27-29 May 2024

imdis

International conference on Marine Data and Information Systems

SeaDataNet AISBL · Havforskningsinstituttet · Ifremer · MARIS · MEDIN · National Oceanography Centre · INGV · iMagine · IODE · eosc Blue-Cloud2026

Interpolation of Continuous Plankton Recorder data using a neural network technique

IMDIS 2024

Charles Troupin // Alexander Barth // Pierre Hélaouët

28 May 2024

our goal

"Interpolation of Continuous Plankton Recorder data using a neural network technique"

the dataset (CPR)

the tool (DINCAE)

"Before you could apply all your **artificial intelligence** *algorithms,* *I had to put* **real intelligence** *to prepare the data."*

How it all began...

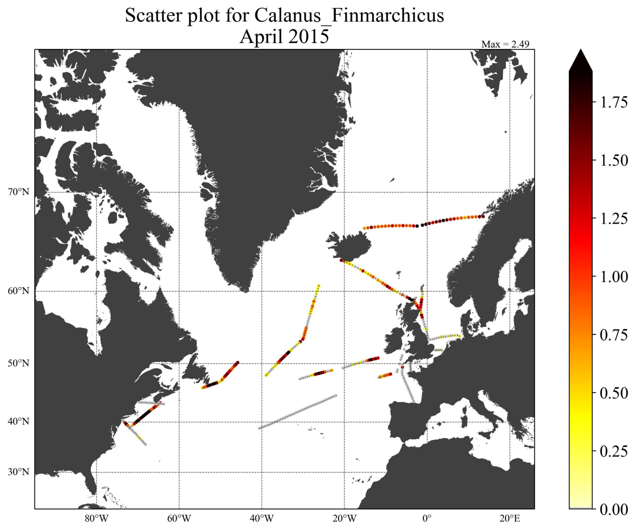
Hi guys, could you create gridded maps from my data? — Pierre

Charles — Yes, sure! Send them to us!

OK, here you go! — Pierre

Charles — Ohh...

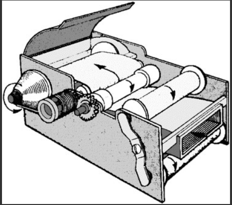
One month of observations...



Continuous Plankton Recorder [CPR] survey

The Continuous Plankton Recorder (CPR) Survey
@CPRSurvey

On this day in 1896, Sir Alister Hardy, the founder of the #CPRSurvey, was born. During his early career, Hardy designed and built the first Continuous Plankton Recorder, and after only a couple of modifications, remains the same design we use today! Happy Birthday! #OnThisDay



1931 Start of the monitoring

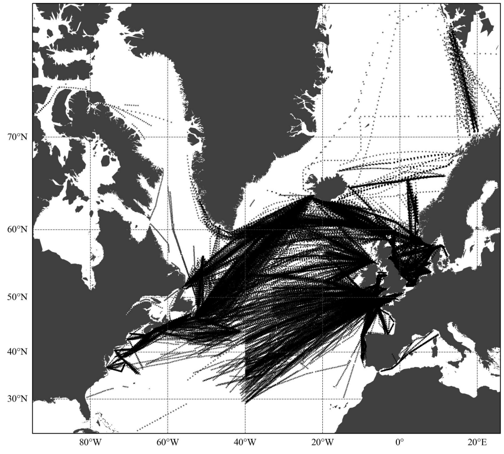
7284725 Total Nautical Miles Towed

659 Taxa Routinely Analysed

282552 Total Samples Analysed

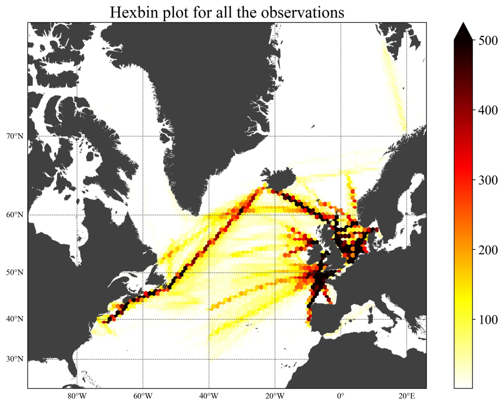
https://www.cprsurvey.org

Continuous Plankton Recorder dataset



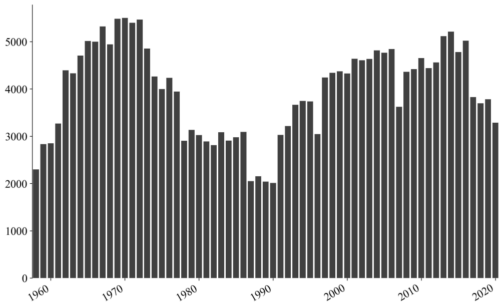
- 250021 observations
- 1958 - 2020 period
- position | time | abundance of different taxa

Continuous Plankton Recorder dataset



- 250021 observations
- 1958 - 2020 period
- position | time | abundance of different taxa
- strong spatial heterogeneity

Continuous Plankton Recorder dataset



- 250021 observations
- 1958 - 2020 period
- position | time | abundance of different taxa
- strong spatial heterogeneity
- variable temporal availability

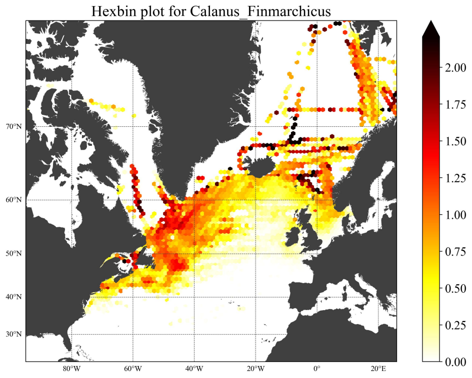
Canalus finmarchicus and Calanus helgolandicus

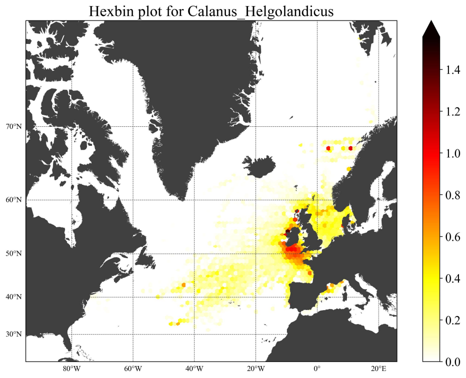
https://www.marinespecies.org/aphia.php?p=taxdetails&id=104464

https://www.marinespecies.org/aphia.php?p=taxdetails&id=104466

- Most studied copepods
- Prey for higher trophic levels
- Different spatial preferences
- Influence of temperature and depth on their distribution

Canalus finmarchicus and Calanus helgolandicus





- Most studied copepods
- Prey for higher trophic levels
- Different spatial preferences
- Influence of temperature and depth on their distribution

The tool: DINCAE

Data
Interpolating } ✔

Convolutional — works on subset of data, trains on local features

Auto
Encoder } — compress/decompress based on its essential features

https://github.com/gher-uliege/DINCAE.jl
doi 10.5281/zenodo.5575066

Want to know more about DINCAE?

Barth *et al.* (2020)

Python version - not maintained

10.5194/gmd-13-1609-2020

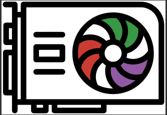
Barth *et al.* (2022)

Julia version

10.5194/gmd-15-2183-2022

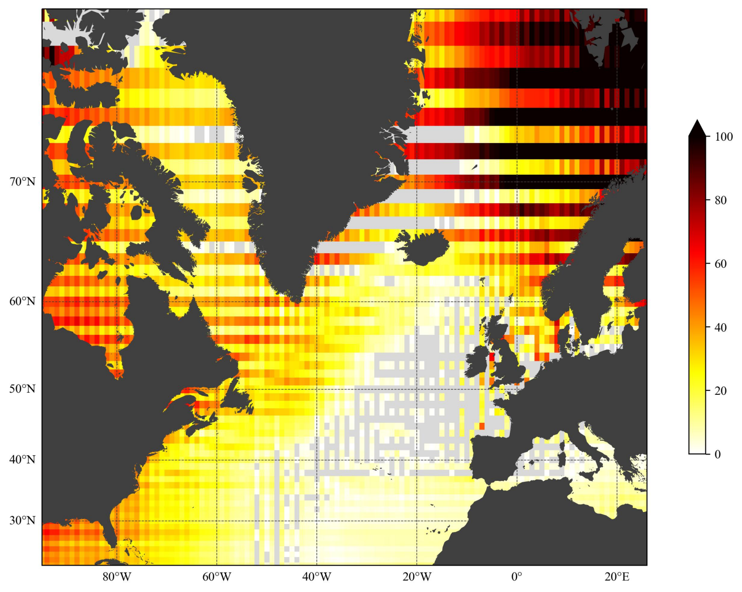
DINCAE is written in Julia and use GPU computing

2018 . . . . Harder, Better, Faster, Stronger
2024 . . . . Faster, more mature



https://juliagpu.org

At first it didn't look that nice



At first it didn't look that nice

Issues:

1. A lot of negative values (grey pixels)
2. Anomalous values near the boundaries
3. Stripes
4. Extreme values

At first it didn't look that nice

Guys what are you doing ?!?

Pierre

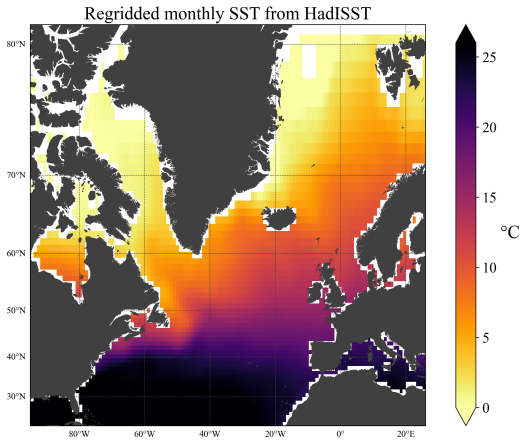
At first it didn't look that nice

Solutions:

1. Apply $\log_{10}(x+1)$ transformation on observations
2. Enlarge domain for computation
3. Perform multivariate analysis
4. Add smoothness constraint in the cost function
5. Tune the hyperparameters

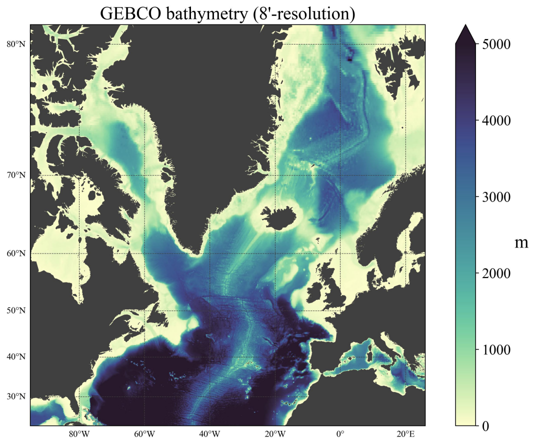
Using environmental variables

Sea surface temperature (Hadley)



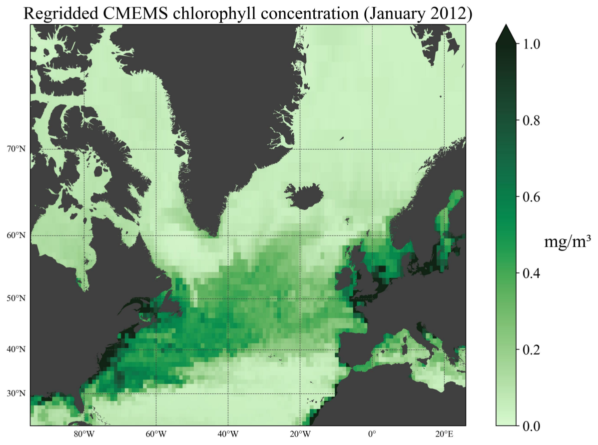
Using environmental variables

Bathymetry (GEBCO, EMODnet)



Using environmental variables

Chlorophyll concentration (CMEMS)



Smoothness constraint in the cost function

Cost function:

$$J_{\text{obs}}(\hat{y}, \hat{\sigma}^2) = \frac{1}{2N} \sum_{ij} \left[\left(\frac{y_{ij} - \hat{y}_{ij}}{\hat{\sigma}_{ij}} \right)^2 + \log(\hat{\sigma}_{ij}^2) \right]$$

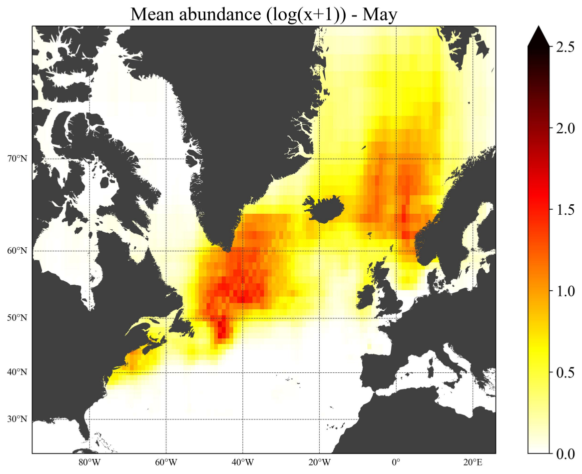
New terms:

$$J_{\text{tot}}(\hat{y}, \hat{\sigma}^2) = J_{\text{obs}}(\hat{y}, \hat{\sigma}^2) + \boxed{\alpha_y} \nabla_d^2 \hat{y} + \boxed{\alpha_{\sigma^2}} \nabla_d^2 \hat{\sigma}^2$$

coefficients to control smoothness constraint

Effect of the Laplacian penalty

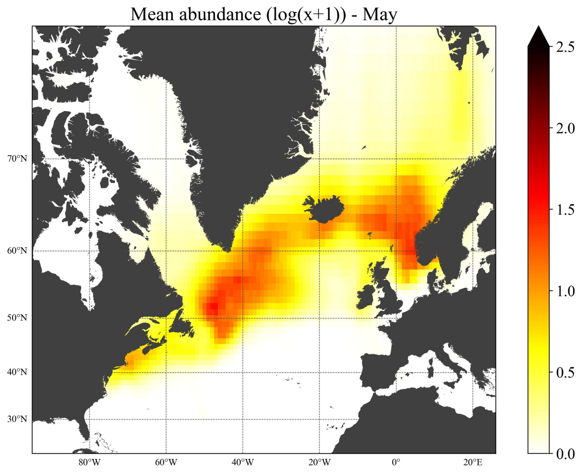
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-6}$$

Effect of the Laplacian penalty

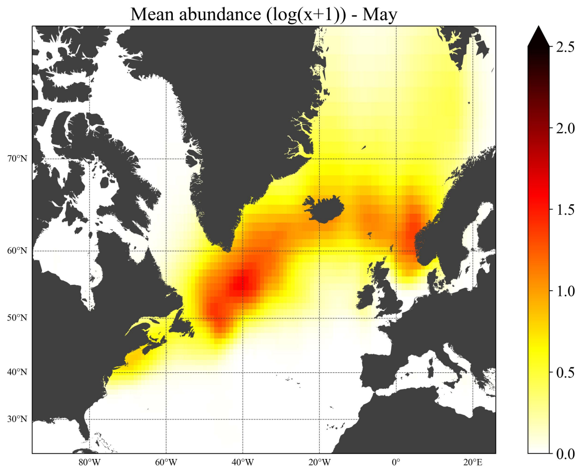
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-5}$$

Effect of the Laplacian penalty

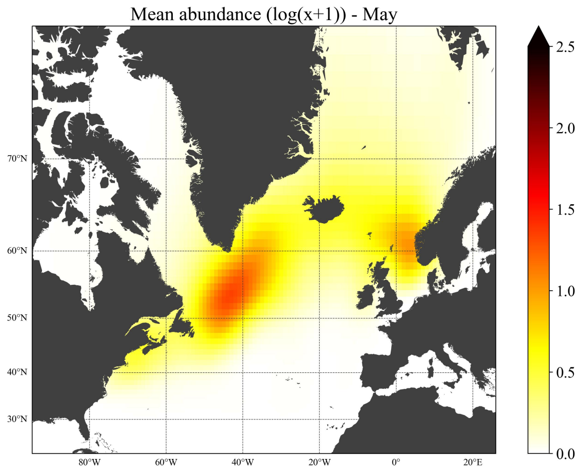
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-4}$$

Effect of the Laplacian penalty

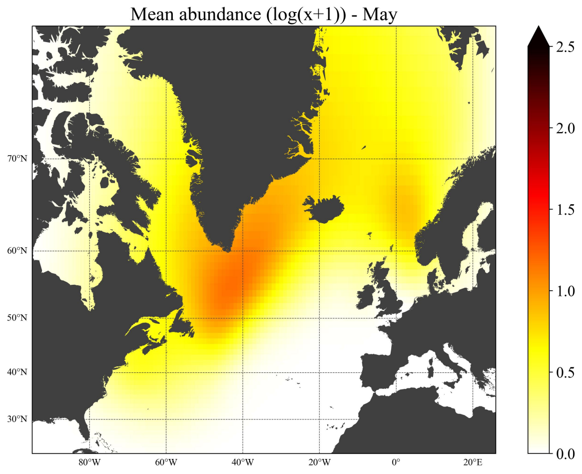
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-3}$$

Effect of the Laplacian penalty

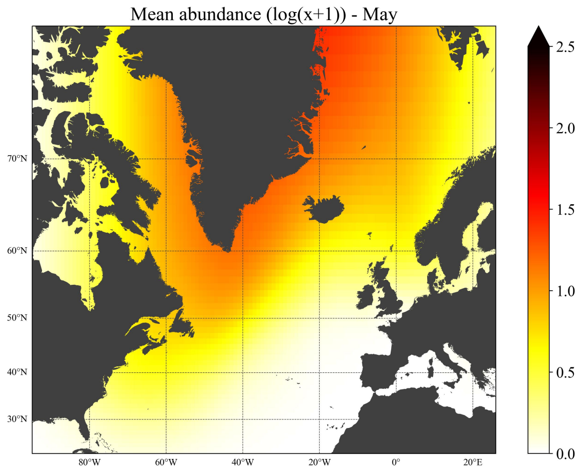
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-2}$$

Effect of the Laplacian penalty

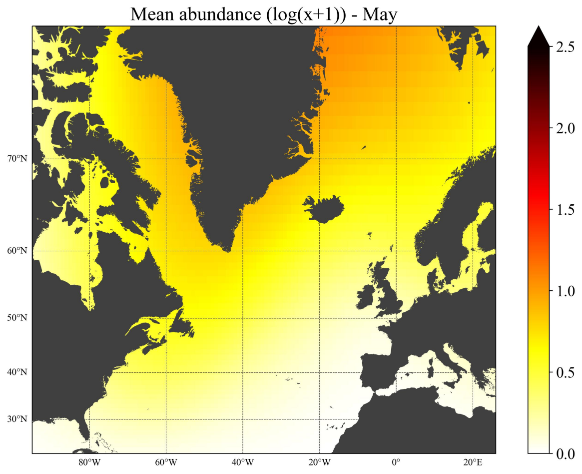
Increasing the coefficient value in the cost function helps smoothing the results



$$\alpha_y = 10^{-1}$$

Effect of the Laplacian penalty

Increasing the coefficient value in the cost function helps smoothing the results



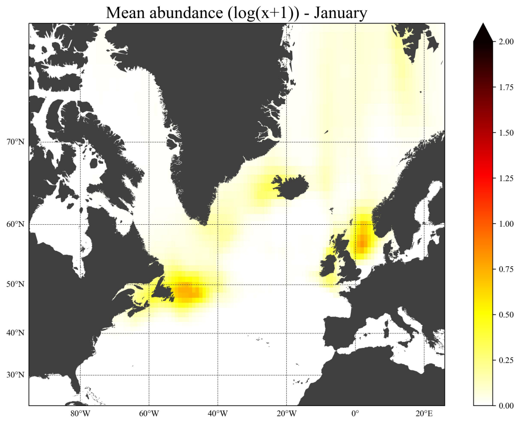
$$\alpha_y = 1$$

Let's look at some results

Seasonal cycle of *C. finmarchicus*

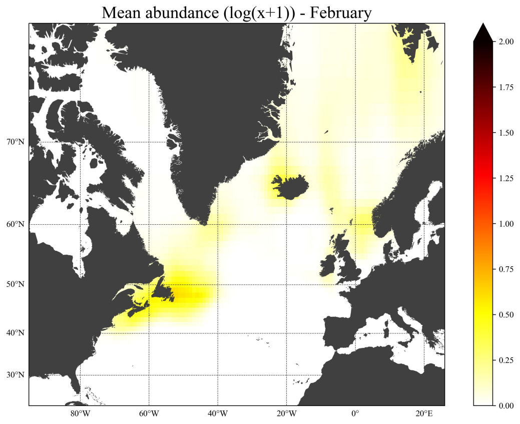
Let's look at some results

Seasonal cycle of *C. finmarchicus*



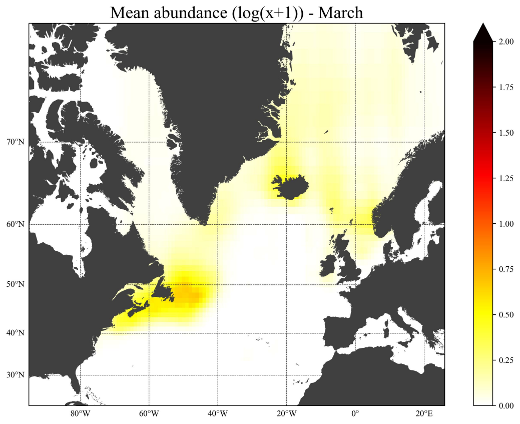
Let's look at some results

Seasonal cycle of *C. finmarchicus*



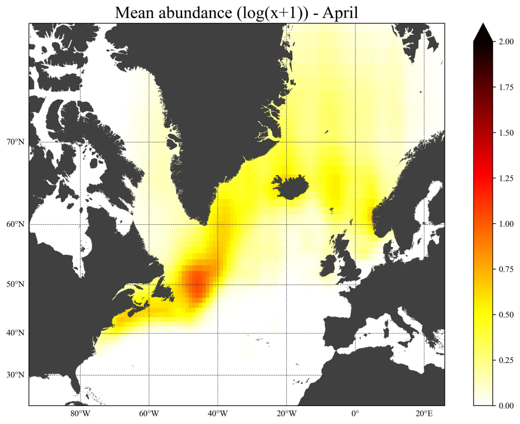
Let's look at some results

Seasonal cycle of *C. finmarchicus*



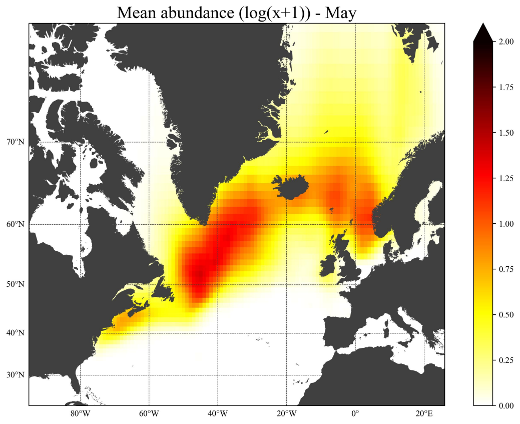
Let's look at some results

Seasonal cycle of *C. finmarchicus*



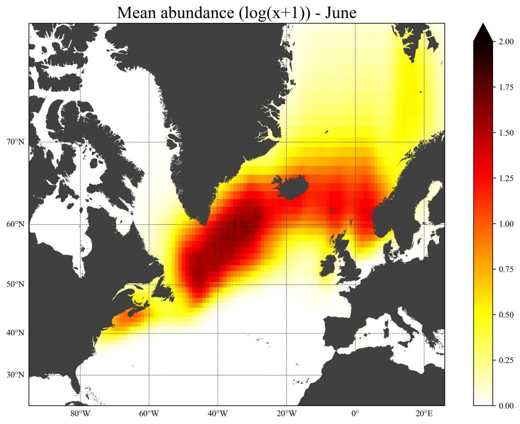
Let's look at some results

Seasonal cycle of *C. finmarchicus*



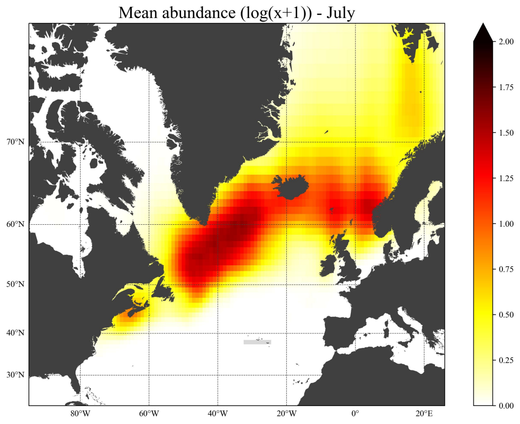
Let's look at some results

Seasonal cycle of *C. finmarchicus*



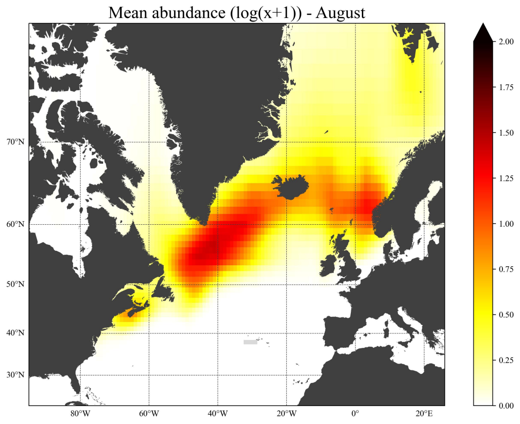
Let's look at some results

Seasonal cycle of *C. finmarchicus*



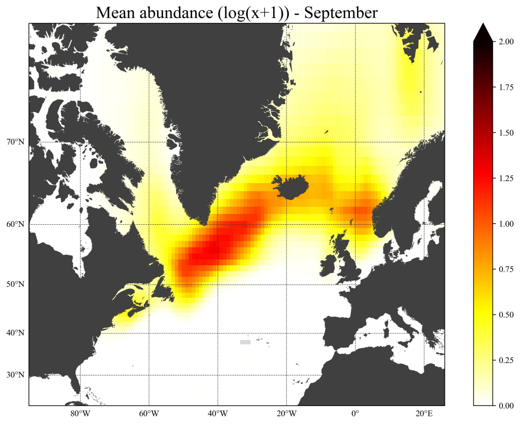
Let's look at some results

Seasonal cycle of *C. finmarchicus*



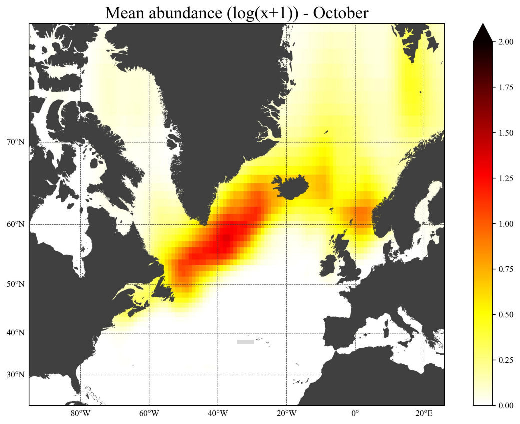
Let's look at some results

Seasonal cycle of *C. finmarchicus*



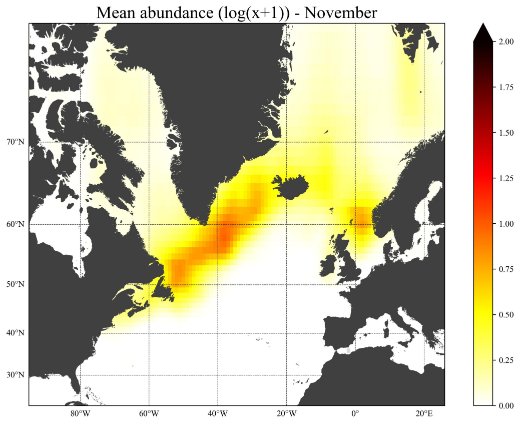
Let's look at some results

Seasonal cycle of *C. finmarchicus*



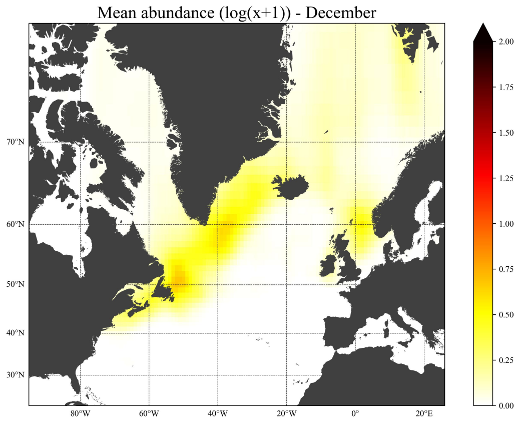
Let's look at some results

Seasonal cycle of *C. finmarchicus*

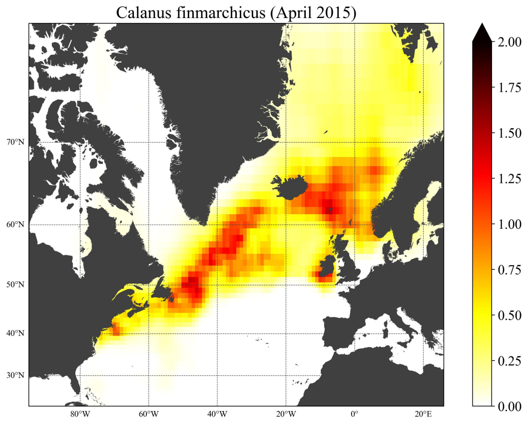


Let's look at some results

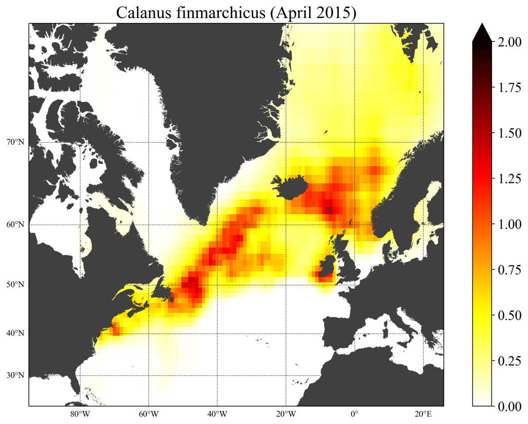
Seasonal cycle of *C. finmarchicus*

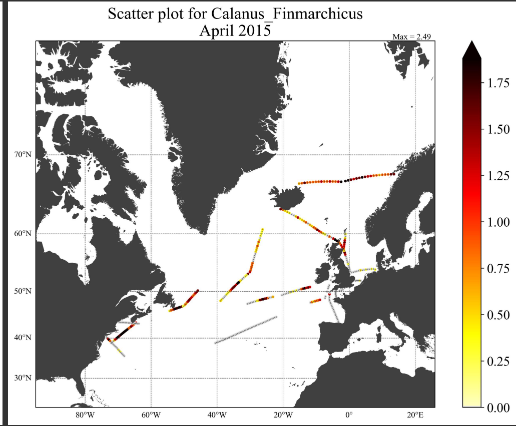


Field for a given month



Field for a given month

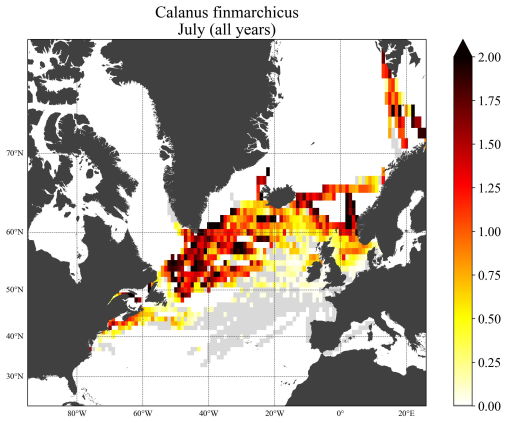




Let's look at some results

Comparison with the observations

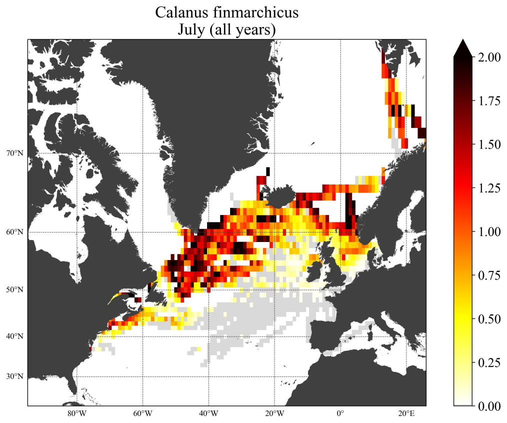
Left = binned observations

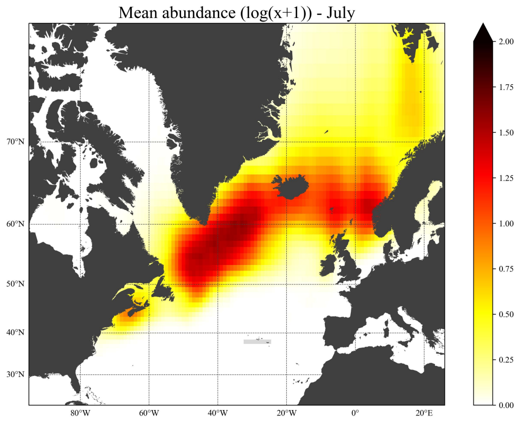


Let's look at some results

Comparison with the observations

Left = binned observations **Right =** gridded fields



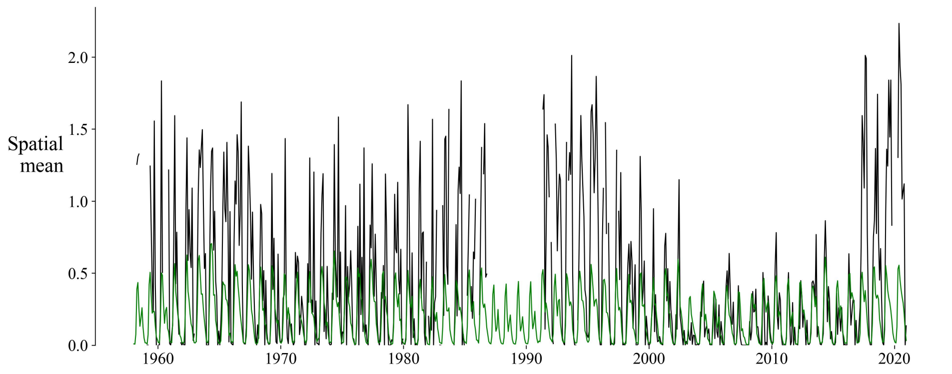


Time series

Spatial average of the C. finmarchicus abundance for each month.

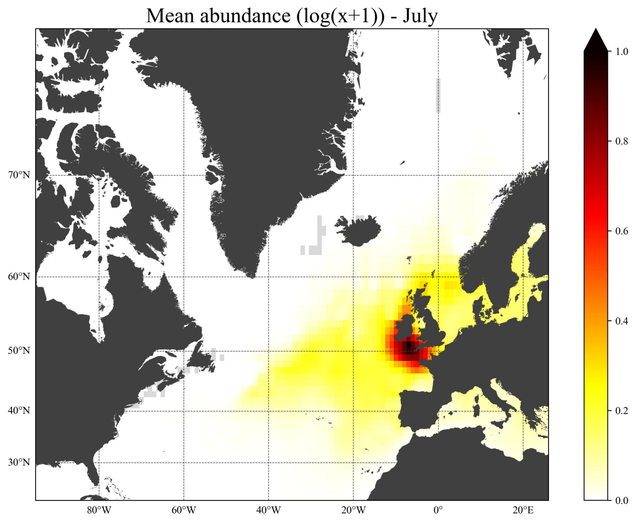
Green = observations

Black = gridded field



Let's have a look at another species

Calanus helgolandicus, July mean



Tuning hyperparameters

Cross-validation:

1. Remove 10% of the observations
2. Perform reconstruction
3. Compare the removed observations with the reconstruction at the same locations
4. Repeat with another set of parameters
5. Minimise the RMS error with the validation dataset

Tuning hyperparameters

Yes, but...

Tuning hyperparameters

number of parameters $\simeq$ 18
number of tested values per parameter $\simeq$ 5
time to run one reconstruction $\simeq$ 2 hours

$18^{5} \times 2$ = 3779136 hours

Tuning hyperparameters

number of parameters $\simeq 18$
number of tested values per parameter $\simeq 5$
time to run one reconstruction $\simeq 2$ hours

$$18^5 \times 2 = 3779136 \text{ hours} = 157464 \text{ days}$$

Tuning hyperparameters

1. Focus on a few, relevant parameters:
   number of epochs, learning rate, Laplacian penalty
2. Test random sets of values in pre-determined ranges
3. Fix the other parameters

How it (almost) ends...

Now that looks pretty good, shall we try to publish the results?

Pierre

Charles

Yes, sure! And maybe present it to IMDIS confer-ence

OK, why not?!

Pierre

Conclusions

1. With a neural network technique,
   we could create realistic monthly gridded fields

Conclusions

1. With a neural network technique,
   we could create realistic monthly gridded fields
2. It was (obviously) necessary
   to tune the hyperparameters for the application

Conclusions

1. With a neural network technique, we could create realistic monthly gridded fields
2. It was (obviously) necessary to tune the hyperparameters for the application
3. Adding environmental variables contribued to improve the results

Let's get FAIR

Findable

Accessible

Interoperable

Reusable

Let's get FAIR

Findable

Accessible

Interoperable

R~~eusable~~ Reproducible

Let's get FAIR

Provide access to:

1. The same input dataset
2. The software code (same version/release)
3. The scripts or notebooks created to use the code on the data
4. The intermediate assets: bathymetry, sea surface temperature, ...

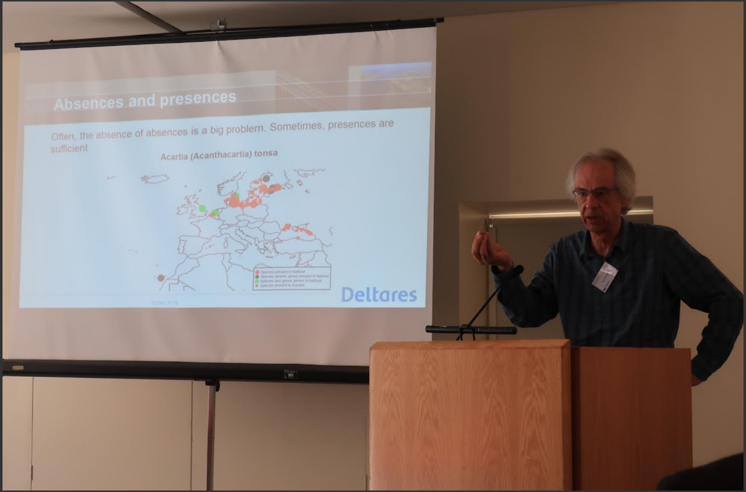
Future work

1. Optimisation of more parameters
2. Increased spatial resolution:
   $1° \times 1° \rightarrow 0.5° \times 0.5°$
3. Testing of other environmental variables?

99 ∞ finish writing the paper and submit it

"Before you could apply all your **artificial intelligence** *algorithms,*
I had to put **real intelligence** *to prepare the data."*

"Before you could apply all your **artificial intelligence** *algorithms,*
I had to put **real intelligence** *to prepare the data."* [Peter Herman]



Credits

DINCAE code	Alexander Barth
Data extraction	Pierre Hélaouët
Data preparation	Charles Troupin
Experiment design	Charles, Alex, Pierre
Bathymetry	GEBCO, EMODnet
Sea surface temperature	Hadley SST
Chlorophyll concentration	Copernicus Marine Service
Computing resources	Lucia, Abacus

27-29 May 2024

imdis

International conference on Marine Data and Information Systems

SeaDataNet AISBL · Havforskningsinstituttet · Ifremer · MARIS · MEDIN · National Oceanography Centre · INGV · iMagine · IODE · eosc Blue-Cloud2026