Inria
BIOCORE

Predicting marine snow abundance with satellite data, a machine learning approach

Directed by

Olivier Bernard

Biocore (future GreenOwl) Team Leader in Inria Sophia

Keywords: modelling, phytoplankton, photobioreactors, nutrient limitation, temperature, ecosystems

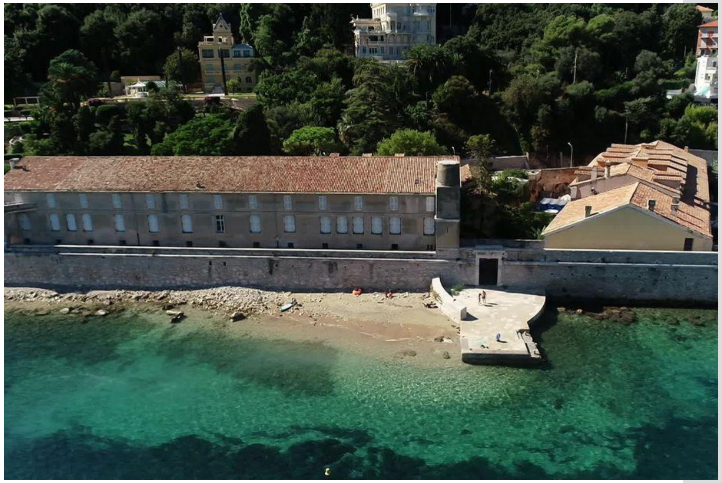
Lionel Guidi

Research Scientist at LOV (Complex team)

Keywords: oceanography, biological oceanography, biogeochemistry, carbon cycle

Where I'm working (LOV)

- Mission : Long-term observation activities to estimate the impacts of climate change and anthropogenic pressure on the marine environment through long term time-series of hydrological, biogeochemical and biological data



- Team COMPLEx: studies the ecology of marine plankton and the oceanic components of biogeochemical cycles

- Stakeholders :

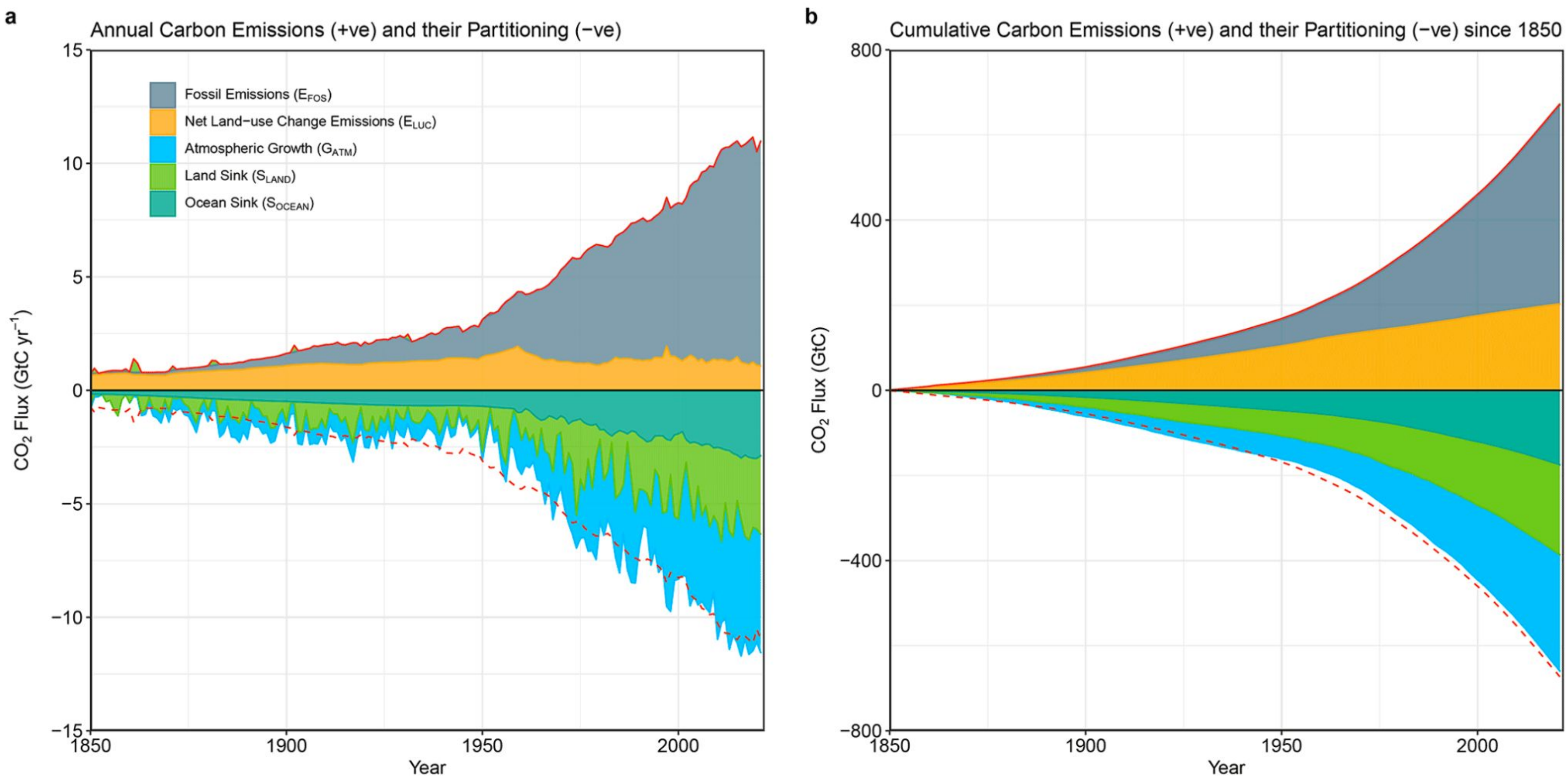
IMEV INSTITUT DE LA MER DE VILLEFRANCHE

SCIENCES SORBONNE UNIVERSITÉ

cnrs

Inria

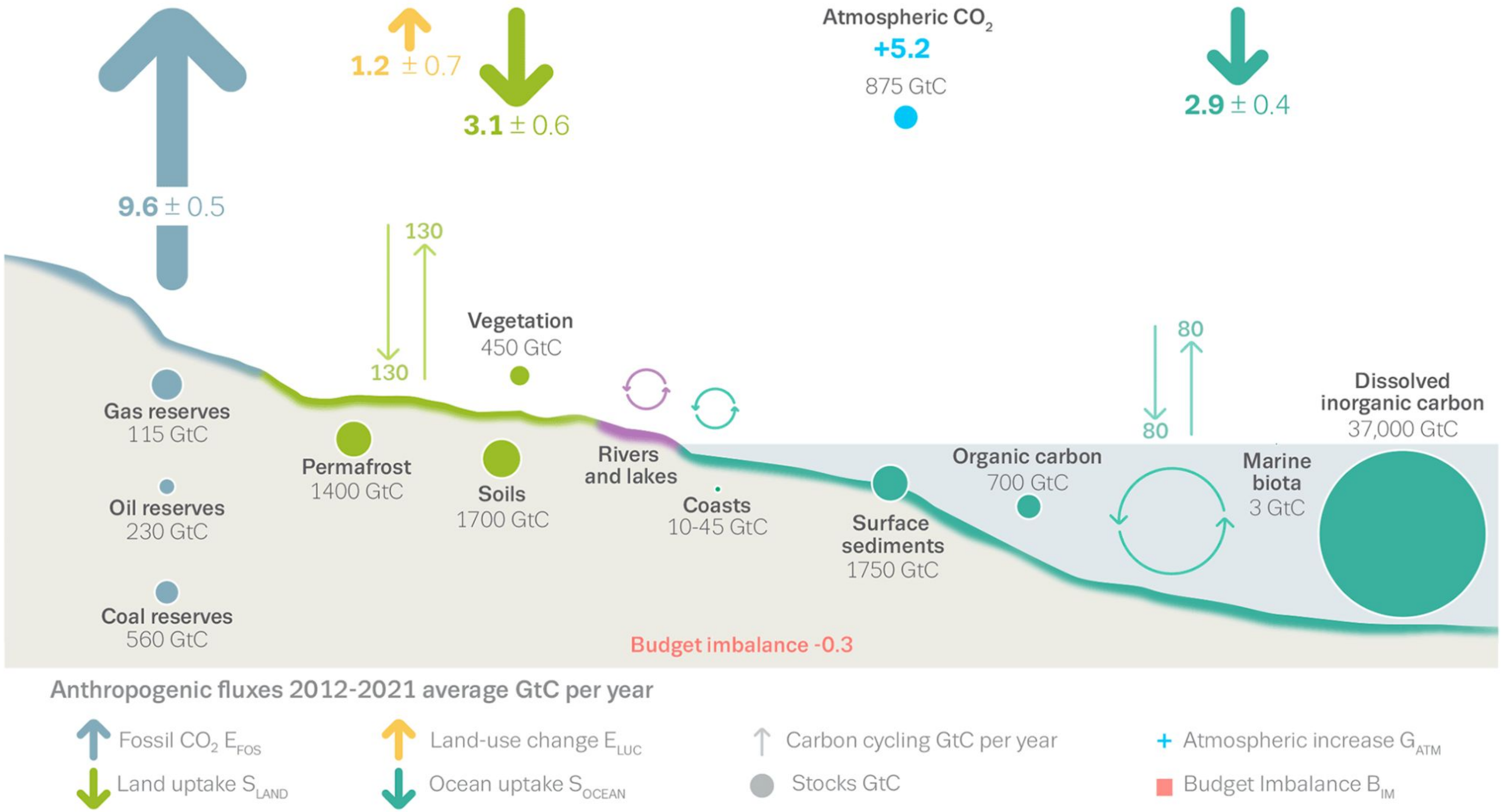
Historic Carbon Emissions and their Partitioning



- The global ocean is one of the main carbon sink, intaking around ⅓ of historic anthropogenic carbon emissions

from "Global Carbon Budget 2022" - Friedlingstein et al. (2022)

Global carbon budget averaged on the decade 2012-2021

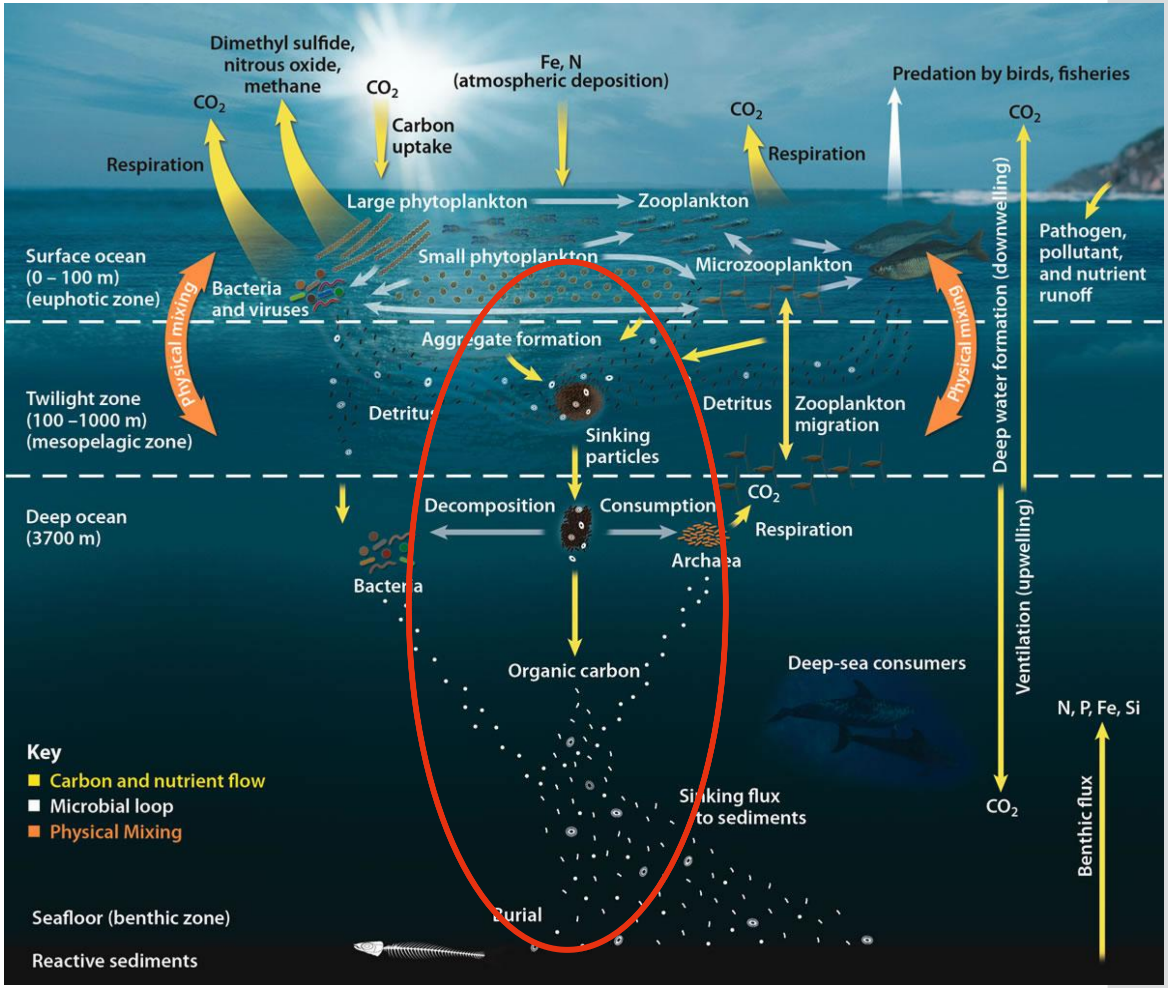


- The global ocean is also the main stock of carbon on Earth, making it a central element in climate modelling

from “Global Carbon Budget 2022” - Friedlingstein et al. (2022)

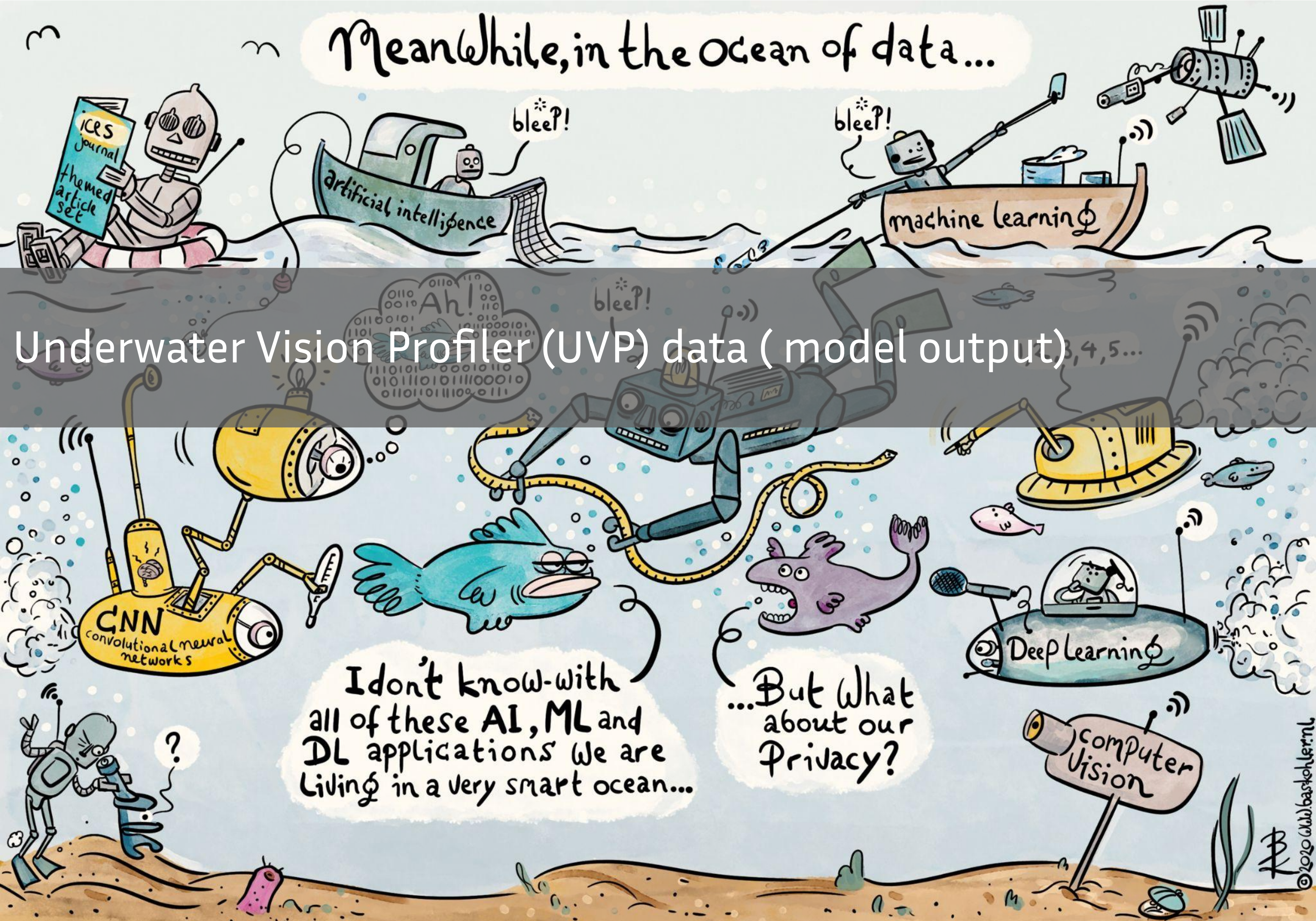
The Biological Carbon Pump

- Marine snow is the main vector of carbon exports to deep water of the biological carbon pump
- The carbon export through marine snow can be seen as a function of its distribution, its velocity and its carbon content
- **Assessing the global distribution can be a precious tool to improve biological carbon fluxes predictions**



© Oak Ridge National Laboratory

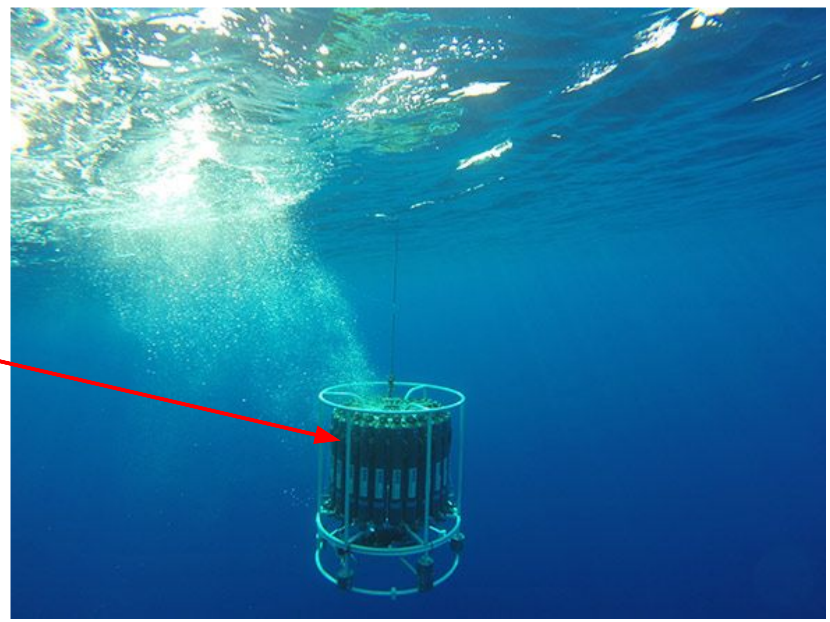
Underwater Vision Profiler (UVP) data (model output)

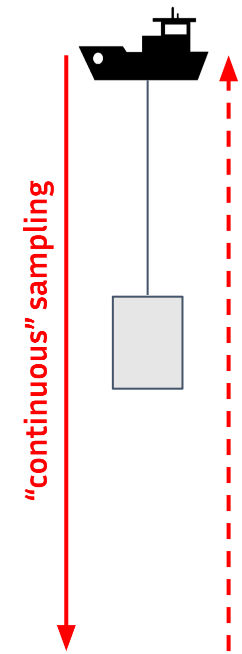


The Underwater Vision Profiller (UVP)

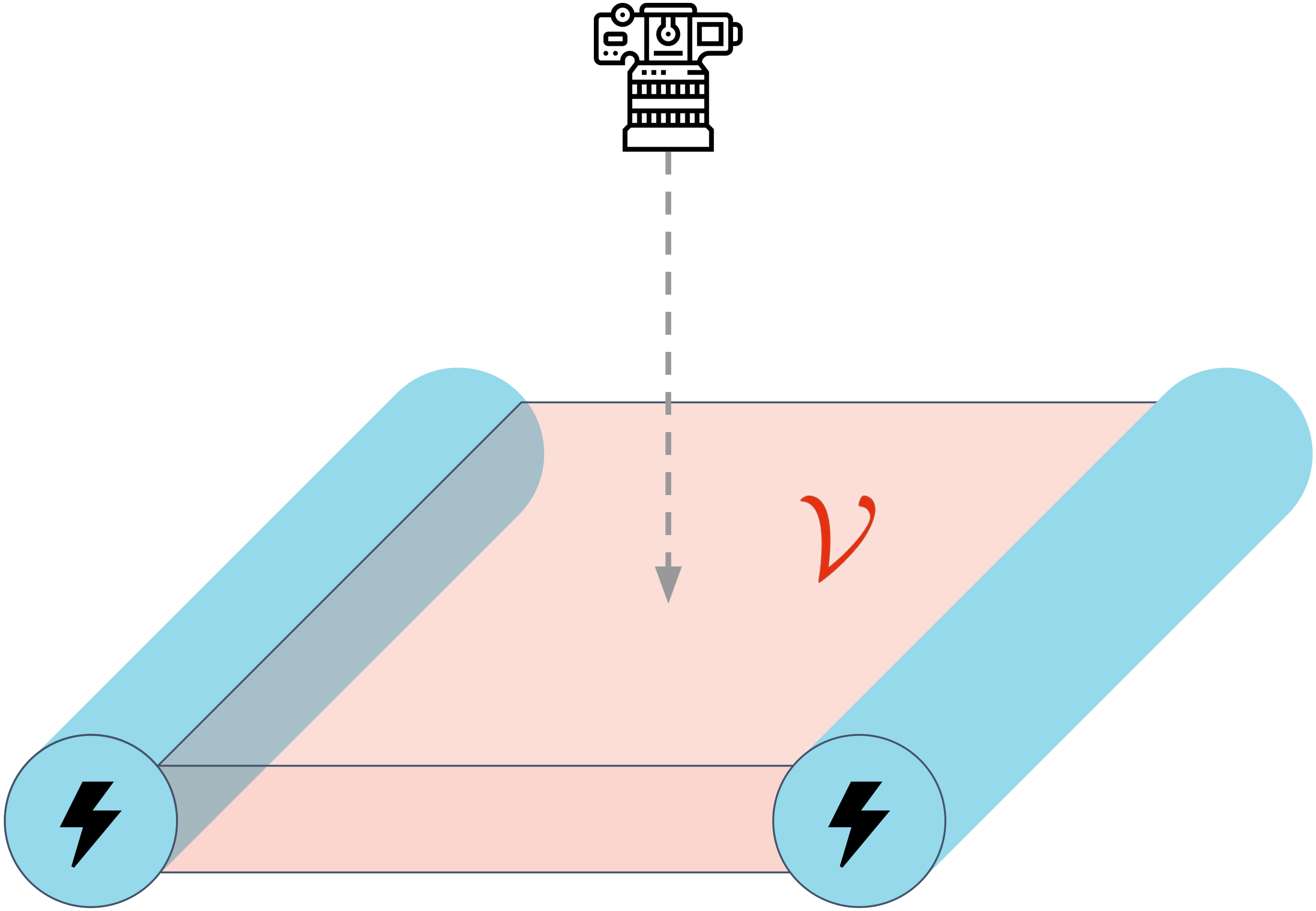
- A **pressure-resistant underwater camera** that can take images of plankton and particles at depths of up to 6,000 m.
- It can take up to 20 pictures per second, and each picture samples a volume of around 1L.

⇒In a 1000 m dive, the UVP can sample up to 20 m^3 of seawater >> 400 liters of waters sampled by Niskin bottle

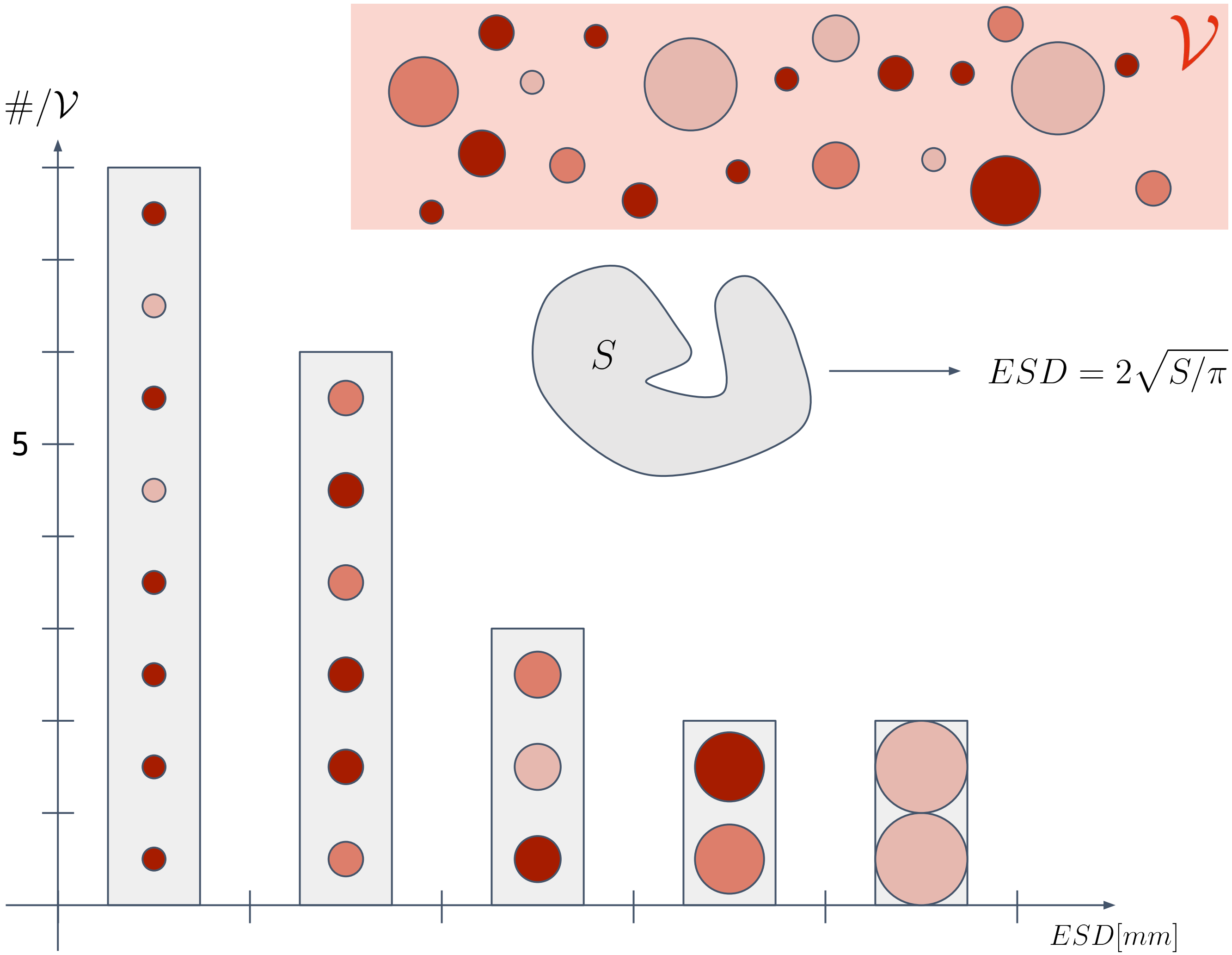




The Underwater Vision Profiller (UVP)



The Underwater Vision Profiller (UVP)



The UVP database (R. Kiko et al., 2022)

Size classes:

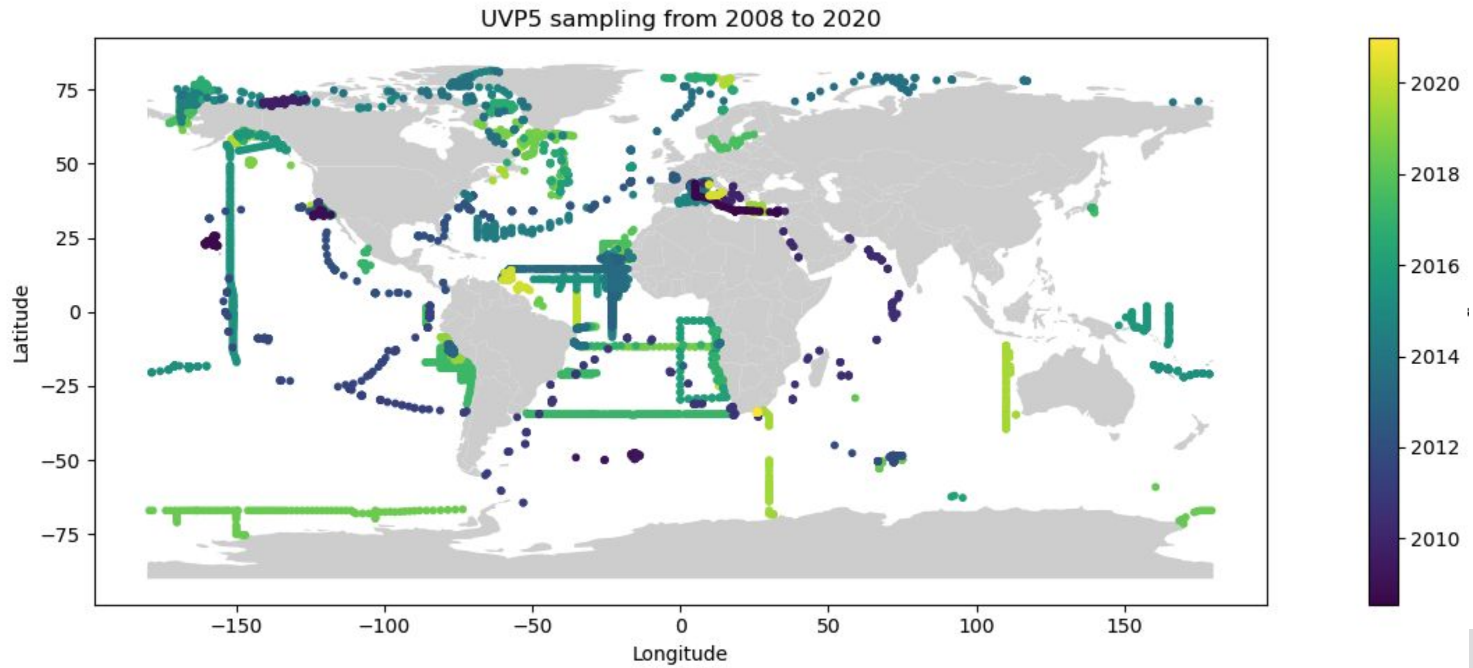
- 28 classes

$\times 2^{1/3}$ $\times 2^{1/3}$

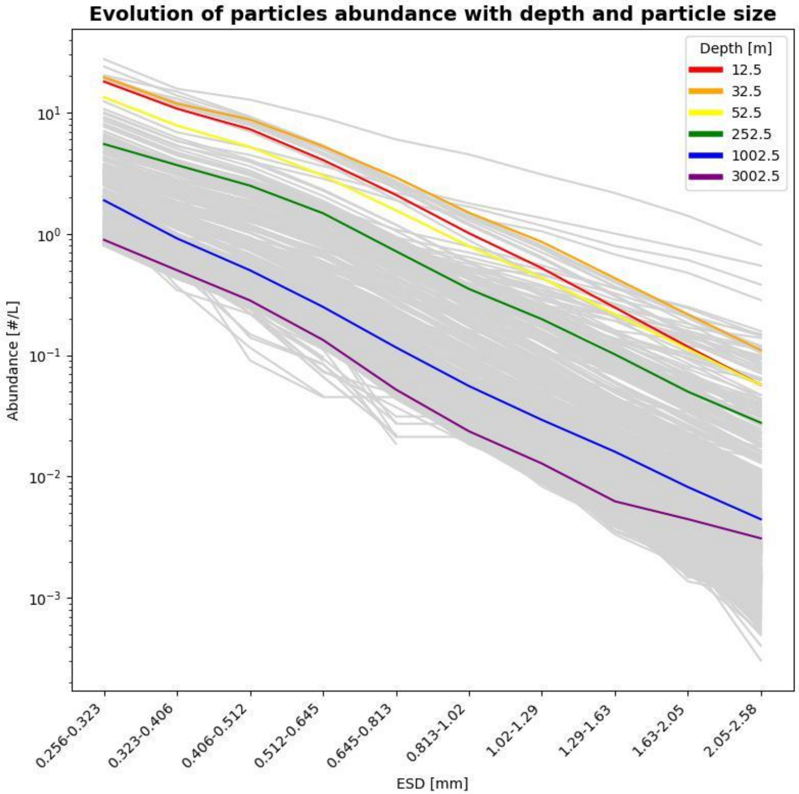
$ESD[mm]$

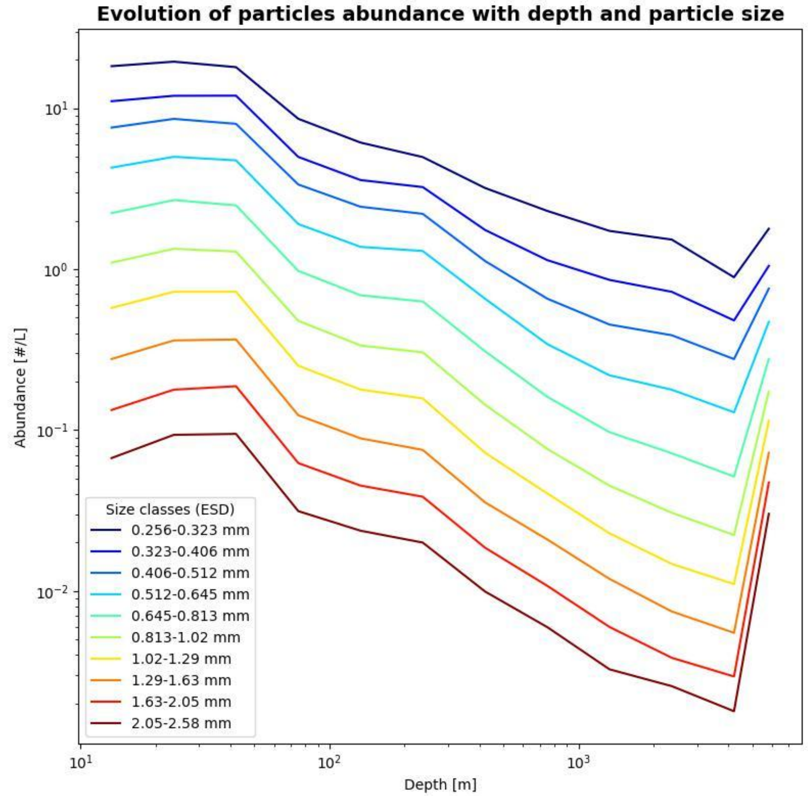
0.04 - 0.05	…	0.41- 0.51	0.51 - 0.65	0.65 - 0.81	…	20.6 - 26

- Pictures are concatenated to create **5 m depth bins**
- **8803 samples**, collected during 139 cruises



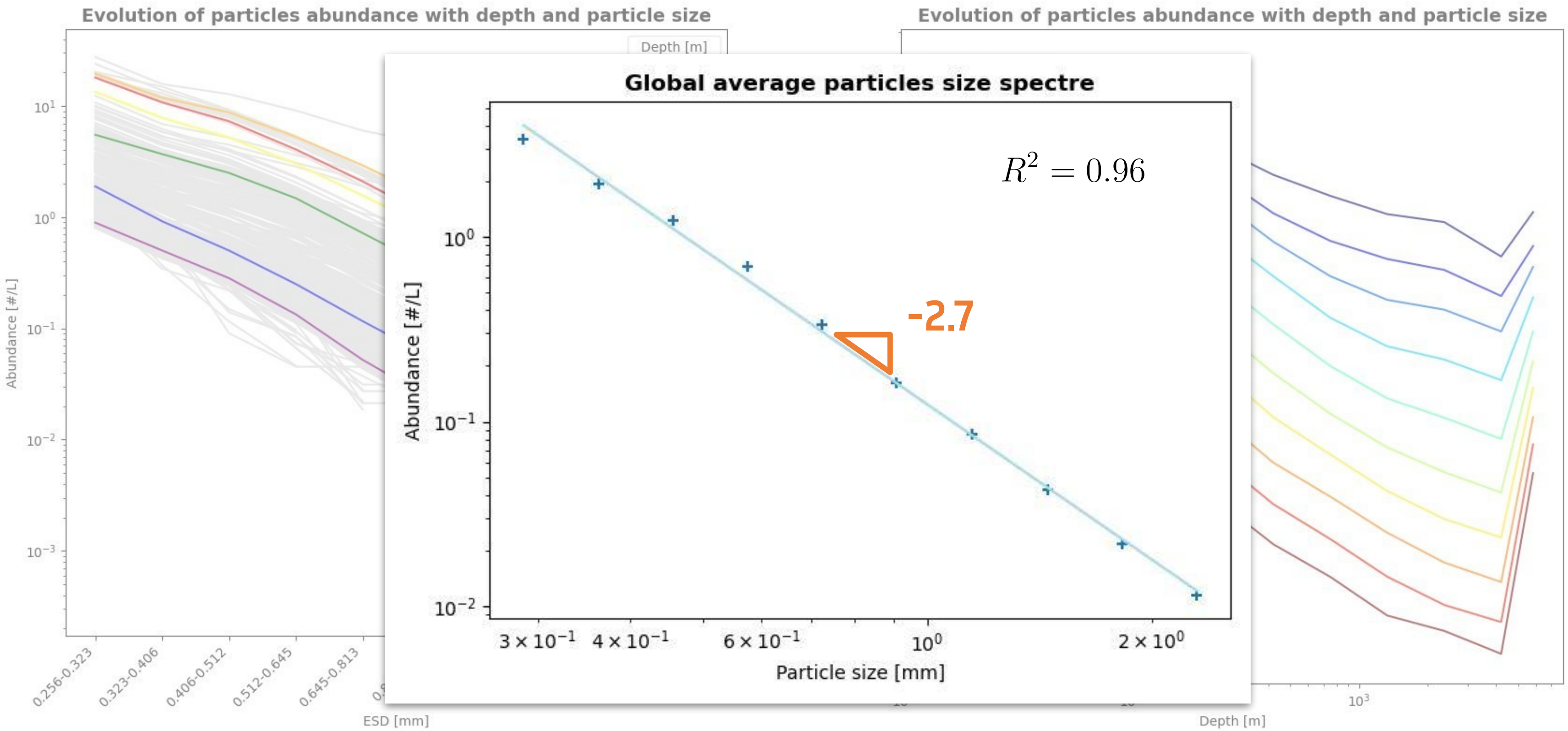
Global averages





- The slope of the particle distribution is quite **independent of the depth** (left panel)
- The particle abundance **decreases with depth under the euphotic zone**, where there is no primary production. The final increase at 6000 m is due to sea floor proximity (right panel)

Global averages

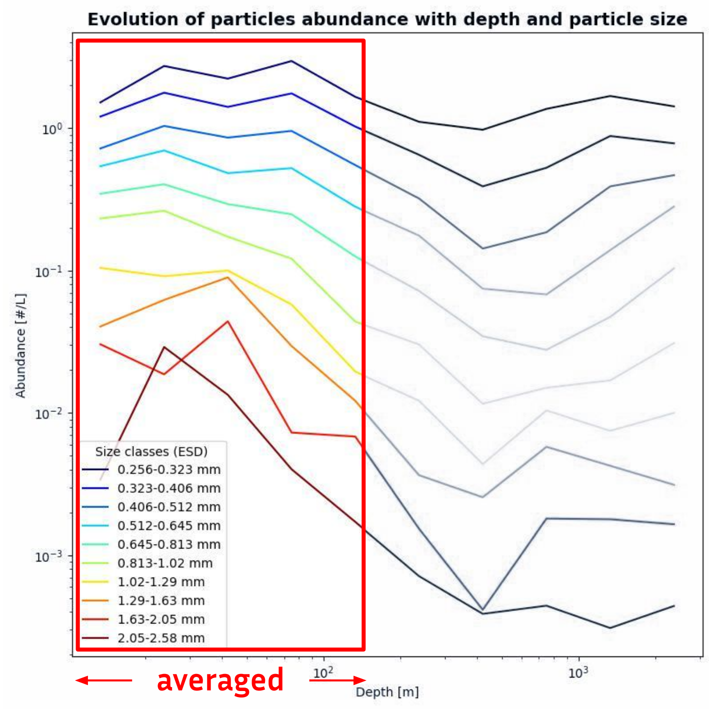


- The slope of the particle distribution is quite **independent of the depth** (left panel)
- The particle abundance **decreases with depth under the euphotic zone**, where there is no primary production. The final increase at 6000 m is due to sea floor proximity (right panel)

The scope of our work (for now)

- Our first models focus on the UVP5 **surface data**: average of particles abundance between 15 and 150 m deep

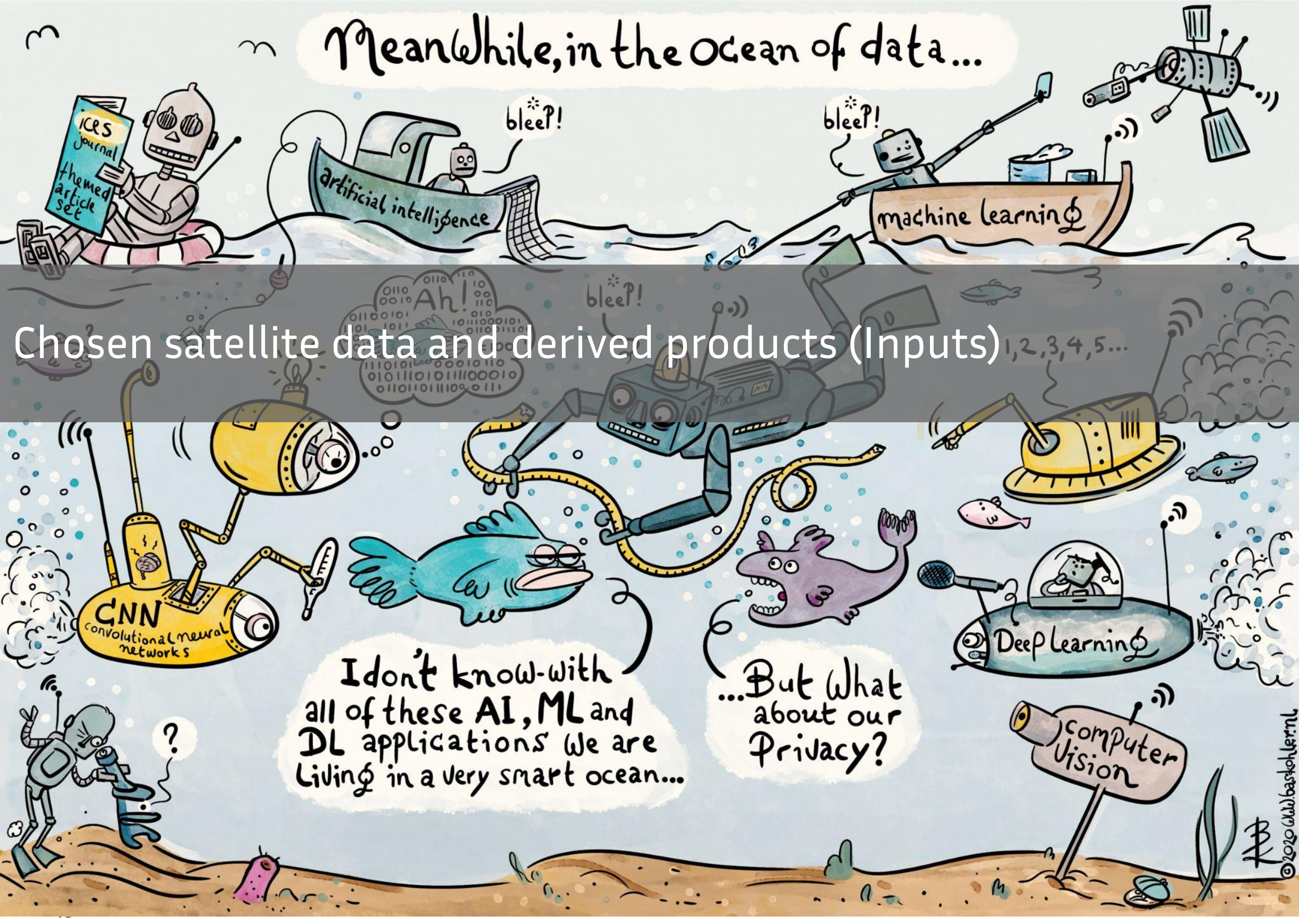
- We only take into account **size classes** between 0.256 mm and 2.58 mm (10 classes)



APERO Campaign

- **40 days oceanographic campaign** in North Atlantic (June 2023). 2 ships and more than 60 scientists on board
- **More than 10 UVPs** were onboard among many other measurement tools.
- The aim of the campaign was to sample **ocean eddies and fronts** at fine scale to better understand **the role of those mesoscale structures in the carbon cycle**

Chosen satellite data and derived products (Inputs)



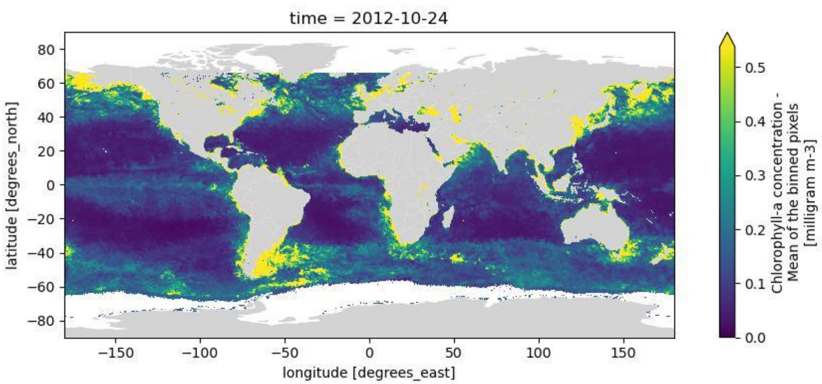
Satellite data

- ESA and other space agencies are putting a lot of effort and resources in remote sensing programs to monitor earth surface
- This allows researchers to access data at a very high space and time resolution

Inria

Context: extrapolation of marine snow dataset with satellite data

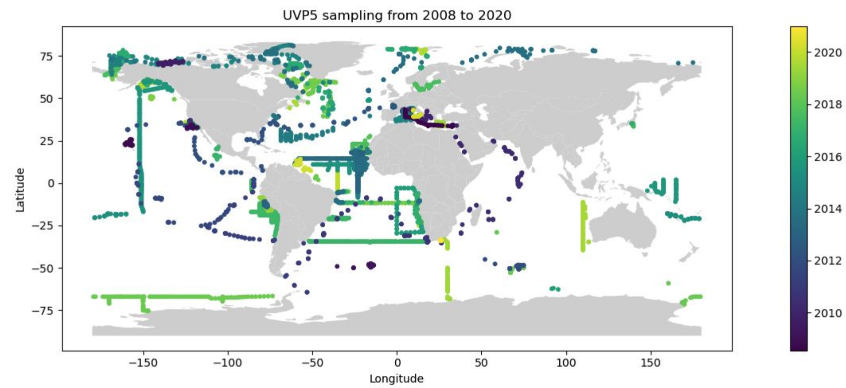
Satellite data



Advantages:
- ➢ High time and space resolution
- ➢ (almost) dense data

Limitations:
- ➢ Data from surface water only
- ➢ Low level of characterization of marine snow

In-situ marine snow data



Advantages:
- ➢ High level of characterization of marine snow particles
- ➢ Deep profiles (up to 6000 m)

Limitations:
- ➢ Sparse data in time and space
- ➢ Geographic and seasonal biases

World Ocean Atlas (Garcia et al., 2019)

- WOA are monthly climatologies of biogeochemistry (BGC) quantities
- Chosen data products for models:

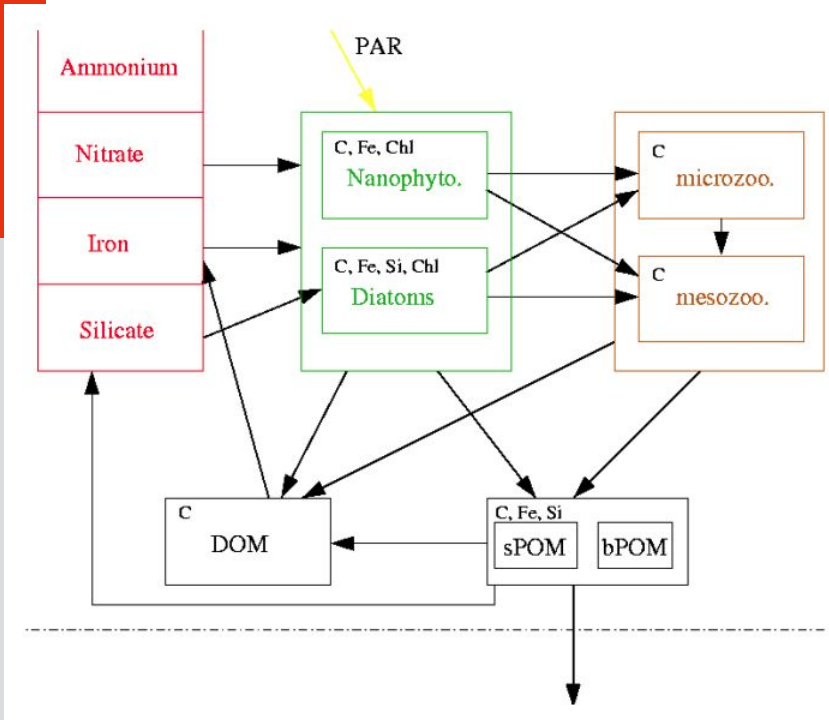
Quantity	Grid size(s)	Time Span
Temperature	1° and 0.25°	1955-2017
Salinity	1° and 0.25°	1955-2017
Density	1° and 0.25°	1955-2017
Conductivity	1° and 0.25°	1981-2010
Nitrate	1°	all available data
Phosphate	1°	all available data
Silicate	1°	all available data
Oxygen (concentration, AOU and saturation)	1°	all available data

Copernicus - Level 4 data

Data source	Grid size(s)	Data Product	Quantities	Frequency
Global Ocean Colour	4 km	Plankton	Chlorophyll concentration	Daily and Monthly
			Biomass of Phyto groups	Monthly
		Reflectance	at 412, 443, 490, 555 and 670 nm	Monthly
		Transparence	KD490, ZSD	Daily and Monthly
			SPM	Monthly
		Optics	BBP, CDM	Monthly
Global Ocean OSTIA SST and Sea Ice	0.05°	SST	-	Daily and Monthly
		SSI	-	Daily and Monthly
Global Ocean SSH And Derived Variables	0.25°	SSH	SLA, ADT	Daily
		Currents	North and West Current velocities and anomalies	Daily

PISCES-v2 (Aumont et al., 2015) coupled with an ocean circulation

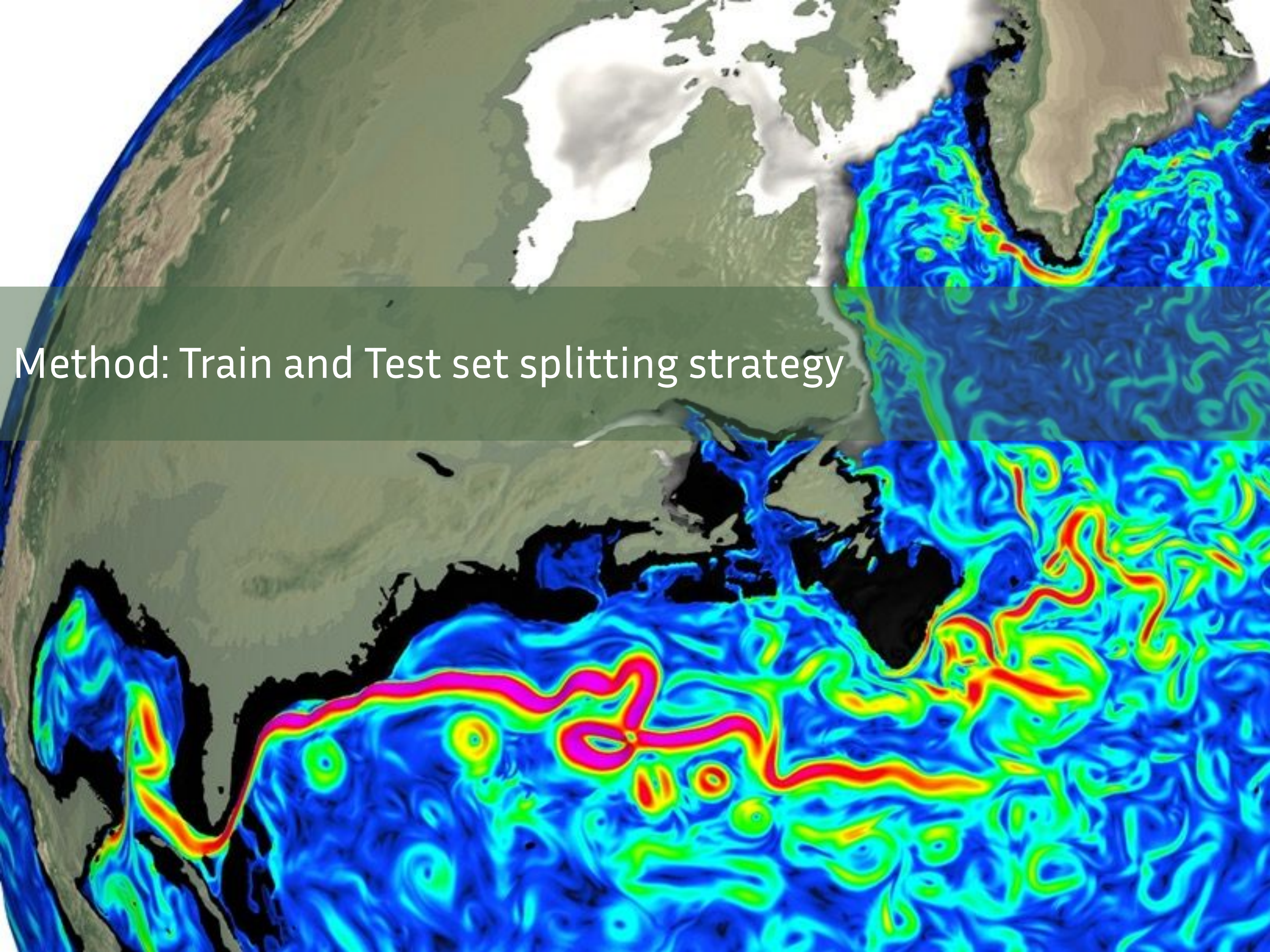
- BGC fields from a global ocean model (*PISCES-v2*) coupled with two physics forcings (*GLORYS2V4-FREE* and *ERA-Interim* atmosphere)
- Grid size: 0.25°



PISCES architecture scheme

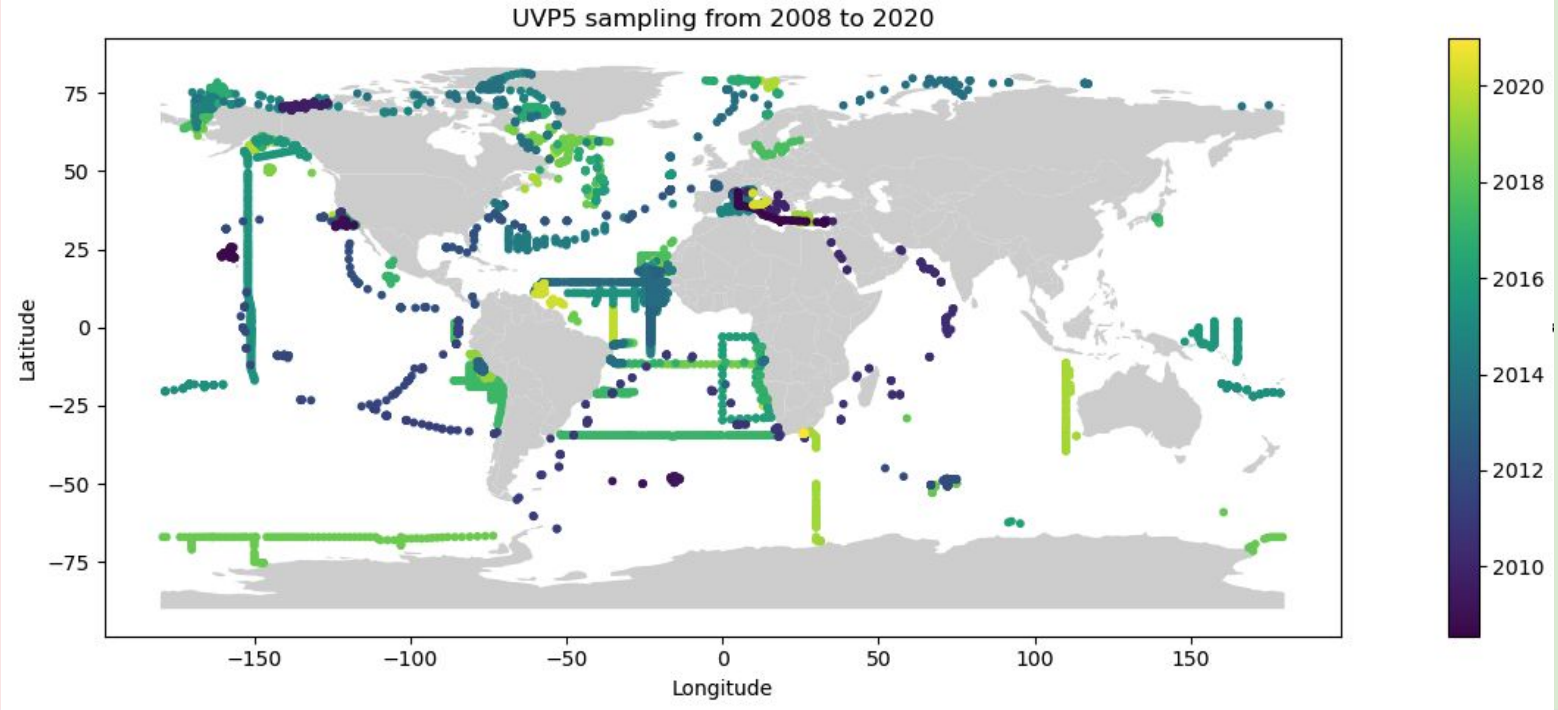
Frequency	Quantities
Daily and Monthly	Chlorophyll
	Nitrate
	Phosphate
	Silicate
	Dissolved oxygen
	Primary production
Monthly	Iron
	Phytoplankton in carbon

Method: Train and Test set splitting strategy

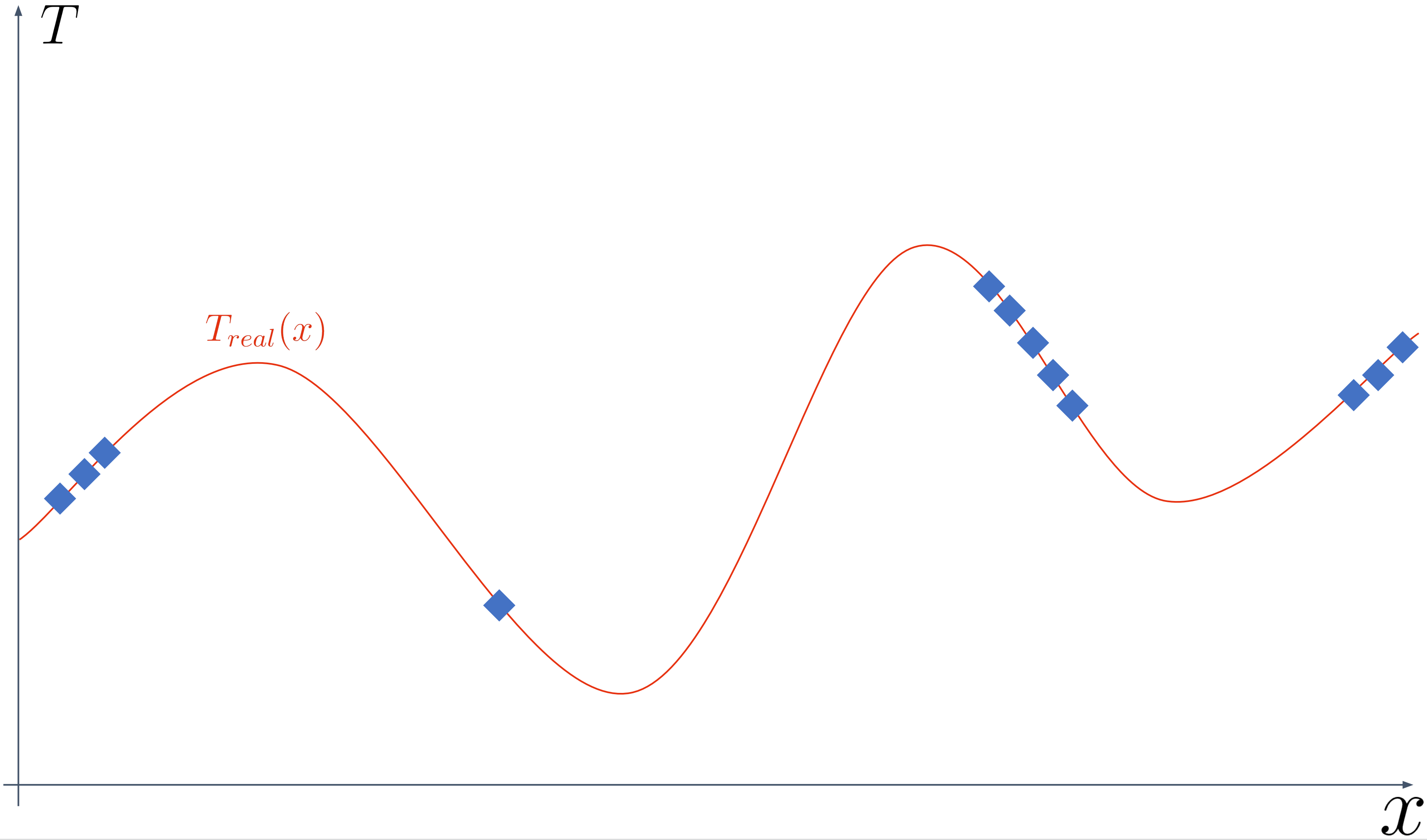


Train/Test splitting strategies specifications

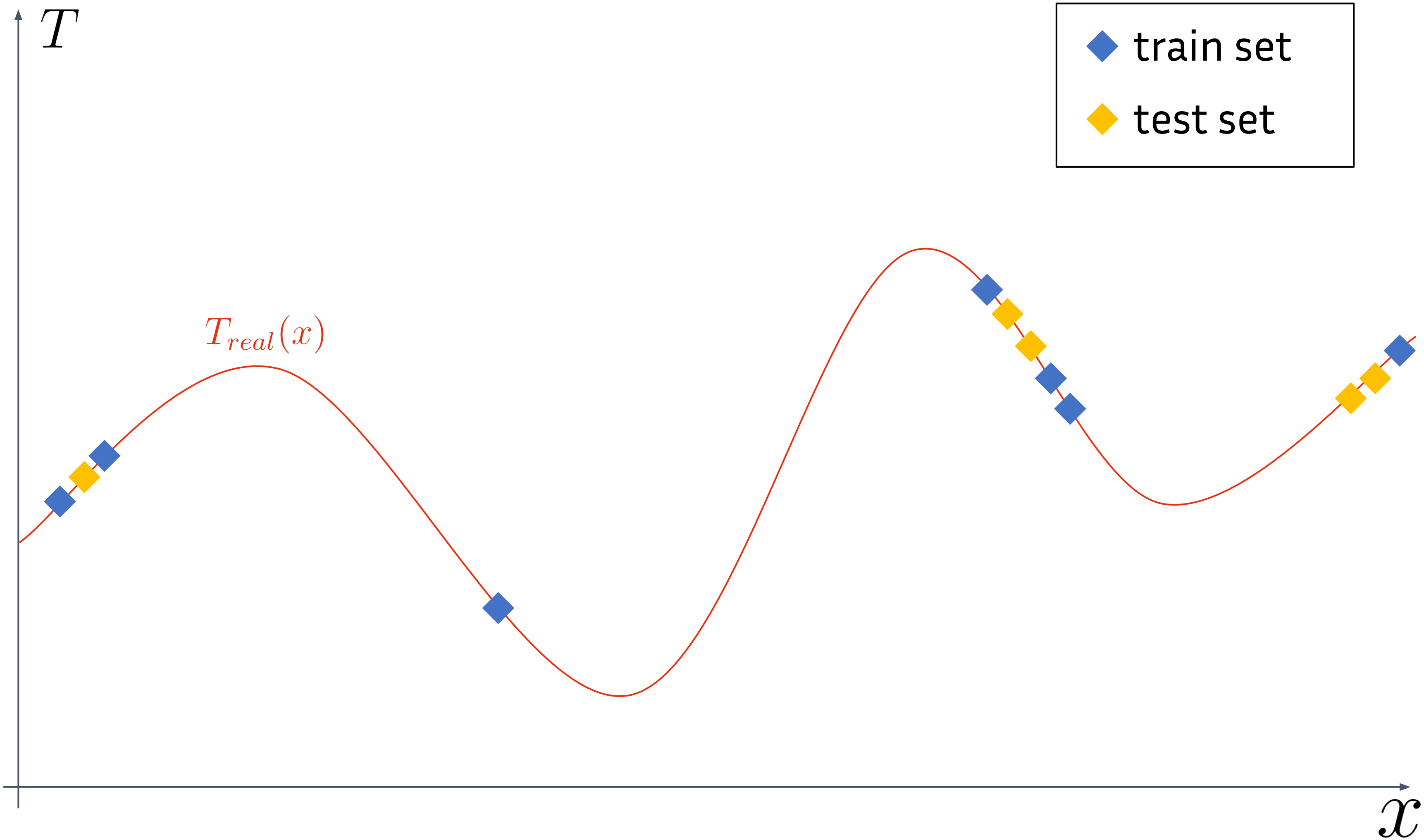
- **The density** of UVP sampling is **very irregular** in time and space



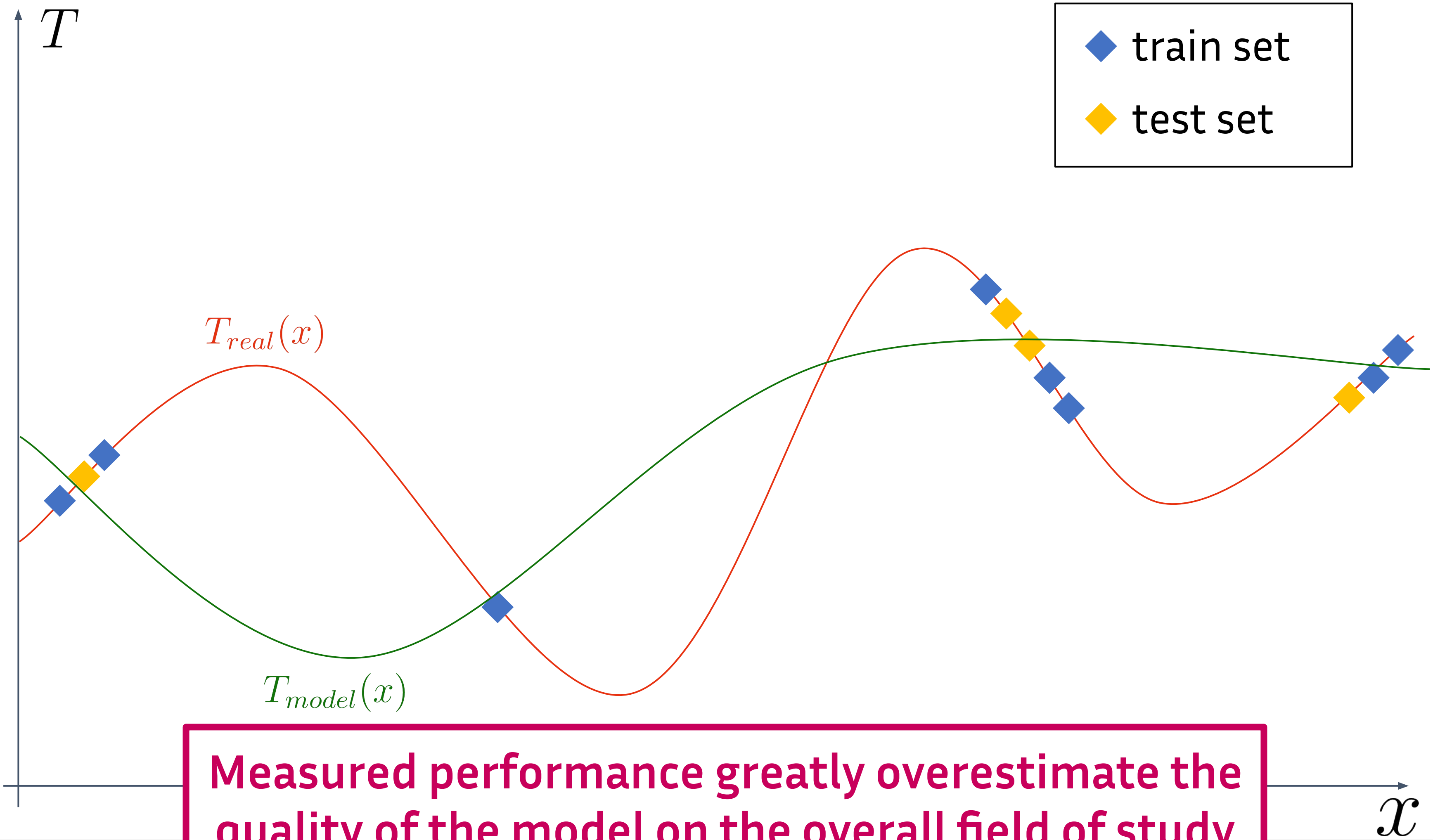
Train/Test splitting strategies specifications



Train/Test splitting strategies specifications



Train/Test splitting strategies specifications



Measured performance greatly overestimate the quality of the model on the overall field of study

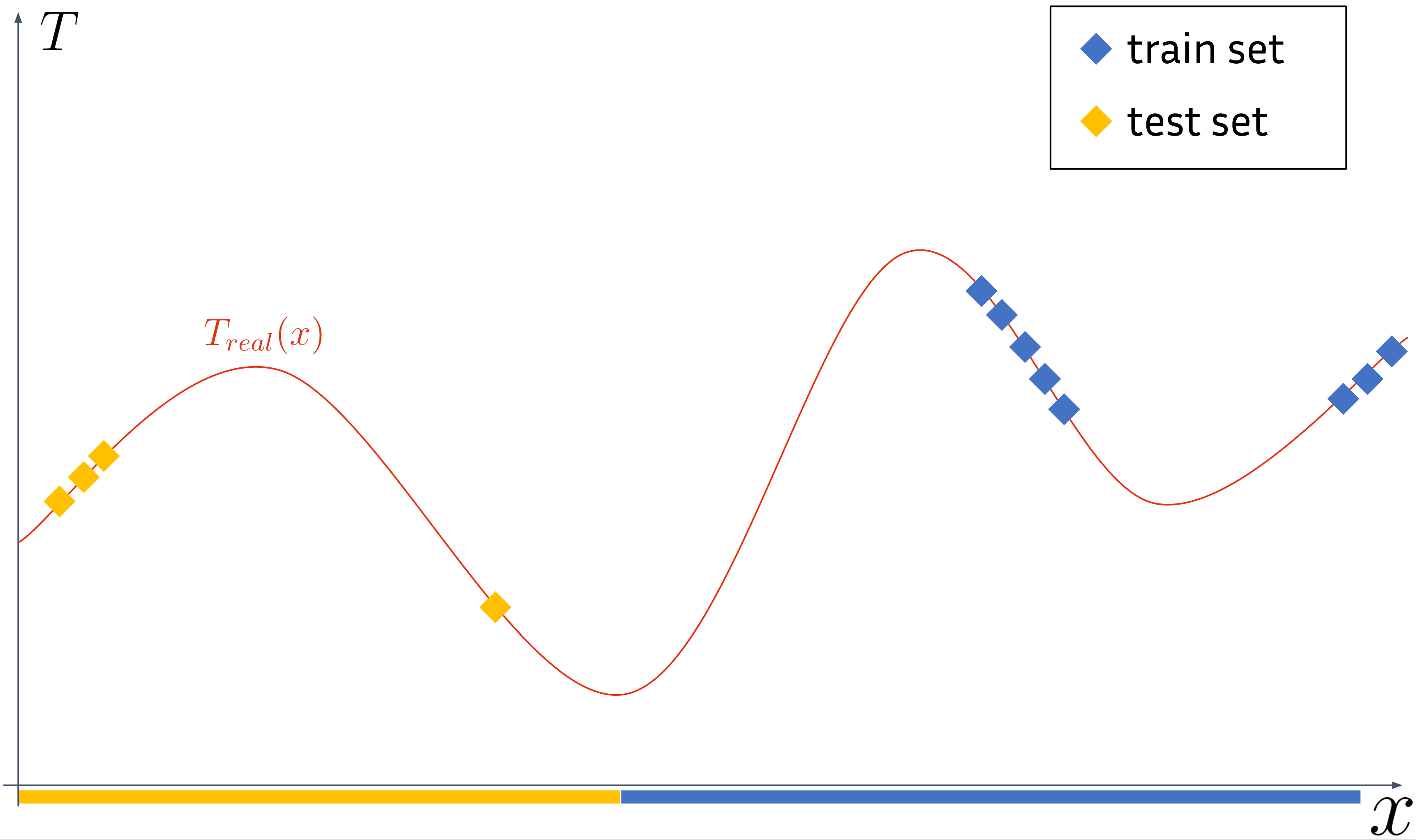
Train/Test splitting strategies specifications

- **The density** of UVP sampling is **very irregular** in time and space

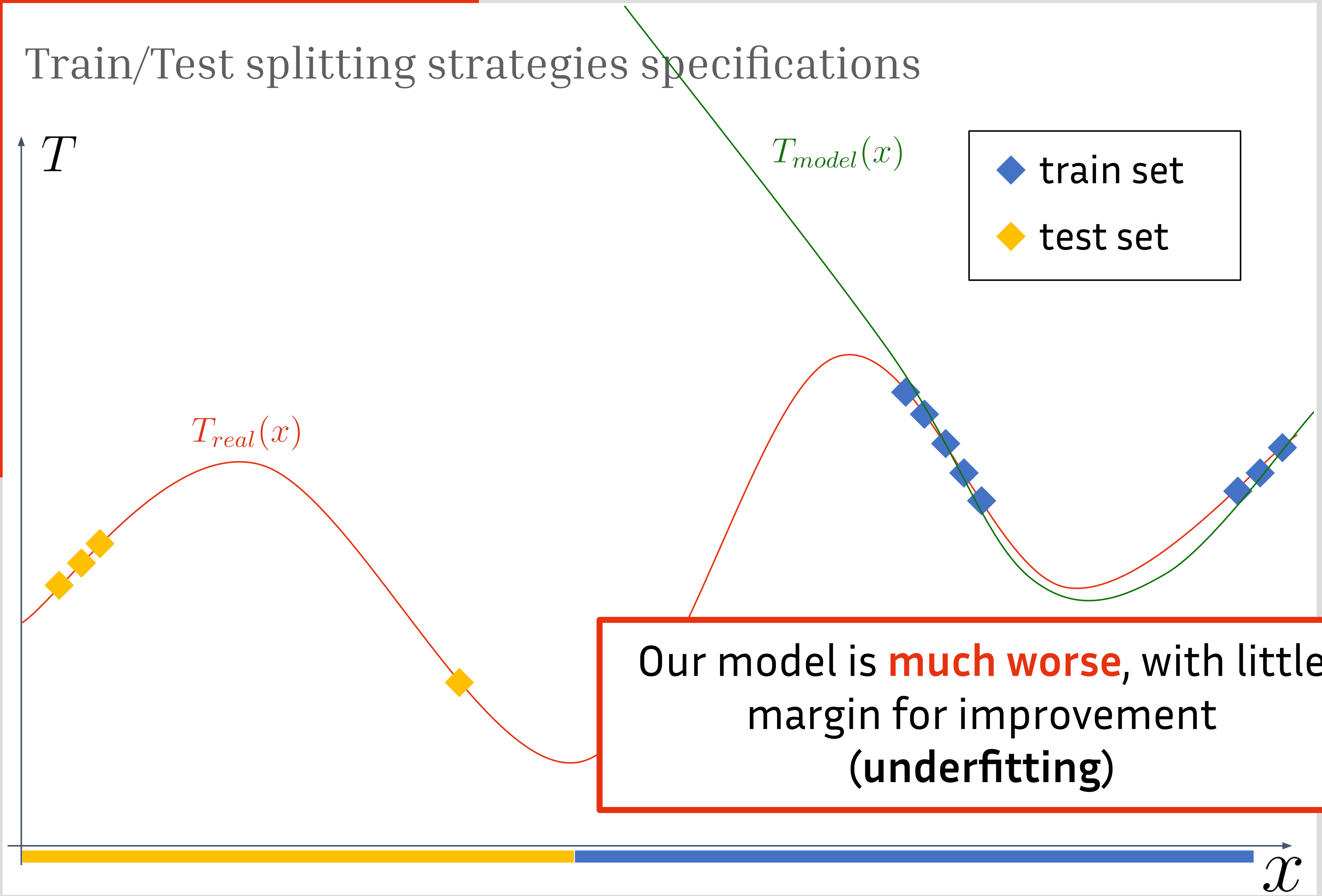
⇒ We need a splitting strategy **resilient to overfitting due to time and space proximity** of samples

- A too conservative approach would make us **lose a lot of information** (seasonality, specificity of each biogeochemical region, local planktonic community, …)

Train/Test splitting strategies specifications



Train/Test splitting strategies specifications



Our model is **much worse**, with little margin for improvement **(underfitting)**

Train/Test splitting strategies specifications

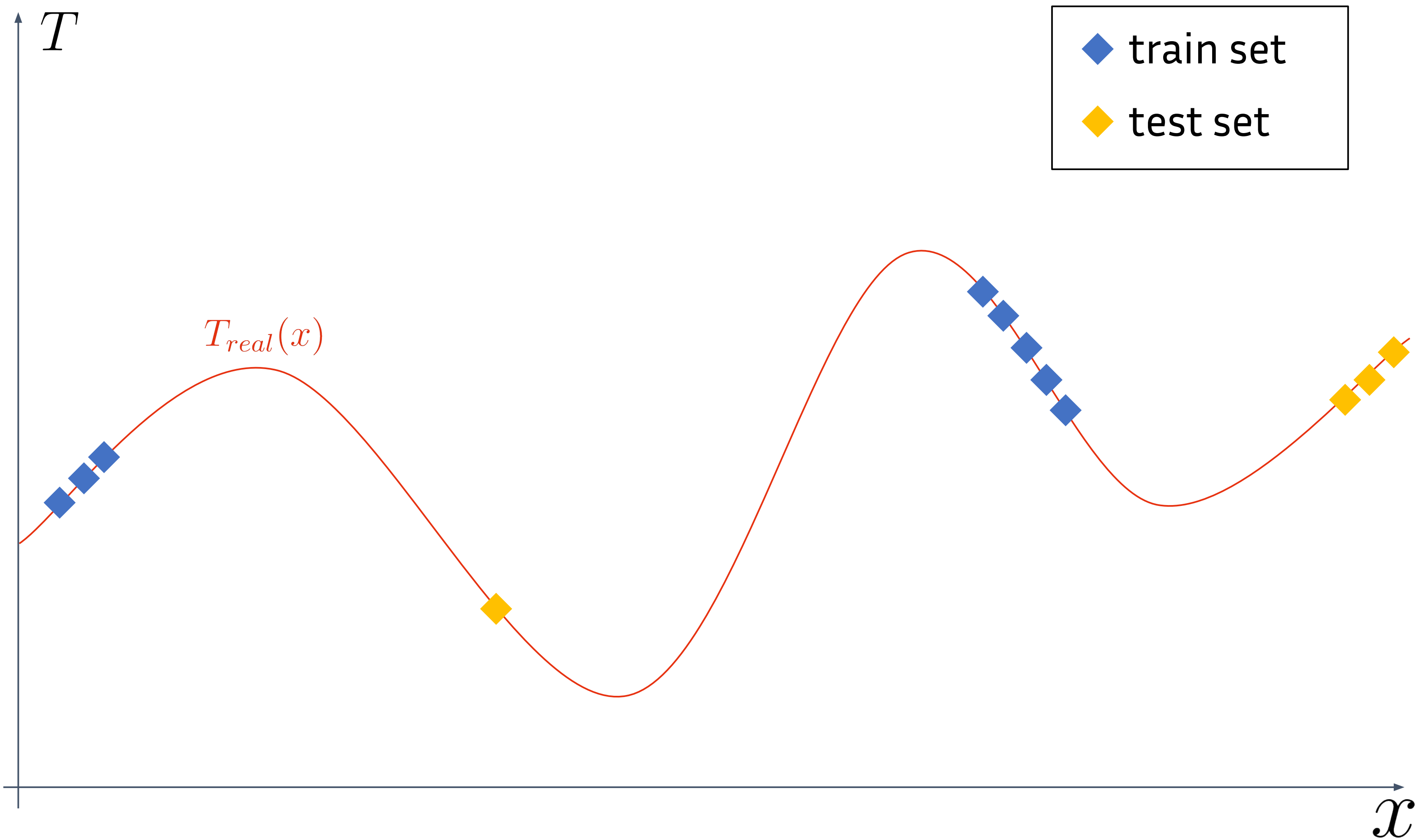
- **The density** of UVP sampling is **very irregular** in time and space

⇒ We need a splitting strategy **resilient to overfitting due to time and space proximity** of samples

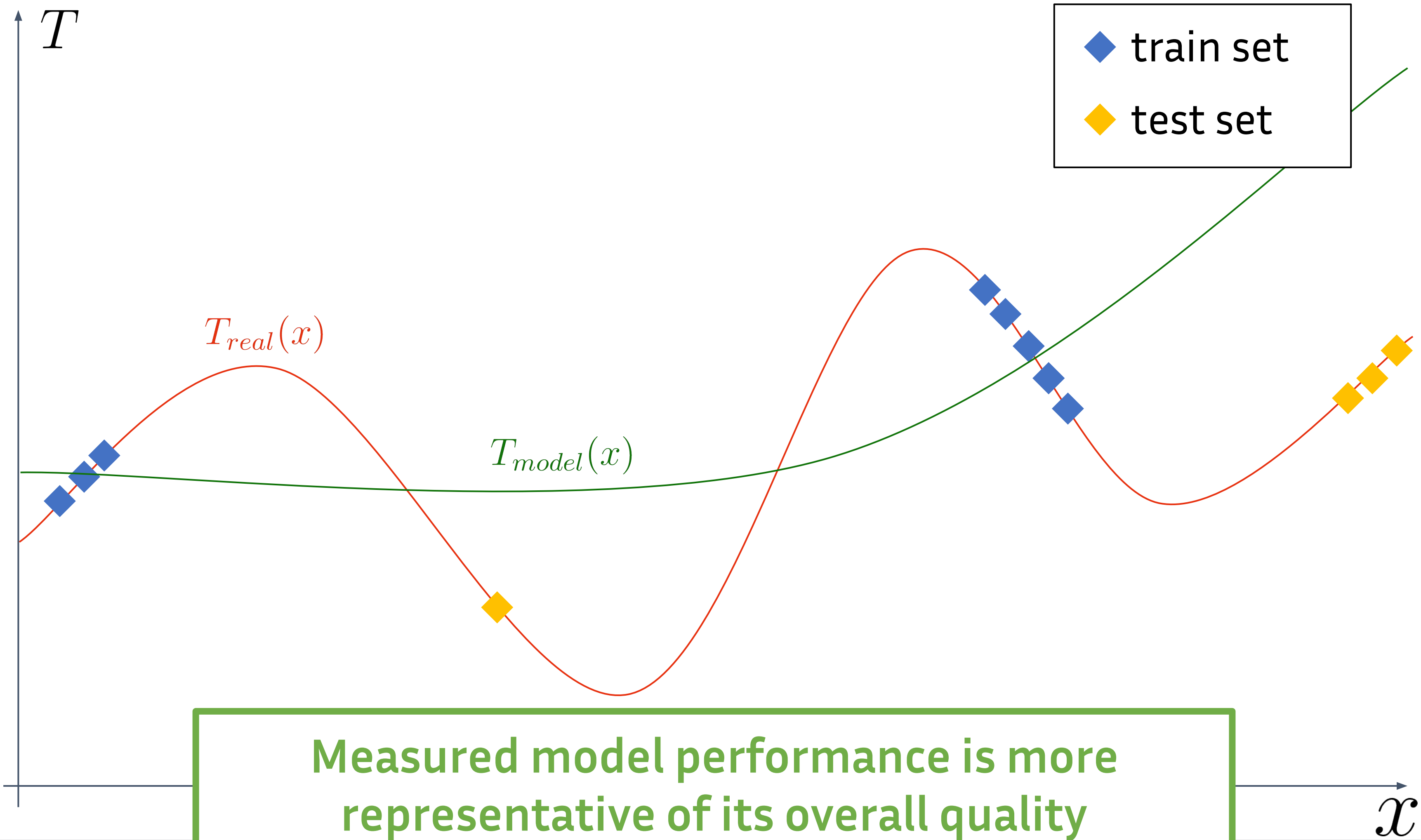
- A too conservative approach would make us **lose a lot of information** (seasonality, specificity of each biogeochemical region, local planktonic community, …)

⇒ We need a splitting strategy that keeps a **good temporal and regional representation** in both datasets

Train/Test splitting strategies specifications



Train/Test splitting strategies specifications

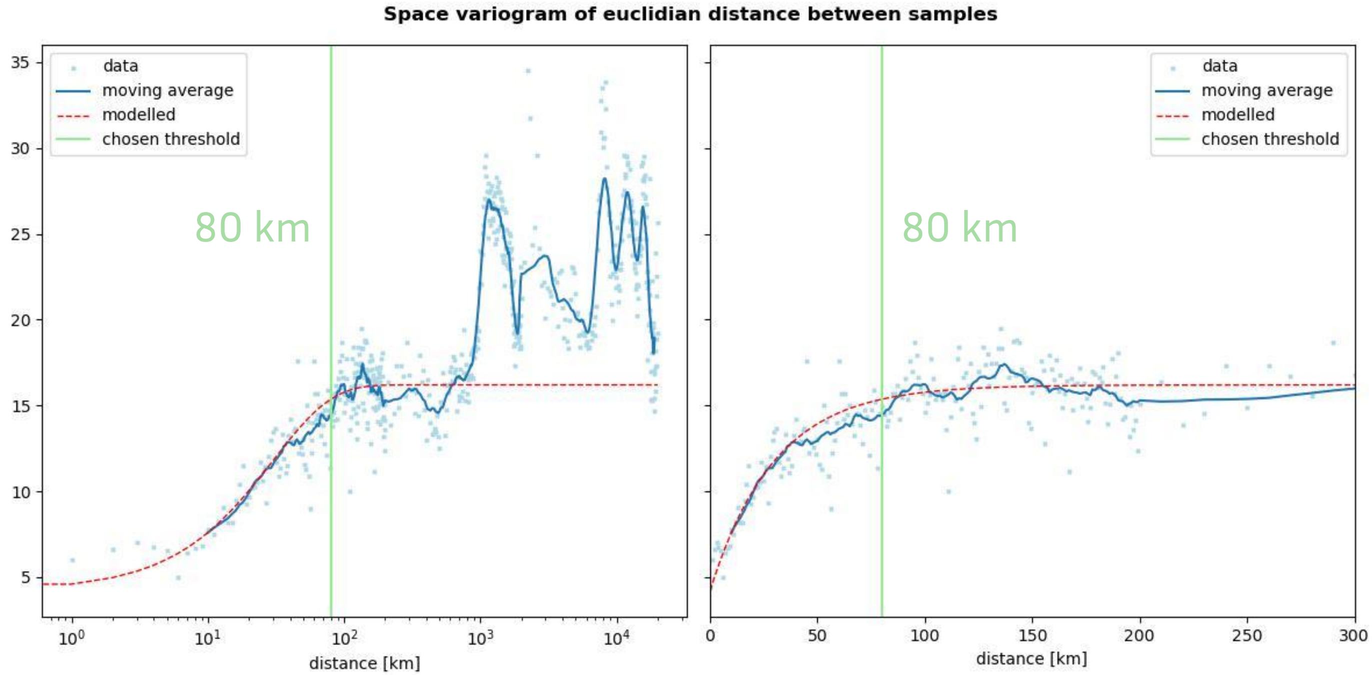


Measured model performance is more representative of its overall quality

Chosen splitting strategy

- We group samples together with the following rule:

$$\forall(x_1, x_2) \in \mathcal{X}, d_\tau(s_1, s_2) < \delta_\tau \wedge d_s(s_1, s_2) < \delta_s \Rightarrow g(s_1) = g(s_2)$$

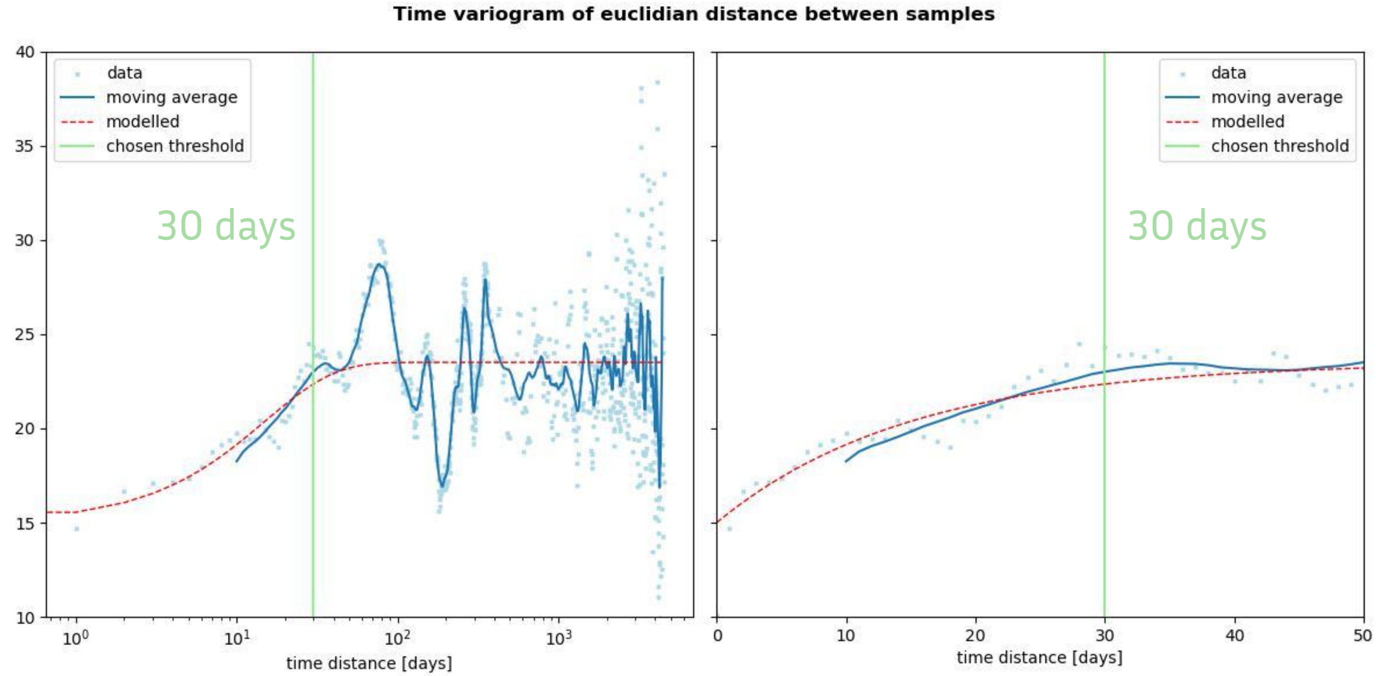


➢ Chosen threshold is at **>60% of global scale variance**, and **>95% of mesoscale variance** (< 300 km)

Chosen splitting strategy

- We group samples together with the following rule:

$$\forall (x_1, x_2) \in \mathcal{X}, d_\tau(s_1, s_2) < \delta_\tau \wedge d_s(s_1, s_2) < \delta_s \Rightarrow g(s_1) = g(s_2)$$



➢ Chosen threshold is at **>60% of global scale variance**, and **>95% of mesoscale variance** (< 300 km)

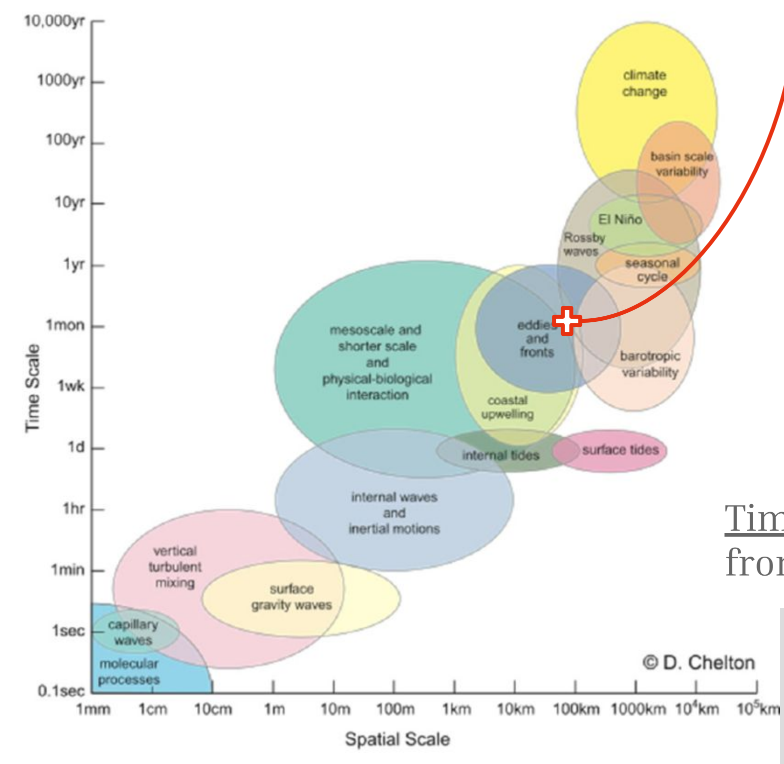
Chosen splitting strategy

- We group samples together with the following rule:

$$\forall(x_1, x_2) \in \mathcal{X}, d_\tau(s_1, s_2) < \delta_\tau \wedge d_s(s_1, s_2) < \delta_s \Rightarrow g(s_1) = g(s_2)$$

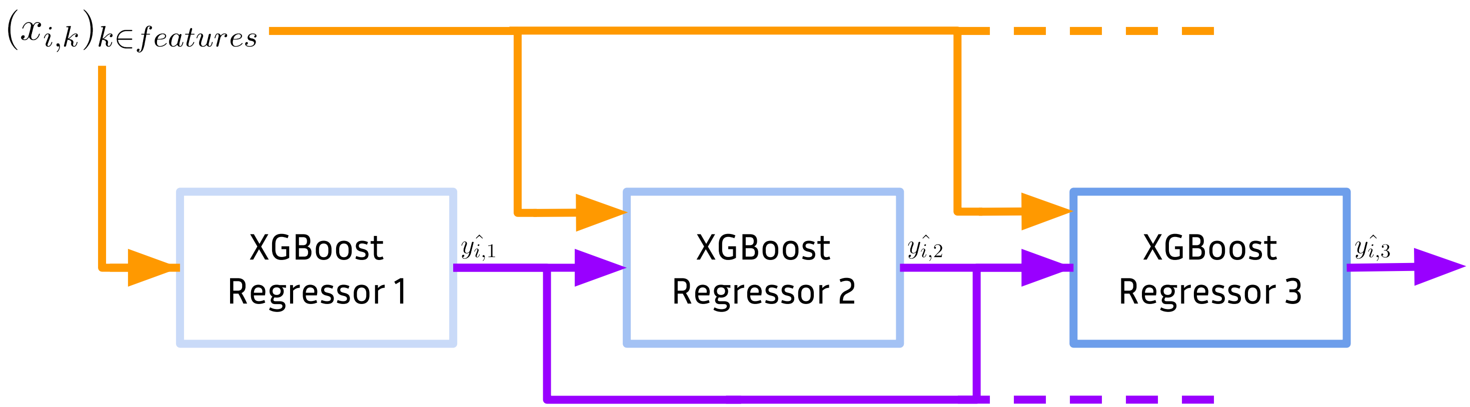
- We chose the following thresholds:

$$\boxed{\delta_\tau = 30\ days,\ \ \delta_s = 80\ km}$$

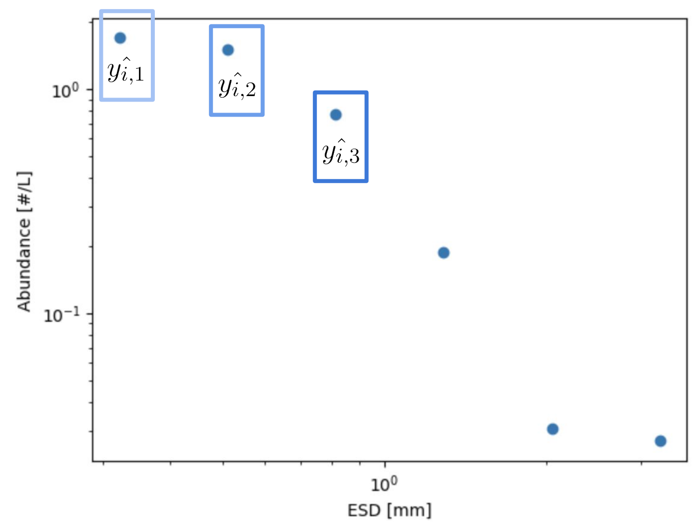


Time and space scales of ocean dynamics
from *Ocean Reference Stations* - Cronin et al. (2012)

"Full spectrum" model



- We build and fit **one model for each particle size class** (10 in total), beginning with the smallest particles
- Each model take as **input** the **selection of surface features and the predictions of previous** (and smaller) **size classes**



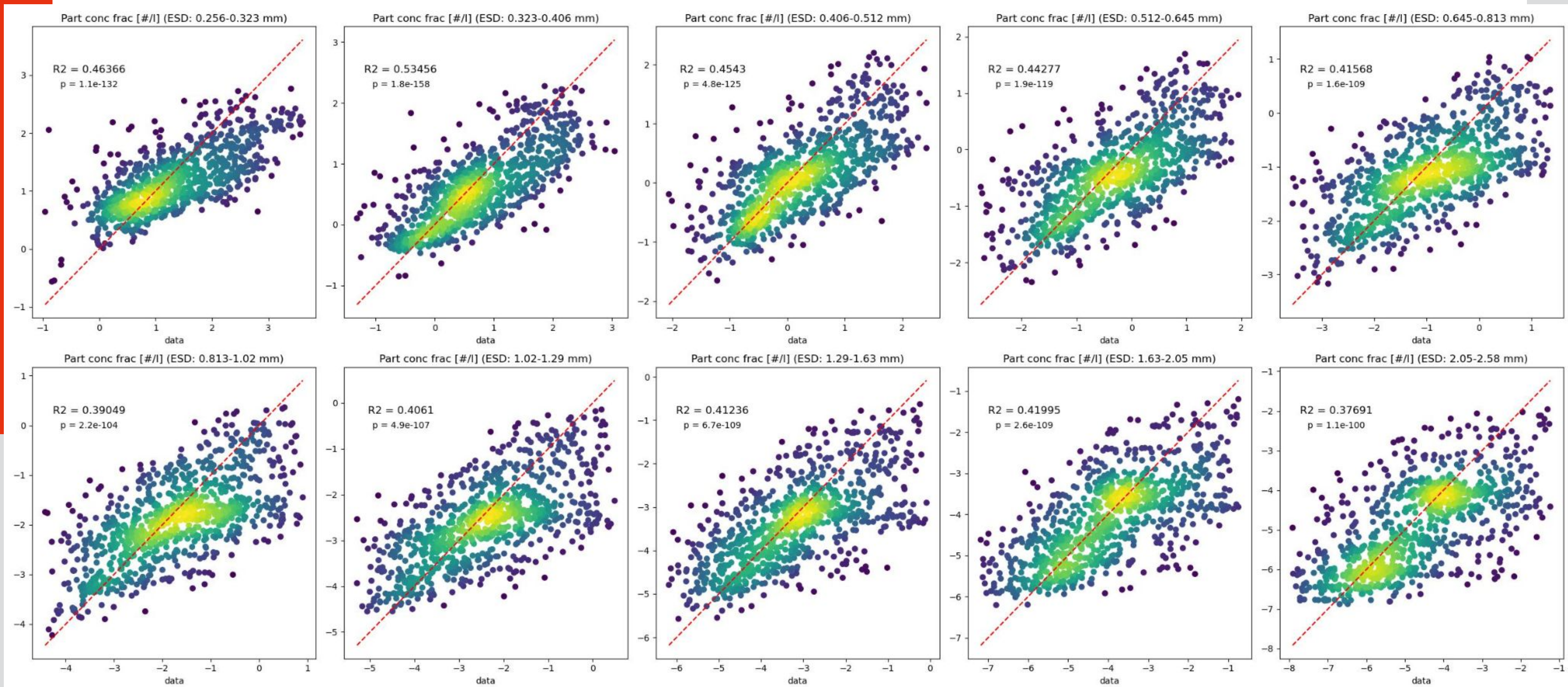
Chosen space of XGBoost hyperparameters

Parameter	Trade of	Chosen range
Trees max depths	Bias / Variance	[2, 7] (q-uniform)
learning rate	Speed / Bias	[7e-3, 4e-1] (log-uniform)
subsample	Bias / Variance & Speed	[0.5, 1] (uniform)
subfeatures	Bias / Variance & Speed	[0.6, 1] (uniform)
lambda (L2-reg)	Bias / Variance	[1, 1.5e2] (log-uniform)
alpha (L1-reg)	Bias / Variance	[5e-5, 1] (log-uniform)
gamma (min loss red.)	Bias / Variance	[0, 10] (uniform)
min child weight	Bias / Variance	[1, 20] (q-uniform)

- To find the best model architecture, I'm using Python *hyperopt* package, equipped with a bayesian algorithm

Results

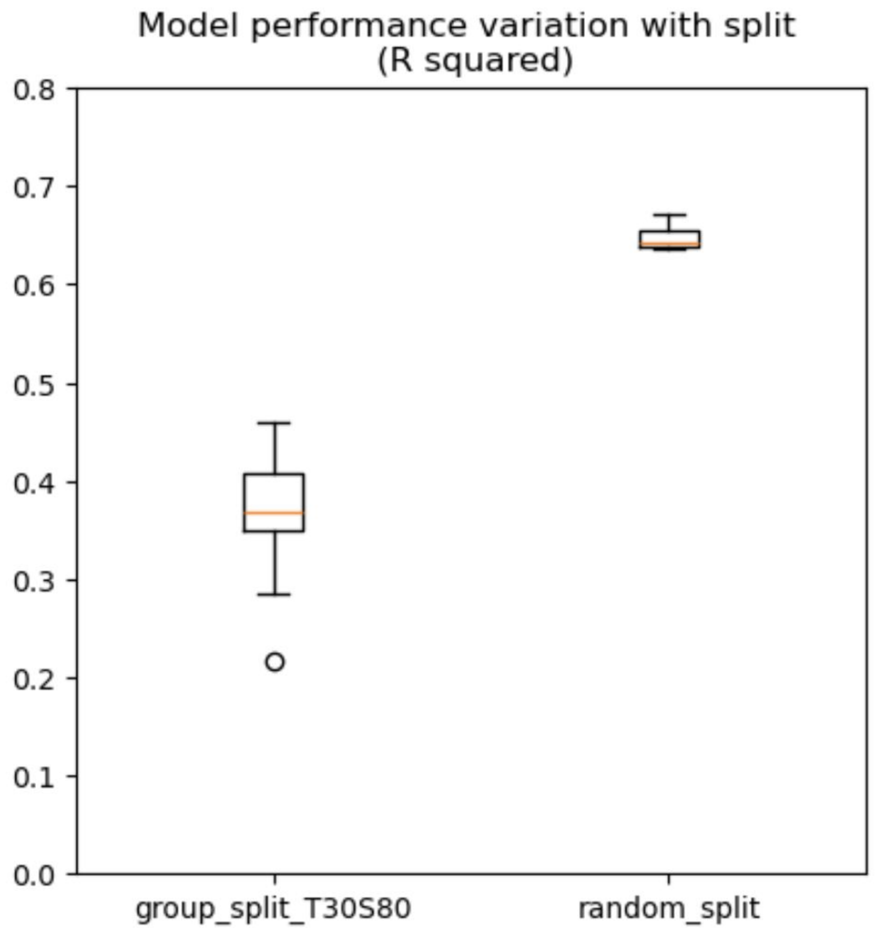
T14S100 models results examples



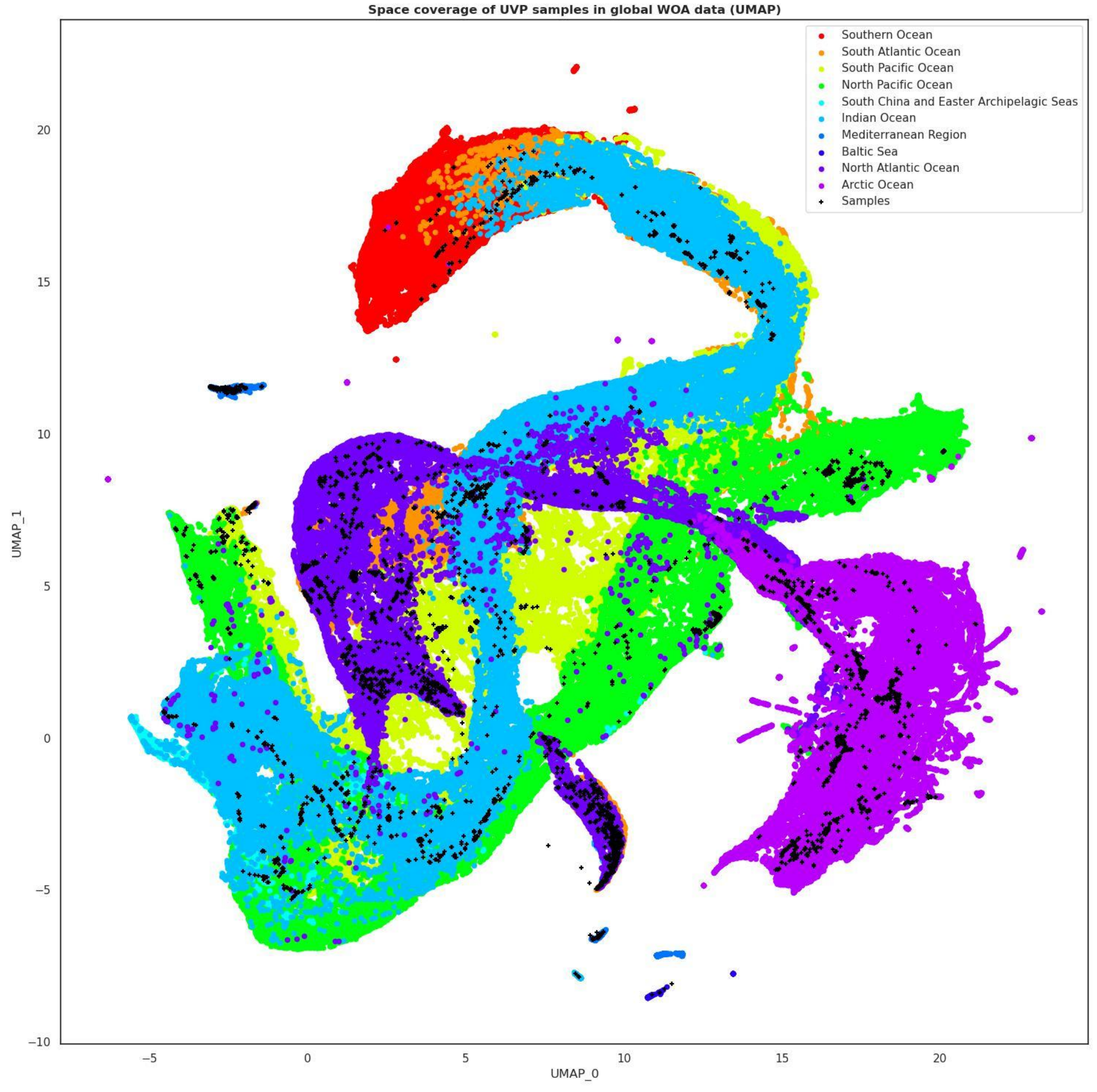
- We observe a slight **decrease of performance for bigger particles**, probably due to the fact that big particles are more dynamics dependent, and less directly correlated to phytoplankton abundance (i.e Chlorophyll concentration)

Performance differences with overfitting random split strategy

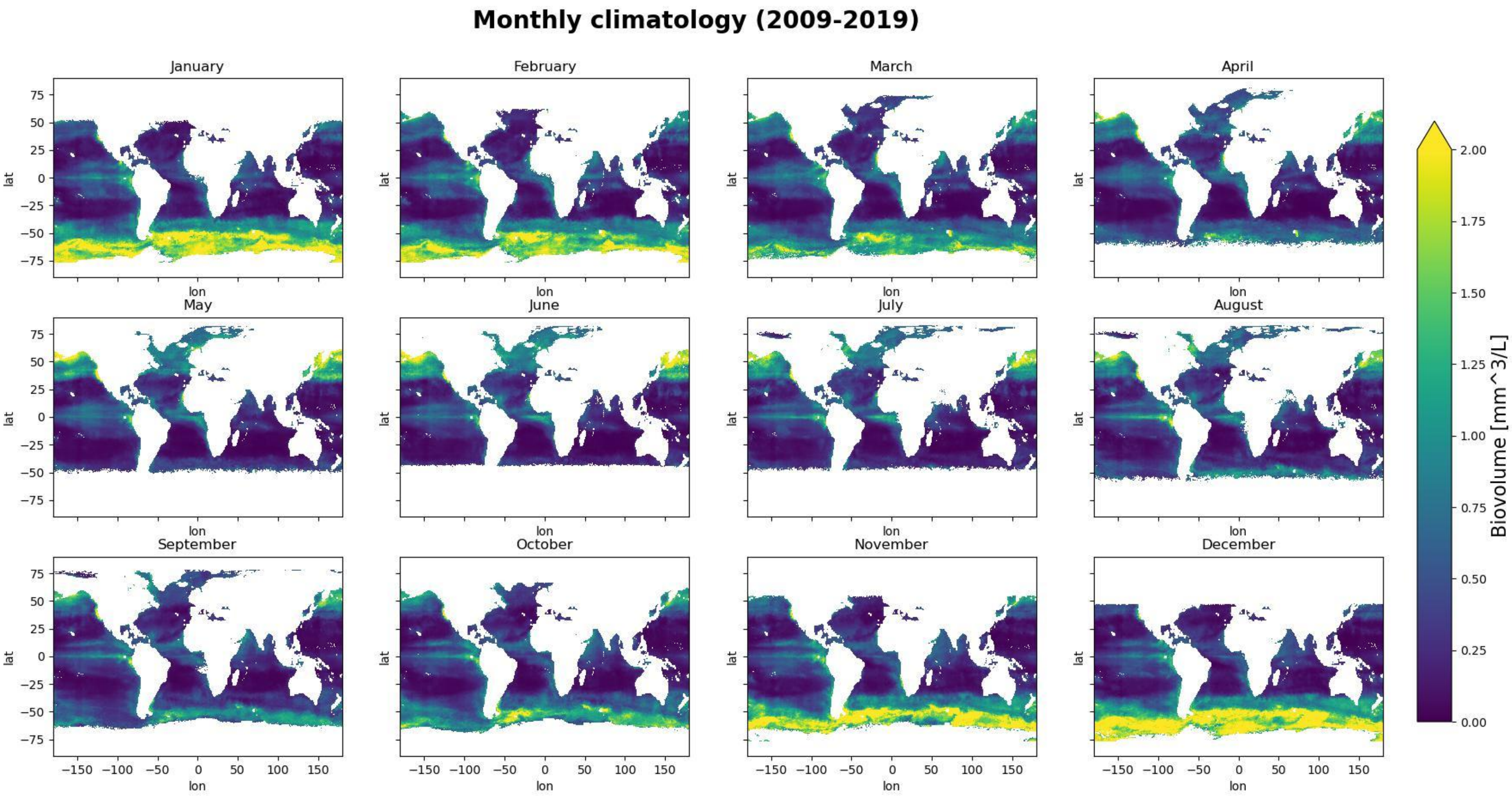
- Overall **displayed performance** of "vanilla" random split strategy is far greater than our conservative approach



Discussion - density of UVP sampling



- In this projection of UVP samples in a nonlinear reduced space of global WOA data, we observe that UVP samples have an **overall good coverage**, but with **wide variations in density**
- This projection, done with seasonal climatologies, doesn't take into account major multiyear phenomenon that can have a global impact on BGC dynamics, like El Nino for example


Monthly climatology (2009-2019)
January
February
March
April
May
June
July
August
September
October
November
December
lat
lon
75
50
25
0
−25
−50
−75
−150 −100 −50 0 50 100 150
Biovolume [mm^3/L]
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

Conclusion

- We propose an **innovative approach in ocean particulate organic matter modelling** to produce a fine scale global product of particles stocks

- Because of the structure of available particles data, we had to propose **an alternative strategy for test and train set split**

- Final results still have **room for improvement**

Perspectives, work to do

- This is still **work in progress**, a few details need to be refined:
  - Even if our train/test splitting strategy is more conservative than actual litterature, we want to improve it with a more formal approach to avoid any overfitting
  - We have yet to use our models to create extrapolated fields

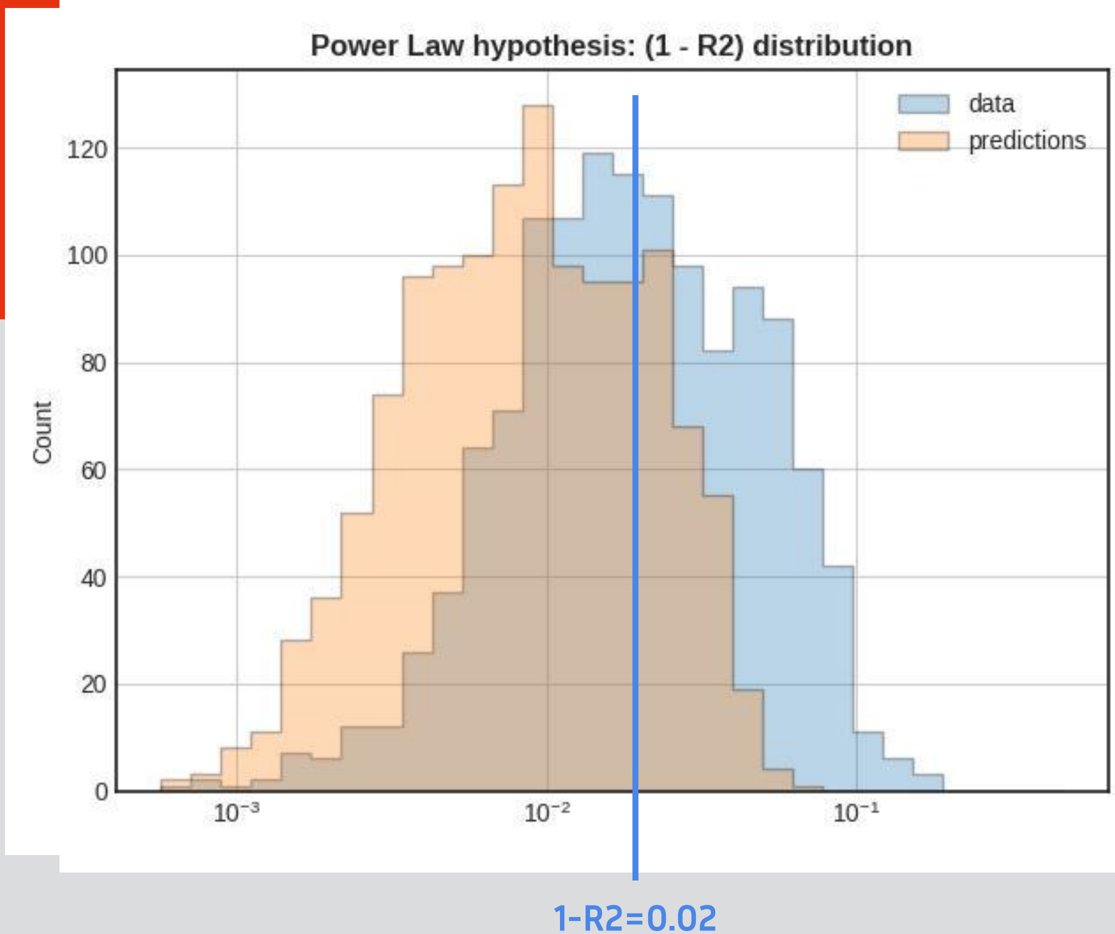
- This is still a zero dimensional model, which probably explain a large part of our bias: the particles are heavily dependent of the history of the water they evolve in.

  We want to take such **dynamics** into account. It could be done by creating a **hybrid lagrangian NPZD model**.

- The **particle distribution in deep water** is yet to be modeled. Deep Learning provide interesting tools like **LSTM neural networks** to simulate time or space series.

Thank you for your attention !

Impact of modelling on the power law hypothesis



hypothesis :
R^2 is lower than R^2 median in data (0.98)

		data	
		+	-
predictions	+	**197**	87
	-	446	**555**

Accuracy = 58%
Precision = 69%
Recall = 31%