A Model for Predicting the Amount of Photosynthetically Available Radiation from BGC-ARGO Float Observations in the Water Column

Frederic Stahl[1], Lars Nolle[1,2], Ahlem Jemai[3], Oliver Zielinski[1,3]

[1]German Research Center for Artificial Intelligence (DFKI), Marine Perception
Oldenburg, Germany
Email: {Frederic_theodor.stahl | Lars.Nolle,Oliver.Zielinski}@dfki.de
[2]Jade University of Applied Science, Department of Engineering Sciences
Wilhelmshaven, Germany
Email: Lars.Nolle@jade-hs.de
[3]Carl von Ossietzky University of Oldenburg, Institute for Chemistry and Biology of the Marine Environment
Oldenburg, Germany
Email: {Ahlem.Jemai | Oliver.Zielinski}@uol.de

KEYWORDS

Machine Learning, BGC-Argo Floats, Underwater light field, PAR, Downwelling Irradiance

ABSTRACT

Modern oceanography uses, amongst other platforms, automated diving devices, which are drifting with the ocean current whilst continuously collecting vertical profiles of environmental parameters. One of the important parameters is photosynthetically available radiation (PAR). It was studied in this work whether the PAR values can be reconstructed by combinations of measurements from the remaining onboard sensors with specific wavelength. If a reconstruction of PAR is possible, this would allow allocating the sensor with a further specific wavelength instead of PAR. Having available more spectral information could for example enable natural scientists to better distinguish phytoplankton or UV radiation. Therefore, data from three different expeditions from different regions of the world were used to model PAR using multiple linear regression and regression trees (RT). Multiple linear regression achieved an R^2 value of 0.970 for the combined dataset and RT achieved an R^2 value of 0.960. Hence, the models are accurate enough to predict the PAR parameter without the need for a dedicated PAR sensor. Thus the PAR sensor reading could be replaced with measurements of an additional wave length.

INTRODUCTION

Modern operational oceanography uses a plethora of different autonomous platforms [1]. Among them, the nearly 4000 Argo floats [2], automated diving devices, drifting with the ocean current and collect continuous vertical profiles from a depth ~ 2000 m, evolved to be a core component. With Argo float data being transmitted via the Iridium or Argos satellite systems, data is publicly and freely available via two global data assembly centers (GDAC) typically within 24 hours (see Argo website https://argo.ucsd.edu).

While Argo started with a three sensor setup aiming at physical oceanographic information, there has been a significant increase in bio-optical instrumentation on Argo, leading to the biogeochemical Argo (short BGC-Argo) initiative [3]. Together with this increase in sensors, accompanied by the data management and quality control processes, demand for machine learning has been on the rise [3,4].

In this context, the BGC-Argo community suggested to re-configurate the Ocean Color Radiometer (OCR) to dismiss the fourth channel, originally designed to record PAR measurement, since this could potentially be reconstructed from the three available distinct channels, measuring wavelengths at 380 nm, 412 nm, and 490 nm. In this study, a machine learning approach is provided, that models the entire wavelength ranges of PAR from the three wavelengths. This enables including a further specific wavelength and thus increase the flexibility of the device [1].

RADIOMETRIC PROFILING FLOAT OBSERVATIONS

The underwater light field is one of the six essential variables measured by so-called BGC-Argo Floats [6]. Featuring the multispectral technology, the OCR-504 from SATLANTIC Inc./Sea-Bird Scientific, USA [7] is used to routinely measure the radiometric observation at four channels. Three channels 380 nm, 412 nm and 490 nm were selected as they are related to the main variations in underwater optical properties. The fourth channel is dedicated to measure PAR. Figure 1 shows the Argo APEX Float WMO7900562, deployment on 27th of September 2019 in the western Mediterranean, with attached sensors, including the radiometer (left) and the radiometer OCR-504 (right).

Communications of the ECMS, Volume 36, Issue 1,
Proceedings, ©ECMS Ibrahim A. Hameed, Agus Hasan,
Saleh Abdel-Afou Alaliyat (Editors) 2022
ISBN: 978-3-937436-77-7/978-3-937436-76-0(CD) ISSN 2522-2414


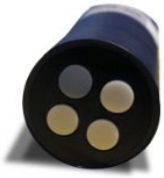

Figure1: Argo APEX platform with attached sensors, including

The PAR parameter is commonly used to disclose the overall light available for the primary production in natural waters and allows for the integration of downward irradiance between 400 nm and 700 nm. Recently, Jemai et al. [4] provided a review of radiometric measurements on Argo floats. They, as well as Organelli et al. [8], emphazised the need for more spectral information, from multi- to hyperspectral instrumentation. This platform provided the data that was used for the modelling as described below.

The data used in this study is publicly available at ftp://ftp.ifremer.fr/ifremer/argo/dac/coriolis. The dataset represents the German contribution within the BGC-Argo program. The data was acquired by four floats deployed at different sites, one (WMO 7900585) in the North Atlantic, one (WMO7900562) in the Mediterranean Sea, and two (WMO7900579 and WMO7900580) in the Baltic Sea. Radiometric observations were collected during the ascent phases every two or five days in the upper layer, and sampling was carried out at 2 dbar vertical resolution for all floats.

PRELIMINARY ANALYSIS AND PROCESSING OF THE DATA

Data from three different expeditions from different regions were used. From the Mediterranean Sea one dataset with 13,068 data instances was used; from the Baltic Sea two datasets were used, one with 1,373 data instances and the other with 1,274 data instances and Atlantic Ocean with 4,403 data instances. Some data instances contained a very small amount of missing values, these data instances were removed. Missing value are caused by malfunction of the float. In total, 20,079 instances were available after deletion of missing values.

In order to establish the correlation between the different sensors, a scatter matrix with all float datesets concatenated was plotted as can be seen in Figure 2. It can be seen that there is generally a good correlation between all sensors, except for P (pressure). What can also be seen is that for P below 100 dbar (equivalent to an approximate depth of 100 m), all sensors produce low values. The reason for this is that light at this depth is fully absorbed by the water. Since this is the case for all sensors, it has no effect on the correlation between all optical sensors. The fact that P does not correlate with the other sensors has subsequently also been confirmed with the Institute for Chemistry and Biology of the Marine Environment. Therefore, it was decided to exclude P from modelling.

MODELLING

Two different methods for modelling were selected, Multiple Linear Regression [9] and Regression Trees [10]. The reason for choosing these two techniques is that they produce predictive models for continuous target variables. All data instances were used for the modelling process with input variables downwelling irradiance at 380 nm, 412 nm, 490 nm and target variable PAR. 30% of the data instances were randomly selected without replacement to be included into a test set and the remaining instances were used to fit the models.

In total, ten models were produced, two for each float location, i.e. one Regression Tree and one Multiple linear regression equation, and two models for all float data combined.

Models based on Muliple Lineare Regression

For the modelling standard Muliple Lineare Regression [9] was used without forcing an intercept. In this paper the Scikit-learn implementation (https://scikit-learn.org/stable/) of Multiple linear regression was used. The results are presented in Figures 3-7. The figures plot the true PAR values versus the predicted PAR values.

Figure 3 shows the result of Muliple Lineare Regression using the combined dataset comprising all float locations.

The R^2 value was 0.97. The resulting model can be found in Equation 1.

$$PAR = 2582.06 * Ed380 - 1715.67 * Ed412 + 1023.94 * Ed490 - 1.57 \quad (1)$$

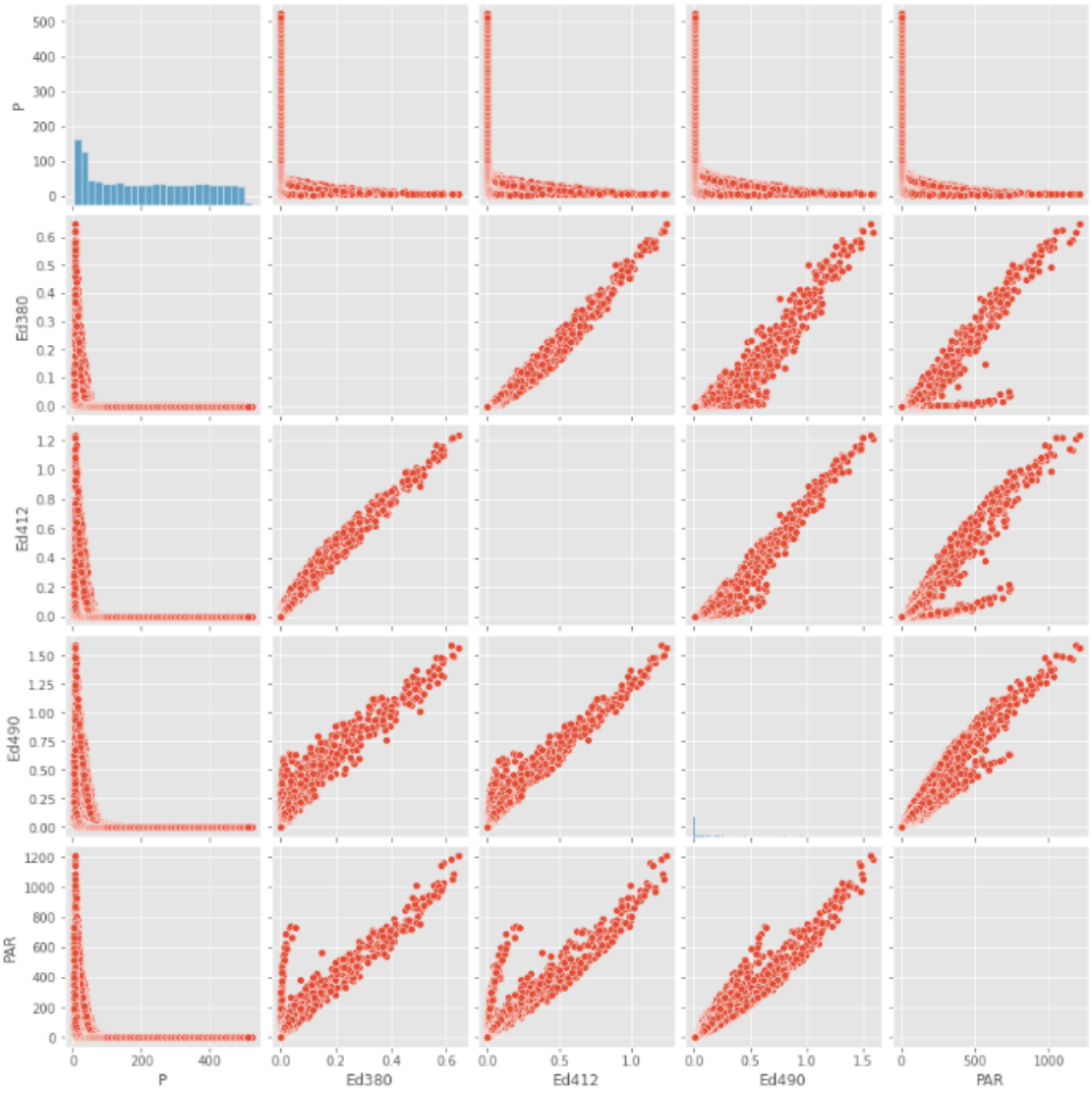

Figure 2: Dependency of sensors of the float

Figure 3: Multiple linear regression on all data

Figure 4 shows the result of multiple linear regression using the dataset comprising data for the Mediterranean Sea float location.

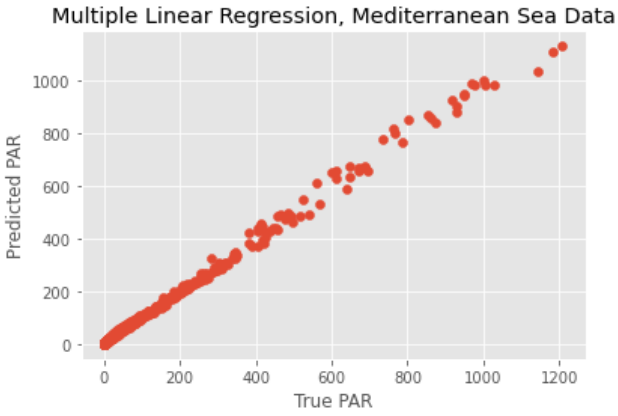

Figure 4: Multiple linear regression on Mediterranean Sea data

The R^2 value was 0.997. The resulting model can be found in Equation 2.

$$PAR = 1744.62 * Ed380 - 726.90 * Ed412 \quad (2)$$

+578.50*Ed490−1.14

Figure 5 shows the result of multiple linear regression using the dataset comprising data for the Baltic Sea float location 1.

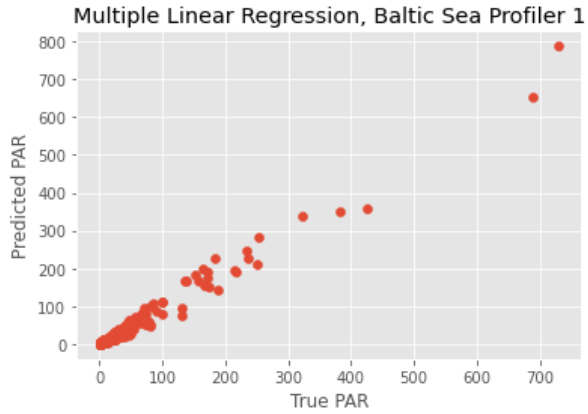

Figure 5: Multiple linear regression on Baltic Sea float 1

The R^2 value was 0.991. The resulting model can be found in Equation 3.

PAR=14321.34*Ed380
−2350.74*Ed412 (3)
+1168.52*Ed490+1.88

Figure 6 shows the result of multiple linear regression using the dataset comprising data for the Baltic Sea float location 2.

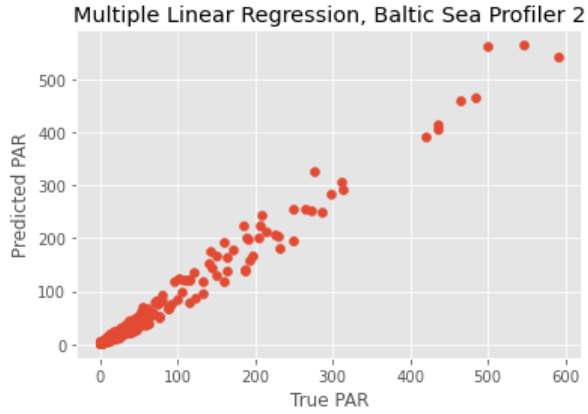

Figure 6: Multiple linear regression on the Baltic Sea float 2

The R^2 value was 0.983. The resulting model can be found in Equation 4.

PAR=3644.03*Ed380
−200.34*Ed412 (4)
+966.75*Ed490+1.44

Figure 7 shows the result of multiple linear regression using the dataset comprising data for the Atlantic Ocean float location.

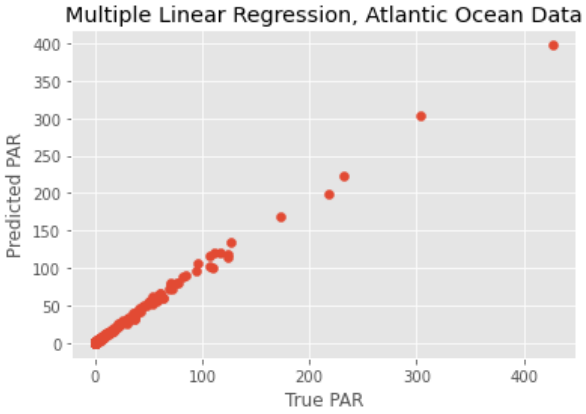

Figure 7: Multiple linear regression on Atlantic Ocean Data

The R^2 value was 0.996. The resulting model can be found in Equation 5.

PAR=805.10*Ed380
+203.80*Ed412 (5)
+494.75*Ed490−0.38

Models based on Regression Trees

A regression tree algorithm generating a binary tree was used in this research. The central task was to find a split that leads to an optimal separation of data [10]. In this paper the Scikit-learn implementation of regression tree was used, which makes use of the Gini Importance [11] to choose an attribute to split on. Figures 9 to 13 plot the predicted PAR values versus the groundtruth. When compared with the results for linear regression, it can be seen that groups of the plotted data points are aligned horizontally. This is because the regression tree predicts value ranges rather than individual values. Figure 8 shows the resulting tree for the combined dataset. Due to the complexity of the tree, the parameters of the subsequent tree based models are omitted.

Figure 9: Regression tree on all data

```
|--- Ed490 <= 0.39
|   |--- Ed490 <= 0.10
|   |   |--- Ed490 <= 0.02
|   |   |   |--- Ed490 <= 0.01
|   |   |   |   |--- Ed412 <= 0.00
|   |   |   |   |   |--- value: [0.08]
|   |   |   |   |--- Ed412 >  0.00
|   |   |   |   |   |--- value: [2.52]
|   |   |   |--- Ed490 >  0.01
|   |   |   |   |--- Ed380 <= 0.00
|   |   |   |   |   |--- value: [12.49]
|   |   |   |   |--- Ed380 >  0.00
|   |   |   |   |   |--- value: [5.54]
|   |   |--- Ed490 >  0.02
|   |   |   |--- Ed380 <= 0.00
|   |   |   |   |--- Ed490 <= 0.04
|   |   |   |   |   |--- value: [32.61]
|   |   |   |   |--- Ed490 >  0.04
|   |   |   |   |   |--- value: [67.39]
|   |   |   |--- Ed380 >  0.00
|   |   |   |   |--- Ed490 <= 0.06
|   |   |   |   |   |--- value: [17.13]
|   |   |   |   |--- Ed490 >  0.06
|   |   |   |   |   |--- value: [41.63]
|   |--- Ed490 >  0.10
|   |   |--- Ed490 <= 0.20
|   |   |   |--- Ed380 <= 0.00
|   |   |   |   |--- Ed490 <= 0.12
|   |   |   |   |   |--- value: [102.96]
|   |   |   |   |--- Ed490 >  0.12
|   |   |   |   |   |--- value: [161.96]
|   |   |   |--- Ed380 >  0.00
|   |   |   |   |--- Ed490 <= 0.15
|   |   |   |   |   |--- value: [59.81]
|   |   |   |   |--- Ed490 >  0.15
|   |   |   |   |   |--- value: [94.03]
|   |   |--- Ed490 >  0.20
|   |   |   |--- Ed380 <= 0.01
|   |   |   |   |--- Ed490 <= 0.28
|   |   |   |   |   |--- value: [231.32]
|   |   |   |   |--- Ed490 >  0.28
|   |   |   |   |   |--- value: [335.38]
|   |   |   |--- Ed380 >  0.01
|   |   |   |   |--- Ed380 <= 0.05
|   |   |   |   |   |--- value: [123.24]
|   |   |   |   |--- Ed380 >  0.05
|   |   |   |   |   |--- value: [184.44]
|--- Ed490 >  0.39
|   |--- Ed380 <= 0.28
|   |   |--- Ed380 <= 0.18
|   |   |   |--- Ed380 <= 0.03
|   |   |   |   |--- Ed490 <= 0.50
|   |   |   |   |   |--- value: [450.25]
|   |   |   |   |--- Ed490 >  0.50
|   |   |   |   |   |--- value: [539.81]
|   |   |   |--- Ed380 >  0.03
|   |   |   |   |--- Ed490 <= 0.55
|   |   |   |   |   |--- value: [255.22]
|   |   |   |   |--- Ed490 >  0.55
|   |   |   |   |   |--- value: [339.21]
|   |   |--- Ed380 >  0.18
|   |   |   |--- Ed490 <= 0.62
|   |   |   |   |--- Ed490 <= 0.52
|   |   |   |   |   |--- value: [344.61]
|   |   |   |   |--- Ed490 >  0.52
|   |   |   |   |   |--- value: [394.69]
|   |   |   |--- Ed490 >  0.62
|   |   |   |   |--- Ed490 <= 0.82
|   |   |   |   |   |--- value: [459.72]
|   |   |   |   |--- Ed490 >  0.82
|   |   |   |   |   |--- value: [508.59]
|   |--- Ed380 >  0.28
|   |   |--- Ed490 <= 1.16
|   |   |   |--- Ed380 <= 0.38
|   |   |   |   |--- Ed490 <= 1.07
|   |   |   |   |   |--- value: [605.62]
|   |   |   |   |--- Ed490 >  1.07
|   |   |   |   |   |--- value: [689.01]
|   |   |   |--- Ed380 >  0.38
|   |   |   |   |--- Ed380 <= 0.41
|   |   |   |   |   |--- value: [688.43]
|   |   |   |   |--- Ed380 >  0.41
|   |   |   |   |   |--- value: [741.71]
|   |   |--- Ed490 >  1.16
|   |   |   |--- Ed490 <= 1.32
|   |   |   |   |--- Ed380 <= 0.47
|   |   |   |   |   |--- value: [802.03]
|   |   |   |   |--- Ed380 >  0.47
|   |   |   |   |   |--- value: [888.95]
|   |   |   |--- Ed490 >  1.32
|   |   |   |   |--- Ed380 <= 0.59
|   |   |   |   |   |--- value: [1021.58]
|   |   |   |   |--- Ed380 >  0.59
|   |   |   |   |   |--- value: [1126.55]
```

Figure 8: Regression tree structure induced on the combined dataset.

Figure 9 shows the result of the regression tree using the combined dataset comprising all float locations. The fit of the regression tree model on all data combined resulted in $R^2 = 0.960$. Figure 10 shows the result of the regression tree using the Mediterranean dataset. The fit of the regression tree model on the Mediterranean Sea data resulted in $R^2 = 0.989$. Figure 11 shows the result of the regression tree using the Baltic Sea dataset float location 1.

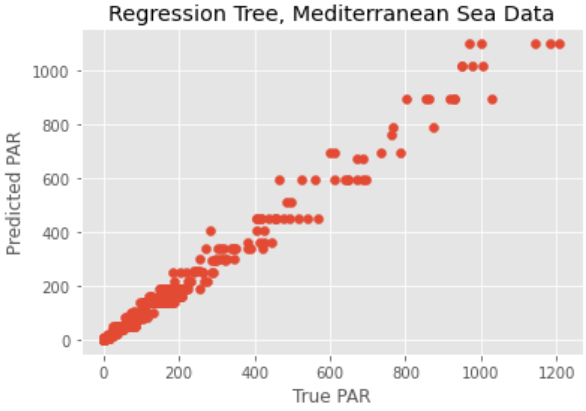

Figure 10: Regression tree on Mediterranean data

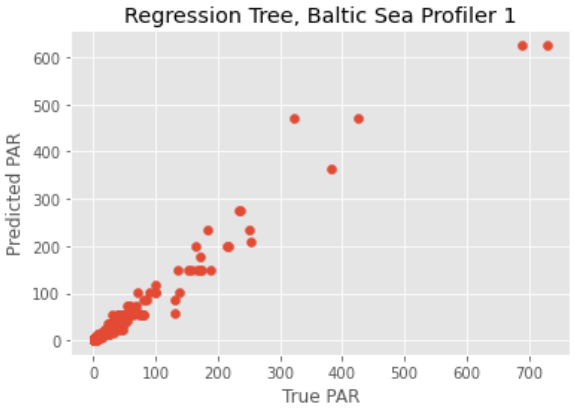

Figure 11: Regression tree on Baltic Sea float 1 data

The fit of the regression tree model on the Baltic Sea location 1 data resulted in $R^2 = 0.973$.

Figure 12 shows the result of the regression tree using the Baltic Sea dataset float location 2.

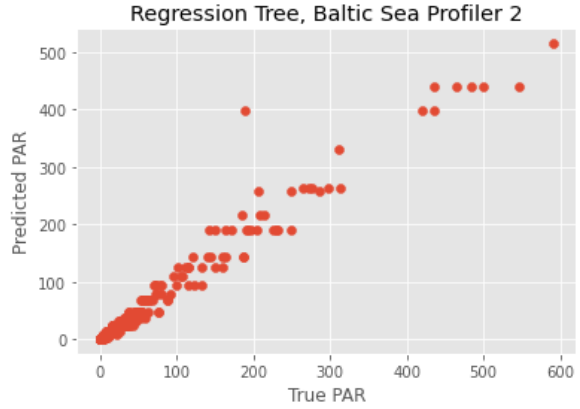

Figure 12: Regression tree on Baltic Sea float 2 data

The fit of the regression tree model on the Baltic Sea location 2 data resulted in $R^2 = 0.963$.

Figure 13 shows the result of the regression tree using the Atlantic Ocean dataset.

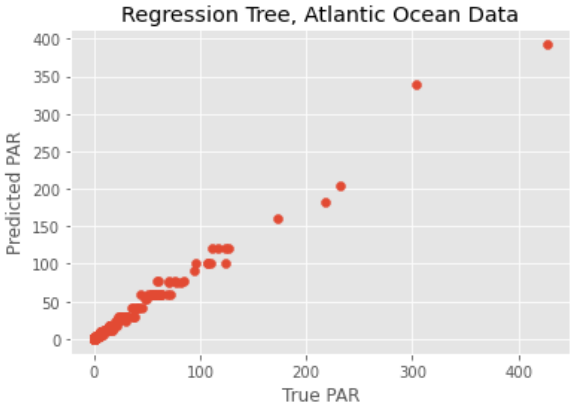

Figure 13: Regression tree on Atlantic Ocean data

The fit of the regression tree model on the Atlantic Ocean data resulted in $R^2 = 0.988$.

RESULTS AND DISCUSSION

In Table 1 the R^2 values for the different models are compared. As it can be observed, the R^2 values for the Multiple linear regression models are marginally better than those for the regression tree. It is assumed that this is caused by inherent discretization of predicted results at the leave nodes of the regression tree.

Table 1: R^2 values for different models using Multiple linear regression (MLR) and Regression Tree (RT)

R^2 values		
Dataset	**MLR**	**RT**
Combined	0.970	0.960
Mediterranean Sea	0.997	0.989
Baltic Sea Float 1	0.981	0.973
Baltic Sea Float 2	0.983	0.963
Atlantic Ocean	0.996	0.988

Furthermore, it can be seen that for both Multiple linear regression and regression trees based models the R^2 values for the combined datasets are slightly worse than models tailored for individual locations. This can be explained by influences of environmental parameters, for example salinity, which are different at the different location. These parameters were not available as input parameters for the models. However, the models are accurate enough to predict the PAR parameter without the need for a dedicated PAR sensor. Thus, PAR can be replaced by a specific wavelength enabling recording of more spectral information.

The next steps in this research is to use non-linear machine learning methods in order to increase the accuracy further.

REFERENCES

1. Roemmich D., Alford M.H., Claustre H., Johnson K., King B., Moum J., Oke P., Owens W.B., Pouliquen S., Purkey S., Scanderbeg M., Suga T., Wijffels S., Zilberman N., Bakker D., Baringer M., Belbeoch M., Bittig H.C., Boss E., Calil P., Carse F., Carval T., Chai F., Conchubhair D.Ó., d'Ortenzio F., Dall'Olmo G., Desbruyeres D., Fennel K., Fer I., Ferrari R., Forget G., Freeland H., Fujiki T., Gehlen M., Greenan B., Hallberg R., Hibiya T., Hosoda S., Jayne S., Jochum M., Johnson G.C., Kang K., Kolodziejczyk N., Körtzinger A., Le Traon P.-Y., Lenn Y.-D., Maze G., Mork K.A., Morris T., Nagai T., Nash J., Naveira Garabato A., Olsen A., Pattabhi R.R., Prakash S., Riser S., Schmechtig C., Schmid C., Shroyer E., Sterl A., Sutton P., Talley L., Tanhua T., Thierry V., Thomalla S., Toole J., Troisi A., Trull T.W., Turton J., Velez-Belchi P.J., Walczowski W., Wang H., Wanninkhof R., Waterhouse A.F.,

Waterman S., Watson A., Wilson C., Wong A.P.S., Xu J., Yasuda I. (2019) On the Future of Argo: A Global, Full-Depth, Multi-Disciplinary Array, *Frontiers in Marine Science*, Vol. 6, Article 439.
2. Sloyan, B. M., Roughan, M., Hill, K. (2018) Global Ocean Observing System, *New Frontiers in Operational Oceanography*, 75-89.
3. Claustre H., Bernard S.., Berthon J, Bishop J., Boss E., Coatanoan C., D`Ortenzio F., Johnson K., Lotliker A., Ulloa O. (2011) Bio-Optical Sensors on Argo Floats, In: Claustre, H. (ed.) *Reports and Monographs of the International Ocean-Colour Coordinating Group*, Dartmouth, Canada, p. 1-89, JRC67902.
4. Jiang Y., Gou Y., Zhang T., Wang K., Hu C. (2017) A machine learning approach to argo data analysis in a thermocline, *Sensors*, Vol. 17, No. 10, Article 2225.
5. Jemai A., Wollschläger J., Voß D., Zielinski O. (2021) Radiometry on Argo Floats: From the Multispectral State-of-the-Art on the Step to Hyperspectral Technology, *Frontiers in Marine Science*, Vol. 8, Article 676537.
6. Claustre H., Johnson K. S., Takeshita Y. (2020) Observing the global ocean with biogeochemical-Argo. Annual Review of Marine Science, 12, 23–48.
7. SATLANTIC (2013) Operation manual for the OCR-504, SATLANTIC Operation Manual SAT-DN-00034, Rev. G, p 66.
8. Organelli E., Leymarie E., Zielinski O., Uitz J., D'Ortenzio F., Claustre H. (2021) Hyperspectral radiometry on Biogeochemical-Argo floats: A bright perspective for phytoplankton diversity, in: Frontiers in Ocean Observing: Documenting Ecosystems, Understanding Environmental Changes, Forecasting Hazards, Kappel E.S., Juniper S.K., Seeyave S., Smith E., Visbeck M. (eds), A Supplement to *Oceanography*, Vol. 34, No. 4.
9. Freedman D.A. (2009) *Statistical models: theory and practice*, Cambridge University Press.
10. Breiman L., Friedman J.H., Olshen R.A., Stone, C.J. (2017) *Classification and regression trees*, Routledge.
11. Nembrini S., König I.R., Wright M.N. (2018) The revival of the Gini importance?, *Bioinformatics*, Vol. 34, Issue 21, pp 3711-3718.

AUTHOR BIOGRAPHIES

FREDERIC STAHL is Senior Researcher at the German Research Center for Artificial Intelligence (DFKI). He has been working in the field of Data Mining for more than 15 years. His particular research interests are in (i) developing scalable algorithms for building adaptive models for real-time streaming data and (ii) developing scalable parallel Data Mining algorithms and workflows for Big Data applications. In previous appointments Frederic worked as Associate Professor at the University of Reading, UK, as Lecturer at Bournemouth University, UK and as Senior Research Associate at the University of Portsmouth, UK. He obtained his PhD in 2010 from the University of Portsmouth, UK and has published over 65 articles in peer-reviewed conferences and journals.

LARS NOLLE graduated from the University of Applied Science and Arts in Hanover, Germany, with a degree in Computer Science and Electronics. He obtained a PgD in Software and Systems Security and an MSc in Software Engineering from the University of Oxford as well as an MSc in Computing and a PhD in Applied Computational Intelligence from The Open University. He worked in the software industry before joining The Open University as a Research Fellow. He later became a Senior Lecturer in Computing at Nottingham Trent University and is now a Professor of Applied Computer Science at Jade University of Applied Sciences. He is also affiliated with the Marine Perception research group of the German Research Centre for Artificial Intelligence (DFKI). His main research interests are AI and computational optimisation methods for real-world scientific and engineering applications.

AHLEM JEMAI is a PhD candidate at *Carl von Ossietzky University of Oldenburg*, Germany. She is working on the "Spectral Argo-N" and the "Deep Argo 2025" projects that are utilising the Argo technology. The projects are focused on the assessment of hyperspectral light conditions in oceanic and coastal waters through ocean color remote sensing and bio-optical models. She is also a co-worker in the project "Meteor Fjord Flux" on Expedition Meteor 179.

OLIVER ZIELINSKI is head of the research group "Marine Sensor Systems" at the Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University of Oldenburg. Since 2019 he is also heading the research department Marine Perception at the German Research Center for Artifical Intelligence (DFKI). After receiving his Ph.D. degree in Physics in 1999 from University of Oldenburg, he moved to industry where he became scientific director and CEO of "Optimare Group," an international supplier of marine sensor systems. In 2005, he was appointed Professor at the University of Applied Science in Bremerhaven, Germany. In 2007, he became Director of the Institute for Marine Resources (IMARE). He returned to the Carl von Ossietzky University of Oldenburg in 2011. His area of research covers marine optics and marine physics, with a special focus on coastal systems, smart sensors, and operational observatories involving different user groups and stakeholders.