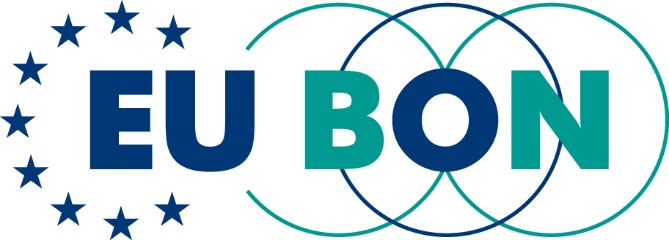
****

**Deliverable 3.2 (D3.2)**

**Report on integrated distributional modelling and associated software**

**M040**

|  |  |
| --- | --- |
| 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 |
| Delivery date from Annex I: | M40 (March 2016) |
| Actual delivery date: | M40 (March 2016) | |
| Lead beneficiary: | NRM, UnivLeeds, | |
| Authors: | Yoni Gavish, Charles J. Marsh, William E. Kunin -  UniLeeds, University of Leeds, School of Biology, UK  Markus Skyttner, Sven Kullander -  NRM, wedish Museum of Natural History, Sweden  Cristina Garilao , Kathleen Reyes -  FIN, FishBase Information and Research Group, Phillipines  Mathias Kuemmerlen, Stefan Stoll, Peter Hasse -  SGN, Senckenberg Gesellschaft für Naturforschung, Germany  Johannes Penner –  MfN, Museum für Naturkunde, Germany | |

|  |  |  |
| --- | --- | --- |
| **This project is supported by funding from the specific programme 'Cooperation', theme 'Environment (including Climate Change)' under the 7th Research Framework Programme of the European Union**  **Dissemination Level** | | |
| PU | Public | 🗸 |
| PP | Restricted to other programme participants (including the Commission Services) |  |
| RE | Restricted to a group specified by the consortium (including the Commission Services) |  |
| CO | Confidential, only for members of the consortium (including the Commission Services) |  |

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

*All intellectual property rights are owned by the EU BON consortium members and protected by the applicable laws. Except where otherwise specified, all document contents are: “© EU BON project“. This document is published in open access and distributed under the terms of the Creative Commons Attribution License 3.0 (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*



**Executive Summary**

**Introduction**

The main objective of *WP3* is to develop and improve tools and methods for biodiversity data analyses, covering habitat classification tools (*task 3.1*), scaling issues (*task 3.2*), species distribution models (*task 3.3*) and data mining (*task 3.4*). This deliverable focuses on task 3.3 that aims to develop enhanced species distribution models and has 3 main objectives:

1. In data poor systems, we aimed to increase map accuracy by allowing expert to edit various aspects of predicted distribution maps.
2. In data rich systems, we aimed to develop hybrid models that account for both niche aspects and spatial aspects when predicting species distributions.
3. We aimed to ensure any tool developed under task 3.3 is available for external users either as a web interface or as an R application.

To meet these objectives we developed five analytical tools, and this deliverable provides background information and implementation code for each. In sections devoted to each tool, we first clarify the tool’s aim and then describe the conceptual rationale behind it. This is followed by a more detailed explanation of the content of each tool along with a clear description of the additional information provided in the appendices or supporting information.

**Progress towards objectives**

We have made considerable progress towards all three objectives. For the first objective, we developed 2 tools based on the *AquaMaps* platform, including the Create your Own Map (*CYOM*) feature of *AquaMaps* (FIN) and an R package implementing the *AquaMaps* algorithm (NRM). For the second objective we developed 2 tools, including a set of hybrid spatial/niche model (UnivLeeds) and freshwater ensemble SDM method (SGN). The last tool is a large scale diversity calculator (MfN) that summarizes distribution patterns of multiple species over wide extent at fine resolution. To meet the third objective, all tools were developed either as R applications or as web-based platforms.

**Achievements and current status**

1. ***CYOM* (section 2, FIN) –** A web-interface tool allowing experts to edit an erroneous map and regenerate a more correct version of it. The expert can alter various stages of the *AquaMaps* algorithm, including: the bounding area, the input occurrence points, the parameters of the environmental envelopes or the map viewing setting. *CYOM* is embedded within the main *AquaMaps* web platform.
2. ***rAquaMaps* (section 3, NRM) –** a standalone R package implementing *AquaMaps*’ algorithm in an R environment. The package reduces computation time significantly relative to the original *AquaMaps* web interface. Furthermore, the package permits users great flexibility in modelling, e.g., modelling a single species or a big batch of species, using own input data and testing with optional environmental parameters. NRM further developed a Shiny-based web application of *rAquaMap* that provide similar interface and features as the *CYOM* tool.
3. **Hybrid spatial/niche model (section 4, UnivLeeds)** **–** a set of four R functions, implementing four different hybrid models, including:
   1. *Moving Windows SDM* – accounting for the mean probability of occurrence (PoO) at user defined window sizes around each cell when modelling species distributions.
   2. *Top X* – Selection of the top X cells with the highest PoO, with X being the predicted occupancy from downscaling models (models that predict fine-scale occupancy from coarse-scale occupancy).
   3. *TopDown PoO* – selection of fine-scale occupancies based on mean PoO at various scales to produce a presence/absence map with the exact number of occupied cells at each scale as predicted by the downscaling models.
   4. *SpaNiche model* – Selection of a single global threshold value that balances fine-scale and coarse-scale accuracy.
4. **Improved freshwater SDMs (section 5, SGN) –** Adaptation of SDMs to freshwater environments through the choice of explanatory variables and spatial configurations. The adapted SDMs were used to model the distribution of local freshwater biodiversity in the Rhine-Main Observatory. A modelling framework for high resolution freshwater SDMs has been developed to serve as a guideline for similar applications to be implemented elsewhere and achieve comparable results.
5. **Diversity calculator (section 6, MfN) –** A free software that calculates alpha and beta diversity on a stack of raster data which have a large number of cells – a task that is not easily achieved in available software (e.g., R) due to computational limits. Utilizes the predicted presence/absence maps of multiple species (i.e., the output of the four former tools)

**Future developments**

The different tools face different challenges ahead:

1. *CYOM* – Assess whether the implemented features are functional and user-friendly
2. *rAquaMaps –* Ensure maintenance of the R package and Shiny web application.
3. *Hybrid models* – Add functions to the R package ‘*downscale*’, developed under task 3.2.
4. *freshwater SDMs* – Finish the development of the freshwater SDM guide and test in additional catchments.
5. *Diversity calculator* – Implement as a standalone R package and allow additional function to be estimated within each window.

Table of Contents

[1: General Introduction 5](#_Toc445210657)

[2: Create Your Own Map (*CYOM*) in *AquaMaps* 10](#_Toc445210658)

[2.a: Aim 10](#_Toc445210659)

[2.b: Introduction 10](#_Toc445210660)

[2.c: Approach 10](#_Toc445210661)

[2.d: Create-Your-Own-Map (CYOM) Manual 13](#_Toc445210662)

[3: *rAquaMaps* Global Modelling Tool 30](#_Toc445210663)

[3.a: Aim 30](#_Toc445210664)

[3.b: Introduction 30](#_Toc445210665)

[3.c: Approach 31](#_Toc445210666)

[3.d: Main functionalities of the R package 31](#_Toc445210667)

[3.e: Versions, installations, guides and recent changes 32](#_Toc445210668)

[3.e: Web-enabled usage of package features 32](#_Toc445210669)

[4: Hybrid Species Distribution Models 33](#_Toc445210670)

[4.a: Aim 33](#_Toc445210671)

[4.b: Introduction 33](#_Toc445210672)

[4.c: Approach 34](#_Toc445210673)

[4.d: Case study 36](#_Toc445210674)

[4.e: Moving Windows SDM 42](#_Toc445210675)

[4.f: Top X occupied cells 45](#_Toc445210676)

[4.g: TopDown PoO 47](#_Toc445210677)

[4.h: The SpaNiche model 53](#_Toc445210678)

[4.i: Comparison of the hybrid models 59](#_Toc445210679)

[5: Improved, high resolution freshwater SDMs 63](#_Toc445210680)

[5.a: Aim 63](#_Toc445210681)

[5.b: Introduction 63](#_Toc445210682)

[5.c: Approach 63](#_Toc445210683)

[5.d: The freshwater SDM framework 64](#_Toc445210684)

[5.e: Applications of the tool 68](#_Toc445210685)

[6: Diversity calculator 69](#_Toc445210686)

[6.a: Aim 69](#_Toc445210687)

[6.b: Introduction 69](#_Toc445210688)

[6.c: Approach 69](#_Toc445210689)

[6.d: Current status 69](#_Toc445210690)

[7: References 71](#_Toc445210691)

[8: Appendices 73](#_Toc445210692)

[Appendix 2.1 – FAO Major Fishing Area 74](#_Toc445210693)

[Appendix 2.2 – AquaMaps five-star rating system 75](#_Toc445210694)

1: General Introduction

Effective management of natural systems at various scales and levels of organization requires a detailed understanding of the states and trends of populations, species, communities and ecosystems. Such understanding cannot be achieved without accumulation and mobilization of biodiversity data. Thus, one of the main objectives of EU BON is to integrate and harmonize various data sources and increase their accessibility to various stakeholders ([Hoffmann et al. 2014](#_ENREF_16)) (#1463). However, an effective management scheme cannot rely on raw data alone. In fact, the three main pillars of EUBON -- data sources & infrastructure, science & application, and policy & dialogue -- follow the transformation of various types of raw data through statistical analysis to a clearer understanding of states or trends that can then be translated to policy. The role of WP3 in this general framework is in the science & application pillar- our aim is to develop new analytical tools and/or improve existing analytical tools for analyses of biodiversity data. The tools developed in WP3 will undergo testing and validation in WP5 and will be used in WP4 to identify biodiversity status and trends. Furthermore, the tools will be disseminated to the wider scientific community, policy makers, and other stakeholders by WP8. Finally, the integration of the tools within EU BON’s biodiversity portal will be explored by WP2.

WP3 covers four of the many aspects of biodiversity, including habitat classification tools (task 3.1), scaling issues (task 3.2), species distribution models (SDMs) (task 3.3) and data mining (task 3.4). In task 3.1 we aim to develop enhanced tools that use fine-scale multi-spectral remotely-sensed imagery along with other earth observation data to map habitats, land-covers or land-uses and to monitor their change. In task 3.2 we aim to develop tools that use readily available data sources at certain scales to predict hard-to-measure properties of biotic datasets such as occupancy or richness at other scales. In task 3.3 we aim to develop enhanced methods for species distribution models for both data-limited and data-rich systems. Finally, in task 3.4 we aim to develop a tool to semi-automatically extract user defined information from published data. Task 3.1 and task 3.2 were reported in deliverable D3.1, while task 3.4 will be reported in D3.3 (due M48), although considerable information on the various tools developed in these tasks can be found in EU BON’s progress reports. In this deliverable we focus on the five different tools developed in task 3.3, jointly led by UnivLeeds and NRM, with considerable contribution from FIN, SGN and MfN. However, we note here that several additional partners (EBCC, MRAC, NHM, and UFZ) contributed considerably to tool development through their suggestions and critiques during web conferences and annual meetings.

The tools developed under task 3.3 can be partitioned to two main groups (Fig. 1.1). The first aims to provide appropriate modelling framework for data-poor systems, with a focus on marine habitat, and is based on the *AquaMaps* algorithm. The second aims to develop enhanced SDMs in data rich systems by accounting for various aspects that may affect the distribution of species (e.g., dispersal barriers, disturbance history and biotic interactions). In the first group, FIN and NRM developed 2 tools, the *Create-Your-Own-Map* (*CYOM*) and *rAquaMaps*, respectively. In the second group, 2 tools were developed. First for terrestrial habitats we developed four hybrid models that account for both spatial and niche when predicting species distribution (UnivLeeds). Second, we explored various ways to adapt SDMs to the special characteristics of freshwater environments (SGN). Finally, we include the diversity calculator tool (MfN) - a tool that relates presence/absence maps of multiple species to patterns of alpha and beta diversity at fine resolution over larger extent (Fig. 1.1).

|  |  |
| --- | --- |
| **Figure 1.1:** The tools developed under task 3.3 of WP3 cover all three main habitat realms and various levels of data availability. | C:\Users\Yoni Gavish\Documents\1-Work\Tree of Knowledge\5-EU-BON\General\Deliverables\WP3\D3_2\Fig 1_1.png |

The *CYOM* tool is an extension of the *AquaMaps* web-interface that allows user to control various aspects of the *AquaMaps* algorithm. For example, the user can add or remove occurrence points, change the boundaries of the species range, edit the parameters of the various environmental envelops or change the map viewing settings. Thus, the *CYOM* interface allows expert to correct erroneous maps generated automatically by the *AquaMaps* algorithm. This feature is extremely useful for data-poor species, where much uncertainty concerns every aspect of the modelling procedure and the most reliable source is an expert. It is designed as a user-friendly web-interface to increase experts’ participation. This tool is covered in Section 2 below, which also includes a comprehensive manual.

The *AquaMaps* algorithm is also central to the work done under task 3.3 by NRM. Although *AquaMaps*’ web-interface have been used to model the distribution of thousands of species, it is relatively inefficient in terms of computation time and is not compatible with open source spatial and analytical software. Therefore, NRM developed ‘*rAquaMaps*’– an R package that implements *AquaMaps* algorithm within an R environment. The package permits users great flexibility in modelling. For example, it allows modelling of a single species or a big batch of species, using own input data, testing with optional environmental parameters and various other adaptations. *rAquaMaps* can show actual distribution (point data), projected distribution using the probabilities of occurrence (suitable habitat), or modelled actual distribution using a combination of probabilities and a system of bounding polygons that constrain the distribution to known areas of occurrence. Furthermore, NRM have developed a Shiny-based web-interface that provides similar features as the *CYOM*. This package is covered in more details in section 3, where links to three vignettes (R manuals) can be found.

For the second group of tools, UnivLeeds developed four hybrid models that aim to incorporate both spatial and niche aspects when modelling the distribution of species. The 4 models differ from one another in the way they modify the SDM’s Probability of Occurrence (PoO) map, in their reliance on downscaling models (models that predict fine-scale occupancy from coarse-scale occupancy) and in the way they translate the continuous PoO map to a binary Presence/Absence (P/A) map. The four models include: **a)** The *Moving Windows SDM* which accounts for the mean PoO at user defined windows size around each cell when modelling species distributions. **b)** The *Top X* model which selects the top X cells with the highest PoO, with X being the predicted occupancy from downscaling models. **c)** The *TopDown PoO* model which selects fine-scale occupancies (based on mean PoO at various scales) such that the generated P/A map have the exact number of occupied cells at each scale as predicted by the downscaling models. d) The *SpaNiche model* which selects a single global threshold value that balances fine-scale and coarse-scale accuracy. All four models also include the option to mask the entire distribution with known atlas data, thus providing a total of nine hybrid models (masked and unmasked version of each model + masking of the original SDM). In this deliverable, UnivLeeds provide examples for the tallow wagtail (*Motacilla Flava*) from Wallonia (Belgium), as part of a larger collaboration with EBCC. Section 4 contains a comprehensive description of each of the four models, as well as its application to the case study.

The hybrid models listed above are highly suitable for terrestrial landscapes, yet their applicability to freshwater environment is limited, mainly since freshwater habitats are arranged hierarchically along dendritic stream networks. For example, it is meaningless to estimate the mean suitability in user defined window sizes around each cell in a stream, since most of the surrounding landscape are terrestrial and are thus unsuitable for stream biota. In addition, freshwater system exhibit stronger spatial directionality than terrestrial systems. That is, conditions in a certain stream location are affected by events and conditions upstream (but less so downstream) and by inflow of energy and materials from the surrounding landscape. To successfully model the distribution of freshwater biota, this spatial aspects need to be accounted for in the modelling framework. As part of task 3.3, SGN created a modelling framework for freshwater systems that account for such aspects by the choice of input variables and by the inclusion of upper sub-catchments predictors in the landscape. As the main contribution here is in the choice of variables, this section does not include a supporting R file. Instead, this tool identify a framework to adapt SDMs to freshwater environment. Section 5 of this deliverable contains a detailed account on the framework with examples from one of EU-BON’s focal observatory sites- The Rhine Main Observatory.

Finally, all four tools listed above provide P/A maps for a single species. However, in many cases, effective management requires information on communities. This can be done by stacking P/A maps of multiple species, as recently shown by *AquaMaps* for the bony fish in the North Sea (CITATION). In such cases, it may be informative to map community level biodiversity indices such as alpha and beta diversity. Unfortunately, mapping alpha and beta diversity over wide extent at fine resolution is not straightforward in commonly used platform such as R, due to computational limitations. To meet this need, MfN developed the diversity calculator- a tool that calculate alpha and beta diversity on a large stack of raster (grid) data, using a moving window approach. The tool is computationally efficient by dividing the work to multiple cores (parallel computation) and restructuring the results thereafter.

To summarize, the five tools developed under task 3.3 may supplement existing tools for both data poor and data rich systems, and allow aggregating the results of multiple species to informative summary maps at the community level. Like all other tools developed under WP3, we have put considerable effort to make the developed tools easily available to the wider audience. Here, we either based the tools on R - the most commonly used analytical platform, or added the tool to existing high-traffic web-interfaces. Table 1.1 below provide direct link to all the tools developed in this task and contact information for further details.

**Table 1.1** List of tools covered in this deliverable, their contact person and relevant supporting files that can be accessed at: ANKE - ADD WEB LINK HERE ONCE THE SI FILES ARE UPLOADED TO EUBON’S SITE

|  |  |  |
| --- | --- | --- |
| Tool | Contact person | Additional information |
| *CYOM*- *AquaMaps* | Cristina Garilao (CG) | ADD LINK TO WEB INTERFACE WITH THE CYOM TOOL |
| *rAquaMaps* | Markus Skyttner (MS) | SI\_3.1.pdf *-- R* Help files for the ‘*rAquaMaps*’ package |
| Hybrid models | Yoni Gavish (YG) | Win\_PoO.R -- R file, the function for the *Moving Window SDM* model |
| TopX.R *--* R file, the function for the *Top X* model |
| TopDown\_PoO.R *--* R file, the function for the *TopDown PoO* model |
| SpaNiche.R -- R file, the function for the *SpaNiche* model |
| Freshwater SDMs | Mathias Kuemmerlen (MK) | MATHIAS - ANYTHING TO ADD HERE? |
| Diversity calculator | Johannes Penner (JP) | Link to tools page -- https://github.com/moritzaugustin/mwmc |

Emails. CG - [cgarilao@geomar.de](mailto:cgarilao@geomar.de) ; MS - [Markus.Skyttner@nrm.se](mailto:Markus.Skyttner@nrm.se) ; YG - [gavishyoni@gmail.com](mailto:gavishyoni@gmail.com) ; MK - [mathias.kuemmerlen@senckenberg.de](mailto:mathias.kuemmerlen@senckenberg.de) ; JP - Johannes.Penner@mfn-berlin.de

2: Create Your Own Map (*CYOM*) in *AquaMaps*

***2.a: Aim***

To improve the existing *CYOM* interface for incorporating expert information in *AquaMaps*.

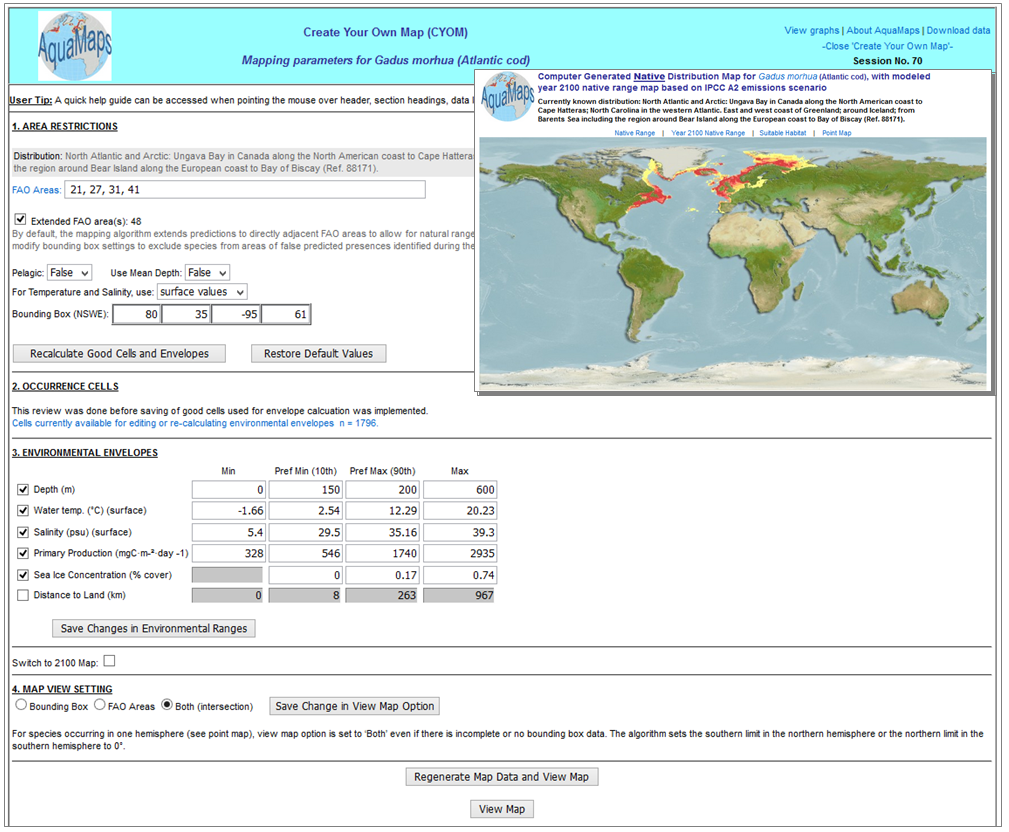
2.b: Introduction

*AquaMaps* is a species distribution modelling approach that combines ecological information, occurrence data and environmental layers to derive a species' environmental tolerances, and predict its natural distribution based on habitat suitability. While this approach enables the generation of a large number of species distribution maps, a map may sometimes be incorrect due to sampling biases, outdated input data or data encoding errors. In such cases, map predictions can be improved if reviewed by an expert.

In *AquaMaps*, *Create-Your-Own-Map* (*CYOM*) is an online tool that allows an expert to edit an erroneous map and regenerate a corrected version of it (tagged as a reviewed map accordingly). The *CYOM* offers several advantages: 1) It makes transparent all parameters and settings used in generating an AquaMap for a species thus making it easier to evaluate and explore; 2) It enables an expert to use/input his own knowledge or data to correct *AquaMaps*’ predictions on the distribution range and relative likelihood of occurrence of a given species; 3) It improves the quality of the distribution dataset behind other *AquaMaps* tools (e.g., country species checklists and species richness maps); 4) Expert-corrected maps are immediately available/downloadable online; 5) Map parameters and settings from expert-corrected maps can serve as high-quality input data for subsequent *AquaMaps* modelling for a species; and 6) All versions of edited maps are compiled by species and thus document the history of corrections and changes made to *AquaMap*s for the species.

2.c: Approach

Create-Your-Own-Map (*CYOM*) is a user interface designed to allow an expert to edit an erroneous species map in four areas: (1) area restrictions that define the known native range of a species; (2) point data used in deriving environmental tolerance of a species, (3) species tolerance threshold estimates (environmental envelopes) for different environmental parameters and which set of parameters to use to predict the distribution of a given species, and, (4) the map display settings (see Fig. 2.1).



**Figure 2.1:** Create-Your-Own-Map (*CYOM*) interface. Shown are four sections of mapping parameters used in modelling the distribution range of Atlantic cod *Gadus morhua*. The interface is interactive and allows a species expert or map reviewer to change the map settings and regenerate an edited species map. Corresponding native distribution map in inset.

Actions/edits that can be performed using the *CYOM* are outlined in table 2.1. An expert can make changes to any or all of these map settings. Built-in commands in the *CYOM* allow the expert to regenerate and plot a new map based on the changes made, and save both the new map and its settings online. The expert can also include remarks about the map and a map rating. An online user's guide detailing the *AquaMaps* methodology, algorithm and data sources has been prepared for further reference but is currently being revised to reflect the latest updates.

Improvements have been implemented to refine the routines behind the *CYOM*. These include:

(1) Incorporation of a new "good cell" rule in the *CYOM* interface (i.e., occurrence cell is within both FAO area AND bounding box limits of a species);

(2) Generation/saving of point map displaying actual good cells used in a reviewed map; and

(3) Saving of images restricted to *CYOM* routine with activity password protection to prevent unintentional map updating on the server by normal users.

**Table 2.1:** Sections and actions enabled in the Create-Your-Own-Map (*CYOM*) interface.

|  |  |
| --- | --- |
| **Section** | **Reviewer actions/edits** |
| **Area Restrictions** | * FAO areas: Add/delete FAO areas according to where a species is native or endemic. * Pelagic flag: Change setting to either TRUE if species distribution is not affected by depth, or FALSE if it is influenced by bottom depth. * Sea temperature and salinity layers: Change to use either surface or bottom values. * Bounding box: Adjust/complete latitudinal or longitudinal extents to encompass area of known native range of species. |
| **Occurrence Cells** | * Adding point data/good cells to be used in generating environmental envelopes (by encoding coordinates or selecting location on the map) * Excluding cells from generation of environmental envelopes |
| **Environmental Envelopes** | * Manual adjustment of species environmental threshold values * Including/excluding use of particular environmental parameters/layers when predicting species distribution (probability of species occurrence) |
| **Map View Settings** | * Specifying area coverage when plotting predicted species distribution range. |

The incorporation of the new ‘good cell’ rule improves the accuracy of computing environmental tolerances of a species, while refinements to the routines for saving point maps and map images ensures mapping parameters and associated images remain intact for any given species.

An ongoing improvement to the routine involves relaxing the number of "good cells" required to model species' distribution range from 10 good cells to 3 or more good cells, and enabling expert-review for these maps through the *CYOM*. Relaxing the good cell requirement enables more species to be mapped and made available for expert-review (currently estimated at over 22,800 marine species as of the September 2015 *AquaMaps* run).

Additional features have also been implemented that would help experts and users in navigating and using the *CYOM*. Among these are the inclusion of text explaining the purpose of the *CYOM* and a pop-up help guide when hovering a mouse over sections, field names and command buttons in the *CYOM* interface.

Establishing links to literature cited by reviewers when checking and editing maps are among the features to be explored in the coming months, along with the setting up of alerts in the *AquaMaps* site and the *FishBas*e and *SeaLifeBase* Facebook pages to advertise whenever a map becomes available for review.

A fully functioning beta version is available for *AquaMaps*’ *CYOM* feature. An example for the computer-generated distribution map of the Atlantic cod *Gadus morhua* can be accessed at: <http://www.aquamaps.org/CreateOwnMap.php?expert_id=0&expert_oc_exists=&what=0&SpecID=Fis-29394&area_res=1&user_session=34&user_session_bef=34&from=premap.>

The next step will be to test, in a case study, whether the implemented features are functional and user-friendly. This is expected to take place towards the end of the first halfof 2016 with the help of selected species experts, after the completion of the latest *AquaMaps* run (Sep-October 2015). Species or family experts will be invited to use the improved *CYOM* tool. Alerts via the *AquaMaps* site and the *FishBase* and *SeaLifeBase* Facebook pages will also advertise the latest maps and the improved *CYOM* tool. Distribution maps for top commercial fishes occurring in Europe will be the priority for map review.

It is asserted however that the case study is intended to test the *CYOM* for ease of use for experts to review, edit and regenerate *AquaMaps*. The aim is to be able to successfully integrate experts’ inputs and assessments through species *AquaMaps* reviewed using the *CYOM*. This, in turn, will be used to inform subsequent mapping and other *AquaMaps* products. The *CYOM* tool is not necessarily nor immediately expected to result in a substantial increase in the number of reviewed maps, yet a more user-friendly interface may increase experts’ participation at the long run. .

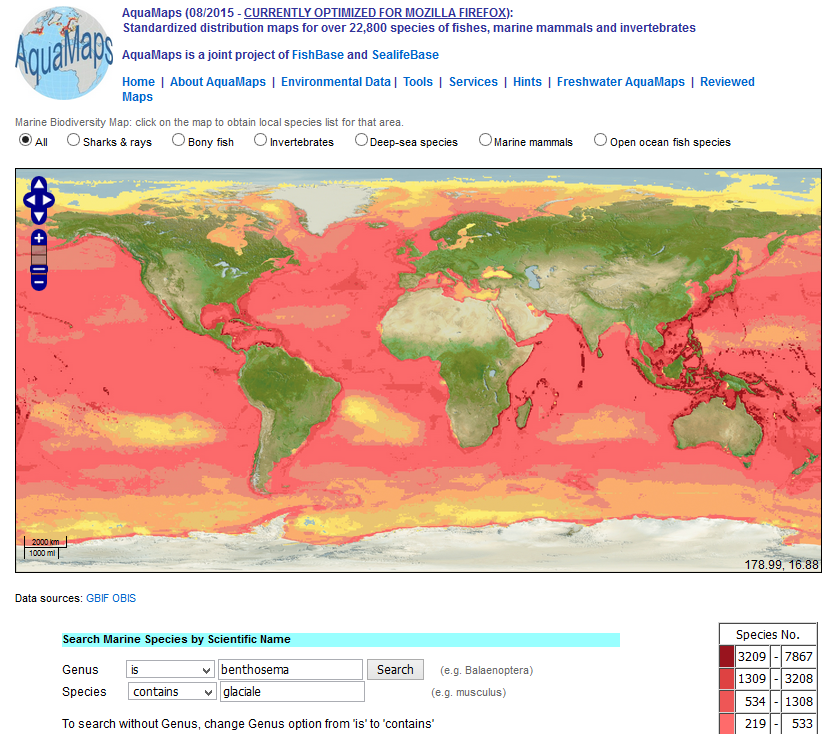
2.d: Create-Your-Own-Map (CYOM) Manual

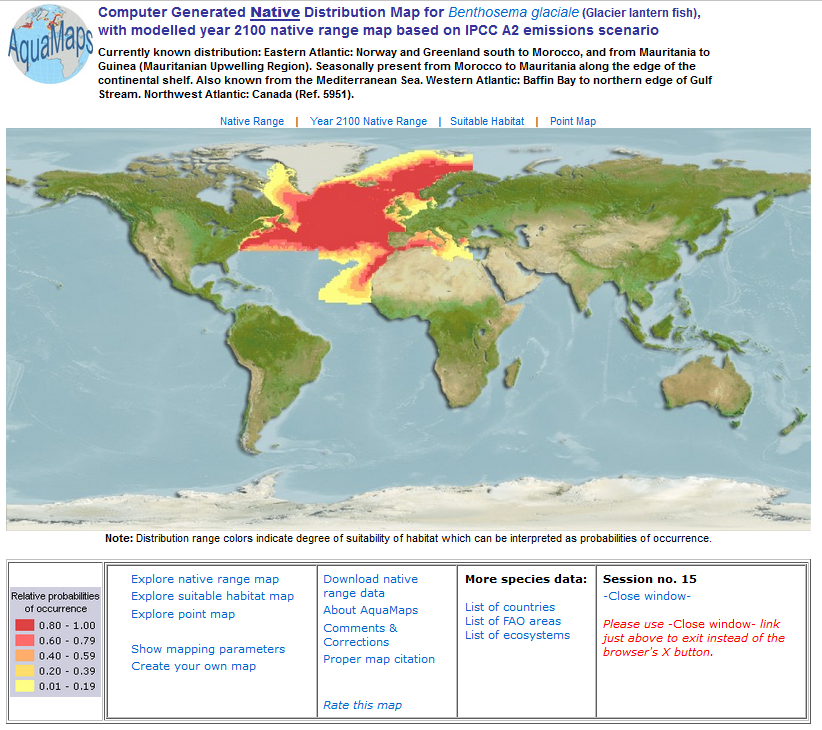
This section contains the manual for using the *AquaMaps* *Create-Your-Own-Map* (*CYOM*) Tool. Topics are arranged following a sequence of steps from calling a species distribution map, editing mapping parameters, regenerating the map, and saving/publishing the new map online.

*2.d.1: Getting started*

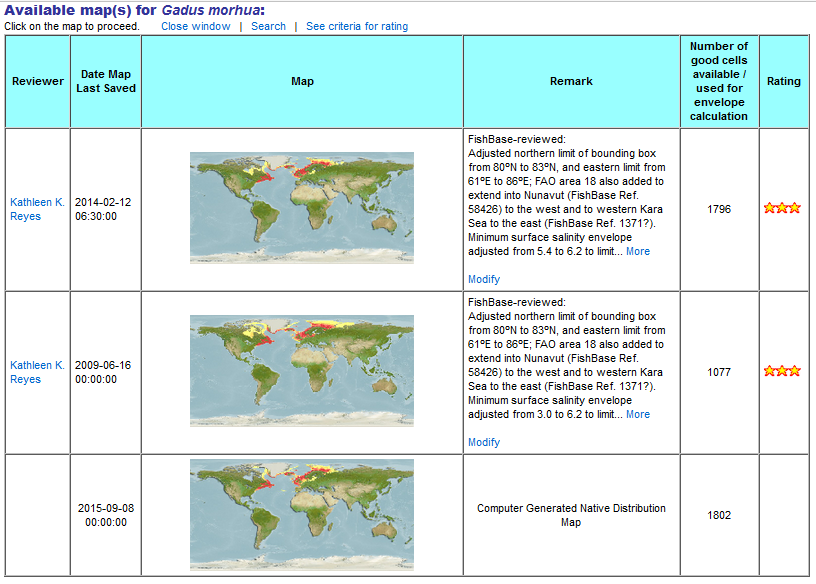
A map is typically reviewed by a species expert or a researcher examining *AquaMaps* predictions against a species’ known distribution. Doing a species search is the first step to checking and editing an AquaMap. Start from the *AquaMaps* Search page at [www.aquamaps.org](file:///C:\Users\Yoni%20Gavish\Documents\1-Work\Tree%20of%20Knowledge\5-EU-BON\General\Deliverables\WP3\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\UJ1I6CVI\www.aquamaps.org) (Fig. 2.2). Then:

1. Go to the section **Search Marine Species by Scientific Name** located below the map in the search page.
2. Specify the scientific name in the **Genus** and **Species** fields. The search accepts current accepted species names and synonyms. A sample search is shown for the Glacier lantern fish ***Benthosema glaciale****.*
3. Click **Search**. This will return a list with one or several records, depending on the scientific name specified in the search.
4. Click on the **Scientific name** of a species on the list. This typically calls the computer-generated native range distribution map for the species (Fig. 2.3a). (Note: Links above the map allow toggling to **Year 2100 Native Range, Suitable Habitat** and **Point Map**.)
5. However, if more than one map exists for the species, a list of maps is shown where the most recent reviewed map version is listed first and the default-computer generated version last. An example is shown for the Atlantic cod ***Gadus morhua*** (Fig. 2.3b). Click on the thumbnail of the map to open in full view.

 **Figure 2.2:** *AquaMaps* species search page, accessed via www.aquamaps.org



**Figure 2.3a:** Computer-generated native range distribution map returned by a map search for the Glacier lantern fish *B. glaciale*. Information on currently known distribution is found above the map while countries, FAO areas and ecosystems of reported occurrences can be accessed through the **More species data** section.



**Figure 2.3b:** Results of the species search for the Altantic cod *Gadus morhua*. The most recent reviewed map version is listed first and the default-computer generated version last.

*2.d.2: Checking and editing the map*

A basic check of an AquaMap involves comparing the predicted native distribution against a species' reported range. Information on the currently known native distribution of a species is provided above the map (Fig. 2.3a). Additional distribution information is also found in the **More species data** section below the map. This includes links to the list of countries, FAO fishery statistical areas, and ecosystems where a species is known to occur. Comparison can also be made by toggling to the **Point Map** which displays species occurrence records. A map may sometimes appear inaccurate due to incomplete or outdated input data, sampling biases, or data encoding errors. In such cases, a map may be corrected by editing mapping parameters and settings in the *CYOM*.

1. Click on the **Create your own map** link found beneath the species map (Fig. 2.3a). This open the *CYOM* tool interface (Fig. 2.4)
2. Examine mapping parameters and settings for the species. Note that the species' map can be improved or corrected by adjusting values or settings in four sections:

* Area Restrictions
* Occurrence Cells
* Environmental Envelopes
* Map View Settings

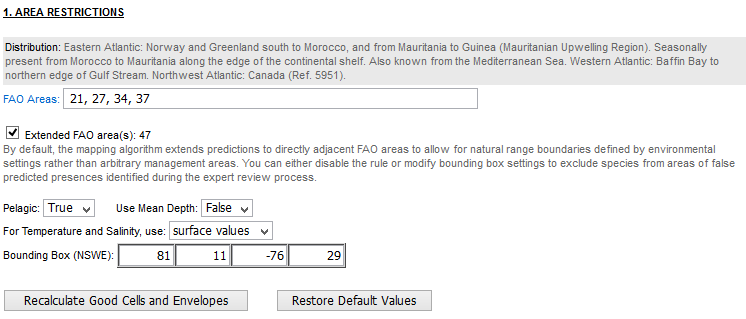
1. See **About AquaMaps** (upper right-hand section) for further information on the *AquaMaps* concept, algorithm and data sources (including other tools).



**Figure 2.4:** Create-Your-Own-Map (*CYOM*) interface showing default mapping parameters and settings for *B.* *glaciale*.

*2.d.3: Working with AREA RESTRICTIONS*

**AREA RESTRICTIONS** is the first section in the *CYOM* interface (Fig. 2.5). It describes the geographic extent of a species' distribution. The information is needed by *AquaMaps* to (1) identify half-degree cells that contain point data that fall within a species’ known natural range (“good cells”), (2) compute the species’ environmental envelopes, and (3) define the area for plotting the relative probabilities of species occurrence. Area restriction settings can be adjusted in any or all of the following sub-sections in Fig. 2.5.



**Figure 2.5:** Close-up of the AREA RESTRICTIONS section of the *CYOM* interface for *B.* *glaciale*.

FAO Areas (test field)

FAO areas are fisheries statistical areas that cover the known natural range of a species. They serve as a proxy for the bounding box in the absence of bounding box coordinates.

1. Review the listed FAO areas.
2. Add or remove FAO areas from the list by typing or deleting area codes, as necessary. FAO areas are two-digit codes representing subdivisions of the world’s oceans used by FAO for reporting fisheries data. (Appendix 2.1).

Extended FAO Area(s) (checkbox)

The model's algorithm extends predictions pole-ward to FAO areas directly adjacent to those listed above. This allowance enables natural range boundaries defined by environmental conditions to emerge, instead of being delimited by arbitrary management areas.

1. Review the extended FAO areas listed.
2. Retain the default setting, or uncheck the box to disable this rule to exclude species from areas of false predicted presences. Alternatively, you can also modify the bounding box settings (see below).

Bounding box (numeric field)

A bounding box defines the latitudinal and longitudinal extent of the species’ known natural range (e.g., based on a map or the literature).

1. Review the bounding box coordinates listed.
2. If necessary, change or complete the bounding box coordinates. The geographic coordinates are listed in the following order: northern limit, southern limit, western limit and eastern limit (i.e., N/S/W/E format).
3. Use whole degrees (although decimal-degrees is also an accepted format).
4. Use the negative sign (-) to indicate latitudes in the southern hemisphere or longitudes in the western hemisphere.

Pelagic (drop-down list)

The Pelagic flag indicates whether or not the distribution of a species is influenced by bottom depth. This information is used by the model when plotting the relative probabilities of species occurrences. “True” indicates the species is found in the water column, well above and independent of the bottom. “False” indicates the probability of occurrence depends on whether the bottom depth is within the depth range of the species.

1. Review the default Pelagic flag.
2. Retain the default setting or set the flag to “True” or “False”, as appropriate for the species.

Use Mean Depth (drop-down list)

The Use Mean Depth flag indicates how the probability of species occurrence with respect to depth is computed. “True” means the probability of occurrence will be based on a comparison of a species’ depth range to the mean depth of an area. “False” means it will be based on a comparison against the minimum and maximum depth of an area.

1. Review the default Use Mean Depth flag.
2. Retain the default setting or set the flag to “True” or “False”, as appropriate for the species.

For Temperature and Salinity, use: (drop-down list)

*AquaMaps* uses either surface or bottom data sets when computing the temperature and salinity tolerance limits of a species. By default, surface values are used when the minimum depth of a species <= 200m, while bottom values are applied to species of deeper minimum depths.

1. Review the default temperature and salinity layer used for the species.
2. Retain the default setting or change the setting to either “surface values” or “bottom values”, as appropriate for the species.

Recalculate Good Cells and Envelopes (command button)

1. If you made changes in any of the Area Restriction settings above, click this button to enable the model to recalculate the mapping parameters using the updated settings. A notification will be displayed when the good cells and envelopes have been recalculated.

Restore Default Values (command button)

1. Click this button if you want to discard the changes you made to the Area Restriction settings and reload the default map values. A notification will be displayed when the default system values have been loaded.

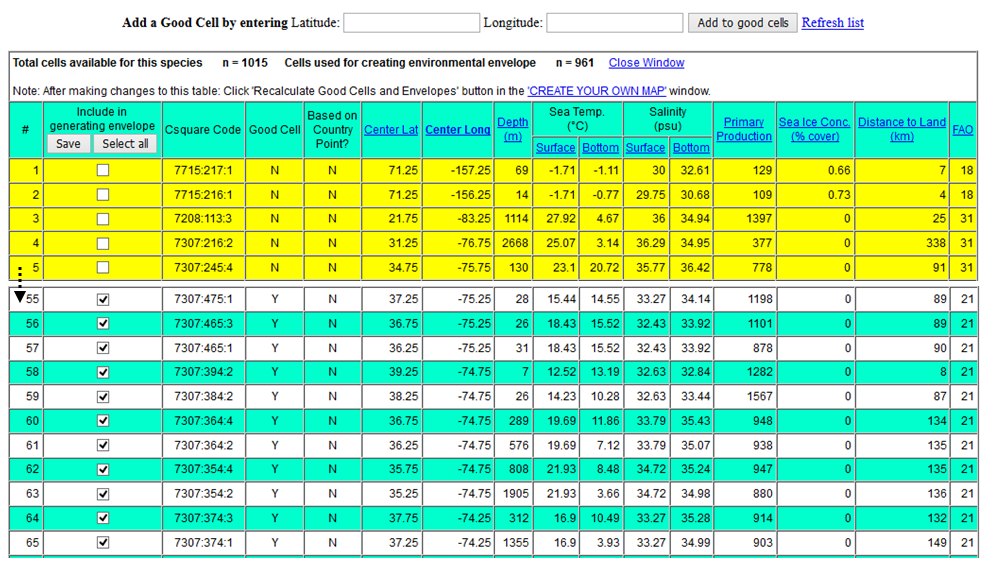
*2.d.4: Working with OCCURRENCE CELLS*

**OCCURRENCE CELLS** is the second section in the *CYOM* interface (Fig. 2.6). *AquaMaps* assigns species point data to a grid of half-degree cells that cover the world’s oceans. Each half degree cell has properties that describe the average depth, sea temperature, salinity, primary production, sea ice concentration, and distance to land of that cell. These are the environmental factors *AquaMaps* uses as predictors of species occurrence. Environmental properties in cells that contain point data within a species’ native range are used to estimate the environmental envelopes (environmental tolerances) of a species. Cells or point data can be included or excluded from the list, or even new ones added, in order to improve the set of environmental parameters from which environmental envelopes are computed. Note that half-degree cells are counted only once regardless of how many point data fall within them. This eliminates any bias from sampling frequency.



**Figure 2.6:** Close-up of the OCCURRENCE CELLS section of the *CYOM* interface for *B.* *glaciale*.

Click the link **Cells** **used for creating environmental envelope n =** (record count) to open the table of half-degree cell used for calculating environmental envelopes (Fig. 2.7).



**Figure 2.7:** List of occurrence cells for *B.glaciale*. Checked records indicate good occurrence cells based on point data found within the known range of the species. Records highlighted in yellow (unchecked) are treated as outliers. In this example, only 961 out of 1015 occurrence cells are considered good data for calculating the environmental tolerances (environmental envelopes) of the species. A user can add point data to the list and recalculate the environmental envelopes of the species.

**Total cells available for this species n = / Cells used for creating environmental envelope n = (record counters) - *Including/excluding occurrence data from the list***

By default, the table of occurrence cells shows the list of half-degree cells that contain point data attributed to a species based on data harvested from GBIF, and those in the FishBase and SeaLifeBase databases (Fig. 2.7). Cells tagged with a check mark contain point data that were used in computing the environmental envelopes of the species. Cells with point data that are out outside of the known distributional range are unchecked and highlighted in yellow. These are excluded from the computation.

1. Review the list of occurrence cells assigned to the species. Toggling to the **Point Map** in the species page (Fig. 2.3a) visualizes these in a color-coded point map indicating good and non-good cells.
2. Use the check boxes to include or exclude more cells from the list. You can also opt to retain the default settings on the list.
3. The summary line above table counts the number of available and used cells.
4. Click on the **Save** button when done selecting/unselecting cells from the list. (Skip this step if you did not make changes in the occurrence list.)
5. Click the **Close Window** link to return to the main *CYOM* page.
6. If you checked or unchecked cells from the list, remember to click the **Recalculate Good Cells and Envelopes** button back in the main *CYOM* interface to re-compute the species’ environmental envelopes based on the changes you made.

**Latitude/Longitude (numeric fields) - *Adding point data to the occurrence list***

Good cells are half-degree cells that contain point data within a species’ known distribution range. You can also add good cells by typing the geographic coordinates in the corresponding **Latitude** and **Longitude** fields found above the occurrence cells table (Fig. 2.7).

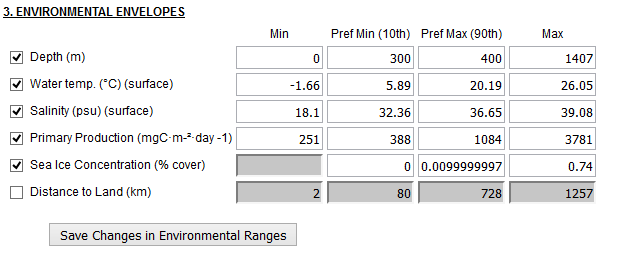
1. Enter coordinates in decimal degree format. Use negative values to indicate latitude in the southern hemisphere and longitude in the western hemisphere.
2. Click on the **Add to good cells** button and a window showing the coordinates you entered and its corresponding half-degree cell and cell properties is displayed.
3. Examine the corresponding cell properties, and click on the link **Add to good cells which will be used for prediction in the ‘Create Your Own Map’ routine** on the right to accept.
4. A dialog box will display "New cell added". Click **OK** to proceed.
5. Another dialog box will then display "In CREATE YOUR OWN MAP - You must now 'Re-calculate Envelope and Good Cells'. You can opt to click the check box to prevent this reminder from creating additional dialogs. Click **OK** to proceed.
6. Click the **Refresh list** link found above the occurrence cells table (upper right) when done. Note that the **Cells used for creating environmental envelope** record count now includes the point(s) added.
7. Click the **Close Window** link to return to the main *CYOM* page.
8. Click the **Recalculate Good Cells and Envelopes** button back in the main *CYOM* interface to re-compute the species’ environmental envelopes based on the changes you made.

*2.d.5: Working with ENVIRONMENTAL ENVELOPES*

**ENVIRONMENTAL ENVELOPES** is the third section in the *CYOM* interface (Fig. 2.8). An environmental envelope describes the range of tolerances of a species for a given environmental factor. These tolerances are presented as minimum (Min), preferred minimum (Pref Min), preferred maximum (Pref Max), and maximum (Max) threshold values. Environmental factors used by the model as predictors of species presence include depth, sea temperature, salinity, primary production, sea ice concentration, and distance to land. With the exception of depth, which is mostly based on the literature, species tolerance thresholds are computed using the average conditions in the half-degree cells enabled (checked) in the OCCURRENCE CELLS section.

Absolute and preferred minima and maxima thresholds are computed as follows:

* Min = 25th percentile - 1.5 × interquartile or absolute minimum in extracted data (whichever is lesser)
* Max = 75th percentile + 1.5 × interquartile or absolute maximum in extracted data (whichever is greater)
* Pref Min = 10th percentile of observed variation in an environmental parameter
* Pref Max = 90th percentile of observed variation in an environmental parameter



**Figure 2.8:** Close-up of the ENVIRONMENTAL ENVELOPES for *B. glaciale*, describing the estimated tolerance thresholds for the six predictors. (Distance to land is unchecked by default as this parameter mostly applies to marine mammals).

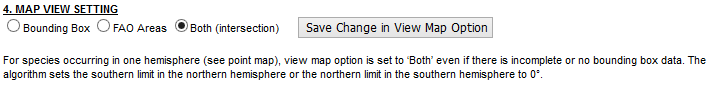
1. Review the environmental factors and threshold values in the species' environmental envelopes.
2. Retain the current settings, or manually change the threshold values and/or use the checkboxes to disable/enable environmental factors to use for predicting species occurrence.
3. If you made any changes, click on the **Save Changes in Environmental Ranges** button.
4. A dialog box will display indicating "Environmental ranges saved". Click **OK** to proceed.
5. A reminder to will show "You must now 'Regenerate Map Data' and then 'View Map' ". Click **OK** to proceed.

*2.d.6: Working with MAP VIEW SETTING*

**MAP VIEW SETTING** is the fourth section in the *CYOM* interface (Fig. 2.9). It determines how the predicted probabilities of species occurrence will be plotted on the map. There are three map view options:

* **Bounding Box** – the predicted probabilities of species occurrence will be plotted only in the area covered by the bounding box which is generally the closest approximation of the known native/endemic range of the species.
* **FAO Areas** – the predicted probabilities will be plotted to the limits of the FAO area(s) that encompass the known/endemic range of the species. This setting typically used when there is either incomplete or no bounding box data for the species.
* **Both (intersection)** – the probabilities of species occurrence will be plotted only over the area common to both bounding box and FAO area(s) of the species. This is the default setting.

Note: For species occurring in one hemisphere (can be verified by viewing the Point map), the map view option is set to ‘Both’ even if there is incomplete or no bounding box data because the algorithm sets the southern limit in the northern hemisphere or the northern limit in the southern hemisphere to 0°.

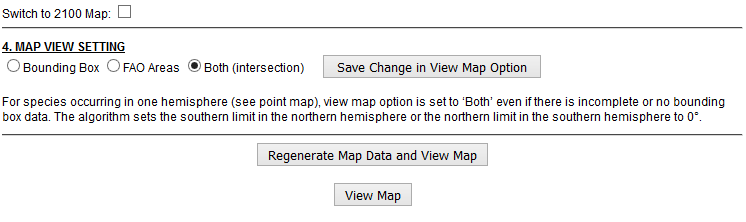
**Figure 2.9:** Close up of the MAP VIEW SETTINGS section for *B. glaciale* set to Both (intersection), indicating that the probabilities of occurrence will be plotted in the area where the both the species' bounding box and FAO area extents overlap.

1. Review the default map view setting for the species.
2. Retain this setting, or change the view setting by clicking on the appropriate radio button.
3. If you changed the setting, click on the **Save Change in View Map Option** button.
4. A reminder to will show "You must now 'Regenerate Map Data' and then 'View Map' ". Click **OK** to proceed.

*2.d.7: Regenerating an edited map*

This step enables you to re-draw the map, applying any changes you have made in the mapping parameters and settings following instructions in the previous sections of this manual. A map can be re-drawn for the predicted distribution range and probabilities of species occurrence in the current period and by the year 2100 using the command buttons at the bottom of the *CYOM* page (Fig. 2.10).

1. If you finished reviewing the mapping parameters and settings and have not made any changes in the *CYOM*, you can proceed to clicking the **View Map** button found at the bottom of the *CYOM* page. This will plot the same map seen in the species page. The resulting map however will be in interactive mode and will allow you to explore the map further.
2. If you have made changes in the *CYOM*, review your map settings to confirm all edits to be applied.
3. Click the **Regenerate Map Data and View Map** button, then click **OK** at the message prompt to proceed. This will re-draw the map of the predicted native range of the species at the current period. The map will be in interactive mode and will allow you to explore the map further.
4. To generate the predictive map for the year 2100, check the box to **Switch to 2100 Map**.
5. A dialog box will display indicating "You must regenerate map data when switching maps ". Click **OK** to proceed. (Check on the box to if you wish to disable this prompt.)
6. Click the **Regenerate Map Data and View Map** button to proceed. The resulting map plots the predicted native range of the species by the year 2100. It is also in interactive mode and will allow you to explore the map further.



**Figure 2.10:** Close-up of the bottom of the Create-Your-Own-Map (*CYOM*) page. Shown are the Switch to 2100 Map, Map View Settings, and the command buttons for regenerating and viewing a map.

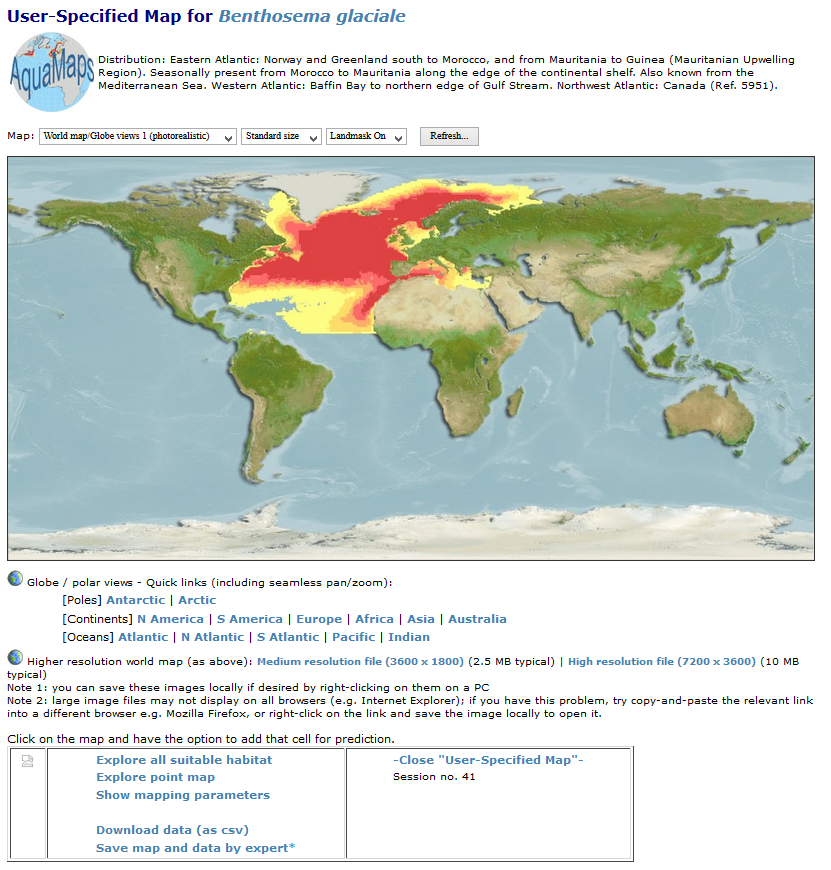
Note: Regenerated maps or maps in interactive mode are labeled **"User-Specified Map for <genus> <species>"**. See Fig. 2.11 for the example of the regenerated map for the Glacier lantern fish *B. glaciale*.

**Revisiting the example of the Glacier lantern fish *B. glaciale***

|  |  |
| --- | --- |
| **Problem with distribution map**   * Hard edges in the plotted distribution range *of B. glaciale* in the western central Atlantic **(A)**, and in the Barents Sea **(B)** |  |
| **Edits/actions applied in *CYOM***   1. Added FAO area 31-Atlantic, Western Central 2. Western and eastern bounding box limits adjusted to 81ºW and 61ºE (to extend further along US east coast & eastward in Barents Sea) 3. Good cells and environmental envelopes recalculated 4. Minimum sea surface salinity threshold increased to 18.4 (to exclude false prediction in Black Sea) 5. Changes in environmental ranges saved 6. Map regenerated |  |
| **Regenerated/Edited Map**  Regenerated native range AquaMap for *B. glaciale*,incorporating edits and adjustments in mapping parameters to improve the predicted distribution for this species |  |

*2.d.8: Saving and publishing an edited map* The distribution map regenerated in the previous section now incorporates the data and/or edits in mapping parameters and settings behind the reviewed map (see example for *B. glaciale* in Fig. 2.11). If the regenerated map meets the approval of the reviewer, the map should now be saved so that this improved version and its associated data and settings can be published and stored in aquamaps.org. Reviewed maps are listed along with the latest computer-generated *AquaMaps* for a species. If a reviewed map exists for a species exists, it is displayed as the default species distribution map in the FishBase and SeaLifeBase Species Summary pages. Otherwise, the default computer-generated map is displayed.

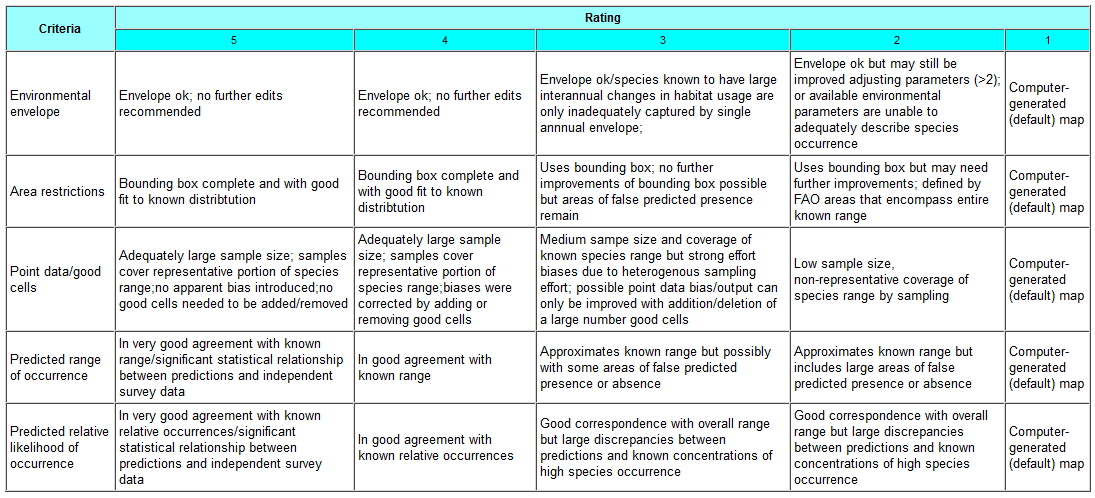
Note that prior registration with *AquaMaps* is required in order to save and publish an edited map in aquamaps.org. Contact Rainer Froese ([rfroese@geomar.de](mailto:rfroese@geomar.de)) for fishes, and Ma. Lourdes Palomares ([m.palomares@fisheries.ubc.ca](mailto:m.palomares@fisheries.ubc.ca)) for non-fish species. An assigned activity password, ExpertID and user password will be provided when the registration has been completed.

 **Figure 2.12:** Regenerated map of *B. glaciale* based on data and edits specified by a reviewer. Link for saving map and associated data encircled in red.

1. Click on **Save map and data by expert\*** (Fig. 2.12).
2. A window will appear. Enter the **activity password** and click Submit.
3. A form for saving a map will appear. Enter your **ExpertID** and **password**.
4. Enter brief notes to document edits/changes applied to the map in the **Remarks** field. Recommended contents could include:

* Problem(s) with the previous version of the map
* Action(s) taken/edits made to improve the map
* References used as basis for corrections made, if any
* Other important comments or notes

1. Give the map a star **Rating**. (See Fig. 2.13 for rating criteria; also Appendix 2.2).
2. Click **Save Expert Map**.
3. The map is saved and now available in aquamaps.org. You will be asked if you would like to announce the completed review through various channels. This step is optional. If you click **Yes**, an announcement of the reviewed map will be posted in the *FishBase* / *SeaLifeBase* and *AquaMaps* Facebook pages, as well as in the EU BON Biodiversity Portal. Click **No**, if you only want to save and publish the map in aquamaps.org without sending out reviewed map alerts.



**Figure 2.13:** Star Rating scheme to guide a reviewer in evaluating the reliability of an *AquaMaps* native range prediction for a given species.

Note: A reviewed/edited map will not necessarily correspond to all conditions under each criterion within a given star rating, and will most likely vary across star ratings with respect to the different criteria. Thus, these criteria for rating are best used as a guide to approximate the degree of reliability of the predicted species distribution in the reviewed/edited map. The final rating is thus left to the discretion of the reviewer.

3: *rAquaMaps* Global Modelling Tool

***3.a: Aim***

To implement the *AquaMaps* algorithm and database in the open source R environment

3.b: Introduction

Species distributions represent the combined effect of historical and ecological factors. Modelling distributions based on environmental data only will show suitable habitats rather than actual distribution which must also take into account physical barriers to dispersal. Species modelling is also dependent on the sampling density and evenness. The *AquaMaps* modelling concept combines spatial and ecological parameters with an option for expert adjustments ([Ready et al. 2010](#_ENREF_26)) (#1632). It also permits direct use of large occurrence databases (GBIF, FishBase) by filtering taxon names through a validator. The power of *AquaMaps* is particularly in predictions of occurrence at a large scale and using sparse distribution records. By implementing *AquaMaps* in R we will make *AquaMaps* more versatile and compatible with other R modelling approaches including the possibility of incorporating algorithms other than the native *AquaMaps* algorithm, but particularly providing greater speed and user programming options for *AquaMaps*. *AquaMaps* modelling is an important tool particularly for large scale modelling with few data points and can be used with alternative environmental layers such as IPCC projections.



**Figure 3.1:** Screenshot of *rAquaMaps* result: probability of occurrence of Atlantic cod *Gadus morhua* in the North Atlantic.

3.c: Approach

*AquaMaps* uses known occurrences along with several environmental predictors to identify the environmental preferences (or limits) of the taxon. These preferences can then be used directly or modified by users/experts to define the environmental envelope of the taxon for each predictor. Based on the environmental envelope, the taxon’s probability of occurrence is presented in a layer of half degree cells covering the Earth. *AquaMaps* relies on open source data: distribution data are taken from GBIF and the environmental parameters are adapted from free global data-sets. A concern with *AquaMap*s has been relatively slow computation time and limited compatibility with open-source spatial software, both relating to the database format. *rAquaMaps* – an R package based on the published *AquaMaps* algorithm and associated datasources – reduces computation time significantly and is built in an open framework, permitting maintenance and development even after the EU BON project. The *rAquaMaps* package permits advanced users great flexibility in modelling, e.g., modelling a single species or a big batch of species, using own input data, testing with optional environmental parameters, etc. *rAquaMaps*implements the *AquaMaps* algorithm used to build an environmental envelope for each species, and uses web services and data harvesting tools implemented in R (e.g., *rGBIF*) to map species distributions and build environmental layers. The *rAquaMaps* package can be built into other applications, run as a server application or as an independent desktop application. This deliverable also includes a complete web application using Shiny.

3.d: Main functionalities of the R package

*rAquaMaps* includes the following components: (a) A global indexed grid of half or quarter degree cells based on the c-squares system of hierarchical numerical identifiers; (b) environmental predictors (e.g., temperature, precipitation, salinity, etc.) for each cell; (c) occurrence frequency for species in each cell; (d) known environmental envelope for each species calculated on the basis of occurrence data (environmental values in cells of occurrence) or using expert information. For each taxon analysed, an envelope is calculated for each environmental predictor, which provides the upper and lower tolerance limits for the taxon. The probability of occurrence is then calculated for each predictor in each cell and may range from 0 to 1. The individual predictor probabilities are then multiplied to give the prediction of occurrence in the cell.

*rAquaMaps* can show actual distribution (point data), projected distribution using the probabilities of occurrence (suitable habitat), or modelled actual distribution using a combination of probabilities and a system of bounding polygons that constrain the distribution to known areas of occurrence. The bounding polygons may be based on Fishery zones (FAO areas), watersheds or other geographical structures that limit taxon distribution in addition to ecological niche, or expert data. By adjusting the environmental data to IPCC models, *rAquaMaps* can also present projected distributions in the future according to global or regional warming scenarios. The tool can also be used for species richness maps, either using raw occurrence data or models showing constrained or absolute suitable habitats. Because *rAquaMaps* has probabilities of occurrence as end product, the graphic system and the underlying statistics tool in *rAquaMaps* can also be used to display results obtained with other modelling tools. *rAquaMaps* works well with marine areas (Fig. 3.1), and has also been used to model continental scale terrestrial and freshwater distributions. The strength of the system is models of large scale patterns using low density datapoints, e.g., at European level, and an obvious use is for evaluating potential invasive species.

3.e: Versions, installations, guides and recent changes

Current releases of *rAquaMaps* are continuously built using Travis CI and made available with install instructions at <https://github.com/raquamaps>. Manuals and documentation are provided for all functions and three vignettes are included – these are instructive tutorials that are provided within the package and gives elaborate usage examples. The three vignettes cover:

1. [raquamaps-overview.Rmd](https://github.com/raquamaps/raquamaps/blob/master/vignettes/raquamaps-overview.Rmd) – Overview of the package.
2. [raquamaps-intro.Rmd](https://github.com/raquamaps/raquamaps/blob/master/vignettes/raquamaps-intro.Rmd) – introduction to *rAquaMaps*.
3. [raquamaps-usage.Rmd](https://github.com/raquamaps/raquamaps/blob/master/vignettes/raquamaps-usage.Rmd) – various usage examples covering some common use cases.

In addition, we provide in Supporting Information SI3.1 the package’s help files, with clear description of each function and examples.

3.e: Web-enabled usage of package features

The development branch of the repository includes a web user interface similar to the AquaMaps.org Create Your Own Map feature, which runs using Shiny, the web application framework for R. After installing the package, this user interface can be launched with a single command. Also, installation can be completely avoided, since the package can be distributed as a no-install web-enabled platform using <https://github.com/raquamaps/mirroreum> which provides a packaged system platform that runs anywhere using Docker (this solution extends <https://hub.docker.com/r/rocker/ropensci>, please find details here: https://github.com/rocker-org/rocker/wiki). Note that any other R packages can be added to this setup in order to provide a similar web-enablement. The only requirement for use then becomes an up-to-date web browser.

4: Hybrid Species Distribution Models

4.a: Aim

To predict species distribution at various grains and extents, while accounting for both environmental and spatial aspects.

4.b: Introduction

Fine scale occurrence data is widely available for many species, either from monitoring programs or from repositories (e.g., GBIF). Despite the fact that only a small fraction of the species’ distributional range is sampled at fine resolution, the occurrence data can provide information on the species distribution via two main methods. First, the fine-scale occurrence data can be coarsened to a large grain size, thereby producing atlas maps. Alternatively, species distribution models (SDMs) that relate the species’ known presences (and absences) to environmental data can be made, thereby producing a fine-scale probability of occurrence (PoO) map.

The advantage of the first method is that each coarse grain cell accumulates information from many potential fine scale samples, such that the probability of correctly assigning a cell as presence or absence increases. On the other hand, in many cases atlas data is too coarse to provide valuable information at a resolution relevant to conservation and management. Downscaling models that predict the proportion of occupied cells at fine resolution based on geometrical scaling properties of the coarse scale distribution ([Kunin 1998](#_ENREF_22), [Hui et al. 2006](#_ENREF_17), [Azaele et al. 2012](#_ENREF_4), [Barwell et al. 2014](#_ENREF_6)) (#1177, #1362, #1031, #764) may perhaps bridge the gap back to finer scales by modelling the occupancy-area relationship (OAR): as grain size increases area of occupancy also increases. Once a model has been fitted to the coarse-scale data it may be extrapolated to predict the total area of occupancy of a species at the fine grain sizes of the SDM. However, these downscaling models do not account for environmental aspects and provide information only on the number of occupied cells at fine resolution, but not on their locations.

Alternatively, SDMs are widely used in conservation and management to extrapolate from known occurrences to unsampled locations, based on the species environmental preference at a fine scale. However, there remain several issues that are unresolved in SDM application. First, SDMs ignore the effect of biotic interactions on the predicted distribution (but see: [Heikkinen et al. 2007](#_ENREF_15), [Baselga and Araújo 2009](#_ENREF_7), [Pellissier et al. 2010](#_ENREF_24), [Guisan and Rahbek 2011](#_ENREF_14), [Calabrese et al. 2014](#_ENREF_9), [Trainor and Schmitz 2014](#_ENREF_28), [D'Amen et al. 2015](#_ENREF_10)) (but see: #1579, #1639, #1536, #1447, #1561, #1640, #1460). Second, spatial biases in sampling locations and variation in detectability rates can lead to unreliable PoO maps, and if these same biases exist in testing data may also lead to unreliable model evaluation ([Guillera-Arroita et al. 2015](#_ENREF_13)) (#1586). Third, SDMs ignore the spatial aspects important for structuring a species distribution that are independent of environmental filtering ([Bahn and McGill 2007](#_ENREF_5)) (#1183). For example, SDMS cannot identify as absences areas of high environmental suitability that remain unoccupied due to dispersal barriers or due to historical effects such as disturbance history. Furthermore, in SDMs each cell is modelled separately and the probability with which it is associated is independent of its neighbourhood, while in reality a cell of certain suitability is more likely to be occupied if surrounded by high suitability cells than by low suitability cells due to dispersal and metapopulation dynamics. Finally, even in absence of sampling bias and detectability issues, SDMs produce maps of probability values and not binary presence /absence (P/A) maps. Translating the PoO values to a binary map requires selecting a threshold, such that PoO larger or smaller than the threshold are converted to presences and absences, respectively. However, there are a plethora of different ways to select a single global threshold, with little consensus over the suggested ones ([e.g., Liu et al. 2005](#_ENREF_23)) (e.g., #1444), although in general a threshold that maximises accuracy to the training data is preferred. Interestingly, once a threshold is selected, the resulting presence/absence map can be up-grained to any resolution, thereby creating a predicted OAR comparable to those created by downscaling atlas data.

As part of task 3.3 of WP3, UnivLeeds aimed to develop enhanced SDMs methods that will simultaneously account for spatial aspects and species habitat preferences. In the next section we first introduce the main idea behind four different hybrid models. Thereafter, we dedicate a section to the breeding bird dataset from Wallonia (in collaboration with EBCC) and to the yellow wagtail (*Motacilla Flava*), which serves as a case study in all the examples. This is followed by four chapters, one for each hybrid model, in which we more thoroughly describe the rationale of the model, its R application and the results obtained for *M. Flava*. We end with sections that compares the different models and final conclusions.

4.c: Approach

As mentioned above, UnivLeeds developed four hybrid models, differing from one another in several aspects (Table 4.1). All models are made available as R functions for ease of application. Each model starts with predicted PoO from an SDM at a fine scale, however each incorporates information at coarser spatial scales. The models differ in the relative emphasis between the fine-scale SDM and the multi-scale spatial information, and the method of thresholding to convert the PoO map to a P/A map. The first model, the *Moving Window SDM****,*** estimates the mean PoO at increasingly large window sizes around every cell. Then, a second SDM is fitted while including the mean PoO at the different windows sizes as additional explanatory variables along with the environmental variables or the original PoO. Thus, the model adjusts the local PoO values according to the habitat suitability context at the landscape scale. In this model, the PoO map is translated to P/A map by selecting a single global threshold that optimize the performance of the model at fine-scale. The model makes no use of information from downscaling models.

Unlike the *Moving Window SDM*, the three additional models incorporate additional information from downscaling models. The simplest of them is the *Top X* model. The model makes no adjustment of the predicted probabilities. but simply identifies the threshold that will create the number of occupied cells predicted by the downscaling model at the SDM resolution.

The third hybrid model, the *TopDown PoO* model, incorporates information from the downscaling models at multiple scales. Although the model does not modify the original SDM’s PoO at the fine-scale, it relies on the mean PoO at larger grain sizes. The algorithm starts at the coarsest resolution, and identifies which *xi* cells have the highest mean PoO (*xi* being the number of cells predicted to be occupied at scale *i* from the downscaling models). Then, the algorithm jumps to the fine-scale SDM resolution and within each selected coarse-scale cell, the fine-scale cell with the highest PoO is assigned as occupied. The occupancy status at all larger resolutions is then updated and the procedure repeated for the second-largest scale and so-on for all scales. Thus, the output of the model is a P/A map at the SDM resolution, which contains the exact number of occupancies according to the downscaling model at all resolutions. Since the *TopDown PoO* model relies on the mean PoO at various resolution, it should account for landscape context affect in a similar way as the *Moving Window SDM* but also for the shape of the OAR.

The fourth model, the *SpaNiche* model, aims to select a threshold that maintains accuracy at fine resolutions, while remaining spatially-consistent with the spatial patterns at coarse-grain size (i.e., at scales in which we have less uncertainty regarding the ‘true’ distribution). It does not modify the original SDM PoO map. The model takes as input the PoO map and the predicted occupancy at various scales from downscaling models of the coarse-scale atlas data. First, the model applies various threshold values to the PoO maps to generate P/A maps at the SDM scale. Each generated P/A map is then evaluated in two ways: a) we estimate the fine-scale accuracy against the test data. ;; and b) we assess the ‘spatial consistency’ by converting them to multiple larger scales to create an OAR and comparing them to the OAR of the downscaling models. Fine-scale accuracy is then plotted against spatial consistency for all thresholds, and the threshold that yields an optimal balance between the two is selected.

Among the four hybrid models developed in tasks 3.2, three make use of coarse resolution atlas data, while assuming that it contains reliable absence data. If indeed the atlas data is reliable, then masking the PoO with the atlas data by setting the probabilities of any fine-scale cell that falls within an atlas scale absence to 0 should increase the models performance by removing false presences. Thus, for each of the four models and for the original SDM we also explored the effect of atlas masking, for a total of nine hybrid models (table 4.1).

Most models covered in this section rely on the ‘*downscale*’ R package, developed by UnivLeeds under task 3.2 of WP3. Although this package is described in details in deliverable D3.1, we note here that it covers 10 leading downscaling models along with an ensemble model. The package also includes an upgraining function to create atlas data from occurrence data, diagnostic tools for upgraining, and additional plotting and predicting functions. The package is available for download from CRAN and contains 2 vignettes on top of the regular R help files. Deliverable D3.1 contains a detailed manual on the package.

**Table 1:** Main properties of the hybrid models developed by UnivLeeds under task 3.3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Moving Windows** | **Top X** | **TopDown PoO** | **SpaNiche model** | **Masked-SDM** |
| Modify SDMs PoO map? | yes | no | yes | no | no |
| Utilises downscaling predictions | no | yes | yes | yes | no |
| Thresholding method | Single threshold that optimizes fine-scale accuracy | Threshold that matches the predicted prevalence | No threshold | Single threshold that balances fine-scale accuracy and spatial consistency | Single threshold that optimizes fine-scale accuracy |
| Number of occupied cells in P/A output | Threshold dependent | As predicted by the downscaling models for the finest resolution | As predicted by the downscaling models at all resolutions | Threshold dependent | Threshold dependent |
| Atlas masking | optional | optional | optional | optional | inherent |
| R functions | Win\_PoO.R | TopX.R | TopDown\_PoO.R | SpaNiche.R | TopX.R |

4.d: Case study

*4.d.1: The Wallonia breeding bird dataset*

In collaboration between EBCC and UnivLeeds, two extensive datasets for birds from Wallonia (Belgium) where made available for exploration of hybrid models. The first data set contained transect data at 1×1 km scale, which was used as the main modelling dataset and for accuracy assessment at the local scale. The second dataset is an independent atlas dataset at 5×8 km scale, which was used to assess the models’ performance at coarser resolution. As far as we know, this is the first attempt to explore SDMs accuracy at more than one scale. The transect data contains 2800 cells, covering approximately 17% of Wallonia. Each transect was sampled at least twice a year with two 1 hour long transects- one in the morning and one in the evening. The atlas data contained 514 cells. The presence of all species in each cell was assessed by multiple visits, each covering all main habitats in a collaborative effort of approximately 650 volunteer fieldworkers. A more detailed account on the sampling protocols and metadata can be found in [Aizpurua et al. (2015)](#_ENREF_1) (#1520) and the citations within.

Although we have applied all the models to all 87 species with at least 50 occurrences in the transect data, here we report on a single species as a case study.to illustrate the methods developed in WP3 rather than the results. We note, however, that a full report on all 87 species, which currently include around 34,800 analyses, is likely to be included in deliverable D4.1 of WP4. Here, we for illustration purposes we have selected the yellow wagtail (*M. flava*) that occurred in 256 of the 514 atlas cells (Fig. 4.1a) and in 577 of the 2800 transect cells (Fig. 4.1b) since:

1. Its basic SDM performed relatively well (fine resolution accuracy: TSS ≅ 0.717).
2. There was a good overall spatial consistency between the transect data and the atlas data (coarse resolution accuracy: TSS ≅ 0.714)
3. Its results seem to qualitatively represent the results we obtained for other species.

*4.d.2: Explanatory variables*

In addition to the transect data and the atlas data, *EBCC* also supplied various environmental variables at a 0.2×0.2 km resolution. This data was aggregated to a 1×1 km resolution using bilinear interpolation with the *resample* function of the R package ‘*raster*’. Categorical values were translated to multiple binary variables, such that the bilinear interpolation yielded a value closely related to the relative cover of the class. The explanatory variables contained topographic variables, climatic variables, land cover and habitat variables and soil type variable. A full list of variables is given in table 4.2.

**Table 4.2:** Explanatory variables used to model the distribution of bird species from Wallonia.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Category | Variable | Description |
| 1 | topographic | AVGALT | Average altitude |
| 2 | AVGORI | Average orientation |
| 3 | AVGSLO | Average slope |
| 4 | TMI | Topographic moisture index |
| 5 | climatic | PAMJ | Mean precipitation April-May-June (mm) |
| 6 | TAMJ | Mean temperature April-May-June (Celsius) |
| 7 | Habitat / land cover | URBA | Surface of urban areas (ha) |
| 8 | SSTW | Surface of standing water (ha) |
| 9 | ORCH | Surface of orchards (ha) |
| 10 | WETL | Surface of wetlands (ha) |
| 11 | PERG | Surface of permanent grassland (ha) |
| 12 | TMPG | Surface of temporary grassland (ha) |
| 13 | NATG | Surface of natural grassland (ha) |
| 14 | SBF | Surface of broadleaved forest (ha) |
| 15 | SCF | Surface of coniferous forest (ha) |
| 16 | SMF | Surface of mixed forest (ha) |
| 17 | SPRCER | Spring cereal (ha) |
| 18 | WINCER | Winter cereal (ha) |
| 19 | FODCUL | Forage culture (fodder) (ha) |
| 20 | SSHCUL | Spring summer hoed culture (ha) |
| 21 | LLVS | Length of linear vegetation structures (m) |
| 22 | NPVS | Number of punctual vegetation structures |
| 23-43 | Soil | SOIL\_CAT | 20 variables, each a binary with the dominant soil type |

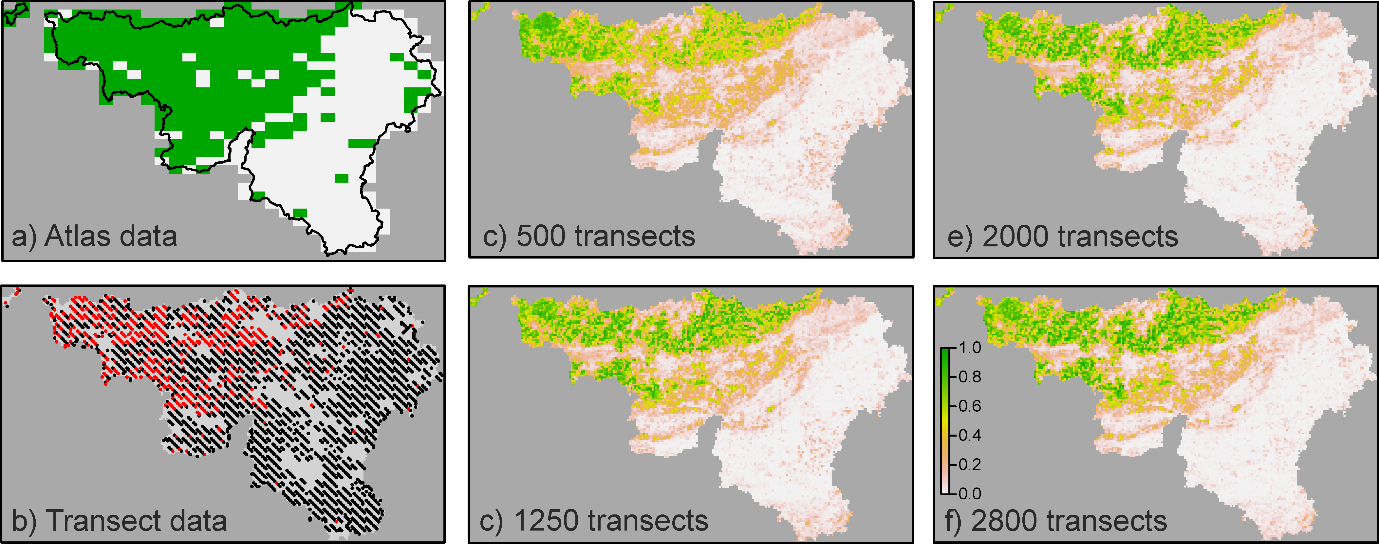
*4.d.4: Data analysis*

Basic SDM

We based the SDM analyses on the randomForest algorithm ([Breiman 2001](#_ENREF_8)) (#1188). RandomForest is a machine learning method designed to produce accurate classification of multiple cases (e.g., grid-cells) to predefined classes (e.g., presence or absence). Each application of randomForest takes as input a set of training cases (e.g., known presences and absences) along with relevant explanatory variables. The algorithm then ‘learns’ the rules by which the explanatory variables can distinguish between the different classes. The learning procedure is based on fitting a ‘forest’ of classification trees, with each tree being based on a different subset of the original training cases (“in-bag” data) and each branching in any classification tree using only a small and random subset of the available list of explanatory variable. Therefore, each classification tree is unique, and can provide independent prediction for each case that was not included in tree growing (i.e., “out-of-bag” data). Thus the output of randomForest is a list containing the predicted class for each out-of-bag training case according to each tree (known as a vote). The results are then translated into PoO, by estimating for each case, the proportion of out-of-bag votes that assigned the case to any of the classes. The reliance on out-of-bag votes (i.e., only trees in which the case was not in the in-bag subset) results with a method robust to over-fitting. The randomForest algorithm also returns a variable importance value for each explanatory variables, based on the decrease in model accuracy when the variable is permutated. Here, we used the R package ‘*randomForest*’, keeping the default parameters values with the exception of the number of classification trees that was set to 10,000 trees.

Random datasets

To examine the robustness of the hybrid models to data quantity we ran the SDM and hybrid models on all 2800 transects (16.8% of total area), as well as randomly generated subsamples of 2000 (12%), 1250 (7.5%) and 500 (3%) transects. We repeated the analysis five times in each case. The effect of sample size is likely to affect all aspects of hybrid modelling, including the SDMs themselves, the accuracy of the atlas data and the accuracy of the downscaling models. An example of the PoO generated by one of the randomForest runs for each sample size is given in Fig 4.1 c-f. Throughout this deliverable we use the results of the dataset S1250\_R1 (i.e., first run with 1250 random transects) in all examples.

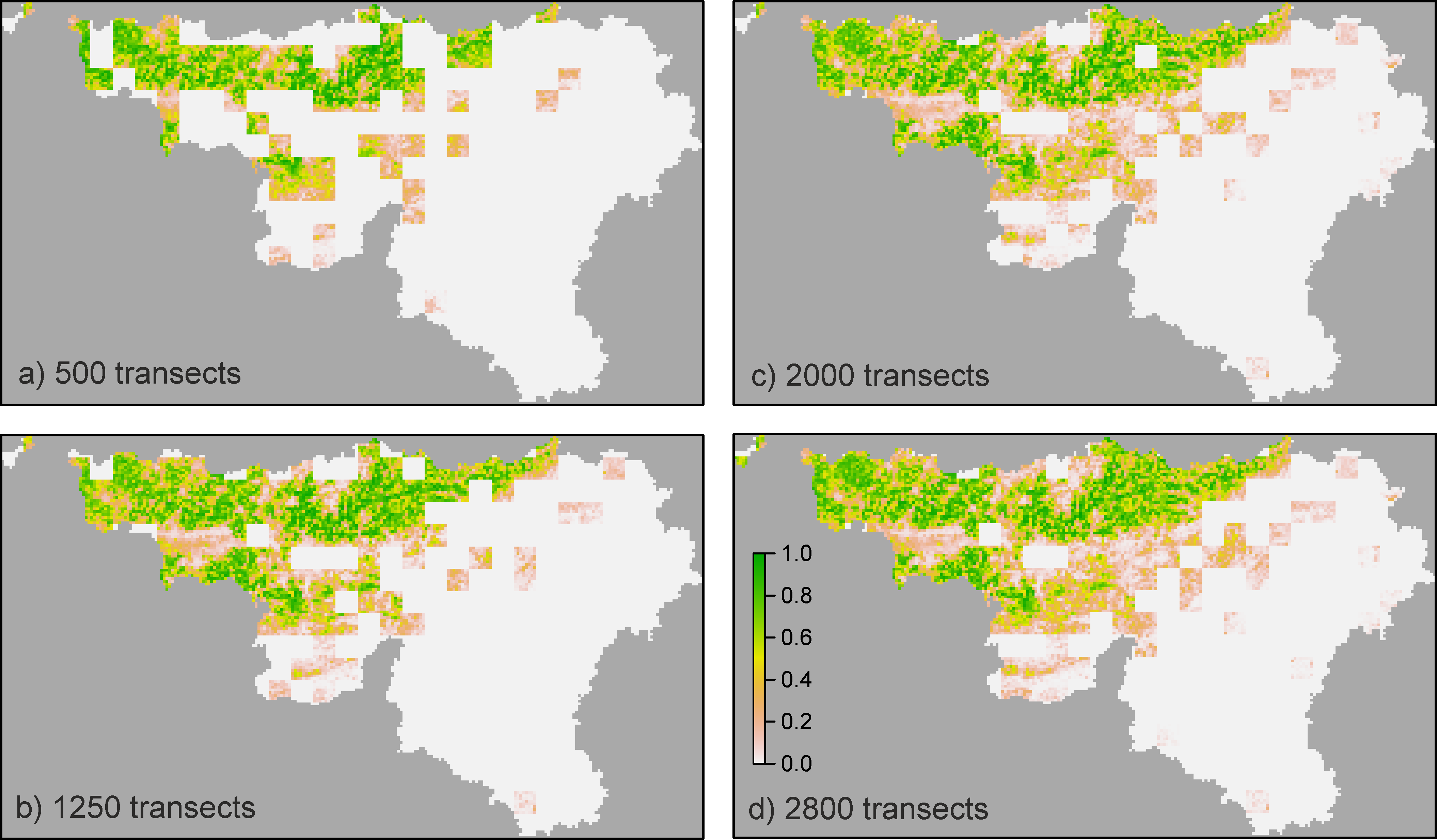


**Figure 4.1:** The distribution and modelled probability of occurrence (PoO) of *M. flav*a in Wallonia. ***a)*** Presence (green) / absence (white) in the 5×8 km atlas data. ***b)*** Presence (red) /absence (black) in the 1×1 km transect data. The PoO according to a randomForest model for samples sizes of ***c)*** 500 transect, ***d)*** 1250 transects, ***e)*** 2000 transects and ***f)*** all 2800 transects.

Atlas scale and SDM masking

We created atlas data from the transect data at 8×8 km resolution using the *upgrain* function of the R package ‘*downscale*’ (using the ‘*All\_Sampled*’ option). This resolution was used since it is the closest we could get to the 5×8 km resolution of the external atlas data while keeping full nestedness of scales at other resolutions (i.e., 1×1, 2×2 and 4×4). When all 2800 transects are considered, at this resolution 98% of cells contained at least one transect, 92% of cells contained at least two transects, and 90% at least three transects. Sample size has a strong effect on the distribution pattern at the atlas scale (Fig. 4.2); with smaller sample sizes there are a greater number of false absences in the atlas data. The atlas data was used for downscaling and also to mask the output of the SDMs (*Masked-SDM*) as well as each hybrid model.

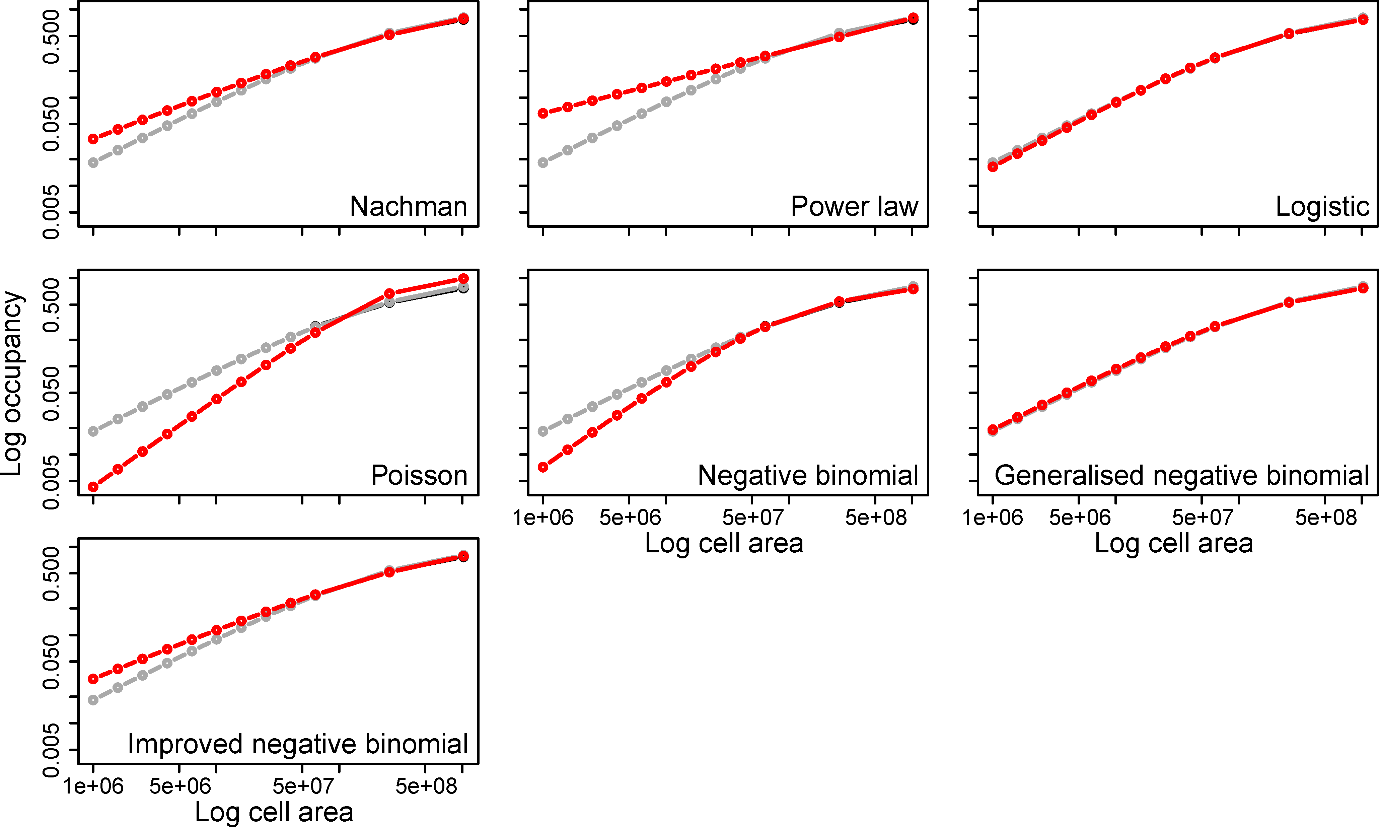
.



**Figure 4.2:** The predicted SDM PoO with ***a)*** 500 transects ***b)*** 1250 transects, ***c)*** 2000 transects and ***d)*** 2800 transects masked by upgrading the training data to a 8×8 km resolution and setting the PoO to zero in all cells that are absent at the atlas scale. Note the difference from Fig 4.1c-f.

Downscaling models

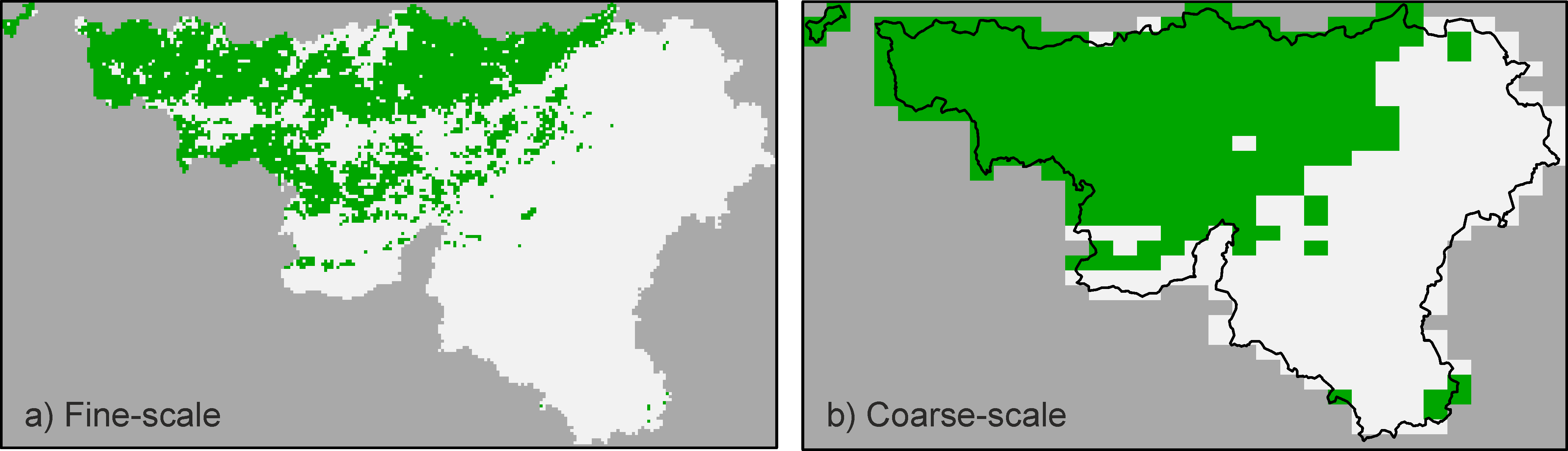
The R package ‘*downscale*’ covers 10 different downscaling models (see deliverable D3.1 for a full list and details on each model). Throughout the analysis we have used the ensemble predictions of 7 models that extensive analyses of approximately 1000 species (carried out by partners from *MRAC*) proved both accurate and time efficient. The seven models include: the Nachman, power law, logistic, Poisson, negative binomial, generalised negative binomial and improved negative binomial. We used the default starting values for all models with the exception of the generalized negative binomial (C=0.001, z=0.1 and k=0.01) and improved negative binomial (C=1, r=1 and b=2). An example of the predicted occupancy at various grain sizes of each of the seven model is given in Fig. 4.3.



**Figure 4.3:** The predicted occupancy at various grain sizes according to the 7 downscaling models used in all examples. The red lines are the predictions for each model, and the grey lines the predictions for the ensemble results, i.e., the mean over all seven models. The 3 largest grain sizes were used to fit each curve. Results from the S1250\_R1 dataset.

Evaluation of model performance

Although various indices to evaluate the performance of SDMs are available, we focused here on the True Skills Statistic ([TSS, Allouche et al. 2006](#_ENREF_2)) (TSS, #1641) calculated as the sensitivity + specificity – 1. We have chosen TSS since it is robust to difference in species prevalence, thereby allowing comparison of values among different species. For the models that provide a PoO map as output (the original SDM, *Masked-SDM*, and *Moving Window SDM*), we created 100 P/A maps, using 100 equally-spaced thresholds between 0 and 1. We calculated the TSS for each threshold using the training transect data (e.g., the 500 transects). We then selected the threshold that maximized the TSS and applied it over the entire extent. After that, we estimated the test data TSS by recalculating the TSS with the selected threshold for the set-aside test data (e.g., 2300 set-a-side transects for the 500 transects datasets). Off-course, we could not estimate the test TSS for the 2800 transects datasets, as all transects were included in the training set. Thus for the 2800 we can only compare the accuracy of the training set. The predicted P/A was upgrained to the 5×8 km resolution, where again TSS was calculated against the external atlas data. For example, for the S1250\_R1 dataset, the threshold that returned the maximal fine scale TSS against the training data (0.73) for the original SDM was 0.2, resulting with 5587 occupied cells at the 1×1 km resolution (Fig. 4.4). Applying the same threshold on the test data yielded a TSS value of 0.696 at the fine scale and the overall predicted P/A map had a TSS of 0.735 against the external atlas data. For the hybrid models that return a P/A table (*Top X*, *TopDown PoO* and *SpaNiche* models), we simply estimated the SDM scale TSS against the training and test data, upgrained the P/A map to the 5×8 km resolution, and estimated the TSS against the independent atlas data.



**Figure 4.4:** The predicted presence (green) / absence (white) map at ***a)*** the 1×1 km and ***b)*** the 5×8 km resolution according to the original, unmasked SDM. Results from the S1250\_R1 dataset.

4.e: Moving Windows SDM

*4.e.1: Main rationale*

In the context of SDMs, dispersal is expected to affect the distribution of species in two main ways. First, some sites that contain suitable habitat will remain unoccupied if they are too far from other suitable habitats. Second, unsuitable sites that are in great proximity to large patches of suitable habitat may be occupied due to continuous inflow of dispersing individuals (i.e., propagule rain, rescue effects and sinks population). When focusing on the PoO map, the cells representing the first case may be characterized by high local Poo but low mean PoO in the surrounding landscape. A cell from the second case may have the opposite pattern- low local scale PoO and high mean PoO in its surrounding landscape. Thus, incorporating in an SDM information on habitat suitability at various scales may allow the model to pick up these cases. The *Moving Windows SDM*does so by calculating the mean PoO in windows of various sizes around each local cell. The mean PoO are then added as additional explanatory variables to a new SDM.

*4.e.2: R application*

As part of task 3.3, we codified an R function which estimate the mean PoO value for each cell in each user-defined window sizes. The function, named *Win\_PoO.R* takes as input the parameters listed in table 4.3. Information on the output is given in the table as well.

**Table 4.3:** the input parameters to the *Win\_PoO.R* function provided in the supporting information.

|  |  |
| --- | --- |
| Parameters | Description |
| ***Input*** | |
| *Stack* | Object of class ‘*RasterStack*’ that contain two layers- cell IDs in the first layer and the PoO values in the second layer. |
| *Win\_Size* | Vector of odd integers with the windows size. The values represent the number of cells that will be included in the window around each focal cell. For example, a window size of 3 will calculate the mean PoO in the 9 cells cantered around every focal cell. |
| *Out* | Character, either ‘*DataFrame*’, ‘*Stacked*’ or ‘*Both*’. see details below. |
| *Plot* | Logical, if *TRUE*, the mean PoO at each window size is plotted. |
| *verbose* | Logical, if *TRUE*, progress information will be printed in the R console. |
| **Output** | |
|  | If ‘*Out’* is set to *DataFrame*’, the function will return a dataframe with all the information of Stack (IDs and PoO), along with the mean PoO at each window size. If set to ‘*Stacked*’, the function will return a *RasterStack* object with the original IDs and PoO along with additional raster layers for each window size. If set to ‘*Both*’, the function will return a list containing both the dataframe and the *RasterStack*. |

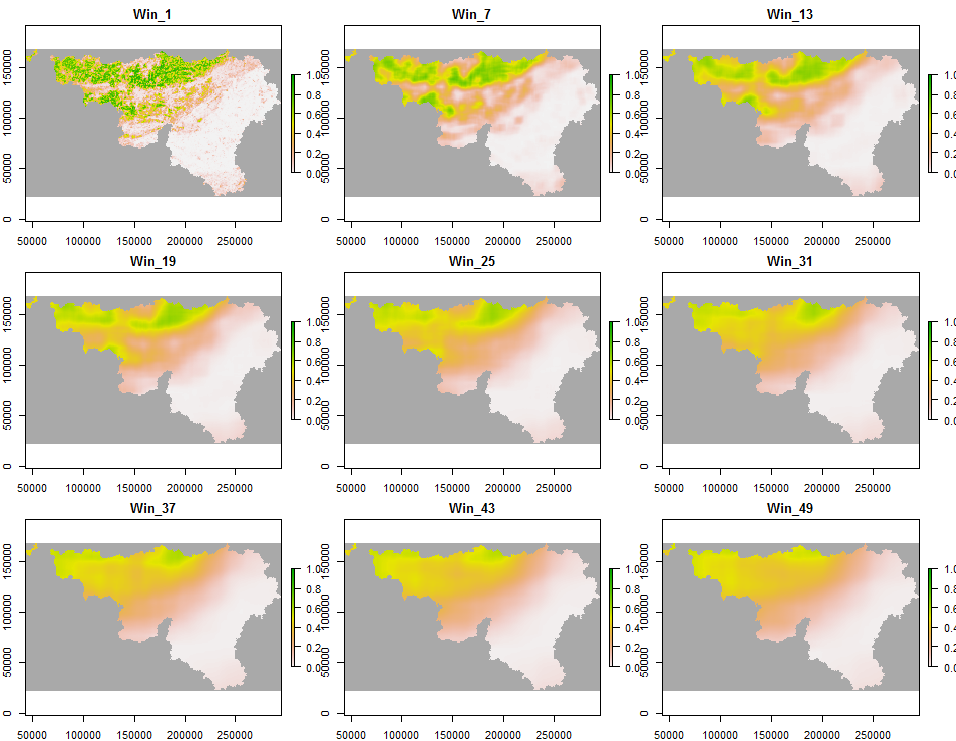
*4.e.3: Additional details*

After applying the *Win\_PoO.R* function, the mean PoO at every window size can be introduced as additional explanatory variables to a new SDM model. Here we explored two different ways in each the new mean PoO variables are included in the SDM. In the first case, (Expl\_MW), the mean PoO values are added to all the other raw explanatory variables used to train the original model. In the second case (named Add\_MW), the mean PoO values are modelled along with the original SDM PoO, without including the raw explanatory variables. For example, assume the original SDM used Envi1, Envi2 and Envi3 as its raw explanatory variables to produce Win1 (the predicted PoO at a window size of 1), and windows size 3, 5, and 7 are called in *Win\_PoO.R*. In this case, the Expl\_MW option will model the training P/A data against Envo1, Envi2, Envi3, Win3, Win5 and Win7. The Add\_MW option will model the training P/A data against Win1, Win3, Win5 and Win7. Thus, the Add\_MW option aims to modify the original PoO by adding information on the landscape suitability context of each cell. It may also allow the identification of important scales that affect the species distribution pattern by comparing the variable importance values of the different window sizes. The first option probably allows the original PoO to change more considerably, but may be more robust to overfitting since the original PoO values are not used directly as explanatory variables.

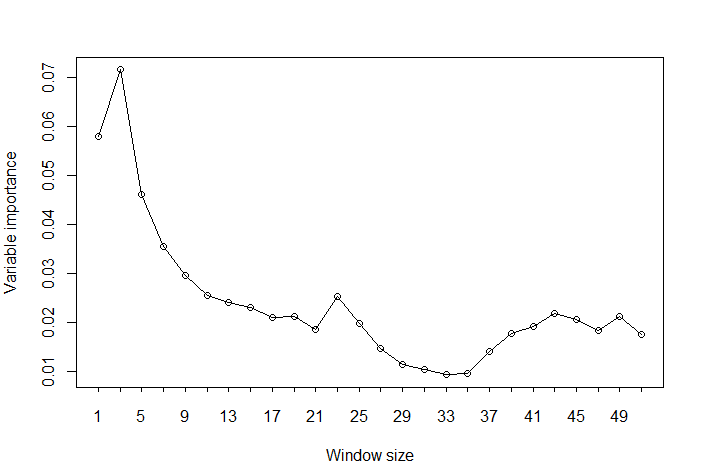
Regardless of the option chosen, the *Moving Windows SDM* result with a new, modified PoO maps, whose accuracy can be assessed against an external validation data set. In fact, to translate the PoO map to P/A values we used here the threshold that maximizes the value of the TSS. We note here, that although external atlas data is not needed to run this hybrid model, known distribution pattern at coarse resolution can be used to mask the generate PoO maps before selecting an optimal threshold.

*4.e.4: Example*

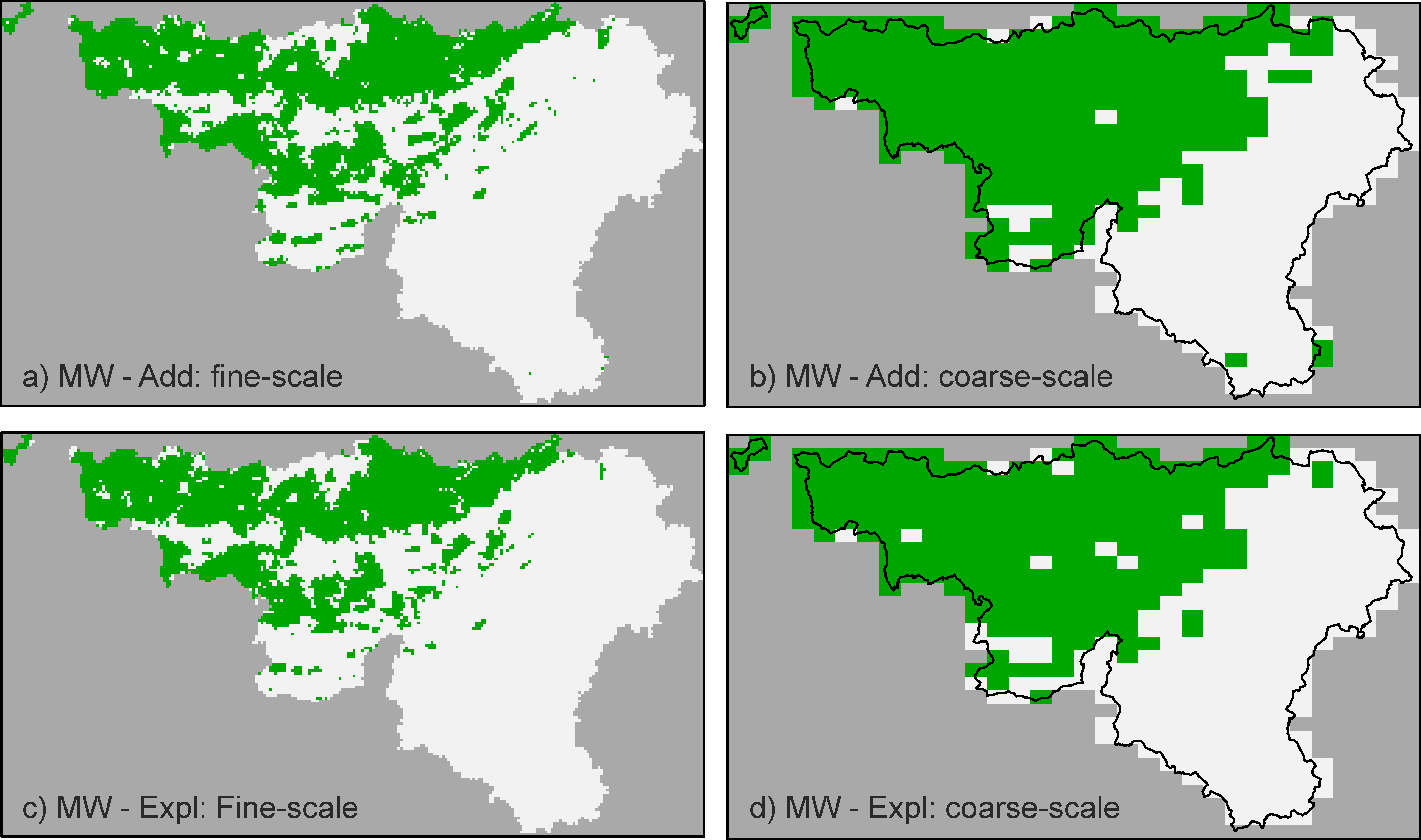
When fitting the *Moving Window SDM* to *M. flava*, we used window size between 3 and 51, with 2 cells increments (e.g., 3, 5, 7, …, 51). The mean PoO tended to become smaller as grain size increased, yet the north-west corner of Wallonia, which was tagged as area of high PoO in the original SDM, remained higher than the rest of the regions in all scales (Fig. 4.5). We explored both methods to use the mean PoO in the second SDM model- along with the original explanatory variables (Expl\_MW) or along with the predicted PoO at the finest resolution (Add\_MW). For the second option, the variable importance from the randomForest may be used to explore the scales at which the species is affected by its environment. For example, *M. flava* seem to have a pick in scales of 3×3, 23×23 and 43-49 (Fig. 4.6). In fact, the mean PoO at a 3×3 km resolution is more important than the original 1×1 resolution. The predicted distribution of M. flava at the 1×1 and 5×8 km resolution, according to both options is given in Fig. 4.7.



**Figure 4.5:** The mean PoO at increasing window sizes around each cell. Results from the S1250\_R1 dataset.



**Figure 4.6:** The change in variable importance with window size (in km). A window size of 5 implies that the mean PoO was estimated at a 5×5 km area cantered around each 1×1 km cell. Results from the S1250\_R1 dataset.

****

**Figure 4.7:** The predicted distribution of *M. flava* according to the Moving Window SDM approach, using either the ‘add to original scale’ (***a*** and ***b***) or the ‘raw explanatory variables’ (***c*** and ***d***) options. Panels ***a*** and ***c*** are at the 1×1 km resolution and panels ***b*** and ***d*** are at the 5×8 km resolution, based on unmasked SDM. Results from the S1250\_R1 dataset.

4.f: Top X occupied cells

*4.f.1: Main rationale*

The *Top X* model is the simplest hybrid model to incorporate information from both the SDM and from downscaling models. The model does not alter the underlying PoO map, but instead finds the threshold that produces the same number of occupancies as that predicted by downscaling. The model first creates atlas data from the training P/A (or presence only) data. Next the training data are upgrained to create a distribution pattern at coarse resolution. The coarse scale distribution pattern is used to fit one or more of the published downscaling model, and the fitted curve is extrapolated back to the resolution of the SDM. Next, thresholds are applied to the PoO data, and the occupancy calculated for each. The selected threshold is that where occupancy is closest to the predicted to be occupied at the SDM resolution by the downscaling models, therefore using coarse thresholds or where the PoO map contains many cells with the same values will likely lead to a slightly different occupancy from that predicted. In a way, this method can be considered as an additional method to select an optimal threshold ([see Liu et al. 2005 for other methods](#_ENREF_23)) (see #1444 for other methods).

*4.f.2: R application*

The R function ‘*TopX.R’* runs the Top X model. The input and output of the function are described in table 4.4.

**Table 4.4:** the input parameters to the *TopX.R* function provided in the supporting information.

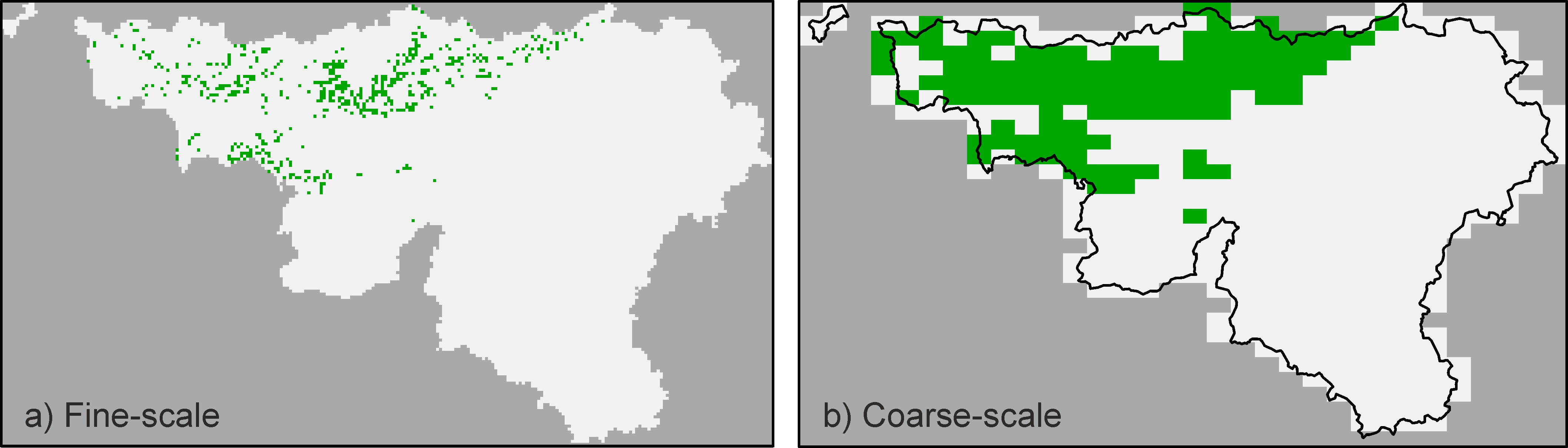
|  |  |
| --- | --- |
| Parameters | Description |
| ***Input*** | |
| *poo* | Object of class ‘raster’ of the predicted probability of occurrence values from the SDM model. |
| *pa\_data* | Data frame of presence-absence data. The data frame must contain the columns ‘X’, ‘Y’ and ‘Presences’ for the data xy coordinates and presence (1) or absence (0) respectively. |
| *atlas.scale* | The cell area for the desired atlas data. This must be a multiple of the cell area of the PoO map. |
| *upgrain.scales* | The number of scales to upgrain the atlas data and therefore for fitting the downscaling models. |
| *thresholds* | A vector of thresholds from which to extract the threshold that maximises occupancy to that predicted by the downscaling models. The longer the vector the greater the processing time (default = from 0 to 1 with increments of 0.01). |
| **Output** | |
| *poo\_map* | Object of class ‘raster’ of the probability of occurrence values masked by the atlas data. |
| *pa\_map* | Object of class ‘raster’ of the presence-absence map defined using the Top X threshold. |
| *pa\_map\_mask* | Object of class ‘raster’ of the presence-absence map defined using the Top X threshold after applying the atlas mask. |
|  | The function also produces four plots upon completion: a) the original PoO map from the SDM, b) the PoO map masked by the atlas data (*poo\_map*), c) the presence-absence map created using the Top X threshold (*pa\_map*), and d) the presence-absence map created using the Top X threshold after applying the atlas map (*pa\_map\_mask*). |

*4.f.3: Additional details*

Similar to other methods, the *Top X* method can make use of reliable atlas data by masking the SDM’s PoO map (i.e., setting all PoO values of absence atlas cells to 0) before selecting the top x cells. The model is highly dependent upon the accuracy of the downscaling model.

*4.f.4: Example*

The ensemble downscaling model predicted that 586 cells should be occupied at the SDM resolution (1×1 km). Fig. 4.8 presents the predicted distribution of *M. flava* using the threshold identified using the Top X method, producing 537 occupied cells. The figure also shows the predicted distribution at the 5×8 km resolution. Note, that the number of occupancies predicted by the downscaling model (586) is considerably smaller than those expected if the optimal TSS is selected (5587). Such differences in occupancy should translate to considerable differences in performance.



**Figure 4.8:** The predicted presence (green) / absence (white) map at ***a)*** the 1×1 km and ***b)*** the 5×8 km resolution according to the unmasked Top X model. Results from the S1250\_R1 dataset.

4.g: TopDown PoO

*4.g.1: Main rationale*

The *TopDown PoO* model is a hybrid model that selects fine-scale occurrences while simultaneously accounting for the following guiding rules:

1. Landscapes with higher mean PoO values are more likely to be occupied than landscapes with lower mean PoO.
2. Within occupied landscapes, cells with high local PoO are more likely to be occupied than cells with low PoO.
3. The number of local occurrences equals the number of predicted occurrences at the SDM scale from downscaling models.
4. The distribution of local occurrences is such that when upgrained, it will exactly equal the number of predicted occurrence according to the downscaling models at all scales.
5. When masked, the distribution of local occurrences is such that when upgrained it will exactly fit the known atlas distribution.

Although the *TopDown PoO* model does not change the SDM’s PoO values, it uses them to estimate the mean PoO at various grain sizes, similar to the *Moving Windows SDM* approach. However, unlike the *Moving Windows SDM*, which overlay a set of windows around each cell, the *TopDown PoO* works on pre-defined set of nested scales, following the regular aggregation rules (i.e., four neighbouring 1×1 cells are aggregated to create a single 2×2 cell, four 2×2 cells are aggregated to create a single 4×4 cells, etc.). The algorithm follows the following steps:

1. Use known occurrences to create atlas data using the upgrain function of the R package downscale (task 3.2, deliverable D3.1).
2. Calculate the mean PoO in each cell of all scales created by the upgrain function. See below for the usage of alternative aggregating functions.
3. Apply a single (or an ensemble) downscaling model using the downscale (or downscale.enemble) function of the package downscale. Save the predicted number of occupied cells at each scale.

After these 3 steps, the algorithm is slightly different when atlas masking is applied or not.

If no atlas masking is applied:

1. rank all cells in coarsest resolution (from the upgrain function) according to their mean PoO, and then:
   1. Select the top xi cells, with xi being the number of cells predicted to be occupied at this resolution according to the downscaling model.
   2. For each of the top xi cells, select the local cell with the highest PoO and tag it as presence.
   3. Update the occupancy status of all occupied cells at all scales. i.e., for each newly tagged presence, tag all cells within which it is nested, at all larger grain sizes, as presence.

If atlas masking is applied:

1. Start with the atlas scale (i), and select all the ki cells that are occupied at the atlas map, regardless of their mean PoO (i.e., ki being the observed number of occupied cells in the atlas data). Then:
   1. For each of the ki cells, select the local cell with the highest PoO and tag it as presence.
   2. Update the occupancy status of all occupied cells at all scales.

After step 4, the masked and unmasked option are re-joined and these steps are followed:

1. For each of the remaining scales, starting from the second coarsest resolution (i-1):
   1. Identify all the cells in scale i that are occupied (ki).
   2. Identify all cells in scale i-1 that fall within occupied scale i cells.
   3. From the cells of 5.b, create two lists of cells- those already occupied in scale i-1 (ki-1) and those that are currently unoccupied, but falls within occupied cells in scale I (ji-1). The second list is the list of potential cells for scale i-1.
   4. Extract from the downscaling model the predicted number of occupied cells at scale i-1 and calculate the number of cells that needs to be added at this scale (yi-1 = xi-1 – ki-1)
   5. If the number of needed cells is larger than 0 (yi-1 >=1), rank the potential i-1 cells (ji-1) according to their mean PoO.
   6. For each of the top yi-1 cells of ji-1, select the local cell with the highest PoO and tag it as presence.
   7. Update the occupancy status of all occupied cells at all scales.
   8. Move to the next lower scale.

*4.g.2: R application*

We have codified an R function named ‘*TopDown\_PoO.R*’ that follows the above algorithm. The input and output parameters of the function are given in table 4.5 below.

**Table 4.5:** the input parameters to the ‘*TopDown\_PoO.R’* function provided in the supporting information.

|  |  |
| --- | --- |
| Parameters | Description |
| ***Input*** | |
| *Stack* | Object of class ‘*RasterStack*’ with three layers at the following order:   1. Cell IDs for all valid cells, with *NA* for non-valid cells. 2. Presence (1) or absence (0) data in some of the cells, with *NA* for unsampled cells. Used to create the atlas data. 3. PoO values in all valid cells, with *NA* for non-valid cells. |
| *scales* | Positive integer, number of scales to upgrain in the *downscale::upgrain* function. larger or equal to *Atlas* below |
| *Atlas* | Positive integer, with a minimum value of 3. The number of top scales to be used when fitting the downscaling models. |
| *Mask* | Logical, if TRUE, the atlas mask will be used. |
| *FUN* | Character, controlling the function that will be used to aggregate PoO to coarser grain size. the following options are supported:   * 1. ‘*mean*’ – the mean PoO will be taken at each grain size.   2. ‘*median*’ – the median PoO will be taken at each grain size   3. ‘*Quan\_75*’ – the 75 quantile of the PoO is taken. |
| *models* | Character or vector of characters with the names of the downscaling models to be used. See *downscale:downscale* and *downscale::downscale.ensemble* for details. |
| *method* | Character, the method that will be used for upgraining (see *downscale::upgrain*). Currently, we suggest using the ‘*All\_Sampled*’ option, although the code was also tested for the ‘*Gain\_equals\_Lost*’ option. |
| *tolerance\_mod* | Numeric, see *downscale::ensemble.downscale.* |
| *tolerance\_pred* | Numeric, see *downscale::ensemble.downscale.* |
| *tolerance\_hui* | Numeric, see *downscale::ensemble.downscale.* |
| *starting\_params* | List, see *downscale::ensemble.downscale* or *downscale::downscale* for details. |
| *Plot* | Logical, if *TRUE*, the mean PoO at each window size is plotted. |
| *verbose* | Logical, if *TRUE*, progress information will be printed in the R console. |
| **Output** | |
| *Stack* | Object of class ‘*RasterStack*’ with the original three layers + the following layers:   1. PoO\_i, – the aggregated PoO at scale i, with i=0 being the original input scale. 2. Pa\_i – the predicted presence absence data at scale i.   As *RasterStack* only support rasters from the same resolution, all layers are returned at the same resolution as the SDM, with all fine resolution cells constituting any single cell of coarser resolution having identical values. |
| *Data* | Dataframe, containing the following columns:   1. ID – The IDs from the first input raster layer. 2. X, Y – The coordinate of each cell. 3. ID\_i, – The new IDs assign for each cell in scale i following the upgraining procedure. 4. PoO\_i – The aggregated PoO at each cell in scale i. 5. PA\_i – The predicted occupancy for each cell at scale i. |
| *DownHyb* | Dataframe, with a row for each scale, containing information on   1. Cell sizes. 2. Standardized and unstandardized extent and occupancy. 3. The observed number of occupied cells. 4. The predicted occupancy and area of occupancy for each downscaling model. 5. The mean predicted occupancy (if an ensemble model is used). 6. The predicted number of occupied cells. 7. The aim number of cells – ensuring that the number of occupied cells cannot increase with grain size. 8. Information from the TopDown loop – number of cells occupied, needed, available and added in each scale. 9. The final column contains the number of occupied cells at each scale as obtained from upgraining the output PA\_0 raster. |
| *UpGrained* | Object of class upgrain, the output of the *downscale::upgrain* function. |
| *DownScaled* | Object of class downscale, the output of the downscale::downscale function |

*4.g.3: Additional details*

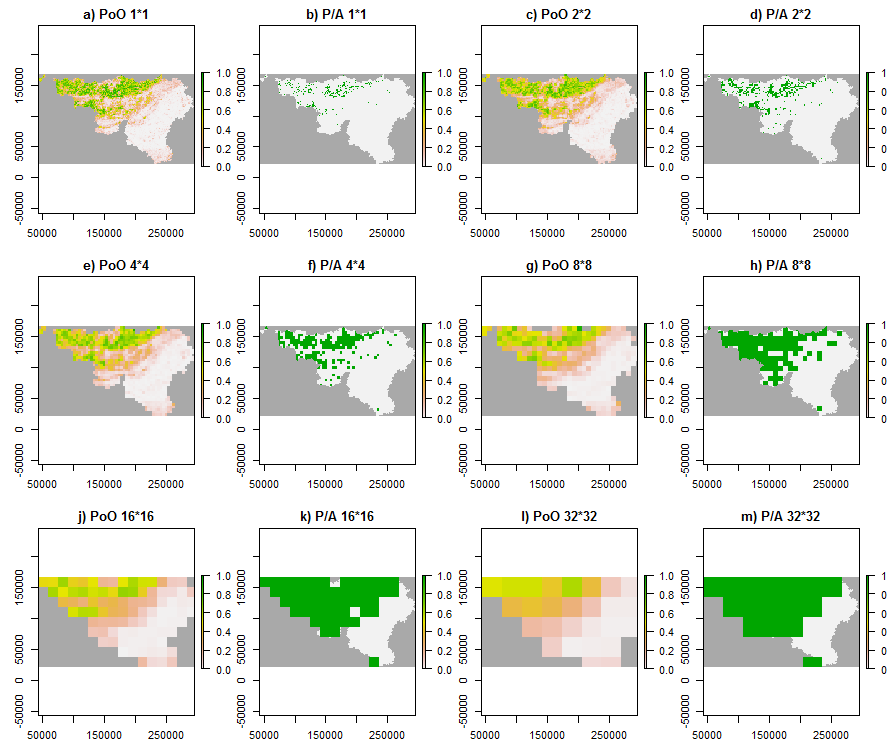
After repeating the procedure outlined in section 4.f.1, a fully nested distribution pattern emerges that satisfies rule D above. If masking is applied, then rule E is satisfied as well. Thus, the *TopDown PoO* method creates a spatially consistent P/A map from the original PoO without explicitly selecting a single global threshold. The algorithm jumps back and forth between coarse and fine scale, thus accounting for both fine-scale and landscape-scale suitability values. It adaptively constrain the fine-scale distribution to be nested within landscapes with favourable conditions. However, since the method is entirely constrained by the coarse scale distribution pattern, its performance should be highly dependent upon the accuracy of the atlas data and the performance of the downscaling model. Further note the selection of sites at coarse scale restrict the number of potential sites at fine resolution, especially if the set of valid cells in the original input raster results with irregular polygon. Thus in some cases, the algorithm cannot find enough potential cells at some scales. in this rare cases, the algorithm will under predict occupancy at some scales, relative to the downscaling models. The function allows an independent atlas data to be insertedas the second raster of ‘*Stack’*. In all cases, the atlas data is entered at the SDM resolution, to ensure full nestedness of all cells.

*4.g.4: Example*

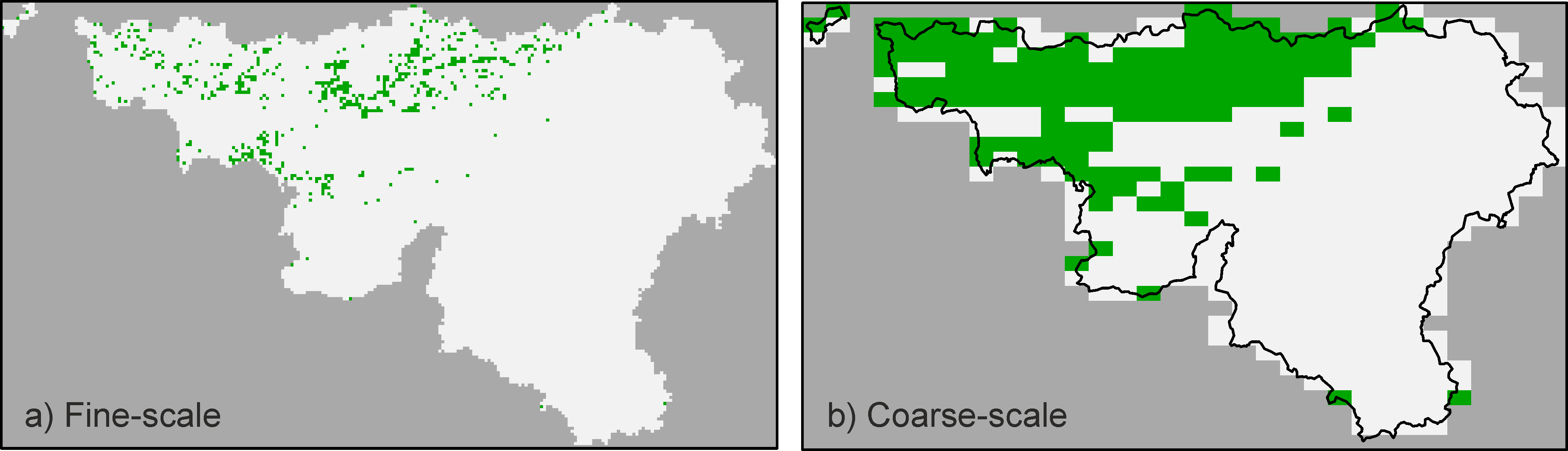
Below we give examples for the application of the *TopDown PoO* model to *M. flava* using the S1250\_R1 dataset with atlas masking. The ensemble downscaling model predicted that 586 cells should be occupied at the SDM cells (Table 4.6). The *TopDown PoO* algorithm distributed this exact number of occupancies, while ensuring that when upgrained, it will exactly fit the predicted number of occupancies at all scales. As atlas masking was employed here, the aim for cells smaller than 8×8 km are similar to the ensemble model, while the aim for cells equal or larger than 8×8 km are as the observed (table 4.6). Thus, the predicted distribution perfectly fit the input atlas data at scales of 8×8 and above. The selection of sites to tag as present at each scale was based on the mean PoO (Fig. 4.9) following the algorithm described above. The predicted distribution of the unmasked model at the 1×1 and 5×8 resolution is given in Fig. 4.10.

**Table 4.6:** the predicted proportion of occupied area (out of a standardized extent of 27,648 km2) observed and predicted by the downscaling models in each grain size. This downscaled prediction are translated to number of cells and the TopDown PoO model distribute the exact number of required occupancies at each scale.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cell area (km) | 1×1 | 2×2 | 4×4 | 8×8 | 16×16 | 32×32 |
| Occupancy | Observed | 0.009 | 0.036 | 0.119 | 0.266 | 0.481 | 0.593 |
| Nachman | 0.059 | 0.101 | 0.17 | 0.278 | 0.435 | 0.632 |
| Power law | 0.085 | 0.128 | 0.19 | 0.284 | 0.423 | 0.632 |
| Logistic | 0.038 | 0.078 | 0.152 | 0.274 | 0.445 | 0.629 |
| Poisson | 0.004 | 0.015 | 0.059 | 0.214 | 0.619 | 0.979 |
| Negative binomial | 0.009 | 0.035 | 0.113 | 0.271 | 0.458 | 0.615 |
| Generalized Negative binomial | 0.006 | 0.026 | 0.103 | 0.269 | 0.459 | 0.614 |
| Improved Negative binomial | 0.051 | 0.094 | 0.166 | 0.279 | 0.438 | 0.633 |
| Ensemble | 0.021 | 0.054 | 0.128 | 0.266 | 0.465 | 0.667 |
| N cells | Observed | 261 | 250 | 206 | 115 | 52 | 16 |
| Ensemble | 586 | 372 | 221 | 115 | 50 | 18 |
| Top down - aim | 586 | 372 | 221 | 115 | 52 | 16 |
| Top down - results | 586 | 372 | 221 | 115 | 52 | 16 |



**Figure 4.9:** The mean probability of occurrence (PoO) and predicted presence / absence (P/A, green as presence and white as absence) maps at all resolutions according to the masked *TopDown PoO* model. Results from the S1250\_R1 dataset.



**Figure 4.10:** The predicted presence (green) / absence (white) map at ***a)*** the 1×1 km and ***b)*** the 5×8 km resolution according to the unmasked *TopDown PoO* model. Results from the S1250\_R1 dataset.

4.h: The SpaNiche model

*4.h.1: Main rationale*

As mentioned above, different thresholds will translate the same PoO map to multiple nested P/A maps. Every such P/A map can then be upgrained to create a different OAR. The different OARs created by upscaling a thresholded PoO map and by downscaling atlas data lies at the basis of the *SpaNiche* model (Spatial-Niche model) developed under task 3.3 of Wp3.

The *SpaNiche* model assumes that the user have three types of input data:

1. A large enough number of occurrence data at a fine grain size to fit an accurate SDM.
2. A set of environmental or remotely sensed variables at a fine grain size for the SDM.
3. Atlas data at coarser resolution, whose probabilities of detection are assumed to be 1 (i.e., all coarse grained cells are correctly assigned as ‘presence’ or ‘absence’).

The *SpaNiche* model starts by applying a SDM model at the finest resolution, and predicting the PoO over the entire extent (Fig. 4.11A). At this step a single SDM algorithm may be used or alternatively the predictions of multiple algorithms can be averaged using available ensemble approaches. Next, if desired, the probability of occurrence map can be masked by the atlas data (in Fig. 4.11, arrow between F and A), such that all fine grain cells within atlas cells assigned as absences are assigned a PoO of 0. Next, multiple thresholds are applied on the PoO map, to produce nested binary maps (Fig. 4.11B). These fine-scale binary maps are used in two ways. First, fine scaled performance is estimated by calculating the TSS between the predicted P/A map and known presences and absence at fine resolution (Fig. 4.11C). The accuracy is plotted against the threshold to produce the niche consistency curves (Fig. 4.11D). If we are to account only for environmental aspects, we would select the threshold that provides the greatest accuracy in this curve. However, this may lead to over- or under-predicting occupancy at coarse resolutions. Thus, the binary maps for each threshold are also used to assess the spatial consistency by upgraining them and comparing them to the OAR of the downscaling models.

To assess spatial consistency we first fit the atlas data with one or more downscaling models (Fig. 4.11F) using the *downscale* or *ensemble*.*downscale* functions of the ‘*downscale*’ R package (see deliverable D3.1). Next we compare the downscaling OAR (Fig. 4.11F) to each thresholded OAR (Fig. 4.11E) and estimate a divergence value for each threshold (Fig. 4.11G). Divergence is based on summing the absolute difference between the downscaled OAR and threshold OAR over all grain sizes smaller that the atlas data. Divergence includes grain-dependent weighting such that differences at coarse resolutions, where we are more certain on occupancy patterns, have a larger influence on divergence than differences at small grain sizes for which the accuracy of the downscaled OAR is less certain. We used Eq 4.1- Eq.4.4 below to calculate the divergence, with *Gi* being the area of a single cell at scale *i* and *ASDM,i* and *Adown,i* being the area of occupancy at scale *i* according to the upgrained SDM and the downscaling models respectively. We then plot the divergence against the threshold to produce the spatial consistency curve (Fig. 4.11H). The threshold with the lowest divergence values is the one with the highest spatial consistency.

Eq. 4.1

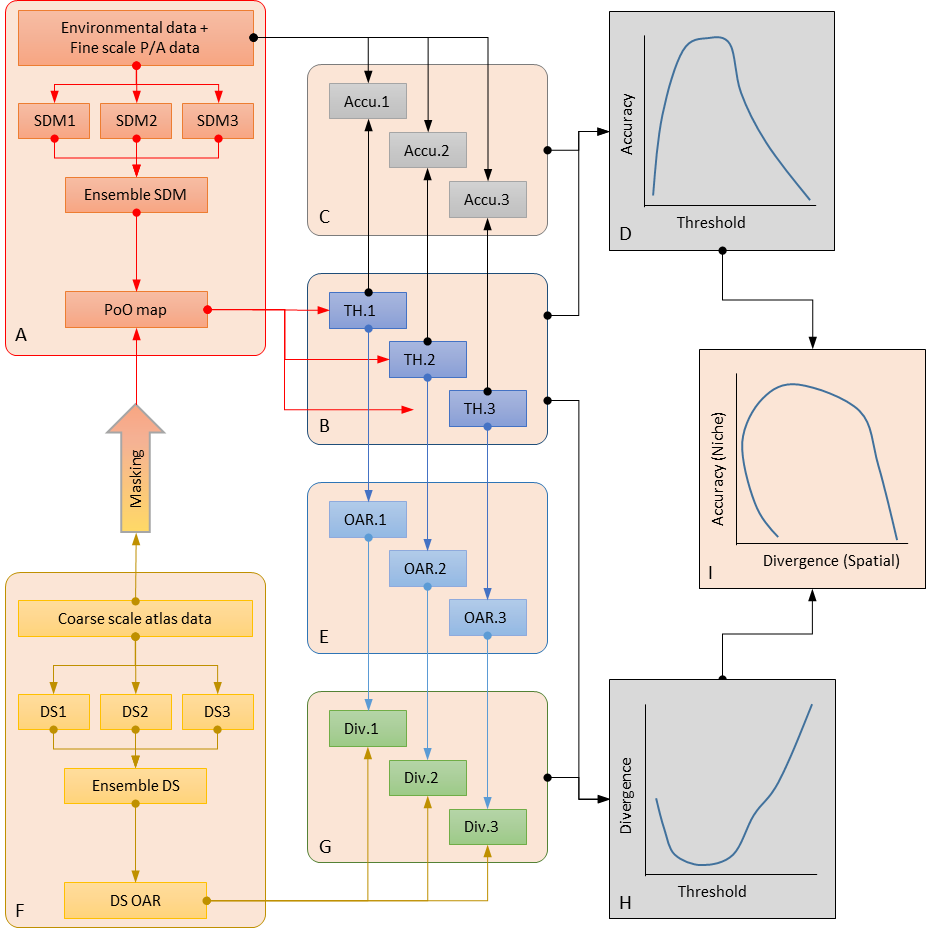
Eq. 4.2

Eq. 4.3

Eq. 4.4

Eq. 4.5

In most cases the maximal spatial consistency threshold is not necessarily the one that provides the highest niche consistency. Thus, the challenge is to find a threshold that provides a good balance between spatial and niche consistency. To do so, we plot divergence against accuracy to create the *SpaNiche* consistency trade-off plot (Fig. 4.11I). To ensure equal weighting between niche and spatial consistency both accuracy and divergence values are standardised between 0 and 1. Note, that TSS range from -1 to 1, yet we rarely observed values smaller than zero for any modelled species. Each point in this plot corresponds to a single threshold, and we select the point that is located at the shortest distance from the top left corner as the optimal combination of high niche accuracy and low spatial divergence. This selection criteria is similar to the procedure of selecting optimal threshold based on Receiver-Operating Characteristic curves ([ROC curve, see Liu et al. 2005](#_ENREF_23)) (ROC curve, see #1444).



**Figure 4.11:** Flow diagram of the *SpaNiche* model. For the niche modelling, ***(A)*** species distribution models are generated through ensemble modelling and ***(B***  thresholds applied to generate presence-absence maps. ***(C)*** Accuracy is measured for each threshold and ***(D)*** the accuracy-threshold curve plotted (niche consistency). For the spatial modelling, ***(F)*** atlas data at a coarse-grain size is downscaled to create an occupancy-area relationship curve (OAR). ***(E)*** Each map from thresholded map is upscaled to create a threshold-OAR, and ***(G)*** the divergence measured from the downscaled OAR, and ***(H)*** the divergence-threshold curve plotted (spatial consistency). ***(I)*** Finally, the balance between niche- and spatial-consistency is explored by plotting accuracy against divergence to determine the optimal threshold.

*4.h.2: R application*

We codified the *SpaNiche* model into the R function ‘*SpaNiche.R*’, available in the supporting information. Details on input and output parameters are given in table 4.7.

**Table 4.7:** the input parameters to the *SpaNiche.R* function provided in the supporting information.

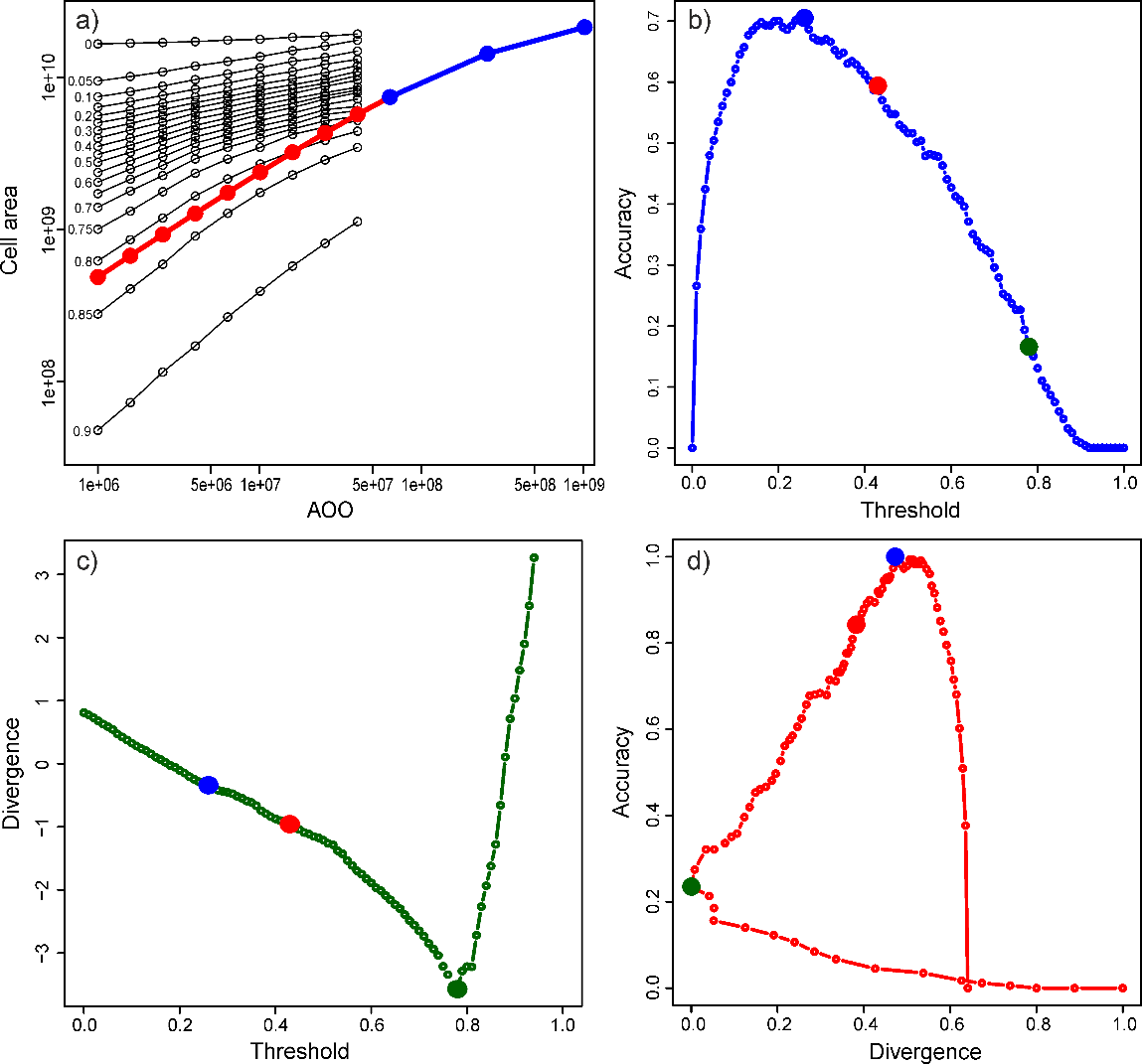
|  |  |
| --- | --- |
| Parameters | Description |
| ***Input*** | |
| *poo* | Object of class ‘raster’ of the predicted probability of occurrence values from the SDM model. |
| *pa\_data* | Data frame of presence-absence data. The data frame must contain the columns ‘X’, ‘Y’ and ‘Presences’ for the data xy coordinates and presence (1) or absence (0) respectively. |
| *atlas.scale* | The cell area for the desired atlas data. This must be a multiple of the cell area of the PoO map. |
| *upgrain.scales* | The number of scales to upgrain the atlas data and therefore for fitting the downscaling models. |
| *thresholds* | A vector of thresholds from which to extract the threshold that maximises occupancy to that predicted by the downscaling models. The longer the vector the greater the processing time(default = from 0 to 1 with increments of 0.01). |
| *divergence.scales* | A vector of cell areas for which to calculate divergence of the upgrained OARs from the predicted downscaled OAR. |
| *masking* | Whether to apply atlas masking (default = FALSE). |
| **Output** | |
| *acc\_div* | A data frame containing five columns:  *Threshold*: the thresholds applied to the PoO map.  *Accuracy*: the accuracy measured using TSS to the presence-absence data of the PA map generated by applying each threshold.  *Accuracy.stand*: the accuracy values standardised as a proportion of the highest accuracy value.  *Divergence*: the weighted difference values between the upgrained OARs created from the PA map generated by applying each threshold and the predicted downscaled OAR measured for each cell size specified by *divergence.scales.*  *Dists*: The distance to the top-left (0, 1) corner of the accuracy-divergence plot for each threshold. |
| *pa\_map* | Object of class ‘raster’ of the presence-absence map defined using the SpaNiche threshold. |
|  | The function also produces six plots upon completion:   1. a log-log plot of cell area against occupancy for occupancies of the predicted downscaled OAR (red), the data for training the downscaling models (blue) and a selection of upgrained OARs for thresholds in increments of 0.05 (black). 2. The niche consistency (accuracy-threshold) plot. 3. The spatial consistency (divergence-threshold) plot. 4. The consistence trade-off (accuracy-divergence) plot.   In these three plots the optimal threshold defined by each modelling stage are plotted as solid blue (optimal accuracy threshold), green (optimal divergence threshold) and red (SpaNiche threshold) points.   1. the original PoO map from the SDM, 2. the presence-absence map generated by applying the SpaNiche threshold (*pa\_map)*. |

*4.h.3: Additional details*

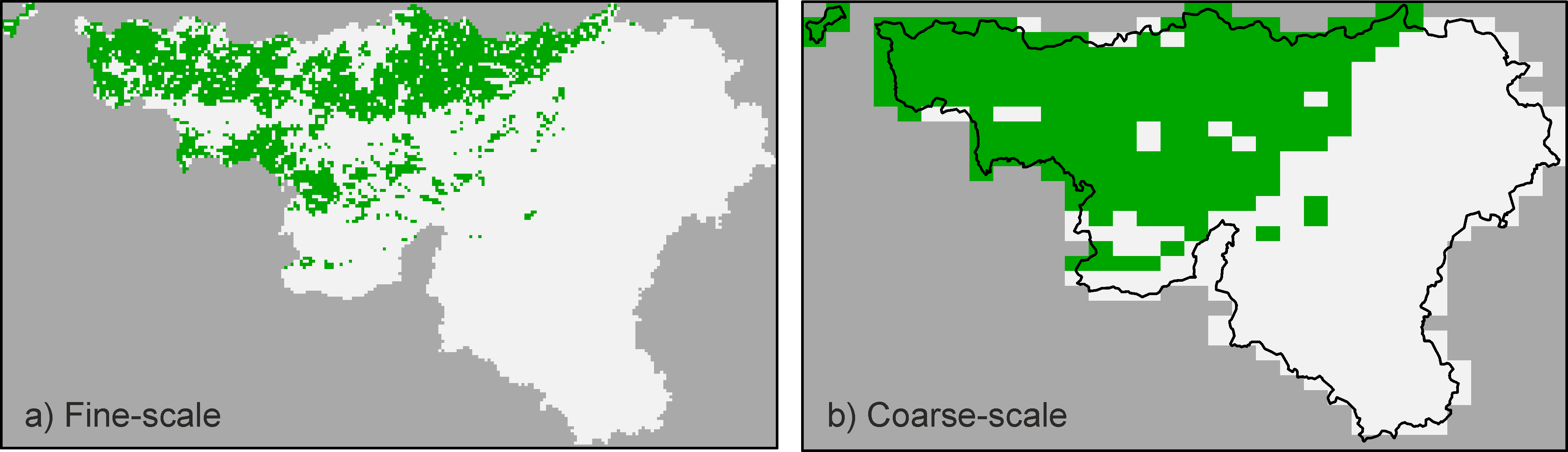
Like all other hybrid model, atlas masking is optional with the *SpaNiche* model as well, yet this is applied before upgraining and estimating the divergence values.

*4.h.4: Example*

When fitting the *SpaNiche* model to *M. flava*, we assessed divergence using 9 grain sizes equally spaced in log-space between the 1×1 km SDM resolution and the 8×8 km atlas resolution. Fine scale accuracy selected a threshold of 0.2 (TSS = 0.731; Fig. 12b, blue dot). This translates to 5587 occupied cells at the 1×1 km resolution. However, this threshold resulted with quite high divergence value (Fig. 12c) due to over-predicting the expected occupancy at all scales (Fig 12a). On the other hand, the minimal divergence was observed using a threshold of 0.76 (Fig. 12c, green dot), which results with only 865 occupied 1×1 km cells- a much closer value to the predicted occupancy of 586 (table 4.6). However, a threshold of 0.76 results in low fine-scale accuracy (TSS = 0.180). As expected, The *SpaNiche* model selected a threshold in between these two extremes, balancing off fine scale and coarse scale accuracy. The threshold with the smallest distance to the top left corner of the spatial-consistency curve was 0.41 (Fig 12d, red dot), resulting with fine scale TSS of 0.662 and only 3710 occupied cells. The predicted distribution according to this threshold at both the 1×1 and the 5×8 km resolution is given in Fig. 4.13.



**Figure 4.12:** Results for the unmasked *SpaNiche* model. a) the Occupancy-Area relationship curves from the ensemble downscaling model at the scales used for fitting (blue) and those extrapolated (red) along with the upgrained occupancy of the SDM for selected thresholds 9black lines). b) the fine-scale TSS values plotted against threshold, with the maximal TSS threshold (blue dot), the minimal divergence threshold (green dot) and SpaNiche model threshold (red dot). c) The standardized divergence values for various threshold (dots colour are as above). d) the fine scale TSS plotted against divergence for all thresholds. The red dot is the optimal threshold (minimum distance from the top left corner) providing relatively high accuracy with relatively low divergence. Results from the S1250\_R1 dataset.



**Figure 4.13:** The predicted presence (green) / absence (white) map at ***a)*** the 1×1 km and ***b)*** the 5×8 km resolution according to the unmasked *SpaNiche* model. Results from the S1250\_R1 dataset.

4.i: Comparison of the hybrid models

*4.i.1: Spatial distributions*

Although caution must be taken interpreting results here from a single species of a single data set some general patterns emerge, although some are likely to be case-specific. Application of the six different models resulted in a wide-range of predicted distributions, varying in their accuracy at both fine-scales and coarse-scales (Fig. 4.14). In general, those models with the greatest emphasis on fine-scale accuracy (original SDM and the *Moving-Window* approaches) predicted wider distributions than those which placed greater emphasis on spatial consistency (*Top X* and *TopDown PoO*). Largely, this is a result of very low prevalence predicted from the downscaling models. If the atlas data is unreliable, with many false absences, this may lead to such underestimation using downscaling models, which rely on accurate atlas data. This can be seen clearly in figure 4.15, where accuracy after masking is lower than without masking when sample sizes are low (500 samples) but higher when sample sizes are higher (2800 samples). In this case, it looks like 2000 samples provides equivalent accuracy to using all 2800 samples. That the atlas data created from the full 2800 transects only has a TSS of 0.635 also indicates that there is imperfect detectability of the species in the transect data.

*4.i.2: Predicted prevalences*

There were large differences in the predicted prevalences (number of cells occupied) of the models (Table 4.8). Models thresholded according to fine-scale accuracy predicted prevalences an order of magnitude larger than those constrained by the downscaling model (e.g. *SDM* = 5587 cells; *Top X* = 537 cells; both unmasked). The *SpaNiche* model falls between the two extremes (unmasked = 3710 cells; masked = 3189 cells). In fact, if the prevalence within the transect data (0.206) were to be extrapolated across the entire extent then we would expect 3428 cells, very close to that predicted by the *SpaNiche* model.

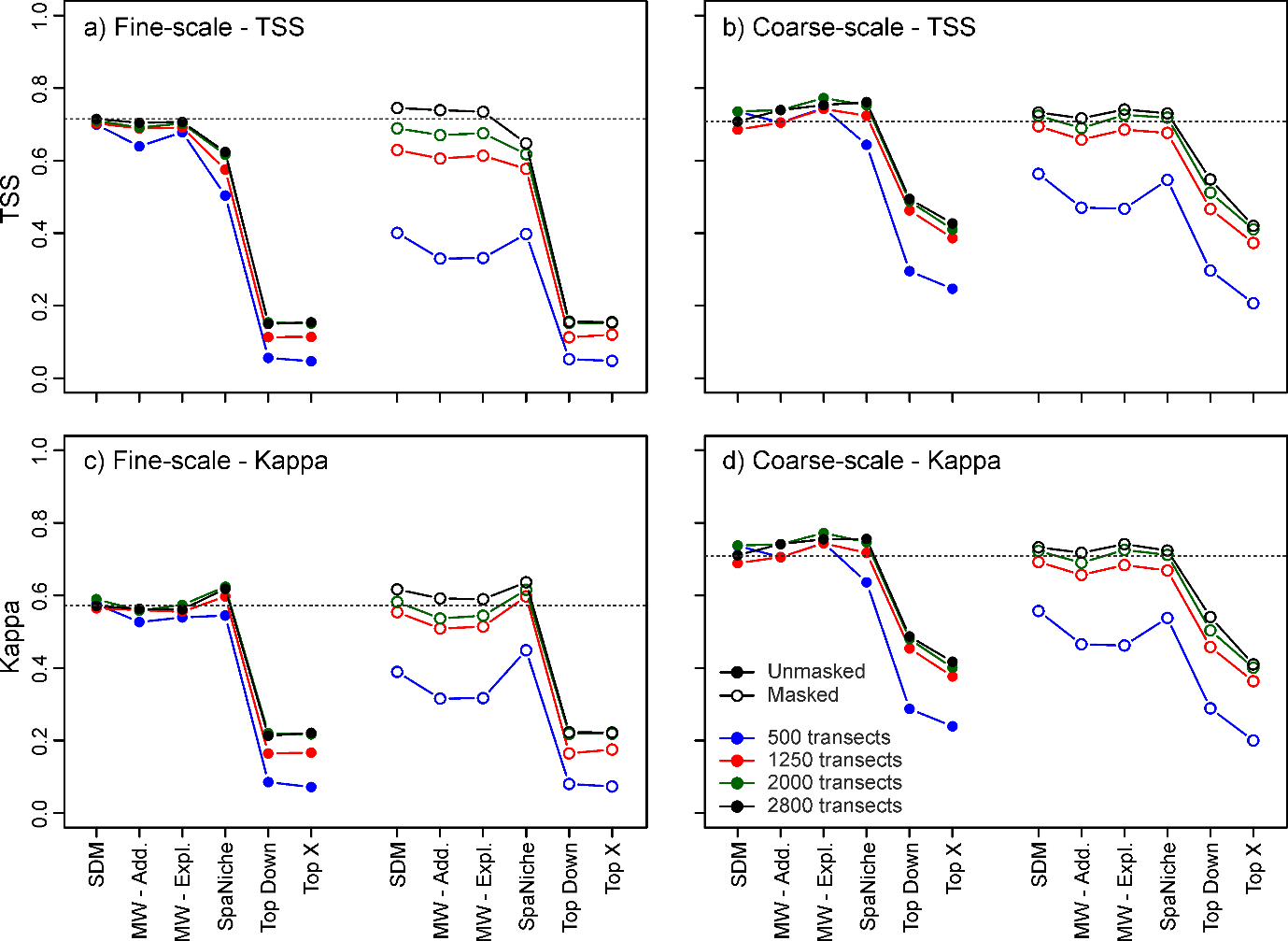
*4.i.3: Fine-scale and coarse-scale accuracy*

Unsurprisingly, fine-scale accuracy when measured using TSS is highest in the original SDM (which finds the threshold that maximises TSS) and the two *Moving Window* approaches (Table 4.8, Fig. 4.15). *TopDown PoO* and *Top X* both perform poorly due to the low prevalence predicted in the downscaling models. If accuracy is measured using Kappa then the *SpaNiche* model performs best (Fig. 4.15). At coarse-scales the two moving window approaches and the *SpaNiche* model all outperform the traditional SDM approach. Simple masking of all models to the coarse-scale atlas data, including the original SDM, can provide significant increases in fine-scale accuracy but only if atlas data is reliable, in this case where sample size is >2000 transects. If atlas data is inaccurate accuracy is greatly decreased.

|  |  |
| --- | --- |
| **Figure 4.14:** The predicted presence (green) / absence (white) map at the 1×1 km (left panels) and the 5×8 km (right panels) resolutions for all unmasked models, the transect data (red = presence, black = absence) and independent atlas data. Results from the S1250\_R1 dataset. | C:\Users\Yoni Gavish\Dropbox\Hybrid figures\Final figures\All maps.png |

**Table 4.8:** The performance of the SDM and hybrid model in fine (1×1 km) and coarse (5×8 km) resolutions. True Skills Statistic (TSS) is given at fine resolution against both the training data (where the threshold was selected) and the testing data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | Fine-scale | | | | Coarse scale | |
|  | |  | TSS  Training data | Threshold | TSS  Test data | No. cells (n = 16635) | TSS | No. Cells  (n = 490) |
| Unmasked | | |  |  |  |  |  |  |
|  | SDM | | 0.731 | 0.20 | 0.696 | 5587 | 0.735 | 277 |
|  | MW – Add. | | 0.704 | 0.196 | 0.687 | 5596 | 0.727 | 275 |
|  | MW – Expl. | | 0.714 | 0.261 | 0.699 | 5067 | 0.810 | 244 |
|  | SpaNiche | | 0.662 | 0.41 | 0.650 | 3710 | 0.792 | 225 |
|  | Top Down | | 0.109 | - | 0.129 | 586 | 0.495 | 143 |
|  | Top X | | 0.100 | 0.80 | 0.119 | 537 | 0.427 | 113 |
| Masked | | |  |  |  |  |  |  |
|  | SDM | | 0.785 | 0.20 | 0.651 | 4489 | 0.707 | 224 |
|  | MW – Add. | | 0.758 | 0.156 | 0.654 | 4768 | 0.680 | 238 |
|  | MW – Expl. | | 0.757 | 0.142 | 0.647 | 5010 | 0.709 | 237 |
|  | SpaNiche | | 0.681 | 0.41 | 0.606 | 3189 | 0.700 | 191 |
|  | Top Down | | 0.122 | - | 0.136 | 586 | 0.475 | 140 |
|  | Top X | | 0.122 | 0.79 | 0.143 | 592 | 0.400 | 106 |



**Figure 4.15:** Comparison of the fine scale (***a***, ***c***) and coarse scale (***b***, ***d***) performance of all models based on the True Skills Statsitics (TSS, ***a***, ***b***) and Kappa statisics (***c***, ***d***). SDM: the original SDM; MW – Add: the *Moving Window SDM* when adding the windows PoO to the original PoO, MW – Expl: the *Moving Window SDM* when adding the windows PoO to the raw explanatory variables. Results for the S1250\_R1 dataset.

5: Improved, high resolution freshwater SDMs

5.a: Aim

To adapt SDM models to freshwater ecosystems based on high resolution monitoring data.

5.b: Introduction

Species distribution models are mainly applied to terrestrial ecosystems where their calibration is straightforward and easy to replicate. However, freshwater ecosystems are structurally different: the habitat is arranged hierarchically along dendritic stream networks, which is influenced by the surrounding landscape within a catchment ([Domisch et al. 2015](#_ENREF_11)) (#1627). These properties need to be considered in freshwater SDMs, in order to produce predictions which truly represent freshwater biota and which can have applications in conservation and management. Applying such a stream-specific, high-resolution SDM approach to individual catchments yields results with a very high spatial resolution, particularly useful for local implementation in regional planning, global change forecasting or biodiversity conservation.

5.c: Approach

This tool modifies existing SDM to suit common stream biodiversity monitoring data by:

1. Including freshwater specific predictors (e.g., hydrology)
2. Considering the effect of the upper subcatchment for predictors in the landscape (e.g., land use and geology).
3. Projecting the predictions on the stream network.

Such models can give accurate predictions of where a species can be expected to occur in the catchment of interest (Fig. 5.1). The tool relies on the R package *biomod2* for the ensemble modelling procedure.



**Figure 5.1**: Predicted distribution of the stonefly *Brachyptera risi* in the Rhein-Main-Observatory (Kinzig catchment).

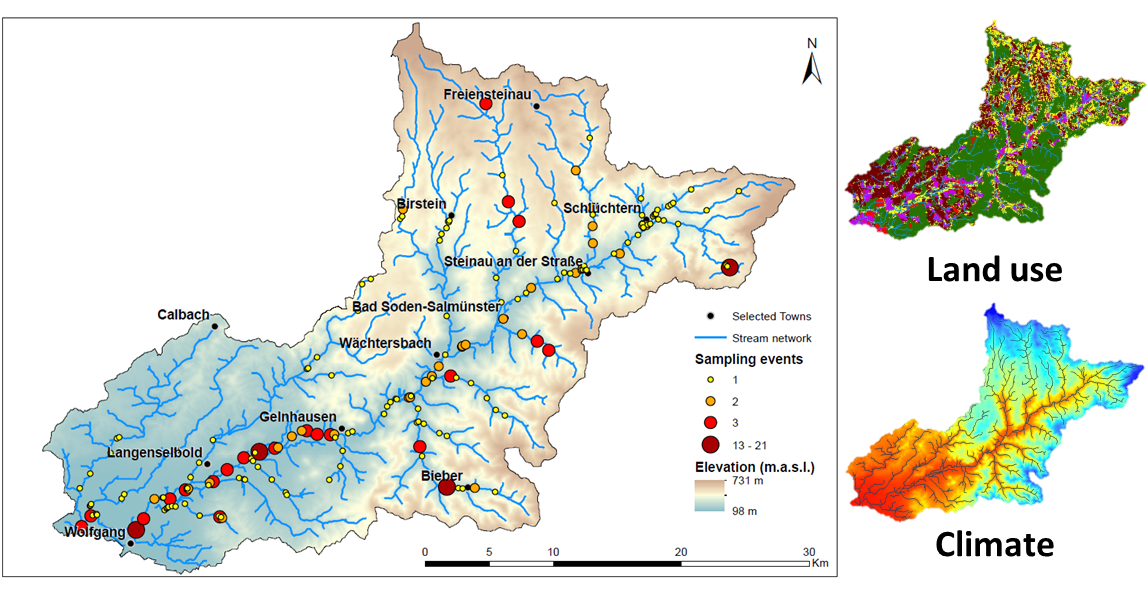
5.d: The freshwater SDM framework

*5.d.1: General instructions*

High-resolution freshwater SDMs are species distribution models (SDMs) adapted to the specific conditions of freshwater ecosystems and have a strong focus on small scale applications. While the procedure is similar to terrestrial SDM applications, there are specific considerations that are key to accurate predictions of freshwater biodiversity ([Domisch et al. 2015](#_ENREF_11)) (#1627). We describe the framework used to apply high-resolution freshwater SDMs as developed and applied in the context of the EU BON project here. However, there is the possibility to modify this framework depending on the particular conditions of the study site. Therefore, this tool consists of a framework to implement a freshwater SDMs, rather than a rigid R code. We exemplify the framework for the tool implementation based on examples from the Rhine-Main-Observatory (RMO), an EU BON test-site.

*5.d.2: Modelling extent*

As the focus lies on freshwater biodiversity, SDMs will be calibrated for the stream network in the study area of interest. It is recommended to work with a single catchment unit, linked by a stream network, thus providing natural boundaries for the model. All data used in the SDM will stem from this catchment (Fig. 5.2). Further, within the catchment, the area of interest is represented by the stream network, as it is here that freshwater biodiversity are mostly distributed. Therefore, data selection and prediction projections will take place exclusively on the stream network ([Domisch et al. 2013](#_ENREF_12)) (#1629). Further, the use of raster layers is recommended.



**Figure 5.2:** Modelling extent of the RMO: the Kinzing catchment; land use and climate predictors for the RMO

*5.d.3: Data preparation*

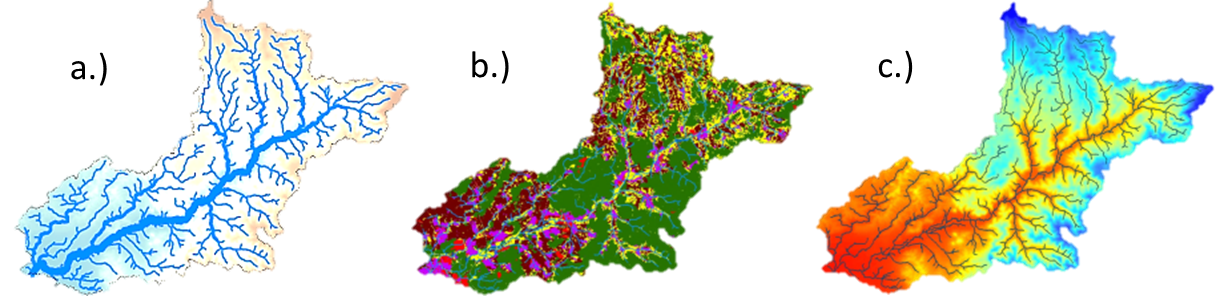
Biological data

Occurrence data for individual species is required for any SDM. For freshwater organisms, the occurrence data will not always be located on the stream network and should be snapped to the closest stream in the network. As in any SDM, repeated observations are converted to simple occurrences and these are then assigned to the nearest raster cell.

Environmental predictors

Commonly used predictors such as climate and topography can be used in the same way as in terrestrial SDMs and have been proven to be useful in stream macroinvertebrate predictions. Freshwater specific predictors are particularly interesting, as these describe conditions unique to freshwater ecosystems (e.g. discharge) and that have been proven to play a leading role in defining the distribution of stream macroinvertebrates ([Jaehnig et al. 2012](#_ENREF_18)) (#1630). Furthermore, some predictors require an adaptation to the freshwater realm, as is the case for land use ([Kuemmerlen et al. 2014](#_ENREF_20)) (#1628). Such predictors have a cumulative effect that extends upstream from any point of interest on the stream network and has an impact on the biota there. An example could be a high nutrient and sediment load generated by intensive agriculture several kilometers upstream from a site of interest, but within the relevant upper subcatchment and connected downstream through the river network.

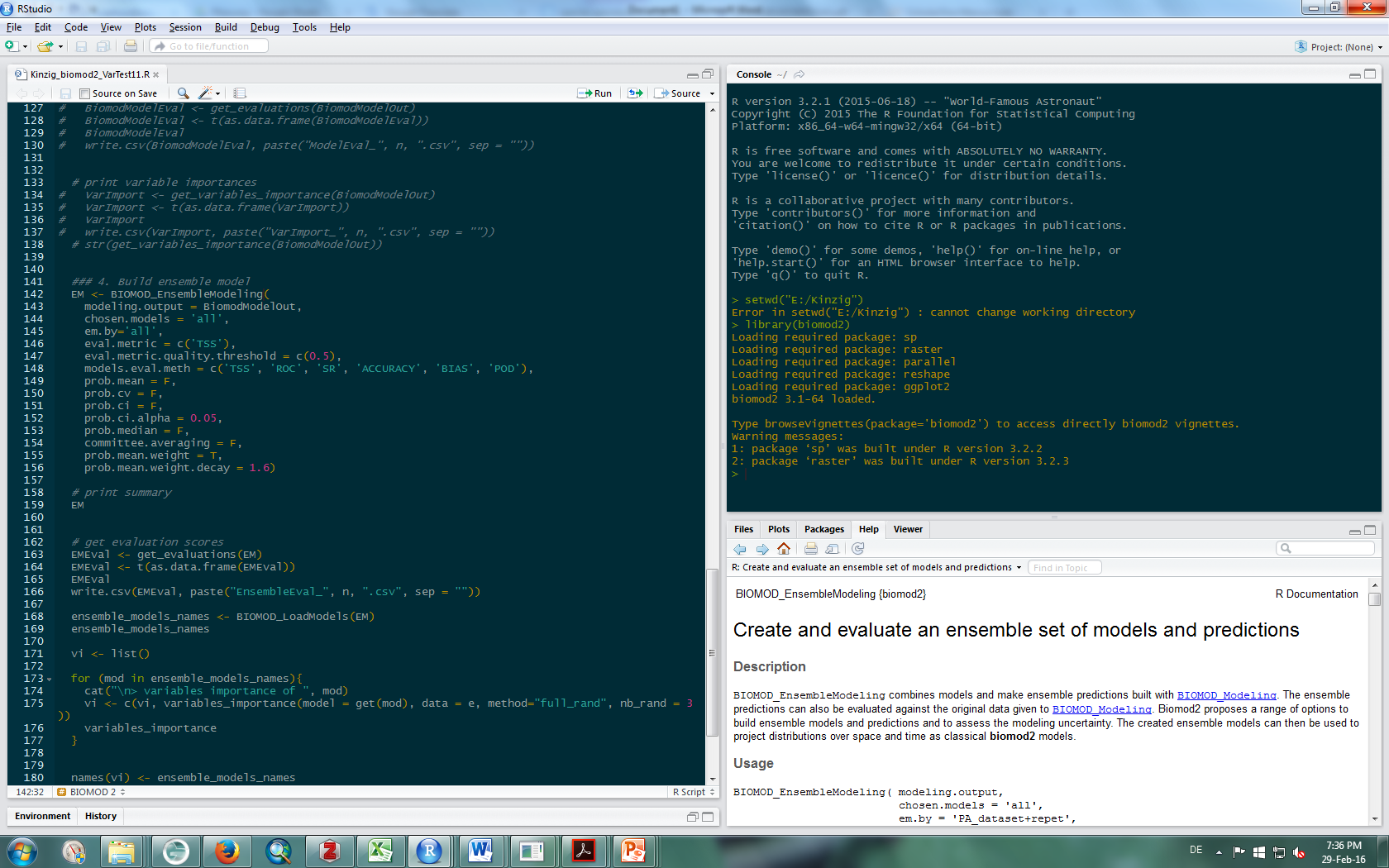
The data of the environmental predictors is homogenized to fit the spatial resolution of the stream network raster layer. Depending on the type of environmental data, different treatments should be observed. If hydrological data is available, then this is highly likely to be provided for few locations and should be extrapolated ([Kuemmerlen et al. 2012](#_ENREF_19)) (#1642) to the entire stream network (Fig. 5.3a). For land use data, it is necessary to consider the area in the landscape upstream of each raster cell to assign the corresponding value in the stream network ([Kuemmerlen et al. 2014](#_ENREF_20)) (#1628) The calculated values then reflect the relative proportion of a certain land use for every raster cell (Fig. 5.3b). A similar procedure has been implemented for geological predictors ([Kuemmerlen et al. 2015](#_ENREF_21)) (#1631). Predictors such as climate and topography do not require particular treatment and can be applied by a simple extraction of the values for each raster cell in the stream network (Fig. 5.3c).



**Figure 5.3:** Environmental predictors for the RMO: **a.)** hydrology, **b.)** land use, **c.)** climate

*5.d.4: Model calibration*

There are several alternatives to calibrate an SDM, which depends on the preferences and expectations of each application. Models presented here are built with the R package biomod2 ([Thuiller 2003](#_ENREF_27), [R Development Core Team 2014](#_ENREF_25)) (#1643, #902), which has several advantages over other alternatives: it implements numerous algorithms, uses several evaluation methods and allows to build ensemble models ([Araújo and New 2007](#_ENREF_3)) (#1644). Data can be provided as tables or as raster to biomod2. The model calibration allows for the modifications of many settings (Fig. 5.4). Models referred to here, were built using standard settings with the exception of the following: algorithms ('GLM', 'GBM', 'CTA', 'ANN', 'MAXENT'); evaluation methods ('TSS', 'ROC'), pseudo absences (7000), pseudo absence repetitions (3x), repetitions (10x) and a weighted average for the evaluation of the ensemble model. With these settings, 150 models are computed per species and summarized in one ensemble model.



**Figure 5.4:** The biomod2 package for R, run using the in R Studio GUI.

*5.d.4: Model projection*

Predicted distributions obtained from biomod2 can be easily mapped using standard GIS software. Among the possible results are predicted probabilities and uncertainties obtained from the variance in the ensemble model. These can be stacked and added across a community (Fig. 5.5). For individual species, a predicted presence/absence prediction can be produced (Fig. 5.1).

*5.d.6: Conclusions*

This framework for high-resolution freshwater SDMs allows to adapt SDMs to freshwater ecosystems and obtain predictions that better depict the conditions unique to these habitats. While we it is recommended to follow the steps as generally described here, it may be necessary to modify the framework because of data restrictions. Further, alternative or additional algorithms may be used. This framework is not a finalized tool, but rather one that is constantly developing. By documenting it, we hope to promote its wide application and improvement in the future.

|  |  |
| --- | --- |
|  |  |

**Figure 5.5:** Projection of results on the stream network. Excerpt from the RMO catchment.

5.e: Applications of the tool

The most recent and complete application of this tool has been performed for the Rhein-Main- Observatory, a EU-BON’s focal observatory site. In a first study, 175 stream macroinvertebrate taxa have been modeled for current conditions using predictors from the following categories: climate, topography, hydrology, land use and geology ([Kuemmerlen et al. 2015](#_ENREF_21)) (#1631). Current efforts are centered in similar predictions of 20 fish species and in the future, freshwater macrophytes will be modelled. As SDMs are increasingly being applied to freshwater ecosystems, it is encouraged to apply this tool in other test sites within the EU-BON consortium and beyond.

6: Diversity calculator

6.a: Aim

To calculate alpha and beta diversity on a large stack of raster (grid) data.

6.b: Introduction

In order to measure biodiversity on a large scale, species distributions are often derived from models (e.g. species distribution models or environmental niche models). However, in many cases the generated presence/absence maps of multiple species can be used to inform policy at the community level through various indices of alpha, beta and gamma diversities. Calculating diversity (alpha or beta), means one has to process large stacks of grid data which also often come with large extent and high resolution. However, current existing software (e.g. R or ArcGIS) often face severe challenges due to the size of the data.

6.c: Approach

The main idea was to split the data into smaller units, such that each can be processed in a separate core (i.e., multiple core machines). Thus, similarity is only calculated between cells that fall within the same unit, but not among units. In addition, within each unit, a moving window approach is employed, with user defined window size. For a focal cell located at the center of the window, similarity is estimated against all other cells that falls within it, but not outside the window. The smaller subsets are merged again after calculations.

6.d: Current status

The software has been developed and several tests were also successful. Currently no user friendly user interface exists. The idea and next steps are currently to develop a user friendly interface which allows the specification of the biodiversity measure (similarity index), the moving window size and the number of cores available. First efforts are conducted in that direction on collaboration with the team of Lifewatch Greece based at HCMR on Crete under the lead of Anastasis Oulas. The goal is to include the diversity calculator as R package under the virtual R lab (R VLab): https://portal.lifewatchgreece.eu/

The functionalities will allow the user the employ the multiple core machine underlying the r vLab. This will ensure that large scale calculations are possible in this remote environment. First programming steps are conducted and tests are in progress.



**Figure 6.1:** Examples of mapped amphibian diversity for West African amphibians: beta diversity with Mountford index (a) and with Jaccard index (b). The latter two were calculated for a moving window size of 51

7: References

Aizpurua, O., J. Y. Paquet, L. Brotons, and N. Titeux. 2015. Optimising long-term monitoring projects for species distribution modelling: how atlas data may help. Ecography **38**:29-40.

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

Araújo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution **22**:42-47.

Azaele, S., S. J. Cornell, and W. E. Kunin. 2012. Downscaling species occupancy from coarse spatial scales. Ecological Applications **22**:1004-1014.

Bahn, V., and B. J. McGill. 2007. Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography **16**:733-742.

Barwell, L. J., S. Azaele, W. E. Kunin, and N. J. B. Isaac. 2014. Can coarse-grain patterns in insect atlas data predict local occupancy? Diversity and Distributions **20**:895-907.

Baselga, A., and M. B. Araújo. 2009. Individualistic vs community modelling of species distributions under climate change. Ecography **32**:55-65.

Breiman, L. 2001. Random forests. Machine Learning **45**:5-32.

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecology and Biogeography **23**:99-112.

D'Amen, M., J.-N. Pradervand, and A. Guisan. 2015. Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. Global Ecology and Biogeography **24**:1443-1453.

Domisch, S., S. C. Jaehnig, J. P. Simaika, M. Kuemmerlen, and S. Stoll. 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 **186**:45-61.

Domisch, S., M. Kuemmerlen, S. C. Jaehnig, and P. Haase. 2013. Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. Ecological Modelling **257**:1-10.

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

Guisan, A., and C. Rahbek. 2011. SESAM – a new framework integrating macroecological and species distribution models for predicting spatio‐temporal patterns of species assemblages. Journal of Biogeography **38**:1433-1444.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography **16**:754-763.

Hoffmann, A., J. Penner, K. Vohland, W. Cramer, R. Doubleday, K. Henle, U. Kõljalg, I. Kühn, W. E. Kunin, J. Negro, L. Penev, C. Rodríguez, H. Saarenmaa, D. Schmeller, P. Stoev, W. Sutherland, É. Ó. Tuama, F. Wetzel, and C. Häuser. 2014. The need for an integrated biodiversity policy support process – Building the European contribution to a global Biodiversity Observation Network (EU BON). . Nature Conservation **6**:49-65.

Hui, C., M. A. McGeoch, and M. Warren. 2006. A spatially explicit approach to estimating species occupancy and spatial correlation. Journal of Animal Ecology **75**:140-147.

Jaehnig, S. C., M. Kuemmerlen, J. Kiesel, S. Domisch, Q. Cai, B. Schmalz, and N. Fohrer. 2012. Modelling of riverine ecosystems by integrating models: conceptual approach, a case study and research agenda. Journal of Biogeography **39**:2253-2263.

Kuemmerlen, M., S. Domisch, B. Schmalz, Q. Cai, N. Fohrer, and S. C. Jähnig. 2012. Integrierte Modellierung von aquatischen Ökosystemen in China: Arealbestimmung von Makrozoobenthos auf Einzugsgebietsebene. Hydrologie und Wasserbewirtschaftung **56**:185–192.

Kuemmerlen, M., B. Schmalz, B. Guse, Q. Cai, N. Fohrer, and S. C. Jaehnig. 2014. Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. Ecological Modelling **277**:77-86.

Kuemmerlen, M., S. Stoll, A. Sundermann, and P. Haase. 2015. Long-term monitoring data meet freshwater species distribution models: Lessons from an LTER-site. Ecological Indicators.

Kunin, W. E. 1998. Extrapolating species abundance across spatial scales. Science **281**:1513-1515.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography **28**:385-393.

Pellissier, L., K. A. Bråthen, J. Pottier, C. F. Randin, P. Vittoz, A. Dubuis, N. G. Yoccoz, T. Alm, N. E. Zimmermann, and A. Guisan. 2010. Species distribution models reveal apparent competitive and facilitative effects of a dominant species on the distribution of tundra plants. Ecography **33**:1004-1014.

R Development Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.

Ready, J., K. Kaschner, A. B. South, P. D. Eastwood, T. Rees, J. Rius, E. Agbayani, S. Kullander, and R. Froese. 2010. Predicting the distributions of marine organisms at the global scale. Ecological Modelling **221**:467-478.

Thuiller, W. 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology **9**:1353-1362.

Trainor, A. M., and O. J. Schmitz. 2014. Infusing considerations of trophic dependencies into species distribution modelling. Ecology letters **17**:1507-1517.

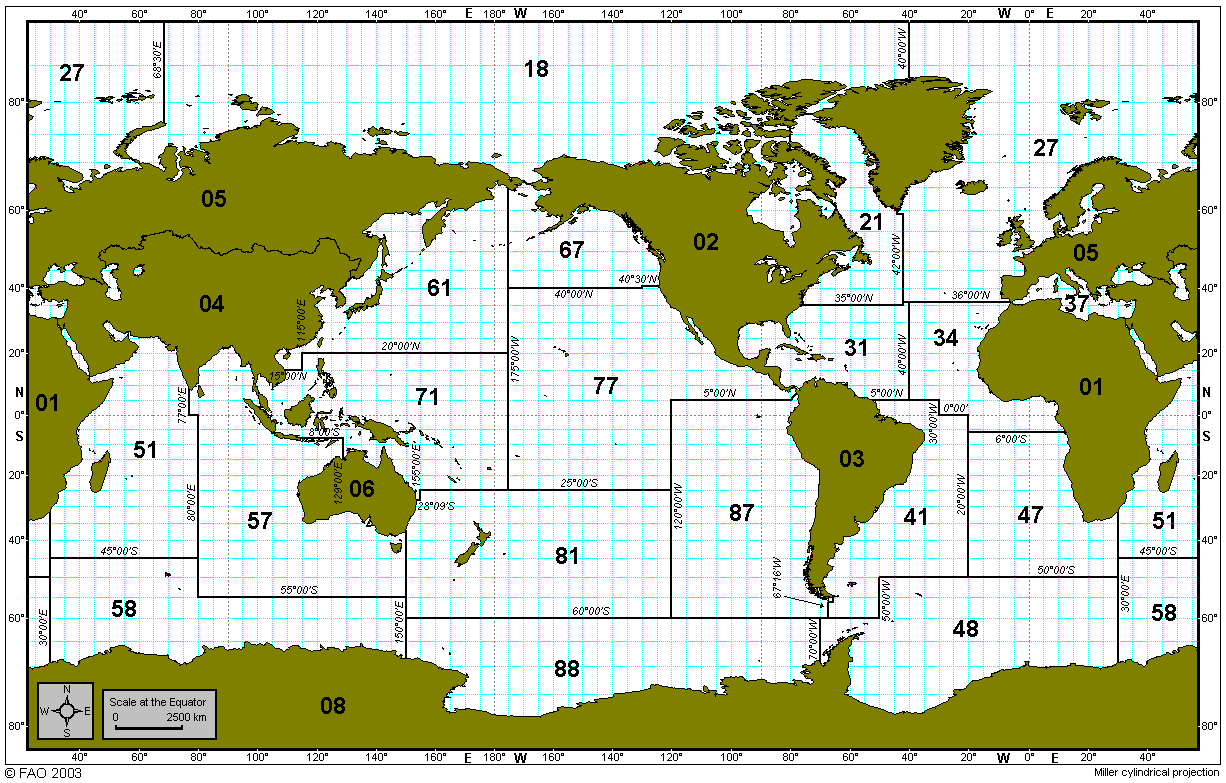
8: Appendices

The following appendices are added at here

* Appendix 2.1 – FAO Major Fishing Area
* Appendix 2.1 – AquaMaps five-star rating system

Appendix 2.1 – FAO Major Fishing Area

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Atlantic Ocean and adjacent seas | | Indian Ocean and adjacent seas | | Pacific Ocean and adjacent seas | |
| 21 | Atlantic, North-West | 51 | Indian Ocean, Western | 61 | Pacific, North-West |
| 27 | Atlantic, North-East | 57 | Indian Ocean, Eastern | 67 | Pacific, North-East |
| 31 | Atlantic, Western Central | 58 | Indian Ocean, Antarctic | 71 | Pacific, Western Central |
| 34 | Atlantic, Eastern Central |  |  | 77 | Pacific, Eastern Central |
| 37 | Mediterranean and Black Sea | Southern oceans and adjacent seas | | 81 | Pacific, South-West |
| 41 | Atlantic, South-West | 48 | Atlantic Antarctic | 87 | Pacific, South-East |
| 47 | Atlantic, South-East | 58 | Indian Ocean, Antarctic |  |  |
| 48 | Atlantic, Antarctic | 88 | Pacific, Antarctic |  |  |
|  |  |  |  |  |  |
| Arctic Ocean | |  |  |  |  |
| 18 | Arctic Sea |  |  |  |  |



Source: <http://www.fao.org/fishery/area/search/en>

Appendix 2.2 – AquaMaps five-star rating system

This appendix contains the Five-Star Rating Scheme to guide an expert-reviewer in evaluating the reliability of an AquaMaps native range prediction for a given species.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Star Rating | | | | |
| 5 | 4 | 3 | 2 | 1 |
| Environmental envelopes | Envelopes ok; no further edits recommended | Envelopes ok; no further edits recommended | Envelopes ok although species known to have large inter-annual changes in habitat usage are only inadequately captured by single annual envelope | Envelopes ok but may still be improved by adjusting parameters (>2); or available environmental parameters are unable to adequately describe species occurrence | Computer-generated map |
| Area restrictions | Bounding box/basins complete and with good fit to known distribution | Bounding box/basins complete and with good fit to known distribution | Uses bounding box/basins; no further improvements of bounding box/basins possible but areas of false predicted presence remain | No bounding box/basins; defined by FAO areas that encompass entire known range | Computer-generated map |
| Point data/good cells | Adequately large sample size; samples cover representative portion of species range; no apparent bias introduced; no good cells needed to be added/ removed | Adequately large sample size; samples cover representative portion of species range; biases were corrected by adding or removing good cells | Medium sample size and coverage of known species range but strong effort biases due to heterogeneous sampling effort; possible point data bias/output can only be improved with addition/deletion of a large number of good cells | Low sample size, non-representative coverage of species range by sampling | Computer-generated map |
| Predicted range of occurrence | In very good agreement with known range/significant statistical relationship between predictions and independent survey data | In good agreement with known range | Approximates known range but possibly with some areas of false predicted presence or absence | Approximates known range but includes large areas of false predicted presence or absence | Computer-generated map |
| Predicted relative probabilities of occurrence | In very good agreement with known relative occurrences /significant statistical relationship between predictions and independent survey data | In good agreement with known relative occurrences | Good correspondence with overall range but large discrepancies between predictions and known concentrations of high species occurrence | Good correspondence with overall range but large discrepancies between predictions and known concentrations of high species occurrence | Computer-generated map |