PML | Plymouth Marine Laboratory

Listen to the ocean

The potential benefit of ML/AI for biogeochemical modelling and data assimilation

Jozef Skakala (PML, NCEO)

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Overview:

- introduction to what are our models and open issues related to them
- how these issues are addressed using ML/AI in the PML modelling group
- other ML/AI work in the PML modelling group

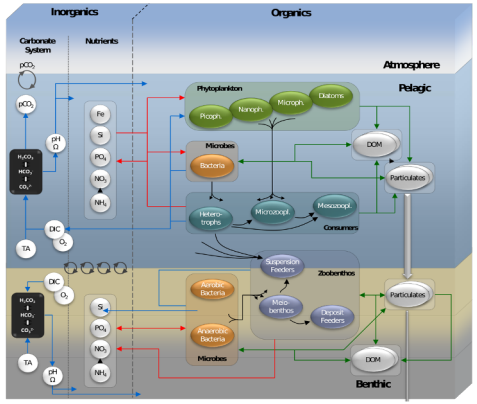
What precisely are our marine models?

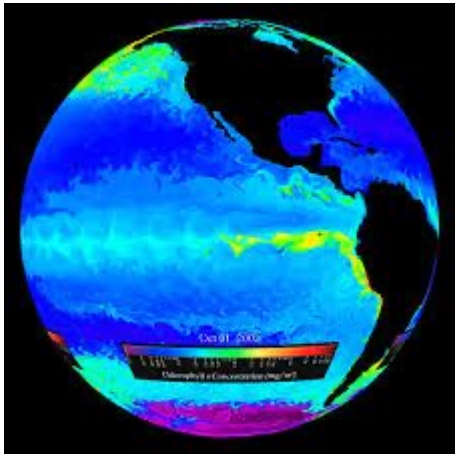
Fundamental principles

$$\begin{cases} \rho\dfrac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \nabla\cdot\boldsymbol{\sigma}(\boldsymbol{u},p) = \boldsymbol{f} \\ \nabla\cdot\boldsymbol{u} = 0 \\ \boldsymbol{u} = \boldsymbol{g} \\ \boldsymbol{\sigma}(\boldsymbol{u},p)\hat{\boldsymbol{n}} = \boldsymbol{h} \\ \boldsymbol{u}(0) = \boldsymbol{u}_0 \end{cases}$$

- physics
- chemistry
- biology

Complex model forced by data running on the supercomputer





Marine models enable us to:

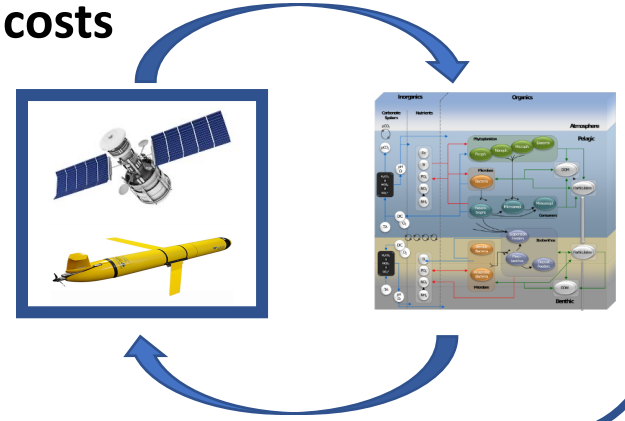
- better understand the present and past states of the ocean (data are limited and sparse!)
- forecast ocean state into the future
- understand how certain processes and drivers impact the ocean state and use that in what-if scenarios

Number of potential issues associated with marine models:

- initialization (chaos etc …)
- cannot represent all the relevant processes, world is too complex
- has to use spatial resolution which is often quite coarse: sub-grid processes have to be included ``indirectly''
- uses a number of parameters that are often uncertain
- relies on external inputs (atmospheric forcing, initial data) that are uncertain

What can we do?

Initialization: One way to constrain our models with observations (re-initialize) is data assimilation, but this is often limited and depends on several assumptions to reduce computational costs



Otherwise one can try to improve the model resolution, add more processes, or at least try to handle model uncertainties using statistical Monte Carlo (ensemble) methods…

but computationally too costly !!

One important idea how to reduce cost: replace for specific purposes model with its simpler but much faster version – an emulator

Emulators can be constructed by reducing the number of model variables and processes (reduced complexity model), or by reducing the model spatial domain (e.g. we frequently use at PML 1D versions of our model).

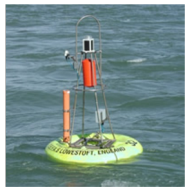
A huge help can be to use statistical emulators based on ML/DL!

A fairly new line of research at the PML modelling group:

- a 6 month PML RP (ongoing)
- 1 PhD studentship (ongoing) and another PhD studentship advertised
- ideas put into a new H2020 proposal (1 full WP), with similar ideas already used in 1 previous unsuccessful DTO proposal

Predict unobserved variables from observations

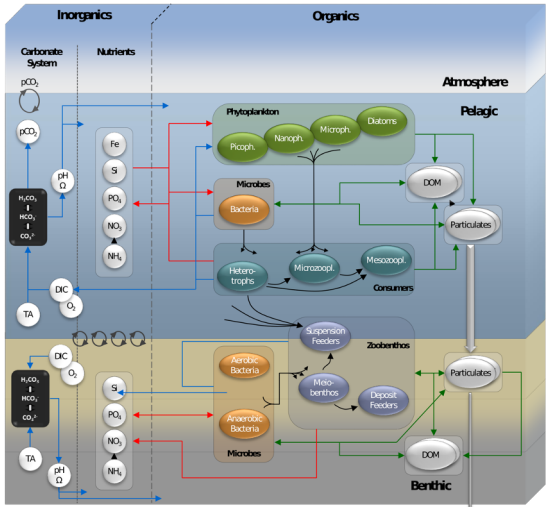
Inputs



observations for SST, chlorophyll …

atmospheric, riverine data

Model



Can be emulated with machine learning

Outputs

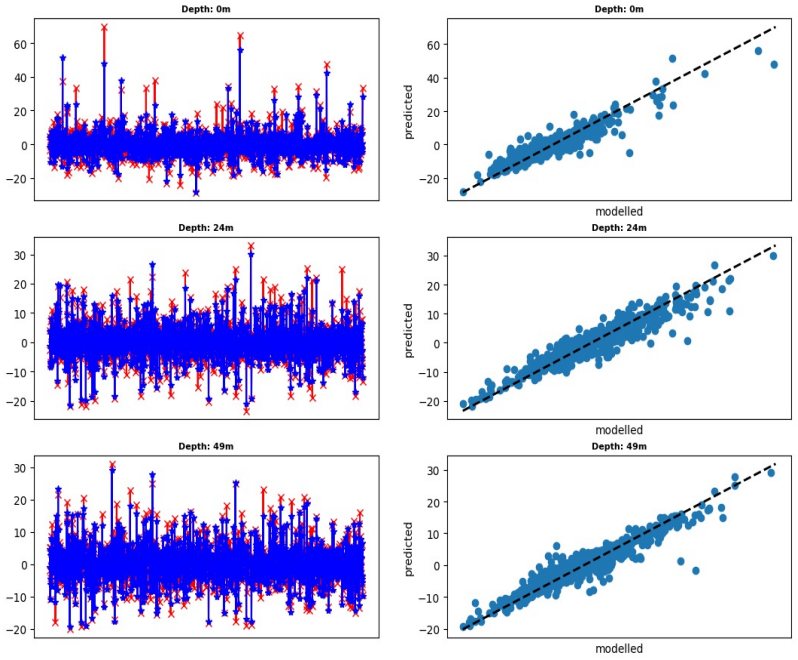
\+

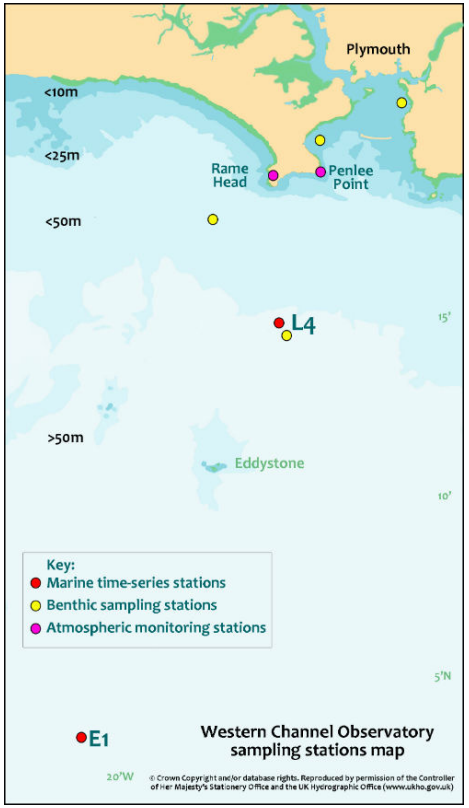
There can be plenty of existing model outputs to train the emulator

RP project REPLACE, which is collaboration between PML modelling, EOSA and Uni of Exeter

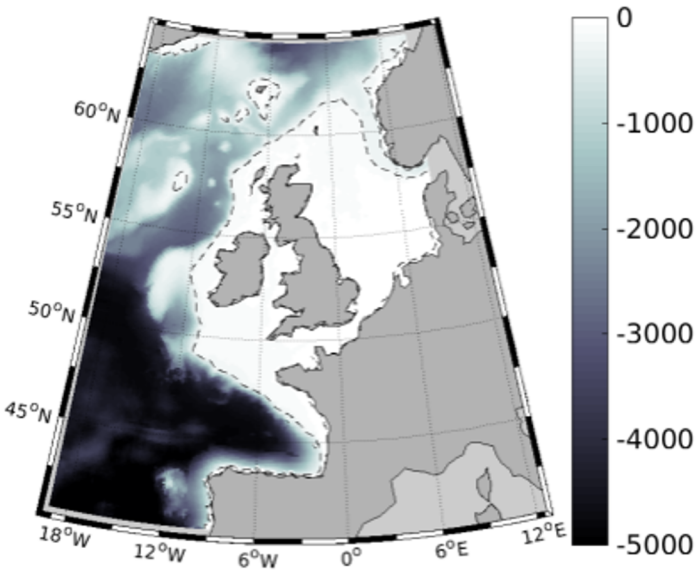
Predict oxygen throughout the water column from SST, surface chl, atmospheric and riverine data for the L4 location in English Channel.

tot chl + SST + phosphate riverine → oxygen





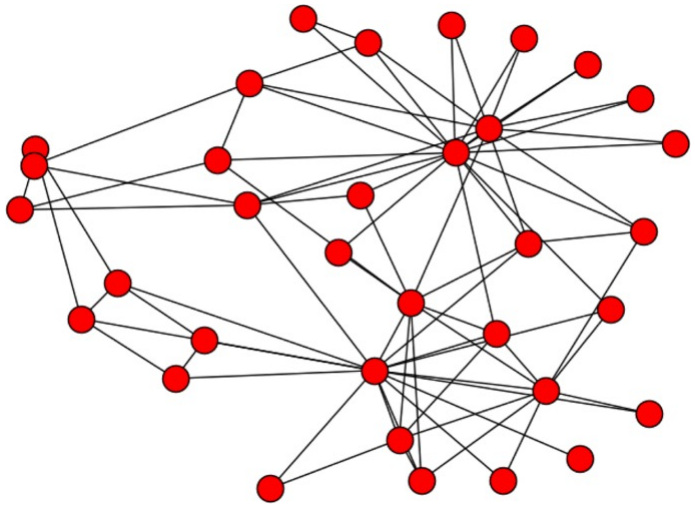
Planned: similar prediction for a 3D NWES domain based on 1D training simulations. The idea is to focus on predicting unobserved variables with high importance for carbon cycle (like POC, DOC).



PhD studentship CONECDA: to propagate information from assimilated variables to non-assimilated variables

An important problem: our operational system updates assimilated variables, but how to best update non-assimilated variables? Idea here is to train a ML emulator to map observed variables to unobserved variables within the ecosystem model. This allows us to calculate increments for non-assimilated variables from the increments of assimilated variables.

The point is to use complex networks to identify connectivity between model variables and spatial regions, and reduce complexity of a ML emulator.



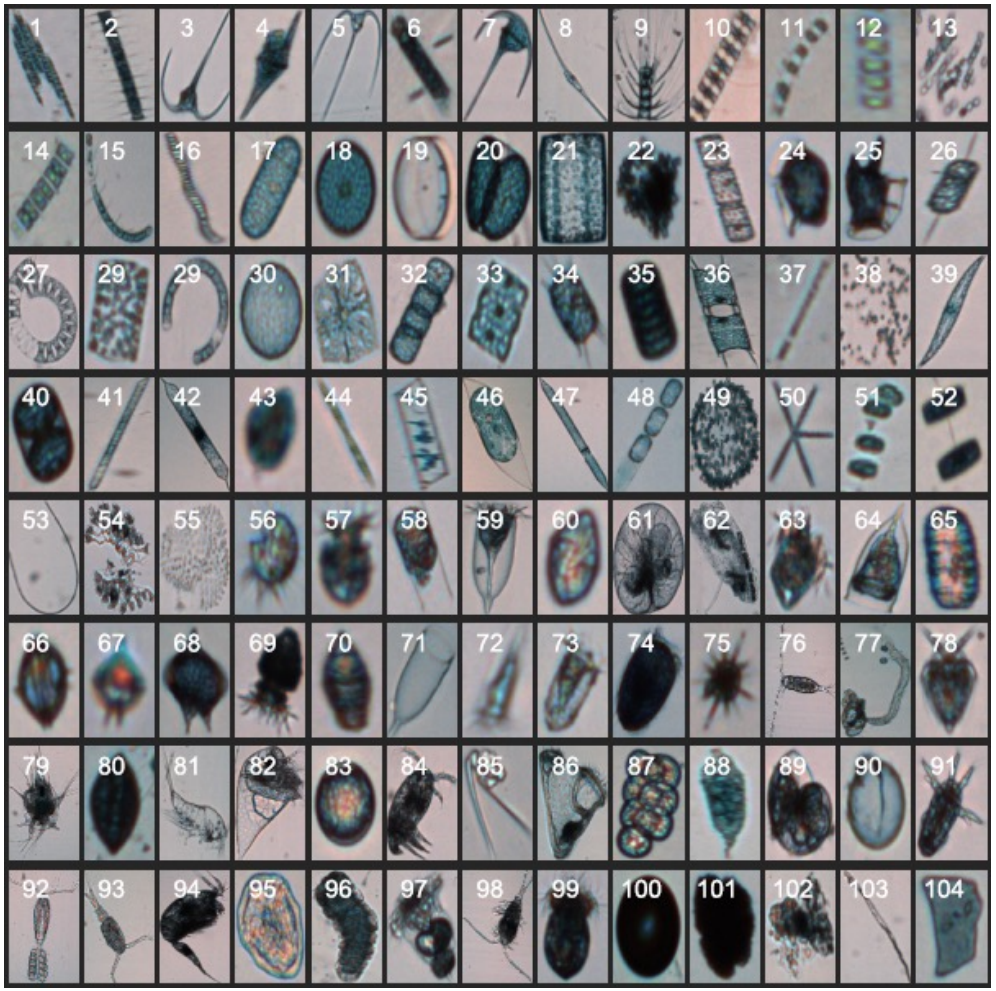
Planned: correcting model biases for eutrophication studies.

Training ML model on observations near the river mouth to predict nutrients. The model will be used to correct model biases in nutrients, a key indicator for eutrophication studies.

Using Machine Learning to automatically classify plankton in image data

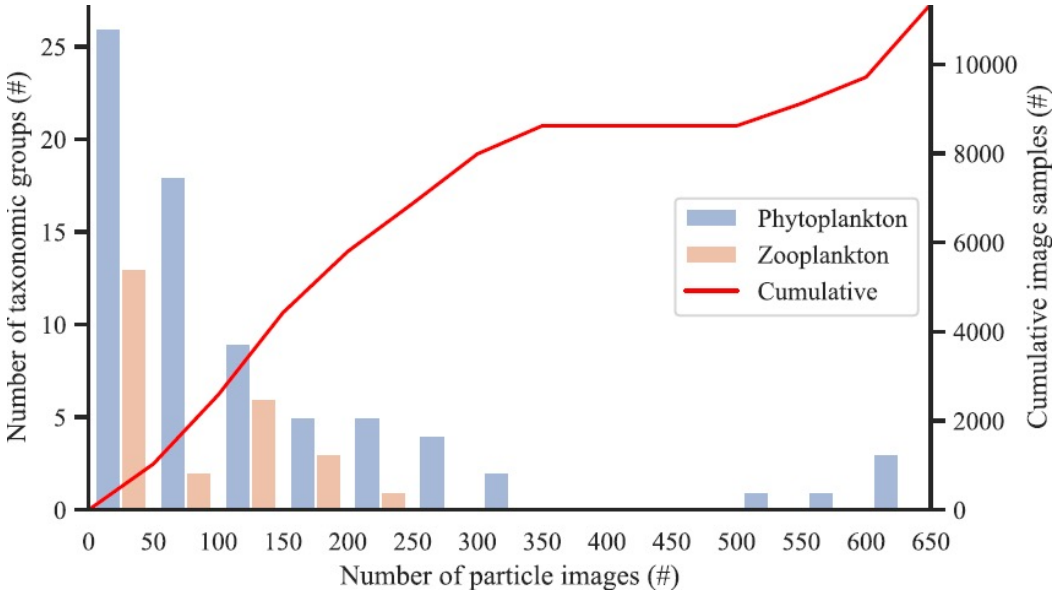
(work by T. Kerr, J. Clark and others)

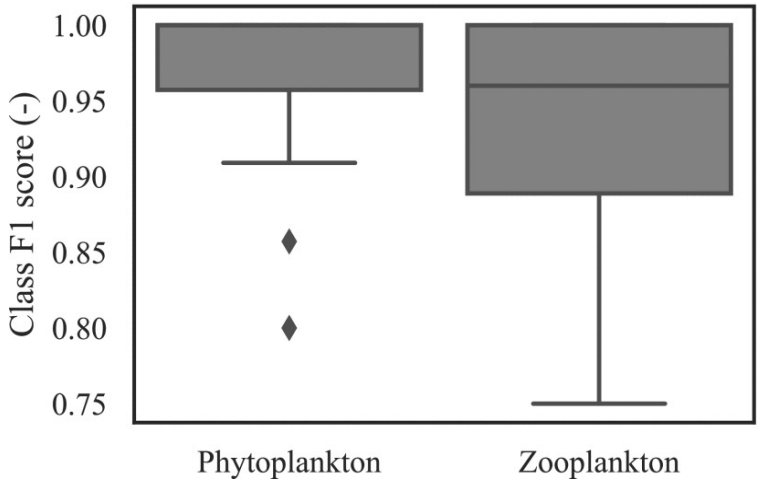
- Automated plankton imaging devices are capable of generating vast quantities of image data.
- These data records can be used to study plankton community and population dynamics in previously unprecedented levels of detail.
- However, given the volume of data, such records only become useful if automated classification algorithms can be developed to identify individual taxa.
- While some progress has been made in this field, the classification or rare taxa remains challenging.



Kerr et al IEEE Access (2020)

Using Machine Learning to automatically classify plankton in image data





- We investigated approaches for using *mu*ltiple deep learning models in collaboration to address the problem of class imbalance.
- The best performing model achieved an overall accuracy score of 97.4 % on FlowCam data.
- However, there remains room for improvement. For example, there was a greater spread in scores for zooplankton taxa, which reflects the fact they include more minority groups.

Kerr et al IEEE Access (2020)

Take home messages:

- ML has a major importance (in specific situations!) in replacing and/or improving the complex marine models: such work is ongoing and planned in the PML modelling group

- other ideas can be discussed in the future, e.g. using ML to improve model parametrization, to downscale the models, to learn model equations from data, ways how to integrate ML and data assimilation…

- other work on ML is ongoing at the PML modelling group

- looking to establish stronger links with ML experts !!