Developing daily gap-filled chlorophyll-a datasets using ensemble (tree) models and deep neural networks that incorporate co-located environmental variables

Shridhar Sinha – University of Washington, Paul G. Allen School of Computer Science & Engineering, Seattle WA
Yifei Hang– University of Washington, Applied & Computational Mathematical Sciences, Seattle WA
Dr. Elizabeth Eli Holmes– NOAA Fisheries, Northwest Fisheries Science Center, Seattle WA

Introduction

There are numerous challenges in estimating ocean chlorophyll (Chl-a), a key indicator of ocean productivity that supports ecosystems and fisheries. Our study focuses on the North Indian Ocean, a region where seasonal upwelling zones drive significant Chl-a blooms. Chl-a estimates are derived from ocean color remote-sensing, but cloud cover poses a major obstacle, as sensors cannot penetrate clouds, leading to large gaps in the data. This issue is particularly pronounced during the summer monsoon when upwelling peaks, and cloud cover is most prevalent.

Traditional gap-filling methods, such as spatial interpolation from neighboring non-missing Chl-a values, often fail when faced with large areas of missing data. In our research, we explore the use of deep-learning and ensemble models to improve Chl-a estimations by incorporating co-located environmental data. Environmental variables, such as sea surface temperature and wind speed, are often correlated with Chl-a due to the processes that influence its growth and distribution. By utilizing these relationships, our approach reduces the reliance on spatially adjacent Chl-a observations and aims to enhance the accuracy of Chl-a estimates in regions with substantial cloud cover.

Acknowledgements: *Thank you to the University of Washington School of Fisheries and Aquatic Sciences Varanasi Endowment which supports summer interns in Quantitative fisheries research at the University of Washington and financial support from the NOAA Fisheries Northwest Fisheries Science Center and the Office of Science and Technology.*

GitHub: *https://github.com/SAFS-Varanasi-Internship*

Methods

PROBLEM: When testing our gap-filling algorithms, evaluating performance is challenging because the true values in the data gaps are unknown. Without a clear metric to compare against, it becomes difficult to determine how well the algorithm is performing or what constitutes a successful result.

SOLUTION: To address this, we create "fake cloud cover" by superimposing cloud cover from the recent past and future onto test data, simulating gaps for the day in question. The withheld data is then gap-filled by the algorithm, allowing us to compare the results with the known ground-truth values and effectively evaluate performance.

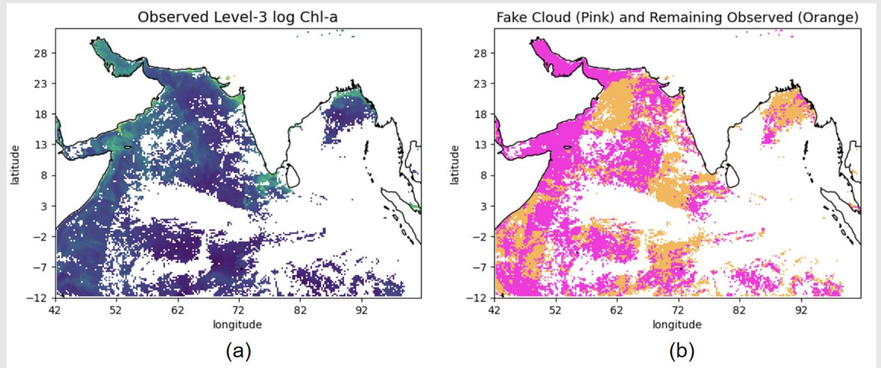

Figure 1. (a) observed chlorophyll of a certain day from the Global Ocean Colour (Copernicus-GlobColour) L3 dataset; (b) cloud masks of fake cloud covers (pink) and the remainder of observed (orange) which the predictions are based on during training.

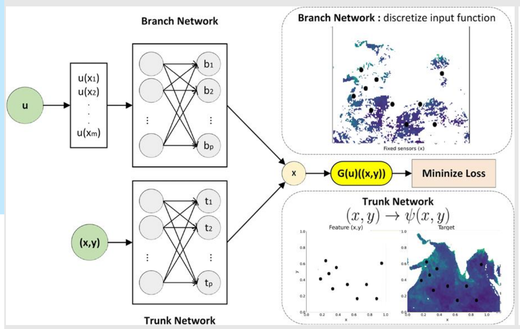

Figure 2. The DeepONet PINN architecture. It consists of two neural networks: the branch, and the trunk. The branch encodes the input function at a fixed number of sensors(remote-sensed data without gap-filling), and the trunk encodes the locations for the output functions(ap-free product).

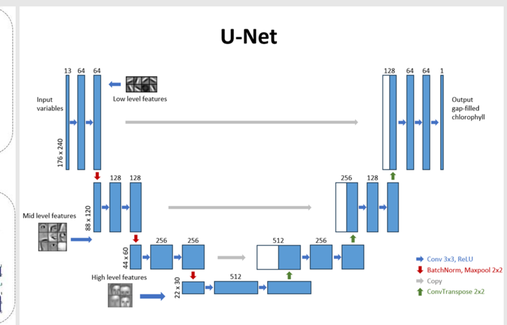

Figure 3. The U-Net architecture which follows an auto-encoder structure and learns hierarchical image features with convolutional layers. The feature images of facial features present an intuitive visualization for U-net's feature extraction ability.

Results

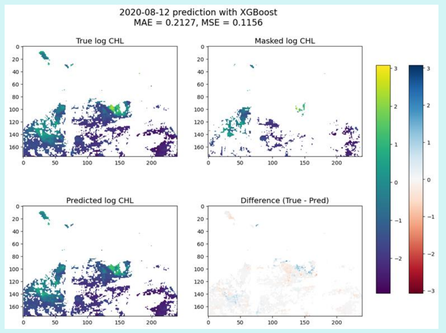

Figure 4. XGBoost Gradient Boosted Tree implementation evaluation on Copernicus L3 data. Bottom right shows the percentage difference between observed and predicted

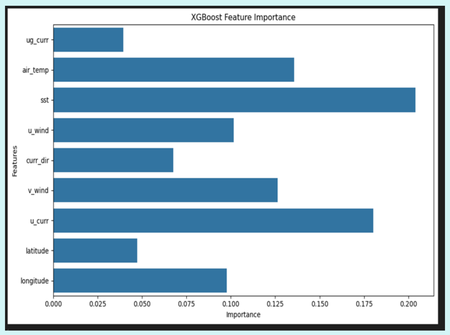

Figure 5. The feature importances of co-located environmental data in the XGBoost model

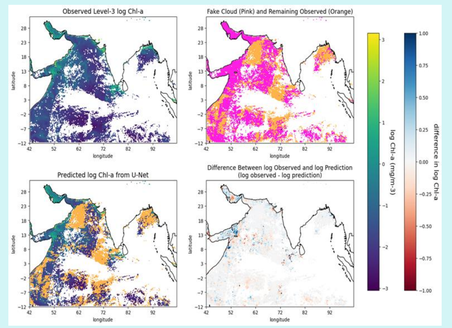

Figure 6. U-Net evaluation on Copernicus L3 data. Bottom right shows the percentage difference between observed and predicted

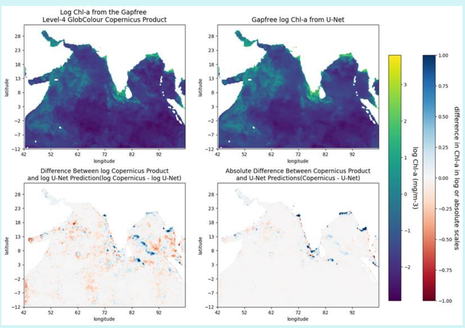

Figure 7. U-Net evaluation on Copernicus L4 (gapfree) data. Bottom left shows the percentage difference between L4 and U-Net predictions while the bottom right shows absolute difference. Main differences can be observed at the coastlines

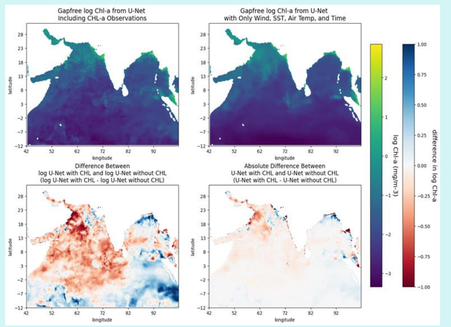

Figure 8. Comparison between U-Net models trained with/without CHL information as model features. Note the no-CHL model can capture general patterns and achieves relatively low absolute differences

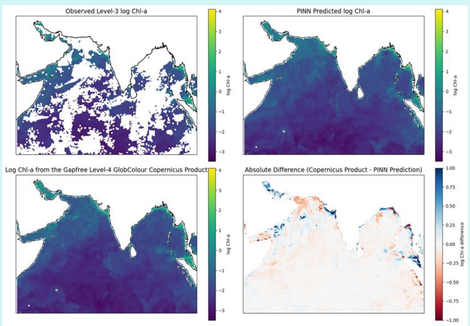

Figure 9. DeepONet evaluation on Copernicus L4 (gapfree) data. Bottom right shows the percentage difference between observed and predicted

Discussion

Deep-learning models designed for image problems demonstrate strong performance, producing Chl-a gap-filled predictions consistent with the Level-4 Copernicus-GlobColour gap-free product, highlighting their effectiveness as comparable to science-grade outputs. The fake cloud metric provides us an estimate of true error that is absent from products such as Copernicus-Globcolour gap-free. When compared to commonly used non-linear scientific machine learning regressors such as XGBoost, we see a more accurate performance from Deep Neural nets, and specifically the U-Net model. These deep-learning algorithms for gap-filling provide solutions that can be applied by researchers to their regional gap-filling problems and that can be tested against their local data---unlike algorithms used in Level-4 products which are not easily available (open source) and available for customization and testing. The current Level-4 gap-filling algorithms are optimized for global performance and locally trained models might be better and can be designed for specialized objectives, like coastal predictions.