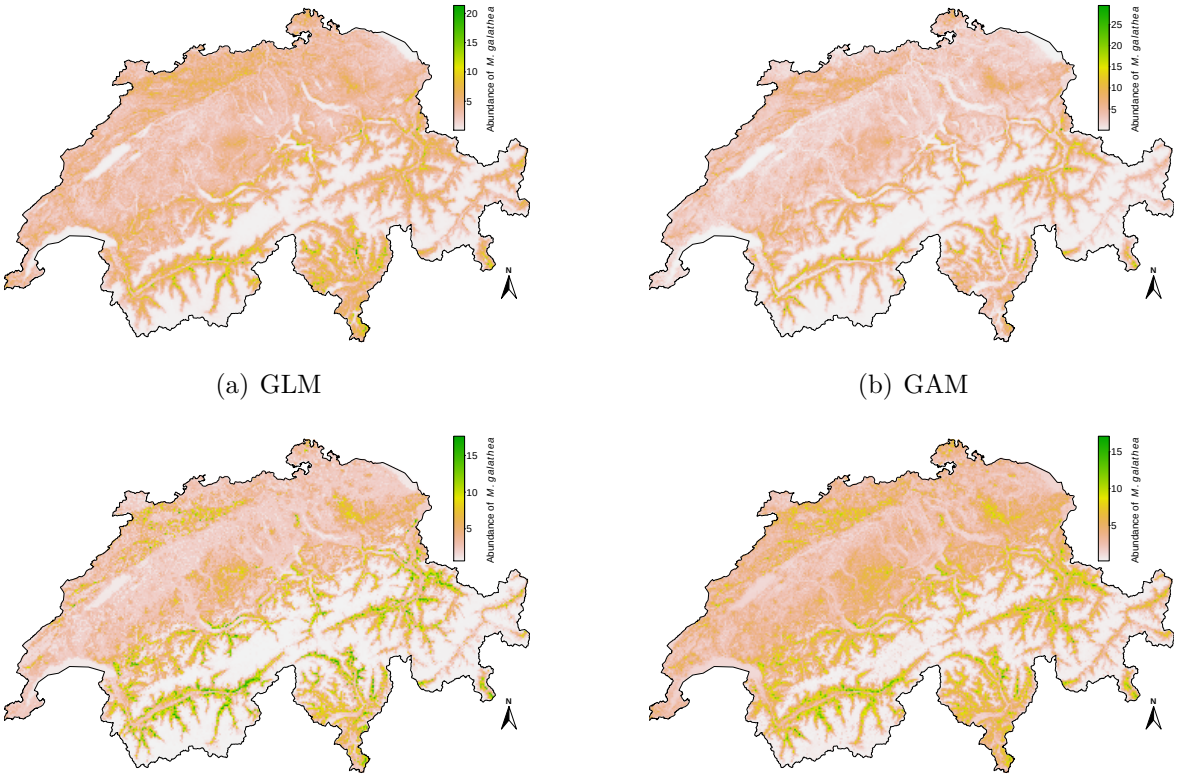
(a) GLM

(b) GAM

(c) RPART

(d) RF

Figure 25: The variance of the spatial projection of the four ensemble models, each has all variable combinations with the OOB-RMSE as weights.



(a) GLM

(b) GAM

(c) RPART

(d) RF

Figure 26: The spatial projection of four ensemble models, each has all variable combinations with no weights.

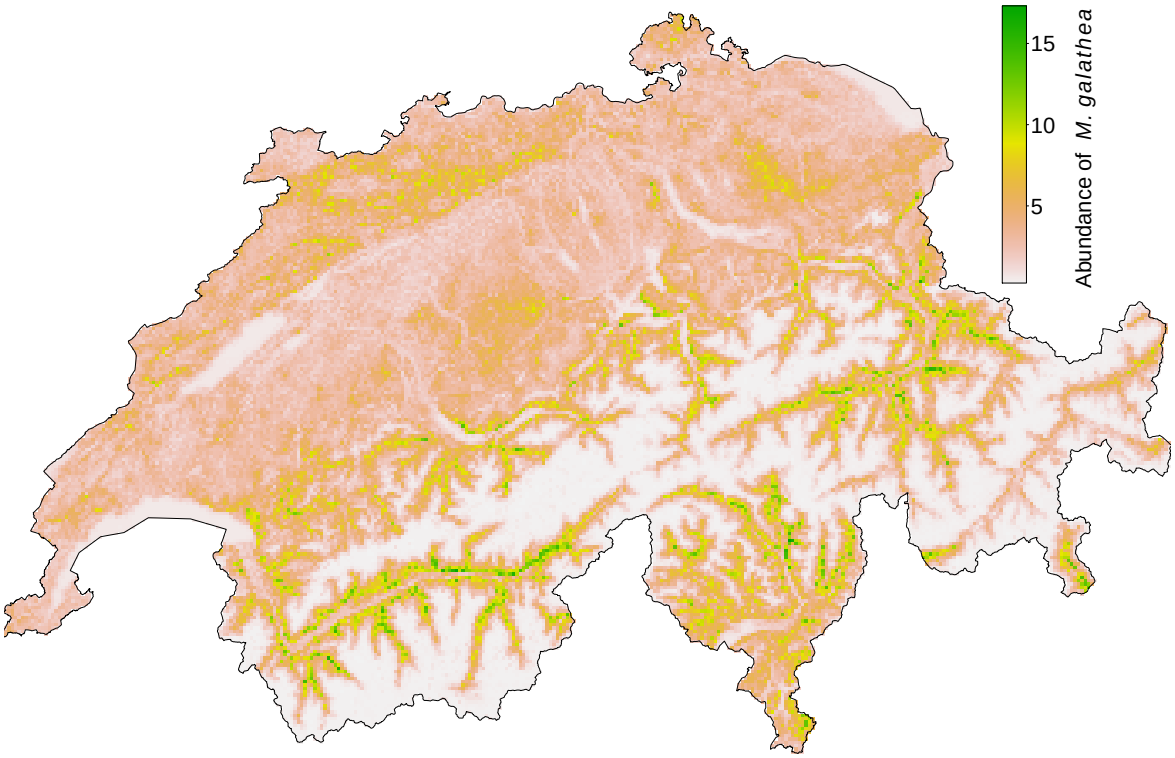
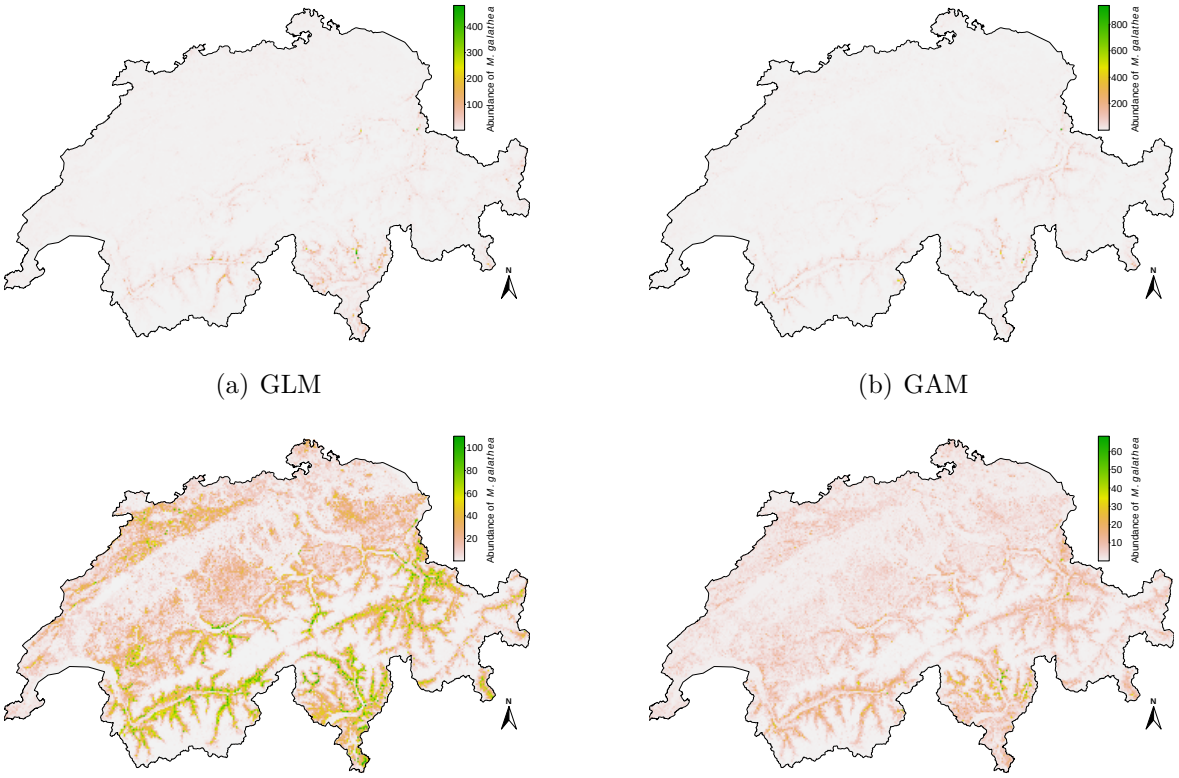


Figure 27: The mean of the four ensemble models from Figure 26, each with the complete variable combination set.



(a) GLM

(b) GAM

(c) RPART

(d) RF

Figure 28: The variance of the spatial projection of the four ensemble models, each has all variable combinations with no weights.

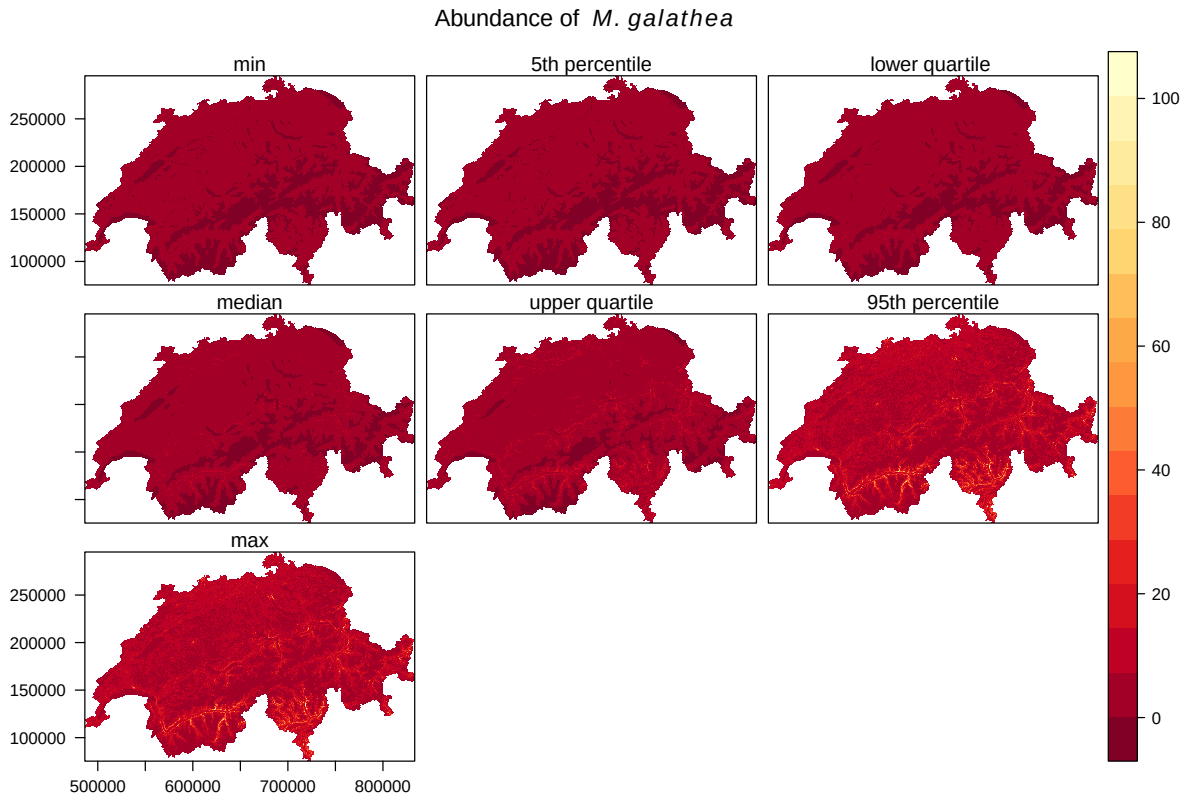


Figure 29: The min, quartile, percentile and max for GLM from the ensemble model with the complete variable combination set for the 500 m buffer zone.

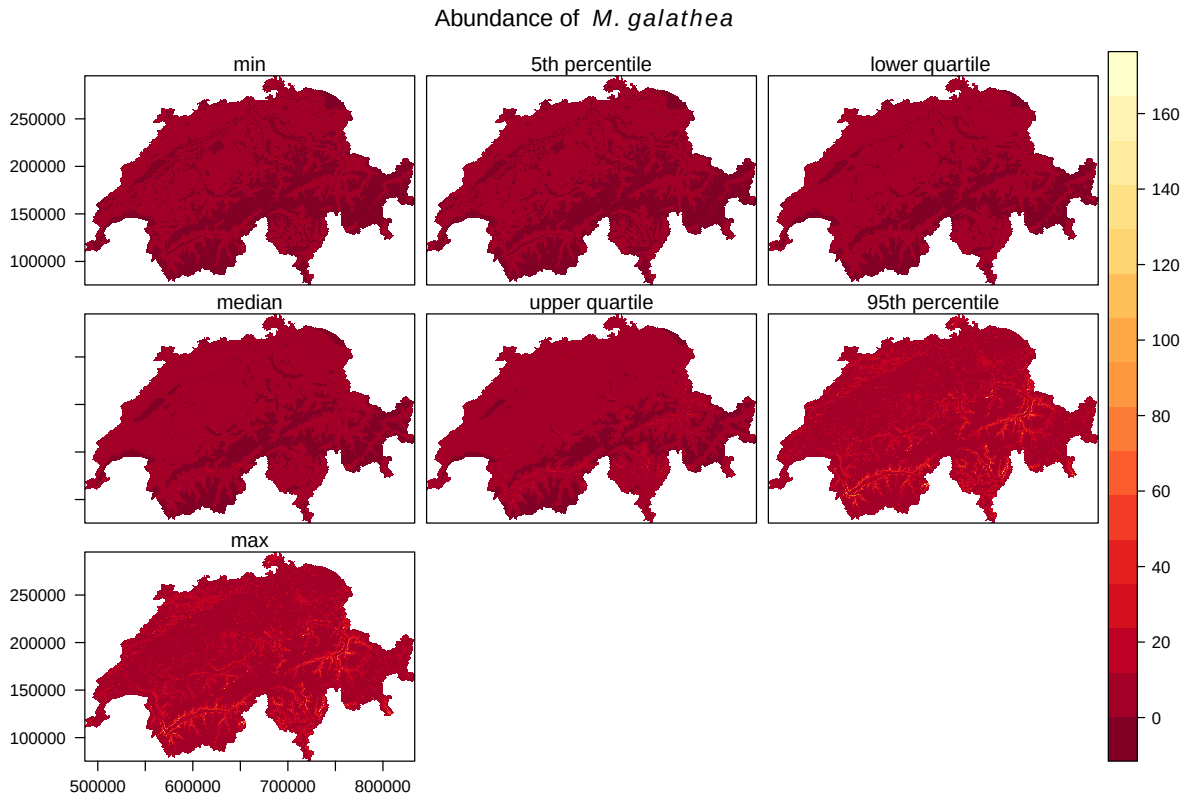


Figure 30: The min, quartile, percentile and max for GAM from the ensemble model with the complete variable combination set for the 500 m buffer zone.

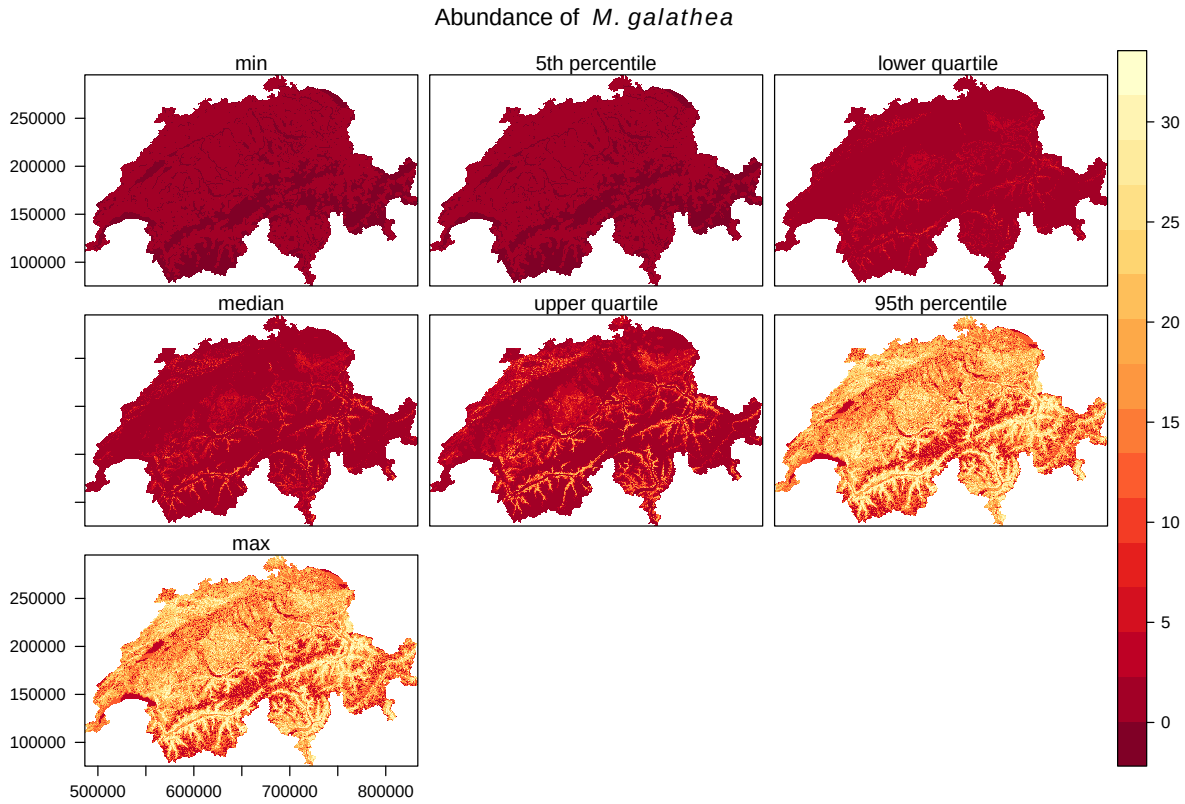


Figure 31: The min, quartile, percentile and max for RPART from the ensemble model with the complete variable combination set for the 500 m buffer zone.

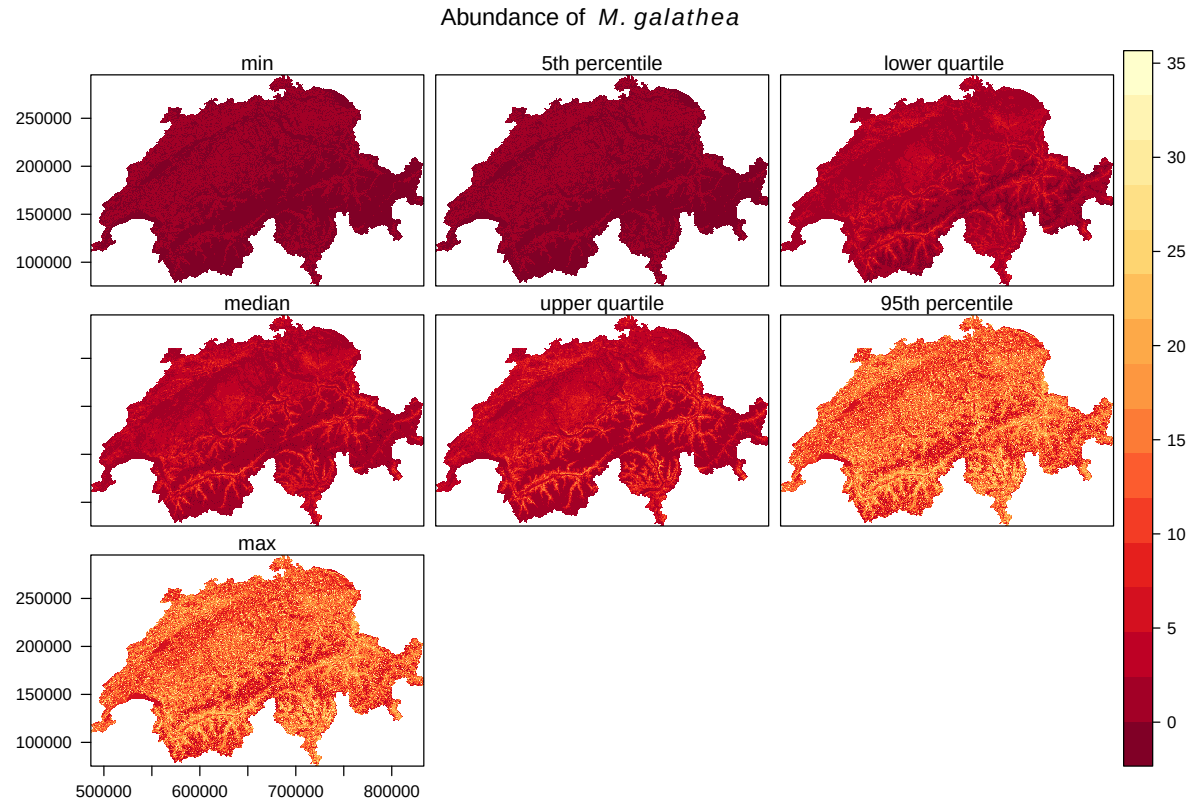


Figure 32: The min, quartile, percentile and max for RF from the ensemble model with the complete variable combination set for the 500 m buffer zone.

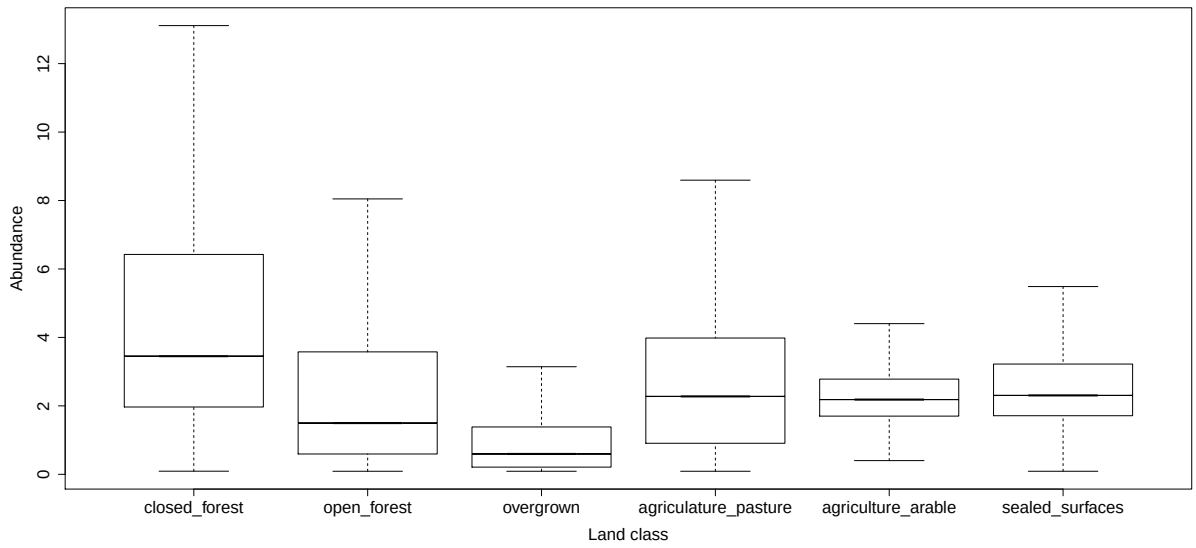
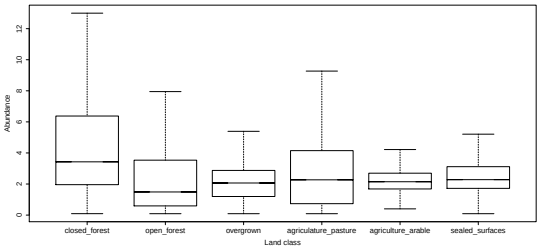
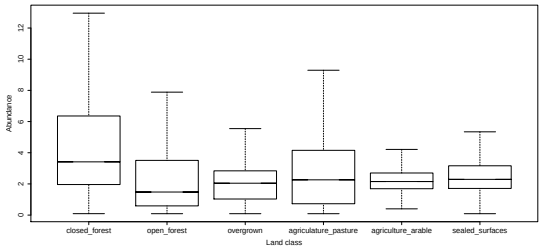


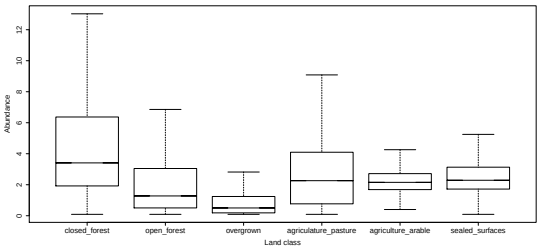
Figure 33: The boxplot with the abundance of *M. galathea* from the ensemble model 1BM for every land class from the year 2009. Outliers are not included, because they can be a few ten thousands, depending on the land class.



(a) A1; RCP4.5

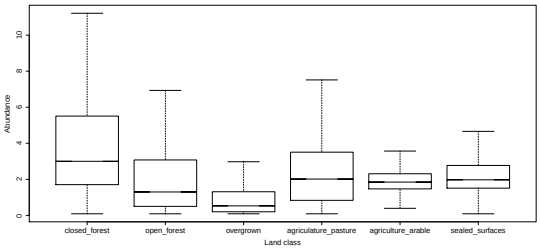


(b) B2; RCP4.5

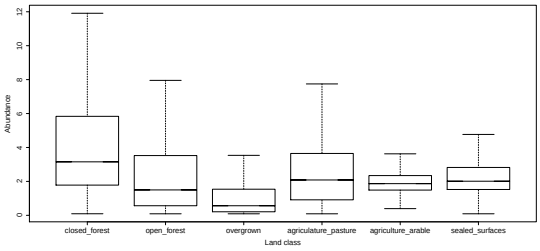


(c) B2; RCP4.5

Figure 34: The boxplots with the abundance of *M. galathea* from the ensemble model 1BM for every land class from the land-use scenarios Trend, A1 and B2. Outliers are not included, because they can be a few ten thousands, depending on the land class.

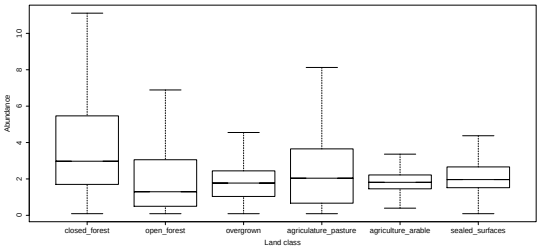


(a) A1; RCP4.5

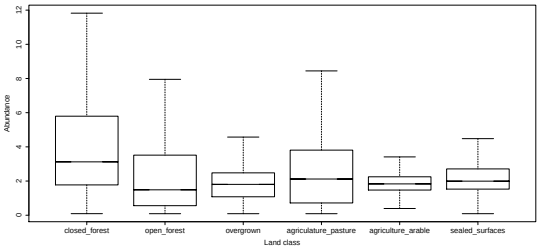


(b) B2; RCP4.5

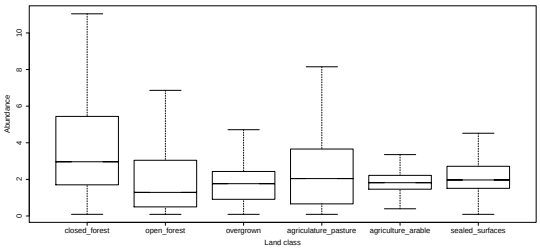
Figure 35: The boxplots with the abundance of *M. galathea* from the ensemble model 1CS with both climate scenario for every land class from the land-use in 2009. Outliers are not included, because they can be a few ten thousands, depending on the land class.



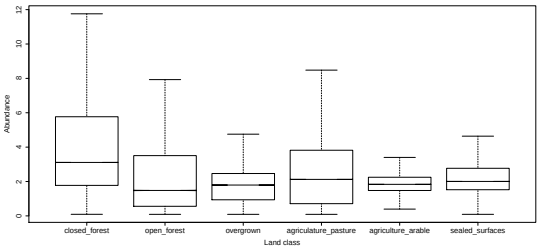
(a) Trend; RCP4.5



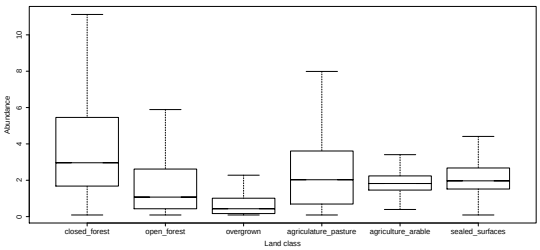
(b) A1; RCP4.5



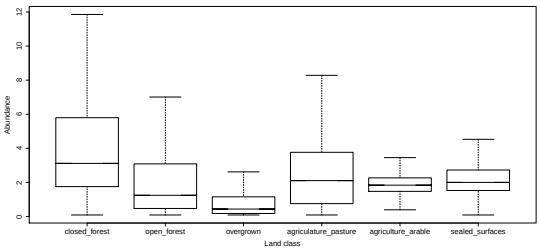
(c) B2; RCP4.5



(d) B2; RCP4.5



(e) B2; RCP4.5



(f) B2; RCP4.5

Figure 36: The boxplot with the abundance of *M. galathea* from the ensemble model 1CS with both climate scenarios for every land class from the land-use scenarios Trend, A1 and B2. Outliers are not included, because they can be a few ten thousands, depending on the land class.