Ocean Carbon From Space Workshop
14–18 February 2022 | Online event

Estimation of Phytoplankton Levels in Global Waters Using Supervised Machine Learning

Submission ID: 148

Surya Prakash Tiwari[1,*], <u>Subhrangshu Adhikary</u>[2,#], Saikat Banerjee[3]

* Corresponding Author
# Presenting Author

[1,*] Center for Environment & Marine Studies, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia

[2,#] Dr. B.C. Roy Engineering College, Durgapur-713206, West Bengal, India

[3] Department Of Remote Sensing, Wingbotics, Kolkata-700086, West Bengal, India

esa NASA PML Plymouth Marine Laboratory surface.ocean solas lower.atmosphere.study

Outline

- Introduction
- Motivation
- Methodology
- Results
- Knowledge gaps & priorities for next steps
- Conclusion

Introduction

- The phytoplankton is driving the marine food chain and plays a crucial role in balancing the oceans, seas and freshwater ecosystems.
- The imbalance of certain biochemical and physical parameters affects phytoplankton growth and these parameters can be obtained from remote sensing sensors [1-2].

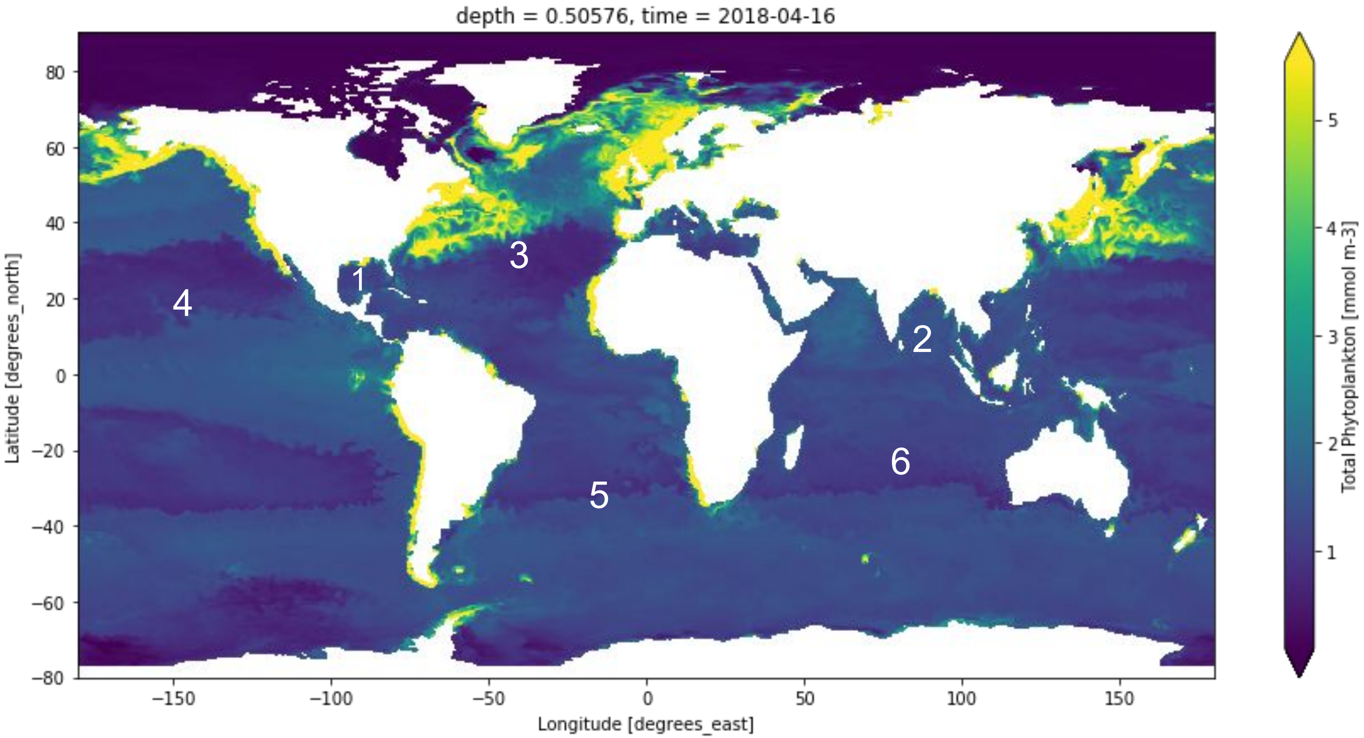
- The current ocean colour sensors provide an excellent capability to monitor the global phytoplankton distribution, which helps to understand the various physical and biogeochemical, processes at local and global scales.
- Changes in the water colour can be seen from the space and contributing factors can be remotely monitored [3-4].

- Several works for remote monitoring of ocean colour have been performed but machine learning (ML) and deep learning (DL) techniques are not much explored to derive the phytoplankton levels.
- Since ML and DL based approaches have been widely used in various research fields; hence, these approaches can also be used to study the ocean colour of the water bodies [5-6].

This study aims:

- i) to develop learning models utilizing the Copernicus Global Ocean Biogeochemistry Hindcast and Physical GLOBAL REANALYSIS dataset for 6 different locations and
- ii) to remotely predict phytoplankton volumes based on other parameters using ML algorithms. We have used 4 algorithms, namely Random Forest, Bagging, Extra Trees and Histogram based Gradient Boosting Regressor (HGBR) [7-8].

Introduction



1. Gulf of Mexico
2. Bay of Bengal
3. North Atlantic
4. North Pacific
5. South Atlantic
6. Indian Ocean

Motivation

- Earlier, several works have been performed to estimate phytoplankton quantities with remote sensing utilizing electromagnetic reflectance property for different wavelengths.
- Other process includes in-situ observations but that limits the scalability of the observation and also the process is very slow and resource consuming.
- Other process includes threshold-based decision-making systems making the algorithm fail to predict when exposed to an uncontrolled environment.
- To best of our knowledge and the literature survey, we found that there is no study available to derive the phytoplankton levels using the Copernicus reanalysis datasets (Global Ocean Biogeochemistry Hindcast).
- Therefore, to solve this gap, we have introduced supervised machine learning regression algorithms to estimate phytoplankton levels using reanalysis data of oceanographic properties.

Methodology

01 Data Collection

We have used a widely accepted Copernicus open-source datasets named Global Ocean Biogeochemistry Hindcast GLOBAL REANALYSIS BIO 001 029 monthly, which contained biochemical records, and Global Re-analysis Phy 001 030 monthly [16].

02 Study Location

- 22 °N to 28°N and 95°W to 85°W (Gulf of Mexico)
- 5°N to 15°N and 82°E to 92°E (Bay of Bengal)
- 20°N to 30°N and 60°W to 50°W (North Atlantic)
- 30°N to 40°N and 160°W to 150°W (North Pacific)
- 50°S to 40°S and 0°E to 10°E (South Atlantic)
- 20°S to 10°S and 80°E to 90°E (Indian Ocean)

03 Parameters

Surface CO_2, Dissolved Oxygen, Nitrate, Phosphate, Dissolved Silicate, pH, Salinity, Dissolved Iron, Temperature, etc.

Methodology

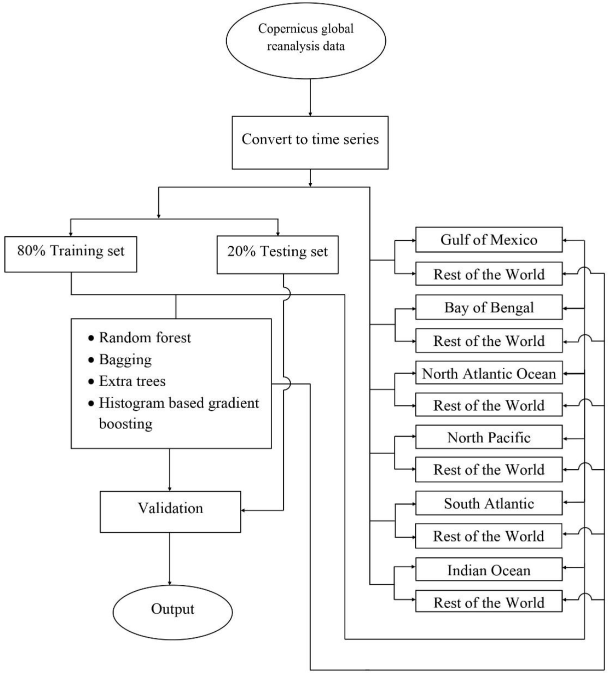
Regression

Estimation

Performance Comparison

Different supervised machine learning algorithm for regression have been used:

- Random Forest Regressor (RFR)
- Bagging Regressor (BGR)
- Extra Trees Regressor (ETR)
- Histogram Based Gradient Boosting Regressor (HBGBR)



Copernicus global reanalysis data

Convert to time series

80% Training set

20% Testing set

- Random forest
- Bagging
- Extra trees
- Histogram based gradient boosting

Validation

Output

Gulf of Mexico

Rest of the World

Bay of Bengal

Rest of the World

North Atlantic Ocean

Rest of the World

North Pacific

Rest of the World

South Atlantic

Rest of the World

Indian Ocean

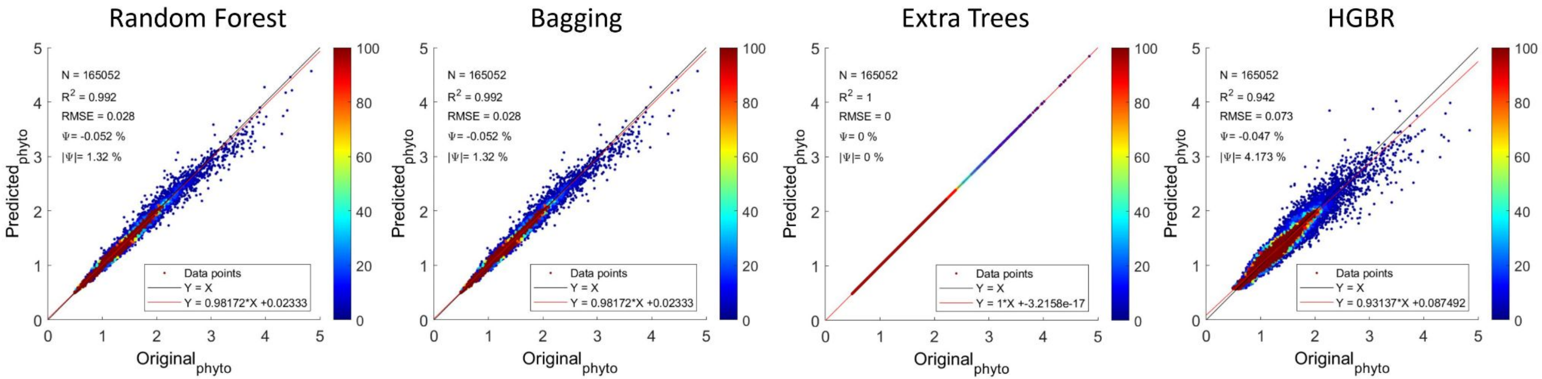
Rest of the World

The performance of these regressors have been measured by different metrics as follows:

- Mean Squared Error (MSE)
- Median Absolute Error (MAE)
- Determinant of Coefficient (R2)
- Explained Variance (EV)
- Max Error (ME)
- Mean Squared Log Error (MSLE)
- Mean Poisson Deviance (MPD)
- Mean Gamma Deviance (MGD)
- Mean Tweedie Deviance (MTD)

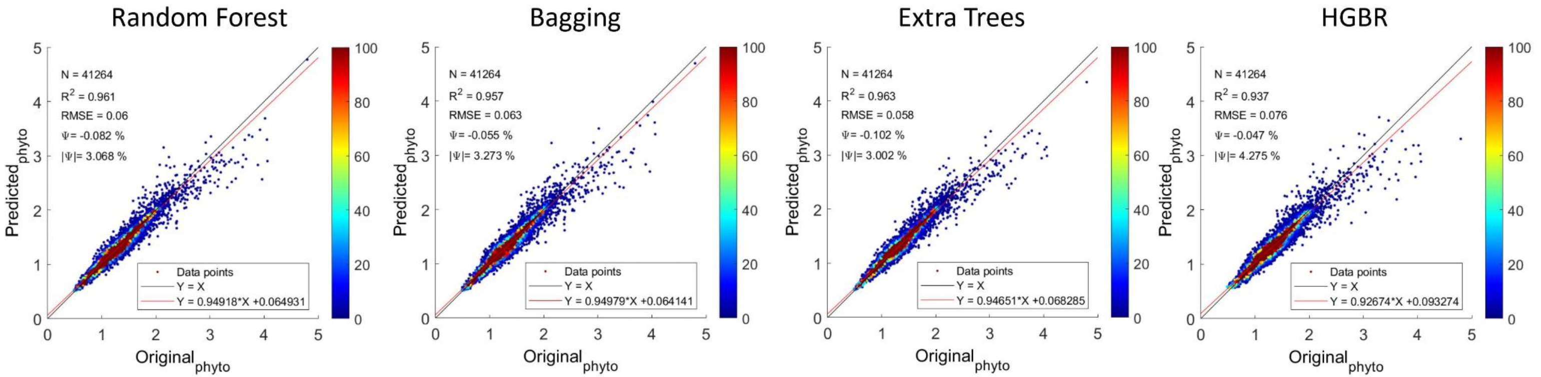
Results

Prediction on seen data



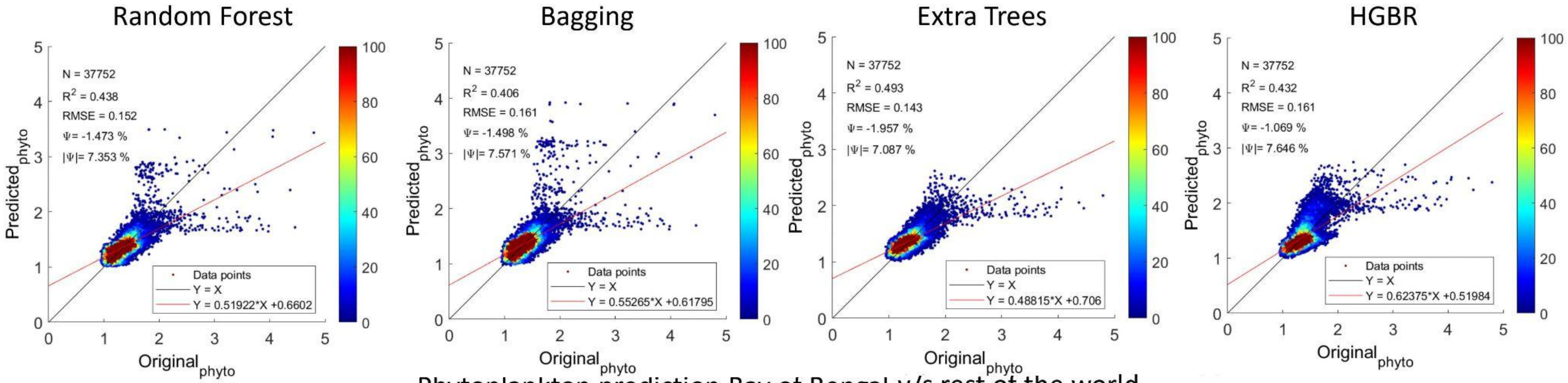
Phytoplankton prediction in global waters

Prediction on unseen data

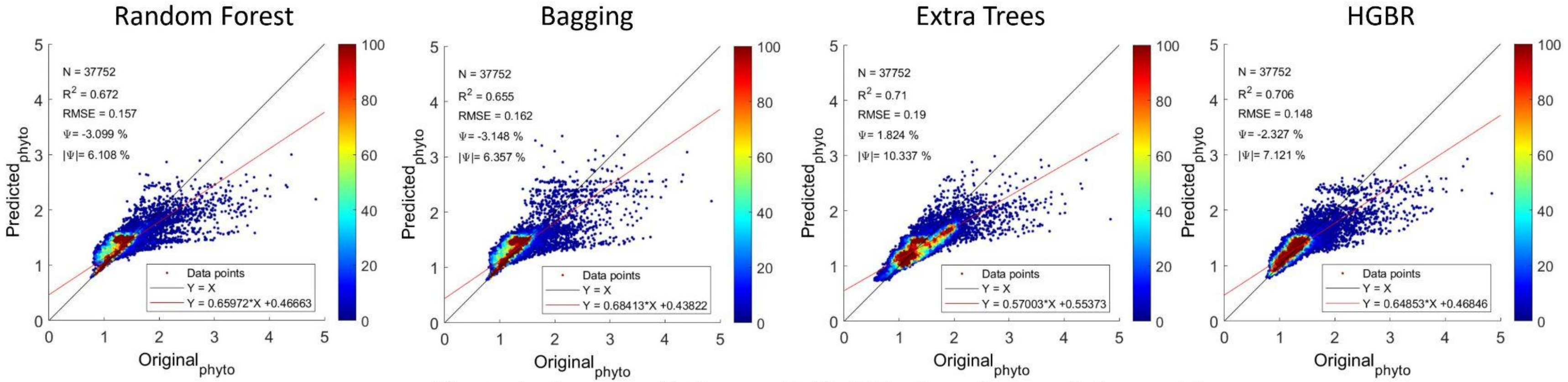


Phytoplankton prediction in global waters

Results

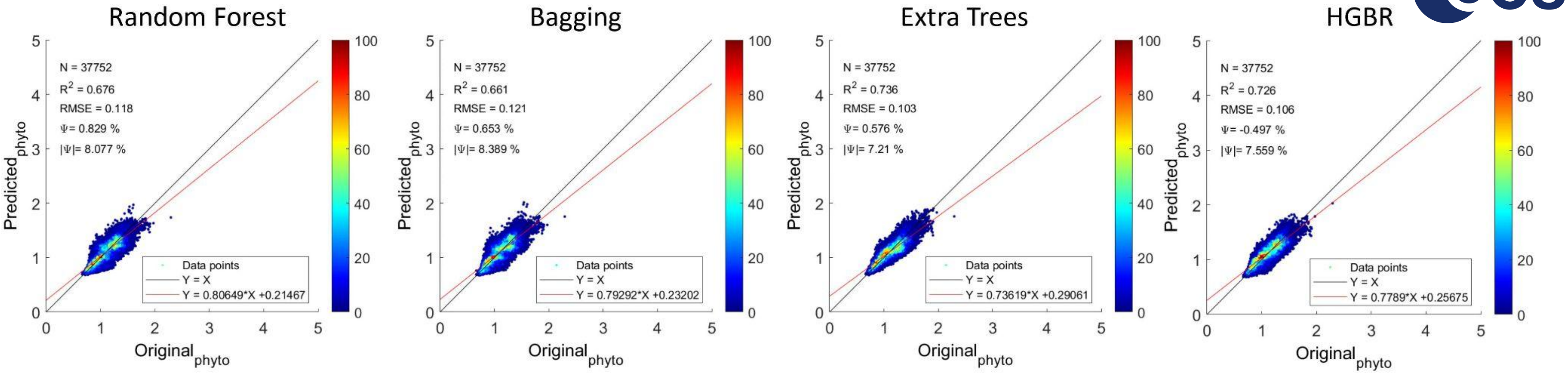


Phytoplankton prediction Bay of Bengal v/s rest of the world



Phytoplankton prediction on Gulf of Mexico v/s rest of the world

Results



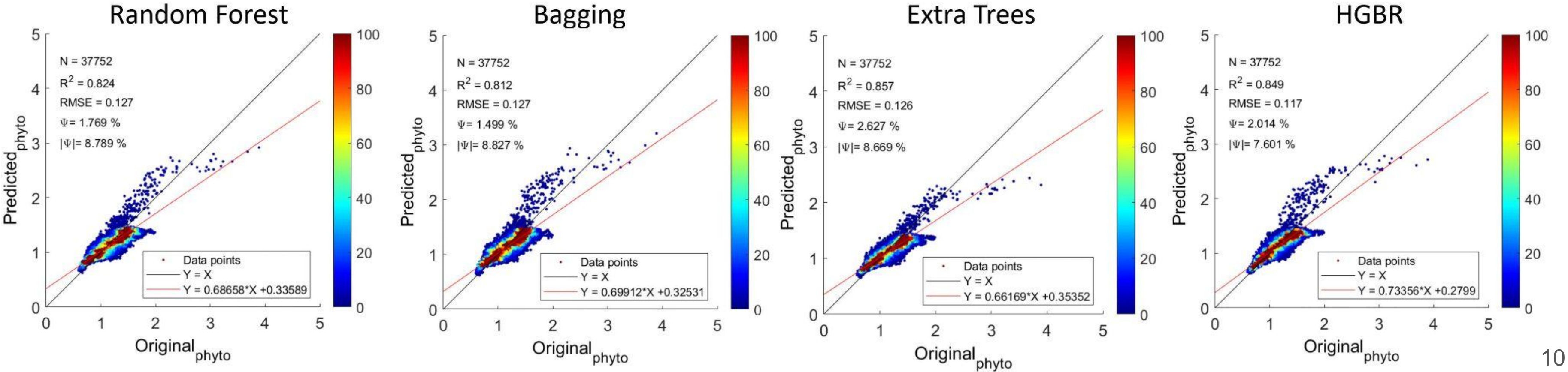
Random Forest

Bagging

Extra Trees

HGBR

Phytoplankton prediction on Indian Ocean v/s rest of the world



Random Forest

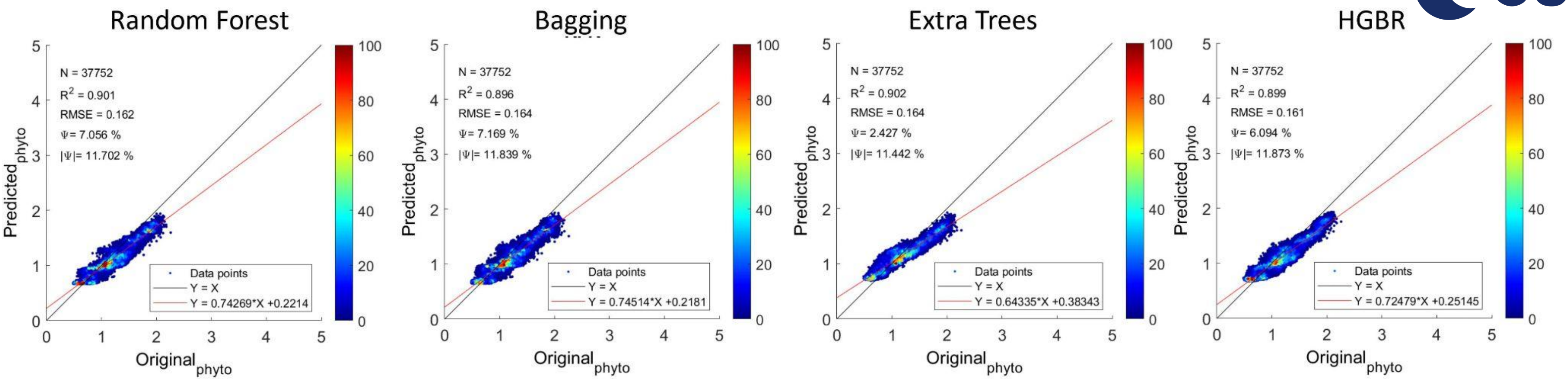
Bagging

Extra Trees

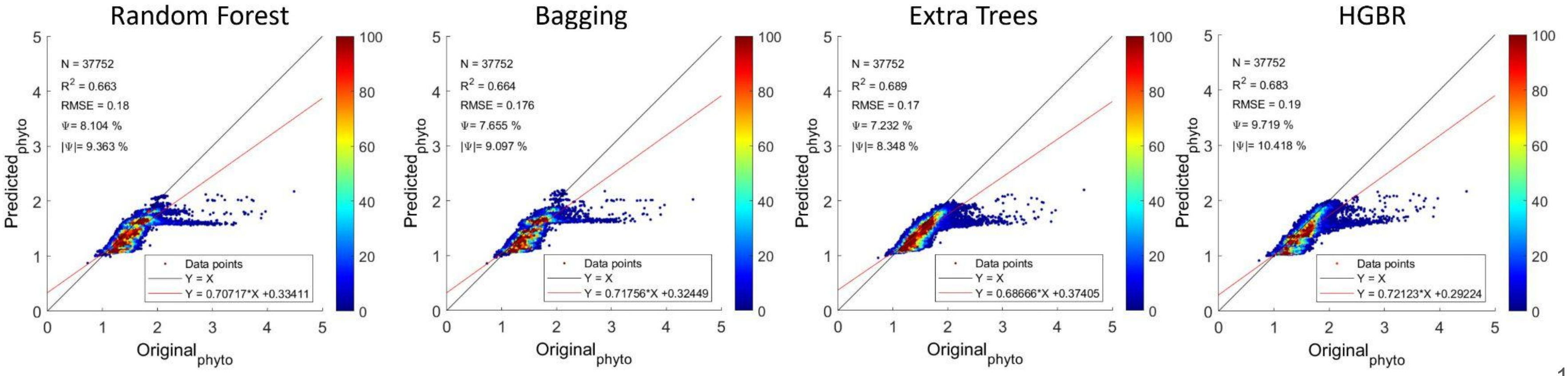
HGBR

Phytoplankton prediction North Atlantic Ocean v/s rest of the world

Results



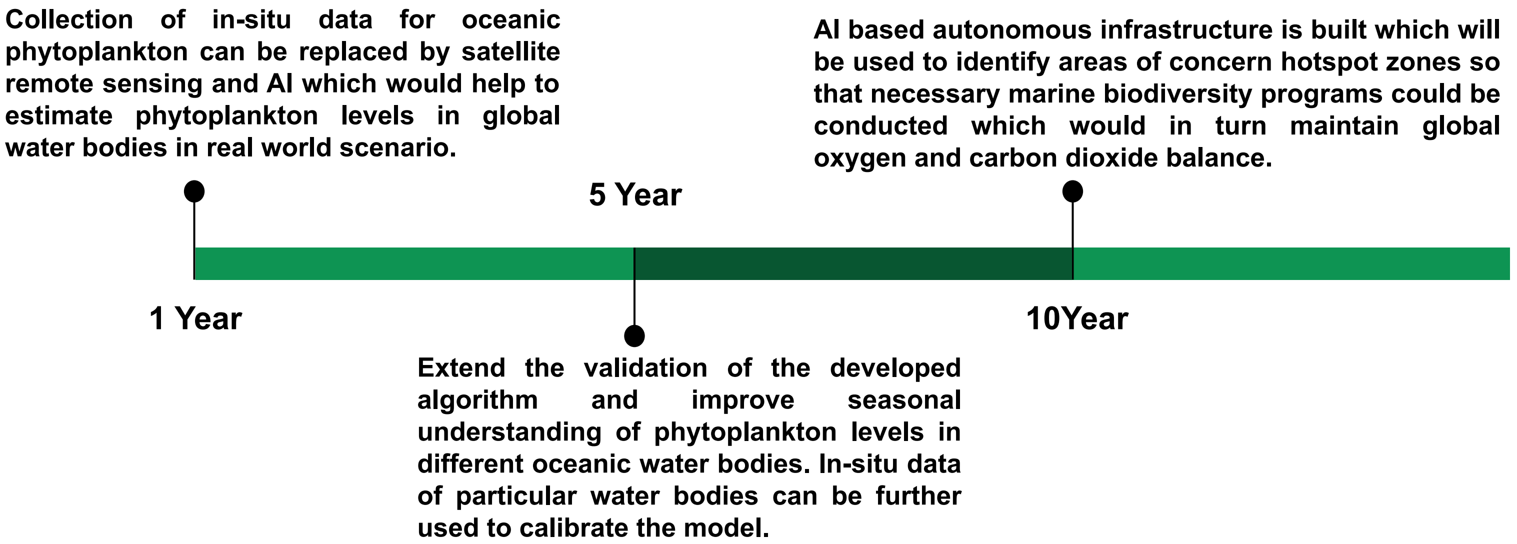
Phytoplankton prediction on North Pacific Ocean v/s rest of the world



Phytoplankton prediction on South Atlantic Ocean v/s rest of the world

Knowledge gap & priorities for next steps

- The study has utilized reanalysis data for training and validating the models, however, in-situ data could be used for the further calibration of the models.
- The work have been conducted using biogeochemical data from 6 parts around the world and therefore more data could be added for enhancing the model performance.



1 Year

Collection of in-situ data for oceanic phytoplankton can be replaced by satellite remote sensing and AI which would help to estimate phytoplankton levels in global water bodies in real world scenario.

5 Year

Extend the validation of the developed algorithm and improve seasonal understanding of phytoplankton levels in different oceanic water bodies. In-situ data of particular water bodies can be further used to calibrate the model.

10Year

AI based autonomous infrastructure is built which will be used to identify areas of concern hotspot zones so that necessary marine biodiversity programs could be conducted which would in turn maintain global oxygen and carbon dioxide balance.

Conclusion

- It has been observed that the Extra Trees regressor performed best for remote estimation of phytoplankton with an R2 score of 0.963 but took a considerable amount of time to train and generate results in 103.4s and 2.42s respectively.
- To understand the effects of variation of biogeochemical distribution pattern worldwide, the trained model has been tested with locations whose data have not been used for training and by this method, it has been understood that some oceans and sea have almost the same properties compared to the remaining world but some seas and oceans have much different biogeochemical distribution.
- To make the model adapt to these variations and overcome underfitting, portions of data from these locations have been included in the training data which effectively addressed the underfitting problem.
- The model would help to understand the depletion of phytoplankton levels where in situ measurements are easily not available.
- Based on the identified knowledge gap & priorities for a decade, this study models performance could be further improved by calibrating with in-situ data from the various parts of the world.

References

[1] K. Blix, J. Li, P. Massicotte, and A. Matsuoka, “Developing a new machine-learning algorithm for estimating chlorophyll-a concentration in optically complex waters: A case study for high northern latitude waters by using sentinel 3 olci,” Remote Sensing, vol. 11, no. 18, 2019.

[2] S. Adhikary, S. K. Chaturvedi, S. Banerjee, and S. Basu, “Dependence of physiochemical features on marine chlorophyll analysis with learning techniques,” in Advances in Environment Engineering and Management (N. A. Siddiqui, K. D. Bahukhandi, S. M. Tauseef, and N. Koranga, eds.), (Cham), pp. 361–373, Springer International Publishing, 2021.

[3] K. Blix and T. Eltoft, “Evaluation of feature ranking and regression methods for oceanic chlorophyll-a estimation,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 5, pp. 1403–1418, 2018.

[4] V. Sagan, K. T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A. Greeling, S. Maalouf, and C. Adams, “Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing,” Earth-Science Reviews, vol. 205, p. 103187, 2020.

[5] H. Soydan, A. Koz, and H. S ¸ ebnem D ¨uzg ¨un, “Secondary iron mineral detection via hyperspectral unmixing analysis with sentinel-2 imagery,” International Journal of Applied Earth Observation and Geoinformation, vol. 101, p. 102343, 2021.

[6] M. Ilteralp, S. Ariman, and E. Aptoula, “A deep multitask semisupervised learning approach for chlorophyll-a retrieval from remote sensing images,” Remote Sensing, vol. 14, no. 1, 2022.

[7] M. Demetriou, D. E. Raitsos, A. Kournopoulou, M. Mandalakis, S. Sfenthourakis, and S. Psarra, “Phytoplankton phenology in the coastal zone of cyprus, based on remote sensing and in situ observations,” Remote Sensing, vol. 14, no. 1, 2022.

[8] J. Shi, Q. Shen, Y. Yao, J. Li, F. Chen, R. Wang, W. Xu, Z. Gao, L. Wang and Y. Zhou, “Estimation of chlorophyll-a concentrations in small water bodies: Comparison of fused gaofen-6 and sentinel-2 sensors,” Remote Sensing, vol. 14, no. 1, 2022.

[9] E. K. Cherif, P. Mozeti ˇc, J. Franc ́e, V. Flander-Putrle, J. Faganeli-Pucer, and M. Vodopivec, “Comparison of in-situ chlorophyll-a time series and sentinel-3 ocean and land color instrument data in slovenian national waters (gulf of trieste, adriatic sea),” Water, vol. 13, no. 14, 2021.

[10] S. Hu, W. Zhou, G. Wang, W. Cao, Z. Xu, H. Liu, G. Wu, and W. Zhao, “Comparison of satellite-derived phytoplankton size classes using in-situ measurements in the south china sea,” Remote Sensing, vol. 10, no. 4, 2018.

[11] J.-E. Park, K. Park, Y.-J. Park, and H.-J. Han, “Overview of chlorophyll-a concentration retrieval algorithms from multi-satellite data,” Journal of the Korean earth science society, vol. 40, no. 4, pp. 315–328, 2019.

[12] H. F. Houskeeper, S. B. Hooker, and R. M. Kudela, “Spectral range within global acdom(440) algorithms for oceanic, coastal, and inland waters with application to airborne measurements,” Remote Sensing of Environment, vol. 253, p. 112155, 2021.

[13] M. Hieronymi, D. M ¨uller, and R. Doerffer, “The olci neural network swarm (onns): A bio-geo-optical algorithm for open ocean and coastal waters,” Frontiers in Marine Science, vol. 4, 2017.

[14] S. Li, K. Song, S. Wang, G. Liu, Z. Wen, Y. Shang, L. Lyu, F. Chen, S. Xu, H. Tao, Y. Du, C. Fang, and G. Mu, “Quantification of chlorophyll-a in typical lakes across china using sentinel-2 msi imagery with machine learning algorithm,” Science of The Total Environment, vol. 778, p. 146271, 2021.

[15] K. Blix and T. Eltoft, “Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval,” Remote Sensing, vol. 10, no. 5, 2018.

[16] C. Perruche, “Product user manual for the global ocean biogeochemistry hindcast global reanalysis bio 001 029.
version 1.,” 2018.

Thank You For Your Kind Attention!