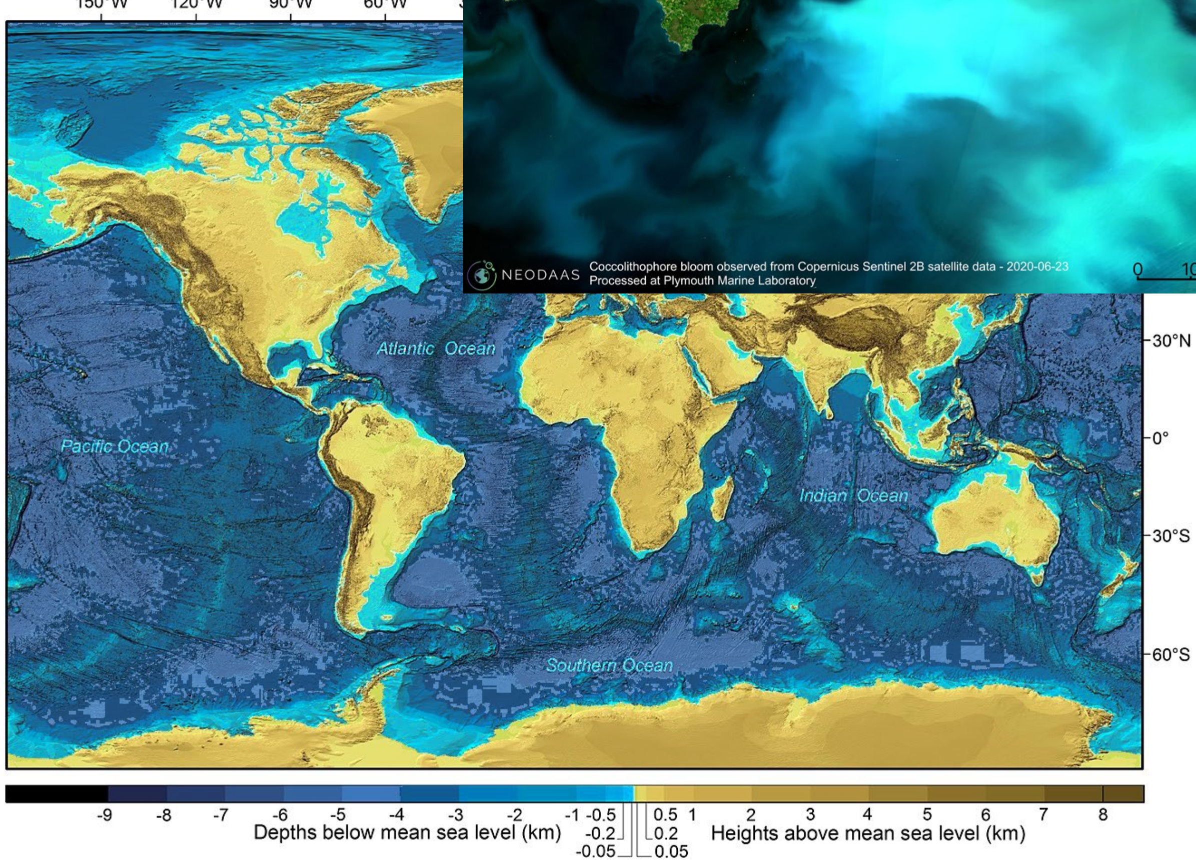
Complex networks and machine learning to improve marine ecosystem modelling and data assimilation

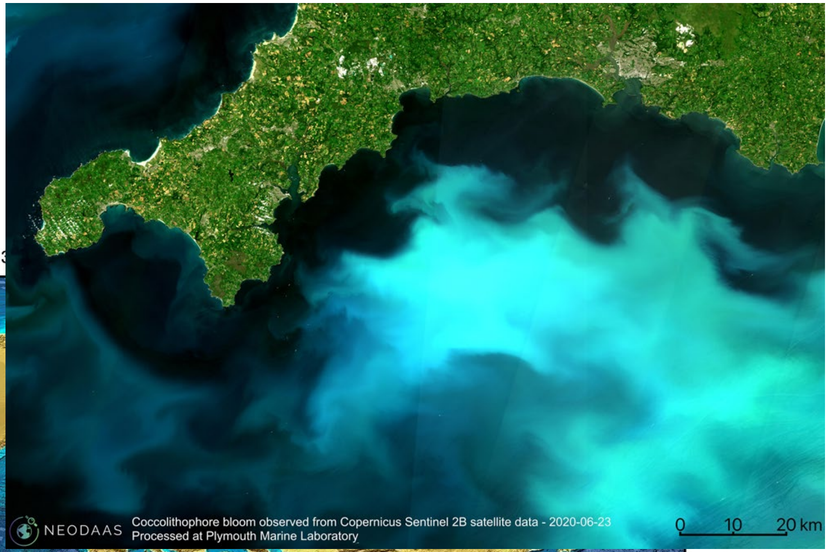
Ieuan Higgs,
Alberto Carrassi, Ross Bannister, Jozef Skakala, Stefano Ciavatta

Shelf seas and marine ecosystems

- Shallow ocean less than 200m depth

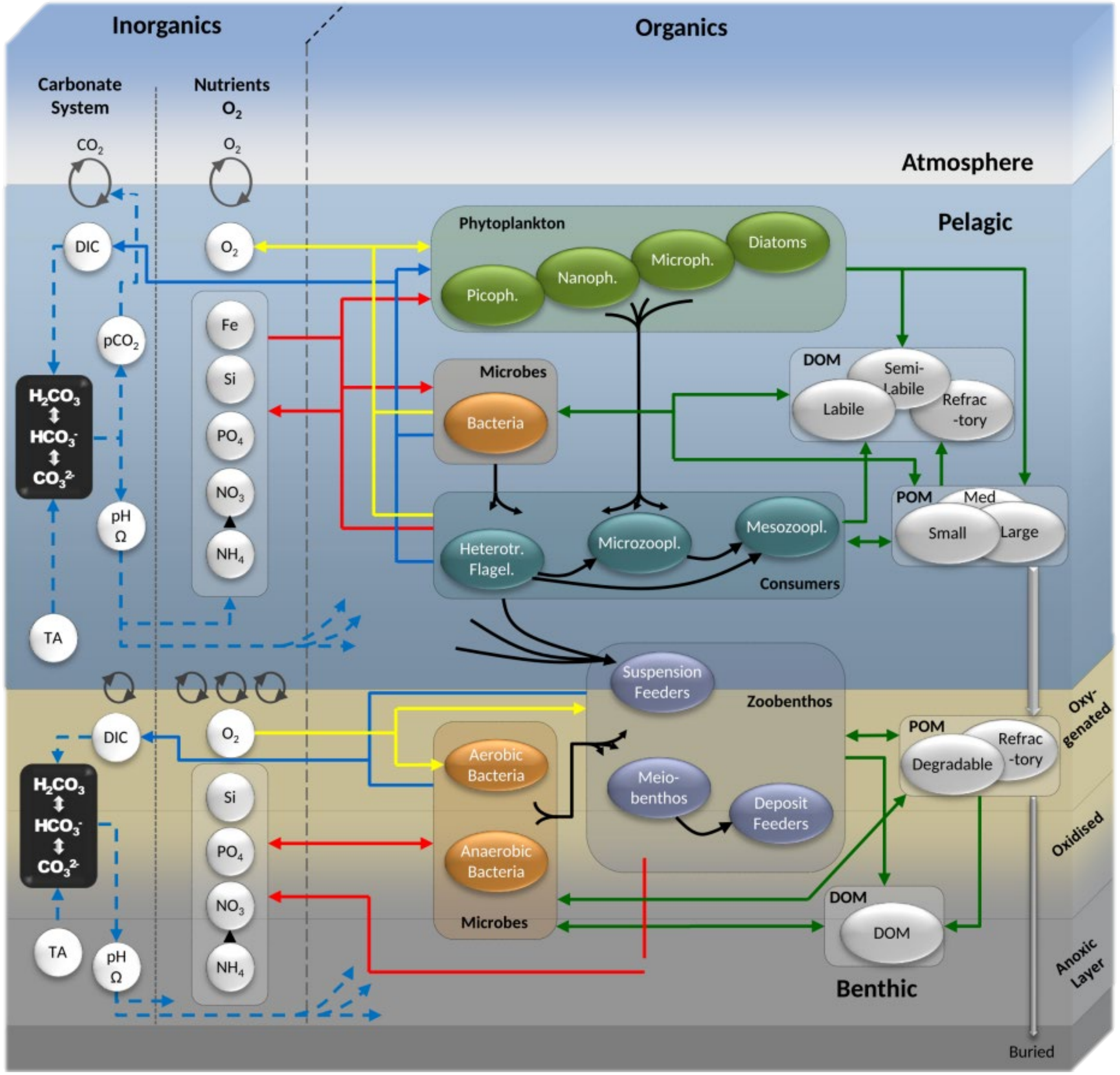
- 7% of global ocean are 'shelf sea'
  - 20% global biological productivity
  - 20% ocean uptake of atmospheric carbon
  - 80% global fish catches
- Need for more detail of key indicators and processes:
  - Refined forecasting capability
  - Improved computational efficiency
  - Better diagnostics and analyses
  - Usable information, relating to management - policy and science





Domain considerations

- Modelling this is a complex task
- Large system of interactions coupled to a physical model
- Huge uncertainties in model parameters & feedback mechanisms
- Observation Space <<<< Model Space (only observe total chlorophyll, derived from ocean colour at surface)



Research questions

Part 1:

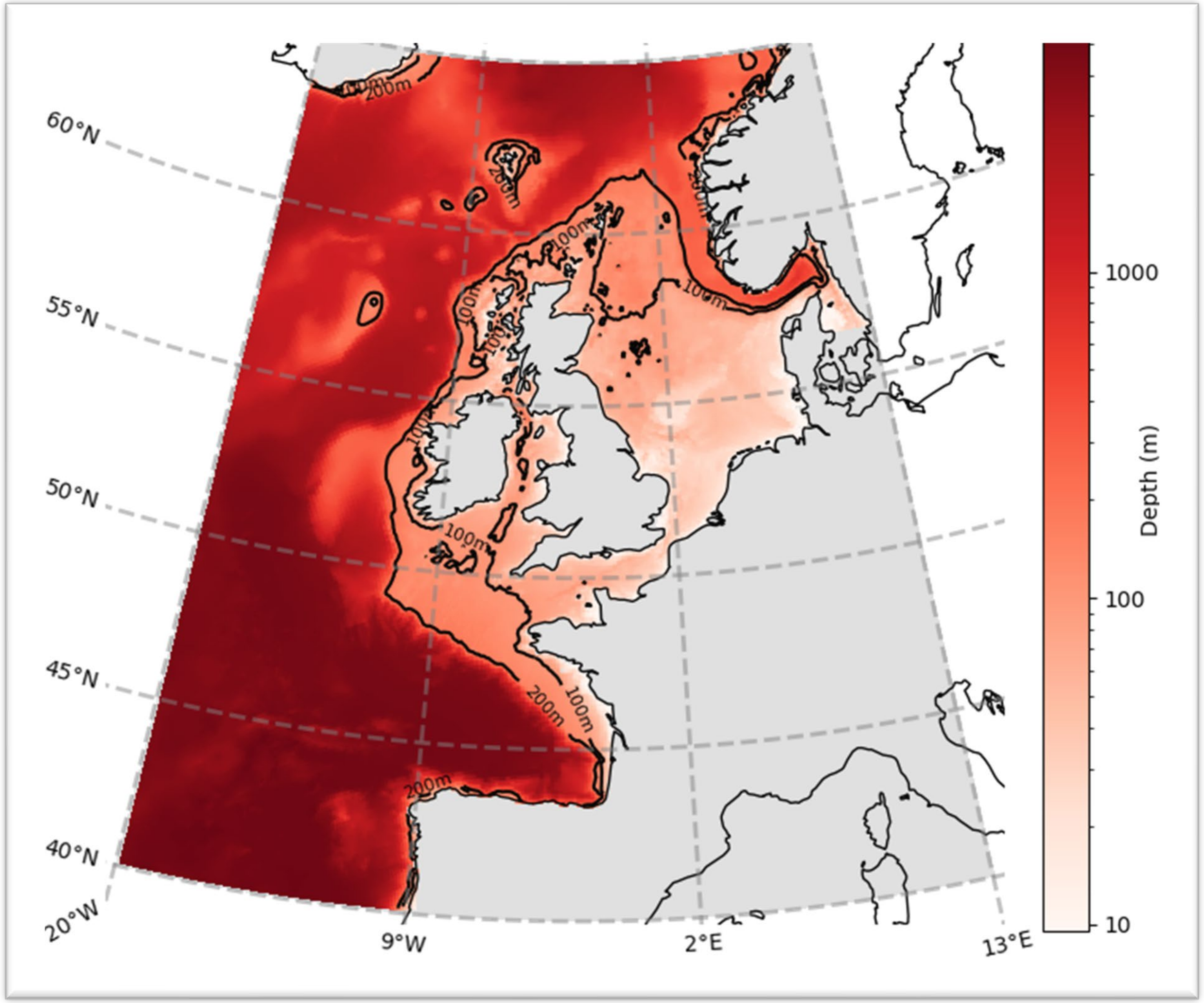
Can we identify potential simplifications for biogeochemical ocean models?

Part 2:

Can we leverage machine learning to improve data assimilation of marine ecosystems?

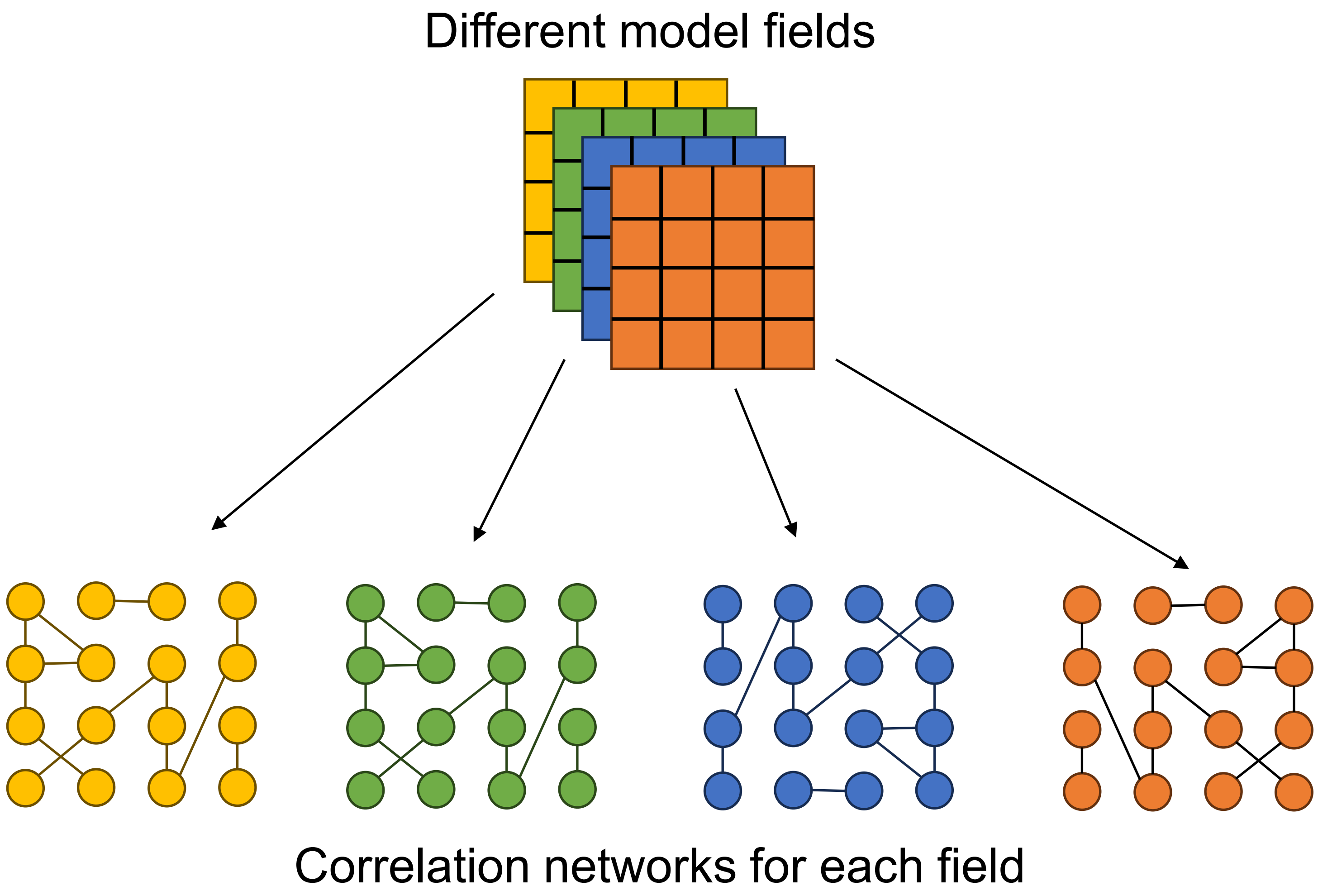
Model and domain

- Surface data from a 2016-18 free run NEMO-FABM-ERSEM
- We focus on the surface since satellite imagery can only observe the ocean's surface
- Pre-process the time series with a highpass filter
- Verified on a subset of variables from 2005-07 run
- We are interested in the lengthscales and interactions of modelled quantities



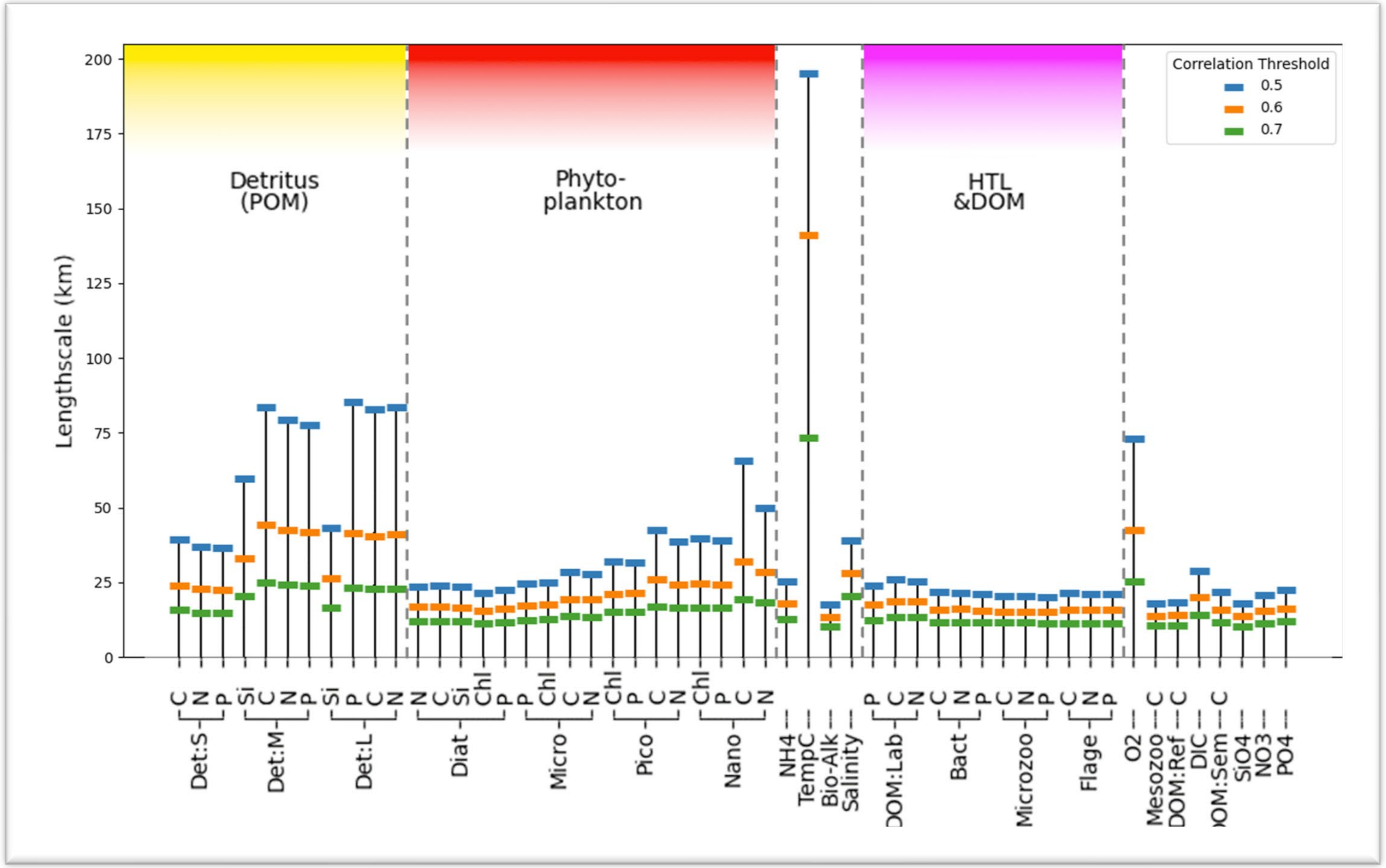
Network construction

- Network links are determined by correlation between any pair of grid points
- This is a flexible structure that can be used for numerous purposes and analyses
- For example, we can approximate length-scales…



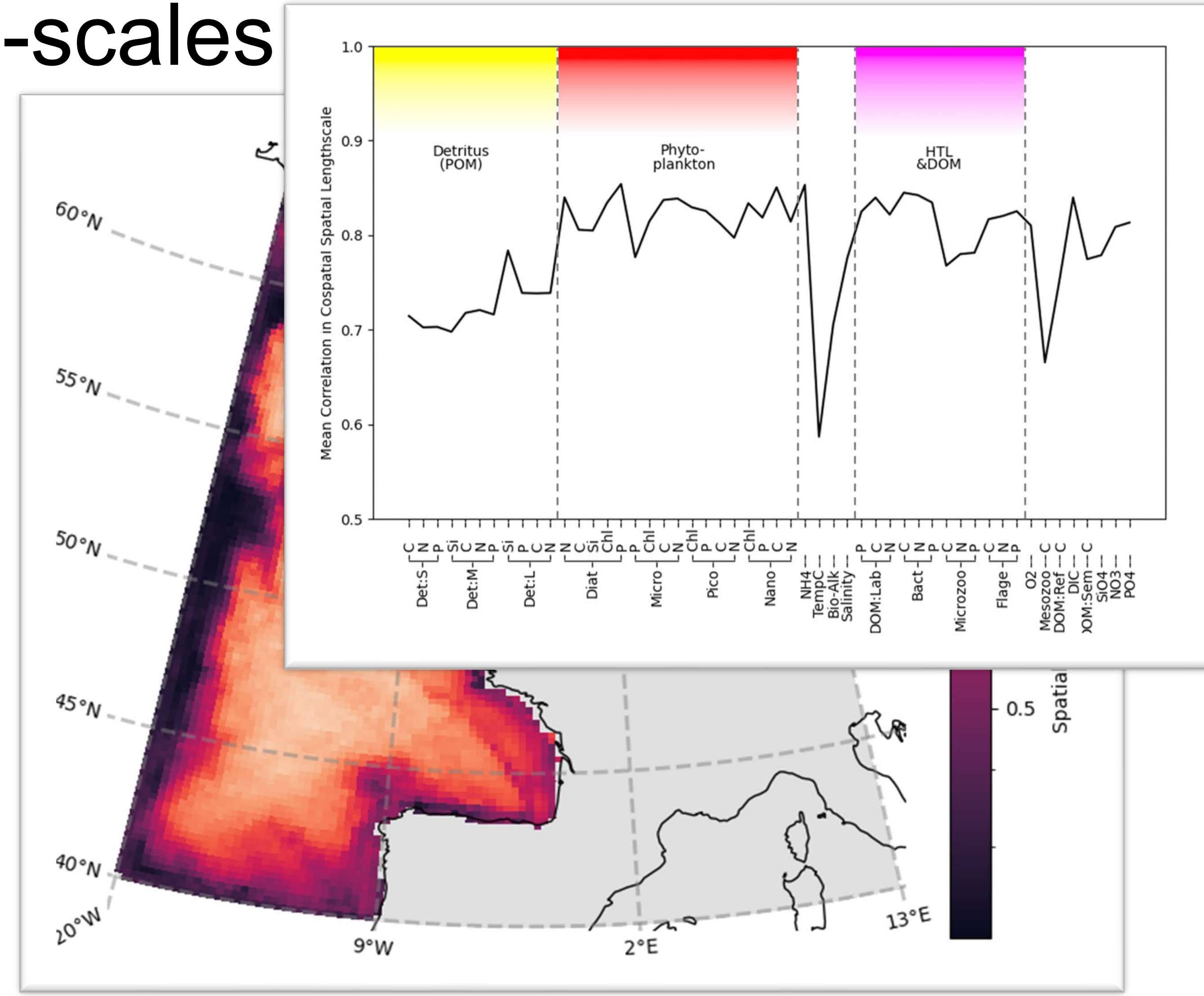
Mean length-scale correlations

- Mean length-scale of each variable
- At different thresholds (0.5, 0.6, 0.7)
- We see some cohesion in the length-scales within groups
- Length-scales are not directly transferable between all variables



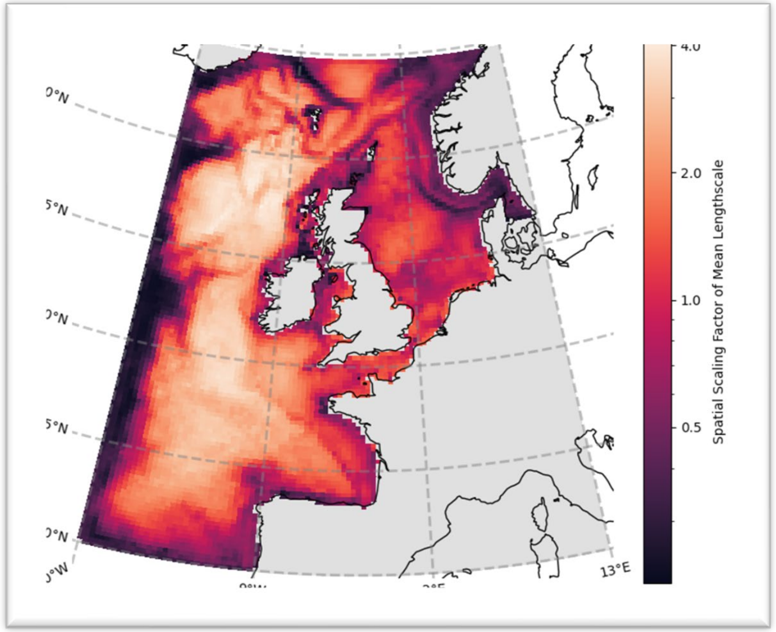
Spatial variation in length-scales

- Dynamically thresholded networks
- Mean length-scale at each point from these networks
- Distinct "cuts" are present in this length-scale map



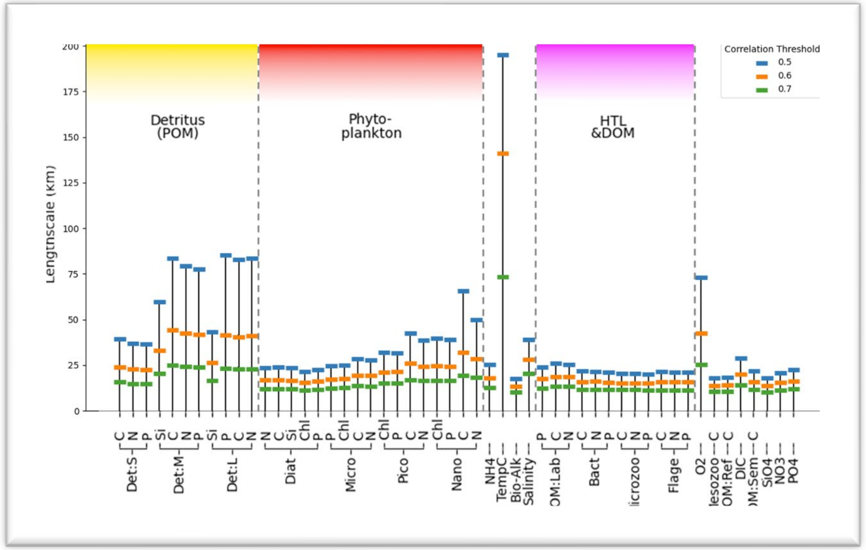
Provides a simple formulation for approximating length-scales

$\ell(x, v) =$



$f(x)$

$\times$



$\bar{\ell}(v)$

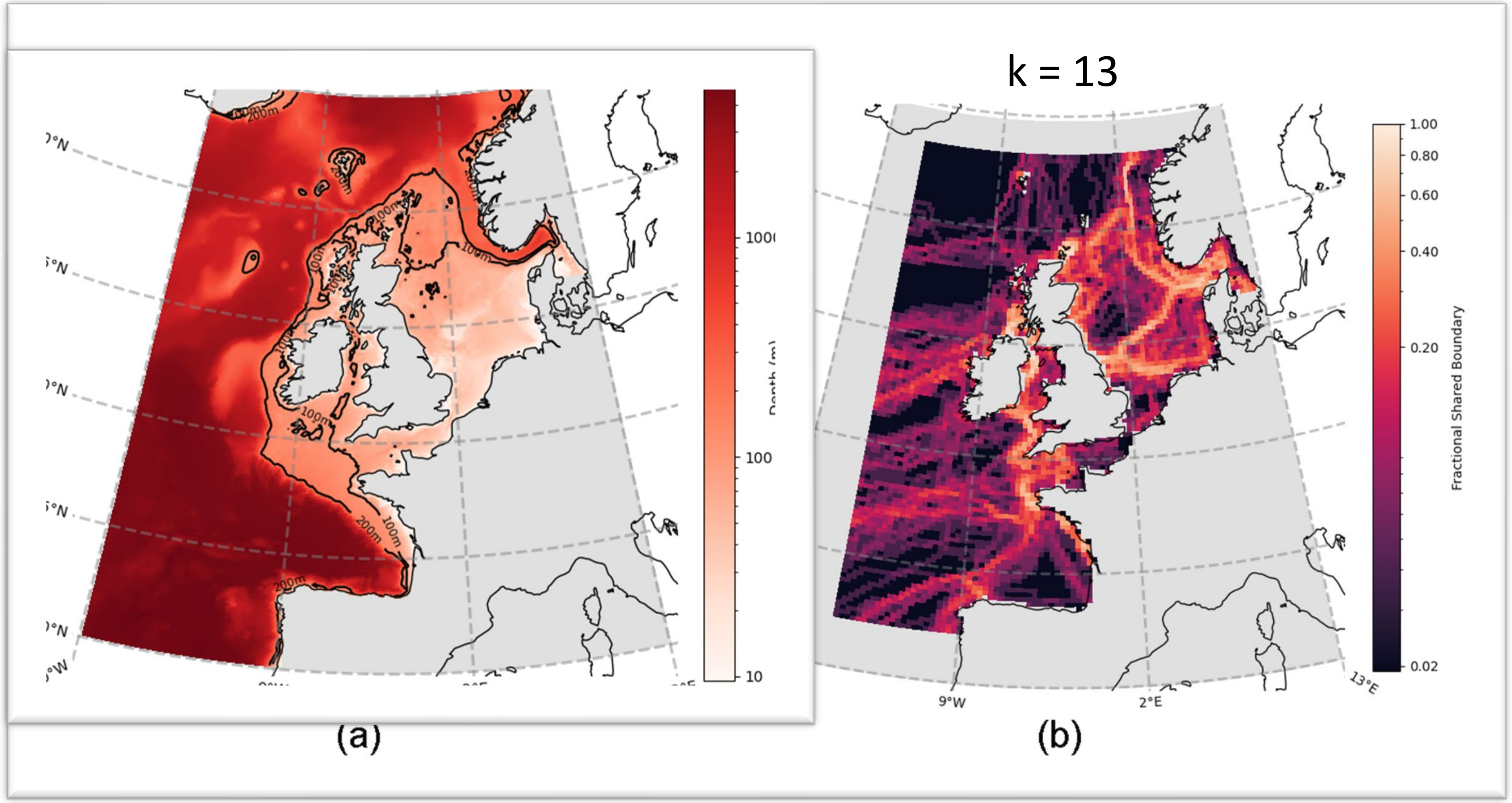
Spectral graph clustering

Aggregated boundary from the 50 ERSEM state variables.

“Brighter” boundaries are more consistent

We see a convergence of boundaries on the shelf (not guaranteed to happen)

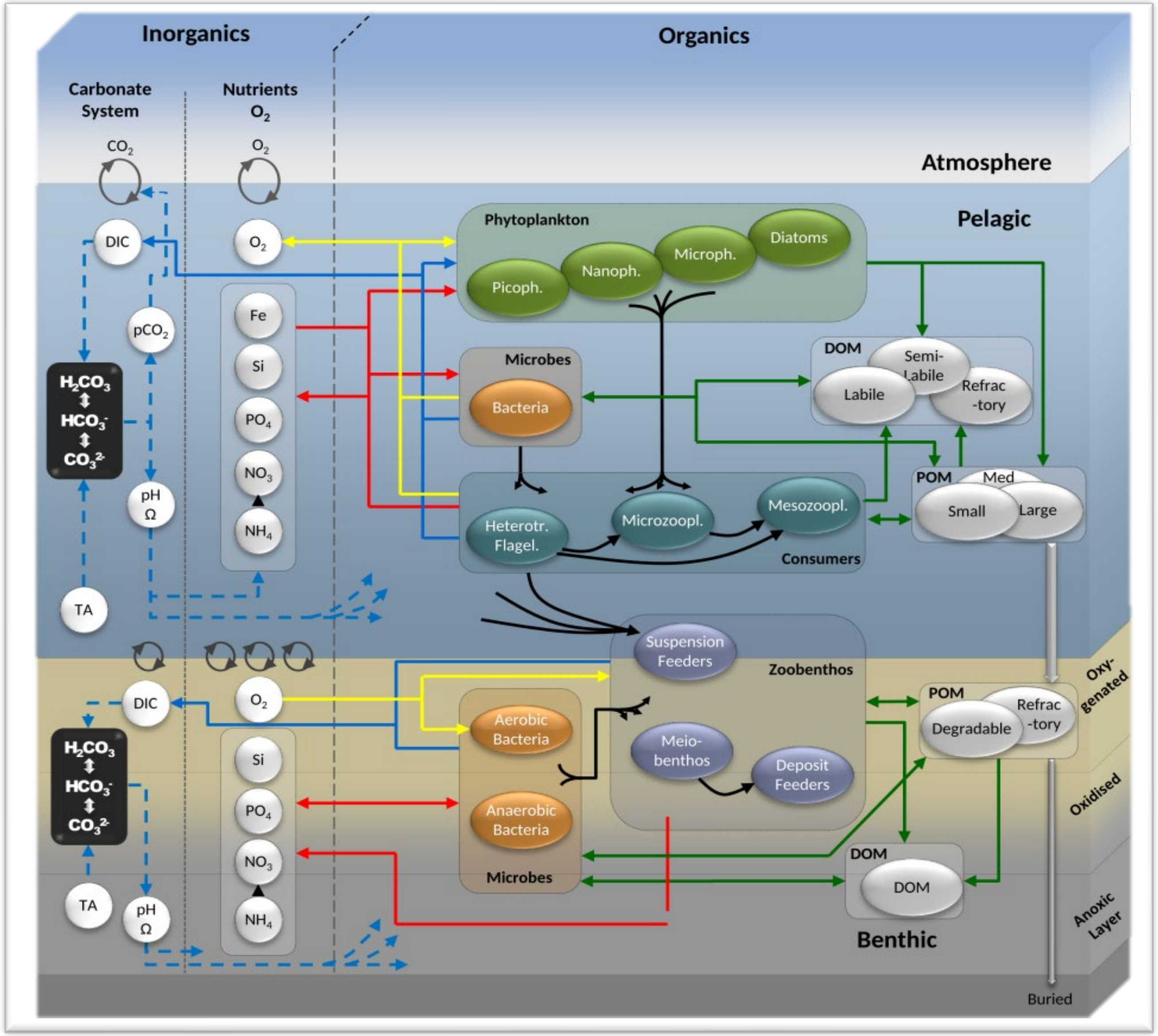
Shelf-sea to open ocean exchange



(a) (b)

Inter-variable network

- Intervariable correlations, with time lag considered
- Distinct clusters of variables form
- These relationships generally match well to the relational ERSEM schematic diagram



Research Questions

Part 1:

Can we identify potential simplifications for biogeochemical ocean models?

Part 2:

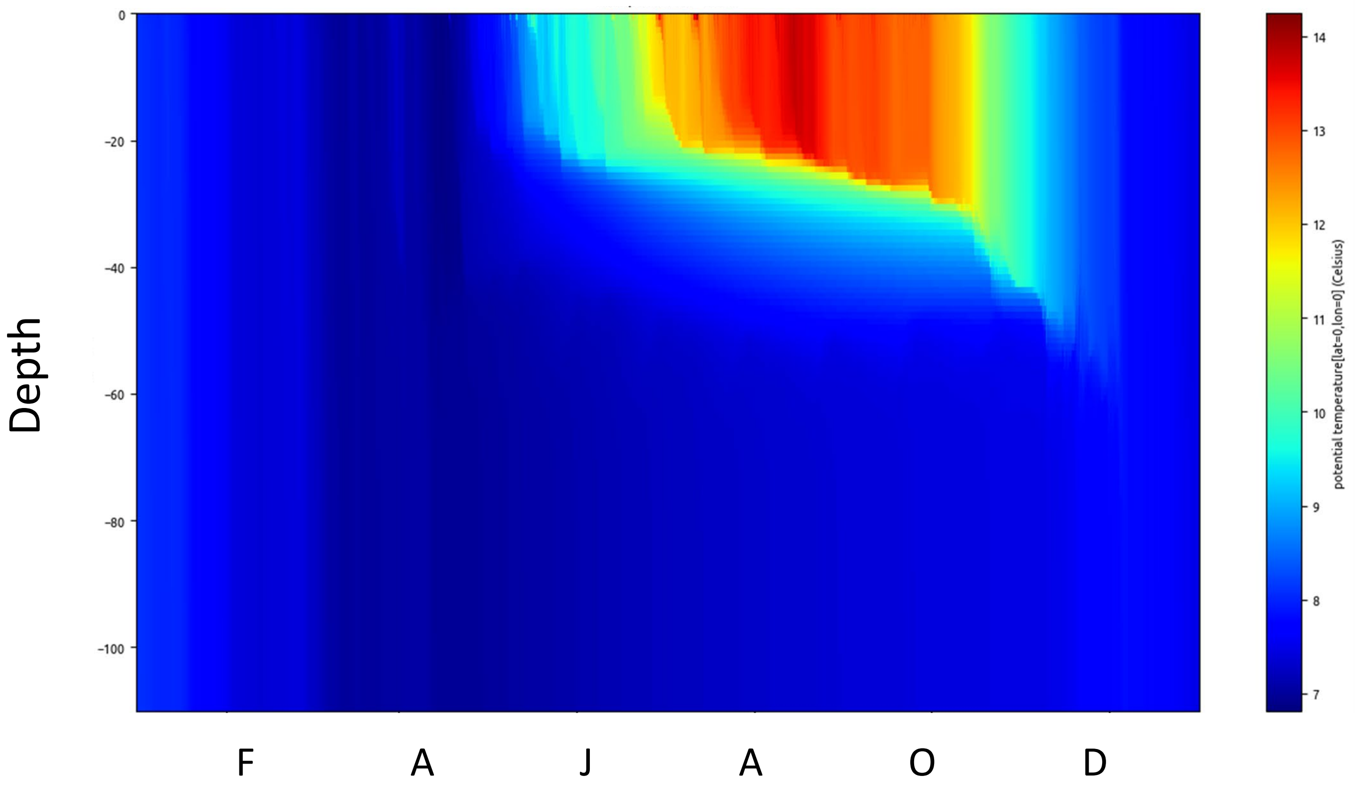
Can we leverage machine learning to improve data assimilation of marine ecosystems?

Model and Domain

- We work on a 1D vertical model
- Good first approx. of error relationships – should have some translational property
- This also affords us several luxuries
  - (i.e. much cheaper computational cost -> faster experimentation)

- Of course, we cannot expect every aspect of the 1D model to scale to 3D
- However, many vertical processes are likely translational

Example 1D Temperature Profile



Problem statement

- Recall – we only observe total chlorophyll on surface of the ocean

Observational space <<< Model space

- We need to to describe accurately the (potentially non-linear) relationship between the few observations and the rest of the unseen domain

- Currently, a univariate EnKF + a balancing scheme is adopted operationally (e.g. at MetOfficeUK)

- A very large EnKF will only partially solve the problem (still based on linear gaussian assumption)

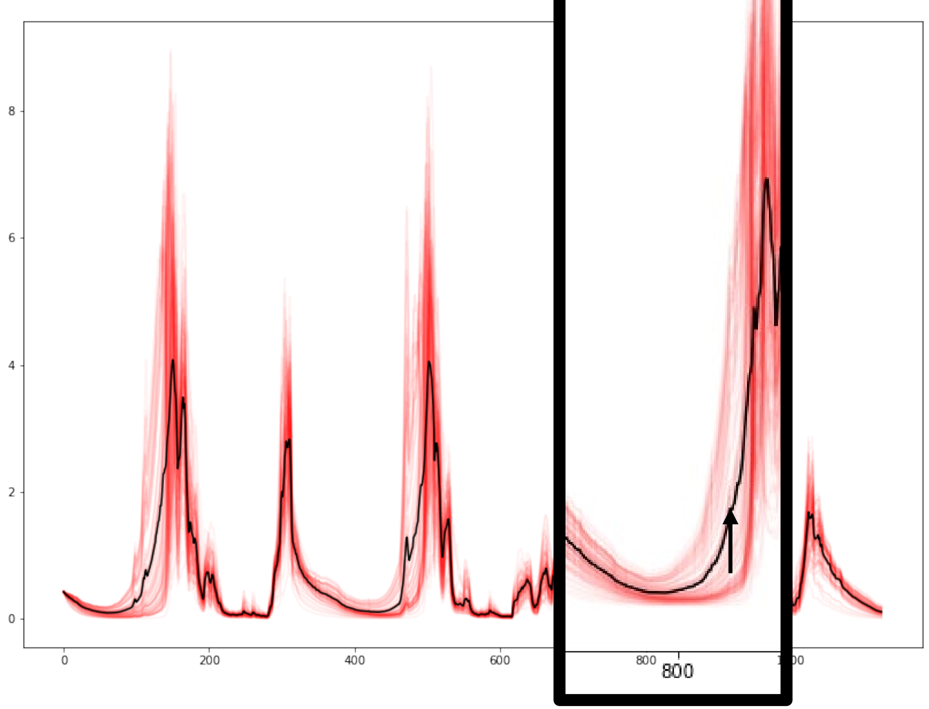
- In principle a particle filter would solve the problem (prevented by the curse of dimensionality)

- We aim at using machine learning to unveil the observed to unobserved relationship, and use it effectively to update the unobserved space

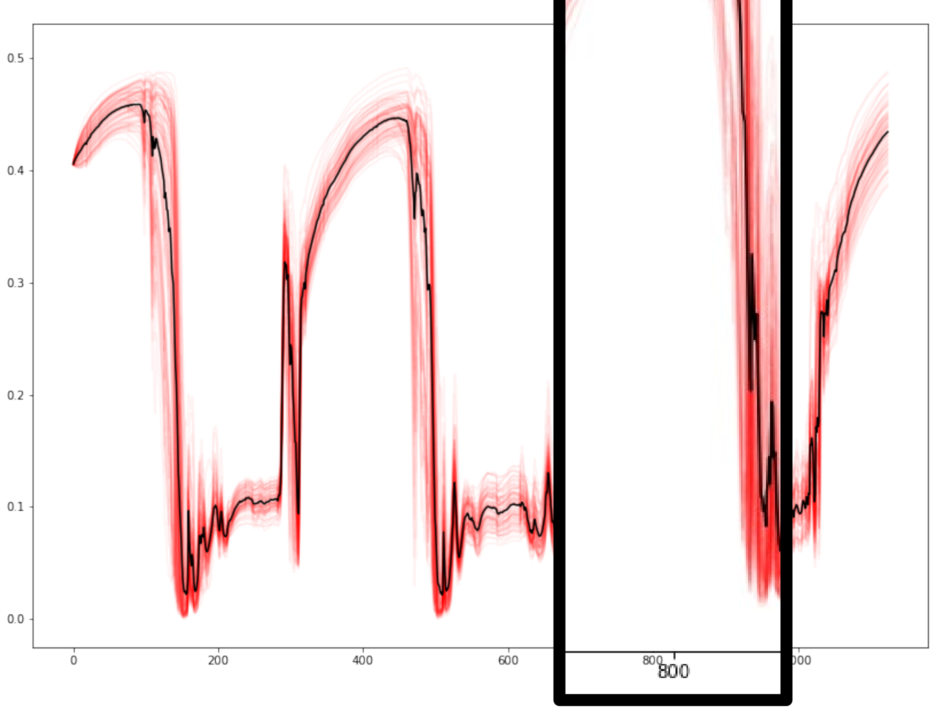
How can we create a training dataset?

- We generated a lo run, with each ensemble me aving some perturbed p ers (which are then fixed for ble run)

Total Chlorop



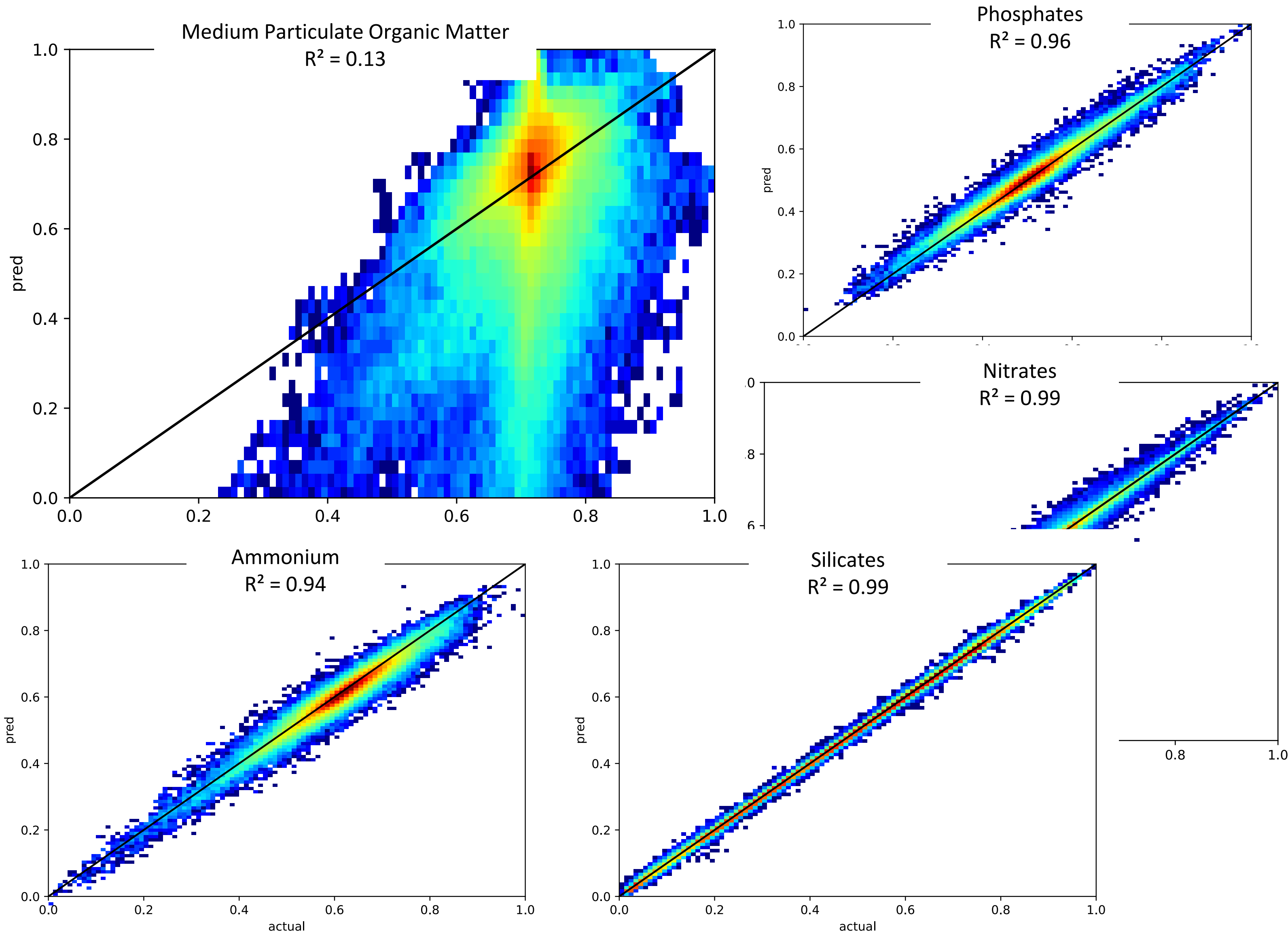
Phosphates (a n



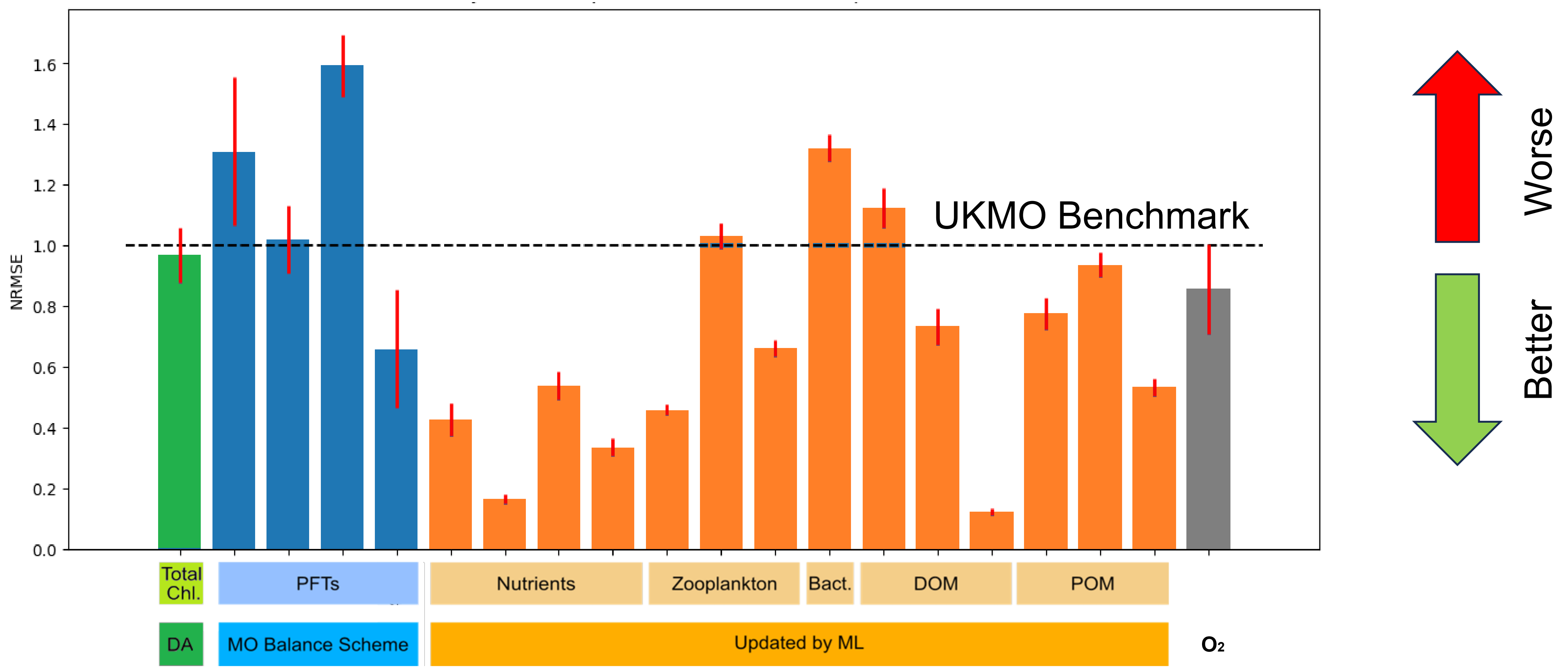
Training (offline)

- The NN learns the relationship well

  (for the most part)
- We prevent ML from updating variables with poor R² score



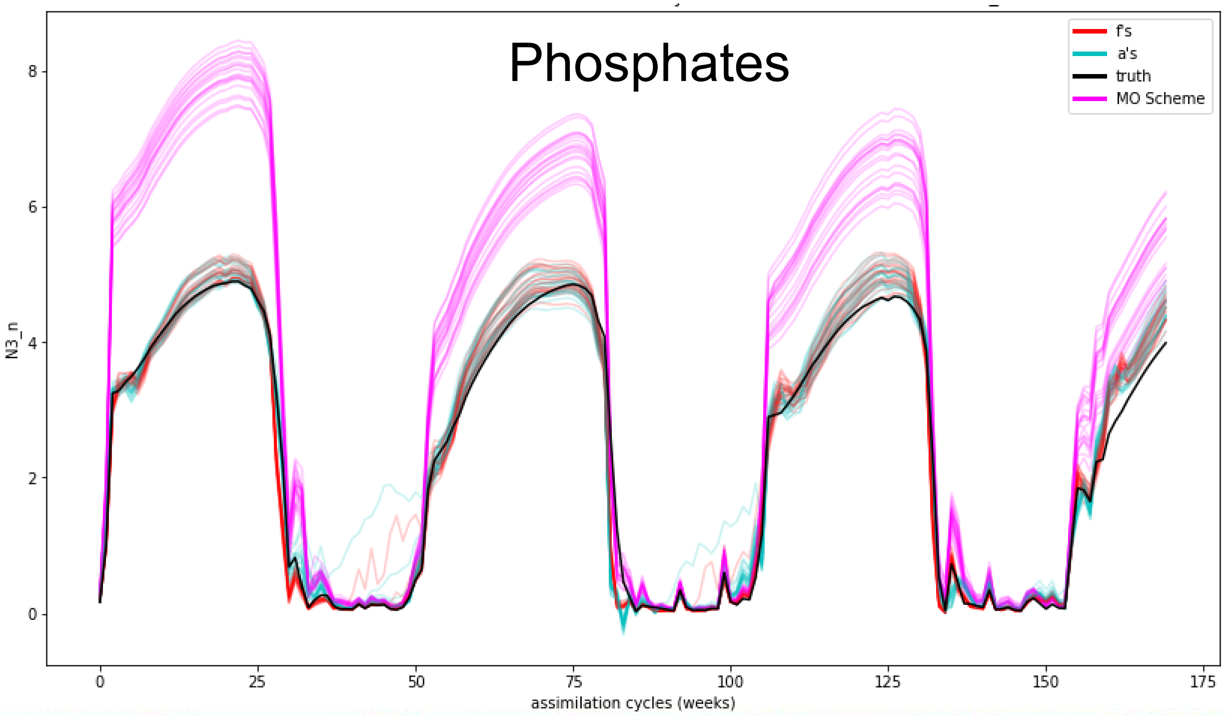
Relative Improvement over existing UKMO scheme

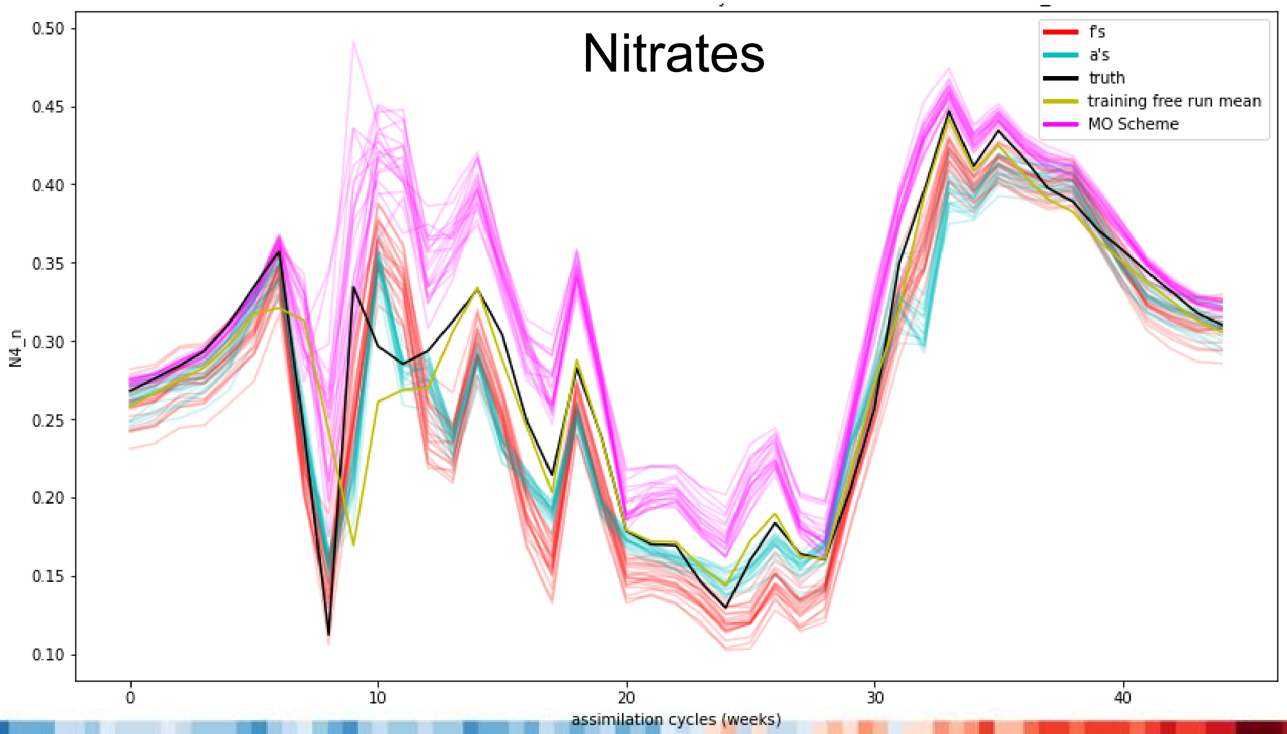


An encouraging result, but…

Closer inspection

- Instead of efficiently correcting the forecast error in the 1 week DA cycle, we instead correct a bias
- A direct consequence of our strategy for constructing the dataset





Our ongoing work and next steps

- Investigating the capability of ML to learn the correction done by a large EnKF (aka a perfect model scenario)

- From a practical aspect, we would like to see if the trained relation in 1D can then applied in the 3D system – translational invariance

- How can we make ML unveil observed to unobserved relationship that are not represented by the Kalman filter?

Conclusions

- Successfully interrogated the structure and interactions at the surface of the 3D model, with simplifications that can be applied across many areas (e.g. DA, ML, modelling…etc)

- Promising results for having ML emulating a fully non-linear multivariate DA

- Thank you for listening, any questions?