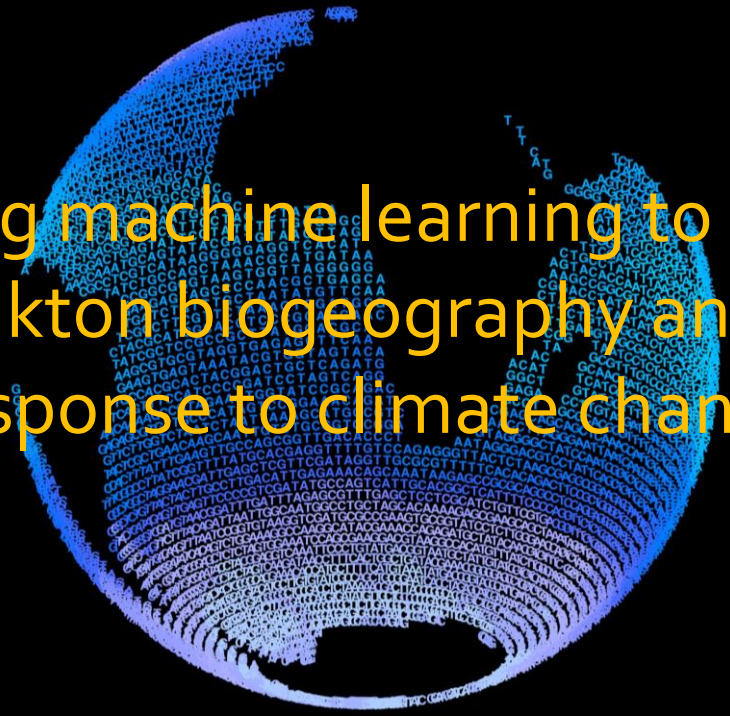
CEODEOS/AtlantEco Workshop
Valparaiso
15-19 May 2023



Using machine learning to infer plankton biogeography and its response to climate change

Paul Frémont
pfremont@gatech.edu
www.paulfremont.com

AtlantECO
Atlantic Ecosystems Assessment, Forecasting & Sustainability

CEODOS Chile

Introduction: Ecological niche

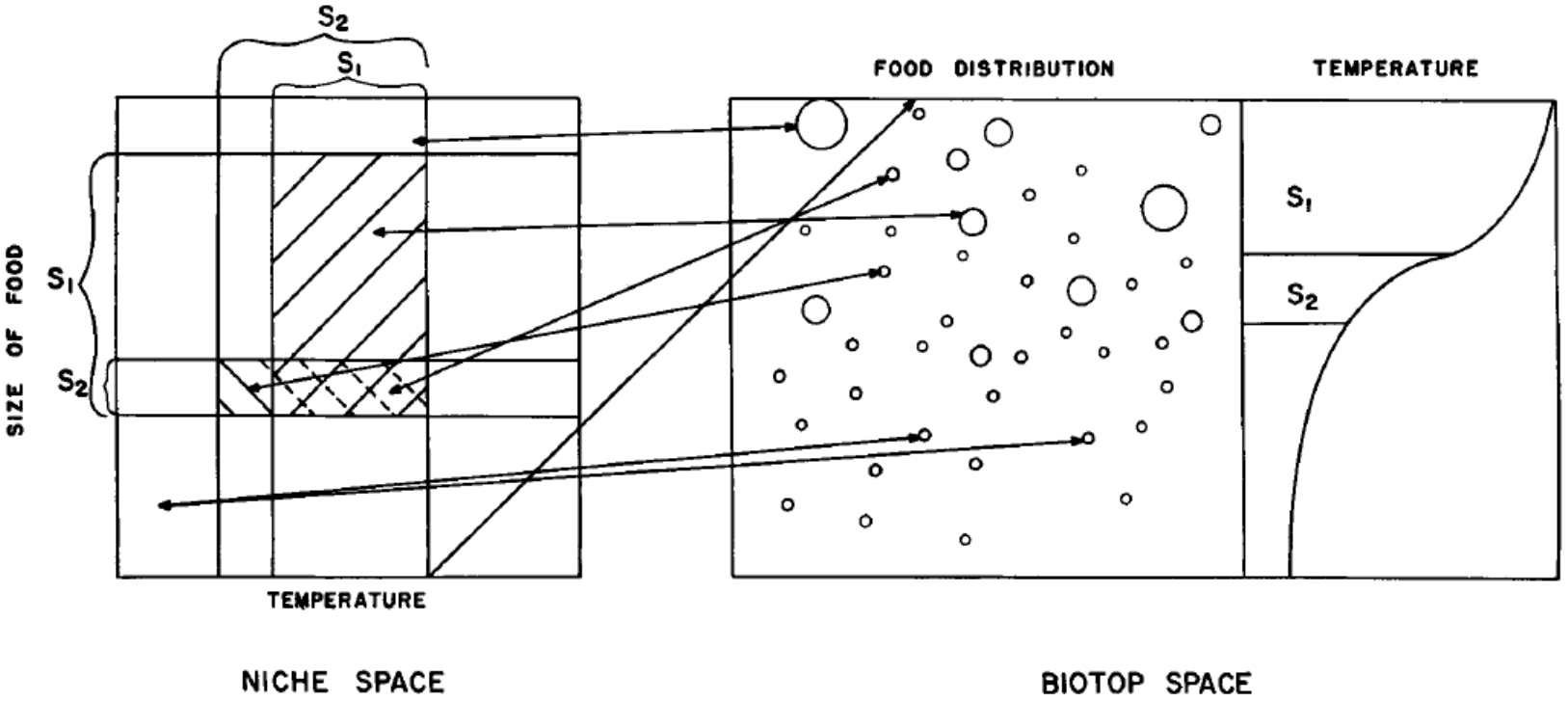


FIGURE 1. Two fundamental niches defined by a pair of variables in a two-dimensional niche space. Only one species is supposed to be able to persist in the intersection subset region. The lines joining equivalent points in the niche space and biotop space indicate the relationship of the two spaces. The distribution of the two species involved is shown on the right hand panel with a temperature depth curve of the kind usual in a lake in summer.

Hutchinson, 1957. *Concluding Remarks*

Introduction: Ecological niche

Fundamental niche:

all the ***theoretical*** environmental conditions where a species can live

Realized niche:

all the environmental conditions where a species is ***observed*** to live

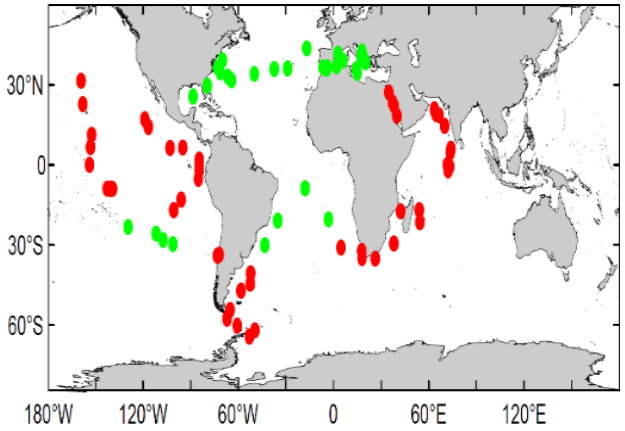
	Fundamental Niche	Realized Niche
Where the organism lives	No	Yes
Size	Large	Small
Competition for resources, predators are present	No	Yes
Other terminology	Precompetitive niche	Postcompetitive niche

$\Rightarrow$ In real life, when sampling an organism, we look at the ***realized*** niche

Introduction: How can we calculate a model of *realized* ecological niche?

How does supervised learning works?
A model is fitted on a **learning set:**
$\boldsymbol{Y}$**: Presence/absence of a species in each *Tara* Oceans stations**
$\boldsymbol{X}$**: Associated physicochemical parameters (predictors)**

⇒ **The model defines the environmental parameters in which the species lives *i.e.* it reflects the combination of parameters in which a species was found**

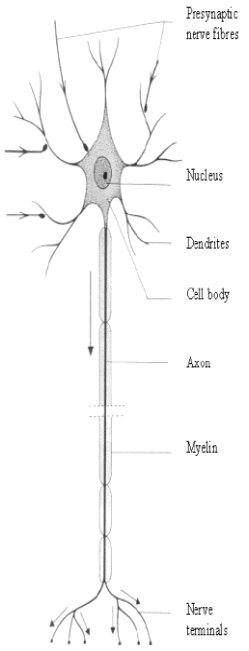


Presence
Absence

$$Y = f(X)$$

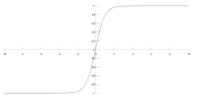
We are looking for $\boldsymbol{f}$

Neural networks

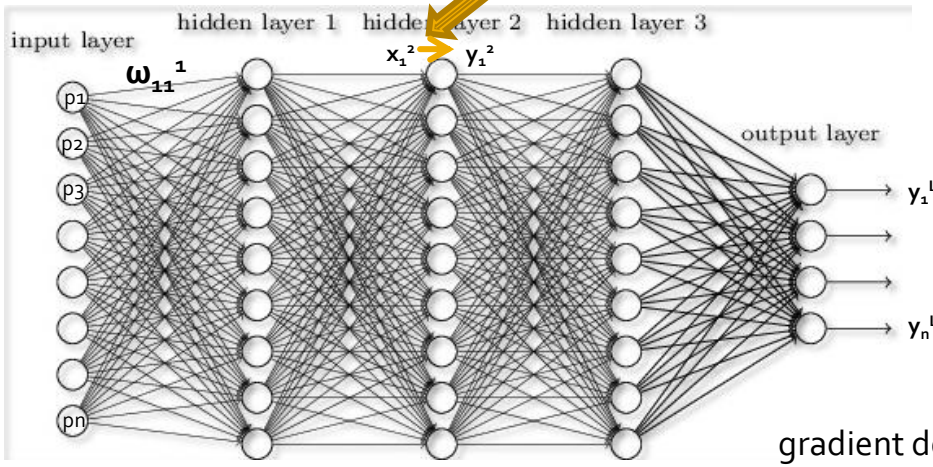


Error function E *e.g.* squared residuals, cross-entropy

Activation function (***e.g*** Softmax***)***



Neural Networks



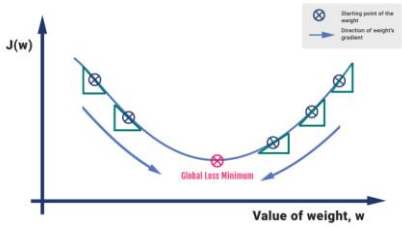
ω_{ij}^{1} ω_{ij}^{2} ω_{ij}^{3}

Forward propagation

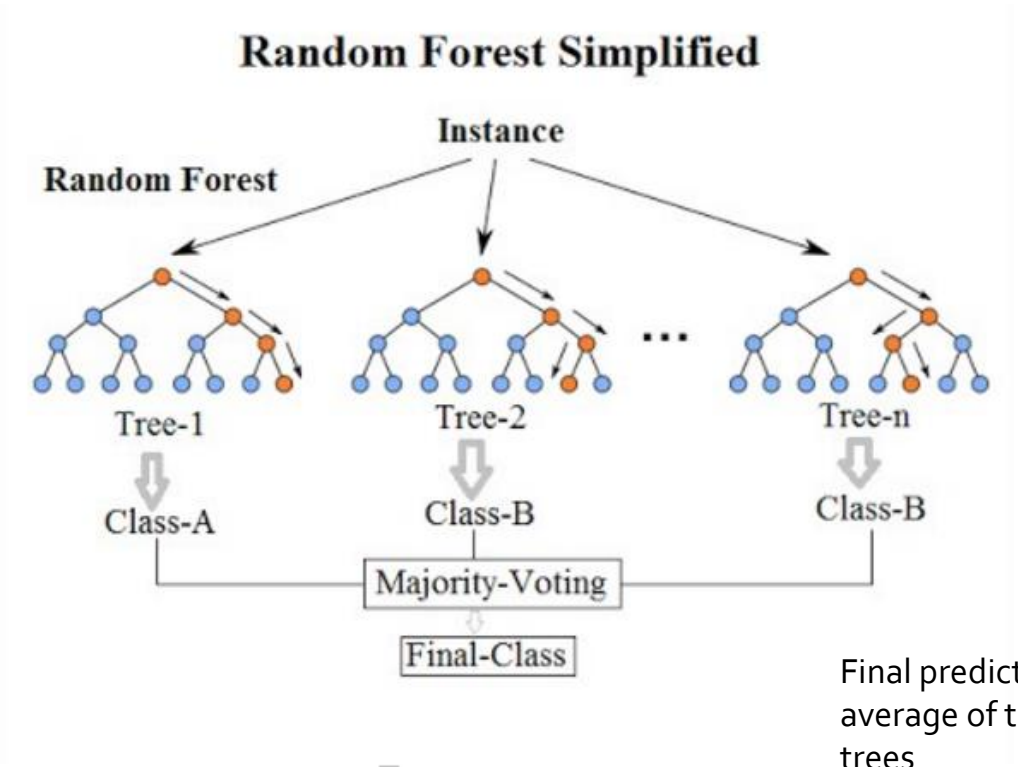
Back propagation (calculate derivative of error with respect to weights)

$$\frac{\partial E}{\partial w_{ij}^{\ell}} = y_i^{\ell} \frac{\partial E}{\partial x_j^{\ell+1}}$$

gradient descent



Random forest

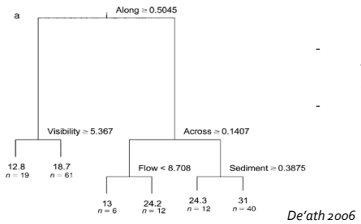


Set of B trees built based on bagging: each decision tree is built based on a subset of the set X, Y
+
Feature bagging: each tree is built on a random subset of the predictors

Final prediction is the average of the decision trees

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Other types of models



De'ath 2006

Example of a single decision tree

- Single trees are poor classifiers and large trees are hard to interpret => idea of combining several trees = **Boosting**
- Each tree is optimized by a least square method => **Regression**

A boosted regression tree is the sum of weighted trees:

$$y = f(x) = \sum_m f_m(x) = \sum_m \beta_m b(x; \gamma_m)$$

y is the **variable to be predicted**
x are the **predictors**
$b(x; \gamma_m)$ represent a tree : γ_m are the **split variables**, x their **values at each node**
β_m is the 'learning rate': weight given at each tree (ranges from 1 to 0,001)

Boosted Regression Trees

- In the regression setting, a generalized additive models has the form:

$$E(Y|X_1, X_2, \ldots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \quad (9.1)$$

Here fj's are unspecified smooth and nonparametric functions.

Instead of using LBE in chapter 5, we fit each function using a scatter plot smoother(e.g. a cubic smoothing spline)

Generalized Additive Models

MAXENT,
Support Vector Machine
Regression,
Lasso regression,
Linear regression
etc

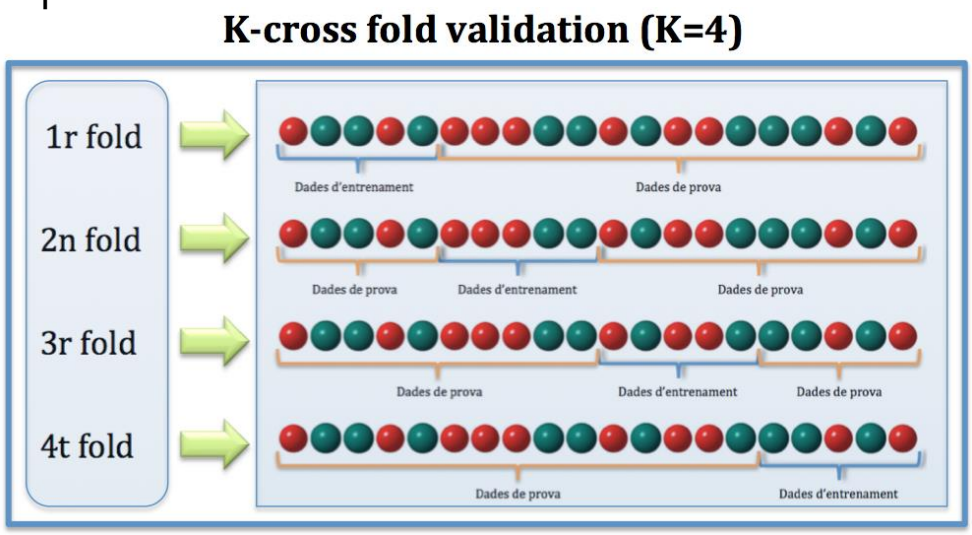
How can we say that a model performs well? Cross validation

Most of the times, models **overfit data** => how can we deal with this?

⇒ Cross validation

⇒ Hyperparameter optimization and model performance assessment

For a given set of hyperparamters:



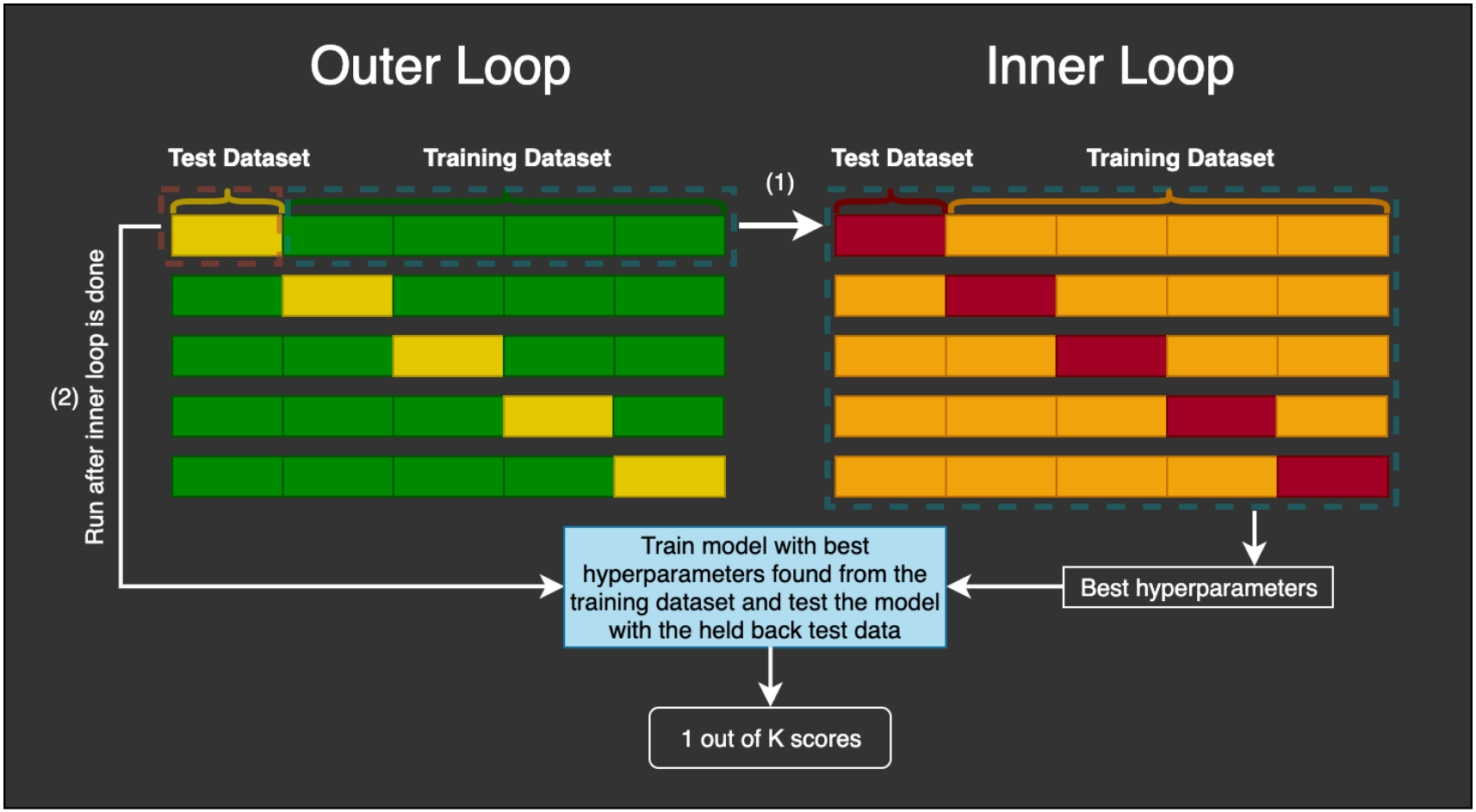
Mean model performance over the k-fold (RMSE, AUC etc)

Choose set of hyperparameter that has the best performance for the final model training

Other possibilities: Monte Carlo cross validation (random subsamples), exhaustive cross validation (leave-p-out cross validation), **nested cross validation**

Nested cross validation

Problem: Flat cross validation uses some points on which it has been trained to estimate model performance => positive bias on the model performance assessment



Overzealous?

Cornell University

arXiv > cs > arXiv:1809.09446

Search... Help |

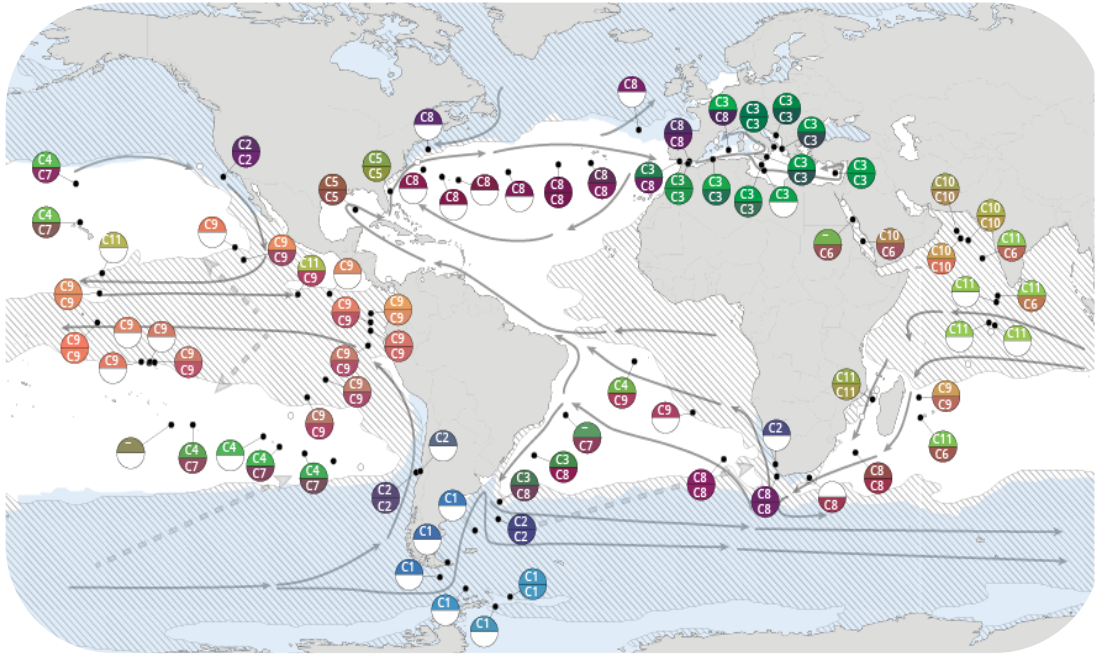
Computer Science > Machine Learning

[Submitted on 25 Sep 2018]

Nested cross-validation when selecting classifiers is overzealous for most practical applications

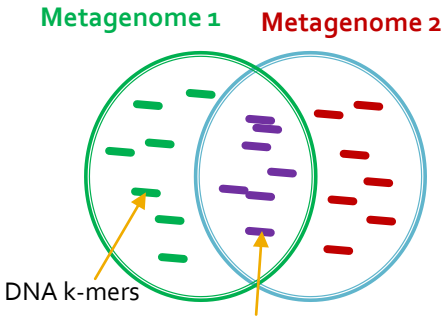
Jacques Wainer, Gavin Cawley

A concrete example: Tara Ocean genomic biogeography Genomic provinces



Richter *et al. 2022*

Metagenomic Biogeography of size fraction 0.8-5 µm definition of **genomic provinces**

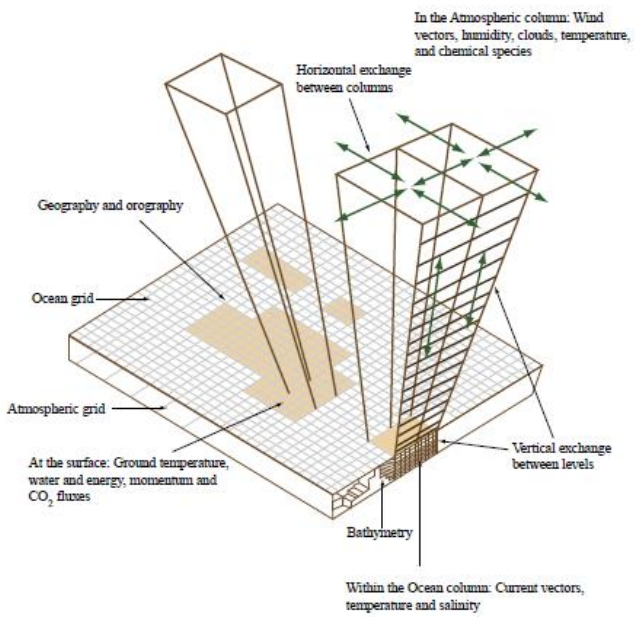


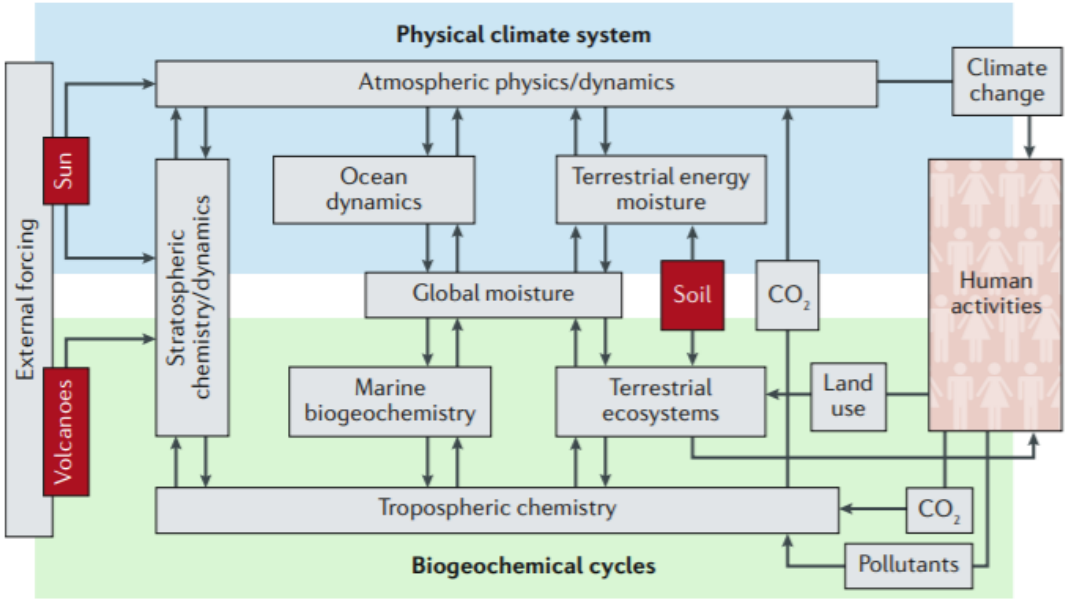
Shared k-mers= **beta diversity**

⇒**Hypothesis of associated environmental niches**

A concrete example: Tara Ocean genomic biogeography
Earth system science

Earth System Science



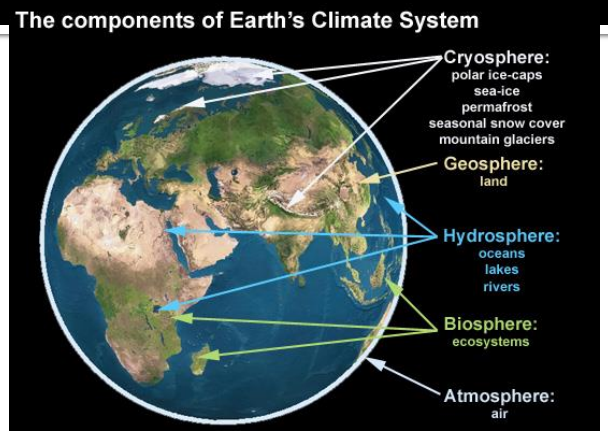


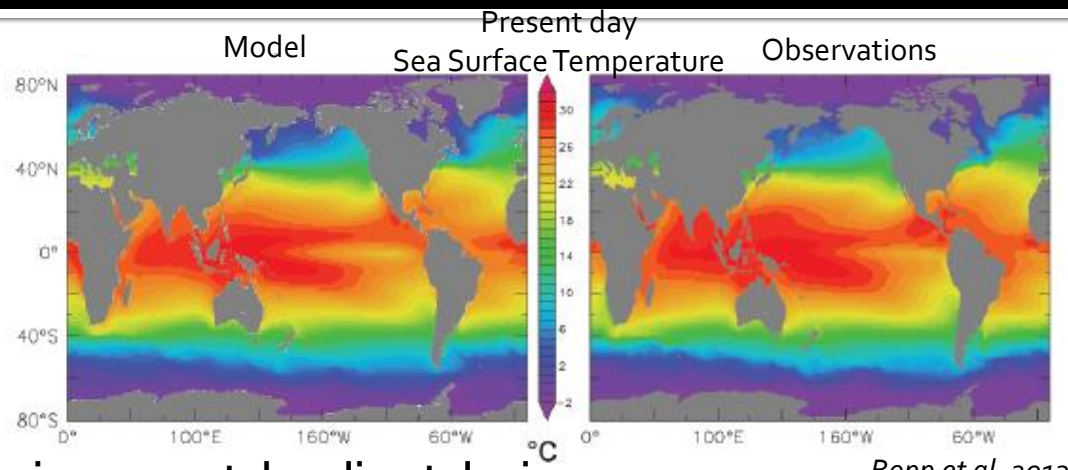
The NASA Bretherton diagram of the Earth System.

Steffen et al. 2020

"The IPSL Climate Model (IPSL-CM) is developed since 1995."

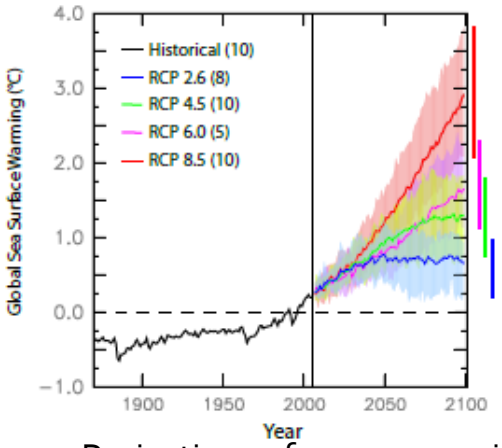
A concrete example: Tara Ocean genomic biogeography
Climate models of the ocean



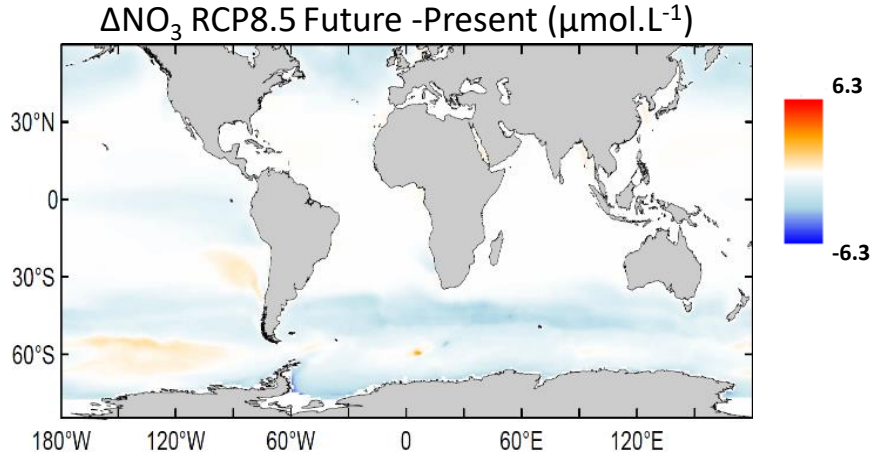


⇒ **Complex climate models reproducing present day climatologies**

Bopp et al. 2013



Projections of ocean warming



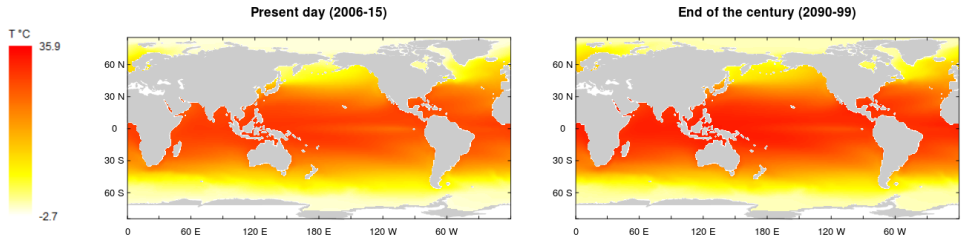
⇒ These models **allow exploring the possible hypothetic futures** depending on different greenhouse gas emission scenarios

⇒ We focus on the '**business as usual**' scenario **RCP8.5**

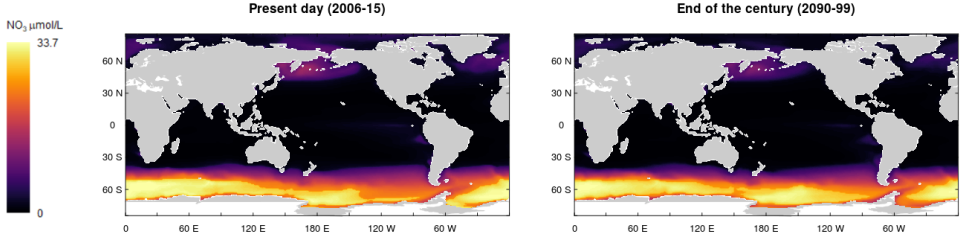
What are ocean climate models good for?

Project **SDMs** on new physicochemical datasets and evaluate climate change effects

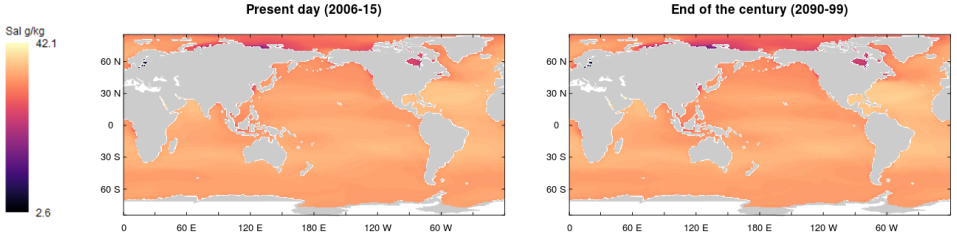
Temperature



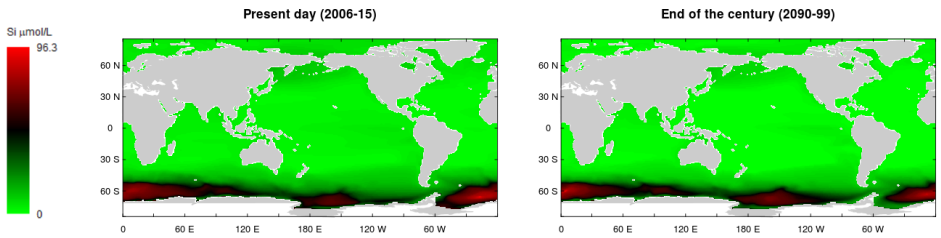
Nitrate



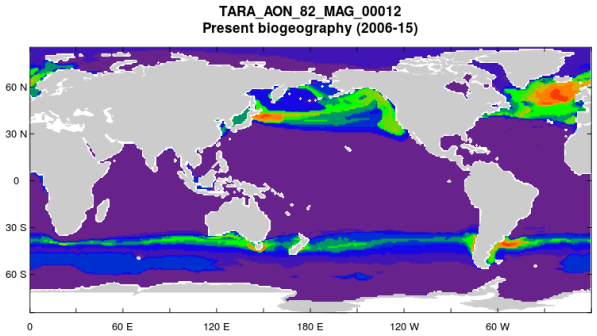
Salinity



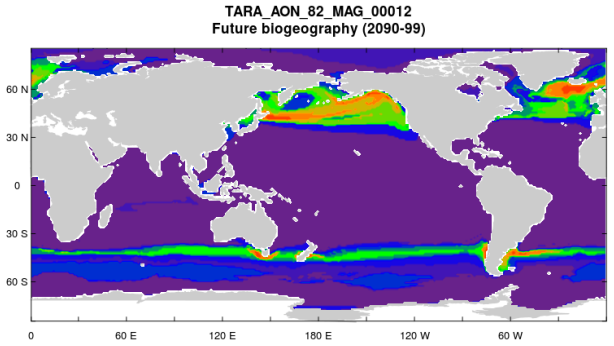
Phosphate



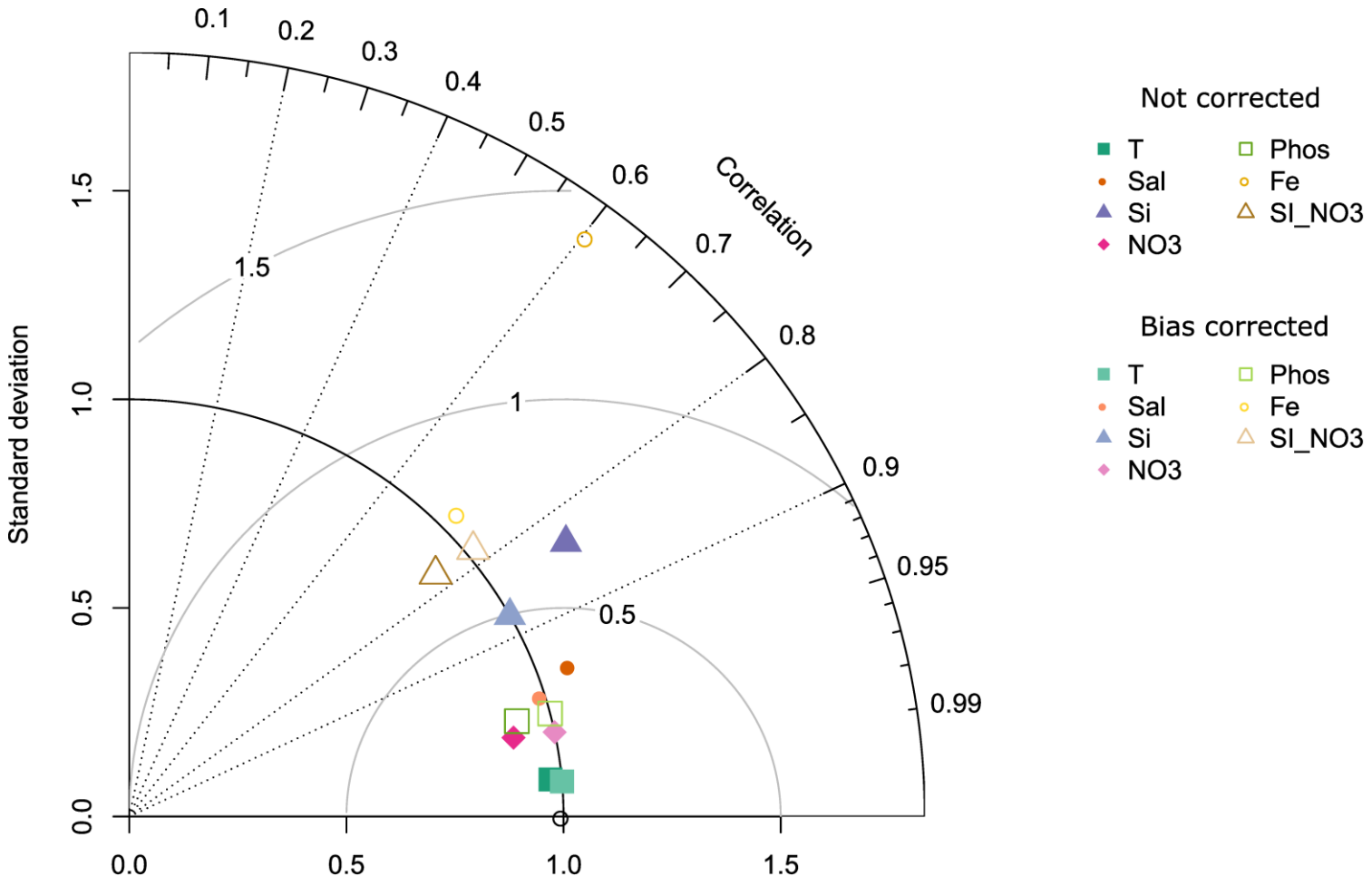
Present day



End of the century

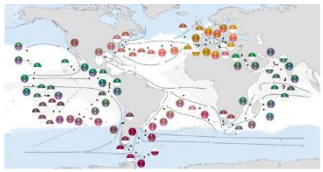


Do they perform well?



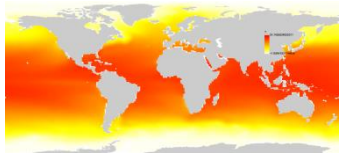
A concrete example: Tara Ocean genomic biogeography

Tara oceans
Metagenomic Provinces

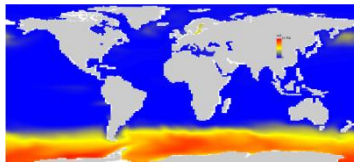


Richter *et al. 2022*

World Ocean Atlas 2013
8 physico-chemical variables**
***in situ* data**



*Earth System Models****
8 physico-chemical variables
model data



Environmental niche modeling
4 **machine learning** methods*

CDFt
Bias correction

invalid models

Environmental niches

Predictions

Bias corrected biogeochemical data

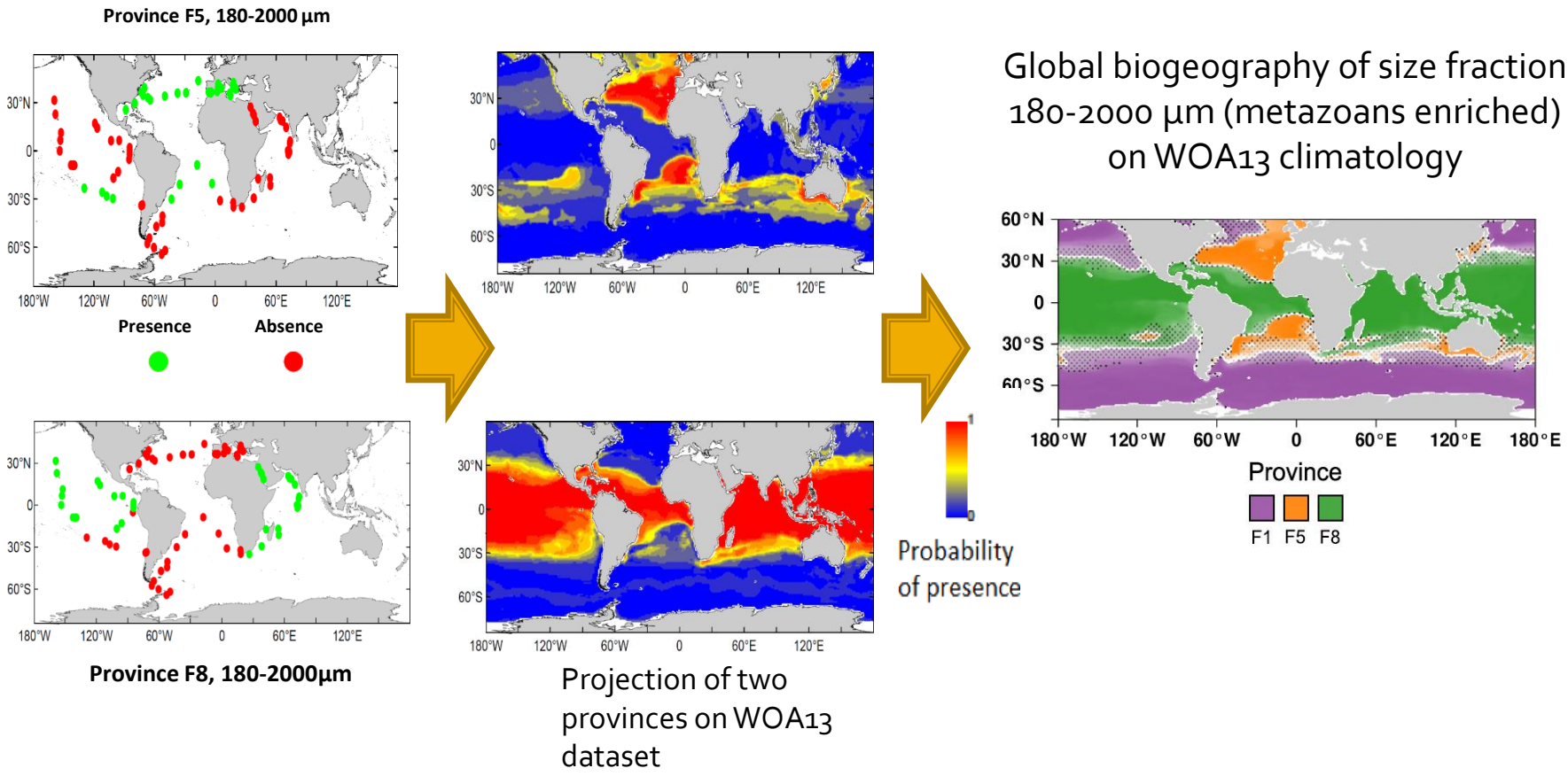
Present and future projections of global plankton biogeography

*Random Forest, Gradient boosted trees, Neural networks and Generalized Additive Models

** Temperature, Salinity, NO_3, PO_4, Si, Fe, SI_NO3

***CESM1-BGC, GFDL-ESM2G, GFDL-ESM2M, HadGEM2-ES , IPSL-CM5A-LR , IPSL-CM5A-MR, MPI-ESM-LR, MPI-ESM-MR, NorESM1-ME

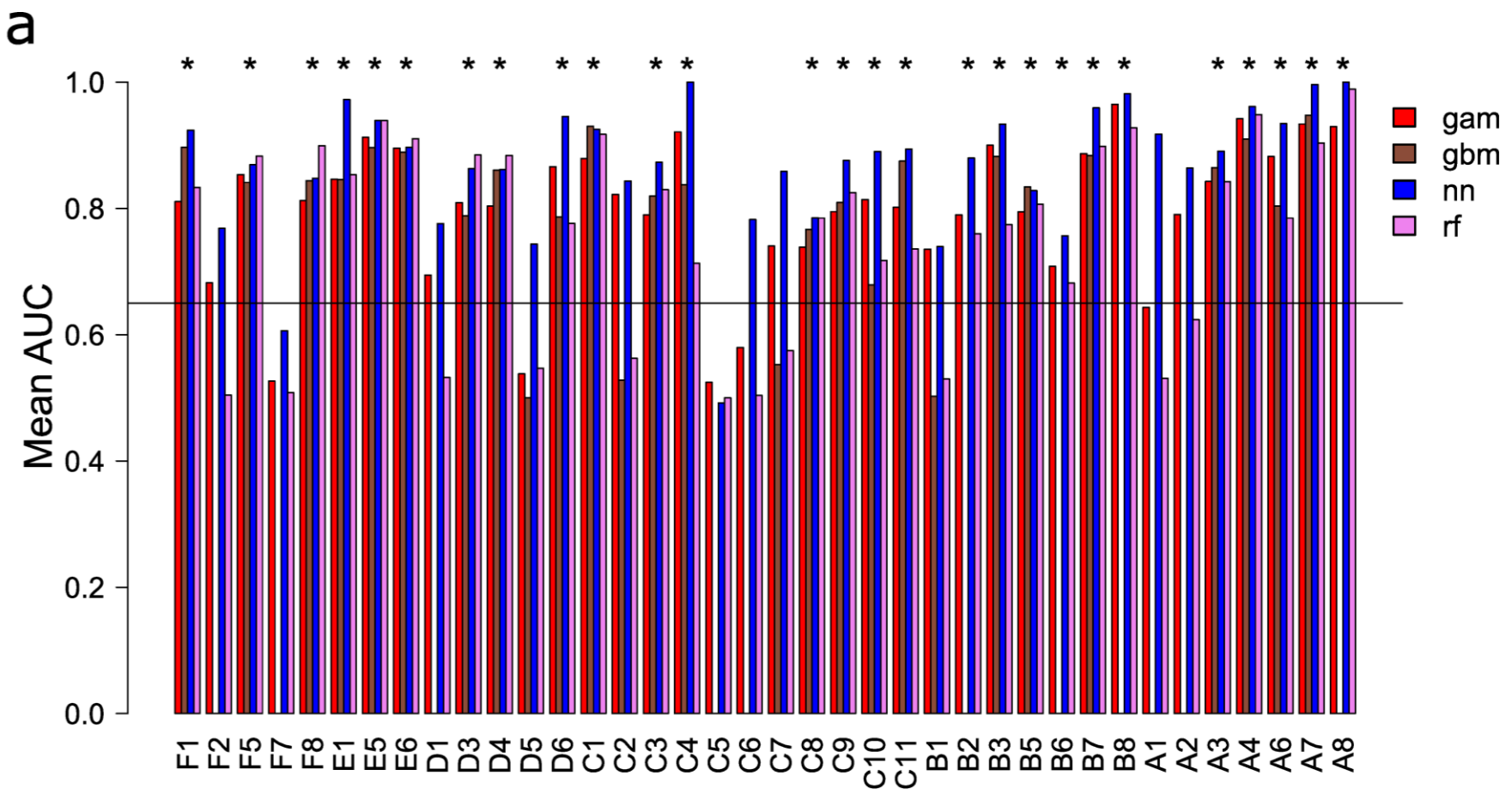
A concrete example: Tara Ocean genomic biogeography Projections



⇒ Multiclass classifier divided in multiple single class classifiers

⇒ Simple biogeography for metazoans (separation of temperate, equatorial + tropical and polar)

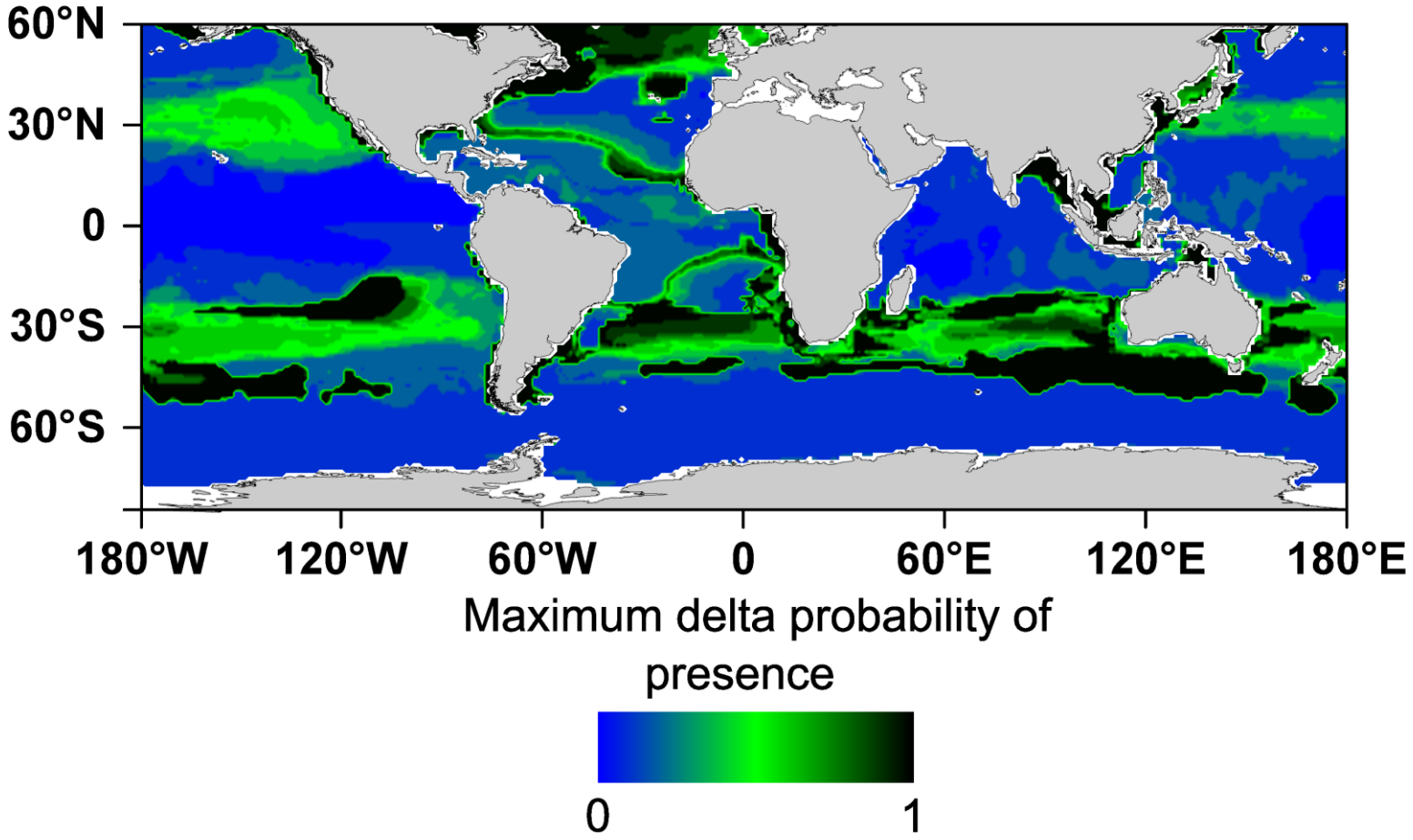
Model performances



27 valid models

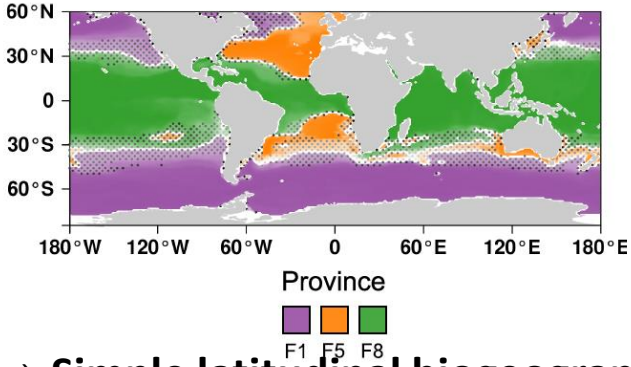
Neural network performs better

Do they agree with each other?

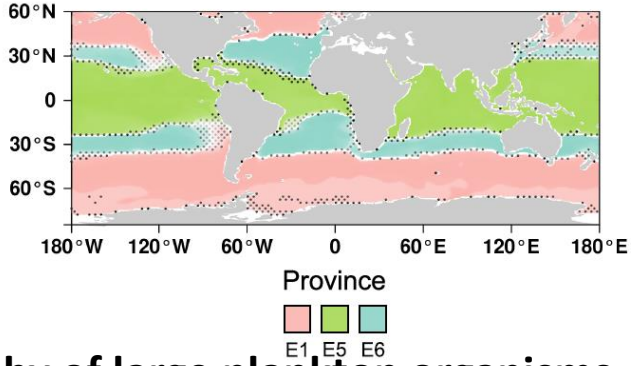


Genomic biogeographies of plankton

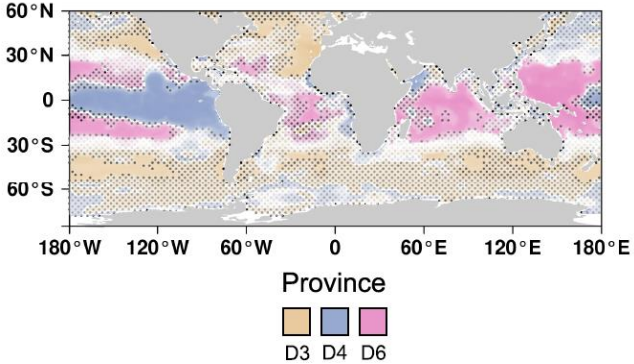
180-2000 µm Small metazoans



20-180 µm Small metazoans

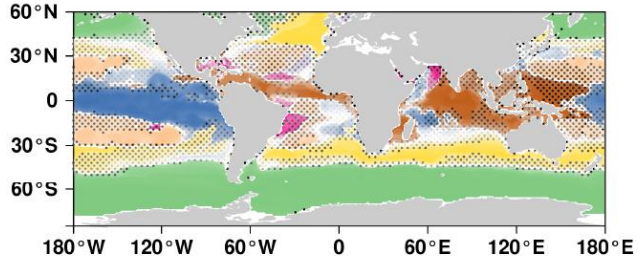


5-20 µm



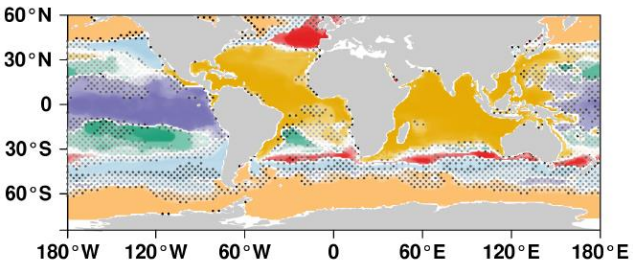
⇒ **Simple latitudinal biogeography of large plankton organisms**

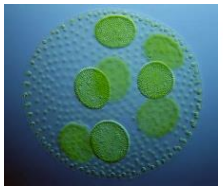
0.8-5 µm



Unicellular algae

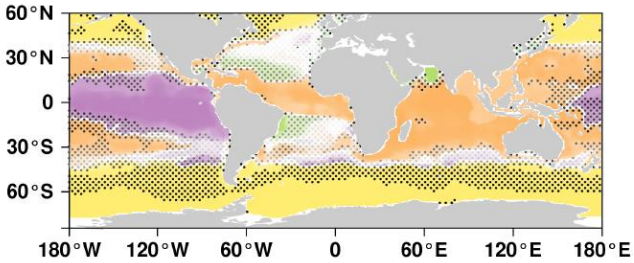
0.22-3 µm

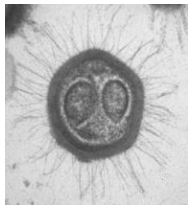




Bacteria

0-0.2 µm

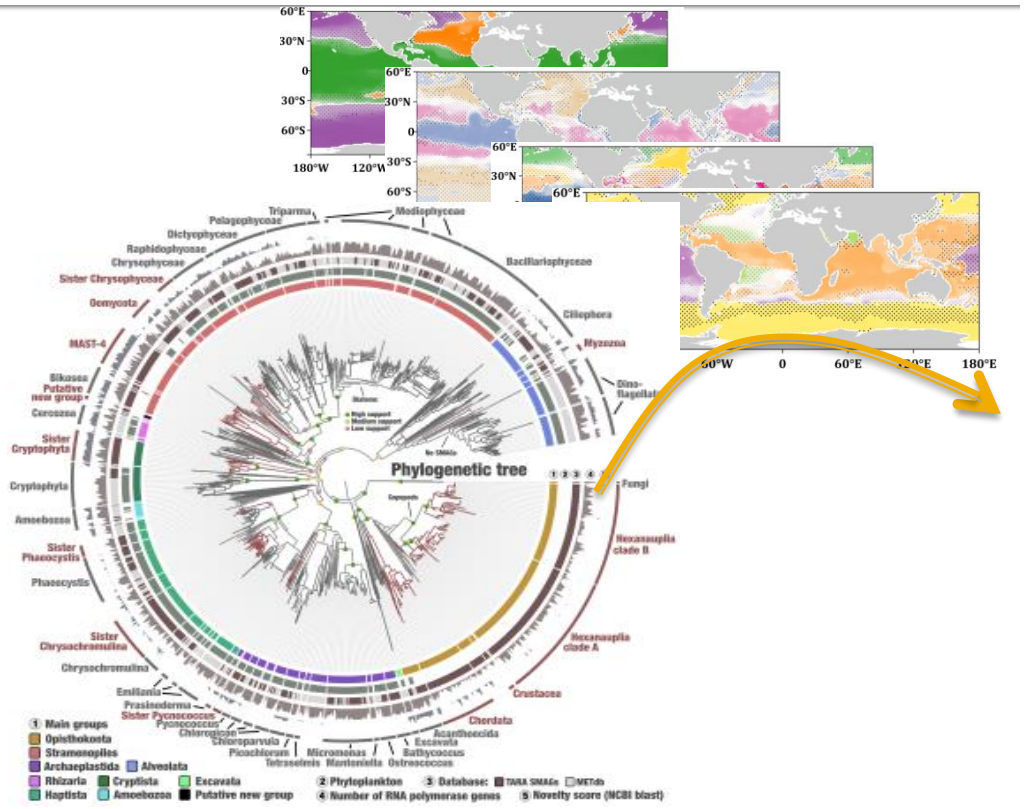




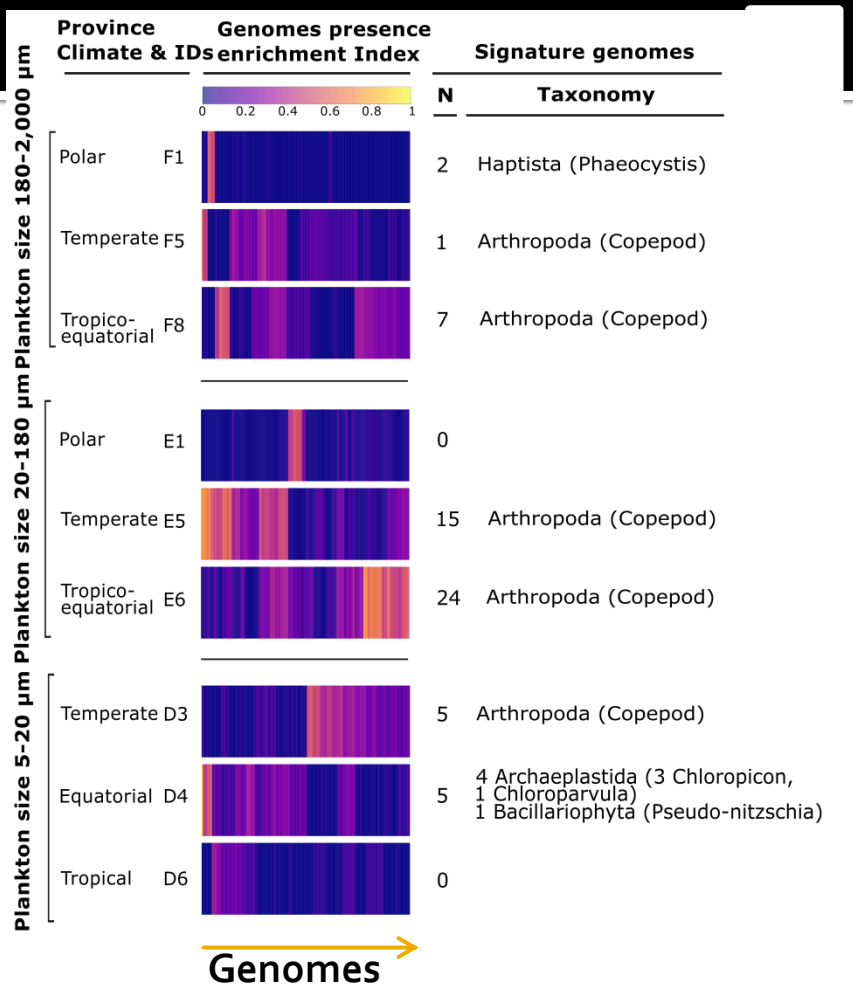
Viruses

⇒ **More complex and patchy biogeography of small plankton organisms**

Signature genomes



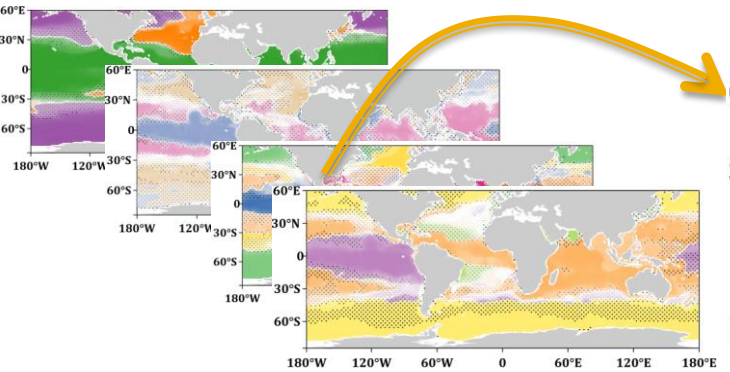
Delmont et al. 2021 *BioRxiv*



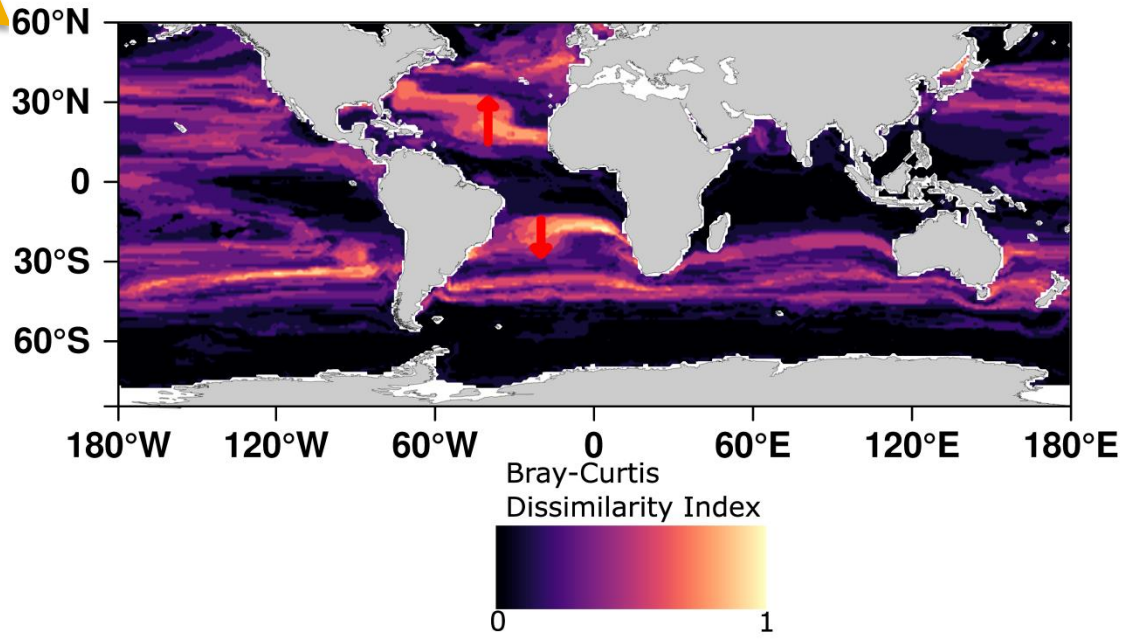
⇒Signature genomes of climato-genomic provinces: highlights species and genomes that structure plankton biogeography at the global scale

⇒ *Climato-genomic* provinces structure plankton biogeography at a higher level than individual genomes

Effects of climate change



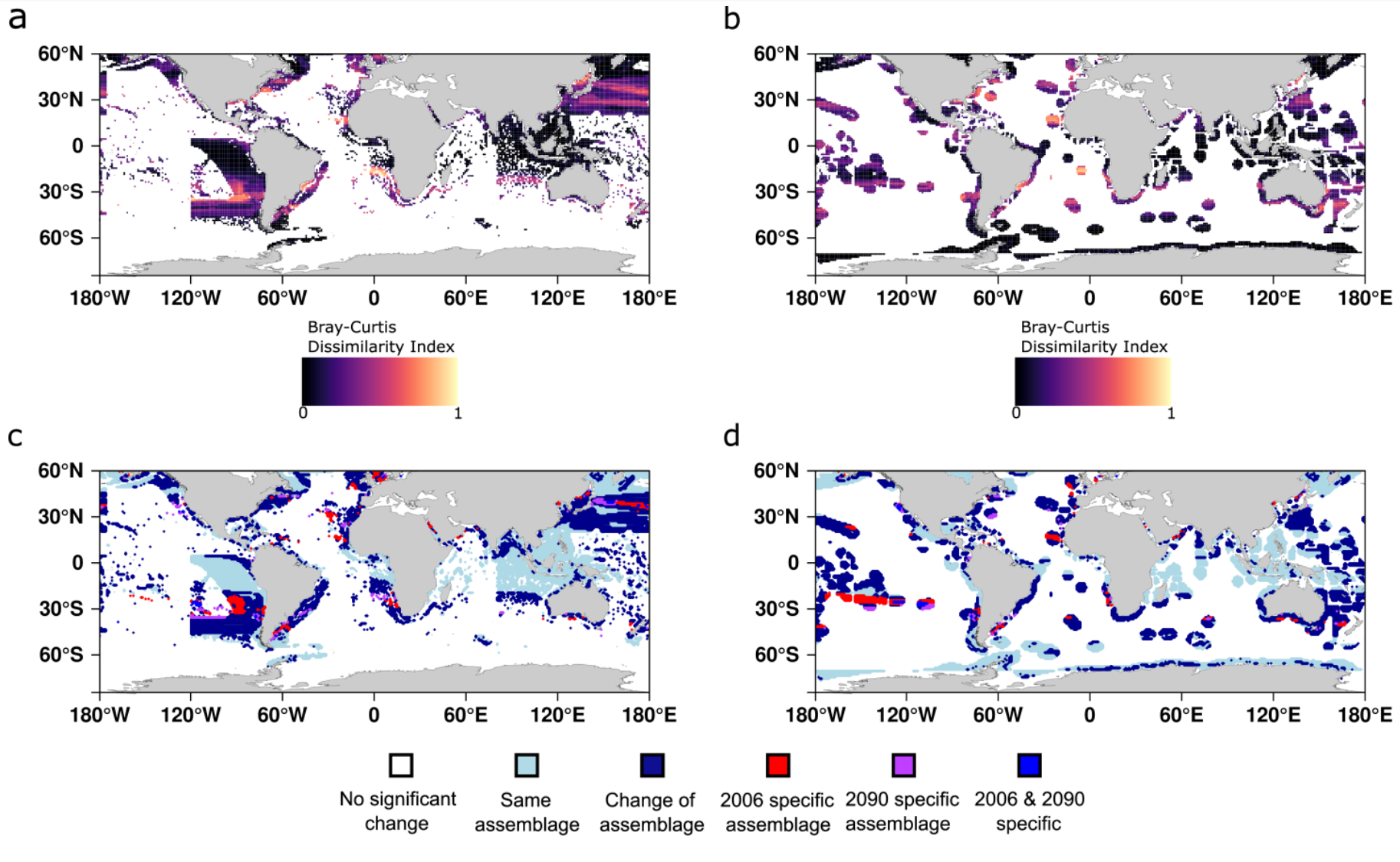
Present versus end of the century



⇒Most important reorganization in temperate regions (25° to 60°): mean dissimilarity of 0.29 (north) and 0.24 (south)

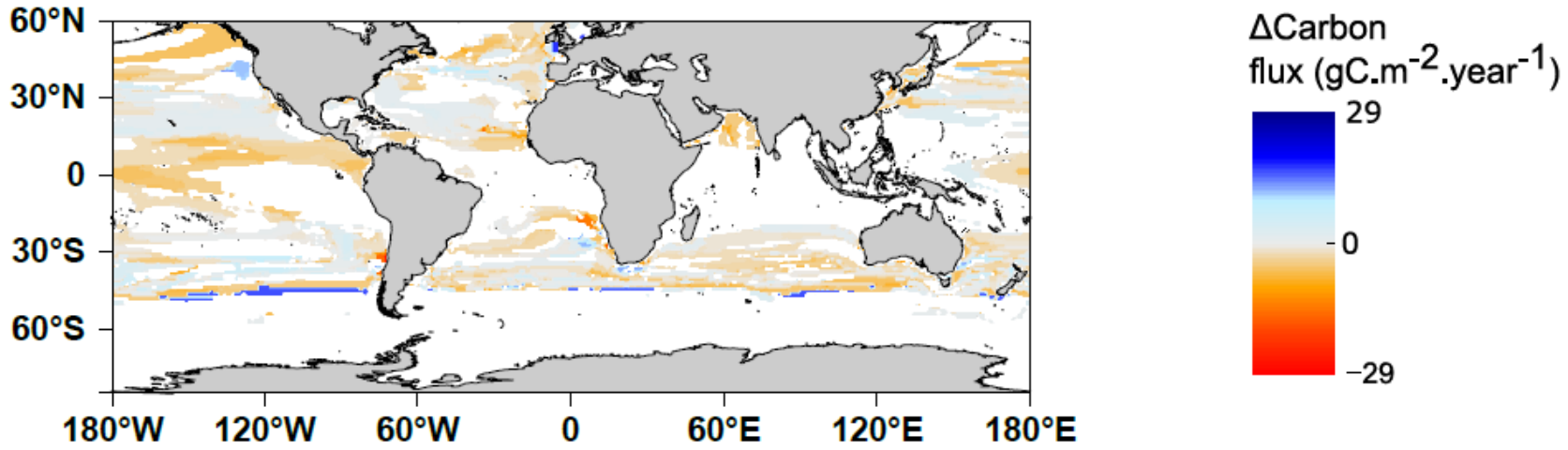
⇒ **45 % to 57 %** of considered ocean area with an important change

Socio-economic impact



Feedback on climate change

d

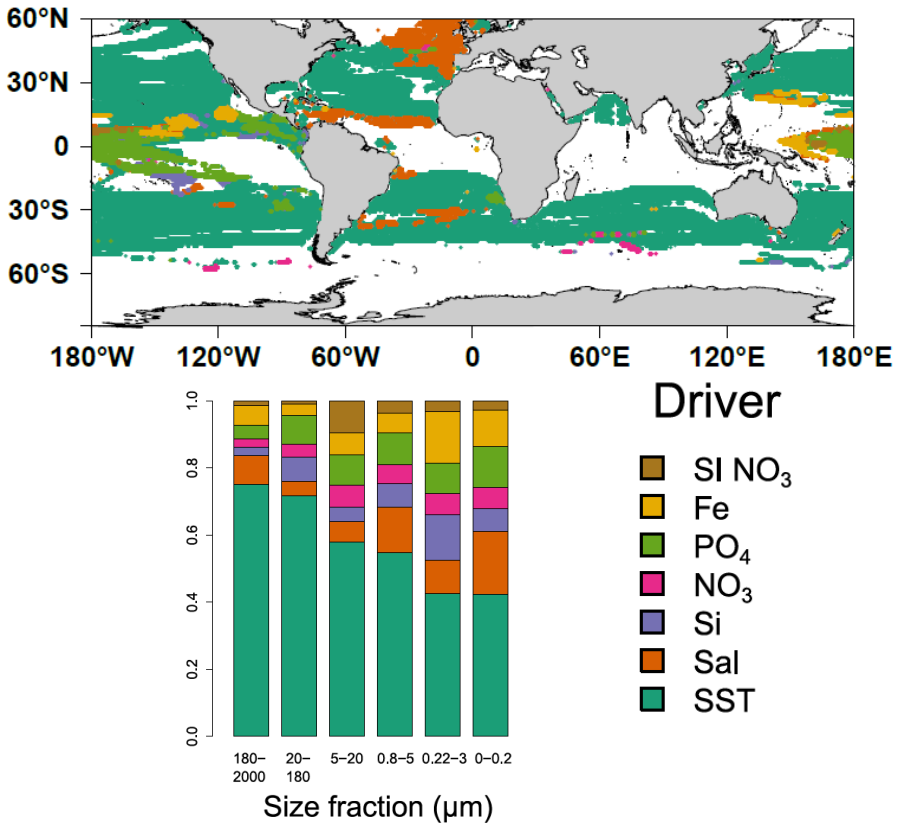


⇒ Based on the provinces' assemblages model: **decrease of 4% on average of POC export fluxes** (based on three extrapolated models of POC export: Eppley et al. 1979, Laws et al. 2000 and Henson et al. 2012)

⇒ Feedback on climate change: **reinforcement**

Drivers

Physico-Chemical Drivers of reorganizations around 2090



⇒ **Temperature changes** explains only **50% of changes** in dominant communities followed by phosphate (11.0%) and salinity (10.3%)

⇒ **Nutrients changes are more important in driving small organisms' changes** (consistent with their trophic modes)

Articles

nature climate change

ARTICLES

https://doi.org/10.1038/s41558-022-01314-8

Check for updates

Restructuring of plankton genomic biogeography in the surface ocean under climate change

Paul Frémont[1,2], Marion Gehlen[3], Mathieu Vrac[3], Jade Leconte[1,2], Tom O. Delmont[1,2], Patrick Wincker[1,2], Daniele Iudicone[4] and Olivier Jaillon[1,2]

The impact of climate change on diversity, functioning and biogeography of marine plankton remains a major unresolved issue. Here environmental niches are evidenced for plankton communities at the genomic scale for six size fractions from viruses to meso-zooplankton. The spatial extrapolation of these niches portrays ocean partitionings south of 60° N into *climato-genomic* provinces characterized by signature genomes. By 2090, under the RCP8.5 future climate scenario, provinces are reorganized over half of the ocean area considered, and almost all provinces are displaced poleward. Particularly, tropical provinces expand at the expense of temperate ones. Sea surface temperature is identified as the main driver of changes (50%), followed by phosphate (11%) and salinity (10%). Compositional shifts among key planktonic groups suggest impacts on the nitrogen and carbon cycles. Provinces are linked to estimates of carbon export fluxes which are projected to decrease, on average, by 4% in response to biogeographical restructuring.

Planktonic communities are composed of complex and heterogeneous assemblages of small animals, single-celled eukaryotes (protists), bacteria, archaea and viruses. They contribute to the regulation of the Earth system through primary production via photosynthesis[1] and carbon export to the deep ocean[2,3] and are at the base of the food webs that sustain the whole trophic chain in the oceans[4].

The composition of planktonic communities varies over time at a given site with daily[5] to seasonal fluctuations[6] following environmental variability[7] (for example, temperature, nutrients). Time series observations have highlighted recent changes in the planktonic ecosystem such as changes in community composition[24] or poleward shifts of some species[25], mainly attributed to temperature increase caused by anthropogenic greenhouse gas emissions. These changes are expected to intensify with ongoing global warming[26] and could lead to major reorganization of plankton community composition[22] with a potential decline in diversity[27–29]. Another major consequence of global reorganization of the seascape on biological systems would be a decrease of primary production at mid-latitudes and an increase at higher latitudes[26].

Codes for SDMs: https://github.com/institut-de-genomique/NCLIM-20102618B

eLife

RESEARCH ARTICLE

Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems

Daniel J Richter[1,2†], Romain Watteaux[3,4†], Thomas Vannier[5,6,7†], Jade Leconte[5,6], Paul Frémont[5,6], Gabriel Reygondeau[8,9], Nicolas Maillet[10], Nicolas Henry[1,6], Gaëtan Benoit[11], Ophélie Da Silva[6,12], Tom O Delmont[5,6], Antonio Fernàndez-Guerra[13,14,15], Samir Suweis[16], Romain Narci[17], Cédric Berney[1,6], Damien Eveillard[6,18], Frederick Gavory[5], Lionel Guidi[6,12], Karine Labadie[19], Eric Mahieu[19], Julie Poulain[5,6], Sarah Romac[1,6], Simon Roux[20], Céline Dimier[1,21], Stefanie Kandels[22,23], Marc Picheral[6,12], Sarah Searson[6,12], Tara Oceans Coordinators, Stéphane Pesant[24,25], Jean-Marc Aury[5], Jennifer R Brum[20,26], Claire Lemaitre[11], Eric Pelletier[5,6], Peer Bork[22,27,28], Shinichi Sunagawa[22,29], Fabien Lombard[6,12,30], Lee Karp-Boss[31], Chris Bowler[6,21,21], Matthew B Sullivan[20,32,33,34], Eric Karsenti[6,21,23], Mahendra Mariadassou[17], Ian Probert[1,6], Pierre Peterlongo[11], Patrick Wincker[5,6], Colomban de Vargas[1,6]*, Maurizio Ribera d'Alcalà[3]*, Daniele Iudicone[3]*, Olivier Jaillon[5,6]*

Cell Genomics

CellPress OPEN ACCESS

Article

Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean

Tom O. Delmont,[1,2,9,*] Morgan Gaia,[1,2] Damien D. Hinsinger,[1,2] Paul Frémont,[1,2] Chiara Vanni,[3] Antonio Fernandez-Guerra,[4] A. Murat Eren,[5] Artem Kourlaiev,[1,2] Leo d'Agata,[1,2] Quentin Clayssen,[1,2] Emilie Villar,[1] Karine Labadie,[1,2] Corinne Cruaud,[1,2] Julie Poulain,[1,2] Corinne Da Silva,[1,2] Marc Wessner,[1,2] Benjamin Noel,[1,2] Jean-Marc Aury,[1,2] Tara Oceans Coordinators, Colomban de Vargas,[2,6] Chris Bowler,[2,7] Eric Karsenti,[2,6,8] Eric Pelletier,[1,2] Patrick Wincker,[1,2] and Olivier Jaillon[1,2]

Plankton biogeography website

TO EN DB Tara Oceans Ecological Niches Database

Tara Oceans niches | IGOB: Interactive Generator Of Biogeography

Visualize theoretical niches (convex hulls)

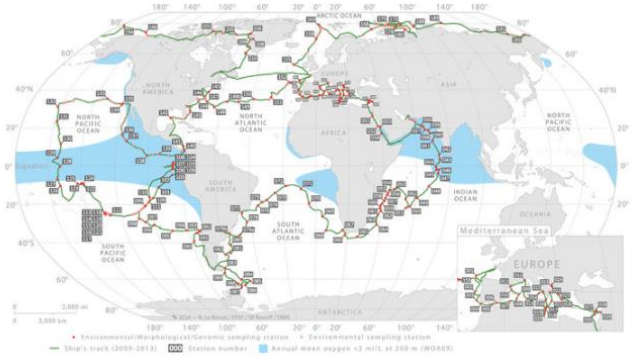
An ecological niche is defined as the envelopp of physico-chemical parameters in which a species lives. The niche concept was popularized by the zoologist G. Evelyn Hutchinson in 1957. Here you can visualize the ecological niches of 374 Metagenome Assembled Genomes (MAGs) built from the Tara Oceans dataset.
Select below the SMAG of interest and visualize the effect of climate change (high emission scenario RCP8.5) on its niche by clicking on 'Severe climate change'
Colors on the maps allows you to visualize the probability of presence at each location of all oceans of the selected MAG

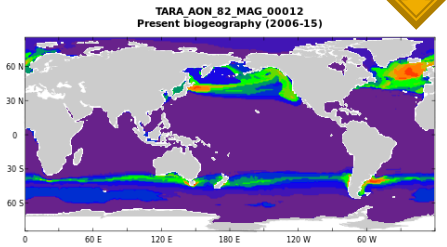
Select kingdom: All | Select phylum: All | Select class: All | Select order: All | Select family: All | Select genre: All

Select SMAG of interest

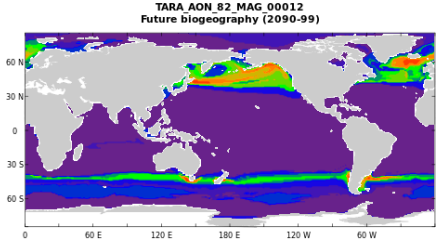
No SMAG Selected

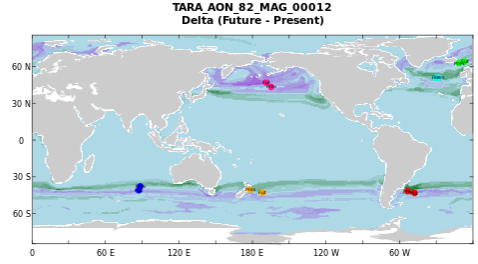
Select a genome and visualize its biogeography and projected impacts of climate change





is map displays the niche of AON_82_MAG_00012_Ostreococcus in projected end of the century conditions under high emisssion scenario RCP





http://end.mio.osupytheas.fr/Ecological_Niche_database/

Good practices in Species Distribution Modeling

• Spatial species distribution models are **powerful tools** to project species distributions or organism or whole community distribution

• **Evaluate model performances** and **optimize hyperparameters** using **cross validation**

• **Don't model species for which there is not enough data** (presence points): equilibrium in presence/absence data is preferrable though rarely achievable

• Be careful in the **choice of predictors** and **extrapolate in predictability zones OR/AND acknowledge uncertainties/biases:**

- **choose predictors based on knowledge**
- **use z-scores (for better performance of neural networks)**
- **spatio-temporal sampling and extrapolation**

• Other types of SDM exists pseudo-absence models, presence only models, models informed by biotic interactions, **quantitative models (very uncertain, strong errors)**

SDM using machine learning: drawbacks and limits

- o/1 SDMs are **Statistical** species distribution models => they remain correlative models, they are not causal (and don't account for currents, adaptation, acclimation…)

- Mechanistic species distribution models exists => predict the fundamental niche of a species based on its physiology and project it on the seascape (competition can be added too)

- Purely statistical models can't predict beyond extremes and training datasets values (can physics do it?)


RESEARCH ARTICLE | BIOLOGICAL SCIENCES

Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*

Pedro Flombaum, José L. Gallegos, Rodolfo A. Gordillo, +9, and Adam C. Martiny | Authors Info & Affiliations

Contributed by David M. Karl, April ; sent for review January 22, 2013)

May 23, 2013 | 110 (24) 9824-9829 | doi.org/10.1073/pnas.1307701110

Bad practice! extrapolation beyond training, no acknowledgement


Article | Open Access | Published: 15 March 2023

Understanding opposing predictions of *Prochlorococcus* in a changing climate

Vincent Bian, Merrick Cai & Christophe ollett

Nature Communications **14**, Article numb 445 (2023) | Cite this article

996 Accesses | **10** Altmetric | Metrics

Warns on the dangers/limits of (quantitative) statistical models

Acknowledgement

My PhD supervisors
Olivier Jaillon (Genoscope), Marion Gehlen


Stazione Zoologica
Daniele Iudicone
Lucia Campese
Bruno Hay Mele



LSCE
Marion Gehlen, Mathieu Vrac

Genoscope (LAGE)
Patrick Wincker, Eric Pelletier, Tom Delmont, Amin Madoui, Adrien Thurotte, Quentin Carradec, Betina Porcel, Julie LeHoang, Romual Laso-Jadart, Jade Leconte, Marie Burel, Janaina Rigonato, Nina Guérin, Margaux Crédeville, Lucas Pavlovic