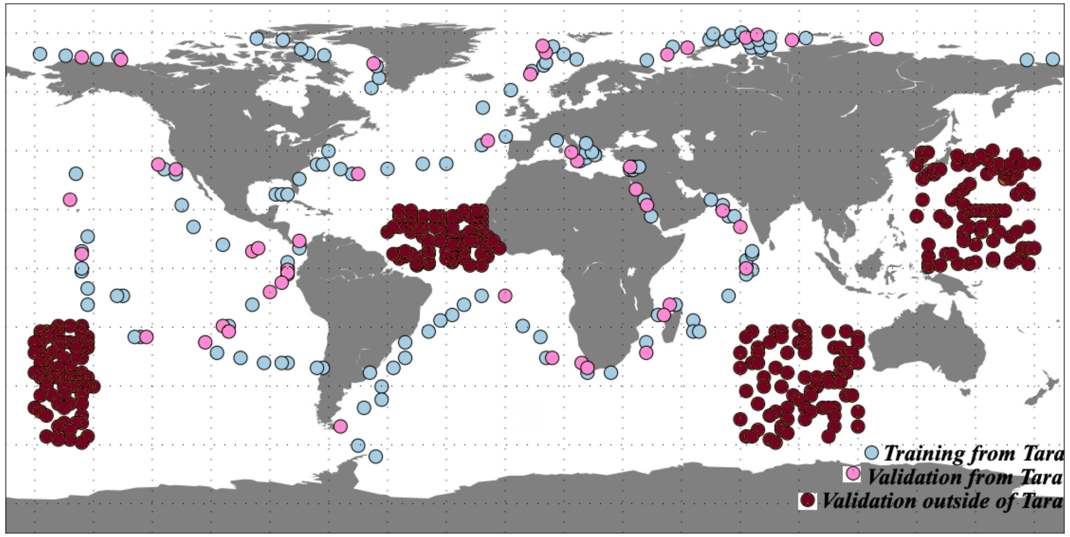
Using new observations and Machine Learning to improve organic sinking processes in the PlankTOM global ocean biogeochemical model

Denvil-Sommer[1] A., Le Quéré[1] C., Buitenhuis[1] E., Guidi[2] L., and Irisson[2] J.-O.
[1]School of Environmental Sciences, University of East Anglia, Norwich, UK, [2]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefanche, Villefranche-sur-mer, France

Aims:

- Reconstruction of small (**POC**) and large particulate organic carbon (**GOC**) as the function of lat, lon, depth, day, Temp, Chl, MLD, NO_3, PO_4 and Plankton Functional Types (**PFTs**).
- Test the impact of sparse observations on the performance of ML techniques using PlankTOM model outputs.

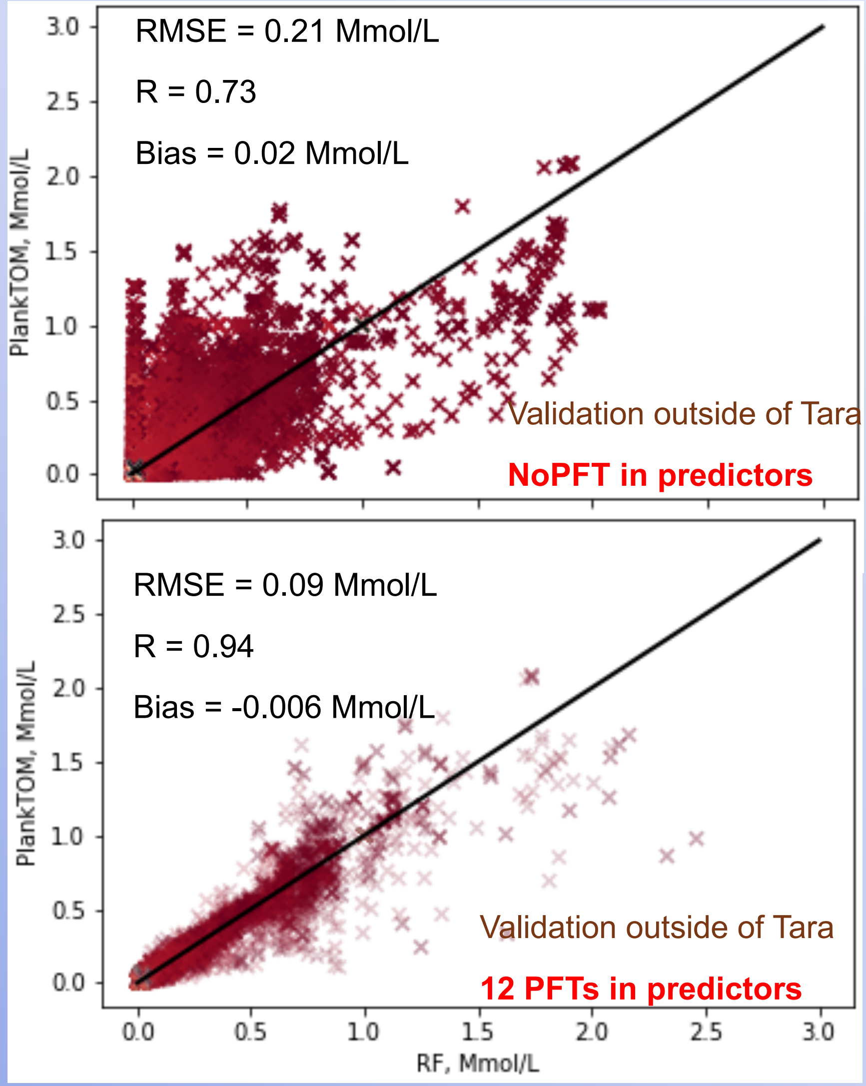
Data distribution



Tara stations' positions (2009-2013) are projected on PlankTOM grid and PlankTOM outputs are used to train and validate ML model.

Validation outside of Tara – PlankTOM outputs from regions where there are not real observations.

Random Forest POC reconstruction based on PlankTOM outputs:



Statistics at Validation Stations (pink dots) when 12 PFTs used as predictors

RMSE = 0.08 Mmol/L

R = 0.98

Bias = 0.005 Mmol/L

Main Results:

- Improvement of results by adding the PFTs as predictors
- The results in regions of Independent Validation (brown dots) are comparable with ones from validation stations (pink dots) when PFTs in predictors

In live chat:

- More about method
- Results for GOC
- Importance of different predictors

EGU General Assembly 2021

Motivation

Improve the parameterization of organic sinking velocity in PlankTOM model.

Small (POC) and large (GOC) particulate carbon concentration represent the concentration of sinking materials in the model. As the first step we reconstruct the concentration of POC and GOC from geographical position, environmental characteristics and ecosystem conditions from observations.

Background

To test the impact of sparse observations on the performance of ML techniques *pseudo-observations* were constructed from PlankTOM model outputs. Pseudo-observations were obtained by co-localizing model output with real-word observation positions.

PlankTOM Global Ocean Biogeochemical model:

Based on Ocean General Circulation model NEMO v3.1

12 Plankton Functional Types

Monthly outputs, 2^{o} spatial resolution

Tara expedition: *in situ* measurements for 2009-2013.

Real plankton and particle size distribution observations from the Underwater Vision Profiler (UVP), plankton diversity data.

Data Pre-Processing

Targets: POC and GOC

Drivers: day of the year, latitude and longitude, depth, Temperature (T), Chlorophyl (Chl), Mixed Layer Depth (MLD), Nitrate (NO3), Phosphate (PO4), Plankton Functional Types (PFTs)

$$POC_{\log n}, GOC_{\log n} = f(day_n, lat_n, lon_{n1}, lon_{n2}, depth_{\log n}, T_n, Chl_{\log n}, MLD_{\log n}, NO3_n, PO4_n, PFTs_n)$$

<u>Normalisation:</u>

$$day_{n=}cos\left(\frac{2\pi * day}{365}\right) \qquad lat_n = sin\left(\frac{\pi * lat}{180}\right) \qquad lon_{n1} = cos\left(\frac{\pi * lon}{180}\right) \qquad lon_{n2} = sin\left(\frac{\pi * lon}{180}\right)$$

$$X_{log} = \log(x) \qquad X_n = \frac{2}{3}\left(\frac{X - mean(X)}{std(X)}\right) \qquad X_{logn} = \frac{2}{3}\left(\frac{X_{log} - mean(X_{log})}{std(X_{log})}\right)$$

Due to the sparse data of Chl, NO3, PO4 in Tara these variables are averaged over MLD to assure their use in ML approach. By analogy with observations we do the same with PlankTOM outputs of Chl, NO3, PO4.

Random Forest (RF)

4253 samples for training.

1448 samples for validation.

8205 samples for independent validation.

We use **sklearn.ensemble.RandomForestRegressor**

The default parameters were applied in presented tests.

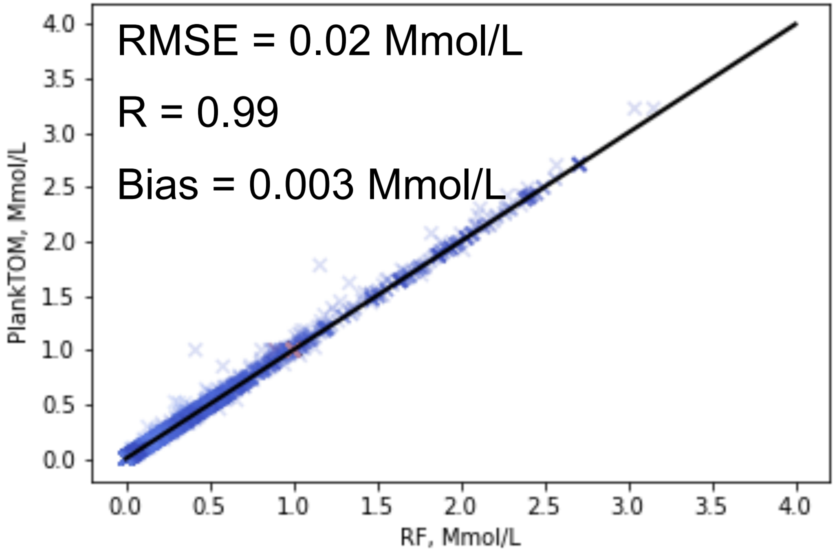
An optimal number of trees is 100. Numbers of 50, 200 and 1000 trees were also tested.

The whole dataset is used to build each tree.

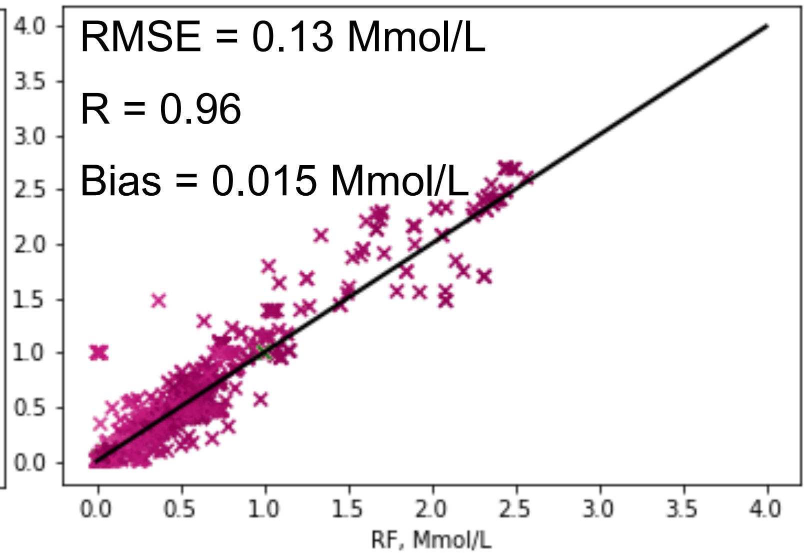
POC Reconstruction by Random Forest

- No PFT in predictors

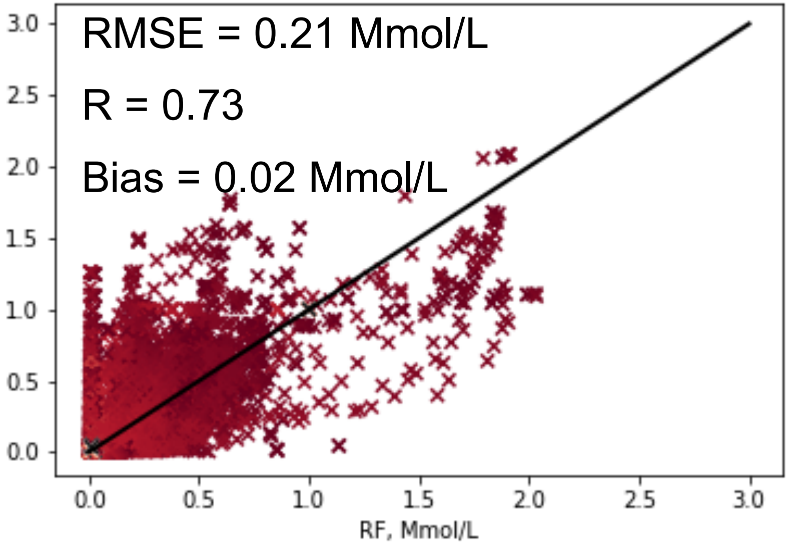
Train



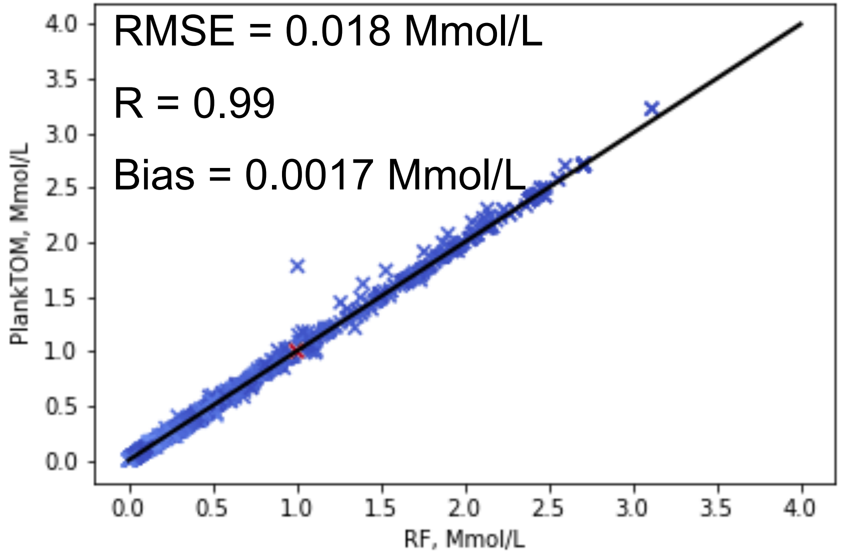
Validation

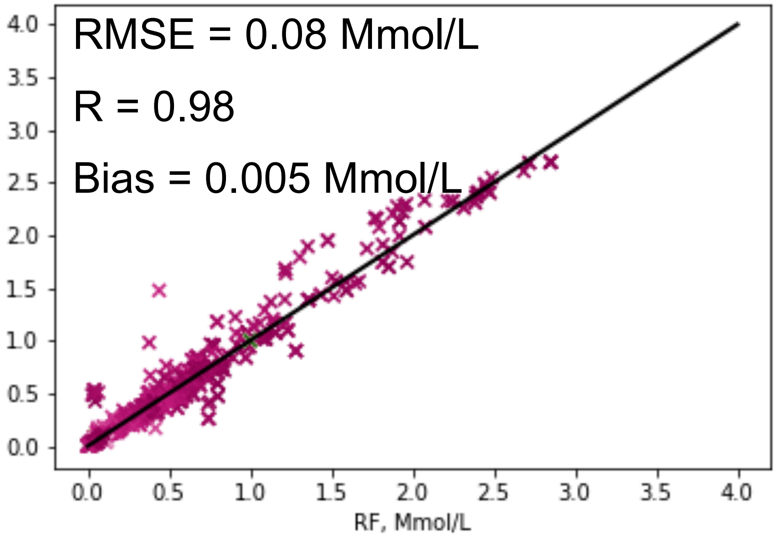


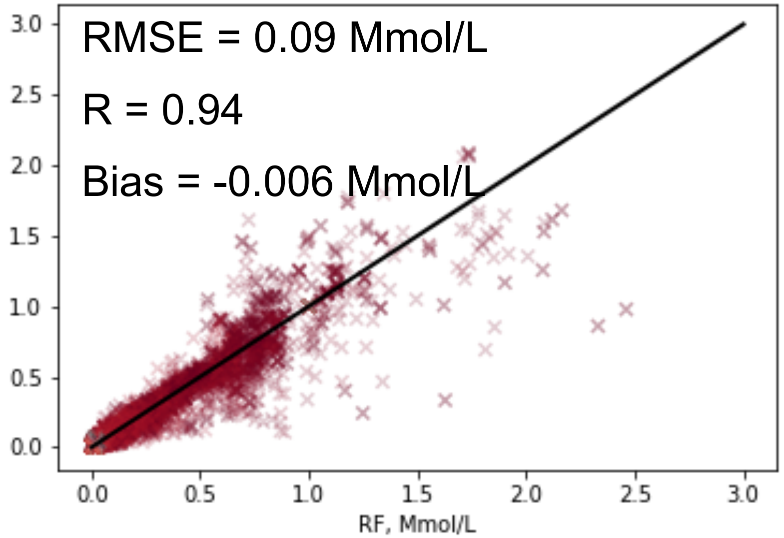
Validation outside of Tara



- 12 PFTs in predictors





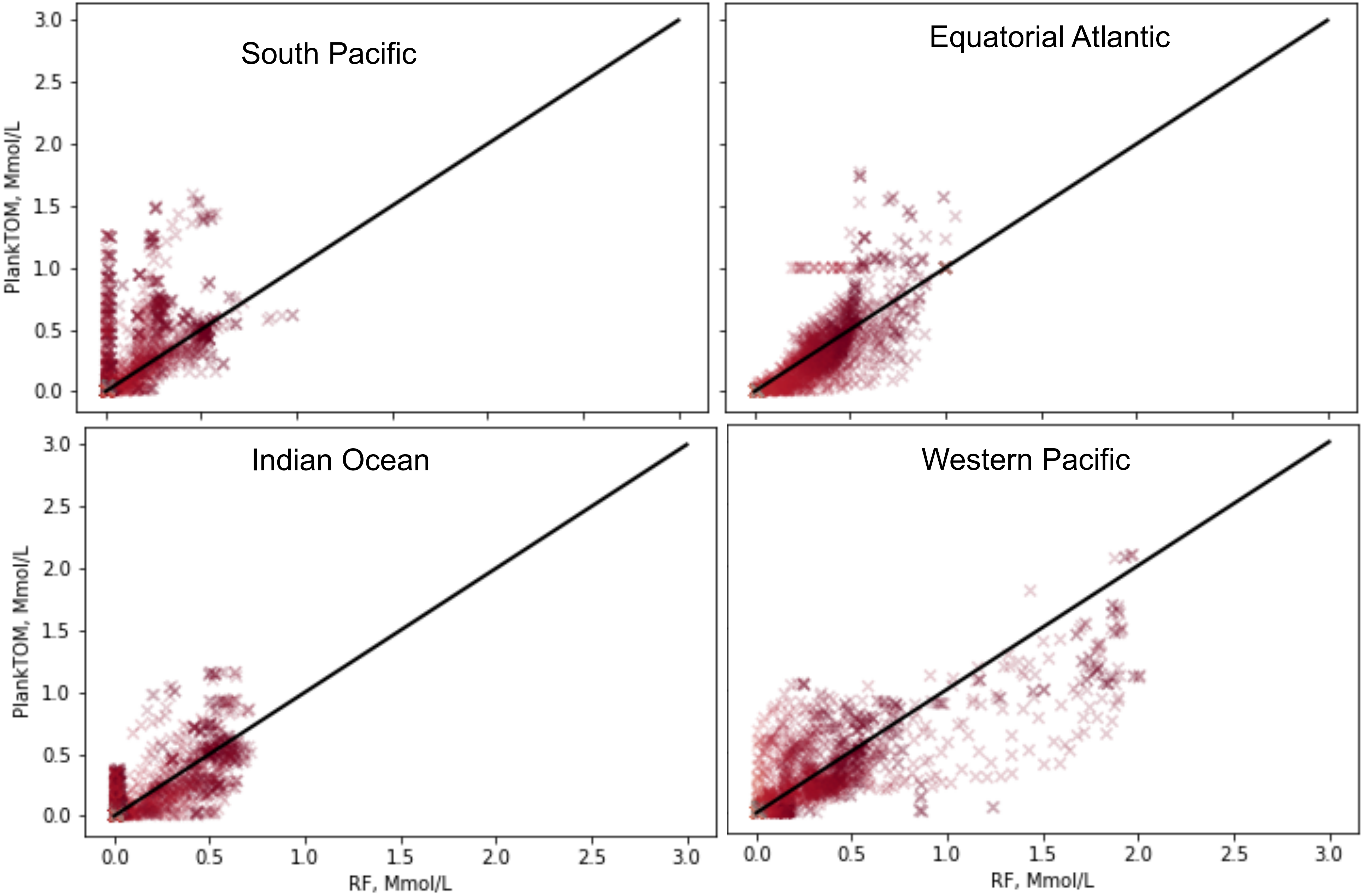


No important difference by using validation data

Large improvement with addition of PFTs in predictors

POC Reconstruction by Random Forest using validation data outside of Tara, per regions

- No PFT in predictors, Validation outside of Tara



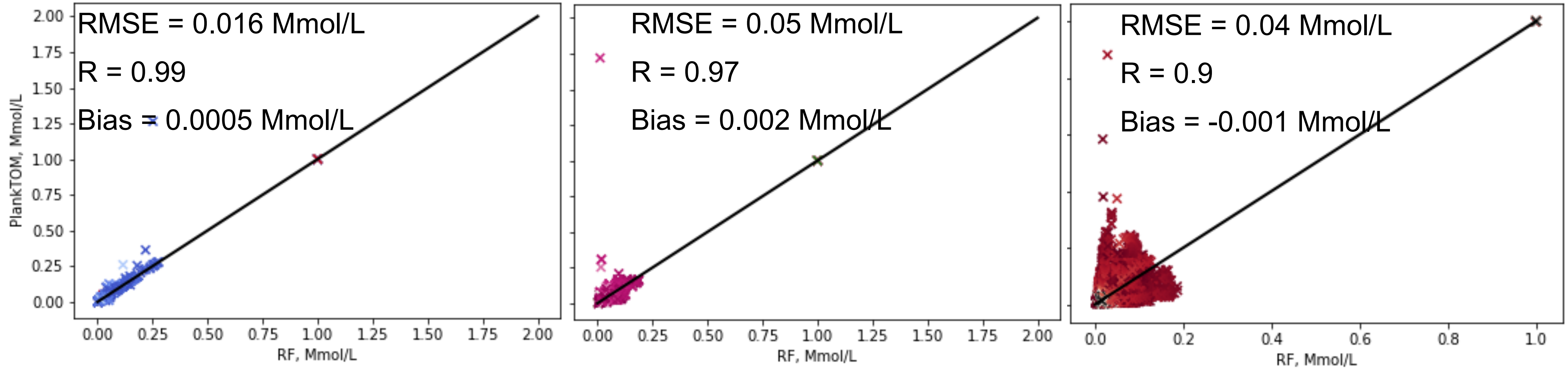
More difficultes to reproduce the values between 0 and 1.5 Mmol/L in the South Pacific and the Indian Ocean.

It can result from the particulate regimes in these zones that are not presented in the training data set.

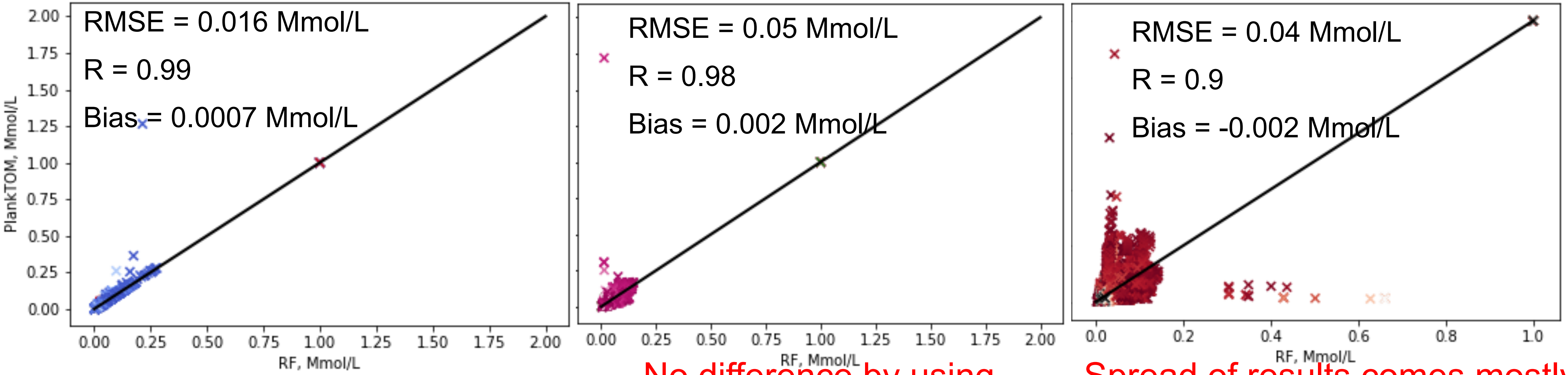
GOC Reconstruction by Random Forest

- No PFT in predictors

Train Validation Validation outside of Tara



- 12 PFTs in predictors



No difference by using validation data

Spread of results comes mostly from Eqiatorial Atlantic and South Pacific

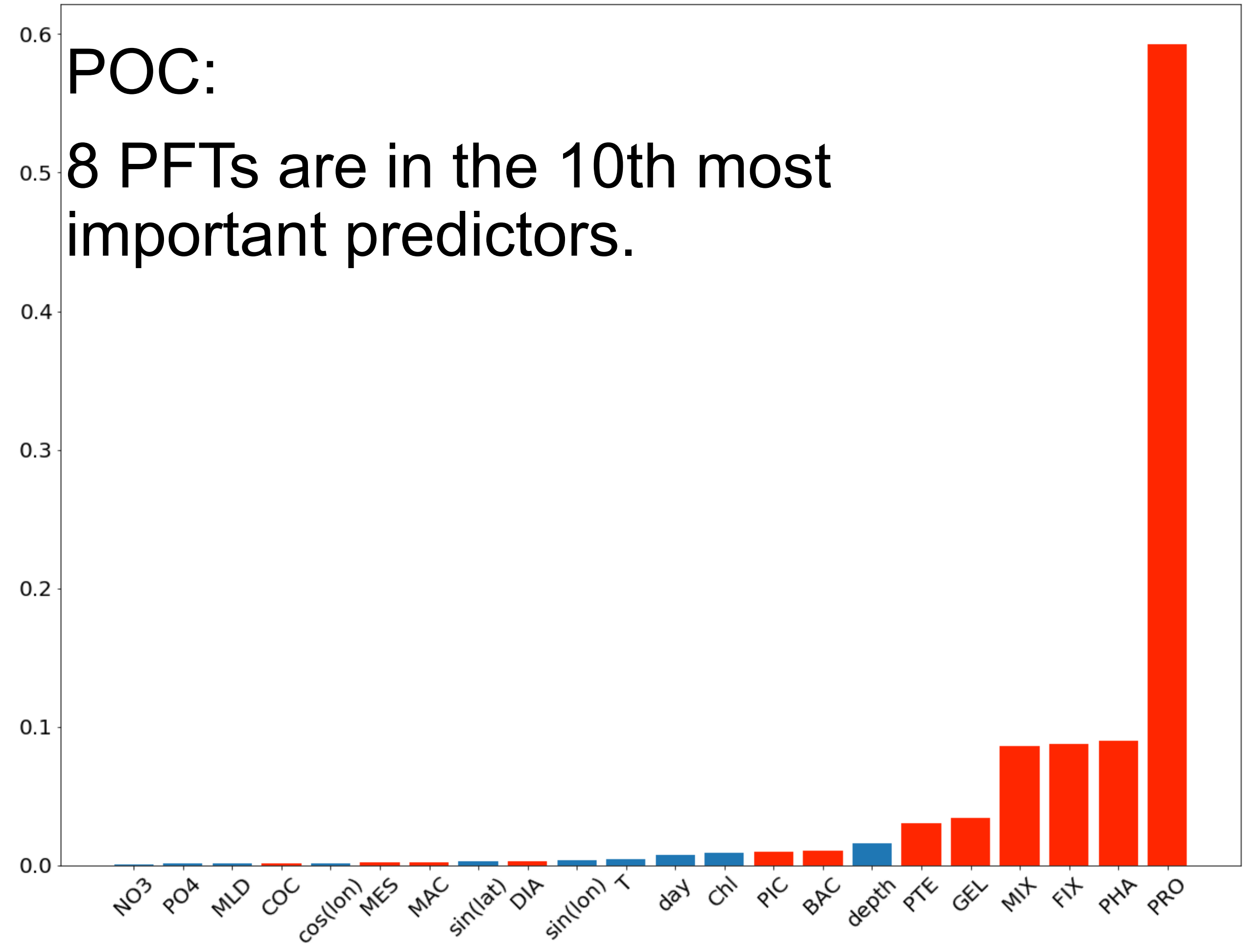
Predictors’ importance

List of PFTs:

- BAC – Bacteria
- PRO – Microzooplankton
- PTE – Pteropod
- MES – Mesozooplankton
- GEL – Jellyfish
- MAC – Macrozooplankton
- DIA – Diatom
- MIX – Mixed Phytoplankton
- COC – Coccolithophore
- PIC – Picophytoplankton
- PHA – Phaeocystis
- FIX – N2-fixers

POC:

8 PFTs are in the 10th most important predictors.

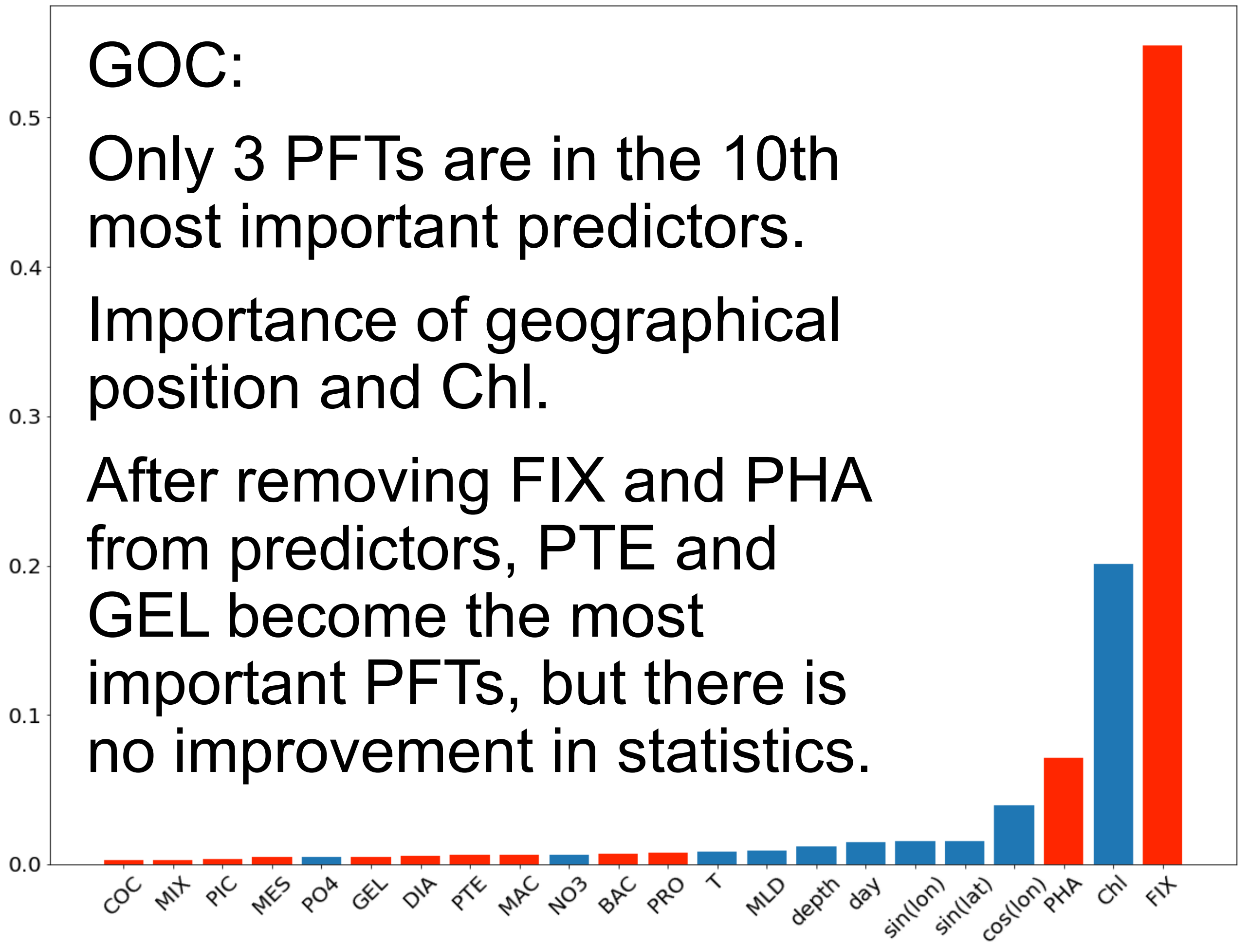


GOC:

Only 3 PFTs are in the 10th most important predictors.

Importance of geographical position and Chl.

After removing FIX and PHA from predictors, PTE and GEL become the most important PFTs, but there is no improvement in statistics.



Conclusion and perspectives

Findings

- Strong influence of PFTs on POC reconstruction.
- Not much influence of PFTs on GOC reconstruction.
- The local high values in GOC affect the training and result in less accuracy.

Next steps

- We need to understand why there is no impact of PFTs information on GOC at the current step of study.
- More *in situ* data will be available soon that will increase the number of training data and will resuts in better ocean cover.

Method development

- The feature importances from RF will be used for Neural Network (NN).
- At the moment we did not find a NN architecture that could at least reproduce the results from RF. We hope to have more data to build a NN.