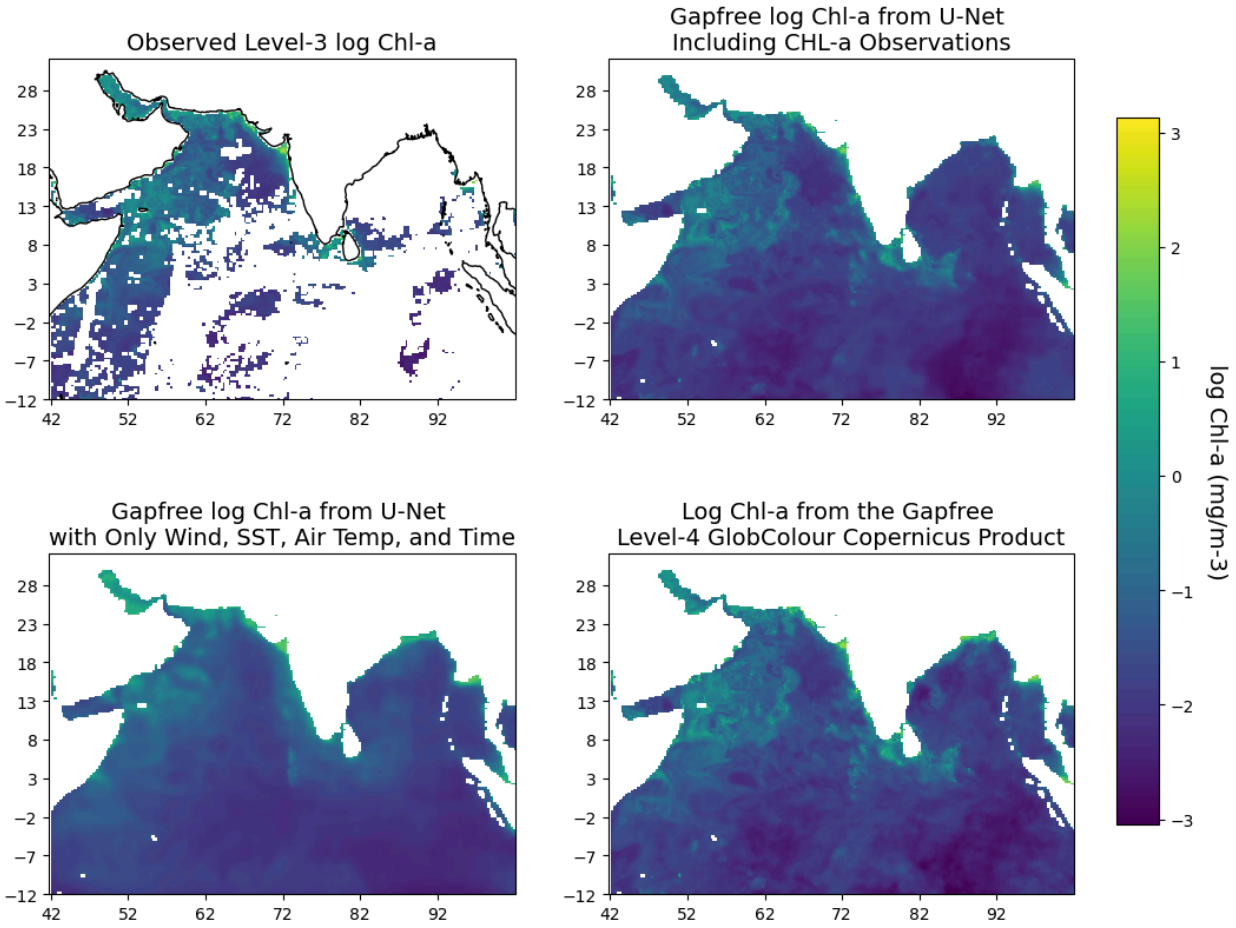
Developing daily gap-filled chlorophyll-a datasets using ensemble (tree) models and deep neural networks that incorporate co-located environmental variables

Shridhar Sinha (1), Yifei Hang (2), Elizabeth Eli Holmes (3)

(1) University of Washington, Paul G. Allen School of Computer Science & Engineering
(2) University of Washington, Applied & Computational Mathematical Sciences
(3) NOAA Fisheries, Northwest Fisheries Science Center

Ocean chlorophyll (Chl-a) is a key indicator for ocean productivity, supporting ecosystems and fisheries. Chl-a products are derived from ocean color remote-sensing, but the sensors cannot penetrate clouds. The result is missing Chl-a estimates when clouds are present. Our study area is the North Indian Ocean, a region with strong seasonal upwelling zones that drive Chl-a blooms. Upwelling peaks during summer monsoon when the region has high cloud cover, and thus Chl-a data are missing. Common methods for gap-filling Chl-a data relying on spatial interpolation from non-missing data do not perform well with large regions of missing values. Our research studies the use of deep-learning and ensemble models that incorporate co-located environmental data to improve Chl-a estimation and reduce the reliance on spatially adjacent Chl-a observations. Physical environmental variables can be highly correlated with Chl-a due to the processes that drive Chl-a growth and movement. The deep-learning models tested were the U-Net architecture of Convolutional Neural Network and Physics-Informed Neural Networks (PINN) with additional co-located physical variables from other remote-sensing sources: sea surface temperature, air temperature, surface winds, and surface currents. We compare these with standard random forest and gradient boosted tree approaches. Our models were trained on the Level-3 (gappy) Chl-a product from CMEMS (Copernicus-GlobColour). Performance was evaluated using two approaches: cross-validation by creating train and test sets from the (gappy) Level-3 Chl-a data and comparison of full-region Chl-a predictions to output from the Level-4 Copernicus-GlobColour product, a gap-filled Chl-a product using a different algorithm. Our cross-validation results show a seasonal pattern of performance across years and models, with the closest predictions in spring and the farthest during summer. The predictions of deep-learning models using available Chl-a observations are consistent with the Level-4 Copernicus-GlobColour gapfree product. This indicates that the models are performing well (comparable to a science grade gapfree product). When no observed Chl-a data were included in the features, the predictions still matched the spatial pattern of the observed Chl-a but the mean-squared error was orders of magnitude higher. This suggests that machine learning approaches might reveal general Chl-a patterns even when ocean color data are unavailable (prior to 1997).


Observed Level-3 log Chl-a
Gapfree log Chl-a from U-Net
Including CHL-a Observations
Gapfree log Chl-a from U-Net
with Only Wind, SST, Air Temp, and Time
Log Chl-a from the Gapfree
Level-4 GlobColour Copernicus Product
log Chl-a (mg/m-3)
28
23
18
13
8
3
−2
−7
−12
42
52
62
72
82
92
3
2
1
0
−1
−2
−3