This is a preprint of the paper published in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*,
DOI: 10.1109/IGARSS53475.2024.10640797

SELECTIVE FILTERING FOR ENHANCING CHLOROPHYLL RETRIEVAL ACCURACY FROM SENTINEL-3 DATA USING RANDOM FOREST MODELS

Pankaj Patidar

AGFE, IIT Kharagpur, 721302 Kharagpur, India
TUM School of Engineering and Design, Germany

Subhadip Dey

AGFE, IIT Kharagpur, 721302 Kharagpur, India
German Aerospace Center (DLR), Germany

Dmitry Efremenko

Remote sensing technology institute (IMF)
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany

Efrain Padilla-Zepeda

Remote sensing technology institute (IMF)
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany

ABSTRACT

This paper presents an investigation into the use of a random forest (RF) model for retrieving chlorophyll content from Sentinel-3 satellite data. We train various RF regression models on available datasets and introduce a classifier to identify instances where predictions may be inaccurate. This classifier aids in filtering out less reliable cases, enhancing the overall accuracy of our models at the expense of reducing the amount of processed data. Additionally, we optimize the hyperparameters of this hybrid model to improve its performance further. Our findings illustrate the effectiveness of combining regression models with a classifier in environmental remote sensing, offering a promising method for improving the accuracy of satellite-derived chlorophyll measurements.

Index Terms— chlorophyll retrieval, neural networks, selective filtering, random forest regression, Sentinel-3

1. INTRODUCTION

Satellite-based remote sensing is a powerful tool for monitoring environmental variables, offering a comprehensive perspective on Earth's vital parameters. Among these, the Sentinel-3 mission stands out for its ability to capture high-resolution multispectral imagery, providing a wealth of information for diverse applications, including aquatic ecosystem monitoring. In particular, the Sentinel-3 data can be used for retrieval of water chlorophyll concentrations. The concentration of chlorophyll-a is regarded as a crucial parameter in assessing water quality, given its significant role in the eutrophication process [1, 2]. Eutrophication can result in serious consequences for aquatic ecosystems, including an escalation in hypoxia, fish mortality, and the emergence of harmful algae blooms [3].

The chlorophyll in water causes the absorption peak near 440 nm and 670 nm, as well as the strong reflection peak at around 550 nm. The chlorophyll content can be obtained using the optimal estimation method [4] that matches the spectral measurements to the bio-optical forward model [5] providing the optical properties of the water containing chlorophyll. This approach may require the linearization of the coupled atmosphere-ocean radiative transfer model equipped with the corresponding bio-optical model [6]. An alternative approach is referred to as the empirical approach and consists of finding regression models between the radiances and chlorophyll concentration. To improve the accuracy of such models, several authors propose to combine radiances (or reflectances) from different bands to make an artificial parameter that is strongly correlated with chlorophyll concentration. For instance, in [7] the normalized difference chlorophyll index (NDCI) was proposed and its formulation includes computing the normalized difference between reflectance values at 708 nm and 665 nm, followed by normalization using the sum of the reflectance values at these wavelengths. In the Mediterranean Ocean Color 4 (MedOC4) models, the logarithm of chlorophyll concentration is expressed as a fourth-degree polynomial of maximum ratios of some bands (see [8] and references therein). The polynomial coefficients are found in the calibration procedure. This approach can be generalized by using artificial neural networks. In [9] a neural network was considered that takes as input different ratios of bands. For their model, authors achieved the root-mean-square error and unbiased percentage difference of 0.13 mg/m3, and 17.31%, respectively. The advantage of this approach is that we do not need the forward model, but rather an extensive training dataset consisting of on-site measurements of chlorophyll concentrations and corresponding sensor sig-

nal onboard a satellite.

In this paper, our goal is to improve the accuracy of the supervised learning models by detecting the cases (pixels) where our random forest model is expected to exhibit suboptimal performance. A classifier is employed that discriminates between "reliable" and "unreliable" pixels, and the random forest model is then applied only for 'reliable' pixels ensuring a more robust and reliable assessment of water quality. Our intention is to investigate the trade-off between the accuracy of retrieval and the amount of pixels that are eligible for processing according to the classifier.

2. DATA COLLECTION

The study region encompasses the Mediterranean Sea, acknowledged as a mid-latitude basin predominantly characterized by oligotrophic and ultra-oligotrophic conditions. Our investigation scrutinized two primary datasets: in situ chlorophyll-a concentrations and satellite data. The in-situ measurements were derived from a dataset provided by the Copernicus Marine Environmental Monitoring Service (CMEMS). This dataset comprises a wealth of ocean bio-optical information, in particular, chlorophyll-a concentration Figure 1 provides a visual representation of the study area, delineating the specific locations of in situ measurements.

In alignment with the in-situ data points from the CMEMS dataset, we gathered the peak radiance values for each band from the Ocean and Land Cover Instrument (OLCI) instrument on the Sentinel-3 (COPERNICUS/S3/OLCI) satellite. Access to these data was facilitated through Google Earth Engine by specifying the acquisition time and location. In the dataset, we include Sentinel-3 OLCI bands. The chlorophyll range over the considered region is between 0 to 1.934 (mg m^3). In total, we have collected 11098 cases.

3. METHODOLOGY

Our methodology is illustrated in Figure 2. It includes a regressor and a classifier to enhance the accuracy of chlorophyll prediction from radiance data. Initially, we use radiance measurements alongside ground truth data on chlorophyll concentrations to train a regression model. This regressor is designed to establish a direct correlation between the input radiance and the corresponding chlorophyll levels. We use the random forest regressor that takes as input all bands of Sentinel 3 available in the dataset.

Then, we validate the trained regressor to assess its performance and select cases where the regressor prediction has an error exceeding a given threshold. In the next step, we train a classifier, that pinpoints instances where the regressor is prone to significant errors. We label these instances as 'unreliable'. Thus, the classifier acts as a filtering mechanism for new data, segregating it into 'reliable' and 'unreliable' categories based on the likelihood of accurate predictions by the regressor. For new data classified as 'reliable', the regressor is then applied to estimate the chlorophyll levels. By selectively applying the regressor only to data classified as 'reliable', we expect to reduce the potential for large errors in our chlorophyll estimations, leading to more accurate and dependable results.

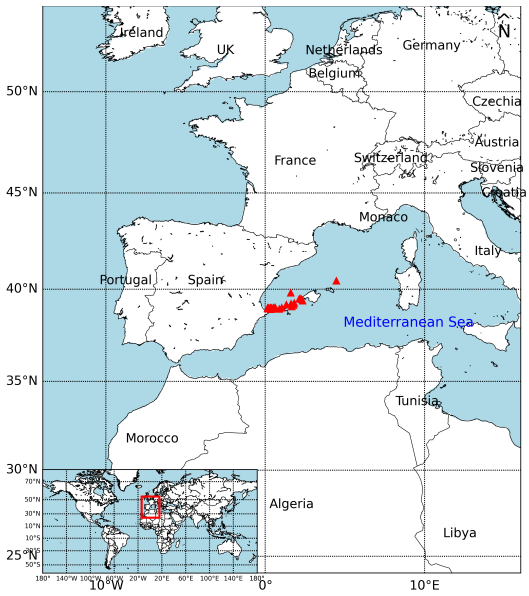


Fig. 1: Location of the Study area and data points

4. RESULTS AND DISCUSSION

We apply the random forest regression model with a secondary classifier. The results are compared with those obtained with NDCI and MedOC4 algorithms. Table 1 shows the Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Mean Bias Error (MBE), and Absolute Percentage Difference (APD). NDCI characterized by a respectable R^2 of 0.7971, presents, however, a relatively higher RMSE (0.0974) and APD of 83.6476%, indicating noticeable errors. Similarly, the MedOC4 model, closely aligned with NDCI, demonstrates a slightly reduced APD of 22.03% with RMSE (0.1004). Notably, the RF model, while not surpassing the performance of NDCI, is also characterized by competitive metrics.

Contrastingly, the RF model combined with a secondary classifier, featuring a confidence level of 0.9 and a threshold of 0.1, outshines all other approaches. It achieves the lowest RMSE of 0.05993, the highest R^2 of 0.8868, and a negligible Mean Bias Error (MBE) of 0.0005.

However, the application of the classifier inevitably reduces the number of cases to which the regressor is applied. In this context, an essential parameter that characterizes the performance of our approach, in addition to accuracy, is completeness, i.e. the proportion of cases that are deemed suit-

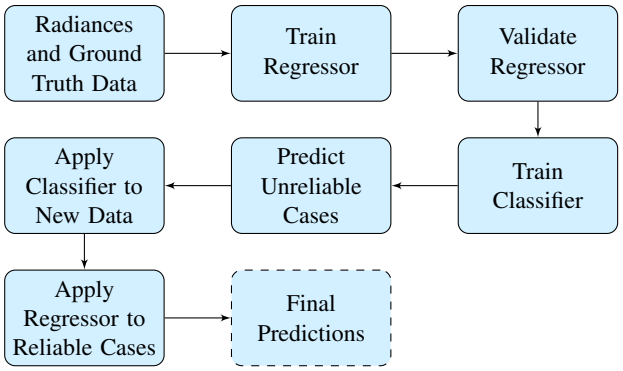


Fig. 2: Schematic representation of the proposed approach: after the regression model is trained, it is evaluated and the classifier is trained to detect the cases when the model is expected to have an error exceeding a given threshold.

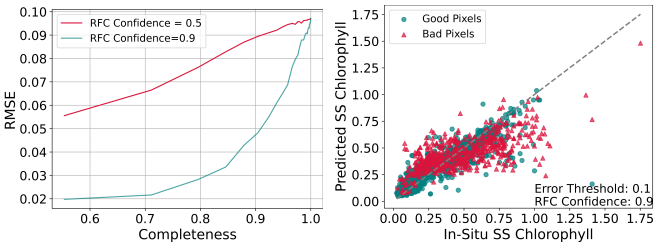
Table 1: Comparative analysis of random forest (RF) regression, RF with a secondary classifier, NDCI, and MedOC4 algorithm

	RMSE	R^2	MBE	APD
RF+Classifier	0.05993	0.8868	0.0005	18.15
RF	0.0992	0.7893	0.0038	22.08
NDCI	0.0974	0.7971	0.0041	83.65
MedOC4	0.1004	0.6850	0.0043	22.03

able by the classifier for regression analysis. To evaluate our model, we consider the error-completeness curve, which illustrates the trade-off between accuracy and dataset coverage. Note that the classifier assigns a probability (confidence) to a pixel, that it belongs to a specific class. We may set a threshold for confidence to adjust the curve, as shown in Figure 3a. It visually demonstrates how adjustments of the confidence level impact both the error rate and the proportion of data suitable for regression analysis. This curve helps us identify an optimal balance between accuracy and completeness. For instance, this figure shows that by setting the confidence level at 0.9 and marking the results with RMSE larger than 0.03 as unreliable we are able to maintain of about 80% of cases for retrieval.

The concept behind utilizing a classifier in conjunction with our regression model is to identify and flag problematic cases where the regression model is anticipated to yield inaccurate results. The root causes of these inaccuracies can vary. They may be due to limitations of the regression model itself (e.g. overfitting, underfitting, etc) or external factors, such as high aerosol loading that can significantly interfere with the signal received by the satellite sensors, leading to distorted or obscured data.

Figure 3b shows the results derived from implementing a pipeline that integrates a classifier with a random forest regressor. The classifier tends to exclude those instances in which the model is likely to incur large errors. This selective process is crucial in maintaining the overall precision of our predictions, as it effectively identifies and sets aside the more problematic cases that could potentially skew the accuracy of the random forest regressor.



(a) (b)

Fig. 3: (a) Completeness vs RMSE curve at the two different confidence level of random forest classifier, and (b) Predicted vs. true chlorophyll concentration values from the testing dataset. Points are color-coded to differentiate cases classified as 'reliable' or 'unreliable' based on the expected error level

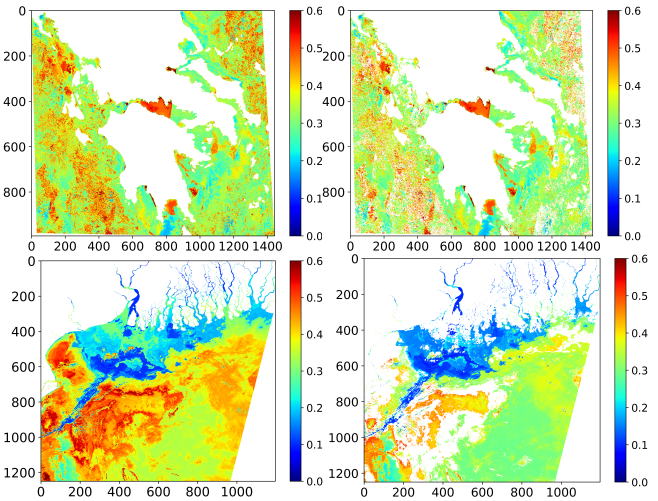


Fig. 4: Chlorophyll Retrieval heat map generated using RFRC algorithm, Top Panel: Mediterranean Sea around Greece. Bottom Panel: Bay of Bengal, Sundarban Delta region. Left Panel: Without filtering, Right Panel: After filtering out non-reliable cases

Finally, we demonstrate the algorithm by applying it to the region in the Mediterranean Sea near Greece and in the Sundarban delta region of Bangladesh and India, as shown in Figure 4. The model's results generally exhibit a spatial pattern similar to those found by other algorithms available in the literature. The distinct edge in the middle of the figure is likely due to the sensor's varying measurement times. Employing a classifier and filtering out unreliable cases de-

creases the overall completeness of the image. Additionally, in this specific instance, the classifier tends to label cases with relatively high chlorophyll values as unreliable. This problem can be eliminated by including in the training dataset more cases with high values of chlorophyll concentration.

5. CONCLUSION

In this paper, we have considered a methodology to elevate the precision of chlorophyll retrieval from Sentinel-3 satellite data. It is based on a regression model with a secondary classifier, designed to identify and exclude instances where the regression model exhibits significant errors.

In the comparative analysis, our methodology, featuring a secondary classifier, emerges as the superior approach, evidenced by a notable reduction in RMSE and a substantial improvement in the coefficient of determination. Despite this success, the introduction of the classifier necessitates a crucial trade-off between accuracy and completeness. Our analysis of the error-vs-completeness curve illuminates the impact of adjusting filtering criteria on both the accuracy of predictions and the coverage of the dataset.

In summary, the proposed methodology combining regression models with selective filtering through a classifier offers a promising tool for enhancing the accuracy of satellite-derived chlorophyll measurements. Further research and refinement of this approach could contribute significantly to the field of environmental remote sensing.

Acknowledgments

We want to acknowledge the Copernicus Marine Environment Monitoring Service (CMEMSE) for providing access to the essential in-situ dataset that served as the cornerstone of our research. Also, freely accessible sentinel 3 data made this research possible. Mr. Patidar thanks the DAAD (German Academic Exchange Service) for providing the fellowship that made this collaborative research possible, and Dr. Dey would like to thank GISEhub IIT Bombay for their support through project receipt grant No. 3/69/2021/NSDI(G)(E.No.36454)/GISE2023/0001018773-01.

6. REFERENCES

[1] Ioannis Moutzouris-Sidiris, Konstantinos Topouzelis, and Evangelia Efi Konstantinidou, "Assessment of chlorophyll-a concentration derived from sentinel-3 satellite images using open source data," in *Seventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2019)*, Giorgos Papadavid, Kyriacos Themistocleous, Silas Michaelides, Vincent Ambrosia, and Diofantos G. Hadjimitsis, Eds. June 2019, SPIE.

[2] Weining Zhu, Qian Yu, Yong Q. Tian, Brian L. Becker, and Hunter Carrick, "Issues and potential improvement of multiband models for remotely estimating chlorophyll-a in complex inland waters," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 8, no. 2, pp. 562–575, Feb. 2015.

[3] João G. Ferreira, Jesper H. Andersen, Angel Borja, Suzanne B. Bricker, Jordi Camp, Margarida Cardoso da Silva, Esther Garcés, Anna-Stiina Heiskanen, Christoph Humborg, Lydia Ignatiades, Christiane Lancelot, Alain Menesguen, Paul Tett, Nicolas Hoepffner, and Ulrich Claussen, "Overview of eutrophication indicators to assess environmental status within the european marine strategy framework directive," *Estuar Coast Shelf Sci*, vol. 93, no. 2, pp. 117–131, June 2011.

[4] Knut Stamnes, Wei Li, R. Spurr, and Jakob J. Stamnes, "Simultaneous retrieval of aerosol and coastal ocean properties by optimal estimation," in *AIP Conference Proceedings*. 2009, American Institute of Physics.

[5] Daniel Odermatt, Anatoly Gitelson, Vittorio Ernesto Brando, and Michael Schaepman, "Review of constituent retrieval in optically deep and complex waters from satellite imagery," *Remote Sens Environ*, vol. 118, pp. 116–126, Mar. 2012.

[6] Robert Spurr, Knut Stamnes, Hans Eide, Wei Li, Kexin Zhang, and Jakob Stamnes, "Simultaneous retrieval of aerosols and ocean properties: A classic inverse modeling approach. i. analytic jacobians from the linearized cao-disort model," *J Quant Spectrosc Radiat Transf*, vol. 104, no. 3, pp. 428–449, Apr. 2007.

[7] Sachidananda Mishra and Deepak R. Mishra, "Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters," *Remote Sens Environ*, vol. 117, pp. 394–406, Feb. 2012.

[8] Rosalia Santoleri, Gianluca Volpe, Salvatore Marullo, and B Buongiorno Nardelli, "Open waters optical remote sensing of the mediterranean sea," in *Remote sensing of the European seas*, pp. 103–116. Springer, 2008.

[9] Guiying Yang, Xiaomin Ye, Qing Xu, Xiaobin Yin, and Siyang Xu, "Sea surface chlorophyll-a concentration retrieval from hy-1c satellite data based on residual network," *Remote Sensing*, vol. 15, no. 14, pp. 3696, July 2023.