



BioVeL training

Matthias Obst, Francisco Quevedo, Elisabeth Paymal

The background image is a collage. It features a white bird in flight on the left, a yellow and black butterfly on the right, and a green forest landscape in the center. Overlaid on these are vertical columns of binary code (0s and 1s) and a test tube with red liquid on the right side.

## BioVeL workflows for taxonomic data processing and ecological niche modelling

TRAINING WORKSHOP  
Paris, FR, March 25, 2014



25 mars 2014 - 9h00- 17h00, MNHN, Salle des logs, bâtiment de géologie, 43 rue Buffon 75005 Paris

Confirmed programme : BioVeL workflows for cleaning an refining data and for studying ecological niche modelling

Note: this training will be offered in English; interprétation possible

Trainers :

Matthias Obst, University of Gothenburg, Sweden

Francisco Quevedo, Cardiff University, UK

Elisabeth Paymal, Fondation pour la Recherche sur la Biodiversité (FRB)

09:00 - About DRW and ENM and scientific examples

Intro to BioVeL and the Taxonomic Data Refinement Workflow (DRW), with examples of scientific applications

9:20 Demo of DRW

9:30 - 10:00 - Sign on the portal and start of hands-on session

10:00 - Download data from GBIF, clean and refine them.

11:30 - Integrate my own data w/ GBIF's, and upload again. Clean the merged set.



12:00 - Lunch

13:00- Introduction to ecological niche modeling, (ENM), and presentation of scientific applications

Showcase and demonstration of the Ecological Niche Modelling workflow

14:00 - Hands-on session of the Ecological Niche Modelling workflow

Use your data set that we prepared for you and run ENM

15:00 - Demonstration and hands-on session of the statistical analysis

- Use your data set that we prepared for you and statistical analysis

15:30 - Demonstration and hands-on session of with the climatic envelop workflow

- Use your data set that we prepared for you with the climatic envelop workflow

16:00 - 16:20 - Pause café/thé

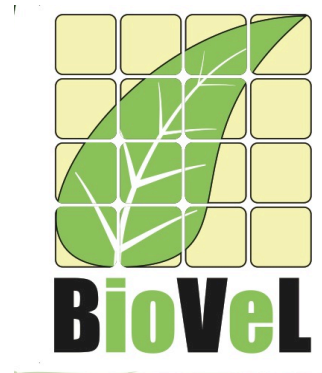
16h:20 - Wrap up discussion (both GBIF and BioVel): feedback and suggestions for research projects and improvements

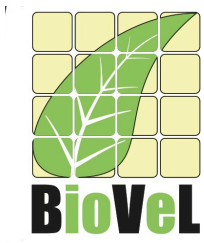
17:00 - Clôture de la session

Tour de table



# Introduction to the BioVeL infrastructure





## Aims of the Biodiversity Virtual e-laboratory (BioVeL)

BioVeL provides functions to:

- Access data from cross-disciplinary resources (Data mining)
- Access analytical methods a range of disciplines (Interoperability)
- Digest large data (Scalability)
- Repeat complex analytical processes (Reproducibility)
- Access to virtual communities (Sociability)

Provide (web)services for the interdisciplinary analysis of biodiversity

Provide analytical pipelines (workflows) based on these services

Taxonomy

Phylogenetics

Genomics

Population modelling

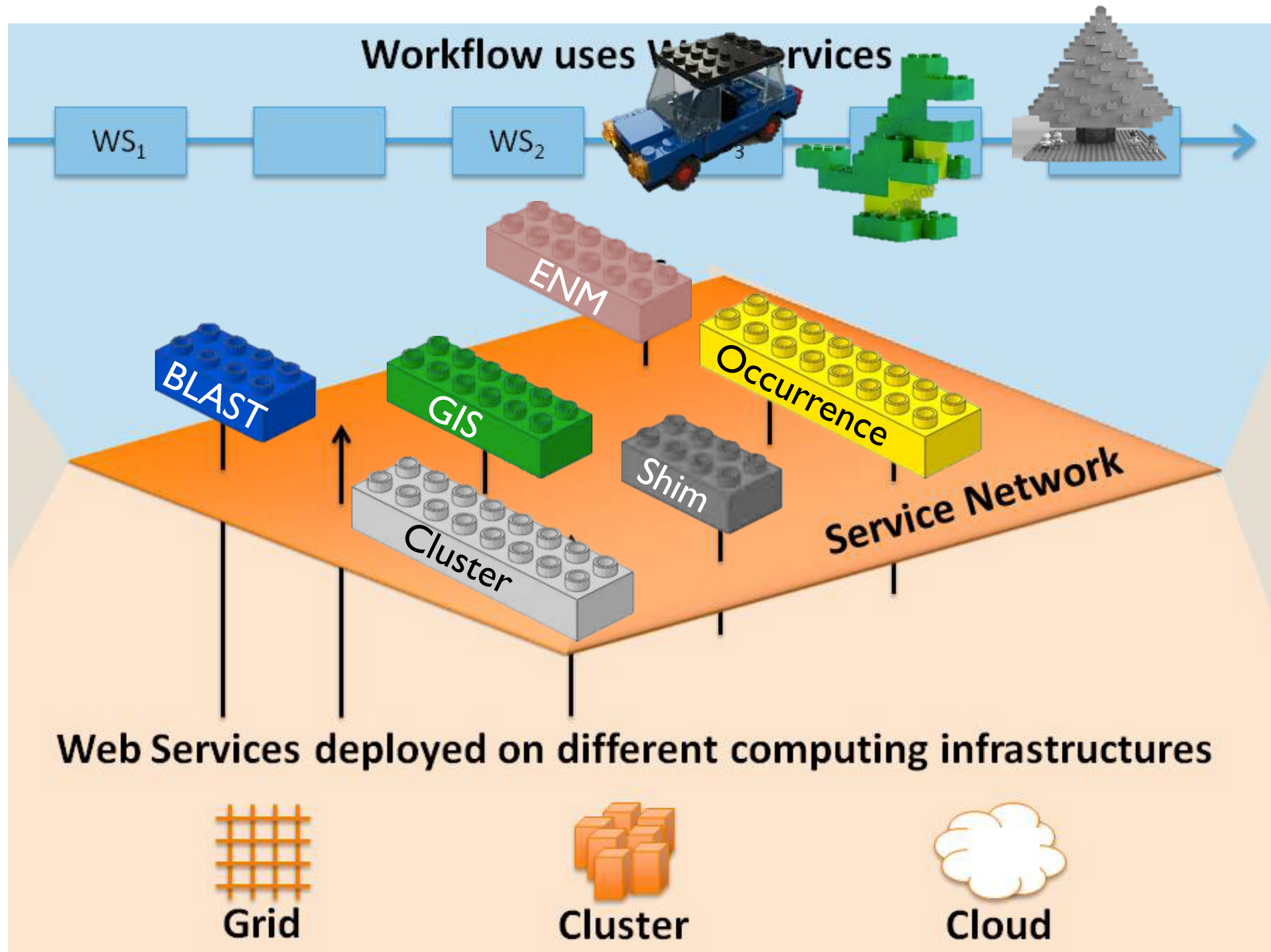
Ecological niche modelling

Ecosystem functioning/  
valuation

...

*Biodiversity research*

## Workflow uses Web Services



# Access web-services at *biodiversitycatalogue.org*

**BiodiversityCatalogue**  
"The Biodiversity Sciences Web Services Registry"

Getting Started | About Us | Contact Us | API Docs

Search:  **Go!** [Home](#) [Services](#) [Register a Service](#) [Providers](#)

[Sign up](#) [Sign in](#)

Home » [Services](#)

**Top 20 tags on BiodiversityCatalogue** (more) [See All Tags](#)

aphiaid | catalogue of life | computation | darwincore | data upload | ecology | gbif | interpolation | marine | modelling | MrBayes | MultiCPU | occurrence | raster | registry | shim | taverna | taxa | taxon | visualization

Displaying **all 17** services **Include archived services?** ☒ **Sort by:** Newest **View:** Grid

**Filtering**

Current Filters Applied: none

Select filters from below...

[Enable tag filters](#)

**Service Types (2)**

[REST](#) (12) [SOAP](#) (5)

**Service Categories (53)**

- [Taxonomy](#) (7)
- [Analysis](#) (3)
- [Modelling](#) (4)
- [Geospatial](#) (1)
- [Data](#) (3)

**PESINameService** SOAP

Taxonomy

As a user or developer you can use the PESI webservice to feed your own application with standard PESI taxonomy.

Provider: [www-eu-nomen-eu](#)

**European Nuc ... ENA) Browser** REST

Categories: Data Retrieval

The European Nucleotide Archive (ENA) Browser provides functionality to view and retrieve data and meta-data archived...

Provider: [European Bioinformatics Institute \(EBI\)](#)

**AphiaNameService** SOAP

Categories: Taxonomy Taxonomic Name Resolution Taxonomic Synonym Resolution

The data is licensed under a Creative Commons 'BY' 3.0 License, see <http://creativecommons.org/licenses/by/3.0/deed.e...>

Provider: [www-marinespecies-org](#)

**openModeller** SOAP

Niche Modelling (Species Distribution)

Functionally equivalent to the openModeller service provided by modeller-cria-org-br, this endpoint is located in Eur...

Provider: [omws-i3m-upv-es](#)

**BGBM CDM Cat ... ife REST** REST

API

Taxonomy Taxonomic Synonym Resolution Taxonomic Diversity Checklist and Classification

This web service namespace is an add-on to the already existing CDM REST API and provides information relating to sci...

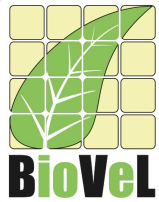
Provider: [BGBM EDIT](#)

**MrBayes 16 CPUs** REST

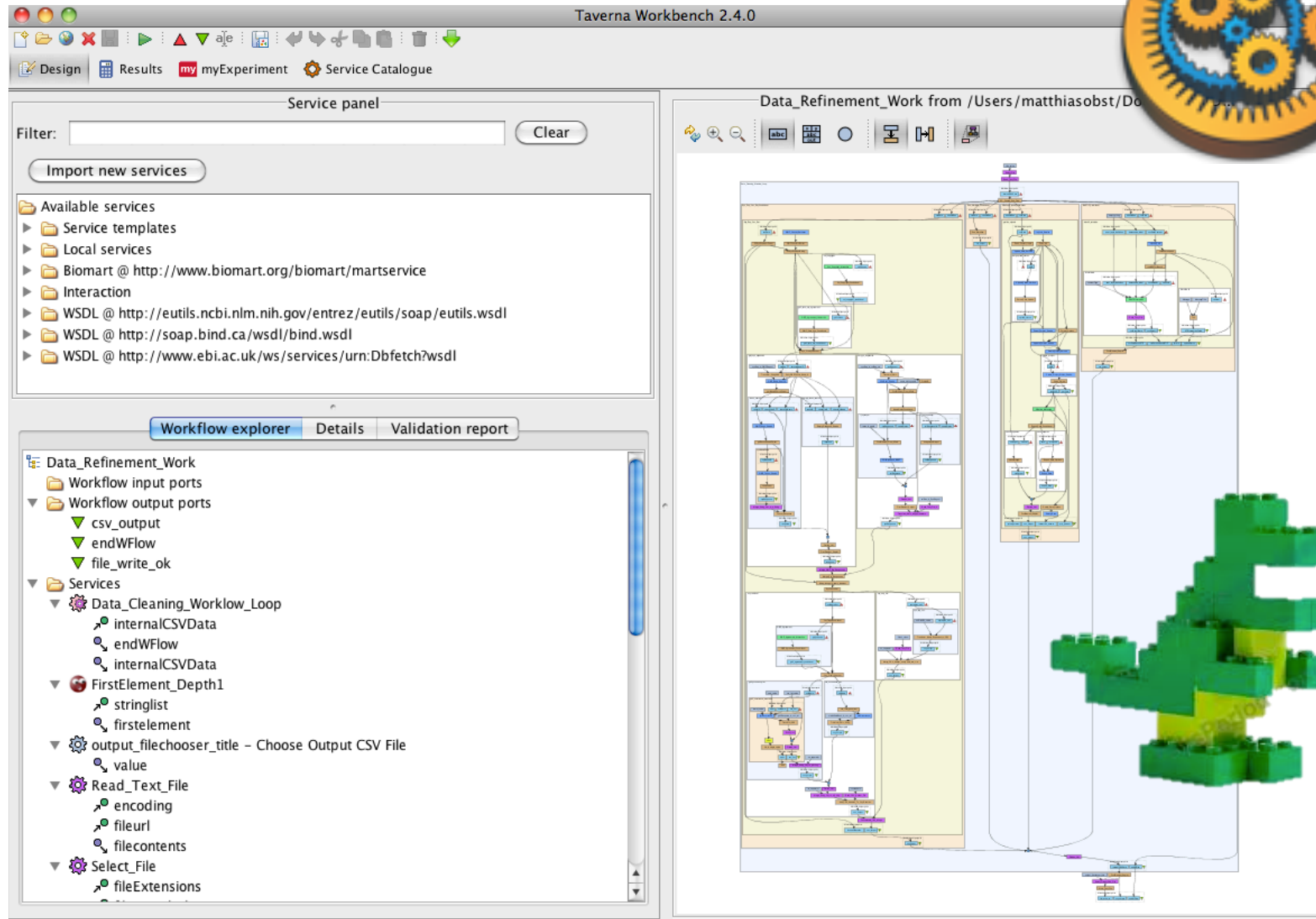
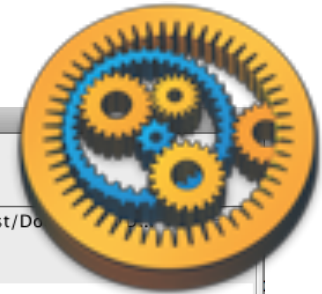
The service launch a bayesian phylogenetic inference with MrBayes 3.2 (<http://mrbayes.sourceforge.net/>) asking as inp...

Provider: [alicegrid17-ba-infr](#)






# Build your own workflow on the Taverna workbench



*Ecological niche modelling workflow*

# Running workflows from the portal

HomeWorkflowsRunsContactMatthias ObstLog out






Photo by Maria Paula Balcázar Vargas

**Welcome to the BioVeL Portal**  
For technical support or questions about the BioVeL Project, please visit the [contact page](#).


**Choose an analysis...**




Taxonomic Refinement




Ecological Niche Modelling




Metagenomics




Phylogenetics




Population Modelling



Ecosystem Modelling



My BioVeL



**BioVeL** is funded by the European Commission 7th Framework Programme (FP7) as part of its e-Infrastructures activity (Grant no. 283359). Under FP7, the e-Infrastructures activity is part of the Research Infrastructures programme, funded under the FP7 'Capacities' Specific Programme. It focuses on the further development and evolution of the high-capacity and high-performance communication network (GÉANT), distributed computing infrastructures (grids and clouds), supercomputer infrastructures, simulation software, scientific data infrastructures, e-Science services as well as on the adoption of e-Infrastructures by user communities.

Portal version: 1.0.0-11417

# The objective of this course

Introduce you to ***scalable analytical methods that integrate data access with analysis*** for species distribution modeling

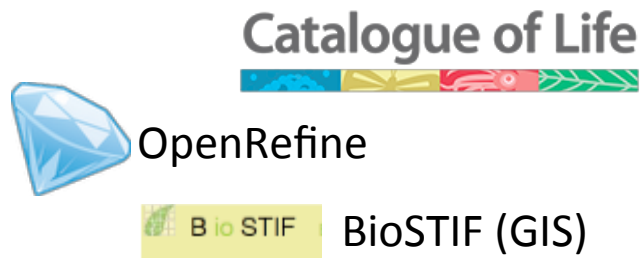
Please note, these are ***not new bioinformatics tools***, but rather existing ones that are seamless connected



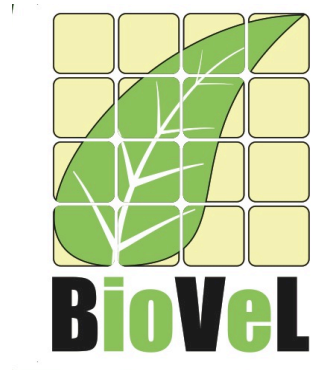
Taxonomy

Species distribution modeling

Statistics

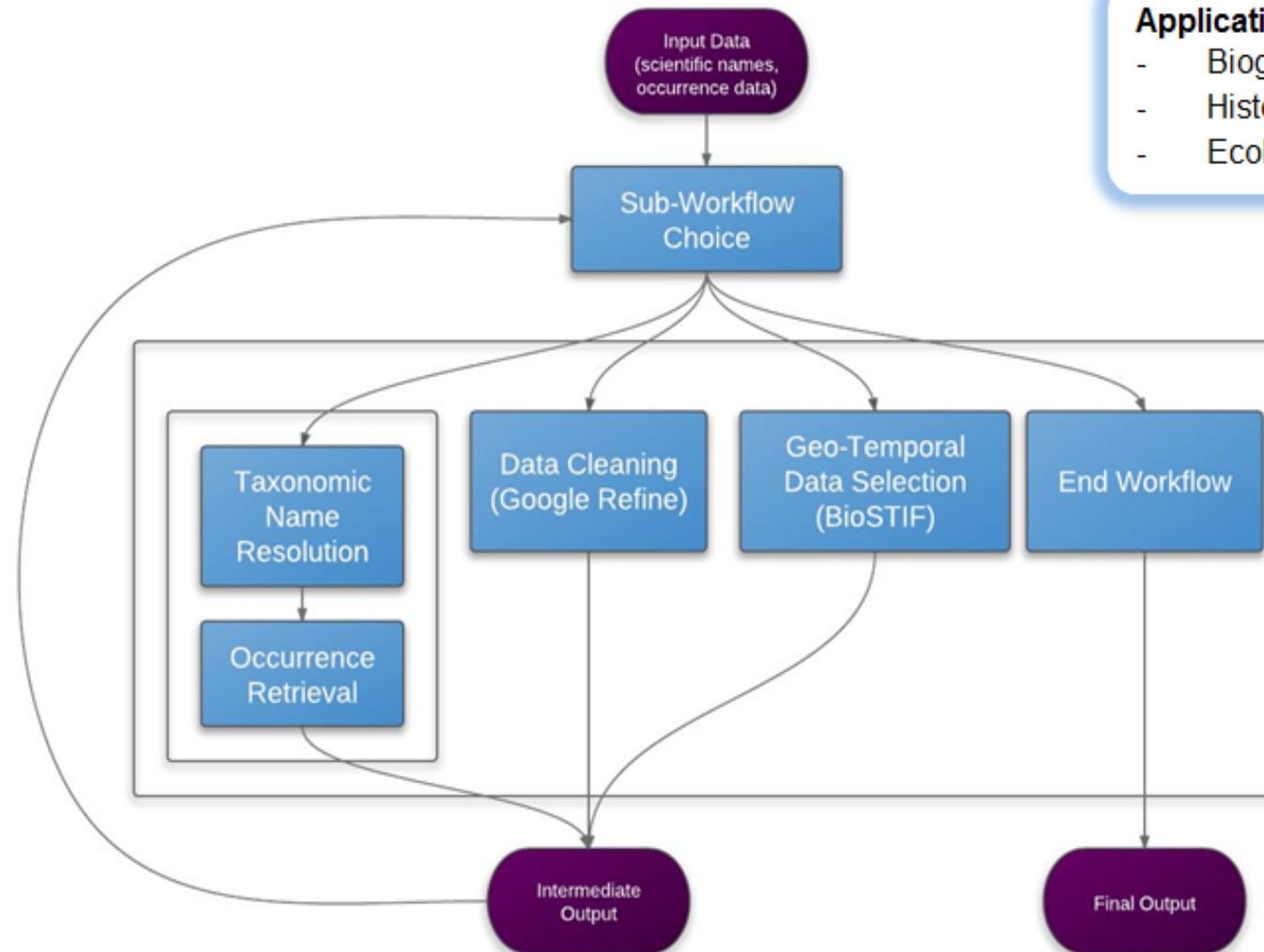


# Introduction to the Data Refinement Workflow and scientific applications





# Data Refinement Workflow



## Application examples

- Biogeographic analysis
- Historical analysis
- Ecological niche modeling

# Taxonomic Data Refinement workflow (DRW)

tavlite1.biovel.eu/runs/969

**BioVeL**

Home Workflows Runs

Running: BioVeL [BETA] Data Refinement Workflow

Action Required

**Choose Sub-Workflow**

- Synonym expansion
- Taxonomic name resolution
- Occurrence retrieval
- Spell checking
- Geographic and taxonomic cleaning
- Temporal refinement
- Data processing log

www.biovel.eu

**Resolve Taxonomic Concept**

Catalogue Of Life World Register of Marine Species

Input Names

Crassostrea gigas

checklist\_url

Accepted Name

☒ Crassostrea gigas (Thunberg, 1793)

Synonym

☒ Crassostrea angulata (Lamarck, 1819)

☒ Crassostrea talienwhanensis Crosse, 1862

☒ Dicoelostrea hispaniola Orton, 1928

☒ Gryphaea angulata Lamarck, 1819

☒ Lopha posjetica (Razin, 1934)

Retrieve (GBIF) species occurrence data CANCEL

Google refine 1366277919951 Permalink

Facet / Filter Undo / Redo

4412 rows

Show as: rows records Show: 5 10 25 50 rows

sample effort	nameComplete	nameAccepted	authorship	rank	Family	Phylum	Class
2	Owenia fusiformis	Owenia fusiformis	della Crijpe, 1844	Species	Oweniidae	Annelida	Polychaeta
2	Echinocardium flavescens	Echinocardium flavescens	(O.F. Mä. ller, 1775)	Species	Loveniidae	Echinodermata	Echinoidea
2	Lepidionotus squamatus	Lepidionotus squamatus	(Linnaeus, 1758)	Species	Polynoidae	Annelida	Polychaeta
2	Alenia gelatinosa	Alenia gelatinosa	(M. Sars, 1835)	Species	Polynoidae	Annelida	Polychaeta
2	Suberites ficus	Suberites ficus	(Johnston, 1842)	Species	Suberitidae	Porifera	Demospongia
2	Halidoria oculata	Halidoria oculata	(Pallas, 1786)	Species	Chalinidae	Porifera	Demospongia
2	Leptochiton asellus	Leptochiton asellus	(Gmelin, 1791)	Species	Leptochitonidae	Mollusca	Polyplocop
2	Pecten maximus	Pecten maximus	(Linnaeus, 1758)	Species	Pectinidae	Mollusca	Bivalvia
2	Ocnus lacteus	Ocnus lacteus	(Forsk. & Goodall, 1839)	Species	Cucumariidae	Echinodermata	Holothuroidea
2	Marthasterias glacialis	Marthasterias glacialis	(Linnaeus, 1758)	Species	Asteriidae	Echinodermata	Asterioida

Selection layer Map selector tools Aggregation type Hide coordinates

80 results with location information

34 results with time information

2006 2007 2008 2009 2010

## *Historical analyses of biodiversity*



Register on the portal  
<http://portal.biovel.eu/>



# Demo - Taxonomic Name Resolution and Occurrence Retrieval

LESSON B (9.20-9.30)

1. Load input file with 6 species names

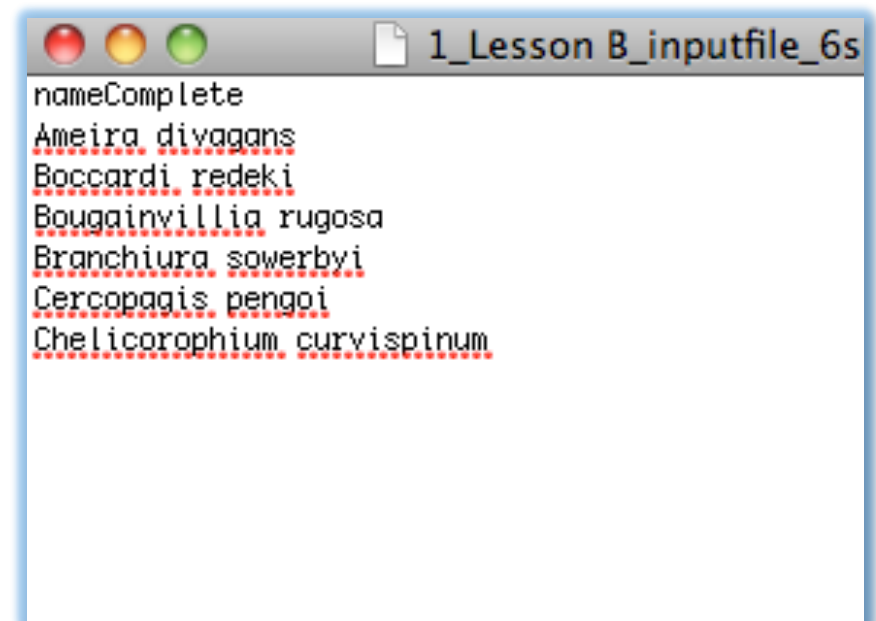
2. Expand names

3. Retrieve records

4. Taxonomic refinement

- data quality
- re-name
- Clustering (NJ, PPM, 5, 8)
- delete recs
- data process log
- Export

5. Geo-temporal refinement



```
nameComplete
Ameira divagans
Boccardi redeki
Bougainvillia rugosa
Branchiura sowerbyi
Cercopagis pengoi
Chelicerophium curvispinum
```

*Input file*

# Practical – taxonomic data refinement

## LESSON A (9.30-11.00, incl. coffee?)

1. Divide in groups of 2
  - 6 groups: <http://portal.biovel.eu/>
  - 6 groups: <https://workshop.at.biovel.eu>
3. Start DRW
  - Load input file LessonA\_Inputfile\_105recs\_v3.csv
  - Choose 'Data Quality'
  - Run tutorial p. 16-38
  - Answer exercises 1-7
4. Short discussion of results

# Assignment (11.00-12.00)

You want to study the potential distribution of the invasive oyster (*Crassostrea gigas*) using species distribution modeling approaches.

You have collected occurrence records of the species in your region (Scandinavia) and want to enrich your records with public data from GBIF, and thereafter create, test, and project an ecological niche model for the species under various climate scenarios.

1. Generate an input file to download GBIF data with the DRWorkflow
2. Retrieve, clean, and refine occurrence data for this species from GBIF
3. Integrate your data with the GBIF records

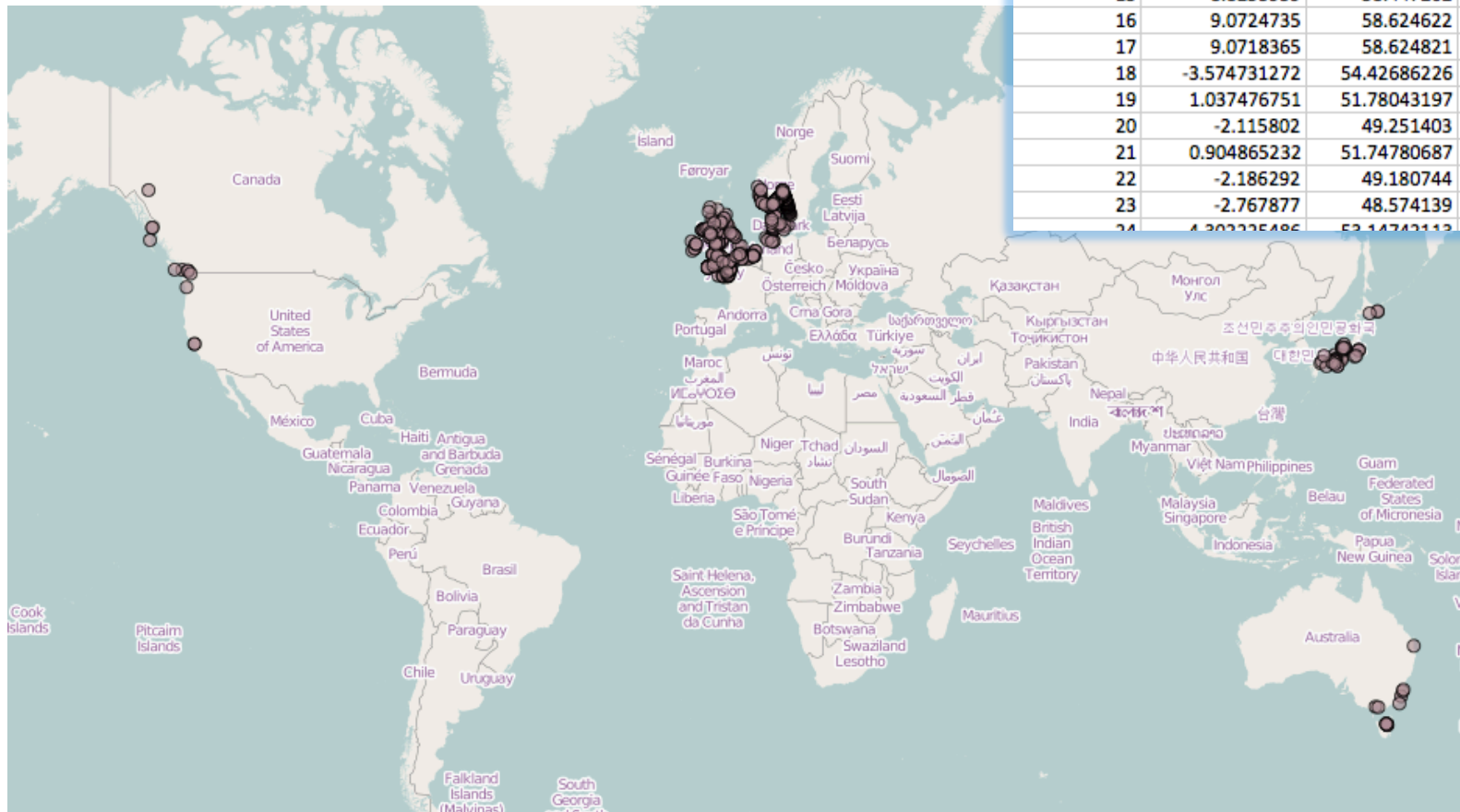
Lunch

4. Create model
5. Test model
6. Project model
7. Statistical analysis of projections

## ENM input files with 4 columns

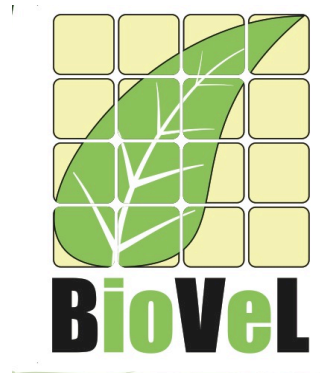
- ID
- Lat
- Long
- Species name

occurrenceID	decimalLongitude	decimalLatitude	nameComplete
1	8.428	55.0315	Crassostrea gigas
2	8.4339	55.0304	Crassostrea gigas
3	8.4314	55.0312	Crassostrea gigas
4	8.4314	55.0312	Crassostrea gigas
5	8.428	55.0315	Crassostrea gigas
6	8.4339	55.0304	Crassostrea gigas
7	8.4172	55.0368	Crassostrea gigas
8	8.428	55.0315	Crassostrea gigas
9	8.4314	55.0312	Crassostrea gigas
10	8.4339	55.0304	Crassostrea gigas
11	8.4314	55.0312	Crassostrea gigas
12	8.4339	55.0304	Crassostrea gigas
13	8.428	55.0315	Crassostrea gigas
14	8.4172	55.0368	Crassostrea gigas
15	8.8255959	58.447262	Crassostrea gigas
16	9.0724735	58.624622	Crassostrea gigas
17	9.0718365	58.624821	Crassostrea gigas
18	-3.574731272	54.42686226	Crassostrea gigas
19	1.037476751	51.78043197	Crassostrea gigas
20	-2.115802	49.251403	Crassostrea gigas
21	0.904865232	51.74780687	Crassostrea gigas
22	-2.186292	49.180744	Crassostrea gigas
23	-2.767877	48.574139	Crassostrea gigas
24	1.202225485	52.14743112	Crassostrea gigas





# Introduction to Species Distribution Modeling



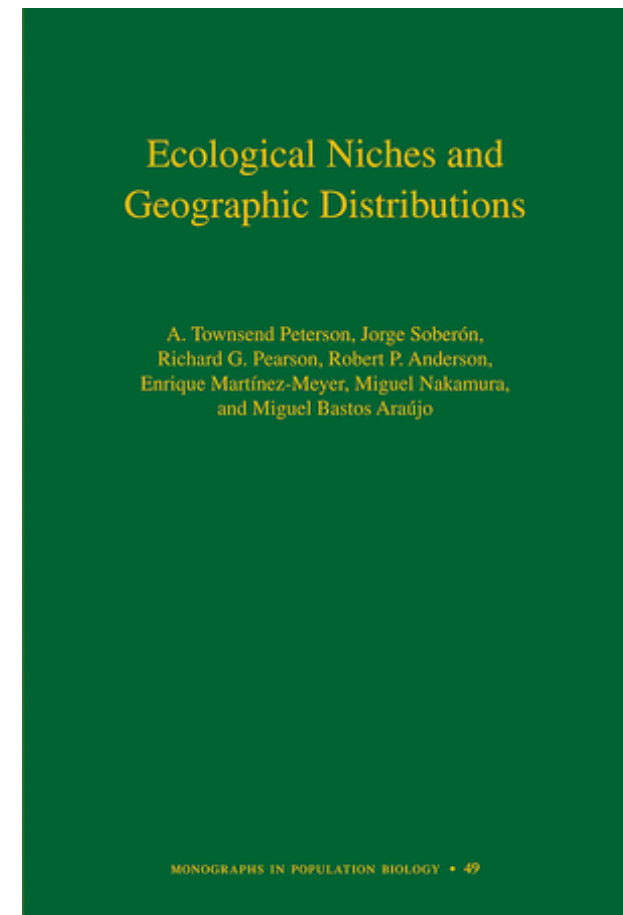
# Literature

**Townsend Peterson, A., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011).** Ecological Niches and Geographic Distributions (Monographs in population biology; no. 49). Princeton University Press. 328 pp. ISBN: 9780691136882 (hard cover), 9780691136868 (Paperback) and 9781400840670 (eBook).

**Pearson, R.G. (2007)** Species' Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History. Available at <http://ncep.amnh.org>

**Elith, J., et al.** Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 2006. 29(2): p. 129-151.

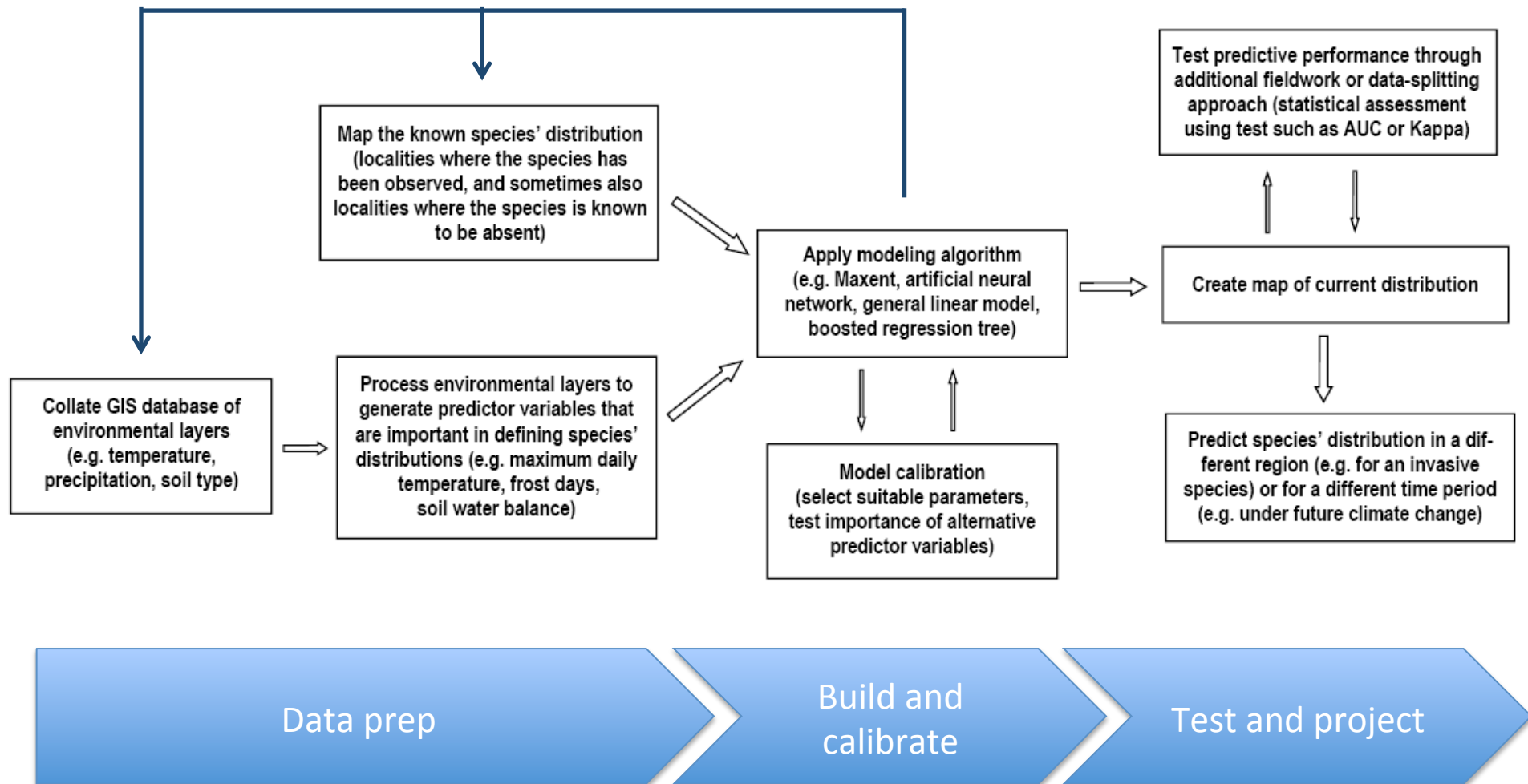
**Soberón, J. and A.T. Peterson.** Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2005. 2: p. 1-10.



# Model types

Model type	Principle	Advantages	Disadvantages
Correlative (statistical extrapolation)	Model distribution by Correlating environmental conditions with a species' occurrence	Only two data types (species occurrence, environmental)  Computational modest	Shows only correlations  Can not predict beyond the observational boundaries
Mechanistic (process-based models)	Model distributions from estimates of responses to environmental conditions	Incl. physiological responses to environment  Projection beyond observed conditions	Resource intense  Require detailed knowledge of physiological responses  Little data available

# Principal steps required for building and validating a correlative Species distribution model



# The basic approach of ENM

1. a study area is modeled as a raster map composed of grid cells at a specified resolution
2. the dependent variable is the known species' distribution
3. a suite of environmental variables are collated to characterize each cell
4. a function of the environmental variables is generated so as to classify the degree to which each cell is suitable for the species



# Important factors influencing the quality of the outputs

## Data

A model is only as good as the data it contains. The data preparation and choice of layers is just as important as the modeling.

## Model extrapolation

‘Extrapolation’ is when you make predictions outside the observational boundaries.

For example, if a distribution model was calibrated within temp. range of 10–20 C, and the model is projected into a temp. range of 10-25 C, then the model is extrapolating and the prediction may be very uncertain.

# The niche concept

Hutchinson defined **the *fundamental niche*** of a species as the set of environmental conditions within which a species can survive and persist.

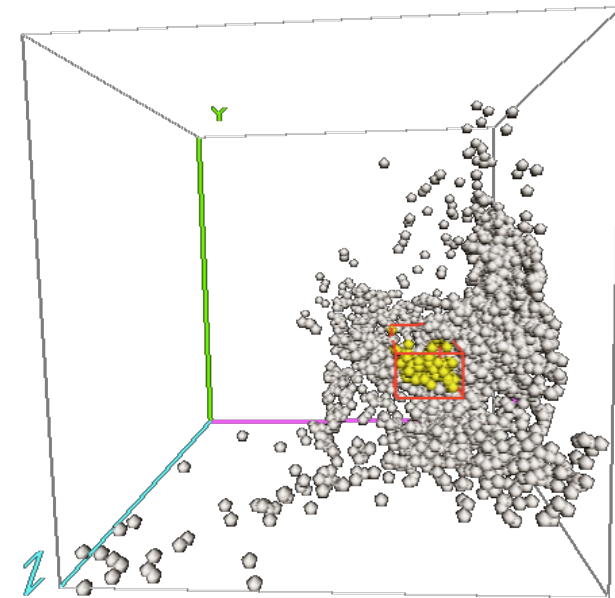
The fundamental niche may be thought of as an ' $n$ -dimensional hypervolume', every point in which corresponds to a state of the environment that would permit the species to exist indefinitely (Hutchinson, 1957)

Fundamental niche (environmental-space) =  
Potential distribution (geographical-space)

Occupied niche (environmental-space) =  
Actual distribution (geographical-space)

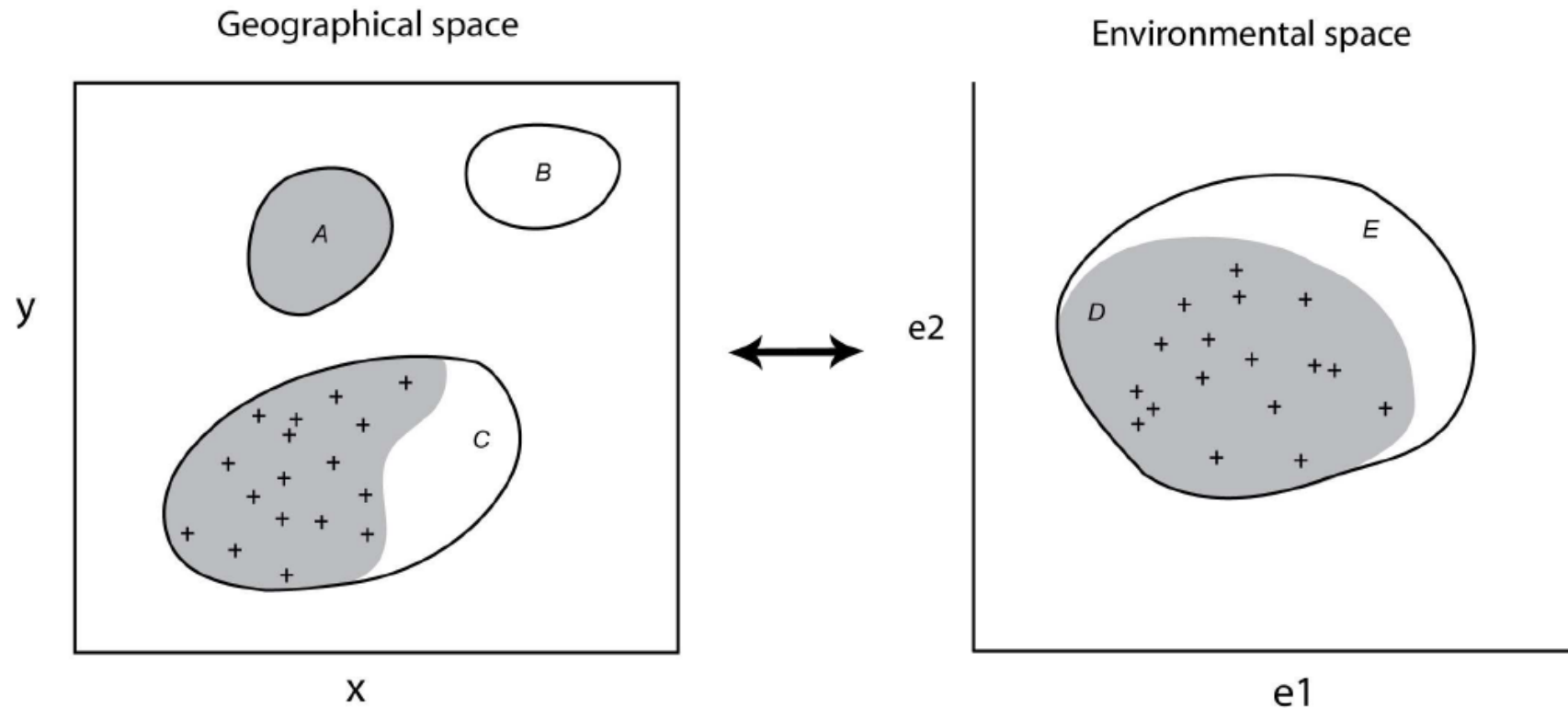
Constraints that influence the potential from  
actual distribution

- Ecological interactions (predators, competition, parasites)
- Dispersal barriers
- Historical



*Fundamental niche depicted in  
environmental-space*

# Geographical versus environmental space



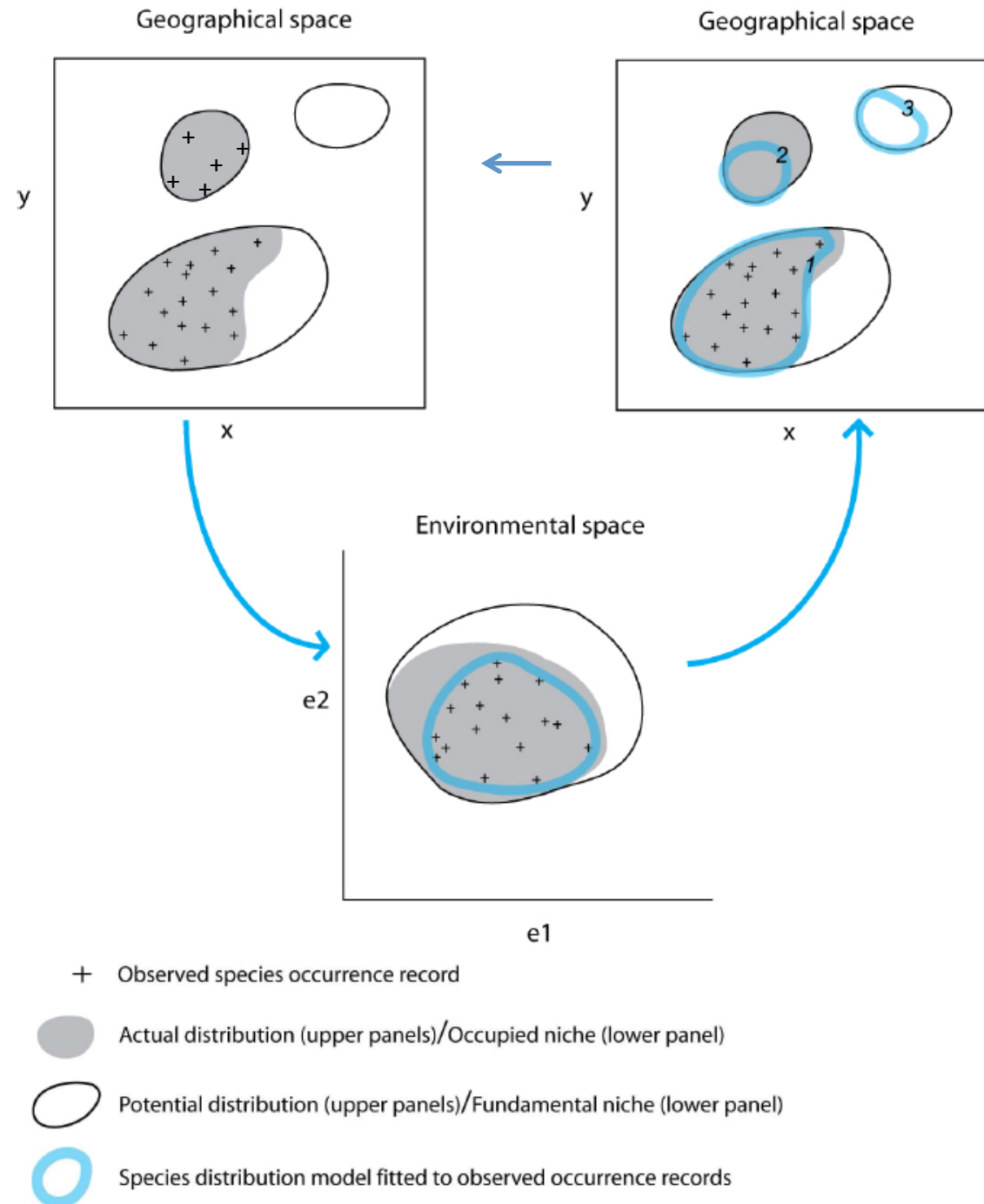
+ Observed species occurrence record

● Actual distribution (left panel)/Occupied niche (right panel)

○ Potential distribution (left panel)/Fundamental niche (right panel)

# Species' distribution models may identify

- 1) Protection areas**, the area around the observed occurrence records that is expected to be occupied (area 1)
- 2) Unobserved populations**, currently unknown distributions (area 2).
- 3) Invasive or re-introduction areas**, potential distribution that is not occupied (area 3)



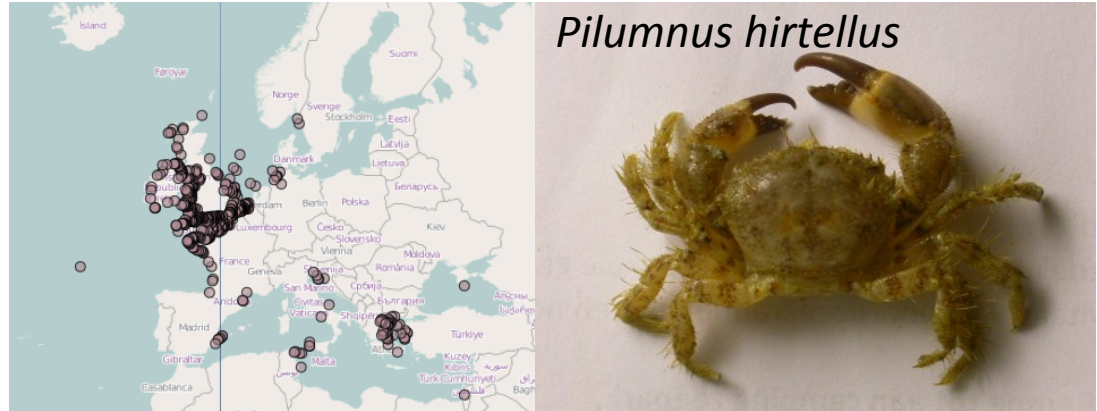
# Species data

## Data types

- Presence only
- Presence/absence
  - only for well studied species
  - False absence lead to serious bias

## Quality criteria

- Should be well distributed in g-space (for projecting the model, enclosed sea problem) and e-space (for building the model)
- Resolution: what resolutions makes sense for the organism and question you investigate
- Species data typically need thorough cleaning/refinement



*Enclosed sea problem (Ready et al., 2010)*

## Species data sources

- Personal
- inventories and museums
- Colleagues and networks
- Online resources
- Literature

## Sources of errors

- Are the species records sustained observations (sink vs source)
- Incorrect identification
- Innacurate spatial reference
- Sampling bias (along roads/rivers)

# Environmental data

## Types

- Categorical (habitat maps)
- Continuous (salinity, temperature)

## Formats

- Point vector data (typically converted to Grid/raster data)
- Polygon vector data (typically converted to Grid/raster data)
- Grid/raster data

## Resolution

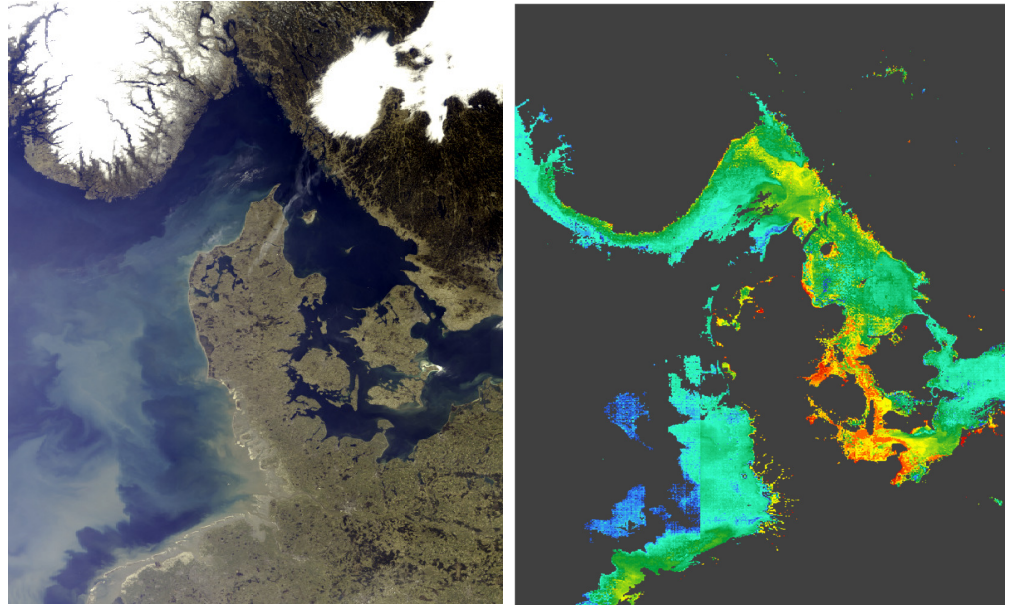
- Should be conform with species data resolution

## Variables

- Climatic (temperature)
- Geographic (altitude)
- Habitat (soil type)
- Ecological (habitat builders)

## Number of layers

- Empirical evidence shows that 4-8 layers are sufficient to generate good models
- Quality before quantity





# Algorithms

The algorithm identifies environmental conditions that are associated with species occurrence

The choice of algorithm depends on your question and what the model should produce, e.g. if you want to protect a species you should identify ***actual distributions***, but if you want to re-introduce a species you should identify ***potential distributions***.

The choice of algorithms also depends on

- the quality/quantity of your species data
- use presence-only, presence-background, presence-absence
- categorical vs. continuous environmental data
- give binary vs. continuous predictions
- causality vs. predictability

Good algorithms are those that minimize predictions of areas that are neither the actual nor the potential distribution

# Algorithms

MaxEnt is very good for predictions inside the observational boundaries, but can generate faulty extrapolations

Algorithms that can incorporate interactions among variables are preferable (Elith et al. 2006), e.g. a more accurate description of a plant's requirements may be that it can occur at localities with mean monthly precipitation between 60mm and 70mm if soil clay content is above 60%, and in wetter areas (>70mm) if clay content is as low as 40%.

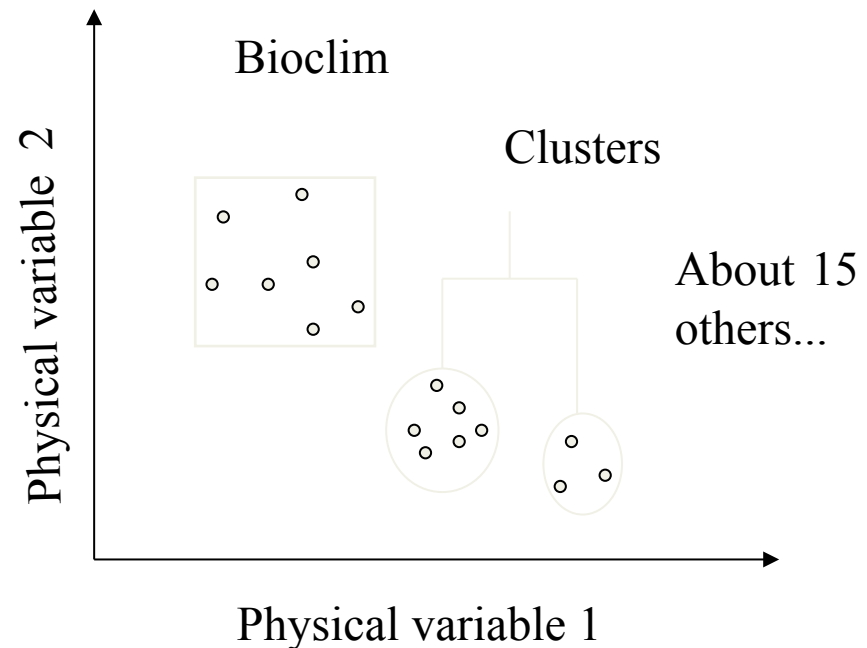


Table 4. Modelling methods implemented.

Method	Class of model, and explanation	Data <sup>1</sup>	Software	Std errors? <sup>2</sup>	Contact person
BIOCLIM	envelope model	p	DIVA-GIS	no	CG, RH
BRT	boosted decision trees	pa	R, gbm package	no	JE
BRUTO	regression, a fast implementation of a gam	pa	R and Splus, mda package	yes	JE
DK-GARP	rule sets from genetic algorithms; desktop version	pa	DesktopGarp	no	ATP
DOMAIN	multivariate distance	p	DIVA-GIS	no	CG, RH
GAM	regression: generalised additive model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
GDM	generalised dissimilarity modelling; uses community data	pacomm	Specialized program not general released; uses Arcview and Splus	no	SF
GDM-SS	generalised dissimilarity modelling; implementation for single species	pa	as for GDM	no	SF
GLM	regression; generalised linear model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
LIVES	multivariate distance	p	Specialized program not general released	no	JLi
MARS	regression; multivariate adaptive regression splines	pa	R, mda package plus new code to handle binomial responses	yes	JE, FH
MARS- COMM	as for MARS, but implemented with community data	pacomm	as for MARS	yes	JE
MARS-INT	as or MARS; interactions allowed	pa	as for MARS	yes	JE
MAXENT	maximum entropy	pa	Maxent	no	SP
MAXENT-T	maximum entropy with threshold features	pa	Maxent	no	SP
OM-GARP	rule sets derived with genetic algorithms; open modeller version	pa	new version of GARP not yet available	no	ATP

*Elith & al. 2006*

# 3 types of presence-only methods

1. Methods that rely ***solely based on presence records*** (e.g. BIOCLIM), e.g. the prediction is made without any reference to other samples from the study area
2. Methods that use ***'background'*** environmental data for the entire study area (e.g. Maxent, ENFA), e.g. focus on how the environment where the species is known to occur relates to the environment across the rest of the study area. Occurrence localities are also included as part of the background.
3. Methods that sample ***'pseudo-absences'*** from the study area. In The aim here is to assess differences between the occurrence localities and a set of localities chosen from the study area that are used in place of real absence data. Pseudo-absence models do not include occurrence localities within the set of pseudo-absences.

# Assessing predictive performance

## Strategies for obtaining test data

Typically split the data into: 30% test data and 70% calibration data

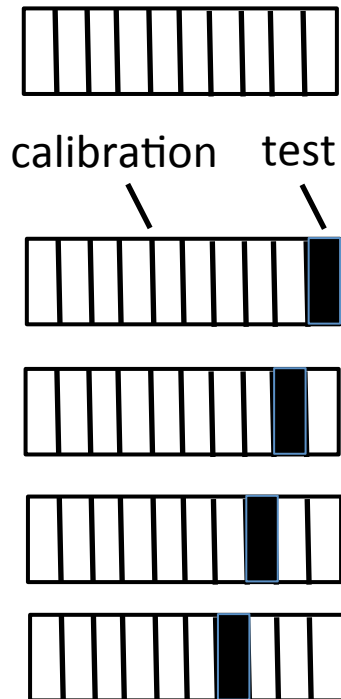
### **Bootstrapping**

- sample the original set of data randomly with replacement
- same occurrence record could be included in the test data more than once
- predictive performance is assessed from multiple re-samplings

### ***k*-fold partitioning**

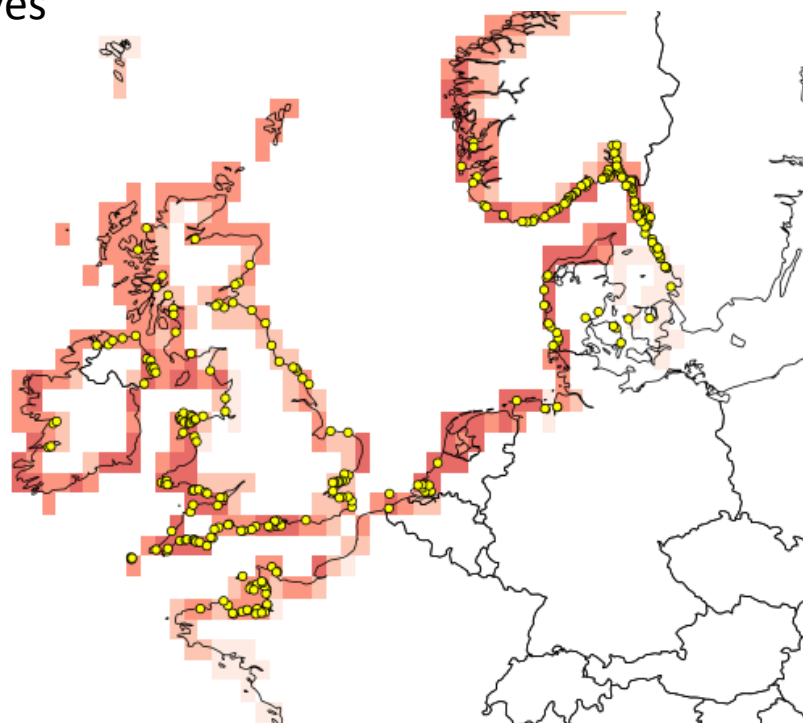
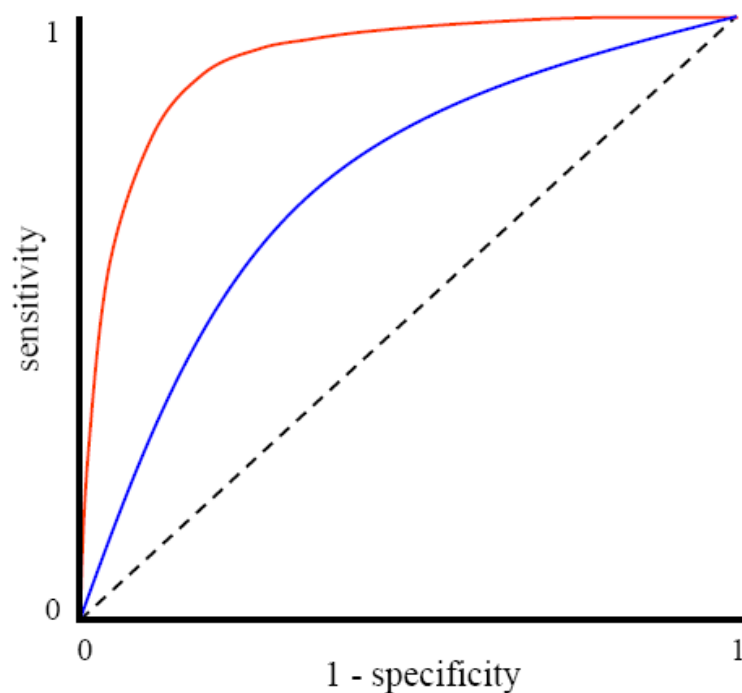
- data are split into  $k$  parts of roughly equal size and each part is used as a test set with the remaining  $(k-1)$  sets used for model calibration.

*10-fold partitioning*



# Assessing predictive performance - validation statistics

Receiver Operating Characteristic (ROC) Curves



## **AUC (area under the ROC curve)**

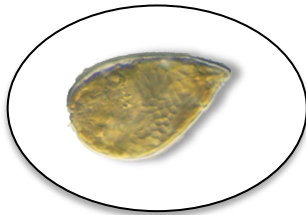
- summarizes the predictive performance across the full range of thresholds
- ranges from 0.5 for models that are no better than random to 1.0 for models with perfect predictive ability



# Examples of species' distribution models in conservation biology

Type of use	Example reference(s)
Guiding field surveys to find populations of known species	Bourg et al. 2005, Guisan et al. 2006
Guiding field surveys to accelerate the discovery of unknown species	Raxworthy et al. 2003
Projecting potential impacts of climate change	Iverson and Prasad 1998, Berry et al. 2002, Hannah et al. 2005; for review see Pearson and Dawson 2003
Predicting species' invasion	Higgins et al. 1999, Thuiller et al. 2005; for review see Peterson 2003
Exploring speciation mechanisms	Kozak and Wiens 2006, Graham et al. 2004b
Supporting conservation prioritization and reserve selection	Araújo and Williams 2000, Ferrier et al. 2002, Leathwick et al. 2005
Species delimitation	Raxworthy et al. 2007
Assessing the impacts of land cover change on species' distributions	Pearson et al. 2004
Testing ecological theory	Graham et al. 2006, Anderson et al. 2002b
Comparing paleodistributions and phylogeography	Hugall et al. 2002
Guiding reintroduction of endangered species	Pearce and Lindenmayer 1998
Assessing disease risk	Peterson et al. 2006, 2007

*Guisan and Thuiller (2005) Ecology Letters*

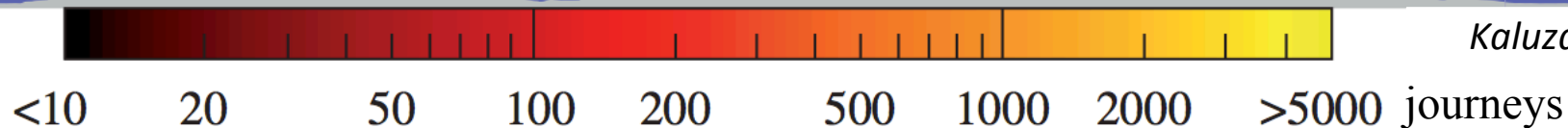
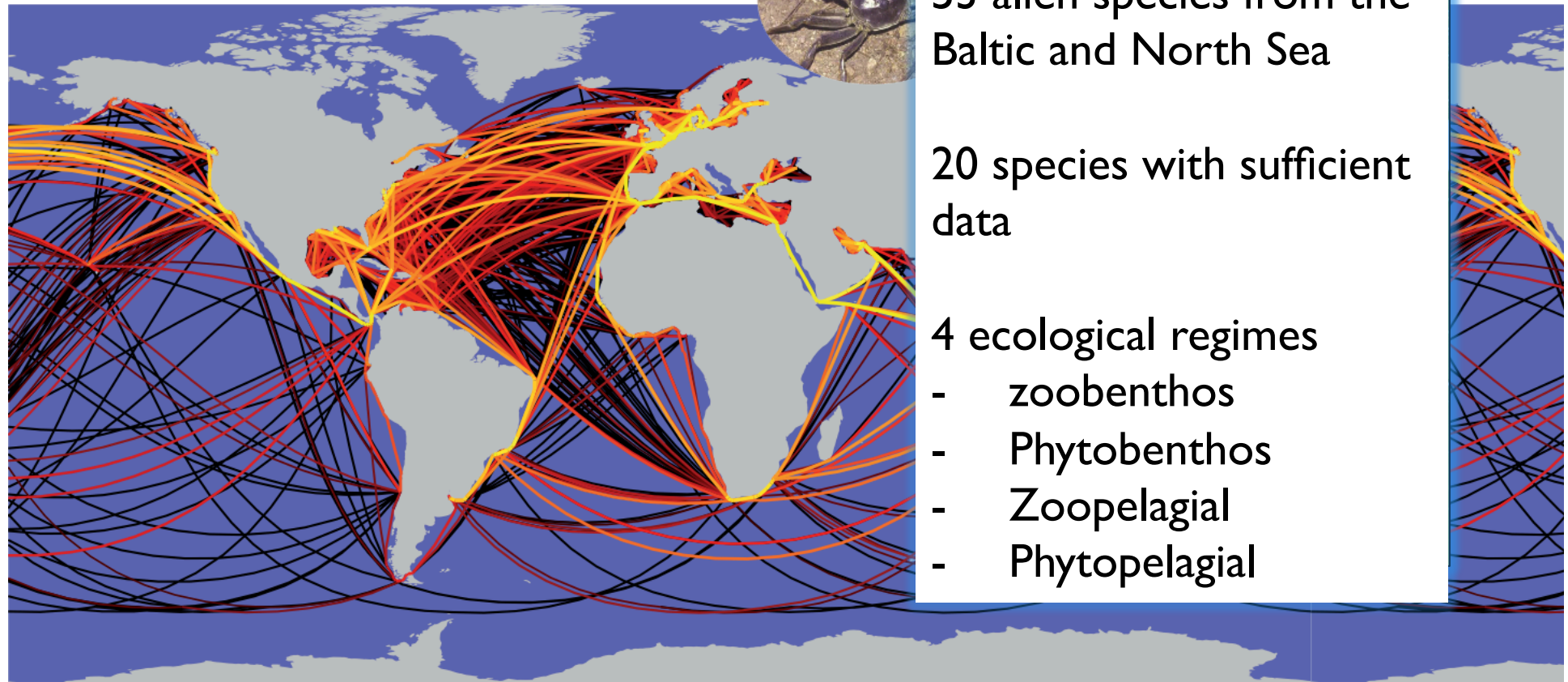


53 alien species from the  
Baltic and North Sea

20 species with sufficient  
data

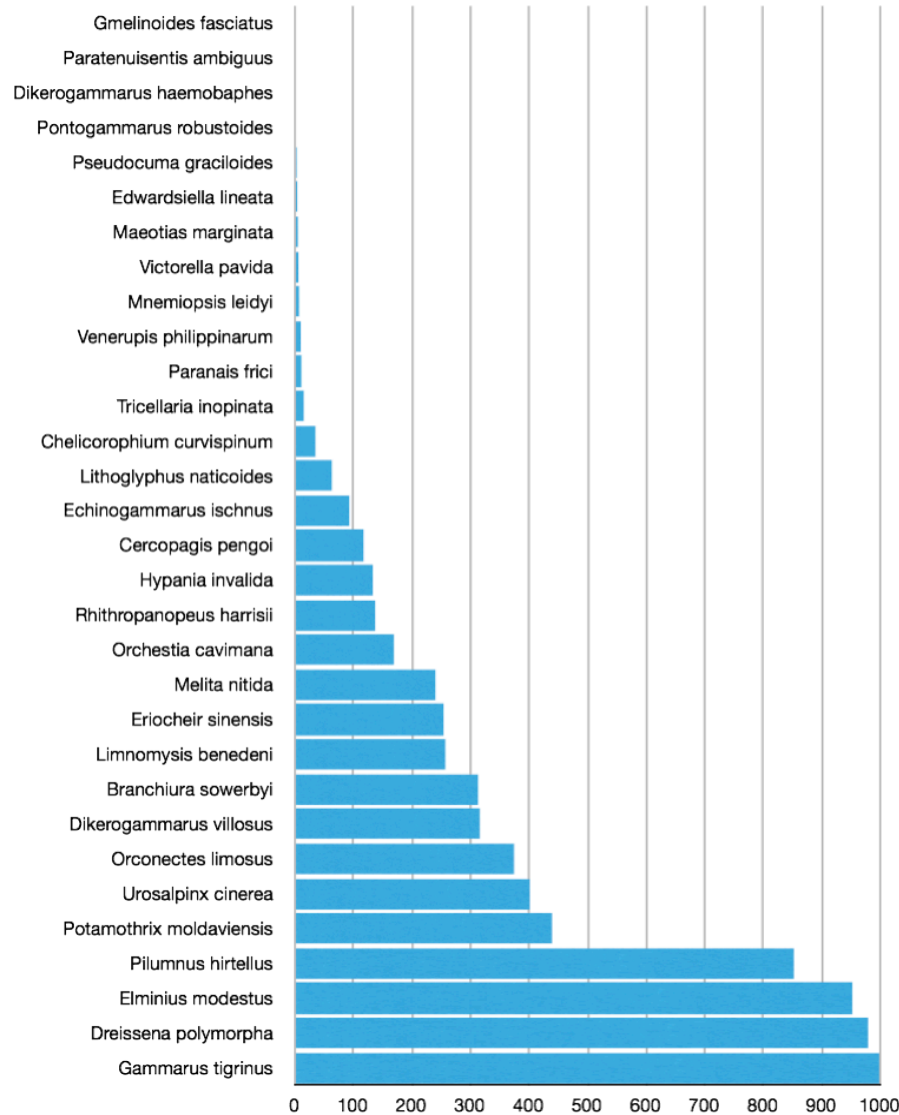
4 ecological regimes

- zoobenthos
- Phytobenthos
- Zoopelagial
- Phytopelagial



*Kaluza et al (2013)*

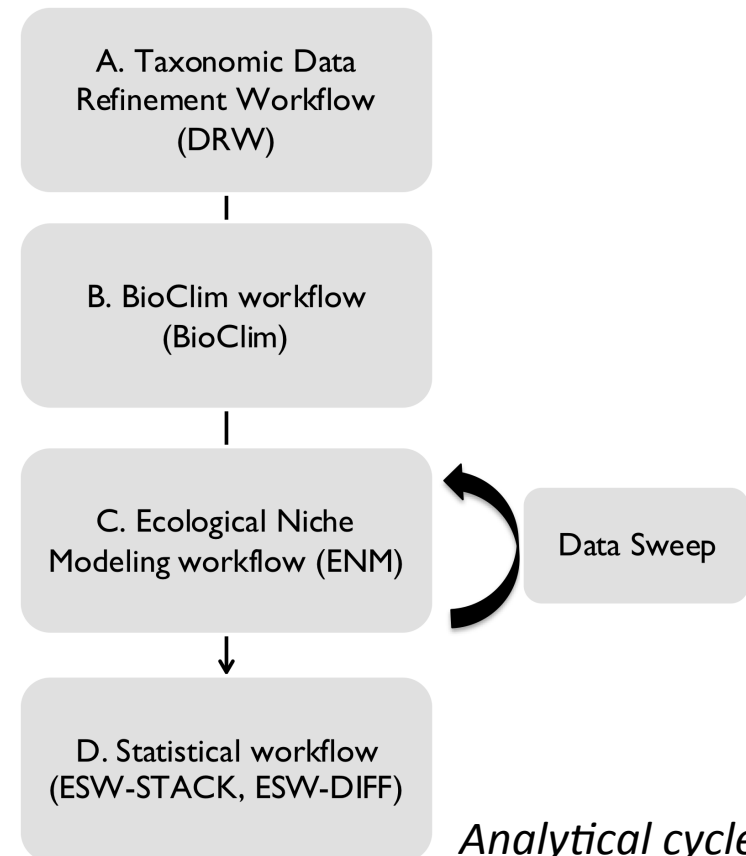
# Data discovery



Occurrence records for 31 species

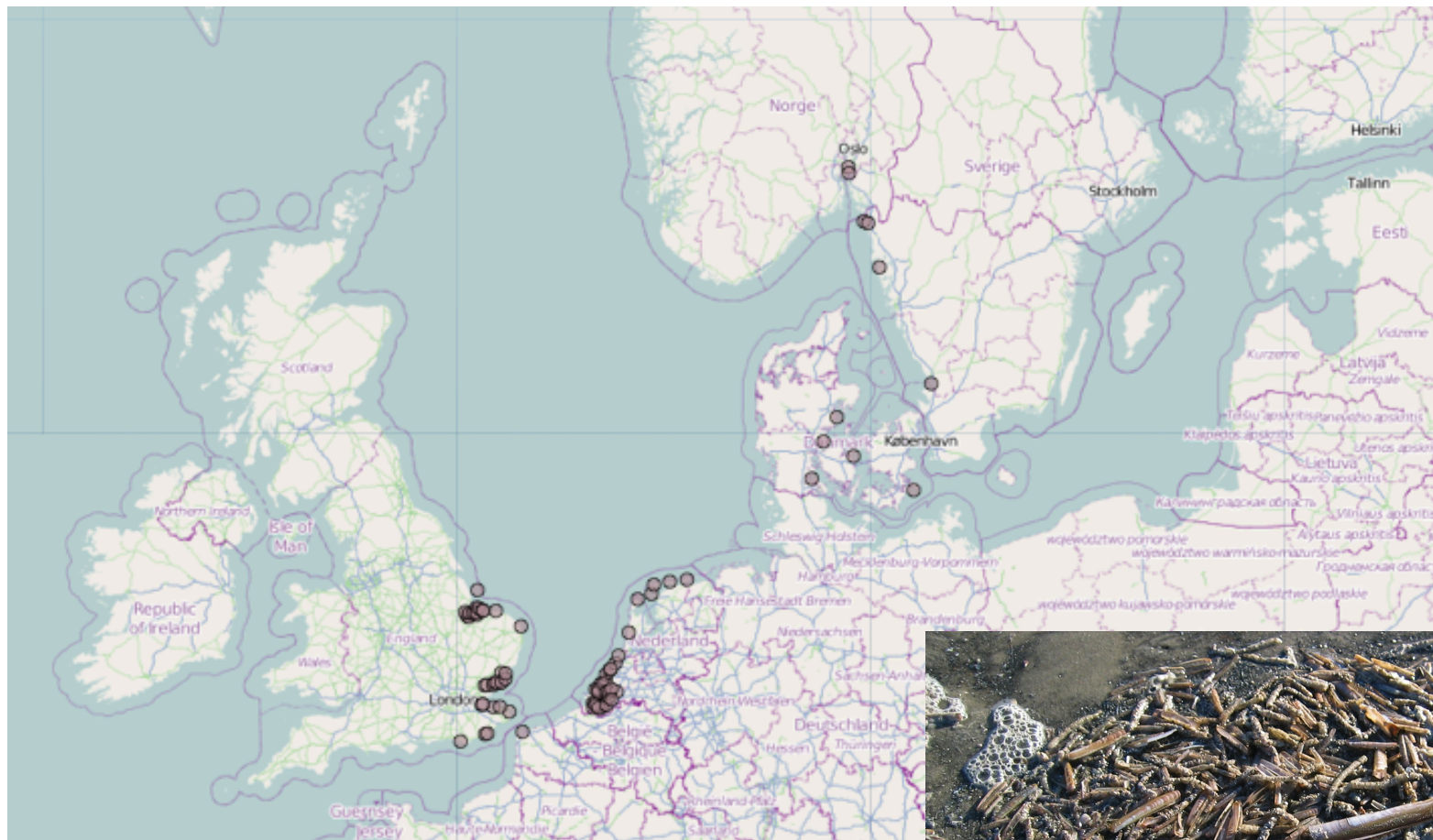
## Sources

- Personal data
- Digital repositories (ICES, GBIF, national environmental agencies)
- Scientific networks
- Literature digitization



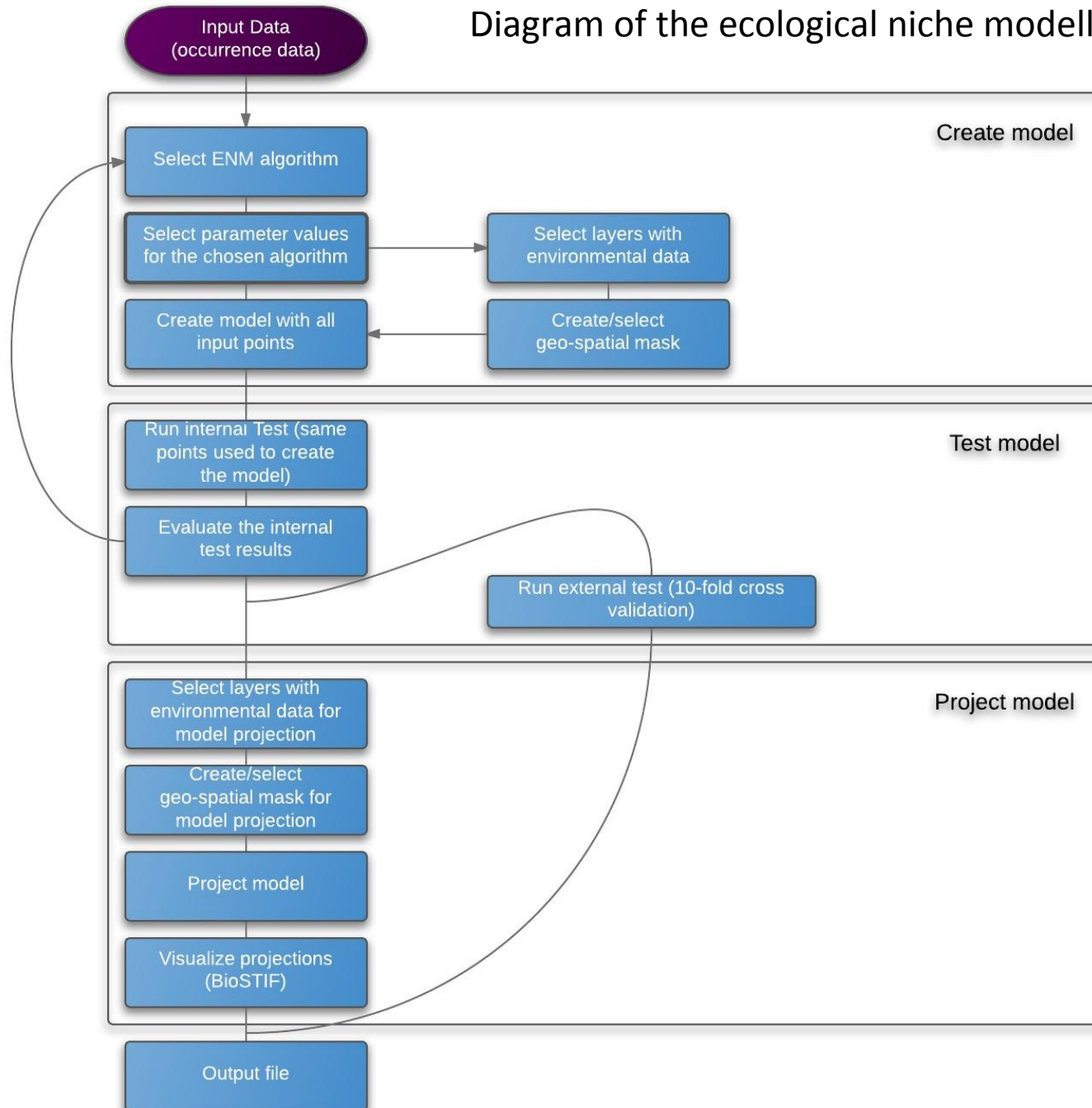
Species, Authors	Eco-group	Origin	Invasion Path	Introduction/ First observation		Total # of records	References of Occurrence data
				North Sea	Baltic Sea		
<i>Austrominius modestus</i> (Darwin, 1854)	ZB	S Pacific	S	1953	-	709	GBIF (709)
<i>Crassostrea gigas</i> (Thunberg, 1793)	ZB	NW Pacific	A, ST	1991	1980s	967	GBIF (967)
<i>Ensis directus</i> (Conrad, 1843)	ZB	NW Atlantic	S	1978/79	1981/1993	817	GBIF (807), Thomsen et al. 2009 (5), <a href="http://www.frammandearter.se">www.frammandearter.se</a> (1), own observations (4)
<i>Eriocheir sinensis</i> H. Milne Edwards, 1853	ZB	NW Pacific	S	1915	1926/1932	740	GBIF (613), Drotz et al. 2010 (46), Normant et al. 2000 (8), Ojaveer et al. 2011 (1), Ojaveer et al. 2007 (68), Otto and Brandis 2011 (4)
<i>Gammarus tigrinus</i> Sexton, 1939	ZB	NW Atlantic	ST	1965	1975/1985	1648	GBIF (1566), Berezina 2007 (2), Guszka 1999 (44), Jazdzowski et al. 2004 (6), Kotta et al. 2013 (26), Strode et al. 2013 (4)
<i>Marenzelleria viridis</i> (Verrill, 1873)	ZB	NW Atlantic	S	1983	2004	789	GBIF (718), Andruliewicz 1997 (4), Bastrop and Blank 2006 (3), Gruszka 1999 (43), Thomsen et al. 2009 (18), Zettler 1996 (3)
<i>Mytilopsis leucophaeata</i> (Conrad, 1831)	ZB	NW Atlantic	S	1835/ <1994	1930s/ <1994/2000	268	GBIF (258), Dziubinska 2011 (1), Laine et al. 2006 (5), Verween et al. 2005 (1), Darr and Zettler 2000 (2), <a href="http://www.frammandearter.se">www.frammandearter.se</a> (1)
<i>Pilumnus hirtellus</i> (Linnaeus, 1761)	ZB	NW Atlantic	S	-	2004	1270	GBIF (1258), Berggren 2012 (10), <a href="http://www.frammandearter.se">www.frammandearter.se</a> (2)
<i>Potamopyrgus antipodarum</i>	ZB	S Pacific	S	1927	1887/1908	990	GBIF (990)





Distribution of the invasive Atlantic jackknife clam (*Ensis directus*) in Europe. Data aggregated from GBIF and scientific networks.

## Diagram of the ecological niche modelling workflow



### Create Model:

- model algorithm
- parameter values
- environmental layer selection
- geospatial mask
- Model created
- background or pseudo-absence points are sampled from the masked region

### Test model:

- statistical evaluation of the model prediction

### Project Model:

- select the layers and masks for model projection
- projections and associated occurrence points are visualized





# Ecological Niche Modeling Workflow (ENM)

**BioVeL Tave**

**BioVeL**

Home Workflows

**Run of Gene**  
**myExperiment**

User: Sarah Bou

Current State: ru

The run was sta

This is the first tim

Cancel Run

**INTERACTION**

Available algorithm

- ☐ AquaMaps (beta)
- ☐ Bioclim
- ☐ Climate Space Model
- ☐ GARP (single run) - DesktopGARP implementation
- ☐ GARP with best subsets - DesktopGARP implementation
- ☐ ENFA (Ecological-Niche Factor Analysis)
- ☐ Envelope Score
- ☒ Environmental Distance
- ☐ GARP (single run) - new openModeller implementation
- ☐ GARP with Best Subsets - new openModeller implementation
- ☐ Maximum Entropy
- ☐ Niche Mosaic
- ☐ Artificial Neural Network

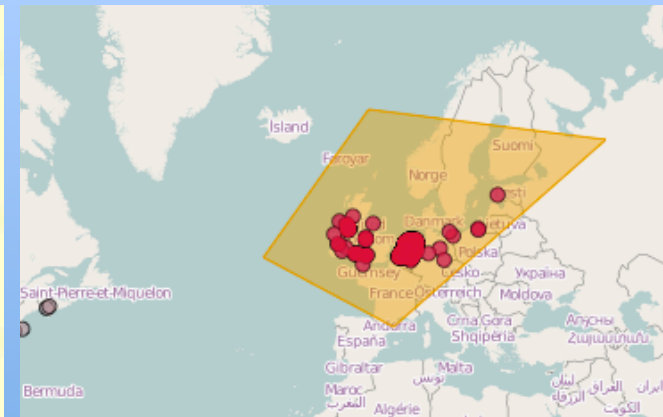
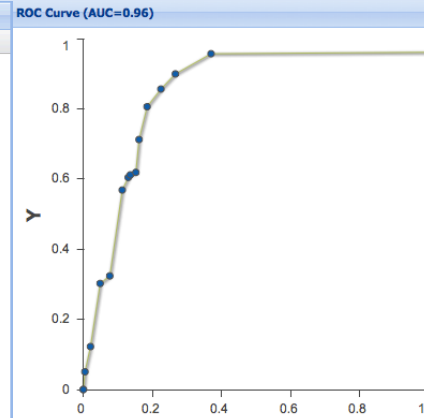
Select layers to create the model

- ☒ HCAFr4
  - ☒ geography
    - ☐ 30arc-minutes
      - ☐ Mean depth in meters
      - ☒ Mean annual distance to land in Kilometers
      - ☐ Minimum depth in meters
      - ☒ Maximum depth in meters
  - ☒ present
    - ☐ 30arc-minutes
      - ☐ Mean annual bottom salinity in psu
      - ☐ Mean annual surface salinity in psu
      - ☒ Mean annual bottom temperature in Celsius
      - ☐ Mean annual surface temperature in Celsius
      - ☒ Mean annual sea ice concentration
      - ☐ Mean annual primary production (chlorophyll A) in mgC/m2/day
- ☐ terrestrial
- ☐ climate

Submit selected layers

**Confusion matrix**

Name	Value
Accuracy	84.32%
CommissionError	-
FalsePositives	422
OmissionError	15.68%
Threshold	0.5
TruePositives	2269



when used with the Gower metric and maximum distance 1, this algorithm should produce the same result of the algorithm known as DOMAIN.

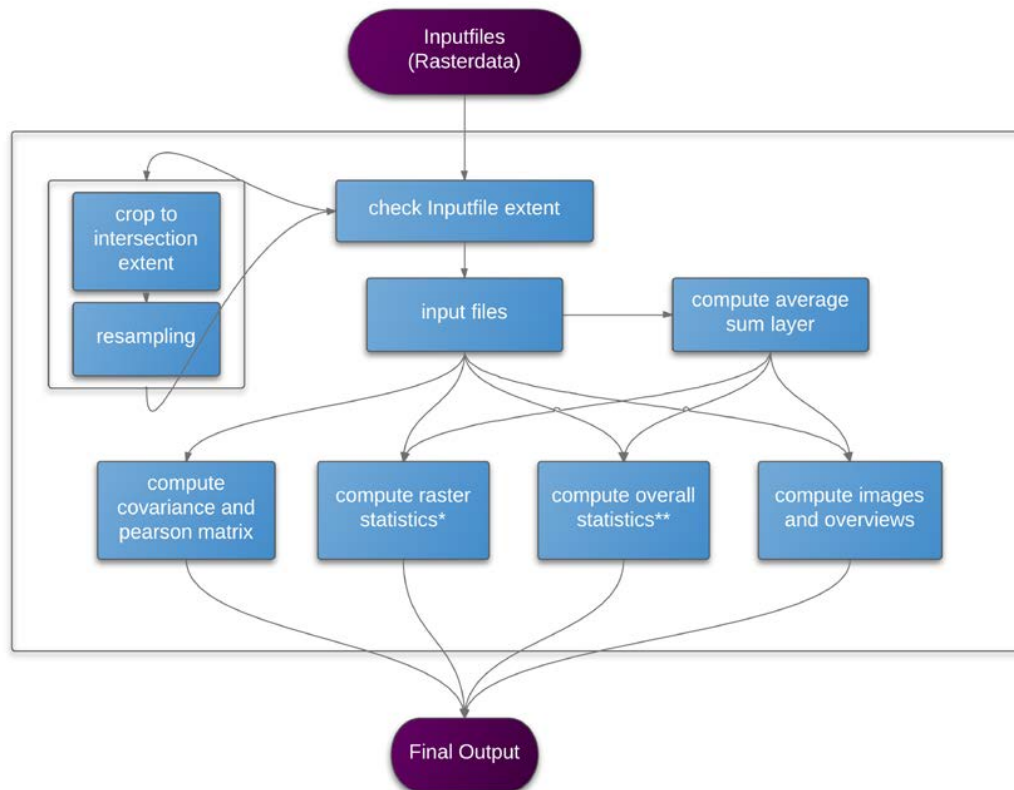
**Author(s)** Mauro E. S. Munoz, Renato De Giovanni, Danilo J. S. Bellini

**Bibliography** Carpenter G, Gillison AN, Winter J (1993) DOMAIN: A flexible modeling procedure for mapping potential distributions of animals and plants. *Biodiversity and Conservation* 2: 667-680. Farber O & Kadmon R 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling* 160: 115&130.

**Developer(s)** Danilo J. S. Bellini, Renato De Giovanni

# ENM Statistical Workflows

## ESW STACK workflow using R

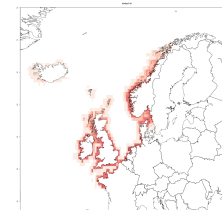


\*computed raster statistics: number of cells, mean, median, coefficient of variation (cv), standard deviation (sd), min, max

\*\*computed overall statistics:

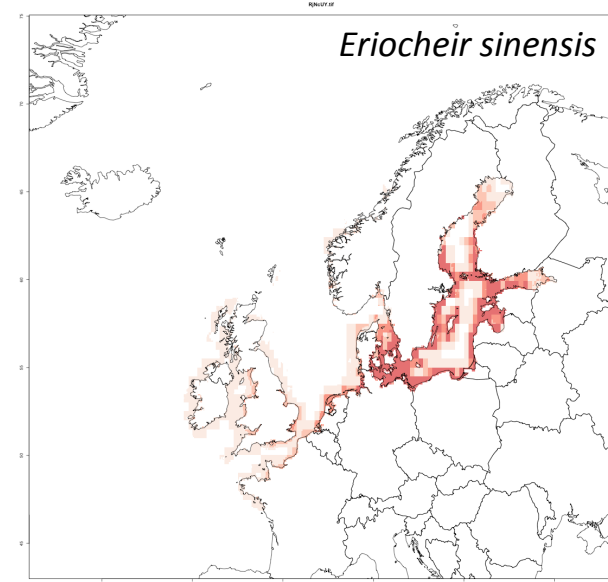
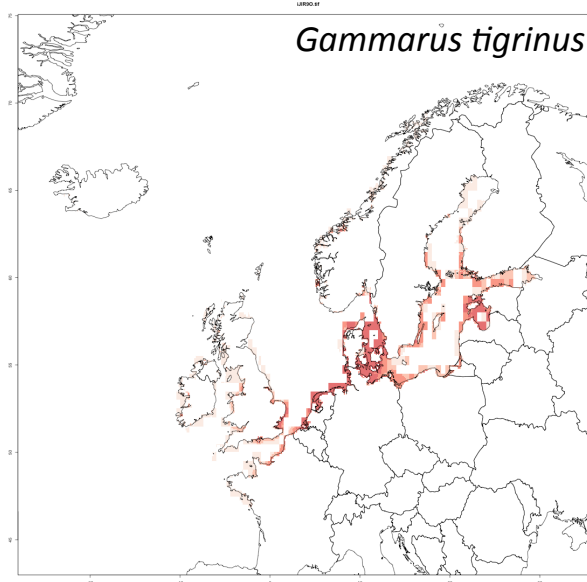
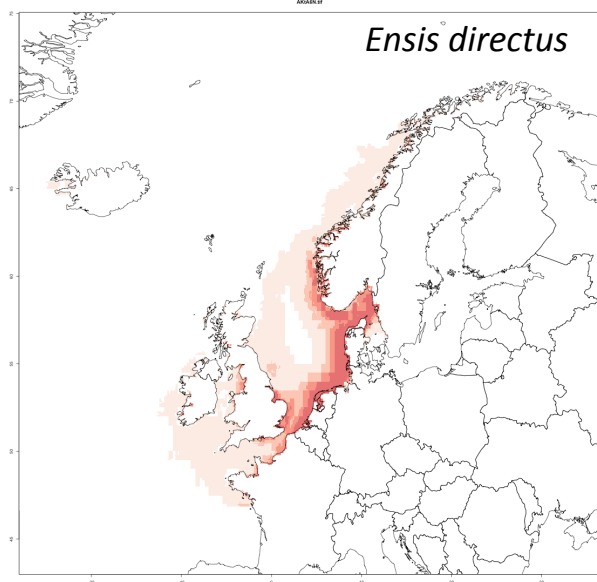
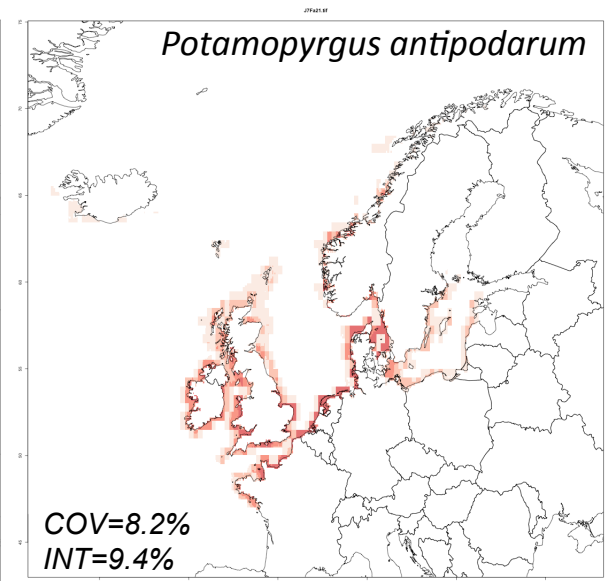
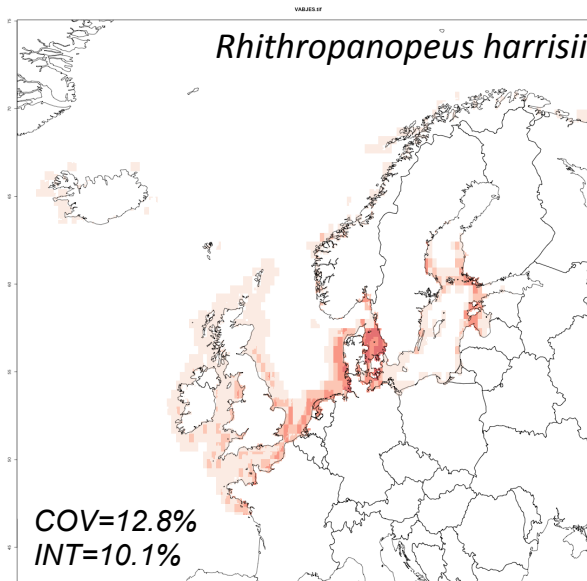
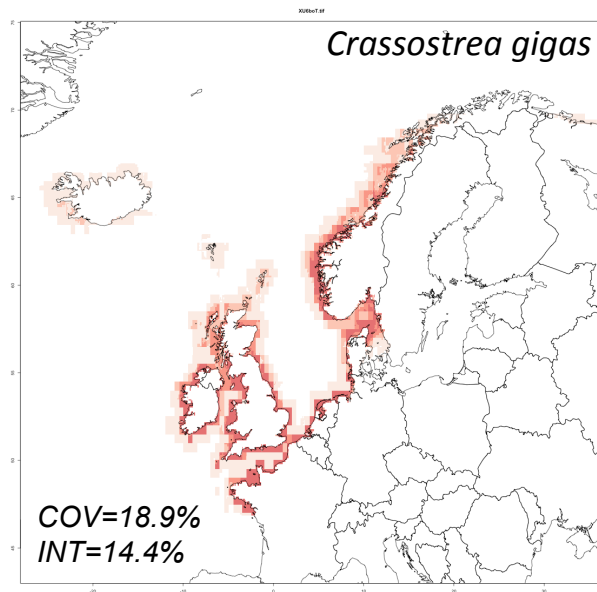
- overall coverage as the percentage of raster cells with values >0,
- overall intensity as the sum of all valued cells divided by the number of raster cells,
- differences between the coverage and intensity between the input files

- 1) ESW DIFF
- 2) ESW STACK
- 3) ESW SHIFT

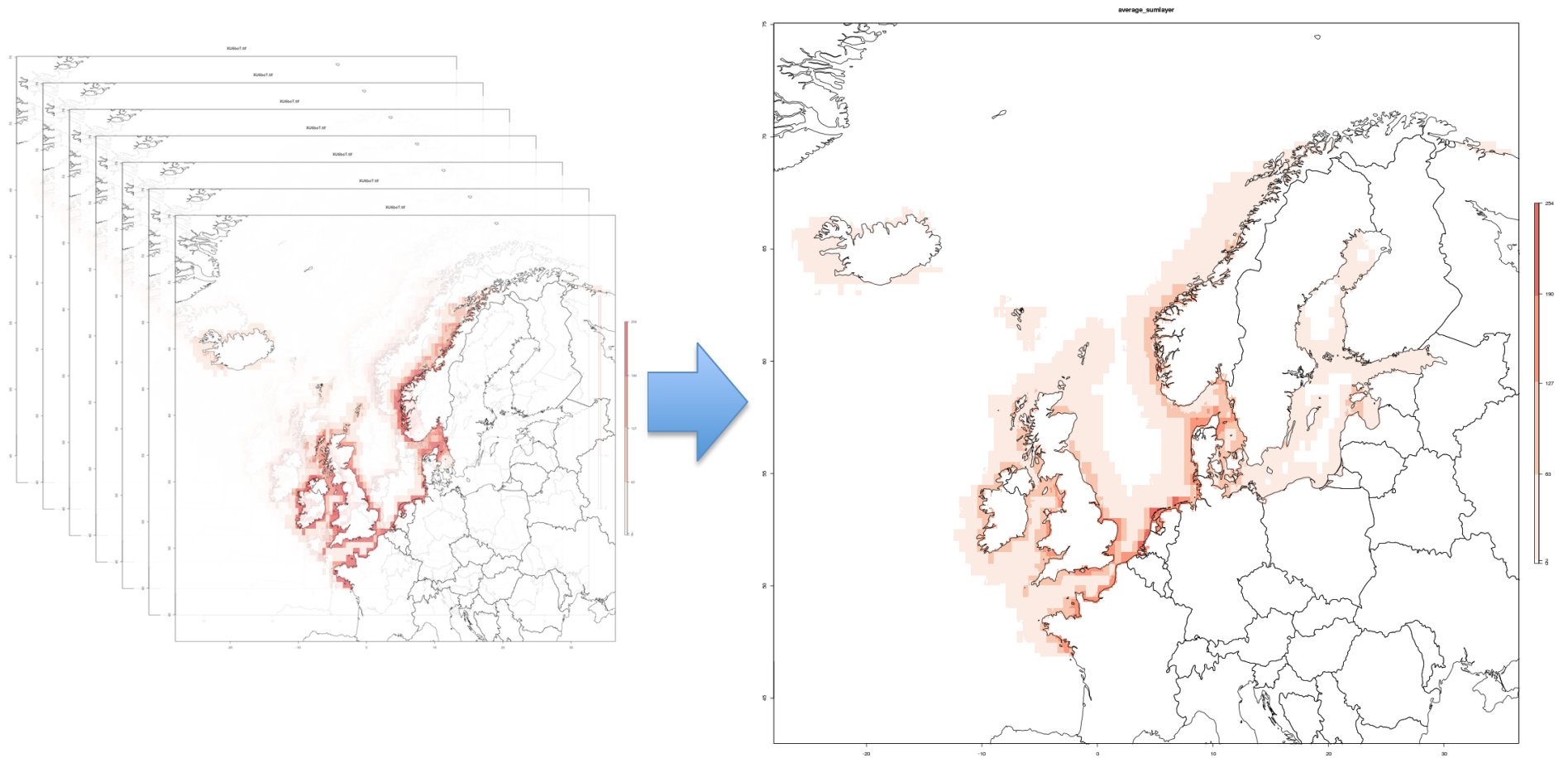


XX	Number of raster cells	Mean of all raster cell values	Median of all raster cell values	Coefficient of Variation
currentLayer	643104	2.889023	0	472.7624667
predictionLayer	643104	2.8119792	0	462.6982272
diffLayer	643104	-0.077043835	0	-11966.1249

# Invasive heatmaps

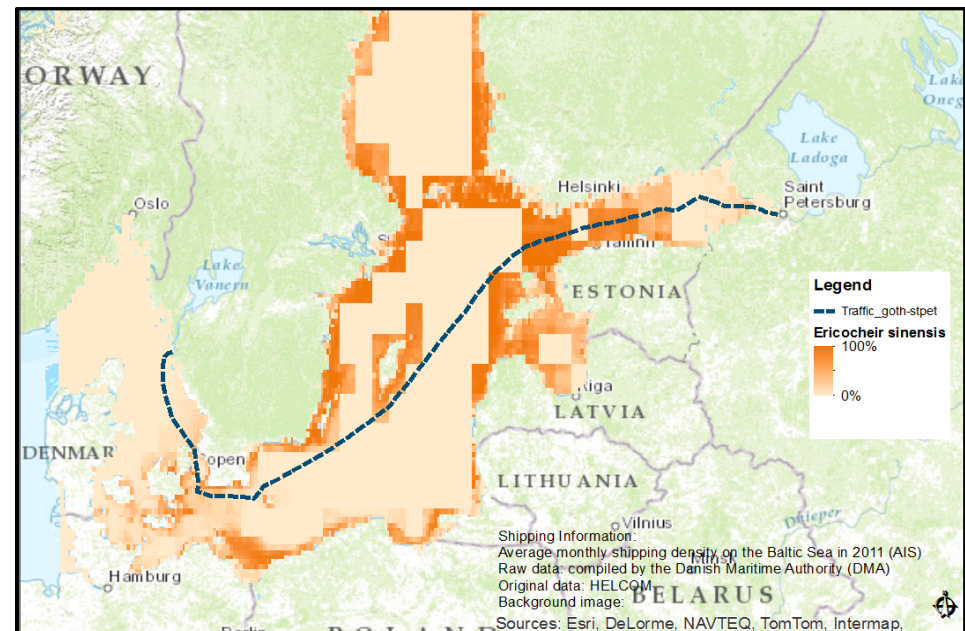


# Accumulated invasive potential for ecological groups



*Stack of macrozoobenthic invasion heatmaps*

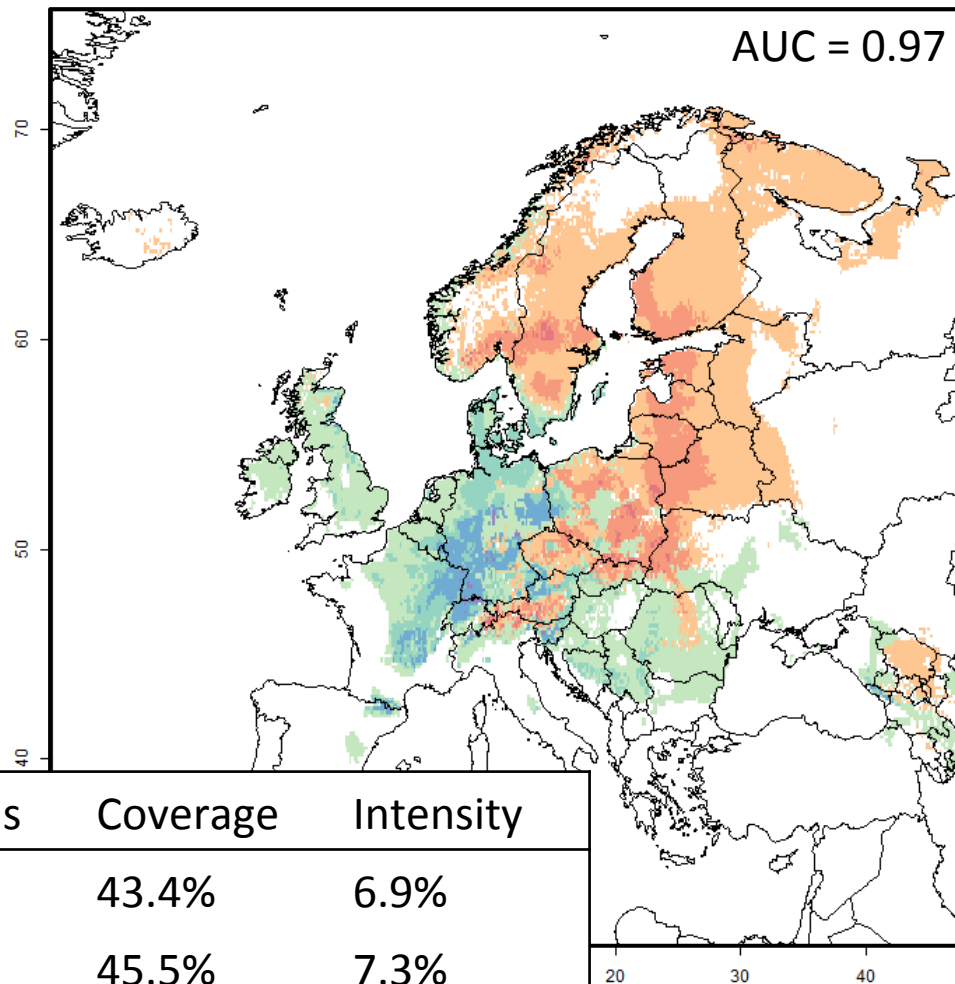
# Accumulated invasive potential







# Dynamic projection of forest pest species based on distribution of host trees



# Assignment (14.00-16.20)

You want to study the potential distribution of the invasive oyster (*Crassostrea gigas*) using species distribution modeling approaches.

You have collected occurrence records of the species in your region (Scandinavia) and want to enrich your records with public data from GBIF, and thereafter create, test, and project an ecological niche model for the species under various climate scenarios.

1. Generate an input file to download GBIF data
2. Retrieve, clean, and refine occurrence data
3. Integrate your data with the GBIF records

Lunch

4. Create model
5. Test model
6. Project model for 2013 and 2050
7. Statistical analysis of difference between projections

occurrenceID	decimalLongitude	decimalLatitude	nameComplete
1	8.428	55.0315	Crassostrea gigas
2	8.4339	55.0304	Crassostrea gigas
3	8.4314	55.0312	Crassostrea gigas
4	8.4314	55.0312	Crassostrea gigas
5	8.428	55.0315	Crassostrea gigas
6	8.4339	55.0304	Crassostrea gigas
7	8.4172	55.0368	Crassostrea gigas
8	8.428	55.0315	Crassostrea gigas
9	8.4314	55.0312	Crassostrea gigas
10	8.4339	55.0304	Crassostrea gigas
11	8.4314	55.0312	Crassostrea gigas
12	8.4339	55.0304	Crassostrea gigas
13	8.428	55.0315	Crassostrea gigas
14	8.4172	55.0368	Crassostrea gigas
15	8.8255959	58.447262	Crassostrea gigas
16	9.0724735	58.624622	Crassostrea gigas
17	9.0718365	58.624821	Crassostrea gigas
18	-3.574731272	54.42686226	Crassostrea gigas
19	1.037476751	51.78043197	Crassostrea gigas
20	-2.115802	49.251403	Crassostrea gigas
21	0.904865232	51.74780687	Crassostrea gigas
22	-2.186292	49.180744	Crassostrea gigas
23	-2.767877	48.574139	Crassostrea gigas
24	4.203335486	53.14743113	Crassostrea gigas



## Create model

**Algorithm:** Environmental distance

**Parameter values:** 2, 0, 1

**Environmental layers:**

- Mean Depth
- Mean Distance to Land
- Mean Surface Salinity
- Mean Surface Temperature
- Mean Sea Ice concentration

**Geographic mask:**

- GBIF\_training\_mask\_1800arcs  
ecs\_25MAR2014

## Test and project model

### Test model

- Inspect AUC
- Cross validation (replicates: 10, measure AUC, threshold LPT)

### Projection model

- Native projection (same layers, same mask)
- Another projection: Press yes
- 2050 projection (2050 layers of same environmental variables, same mask)
- Another projection: Press no
- Finish workflow and download all results

# Homework assignment

Use your own and GBIF data to project habitat suitability for the invasive oyster *C. gigas* into Swedish and Norwegian Exclusive Economic zone. How many percent increase of suitable habitat can we expect until 2050?

1. Mobilize and integrate occurrence data (DRW)
2. Filter environmentally unique points (BioClim)
3. Build and test global model; make local projections into Swedish/Norwegian Exclusive Economic Zones for 2013 and 2050 (ENM)
4. Analyze difference in rather projections between 2050 and 2013 in Norway and Sweden