

Milestone MS412

Version: 1.1
Date: 21.09.2015
Author: FEM, UFZ, RMCA, SGN,
EBCC-CTFC, HCMR, BGBM
Document reference: Milestone MS412



First species distribution models runs using throughput methods, relying on open source algorithms (M34)

STATUS: FINAL

Project acronym: EU BON
Project name: EU BON: Building the European Biodiversity Observation Network
Call: ENV.2012.6.2-2
Grant agreement: 308454
Project Duration: 01/12/2012 – 31/05/2017 (54 months)
Co-ordinator: MfN, Museum für Naturkunde - Leibniz Institute for Research on Evolution and Biodiversity, Germany

Partners: UTARTU, University of Tartu, Natural History Museum, Estonia
UEF, University of Eastern Finland, Digitisation Centre, Finland
GBIF, Global Biodiversity Information Facility, Denmark
UnivLeeds, University of Leeds, School of Biology, UK
UFZ, Helmholtz Centre for Environmental Research, Germany
CSIC, The Spanish National Research Council, Doñana Biological Station, Spain
UCAM, University of Cambridge, Centre for Science and Policy, UK
CNRS-IMBE, Mediterranean Institute of marine and terrestrial Biodiversity and Ecology, France
Pensoft, Pensoft Publishers Ltd, Bulgaria
SGN, Senckenberg Gesellschaft für Naturforschung, Germany
SIMBIOTICA, Simbiotica S.L., Spain
FIN, FishBase Information and Research Group, Inc., Philippines
HCMR, Hellenic Centre for Marine Research, Greece
NHM, The Natural History Museum, London
BGBM, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany
UCPH, University of Copenhagen: Natural History Museum of Denmark, Denmark
RMCA, Royal Museum of Central Africa, Belgium
PLAZI, Plazi GmbH, Switzerland
GlueCAD, GlueCAD Ltd. – Engineering IT, Israel
IEEP, Institute for European Environmental Policy, UK
INPA, National Institute of Amazonian Research, Brazil
NRM, Swedish Museum of Natural History, Sweden
IBSAS, Slovak Academy of Sciences, Institute of Botany, Slovakia
EBCC-CTFC, Forest Technology Centre of Catalonia, Spain
NBIC, Norwegian Biodiversity Information Centre, Norway
FEM, Fondazione Edmund Mach, Italy
TerraData, TerraData environmetrics, Monterotondo Marittimo, Italy
EURAC, European Academy of Bozen/Bolzano, Italy
WCMC, UNEP World Conservation Monitoring Centre, UK
UGR, University of Granada, Spain

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308454.



EU BON

EU BON: Building the European Biodiversity Observation Network
Project no. 308454

Large scale collaborative project

MS412**First species distribution models runs using throughput methods,
relying on open source algorithms**

Milestone number	MS412
Milestone name	First species distribution models runs using throughput methods, relying on open source algorithms
WP no.	WP4
Lead Beneficiary (full name and Acronym)	Fondazione Edmund Mach (FEM)
Nature	Written report
Delivery date from Annex I (proj. month)	2015-09-21 (M34)
Delivered	[yes]
Actual forecast delivery date	2015-09-30
Comments	

Name of the Authors	Name of the Partner	Logo of the Partner
Carol X. Garzon-Lopez	Fondazione Edmund Mach	
Duccio Rocchini		
Mathias Kuemmerlen	Senckenberg Gesellschaft für Naturforschung	
Nicolas Titeux	European Bird Census Council/ Centre Tecnològic Forestal de Catalunya (EBCC/CTFC)	
Lluís Brotons		
Ingolf Kühn	Helmholtz Centre for Environmental Research - UFZ	
Johannes Penner	Museum für Naturkunde	
Patricia Mergen	Royal Museum of Central Africa	
Israel Peer	GlueCAD	
Charles Marsh	University of Leeds	

In case the report consists of the delivery of materials (guidelines, manuscripts, etc)

Delivery name	Delivery name	From Partner	To Partner

Contents

Summary of the Milestone

1. Introduction

2. Achievements and current status

2.1. Data sources

2.1.1. Biodiversity data

- a. GBIF data
- b. Local data sources (test sites)
- c. DNA sequence data

2.1.2. Environmental data

- a. Bioclimatic variables
- b. Land Surface Temperature (LST) datasets
- c. Land use and landcover data (LULC)
- d. Land use changes and trends
- e. Hierarchical structure of habitat classification scheme
- f. Other environmental variables

2.2. Biotic interactions in species distribution models

2.3. Species distribution models developed by the partners

2.3.1. Catchment-scale freshwater species distribution models

2.3.2. The SpaNiche hybrid model

2.3.3. Modelling DNA fungi data

2.3.4. Effect of data sources (environmental and species) on SDM approaches

2.3.5. Within-species spatial niche variation and projections of species distribution under future climate change

2.3.6. Use case of small islands with high mountains

2.4. Representation of uncertainty

- a. Catchment-scale freshwater species distribution models
- b. Effect of data sources on SDM approaches
- c. Within-species niche variation

3. Directions

4. References

Summary of the Milestone

Species distribution models have become a paramount tool to biodiversity assessments. And this tool, in combination with remote sensing products and the current global datasets of species ground observations, results in a powerful approach to monitor biodiversity at multiple spatial and temporal scales.

Even though, great advancement in the development of these tools has been reached, the strength of this combination - modelling methods, remote sensing products and species observations - relies in its informed, and careful, selection and application using integrative modelling approaches.

The goal of this milestone is to provide the first species distribution models using throughput methods, relying on open source algorithms. To accomplish this goal we divided the milestone in four main areas:

1. **Data sources:** Species distribution models require reliable information regarding the environmental conditions at the study area, as well as ground observations of the focal species. There are multiple sources of this information, and they vary in their characteristics and quality, and therefore must be selected depending on the system and aim of the study. In this section of the milestone, there is a description of the sources of environment and species data that includes an assessment of the strengths and weaknesses for species distribution modelling.
2. **Biotic interactions:** Species distribute depending on their physical requirements (i.e. environmental variables) and the distribution of other species (e.g. predators, prey, hosts, parasites, etc.), thus quality of the distribution models also depend on the information regarding the characteristics and biotic interactions of the focal species. In this section, biotic interactions are described in the light of species distribution models, in order to provide a background for their inclusion in the integrative modelling approaches that are being developed as part of EU BON.
3. **Species distribution models:** In this section the first integrative species distribution modelling approaches that have been developed by the partners in task 4.1 are described, providing information on the methods, implementation and the outcomes. Additionally, there is an analysis of sampling bias due to data sources (environment and species) and site characteristics (e.g. small islands with high mountains) on species distribution models.
4. **Representation of uncertainty:** Uncertainty in the outcomes of the species distribution models is due to the variability in the characteristics and quality of the data sources and the modelling approaches. Such uncertainty must be explicitly presented in order to ensure transparency of the methods, and increase the applicability of the resulting models to end-users. This section includes the identification of uncertainty in the modelling approaches applied with a set of strategies to quantify and present it.

Finally, the links to other work packages and the future directions on the work of *task 4.1* are presented.

1. Introduction

Species distribution models are an important tool to predict current species and to project potential (past and future) distribution across landscapes. This can be done based on changing environmental predictors, such as different scenarios on land and marine use and climatic data. In the previous Milestone (MS 411; Scoping out integrative analyses approaches for biodiversity distribution

modelling) we had identified a set of species distribution models that meet the aim of accounting for the variety of data available and can be implemented in Free and Open Source Software.

We focus on **four key challenges in the selection**, implementation and analysis of species distribution models: **i)** the characteristics of the input data (Beck et al. 2014) and the evaluation of the effect of the properties of such data on the outcome, **ii)** the effect of the model algorithm itself, **iii)** the importance of explicitly incorporating spatial scale and biotic interactions in the modelling process and, **iv)** the assessment and representation of the uncertainty contained in the outcome (Rocchini et al. 2011).

In this milestone, and based on the methods previously selected (MS411), we analysed the characteristics of the available data sources, evaluated the effect of these data on the model performance, developed and tested integrative species distribution modelling techniques and scope approaches to represent uncertainty. All these, are using Free and Open Source algorithms.

2. Achievements and current status

2.1. Data sources

2.1.1. Biodiversity data

a. GBIF data

One of the most important and increasingly used, source of species data, that is, species field-observations, is the Global Biodiversity Information Facility (GBIF, 2015), an international open data infrastructure that allows access to more than 500 million records of more than 1.5 million species worldwide (www.gbif.org). Such a vast, and constantly growing, dataset is of high relevance for its potential applications within EU BON, as highlighted by WP2 in their deliverable D2.1.

GBIF brings together data ranging from 1900 (but older records can also be found) to the current date and from museum collections to direct field observations (**Fig. 1**). Such an enormous database provides a huge potential for analysis of past, current and future trends in species distribution, but the ample range of data sources implies a large variation in the methods used/available which in turn affects the characteristics and accuracy of the records (Gaiji et al. 2013).

Data used in species distribution models are prone to several sources of bias, which usually arise during data collection or sampling: (i) a spatial bias due to unequal distribution of sampling sites in space (Bean et al. 2012, Hortal et al. 2008), (ii) taxonomical bias, due to misidentification, and (iii) temporal bias that stems from an unequal sampling frequency of the different sites (Bean et al. 2012, Rota et al. 2011).

Identification of specimens may also play a role in large datasets as taxonomical expertise varies largely among data sources, because there are changes in taxonomy across time, and finally because the newly found species are missing from the records.

Furthermore, GBIF currently only provides presence data, but advancements are being made (in collaboration with EU BON) to include relevant information such as, absences, organism quantities, sampling protocol and sample size (<http://eubon-ipt.gbif.org>). This limits the species distribution models that can be used, and only allows for relative likelihood to be modelled, not true probability of occurrence in cases where the surveying is not intensive enough to assume absence (Guillera-Arroita et al. 2015). This severely limits comparability between species and different models of the same species.

GBIF, undoubtedly, provides an essential database for the estimation of multinational scale trends in species distributions, which is of paramount importance in the development of species conservation strategies at multiple scales, from reserve design to global conservation priorities. Consequently, it is critical to find tools to provide outcomes that contained information on the degree of uncertainty that arise from the biases described.

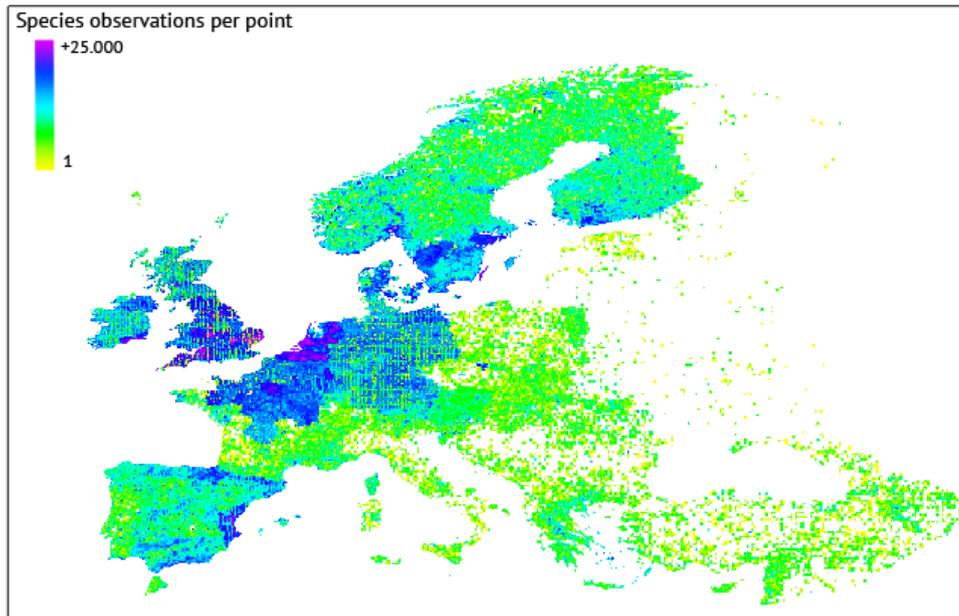


Figure 1. Species observation of taxa plants in the GBIF dataset (website accessed Feb 2015).

b. Local data sources (test sites)

Significant data on biodiversity is frequently recorded by localized, small-scale, research efforts ranging from sporadic monitoring by a local authority to regular standardized sampling from a long term study. Typically, such data sets do not have an extensive geographic coverage; however their strength lies in the repeated sampling through time, producing time series that allow deriving temporal trends.

Data from test sites can be quite different between sites, as sampling protocols tend to differ between countries and regions. Recent efforts aim at the homogenization of sampling and data recording, to allow for comparable analysis and results. In addition, some of these sites are operated as long-term ecological research (LTER) sites, where the sampling method is standardized. Adhering to an existing research framework promotes the formation of research networks that pool their data resources and make them more visible to data users (e.g. LTER Networks).

Furthermore, long-term sampling often evolves through time, either adding or modifying the amount of data sampled. For example, research questions may change and require a new sampling framework (Kuemmerlen et al. accepted). This may hamper the applicability of data, but in general terms improves its quality. However, such issues in datasets may be overcome through the application of statistical methods.

c. DNA sequence data

DNA sequence data are a special case of presence only data. All restrictions mentioned for GBIF data do in principle also apply to sequence data. However, due to sampling strategies, data are even more scarce and hence in danger of having an even stronger spatial and temporal bias. Furthermore, DNA sequence is “translated” into “taxonomic hypothesis” which can result in a taxonomic bias, favouring

those taxa that are preferred for genetic analyses. Hence, using DNA sequence data for “species” distribution modelling hence requires even more care than when using ordinary species data.

2.1.2. Environmental data

In nature species show specific requirements that limit their distribution, thus each species (or set of species) occupies the space that matches the set of conditions that allow them to establish and survive. In this framework, environmental variables act as proxies to predict species distribution from species occurrences, and to project future trends via future environmental scenarios.

However, the strength of the association between species and the physical environment is modified by biotic interactions, dispersal limitation, source-sink dynamics (Pulliam 2000) and the ability of species to cope with the environment, and all these factors, together with the environmental variables, take place at multiple scales (Whittaker et al. 2001). At large scales, species composition varies with global climate patterns (Rhode 1992) while at smaller scales, species composition is affected by local environmental conditions, biotic interactions and dispersal (Duivenvoorden et al. 2002). Consequently, spatial scale constitutes an essential feature when predicting species distribution (Henle et al. 2014, Domisch et al. 2015). Currently, models are available to incorporate dispersal and biotic interactions into SDM (e.g. Kissling et al. 2012).

The environmental data available comes from multiple sources (e.g. meteorological stations, satellite data, etc) and thus vary in accuracy and resolution (Metz et al. 2014). But, the selection of environmental data sources depend on, the ecology of the study species (or set of species), the characteristics of the data in terms of resolution, extent and quality of both environmental data as well as the species occurrences and the research question for which the SDMs are used. Below we provide a brief description of the environmental data sources focusing on the applications for species distribution modelling.

a. Bioclimatic variables

The Worldclim and Climond datasets include 40 climate layers (19 layers in Worldclim and 16 in Climond, **Table 1**) derived from the interpolation of monthly climate data from weather stations (Hijmans et al. 2005), offering an impressive set of environmental estimates, of the terrestrial portion, for entire world at a spatial resolution of 1 km² extracted from a temporal extent of 50 years (from the year 1950 to the year 2000). However, there are a number of concerns in the accuracy of the grid values especially for areas with a lower number of weather stations and, according to the authors, in mountainous areas for some of the variables (Hijmans et al. 2005). Also one needs to take into account that the high resolution data stems from a scaling procedure and may indeed be representative for a much coarser resolution.

Table 1. Bioclimatic variables contained in the Worldclim (1 to 19) and Climond (1 to 35) dataset. Table modified from Hutchinson et al. (2009) and Kriticos et al. (2014).

Variable Code	Description	Minimum temp (°C)	Maximum temp (°C)	Rainfall (mm month-1)	Radiation (W m-2d-1)	Pan evaporation (mm d-1)
Bio01	Annual mean temperature (°C)	x	x			
Bio02	Mean diurnal temperature range (mean(period max-min)) (°C)	x	x			
Bio03	Isothermality (Bio02 ÷ Bio07)	x	x			
Bio04	Temperature seasonality (C of V)	x	x			
Bio05	Max temperature of warmest week (°C)		x			
Bio06	Min temperature of coldest week (°C)	x				
Bio07	Temperature annual range (Bio05-Bio06) (°C)	x	x			
Bio08	Mean temperature of wettest quarter (°C)	x	x	x		
Bio09	Mean temperature of driest quarter (°C)	x	x	x		
Bio10	Mean temperature of warmest quarter (°C)	x	x			
Bio11	Mean temperature of coldest quarter (°C)	x	x			
Bio12	Annual precipitation (mm)			x		
Bio13	Precipitation of wettest week (mm)			x		
Bio14	Precipitation of driest week (mm)			x		
Bio15	Precipitation seasonality (C of V)			x		
Bio16	Precipitation of wettest quarter (mm)			x		
Bio17	Precipitation of driest quarter (mm)			x		
Bio18	Precipitation of warmest quarter (mm)	x	x	x		
Bio19	Precipitation of coldest quarter (mm)	x	x	x		
Bio20	Annual mean radiation (W m-2)				x	
Bio21	Highest weekly radiation (W m-2)				x	
Bio22	Lowest weekly radiation (W m-2)				x	
Bio23	Radiation seasonality (C of V)				x	
Bio24	Radiation of wettest quarter (W m-2)			x	x	
Bio25	Radiation of driest quarter (W m-2)			x	x	
Bio26	Radiation of warmest quarter (W m-2)	x	x		x	
Bio27	Radiation of coldest quarter (W m-2)	x	x		x	
Bio28	Annual mean moisture index			x		x
Bio29	Highest weekly moisture index			x		x
Bio30	Lowest weekly moisture index			x		x
Bio31	Moisture index seasonality (C of V)			x		x
Bio32	Mean moisture index of wettest quarter			x		x
Bio33	Mean moisture index of driest quarter			x		x
Bio34	Mean moisture index of warmest quarter	x	x	x		x
Bio35	Mean moisture index of coldest quarter	x	x	x		x
Bio36	First principal component of the first 35 Bioclim variables	x	x	x	x	x
Bio37	Second principal component of the first 35 Bioclim variables	x	x	x	x	x
Bio38	Third principal component of the first 35 Bioclim variables	x	x	x	x	x
Bio39	Fourth principal component of the first 35 Bioclim variables	x	x	x	x	x
Bio40	Fifth principal component of the first 35 Bioclim variables	x	x	x	x	x

Environmental variables are divided into three groups: Variables 1 to 19 are known as the core variables because they are calculated only based on temperature and precipitation data (**Table 1**). Variables 20 to 27 provide data on solar radiation, and 28 to 35 are calculated from data on soil moisture (e.g. atmospheric moisture content, Hutchinson et al. 2009), and variables 36 to 40 correspond to the first five principal components of the 35 variables, this allows capturing more than 90% of the variance in the full dataset (Kriticos et al. 2014). Variables 36 to 40 are best suited for distribution models of species from which information on their ecology is scarce (Kriticos et al. 2014).

b. Land Surface Temperature (LST) dataset

Derived from satellite data, the high-resolution land surface temperature (henceforth LST) contains bioclimatic variables 1 to 7, 10 and 11 (**Table 2**) at a spatial resolution of 250m extracted from a temporal extent of 10 years (2000-2013) with a daily coverage and with a spatial extent of Europe,

these dataset can provide 4 maps per day using a newly proposed method to reconstruct the LST time series at the continental scale (Metz et al. 2014). And, as of now, Metz and colleagues are working to reconstruct the other 10 variables missing from the Bioclim 'core variables set'.

Table 2. Climatic variables contained in the LST dataset. Table extracted from Metz et al. 2014.

Variable code	Description
BIO1	Annual mean temperature (°C . 10)
BIO2	Mean diurnal range (monthly mean of diurnal range; °C . 10)
BIO3	Isothermality ((BIO2/BIO7) . 100)
BIO4	Temperature seasonality (standard deviation . 100)
BIO5	Maximum temperature of the warmest month (°C . 10)
BIO6	Minimum temperature of the coldest month (°C . 10)
BIO7	Temperature annual range (BIO5–BIO6) (°C . 10)
BIO10	Mean temperature of the warmest quarter (°C . 10)
BIO11	Mean temperature of the coldest quarter (°C . 10)
Monthly mean	Monthly mean temperature (°C . 10)

c. Land use and landcover data (LULC)

Land use and land cover data refers to the information on the type of cover (e.g. forest, wetland, grassland) or use (agriculture, urban, mining) given to and specific land unit. It is derived from satellite imagery and is commonly presented in a raster or grid format using a set of categories used to classify the land.

The large number of ecosystem types and land uses, coupled with the constant changes in the human societies imply a rapid rate of change in the LULC products, which, as a consequence, are commonly developed at local scales, but there are a number of big projects in which the maps are developed, at coarser resolutions in time and space, at the continental or global scales.

Land Cover map (CCI-LC)

The Climate Change Initiative (CCI) of the European Space Agency (ESA) generated Landcover maps for three series of time (1998-2002, 2003-2007 and 2008-2012) using a multi-sensor strategy (MERIS and SPOT-Vegetation time series). The maps have a temporal resolution of 5 years and a spatial resolution of 300m.

CORINE Landcover

The CORINE (Coordination of information on the environment) programme was created in 1985 to compile information on the environment status (e.g. land uses, state of natural areas, distribution and abundance of wild areas and water resources and assessment of hazards) across the European Union (EEA, 1995).

Table 3. Characteristics of CORINE landcover products

	CLC 1990	CLC 2000	CLC 2006	CLC 2012
Satellite data	Landsat-5 MSS/TM	Landsat-7 ETM	SPOT-4/5 and IRS P6 LISS III	IRS P6 LISS III and RapidEye
Time consistency	1986-1998	2000 +/- 1 year	2006 +/- 1 year	2011-2012
Geometric accuracy, satellite data	< 50 m	< 25 m		
Minimum mapping unit/width (MMU)	25 ha / 100 m			
Geometric accuracy	100 m	better than 100 m		
Thematic accuracy	> 85% (probably not achieved)	> 85% (achieved)	> 85% (not checked)	> 85%
Change mapping (CLCC)	not implemented	boundary displacement min. 100 m		
		for existing polygons >5 ha; for isolated changes > 25 ha	all changes > 5 ha are to be mapped	
Thematic accuracy of CLCC	-	not checked	> 85%	
Production time	10 years	4 years	3 years	2 years
Documentation	incomplete metadata	standard metadata		
Access to the data	unclear dissemination policy	dissemination policy agreed from the start	free access for all users	
Number of countries involved	26	30	38	39

After the first inventory of 1990 and since then there has been three updates (2000, 2006 and 2012). The number of countries involved in the project has increase (from 26 to 39 countries, see **Table 3**) as well as the accuracy, while the production time has been reduced (see **Table 3**).

The products include 44 cover classes that span include artificial surfaces (e.g. fabrics, urban areas, mines, construction sites), agricultural areas (e.g. arable land, permanent crops, pastures), forest and semi-natural areas (e.g. forests, shrubs, grasslands), spaces with scarce or no vegetation (e.g. beaches, dunes, glaciers), inland and maritime wetlands (e.g. marshes, peat bogs) and marine and inland water bodies (EEA, 1995).

GlobCover Land Cover

Developed by the European Space Agency in collaboration with a network of partners (EEA, FAO, GOCF-GOLD, IGBP, JRC and UNEP) uses the MERIS satellite imagery (January 2005 to June 2006) to create a land cover map of the world for 2005. The GlobCover products include: bimonthly MERIS full resolution mosaics, annual MERIS full resolution mosaic, Globcover land cover map (December 2004-June 2006), and regional GlobCover land cover maps (Bicheron et al. 2008). In 2010 a second version (GlobCover 2009) was released (Arino et al. 2009). It classifies the land cover in 22 classes that include artificial surfaces, forests, grasslands and croplands (Bicheron et al. 2008).

Globeland30

This global land cover map launched by China in 2010, consist of a 30 meters spatial resolution product with 10 landcover types (e.g. cultivated area, forest, grassland, etc) and for the years 2000 and 2010 within a four years period. It was developed using multispectral images including the TM5 and ETM + of America Land Resources Satellite (Landsat) and the multispectral images of China Environmental Disaster Alleviation Satellite (HJ-1) (Chen et al. 2011).

d. Land use changes and trends

Land use patterns across the landscape are constantly changing. Due to its pervasive effect on the distribution patterns exhibited by biodiversity it is a relevant factor to include in the estimation of biodiversity current distribution and future trends. There are a number of recent projects developed to trace the changes of land use trough space and time.

The CAPRI model

The CAPRI (Common Agricultural Policy Regionalized Impact analysis) model is a “multi-purpose” economic model that consists of a combination of databases from the Farm Accounting Data Network (FADN) and modelling methods, implemented a software that includes about 50 animal and crop-based products in the European Union (Britz & Witzke 2012).

e. Hierarchical structure of habitat classification scheme

Most LULC (e.g., CORINE) or habitat classification schemes (e.g., EUNIS) have a hierarchical, tree-like structure, in which fine categories are nested within more general categories. This hierarchical structure plays an important role in rule-based (knowledge based) classification, in which experts first provide rules that separate general categories from one another, and then move down the hierarchy while providing more specific rules for finer categories. The main advantage of rule-based classifications is in the more process-based understanding they provide and in the ability to apply the model without ground truth data. However, rule-based classifications have several shortcomings. First, expert knowledge is not always available and the labour time of experts is expensive. Second, fine tuning rules for a system are time consuming. Third, rule based classification is static since the rules are specific to the available input-data layers, and cannot be automatically updated when new input data becomes available, or when old data becomes unavailable.

Alternatively, in recent years there is a growing usage of machine-learning algorithms (e.g., Support Vector Machine, Random-Forest) for habitat classification. These methods offer a cheap and dynamic modelling framework, with similar accuracies as knowledge-based classification. However, most machine learning classification methods follow a flat classification approach (sensu Silla & Freitas 2011), in which all terminal nodes are classified simultaneously in a ‘one-against-all’ approach. In addition, the machine learning algorithms are ‘black-boxes’ and unlike rule-based classification are not based on understanding underlining processes that govern the distribution of different categories.

As part of deliverable D3.1 of WP3 (task 3.1), we developed a novel machine-learning based hierarchical classification method, which accounts for the pre-defined hierarchical structure of the habitat classification scheme, while avoiding the need of expert knowledge. The method is based on applying randomForest (Breiman 2001) as the local classifier at each parent node along a pre-defined class hierarchy. The hierarchical randomForest (HRF) tool was codified as an R package (‘HieRanFor’), with clear help files and examples. A tutorial can be found in deliverable D3.1 and fully working version of the package will soon to be submitted to CRAN. First test of the HRF model have shown that it provides higher accuracy in the validation test than a flat classification approach. It also provides a more detailed understanding of the relative importance of different variables in different branches of the hierarchical classification scheme. The HRF method can create fine-scale

LULC or habitat layers that may be plugged as input to SDMs, using either global (e.g., EUNIS) or species-specific classification schemes. Potentially, HRF can also be used as a SDM algorithm, if the distribution of a given species is modelled according to a hierarchical structure (e.g, first level separation of suitable vs. unsuitable habitats, and second level separation to occupied and unoccupied).

f. Other environmental variables

Abiotic properties such as geology, topography, among others, also change across the landscape at multiple spatial scales. These variables are characterized by its slow rate of change and where only abrupt events like anthropic driven land-use or catastrophic phenomena (i.e.volcano, floods, etc) can modify them in the short-term. A very interesting dataset for European soil information is provided by the European Soil Database (2004).

Table 4. Environmental data derived from remote sensing.

Resource	Resolution	Area	Timeframe	Periodicity	Description	Source
ASTER Global Emissivity Database (GED)	100m or 1km	World	2000 to 2008	Mean	Emissivity, mean land surface temperature, NDVI	NASA
AMSR-E/Aqua Daily L3 surface soil moisture	25km	World	2002 to 2011	Daily	Soil moisture/Water content	NSIDC
ECV soil moisture dataset	0.25 degrees	World	1979 to 2010	Daily	Soil moisture based on six sensors	ESA
ERS/MetOp Soil Moisture	25-50km	World	1991 to present	Monthly	Coarse resolution soil moisture data	Tuwien
ASCAT coastal winds	12.5km	World	2011	N/A	Swath grid -Winds	EUMETSAT
DLR-AVHRR	~1km	Europe	2000 to present	Yearly	Various atmospheric variables (temperature, air)	NOAA/AVHRR
NOAA/AVHRR	1.1km	World	1978 to present	Daily	Advanced Very High Resolution Radiometer	NOAA
NASA MEaSUREs	30, 60 and 250m	World	1981 to 2010	Daily	Vegetation phenology and indexes – NDVI	NASA
MODIS Land use	300m	World	2001 to 2012	Yearly	Land cover classification systems	NASA
Landsat Vegetation Continuous Fields	30m	World	2000	N/A	Tree cover layers	University of Maryland
Harmonized World Soil Database	1 km	World	1971-1981	N/A	Soil map of the world – ground observations	FAO/IIASA

2.2. Biotic interactions in species distribution models

The nature of biotic interactions is complex, it can be facilitative (e.g. mutualism) or antagonistic (e.g. predation), direct (predator-prey) or indirect (plant-herbivore-predator), and can also vary in strength depending on the spatial context (i.e. environment, species structure, scale) and characteristics (i.e. traits) of the species involved in the interaction (Morales-Castilla et al. 2015). For example, at local spatial scales, the range of a predator can modify the range of the prey, while at large scales the behaviour of animal seed-dispersers can determine the limits of the distribution of the plant they disperse.

Most of the studies have shown how biotic interactions can affect the response of species to the environment and determine its presence/absence at local scales (Bascompte, 2009), but there is also evidence of their importance in shaping the distribution of species at large scales (Brooker & Callaghan 1998, Tylianakis et al. 2008, Hanspach et al. 2014). Some studies have even demonstrated that climate change affects species interactions which in turn can determine whether a species can persist or adapt to the future climatic conditions (Gilman et al. 2010, Van der Putten et al. 2010). Consequently, it is paramount to identify the type and strength of the biotic interactions in order to accurately model the current and potential distribution of species.

Most species distribution models rely on the assumption that biotic interactions do not influence species spatial patterns or only influence them at small spatial scales, and consequently they are built using only environmental predictors (Zimmermann et al. 2010). However, integrative approaches have been developed to include biotic interactions (e.g. Schweiger et al. 2012), which has resulted in improved models demonstrating the importance of including biotic interactions in SDMs (Kissling et al. 2012, Wisz et al. 2013).

One key ingredient in the application of such integrative approaches is reliable data on the ecology of the species, as well as the co-occurrence of species, with high resolution and at large extents. This is an ongoing effort led by a number of plot monitoring schemes that have collected spatial and temporal data that can be used to infer the characteristics and strength of the interactions among species.

2.3. Species distribution models developed by the partners

2.3.1. Catchment-scale freshwater species distribution models

Partner: Senckenberg Gesellschaft für Naturforschung

a. Description

In this application, species distribution models (SDMs) are used to predict and analyse distributions of freshwater organisms at the catchment scale. Current efforts to improve the application of SDMs to freshwater ecosystems are focused on the Kinzig River catchment of central Germany, encompassed by the Rhein-Main Observatory (RMO), an EU BON test site and a long term ecological research (LTER) site. Although the environmental predictors consider the whole Kinzig catchment, distribution predictions are projected on the stream network only, which represents the freshwater ecosystem (Domisch et al., 2013). The stream network consists of 28, 205 grid cells at a spatial resolution of 25m.

SDMs are built using the R package ‘biomod2’ (Thuiller, Georges, & Engler, 2013) as ensemble models, which provides robust predictions (Araújo & New, 2007). Each model ensemble includes five algorithms (GLM, GBM, CTA, ANN, MaxEnt), three pseudo-absences (PA) runs and ten repetitions for each combination of algorithm and PA run. PAs are necessary because of the presence-only nature of most data used in SDMs. 150 models are built per taxon, but only those surpassing a certain model performance indicator threshold, are taken into account in the ensemble model (in this case TSS = 0.6). To calculate the performance indicators, occurrence data is randomly split in each repetition into a 70% calibration and a 30% testing subset.

Models are evaluated in terms of their performance (TSS and ROC/AUC). The predictor importance for each species determined. Predicted probability of occurrence, predicted binary occurrence and mapped uncertainty are produced for each ensemble model.

b. Implementation

- Species data

Data from the monitoring scheme in the RMO-LTER are used to build the SDMs, providing precise information on the sampling: frequency, exact location, a uniform method and standardized taxonomic identification. Occurrence data is supplemented with sampling carried out by the local authorities. Focal organism groups are stream macroinvertebrates and fish.

- Environmental data

Models incorporate climatic, topographic, hydrological, land use and geological predictors. These environmental predictors are either specific to freshwater ecosystems (e.g. hydrology) or modified to fit the hierarchical structure of river networks (e.g. sub-catchment specific approach (Kuemmerlen et al. 2014)).

- Test site

SDMs are built for the Rhine-Main-Observatory (RMO) test site located in central Germany. The RMO, a long term ecological research site (LTER), encompasses the entire Kinzig River catchment (ca. 1 100km²), a tributary of the Main River, close to its confluence with the Rhein River.

c. Outcomes

First predictions have been made for the current distribution of stream macroinvertebrates, including 175 taxa. Currently, the main focus is on identifying and analysing bias sources within the dataset used. These results will be described in a manuscript that is in preparation.

d. Challenges

Bias arising from sampling is a major problem concerning the occurrence data (Kuemmerlen et al. *in press*). This bias has two main sources: (a) spatially uneven sampling results in the overrepresentation of some areas (e.g. because of easy access, or monitoring targeted at specific goals), while others are underrepresented (e.g. headwaters in freshwater ecosystems); (b) different sampling frequencies across the study area (if data from different sources is used) will lead to varying detection probability of species (most SDM approaches assume homogeneous detection probability at all sites and detectability across species).

The greatest challenge is to fully adapt species distribution predictions into freshwater habitats by including the hierarchical structure of stream networks, as it improves predictions (Domisch et al. 2013). However, large scale approaches still model freshwater biota on a continuous landscape, analogous to terrestrial biota. For catchments small to moderate in size (< 2000 km²), this has been achieved already (Kuemmerlen et al. 2014), but for larger catchments (>2000 km²) this is still a challenge as environmental predictors specific to freshwater ecosystems are largely lacking. Two main issues stem from the physical properties in streams: the hydrology and the temperature. Hydrology, identified frequently as being among the most important predictors, is a complex environmental variable to obtain, as it is obtained from hydrological models which are rare even in Europe. Temperature, as currently SDM approaches rely on atmospheric temperature, rather than on water temperature. More freshwater-relevant predictors that would undoubtedly be of interest include dissolved oxygen and pollutants, however these are significantly more difficult to achieve as measurements are scarce. Making such predictors available for freshwater models and extensive areas would significantly improve model quantity and quality, however producing them is a challenge in itself.

e. Future developments

The next steps will include modelling the current distribution of fish at the catchments scale for the RMO-LTER, under the same conditions and using the same method as from the stream macroinvertebrates. Further efforts include projecting distributions into the future to determine possible changes in taxonomic richness, range size, elevation and species turnover.

2.3.2. The SpaNiche hybrid model

Partner: University of Leeds

a. Description

Species distribution models (SDMs) are commonly applied to infer species distributions to areas that are unsampled based upon their niche space (correlations with environmental predictors). However, there remain several issues that are unresolved in their application.

First, the probability of occurrence needs to be converted to a binary presence-absence map through application of a threshold without information of what the total number of occupancies should be. The choice of criteria for the threshold is problematic, and without current consensus (e.g. Liu et al. 2005), and they may be currently unsuitable to assess trends in status (Guillera-Arroita et al., 2015). Furthermore, a single threshold is applied universally across the SDM. In reality, a cell of high suitability is more likely to be occupied if surrounded by low suitability cells. A single-threshold approach will assign both cells as the same status.

Second, modelling only takes into account the correlation of the species distribution with regards to environmental predictors, but ignores the spatial aspects important for structuring a species distribution. These include the scale of spatial autocorrelation due to dispersal abilities, where areas are of high environmental suitability but remain uncolonised due to dispersal barriers, and often unknown historical factors. Thus species distribution models largely predict the potential, rather than the realised distribution.

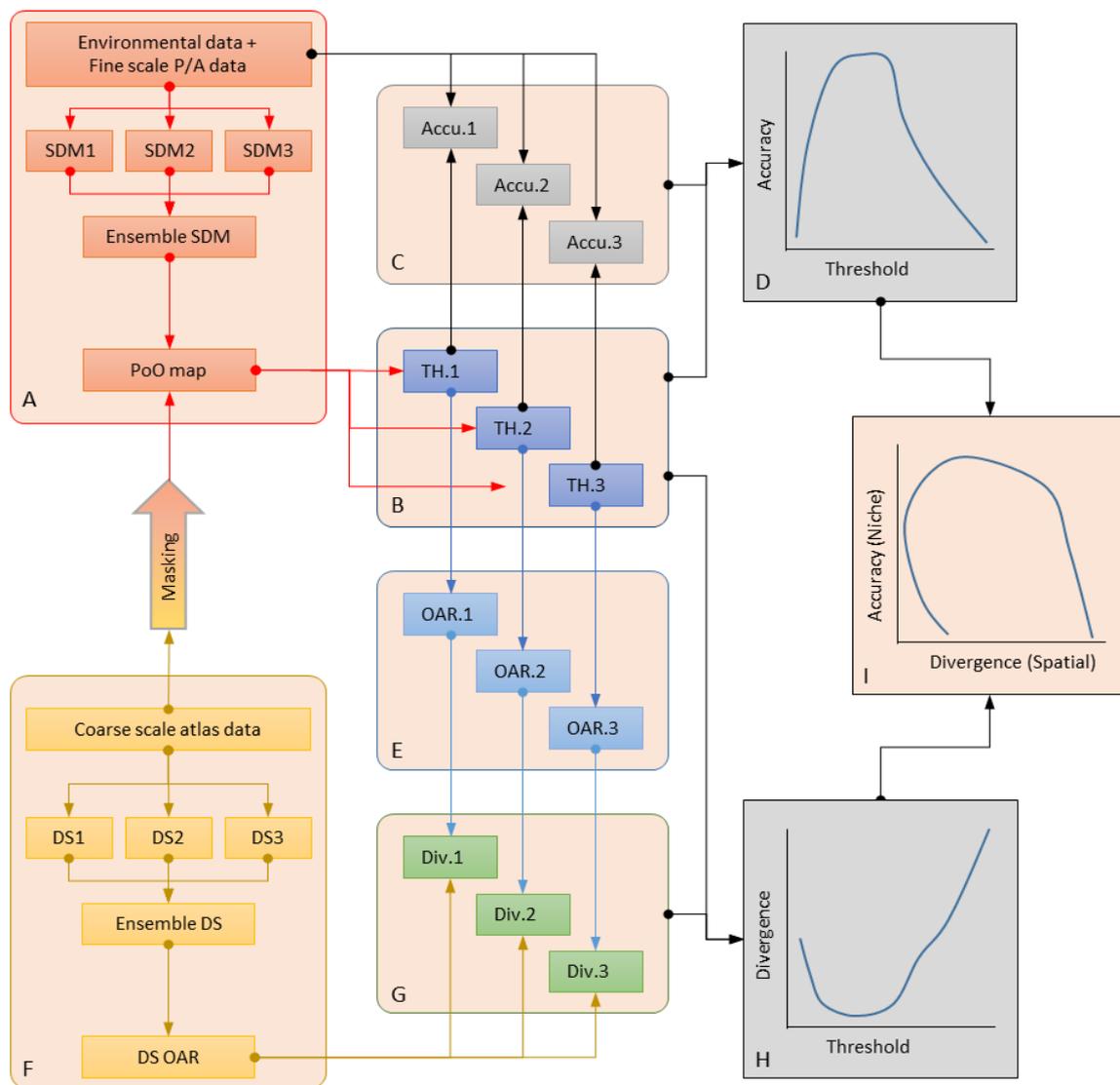


Figure 2. Flow diagram of the SpaFigure 2. Niche model. For the niche modelling, species distribution models are generated through ensemble modelling (A) and thresholds applied to generate presence-absence maps (B). Accuracy is measured for each threshold (C) and the accuracy-threshold curve plotted (D; niche consistency). For the spatial modelling, atlas data at a coarse-grain size is downsampled to create an occupancy-area relationship curve (OAR; F). Each map from B is upsampled to create a threshold-OAR (E), and the divergence measured from the downsampled OAR (G), and the divergence-threshold curve plotted (H; spatial consistency). Finally, the balance between niche- and spatial-consistency is explored by plotting accuracy against divergence (I) to determine the optimal threshold.

On the other hand there are spatial models. At the grain size of the SDM, large areas of the landscape are unsampled and so probability of detection is 0 in these locations. If data is presence-only, we also only have a probability of detection of 1 where a record occurs, but nowhere do we have a certainty of absence. As the grain size is coarsened the probability of detection (and therefore the certainty of our absences) increases. At the atlas scale the probability should approach 1, and therefore we can have high certainty that our presences and absences are correct, although at a grain size much coarser than an SDM. We can downscale the occupancy at the coarse atlas scale to predict the proportion or area of occupancy (AOO) at the finer grain size of the SDM (Barwell et al., 2014). However, such models ignore environmental aspects and so do not reveal the location of those occupancies.

The SpaNiche model, under development in WP3 (Task3.3), attempts to combine these two contrasting approaches in to a hybrid approach. Ultimately, it aims to select a threshold, or range of scale- or cell-specific thresholds, for the SDM that is accurate at the fine grain size the SDM is carried out, but remains spatially-consistent with the spatial patterns at the coarse-grain size.

The SpaNiche model achieves this in two ways (**Fig. 2**). First, if we assume a probability of detection of 1 for all cells at the atlas scale, then we can use the absences at this scale as a mask for the SDM. Therefore, all those areas that are atlas-scale absences are converted to a probability of occurrence of 0 in the SDM (arrow from Fig. 2F to Fig. 2A).

We then create an accuracy-threshold curve for the SDM using a measure of accuracy (e.g. kappa statistic) for the presence-absence maps created after application of all possible thresholds (Fig. 2A-D). This identifies the threshold that is most consistent with the fine scale occurrence (here we refer to this as niche consistency).

Second, we create a distribution at a coarse grain size where probability of detection is high to create an atlas map. This can be an independent data set, or generated from the same data as for the SDM. We then downscale occupancy to the grain size of the SDM to create an occupancy-area relationship curve (OAR) (Fig. 2F). For each of the presence-absence maps created from the SDM thresholding (Fig. 2B) we upgrain the map, therefore creating a threshold-OAR for each threshold (Fig. 2E). We measure the divergence of each threshold-OAR from the original downscaled-OAR, creating a divergence-threshold curve (Fig. 2H). This identifies the threshold that is most consistent with the spatial scaling of occupancy (here we refer to this as spatial consistency).

Finally, we can plot accuracy (niche consistency) against divergence (spatial consistency) in order to identify the threshold that best balances both aspects (Fig. 2I).

b. Implementation

As the model is still under development model-specifics are likely to be altered and adapted from its current state, so plans are still preliminary.

Occupancy downscaling of the atlas data to generate the downscaled-OAR will be carried out using the 'ensemble.downscale' function in the 'downscale' package developed in WP3.2. and awaiting submission to CRAN. Creation of the threshold-OARs will be carried out with the 'upgrain' function in the same package. SDMs will be generated as ensemble models using the R package 'biomod2' (Thuiller, Georges & Engler 2013) in line with other studies in the work package.

As model effectiveness has yet to be tested we will focus on running the model on a set of virtual species which will allow us to test accuracy against the known distributions at fine and coarse scales. These species are currently being generated using the R package 'virtualspecies' (Leroy et al. 2014) for testing with multiple other models (see Section 3: Directions), and so will allow us to directly compare model accuracy against a range of established SDMs under various conditions.

If successful, we can then apply the model to real datasets for which there are, or we can generate, both atlas data and fine-scale presence data, such as the Wallonia birds data set

c. Outcomes

The SpanNiche model is currently undergoing development in WP3.3 in preparation for deliverable D3.2 (due date March 2016). The majority of the model has been finalised and coded in R, however, there still remains some important components to be developed and we expect it's final state to differ in some respects to that described.

2.3.3. Modelling DNA fungi data

Partner: Helmholtz Centre for Environmental Research - UFZ

a. Description

In order to study the relationship between species occurrences and climate and environmental variables and to predict species geographic distribution, certain statistical models and methods bundled under the term species distribution modelling (SDM) are well-established (Franklin, 2009). Instead of species, we use DNA data providing species concepts, provided by TARTU (Urmas Kõljag).

While the locations where the species was found can clearly be interpreted as “true” presences, the rest of locations can hardly be specified as “true” absences. As Hirzel et al. (2002) explained absence data are often difficult to obtain accurately because in many cases the species cannot be detected even though it is present. Moreover, all species seem to be extremely rare species, i.e. the number of pretended absences is much higher than that of presences. Therefore, the first step in the analysis must be a pseudo-absence selection taking the environmental conditions for the species into account. This is done by calculating a surface range envelop. Absence data only outside of this envelop, so-called pseudo-absence data, are incorporated into the analysis (Thuiller, Georges, & Engler, 2013). The main part of the analysis is a multivariate logistic regression. For this purpose, we use a Generalized Linear Model (GLM) for binary response data (McCullagh & Nelder, 1989). Because a stepwise predictor reduction by means of the Akaike’s Information Criterion (AIC) did not lead to better results, we incorporate the full set of environmental variables in the final analysis.

For evaluation of the predictive performance of the model, it is common to split the dataset into 2 parts: data for training and data for testing. According to the high number of predictors, we use the ratio 3:1, i.e. the models are built on 75% of the data (training) and validated on 25% (test). The selection is carried out randomly and repeated 50 times. Four criteria are used for assessing model quality: Kappa, MaxKappa, MaxTSS, and AUC. Kappa as a threshold-dependent measure of accuracy is given here for threshold 0.5. MaxKappa is the maximum value of Kappa for all possible thresholds. MaxTSS is the maximum value of TSS for all possible thresholds. AUC is threshold-independent.

b. Implementation

- “species” data

Instead of species, we use global extent DNA data on fungi relating to species concepts, provided by TARTU (Urmas Kõljag). We selected as examples the most frequent “species”, i.e. those species, which occur at the highest number of different locations. These are: "SH044209.06FU" (frequency=93), "SH052254.06FU" (84), and "SH044721.06FU" (82). All the following analyses are carried out separately for these species.

- *Environmental data*

As environmental information, a set of CliMond variables was used.

- *Site*

The extension of the study was global, but in later runs was restricted to Europe to increase model performance. Resolution was point resolution scaled to the CliMond grid.

c. Outcome

Distribution models for the three modelled species hypothesis were derived (see **Figs. 3-5**). Model quality is fair but can hopefully be improved by adding additional environmental layers, such as land cover, in future steps.

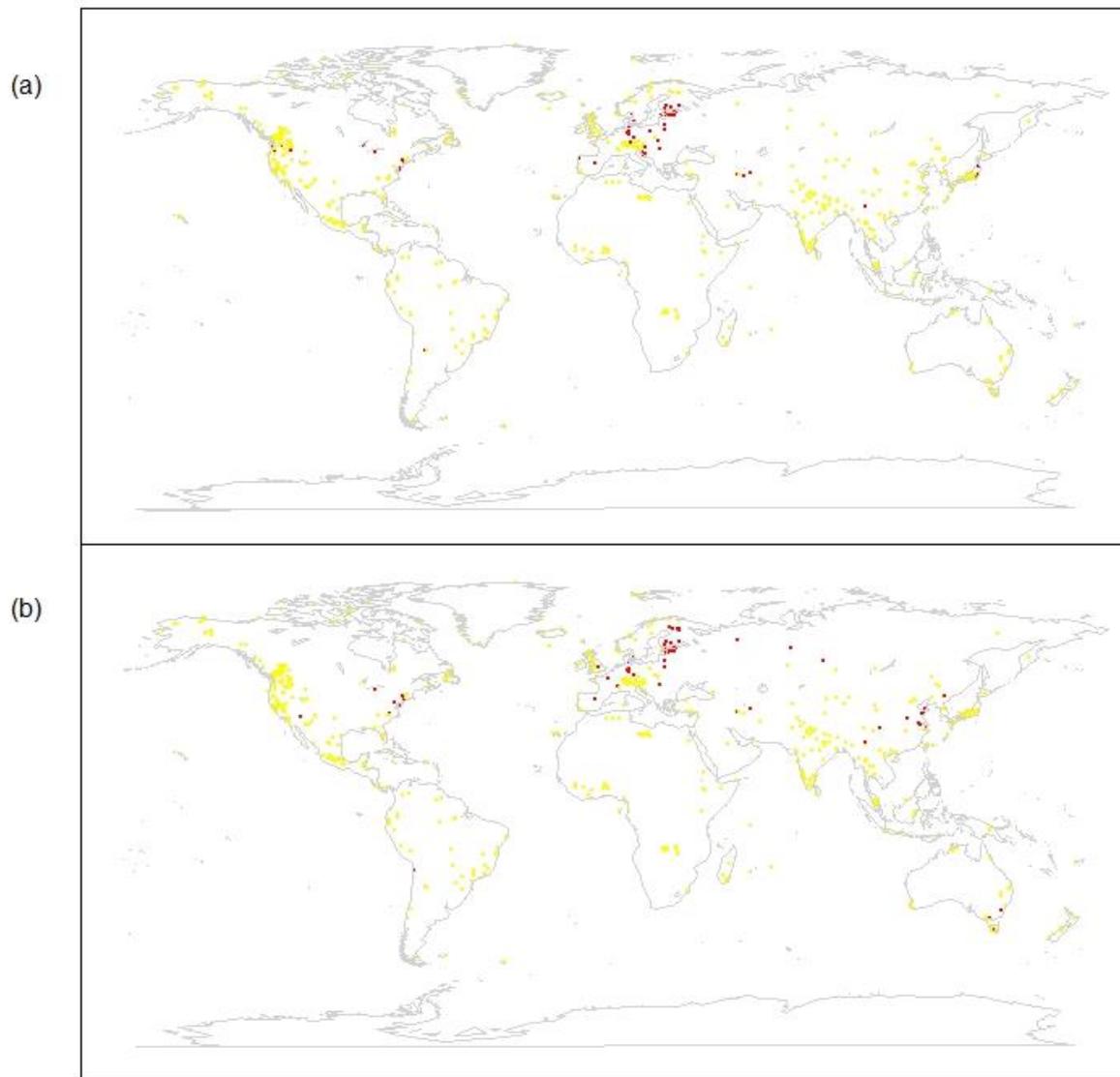


Figure 3. Distribution data for "SH044209.06FU". Data for (a) true presences (red) and estimated pseudo-absences (yellow), (b) fitted values of presences (red) and absences (yellow).

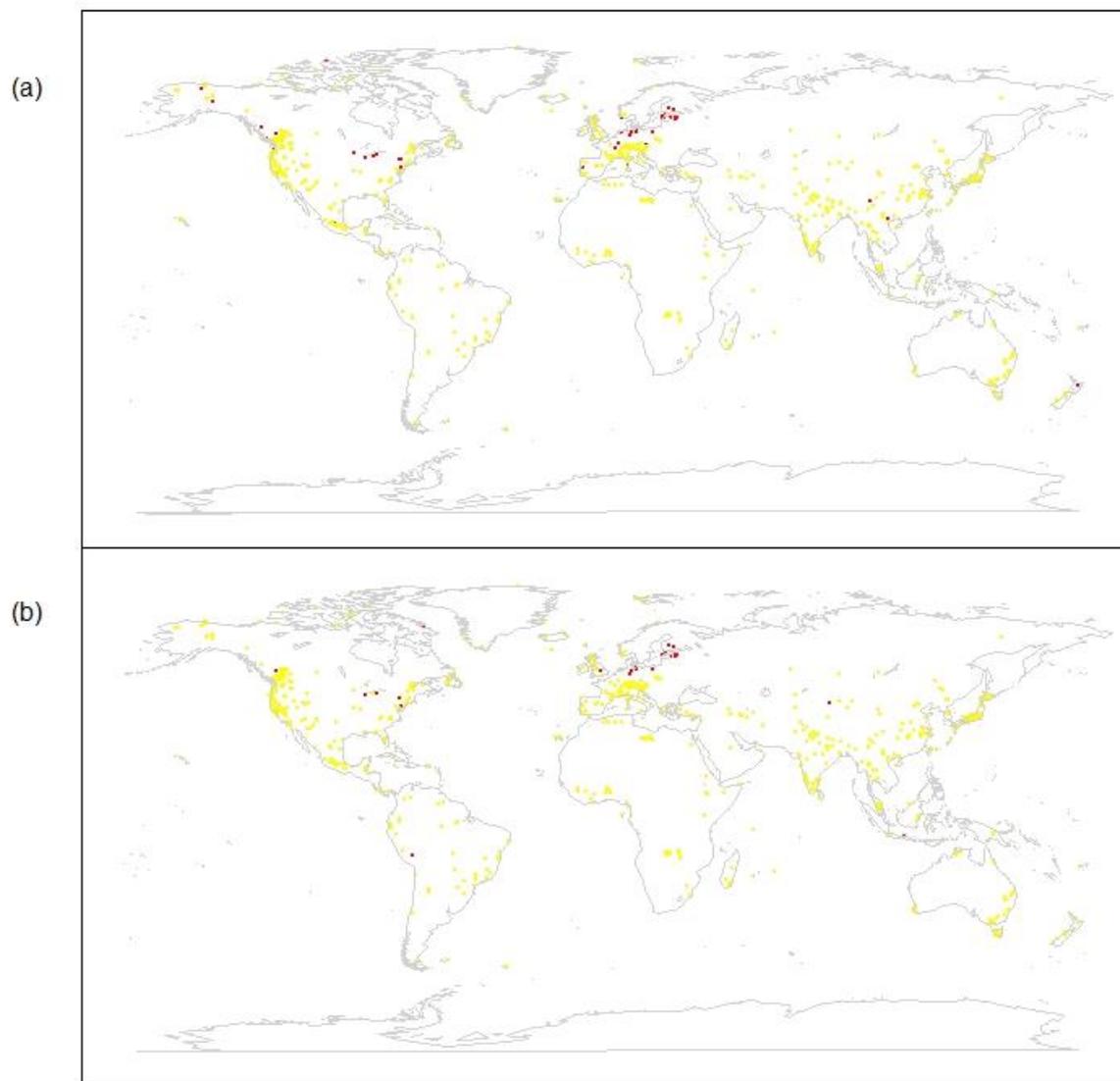


Figure 4. Distribution data for "SH052254.06FU". Data for (a) true presences (red) and estimated pseudo-absences (yellow), (b) fitted values of presences (red) and absences (yellow).

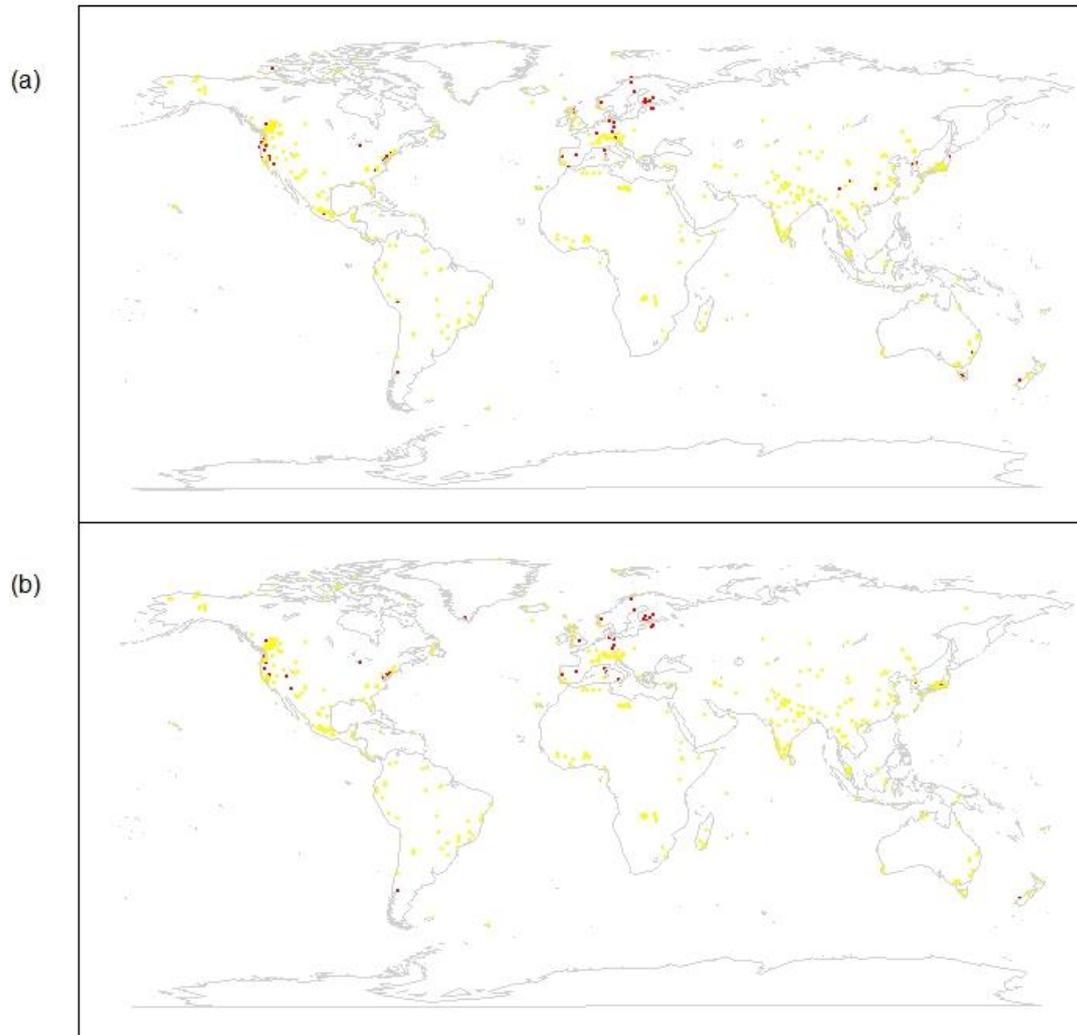


Figure 5. Distribution data for "SH044721.06FU". Data for (a) true presences (red) and estimated pseudo-absences (yellow), (b) fitted values of presences (red) and absences (yellow).

2.3.4. Effect of data sources (environment and species) on SDM approaches

Partner: Fondazione Edmund Mach

a. Description

Species distribution models rely on the available data on species occurrences and environmental measurements. Data on species occurrences is often extracted from the global biodiversity information facility (GBIF), one of the largest sources of data on species occurrences that combines data from multiple sources (museum, field observations, monitoring schemes) into one large global database (Beck et al. 2012). Such a large coverage approach is found in the environmental proxies used to describe the climatic conditions, and that is the case of the Bioclim data, a global interpolation of weather station data used to generate raster layers of temperature and precipitation among other climatic variables, commonly used in the development of bioclimatic models. Consequently, while species occurrences from the GBIF database include biases due to the variation in sampling effort and number of studies per area, environmental data varies spatially depending on the uneven distribution of weather stations in the world.

In this application, the effect of data sources is evaluated for a set of species distribution modelling methodologies, using a multiscale approach. We are extracting data from GBIF for a selected set of plant species (provided by Neil Brummit, NHM) with varying distribution (e.g. clustered, sparse, rare, common, etc). Moreover, to have a known distribution from which to sample occurrences, we have generated virtual species, using the R package 'virtualspecies' (Leroy et al. 2014). The species distribution models are built using the R packages 'biomod2' (Thuiller, Georges, & Engler 2013) and 'ENiRG' (Canovas et al. 2014).

b. Implementation

- Species data

Species data on a set of plant species is extracted from the GBIF facility using the R package 'rgbif' (Chamberlain et al. 2015). The data is processed to quantify the number of sources and sampling approaches used and to filter out occurrences with spatial issues.

- Environmental data

Two climatic dataset are used: the commonly used Bioclim dataset and the recently developed LST dataset. The environmental variables extracted from both dataset remain the same, and the LST dataset is applied using its resolution and a coarser resolution that matches the resolution of the bioclim dataset.

- Site

The tests are performed at multiple scales, starting with Europe as the larger scale and selecting a number of smaller extents.

c. Outcomes

Initial tests have been developed using a combination of data of three tree species, bioclim and LST data, at continental scale. The same combination has been applied in the case of a virtual species. The models have been built in the 'biomod2' (Thuiller, Georges, & Engler, 2013) and 'ENiRG' (Canovas et al. 2014) packages. The work has been focused on the selection of plant species relevant for our study, the development of virtual species, and the analysis of the variation in the models performance (AUC and TSS).

d. Future developments

The next steps will include the development of a Bayesian approach to explicitly show the spatial distribution of uncertainty, strictly linked to Task 4.5, in the species distribution modelling output. This will be done by relying on Markov Chains Bayesian models based on climatic priors

2.3.5. Within-species spatial niche variation and projections of species distribution under future climate change

Partner: European Bird Census Council / Centre Tecnològic Forestal de Catalunya (EBCC/CTFC)

a. Description

Species distribution modelling often ignores that the niche of a species may vary within the geographical space it occupies (Pearman et al. 2008). Non-stationarity in the response of a species to environmental conditions may induce low predictive performance at a local scale and uncertain spatial inference across the studied area (Hawkins, 2012). Surprisingly, there is a lack of modelling approaches that deal with this issue and that are suited to use large samples of species to predict species distributions under future changing conditions (Osborne et al. 2007). Accordingly, it is largely unknown to which extent predictions of future species distributions may be altered when accounting for non-stationarity.

In this application, we used a large sample of species and we compared the predictions of species distribution under future climate change according to (1) a global modelling approach that assumes a unique response of the species to climate conditions across Europe, and (2) a local modelling approach that deals explicitly with within-species, spatial niche variation. The local approach is based on partitioning procedures to split the entire distribution of the species in ecologically relevant subsets. Local models are built at the level of each subset and assembled with each other to capture within-species spatial niche variations across Europe.

b. Implementation

- Species data

We used the dataset from the recent Distribution Atlas of Butterflies in Europe (Kudrna et al., 2011). We randomly selected a representative sample of butterfly species (N=120) to examine the effect of species range characteristics on the modelling outcomes as we expected that widely distributed species are more prone to show a non-stationary response to the environmental conditions than species associated with a geographically restricted range.

- Environmental data

Climate data were obtained from the Climatic Research Unit (CRU) climatological database with 10' resolution (Mitchell et al., 2004). We derived the three most commonly used climate variables that reflected the main constraints on butterfly growth and survival (Heikkinen et al., 2010). Current climate data covered the period 1971-2000. Future climate data simulated climate conditions during the periods 1991-2020, 2021-2050 and 2051-2080 under a set of scenarios reflecting alternative plausible futures in Europe.

- Site

Grid-based distribution data after 1980 were extracted for all butterfly species across Europe using the ETRS89-LAEA reference system (Lambert Azimuthal Equal Area) at 50-km resolution. We rescaled the climate variables to the resolution of the butterfly species distribution data.

c. Outcomes

Based on an innovative modelling framework applied to a representative sample of species at large spatial scale, we provide novel insights into (1) the potential consequences of within-species spatial niche variation when predicting the future of biodiversity under climate change and (2) the importance of dealing explicitly with this issue when using modelling tools in conservation biogeography and global change research. Future research efforts should achieve a better balance between the development of local and global modelling approaches to better evaluate the level of uncertainty due to within-species spatial niche variation in global change impact assessments.

2.3.6. Use Case of Small Islands with high mountains.

Partner: Royal Museum for Central Africa

Based and translated from 'Louette, M., Abderemane, H., Yahaya, I. & Meirte, D. 2008. *Atlas des oiseaux nicheurs de la Grande Comore, de Mohéli et d 'Anjouan. Series 'Studies in Afrotropical Zoology', 294. Tervuren: RMCA. 240 p'* funded in part by the Belgian Cooperation. This Atlas is the result of a cooperative project between the RMCA and several collaborating partners, notably the Convention for Biological Diversity-Comoros and the National Museum of the Comoros.

Translated, compiled and further discussed by Patricia Mergen

a. Description

As reported in previous chapters of this report, BioClim/Worldclim data have a maximum resolution of about 30 arc seconds (about 1 km) for current data; going back into the past this resolution can drop to 1 or even 5 min arc or less. In the case of small islands downscaling is needed, because of the rapid topographic and environmental conditions variations as expressed thorough the literature (see e.g. Keener et al. 2012). Species Distribution modelling studies for terrestrial species are rarely undertaken on small islands because the available data is often biased and skewed towards more accessible areas only. For such a study on the island of Trinidad in the Caribbean, Maharaj and New (2013) warn about unevenly distributed location data and a too small number of occurrence points which induce poor performance of otherwise much used Species Distribution algorithms and associated software. For the environmental and climate data, they also refer to the issue that the too low resolution obtained by Global Climate models cannot make the difference between the conditions on the island and the surrounding oceans. The uneven topography of such islands has similar impacts on the modelling tools tested. However, if sufficient data is available and proper downscaling can be achieved using species modelling tools like Maxent, modelling indeed becomes possible, while still being challenging.

The use case outlined here is based on the findings of the Atlas of breeding birds in the three western Comoro islands: Grand Comoro, Moheli and Anjouan (Louette *et al*, 2008). Data collected by RMCA staff and their collaborators from 1981 to 2006 was used in this study. In this particular case, it was not possible to obtain more data, or to perform sufficient downscaling to use classical species modelling tools. It is therefore interesting to find working alternatives to produce a valid atlas and identify potential ornithological zones of interest with relatively low data availability.

b. Implementation

The intention of the study is to define the areas of importance for the conservation of birds (birds being good indicators for biodiversity in general), with advanced analysis (Louette et al, 1995). The published book is intended both for the decision makers, researchers, amateurs and educators interested in the avifauna of the region.

- Species data

The species data consist of visual and sound observations, made from 1981 to 2006. In terms of breeding birds 59 species were observed, 47 on Grand Comoro, 44 on Moheli and 39 on Anjouan. 15 endemic species and 51 endemic taxa were identified (some being endemic species, while others being endemic sub-species of non-endemic species). Only resident terrestrial birds are concerned, the marine birds along the shorelines were not studied for this purpose. The non-breeding birds were not considered in the analysis and in the Atlas, but were listed in a table for information.

Most observations were done along transects by standardized counting on predefined spots during 5 to 15 minutes. Additional data was mined from field reports and published literature. Taxonomic identifications were cross checked by specialists and doubtful records discarded from the analysis.

All the data was carefully encoded with standard metadata associated:

- a) Team (acronym)
- b) The source (internal code), indicating whether or not it was along a transect for bird counts (a path)
- c) Year, month and (if known) exact date of observation
- d) Which species
- e) Which island
- f) Latitude coordinates provided by the observer
- g) Longitude coordinates provided by the observer
- h) The system of coordinates provided by the observer
- i) The position, and if possible, the observation point, during the sequence of the journey
- j) The description of the observation point
- k) The altitude given by the observer
- l) Other details given by the observer
- m) Number of individuals
- n) Indications of nesting.

In total there are more than 20 000 observations from about 2400 observation points. The georeferencing of these points have been checked, even when GPS information was available. Different background maps have been used including both online available maps, satellite imagery, but also paper maps which were scanned and georeferenced.

- Environmental data

Grid maps were produced featuring the habitats where the birds have been observed. A method of ecological envelopes was used to assess the potential range both in distribution and altitude. Different algorithms and modelling tools like Maxent were tried out, but the available environmental datasets were not sufficient to have conclusive results. Finally, following parameters were measured: altitude, forested/non-forested areas, rainfall, rivers, lakes, and distances to villages or roads; in order to determine the extent of the areas where the birds could potentially breed. For the grids SRTM3 20×20 was used, taking into account only those on the islands or touching them. Ecological parameters were also taken from the SRTM 3 database or inferred from other maps and literature.

- Site

The Comoro islands are an oceanic archipelago of 4 units situated in the Mozambique Channel between 11° 23' and 13° S and 43°13' and 45°18' E more or less half way between the African continent and Madagascar. Additionally to the challenge of their small size these islands are also characterized by volcanic mountains, some still active. Grand Comoro (1148 km²) has two mountains: Mount Karthala, an active volcano rising to 2 361 m and La Grille, an extinct volcano, rising to 1 087 m. Interesting to know is that volcanic eruptions happened in 2000, 2003 and 2005 during the observation period concerned and which had a high short term impact on the vegetation and local environmental conditions. The two other islands studied for the atlas are smaller and have a very different profile with long extinct volcanos. Moheli with its 211 km² is hilly with a maximum of 790 m high, while Anjouan (424 km²) presents steep cliffs rising to about 1 595 m with a central basin in between. The fourth island, Mayotte, was not studied for this atlas.

c. Outcomes

The results showed that on all three islands the endemic birds' preferred habitats are situated at higher altitudes. On Grand Comoro (the island with large tracts of vegetation in a pristine state) these zones represent the main forest area, on Moheli (with a relatively low human population) intermediate altitudes also showed habitats suitable for endemic bird species, while Anjouan with only forest patches left in otherwise agricultural lands, showed that less suitable habitats can be used to a certain extent by endemic species.

The method consists of entering contiguous squares to the ecological envelope in an algorithmic calculation. The superposition of the grid maps for all species considered yields a map of ornithological interest. However, because the contribution of each species depends on the number of squares in which it is supposed to occur, a value is calculated for each. A total index of endemism is calculated, taking into account the number of species present on each island (Louette & Meirte, 2009).

d. Challenges and critical analysis

While the data had not been originally collected in view to compile an Atlas, the extent of visual and acoustically observed bird counts and the abiotic data collected, compared favourably with the type of data used in other atlases. However, a bias was detected in terms of sampling efforts between the different islands and also, to a lesser extent, between different areas within a same island. It was estimated that the efforts have been sufficient to detect all breeding species with a good confidence, but these biases could have had a minor influence on the calculated distribution ranges. The collecting range is spread over 26 years, so the information available about vegetation and rainfall used may have varied during this period. Further biases are that the observations did not all take place during the same season or at the same time during the day. Especially only few counts were made during the night and some night active species presence may have been underestimated. As most observations have been made on land, marine birds could unfortunately not be included in the study. Additionally on these volcanic islands the environment varies a lot and the biotopes change rapidly along a gradient on the slopes of the mountains. Thus (because the observer is always some distance away) some birds might not have been recorded exactly in their preferred breeding habitat. The environmental parameters have mostly been inferred from maps to calculate the ecological envelopes. However the time period in which these maps have been produced, may not correspond exactly with the moment of the observation of the bird. Land use and rainfall can have varied drastically in a short time frame, not forgetting the several volcanos outbursts in the concerned study period.

The observation of very mobile units like birds does also not correspond to a point observation, but of a trajectory, thus the usage of grids of about 1800 m bases seemed anyhow more appropriate than using point data.

The method used by the authors is qualified as not being a modelling in the classical definition of the term as only very few parameters could be used and some findings cannot be statistically verified. They base their method on similar approaches used in similar circumstances and also refer to Seoane et al. (2005), stating that conspicuousness, gregariousness and maximum ecological densities cannot be avoided in such studies and that greater efforts would not manage to yield that much better results in such conditions.

Concerning rainfall data used in such modelling approaches, Staub et al. (2014) identified, in a case study on Mauritius, that the determination between rainfall and landscape properties give accurate estimates with linear regression models; however, they warn that these relationships can be complicated and biased by uneven station distribution and sparsity of available data. This is especially the case for mountains and small islands. These authors argue on the other hand for a need of more sophisticated and intensive algorithms to overcome these issues, to include more rainfall stations and taking into account the individual topographic features like slope orientation and weight the stations

data points to take the effects of the variables better into account. Models of rainfall for small islands were produced that were estimated as quite accurate. Nevertheless, there are still some limitations due to doubts on the quality or accuracy of the source data, which still leaves some uncertainties of potential over estimations in some areas. Some calibration and fine-tuning is still needed.

In another context, but closely related, Martin (2008) identified in the Canary Islands IUCN home range thresholds for species that are equally affected by the uneven distribution of species. Almost all species find so their way on the Red Lists of threatened species of the Canary Islands, which made it in practice difficult to set priorities for most needed conservation areas for the decision makers. Here also an urgent plea was made that IUCN criteria be adapted and downscaled to fit the special environments of oceanic archipelagos.

If we look at the wider picture of sustainability approaches and strategic environmental assessment, including also social-economic aspects and dialogues with decision makers, we find some examples in literature. While not achieving, as they had planned, to come out with a clear view on defining best sustainability approaches, Polido et al. (2014), identified that special priorities and protocols are needed for small islands in general. Recent studies by Van Wynsberge, et al, 2015 in French Polynesia compared the relevance of two approaches for conservation planning in small tropical islands (regular planning units *versus* data driven planning units). They identified that regular planning grids highly affected the models, while data driven planning units rendered better results even on small scale pattern of interest but reduced redundancy on the other hand. They suggested a 3 steps process to identify adequate trade-offs between planning unit size, planning unit redundancy and data loss to properly draw practical recommendations for small islands.

e. Conclusions

More sophisticated and global methods thus exist but were judged as inappropriate in the framework of the atlas of breeding birds of the Comoro Islands. Some squares could not be sampled at all and remain empty, often for accessibility reasons in the field. No probability of presence nor absence or pseudo-absence/present methods could be used. The authors remedied this by identifying the areas on the maps having the most similar ecological parameters as those where the bird was actually observed.

The method was judged sound by the authors because there are sufficient filled squares per island and these are sufficiently widely distributed across each island. Thus a complete atlas of the resident birds of these islands was produced and zones of special ornithological interest could be outlined, despite the relative scarcity of available information.

The approach could be tested in other mountainous oceanic islands areas, which have been unevenly monitored over time and space and where long data series are not available both in terms of observed species numbers, number of observations or specimens collected and associated available environmental and habitat parameters.

2.4. Representation of uncertainty

a. Catchment-scale freshwater species distribution models

Partner: Senckenberg Gesellschaft für Naturforschung

Uncertainty is measured by determining the variance in the predictions from all selected model runs, for single taxa. Therefore, current measures represent the uncertainty that arises from the ensemble model. This includes uncertainty between algorithms, PA runs, repetitions and of the data-splitting

procedure. Uncertainty is given for every grid cell and can be mapped on the stream network to evaluate its spatial distribution.

b. Effect of data sources (environment and species) on SDM approaches

Partner: Fondazione Edmund Mach

Uncertainty from sampling bias is represented as a cartogram that shows a deformation in the size of the country as a function of the differences in sampling effort per unit area (**Fig. 6**). We are working in the development of other tools/methods to represent uncertainty from other sources.

c. Within-species spatial niche variation

Partner: European Bird Census Council / Centre Tecnològic Forestal de Catalunya (EBCC/CTFC)

The range of uncertainty that may arise from within-species spatial niche variation when predicting species distributions under future climate change is assessed through the comparison between the predictions from the local and global models. We used the same kind of approach as when assuming ‘full’ and ‘no’ dispersal in the absence of reliable data on the dispersal abilities of the species.

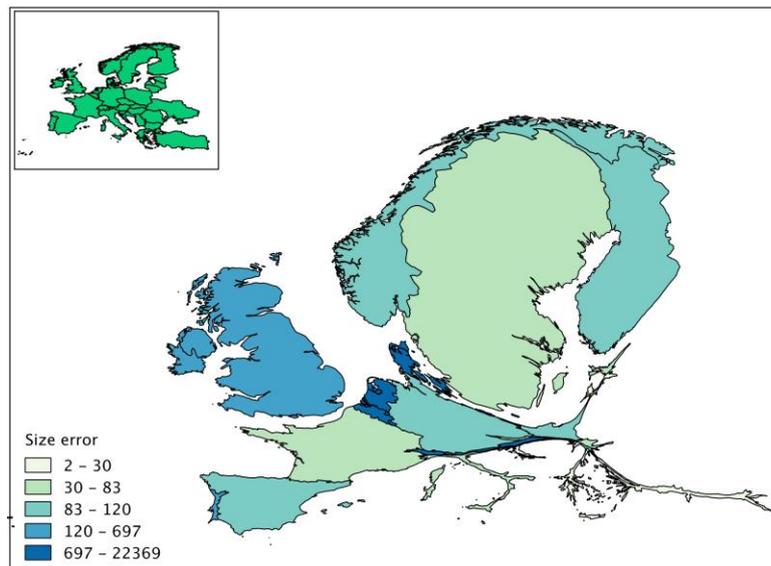


Figure 6. Cartogram, at the per country scale, of species occurrences (all taxa), extracted from GBIF data (website accessed Feb 2015). Error size above 100 indicates oversampling and error size below 100 indicates under-sampling. Polygon size represents the ratio between area of the country and the number of records.

3. Directions

Until now we have identified the properties of each of the elements needed and available to perform species distribution models. In the previous milestone we: (a) examined the characteristics of the various modelling approaches available, (b) selected a set of integrative modelling approaches and (c) highlighted the challenges to the distribution modelling process. In this this milestone we: i) evaluated the sources of environmental and species data available for species distribution modelling, ii) examined the characteristics and identified the advantages and possible limitations of such data sources and iii) started developing integrative approaches that deal with the sources of error in a way that its transparent to other WPs and end-users. We have already advanced in testing species distribution models at various spatial scales, multiple organization levels (e.g. species, communities),

and including biotic interactions. All the modelling work has been done using Open Source software, mostly the R software (R Core Team 2013) and GRASS GIS (Neteler et al. 2012, GRASS Development Team 2014).

In the analyses of data sources we have included some of the outcomes from the gap analysis performed by WP2, and in the evaluation and development of species distribution models, we have incorporated the methods for habitat classification and the Land Surface Temperature variables developed as part of WP3. Additionally, the sources of uncertainty that we have identified, such as the one derived from the properties of the data sources (e.g. islands with mountains study case), are being used by *task 4.5* in the study and quantification of uncertainty.

The next developments will include:

1. Modelling species distribution for a set of virtual species, with varying traits (e.g. rare, common, clustered, dispersed). The virtual species will be developed using the R software package “virtuallspecies” (Leroy et al. 2014). Such package allows the user to generate species habitat suitability maps (based on random or customized parameters) from which species occurrences are drawn. The result is a set of species observations with a known habitat suitability that will be used to evaluate the various modelling outputs. This will be performed with two objectives: i) further examine and identify how species characteristics and sampling biases affect the output, and ii) to evaluate the outputs using each of the integrative SDMs developed and presented in this milestone.
2. The development of a “roadmap” to species distribution modelling with examples using the integrative approaches developed in this task. The main objective of this “roadmap” is to provide information on how to perform species distribution models within a “good practices” framework, in which factors such as data quality, modelling approach, and biotic interactions are incorporated.

4. References

- Araújo, M.B. & New, M. 2007. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* 22, 42–47.
- Arino, O., Ramos, J., Kalogirou, V., Defourny, P. & Achard, F. 2009. GlobCover 2009. ESA Living Planet Symposium, 27 June. 2 July 2010, Bergen, Norway.
- Bascompte, J. 2009. Disentangling the Web of Life. *Science* 325, 416–419.
- Barwell, L.J., Azaele, S., Kunin, W.E. & Isaac, N.J.B. 2014. Can coarse-grain patterns in insect atlas data predict local occupancy?. *Diversity and distributions* 20, 895-907.
- Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. 2014. Spatial bias in the GBIF database and its effect on modelling species' geographic distributions. *Ecol. Inform.* 19, 10–15.
- Bean, W.T., Stafford, R. & Brashares, J.S. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography* 35, 250–258.
- Bicheron, P., Amberg, V., Bourg, L., Petit, D., Huc, M., Miras, B., Brockmann, C., Delwart, S., Ranéra, F., Hagolle, O., Leroy, M. & Arino, O. 2008. GlobCover: Products Description and Validation Report. ESA GlobCover project.
- Breiman, L. 2001. Random forests. *Machine Learning* 45, 5-32

Britz W. & Witzke P. 2012. CAPRI model documentation 2012. http://www.capri-model.org/docs/capri_documentation.pdf [accessed: 14.04.2015]

Brooker, R.W. & Callaghan, T.V. 1998. The Balance between Positive and Negative Plant Interactions and Its Relationship to Environmental Gradients: A Model. *Oikos* 81, 196–207.

Canovas, F., Magliozzi, C., Palazon-Ferrando, J.A., Mestre, F. and Gonzalez-Wanguemert, M. 2014. ENiRG: R-GRASS interface to efficiently characterize the ecological niche of species and predict habitat suitability. *Ecography*. (in review).

Chen, J., Chen, J., Gong, P., Liao, A. & He, C. 2011. Higher resolution GLC mapping. *Geomatics World* 4, 12–14

Chamberlain, S., Ram, K., Barve, V. & Mcglinn, D. 2015. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.8.0. <http://CRAN.R-project.org/package=rgbif>

Domisch, S., Kuemmerlen, M., Jähnig, S.C. & Haase, P. 2013. Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. *Ecol. Model.* 257, 1–10.

Domisch S., Jähnig S.C., Simaika J.P., Kuemmerlen M. & Stoll S. 2015. Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. *Fundamental and Applied Limnology/Archiv für Hydrobiologie* 186, 45–61.

Duivenvoorden, J.F., Svenning, J-C & Wright S.J. 2002. Beta diversity in tropical forests. *Science* 295, 636-637.

EEA. 1995. CORINE Land Cover Project, published by Commission of the European Communities.

Franklin, J. 2010. Mapping species distributions. Spatial inference and prediction. Cambridge University press.

Gaiji, S., Chavan, V., Ariño, A.H., Otegui, J., Hobern, D., Sood, R., & Robles, E. 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. *Biodiversity Informatics* 8, 94-172.

GBIF. 2014. Global Biodiversity Information Facility (GBIF). <http://www.gbif.org> (Accessed February 2015).

Gilman, S.E., Urban, M.C., Tewksbury, J., Gilchrist, G.W., & Holt, R.D. 2010. A framework for community interactions under climate change. *Trends Ecol. Evol.* 25, 325–331.

GRASS Development Team, 2014. Geographic Resources Analysis Support System (GRASS) Software, Version 6.4.4. Open Source Geospatial Foundation. <http://grass.osgeo.org>

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. *Global ecology and biogeography* 24, 276-292.

Hanspach, J., Schweiger, O., Kühn, I., Plattner, M., Pearman, P.B., Zimmermann, N.E. & Settele, J. 2014. Host plant availability potentially limits butterfly distributions under cold environmental conditions. *Ecography* 37, 301-308.

Hawkins, B.A. 2012. Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography* 39, 1–9.

- Heikkinen, R.K., Luoto, M., Leikola, N., Pöyry, L., Settele, J., Kudrna, O., Marmion, M., Fronzek, S. & Thuiller, W. 2010. Assessing the vulnerability of European butterflies to climate change using multiple criteria. *Biodiversity and Conservation*, 19, 695–723.
- Henle, K., Potts, S., Kunin, W., Matsinos, Y., Simila, J., Pantis, J., Grobelnik, V., Penev, L. & Settele, J. (Eds). 2014. *Scaling in Ecology and Biodiversity Conservation*. Advanced Books: e1169.
- Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25, 1965-1978.
- Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data?. *Ecology* 83, 2027-2036.
- Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* 117, 847–858.
- Hutchinson M., Xu T., Houlder D., Nix H. & McMahon J. 2009. *ANUCLIM 6.0 User's Guide*. Australian National University, Fenner School of Environment and Society.
- Keener, V. W., Marra, J. J., Finucane, M. L., Spooner, D., & Smith, M. H. (Eds.). 2012. *Climate Change and Pacific Islands: Indicators and Impacts*. Report for The 2012 Pacific Islands Regional Climate Assessment. Washington, DC: Island Press
- Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McNerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J-C., Zimmermann, N.E. & O'Hara, R.B. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents: Modelling multispecies interactions. *J. Biogeogr.* 39, 2163–2178.
- Kriticos D.J., Jarošik V. & Ota N. 2014. Extending the suite of Bioclim variables: a proposed registry system and case study using principal components analysis. *Methods in Ecology and Evolution* 5, 956-960.
- Kudrna, O., Harpke, A., Lux, K., Pennerstorfer, J., Schweiger, O., Settele, J. & Wiemers, M. 2011. *Distribution Atlas of Butterflies in Europe*. Gesellschaft für Schmetterlingsschutz, Halle, 1-576 pp.
- Kuemmerlen, M., Schmalz, B., Guse, B., Cai, Q., Fohrer, N. & Jähnig, S.C., 2014. Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. *Ecol. Model.* 277, 77–86.
- Kuemmerlen, M., Stoll, S., Sundermann, A., Haase, P. 2015. Long-term monitoring data meet freshwater species distribution models: Lessons from an LTER-site. *Ecological Indicators*. *In press*.
- Leroy, B. with help from C. N. Meynard and C. Bellard & F. Courchamp. 2014. *virtualspecies: Generation of Virtual Species Distributions*. R package version 1.0. <http://CRAN.R-project.org/package=virtualspecies>
- Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28, 385–393
- Louette, M., Abderemane, H., Yahaya, I. & Meirte, D. 2008. *Atlas des oiseaux nicheurs de la Grande Comore, de Mohéli et d'Anjouan*. Series 'Studies in Afrotropical Zoology', 294. Tervuren: RMCA. 240 p.

- Louette, M., Bijmens, L., Agenon'ga Upoki, D. & Fotso, R.C. 1995. The utility of birds as bioindicators: case studies in equatorial Africa. *Belg. J. Zool.* 125, 157-165
- Louette, M. & Meirte, D. 2009. 'Birds from the Albertine Rift'. *Mountain Forum Bulletin* 9 (2), 50
- Maharaj, S.S & New, M. 2013. modeling individual and collective species responses to climate change within Small Island States. *Biological Conservation* 167, 283-291
- Martin J. L. 2009. Are the IUCN standard home-range thresholds for species a good indicator to prioritise conservation urgency in small islands? A case study in the Canary Islands (Spain). *Journal for Nature Conservation.* 17, 87-98.
- McCullagh P, Nelder JA (1989) *Generalized linear models*. London: Chapman and Hall. 511 p
- Metz, M., Rocchini, D. & Neteler, M. 2014. Surface temperatures at the continental scale: Tracking changes with remote sensing at unprecedented detail. *Remote sensing.*
- Mitchell, T.D., Carter, T.R., Jones, P.D., Hulme, M. & New, M. 2004. A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100). Working Paper No. 55, Tyndall Centre for Climate Change Research, UK, 1-30 pp.
- Morales-Castilla, I., Matias, M.G., Gravel, D., & Araújo, M.B. 2015. Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30, 347–356.
- Neteler, M., Bowman, M.H., Landa, M. and Metz, M. 2012. GRASS GIS: a multi-purpose Open Source GIS. *Environmental Modelling & Software* 31, 124-130
- Osborne, P.E., Foody, G.M. & Suárez-Seoane, S. 2007. Non-stationarity and local approaches to modeling the distributions of wildlife. *Diversity and Distributions*, 13, 313–323.
- Panagos, P. 2006. The European soil database. *GEO: connexion* 5, 32-33.
- Pearman, P.B., Guisan, A., Broennimann, O., & Randin, C. 2008. Niche dynamics in space and time. *Trends in Ecology and Evolution*, 23, 149-158.
- Polido, A., João, E. & Ramos T. B. 2014. Sustainability approaches and strategic environmental assessment in small islands: An integrative review. *Ocean & Coastal Management* 96: 138-148.
- Pulliam, H.R. 2002. On the relationship between niche and distribution. *Ecology letters* 3, 349-361.
- R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., & Chiarucci, A. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Prog. Phys. Geogr.* 35, 211–226.
- Rohde, K. 1992. Latitudinal gradients in species diversity: the search for the primary cause. - *Oikos* 65: 514-527
- Rota, C.T., Fletcher, R.J., Evans, J.M. & Hutto, R.L., 2011. Does accounting for imperfect detection improve species distribution models? *Ecography* 34, 659–670.
- Sandel, B. 2014. Towards a taxonomy of spatial scale-dependence. *Ecography* 37, 001-012.

- Silla, C. N. & Freitas, A.A. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22, 31-72.
- Seoane, J. L. M., Carrascal, C. L. A. & Palomino, D. 2005. Species-specific traits associated to prediction errors in bird habitat suitability modeling. *Ecological modeling* 185, 299-308.
- Schweiger, O., Heikkinen, R.K., Harpke, A., Hickler, T., Klotz, S., Kudrna, O., Kühn, I., Pöyry, J. & Settele, J. 2012. Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecol Biogeography* 21, 88-99.
- Staub, C. G., Stevens, F. R. & Waylen, P.R. 2014. The geography of rainfall in Mauritius: modeling the relationship between annual and monthly rainfall and landscape characteristics on a small volcanic island. *Applied Geography* 54, 222-234.
- Thuiller, W., Georges, D. & Engler R. 2014. biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64. <http://CRAN.R-project.org/package=biomod2>
- Tylianakis, J.M., Didham, R.K., Bascompte, J. & Wardle, D.A. 2008. Global change and species interactions in terrestrial ecosystems. *Ecol. Lett.* 11, 1351–1363.
- Van der Putten, W.H., Macel, M. & Visser, M.E. 2010. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 2025–2034.
- Van Wynsberge, S., Andréfouët, S., Gaertner-Mazouni, N. & Remoissenet G. 2015. Conservation and resource management in small tropical islands: Trade-offs between planning unit size, data redundancy and data loss. *Ocean & Coastal Management* 116, 37-43.
- Whittaker, R.J., Willis, K.J. & Field, R. 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of biogeography* 28, 453-470.
- Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning J.C. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* 88, 15–30.
- Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B. & Svenning, J.-C. 2010. New trends in species distribution modelling. *Ecography* 33, 985–989.