Estimating global POC fluxes using ML and data fusion on heterogeneous and sparse *in situ* observations

Abhiraami Navaneethanathan[1], B.B. Cael[2], Chunbo Luo[1], Peter Challenor[1], Adrian Martin[2], Sabina Leonelli[1]

[1] *University of Exeter, Exeter, United Kingdom* [2] *National Oceanography Centre, Southampton, United Kingdom*

an498@exeter.ac.uk

Sharing is encouraged

1) Background

- The **ocean biological carbon pump** drives the vertical transport of particulate organic carbon (POC).
- Accurate estimates of POC fluxes can shed light on the underlying mechanisms of carbon transport that influence **ocean carbon sequestration** and the **distribution of nutrients** to marine ecosystems.
- POC fluxes can be derived from ***in situ* observations**, with the main sources being:
  1. **Sediment traps**
     - *Directly collects POC over time (a)*
  2. **234-Thorium radioactive tracers**
     - *Derived from 238-U and 234-Th disequilibrium (a)*
  3. **Underwater Vision Profilers (UVPs)**
     - *Images → particle size distribution → POC fluxes (b)*
- However, the resulting datasets are often **globally sparse**, leading to large model uncertainties in under-sampled areas.

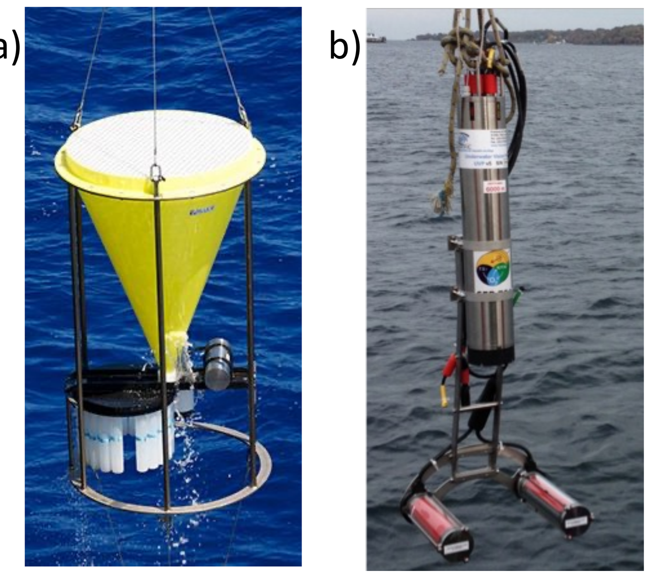

Figure 1: a) Sediment trap b) Underwater Vision Profiler

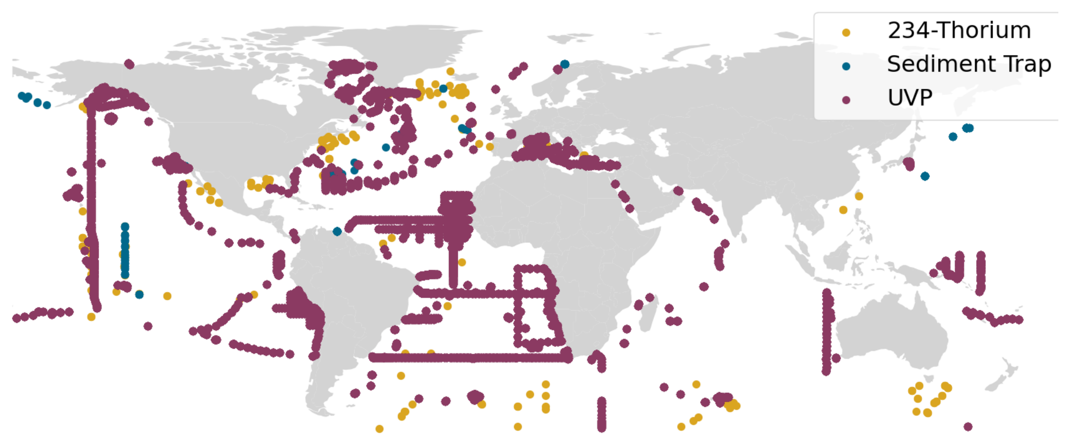

Figure 2: Observations from sediment trap, 234-Thorium tracer and UVP datasets between 100-300m depths, 60N-60S and aggregated over 1987-2020

2) Aims

1. **Estimate global POC fluxes** with well-sampled environmental driver datasets, testing several types of models.
2. **Combine *in situ* heterogeneous POC flux datasets** to address the sparsity in measurements via data fusion methods.
3. Identify the **importance of each environmental driver** for predicting POC fluxes.

3) Methods

- **Predictors:** well-sampled global environmental driver datasets (monthly climatologies from the World Ocean Atlas).
- **Training data:** *in situ* POC flux observations from three different sources.
- **Data fusion:** stacking predictions from models trained on multiple samples of the 3 *in situ* datasets.

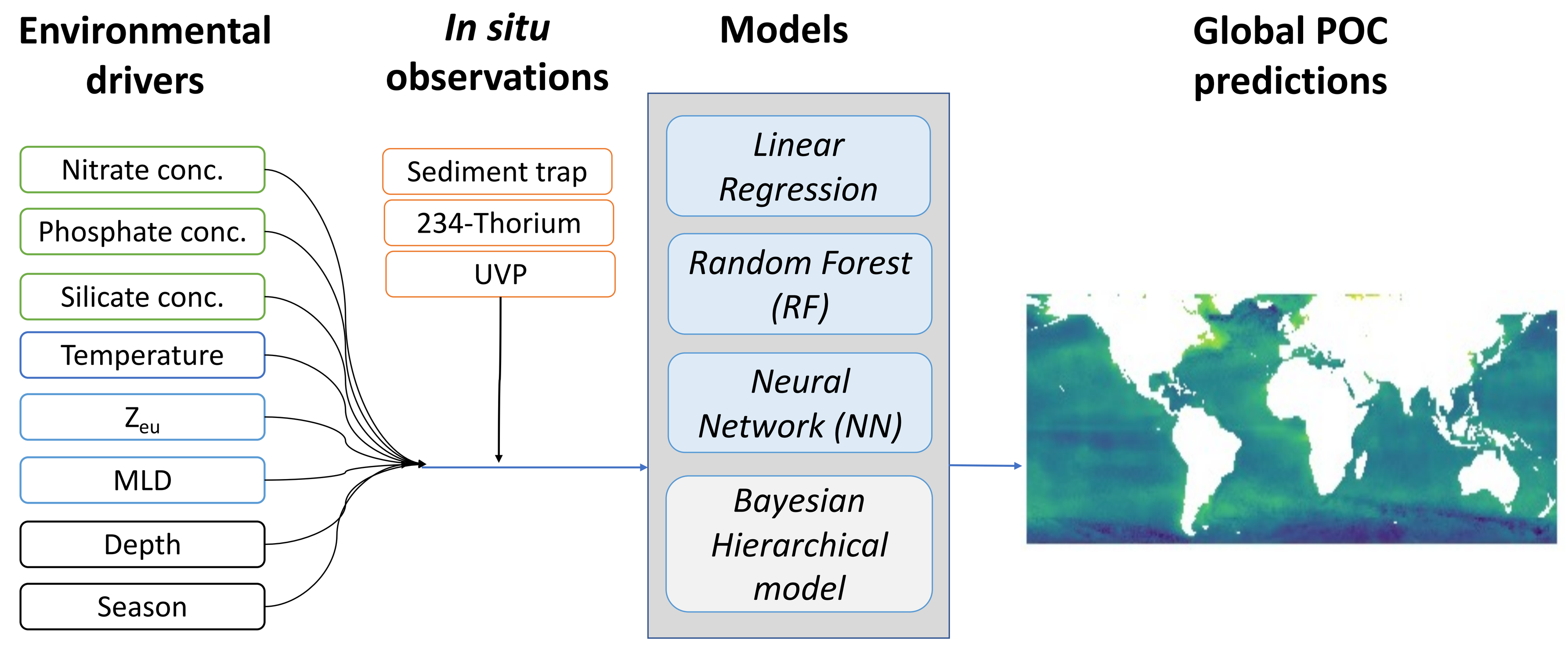

Figure 3: General model workflow for predicting global POC fluxes (1 degree resolution)

Bayesian hierarchical model (BHM)

- Hierarchical models group the POC flux observations by their instruments to appropriately account for their uncertainties with domain knowledge.
- This fusion of the 3 datasets gives the true process model
- Trained using MCMC sampling.

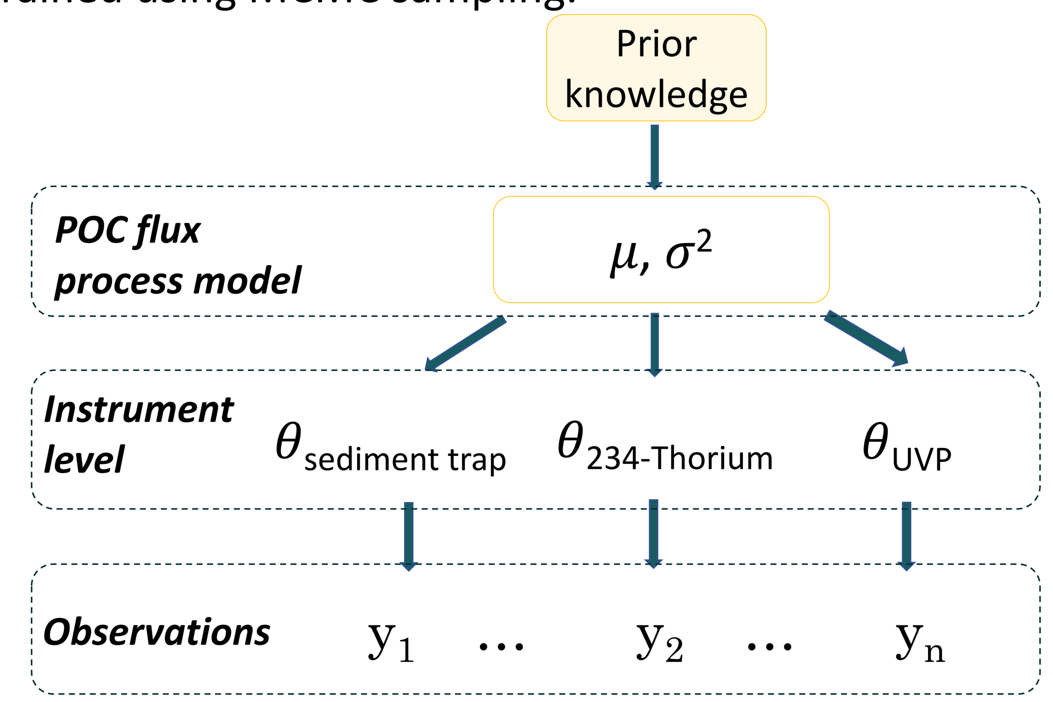

Figure 4: Figure 4: Schematic diagram of the Bayesian hierarchical model

4) Preliminary Results

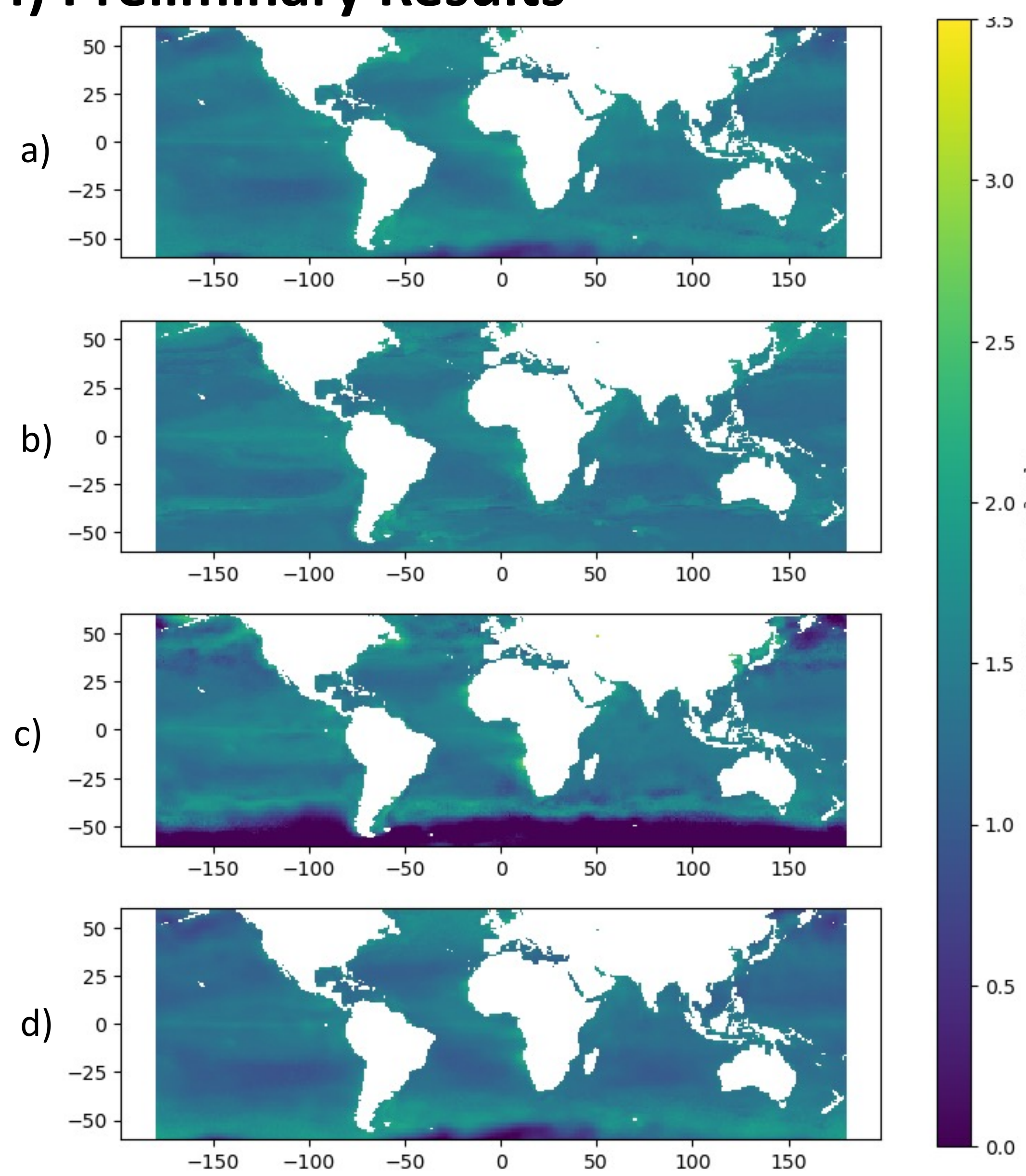

Figure 5: Global estimates of POC fluxes from a) linear regression, b) Random Forest, c) Neural Networks and d) Bayesian hierarchical models.

Model	RMSE	R2
Linear Regression	0.3414	0.3964
Random Forest	0.2282	0.7766
Neural Network	0.2812	0.6711
Bayesian hierarchical model	0.3173	0.4848

Table 1: Model performance across the 3 data sources

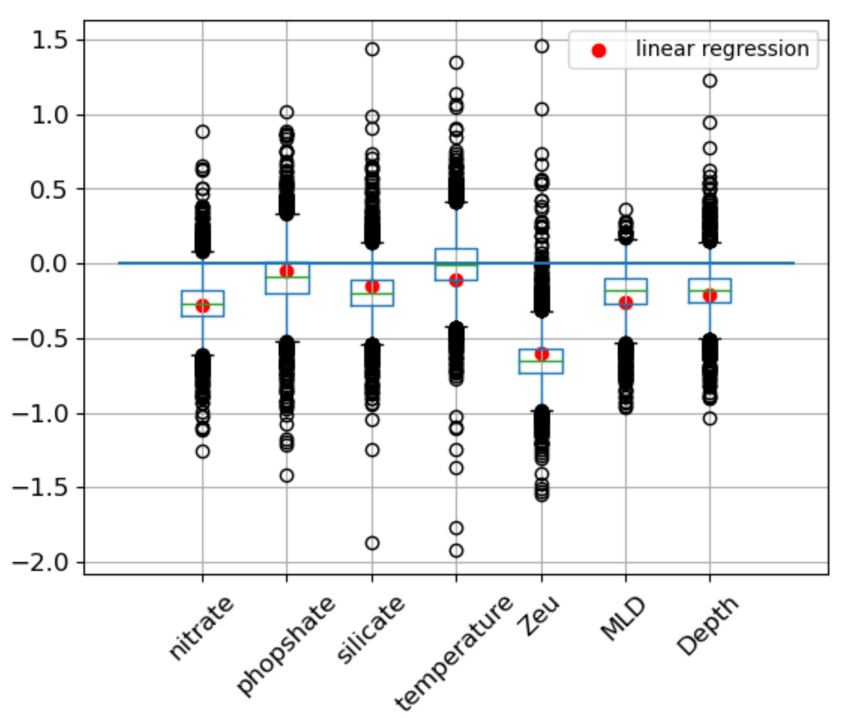

Figure 6: Bayesian hierarchical model parameter posterior distributions with corresponding linear regression coefficients

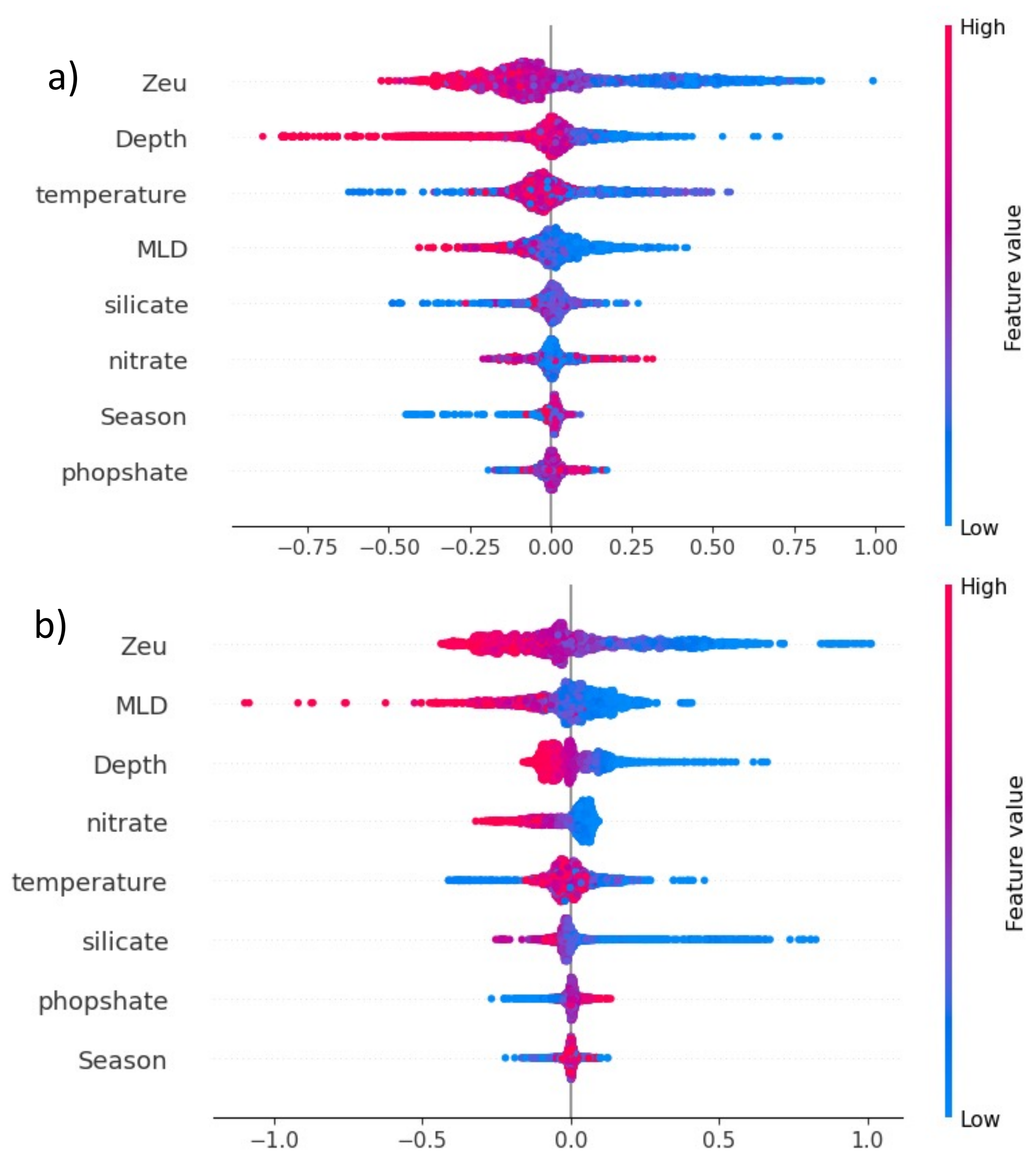

Figure 7: SHAP values of the environmental drivers for the a) Random Forest and b) Neural network model.

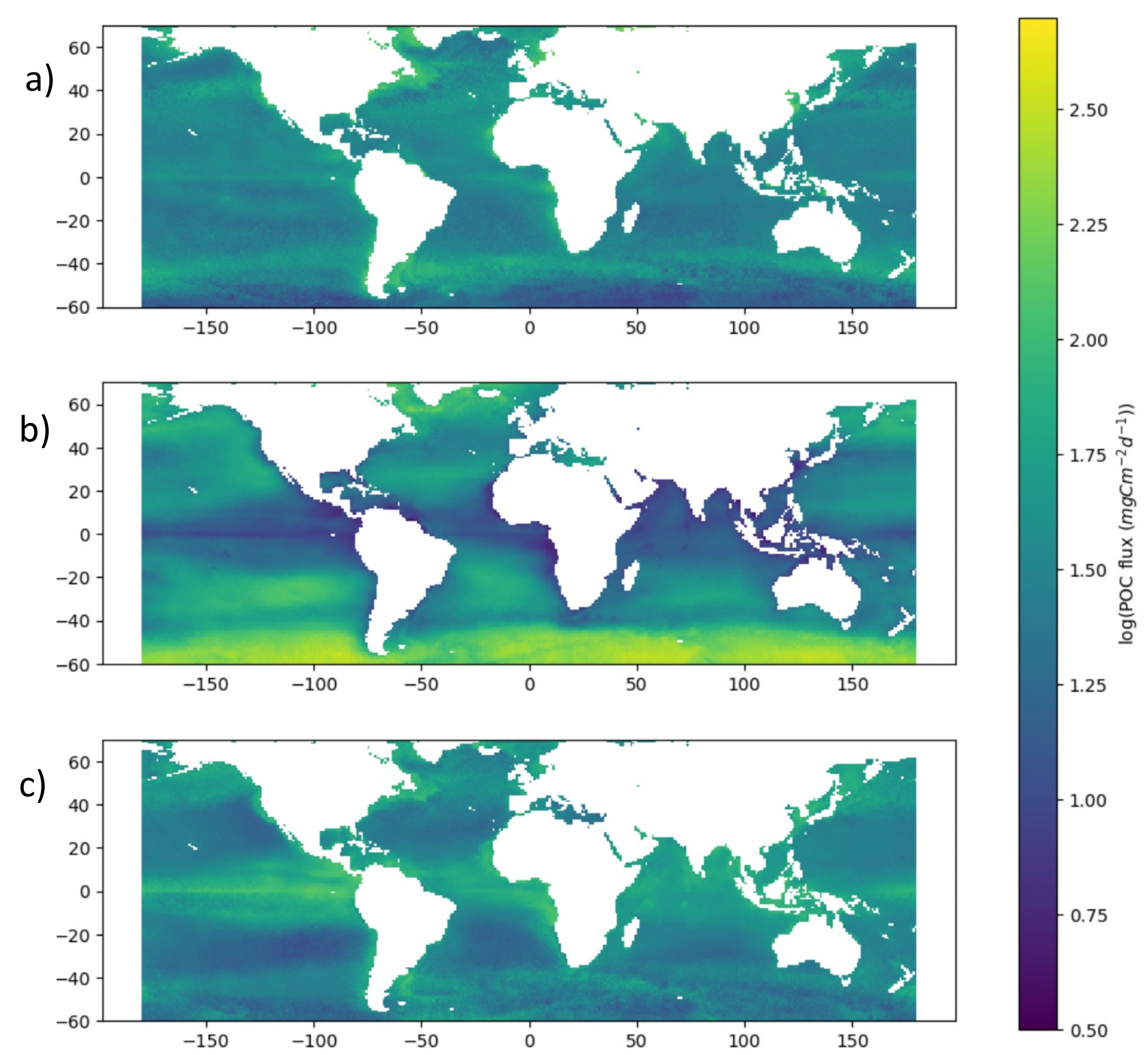

Figure 8: Global predictions of POC fluxes from models trained on a) UVP b) 234-Thorium c) sediment trap observations averaged over all model types

5) Discussion

- *Non-linearity*: The RF and NN can model complex non-linear relationships well.
- *Uncertainty*: The BHM can give uncertainties for each parameter via the posterior distributions.
- *Measurement error:* Only accounted for in the BHM.
  - Predicting instrument measurements ≠ predicting true POC fluxes.
- *Interpretability*: The BHM and linear regression give numerical relationships between drivers and POC fluxes.

6) Conclusions

- ML and statistical models trained on fused *in situ* POC flux observations and environmental driver datasets can **effectively estimate global POC fluxes.**
- The **random forest** model performs the best.
- The most important environmental drivers for estimating global POC fluxes were found to be the depth parameters (**euphotic depth (Zeu), MLD and depth**).

7) Future work

- **Regional** and **seasonal** analysis
- Include **interaction** terms

References:

Brewin, R. J. W., et al., (2021). Sensing the ocean biological carbon pump from space: A review of capabilities, concepts, research gaps and future developments. doi: 10.1016/j.earscirev.2021.103604
Garcia, H. E., et al., (2019). *World Ocean Atlas 2018.*. https://data.nodc.noaa.gov/woa/WOA18
Britten, G. L., et al., (2021). Evaluating the Benefits of Bayesian Hierarchical Methods for Analyzing Heterogeneous Environmental Datasets: A Case Study of Marine Organic Carbon Fluxes. doi: 10.3389/fenvs.2021.491636
Kiko, R., Hauss, H., & Mehrtens, H. (2021). *SFB754 Zooplankton and particle distribution from UVP5 measurements.* https://doi.org/10.1594/PANGAEA.927040
Mouw, C. B. et al., (2016). Global ocean particulate organic carbon flux merged with satellite parameters. https://doi.org/10.5194/essd-8-531-2016
Ceballos-Romero, E., et al., (2021). *More than 50 years of Th-234 data: a comprehensive global oceanic compilation.* https://cafethorium.whoi.edu/thorium-data/