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

Abhirami Navaneethanathan1, B.B. Cael2, Chunbo Luo1, Peter Challenor3, Adrian Martin1, Sabina Leonelli1

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

an498@exeter.ac.uk

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 234 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.

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.

Environmental drivers

POC fluxes

Models

Bayesian Hierarchical model

Global POC predictions

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.

4) Preliminary Results

Table 1: Model performance across the 3 data sources

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.3413</td>
<td>0.964</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.2828</td>
<td>0.7766</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.2812</td>
<td>0.6711</td>
</tr>
<tr>
<td>Bayesian Hierarchical model</td>
<td>0.3173</td>
<td>0.4848</td>
</tr>
</tbody>
</table>

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 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


