On the Optimisation of Machine Learning Models for Predicting the Photosynthetically Available Radiation in the Water Column

Frederic Stahl[1], Lars Nolle[1,2], Martin Maximilian Kumm[2] and Christoph Tholen[1]

[1]German Research Center for Artificial Intelligence
Marie-Curie-Straße 1
26129 Oldenburg, Germany
Email: {christoph.tholen|lars.nolle|frederic_theodor.stahl}@dfki.de

[2]Jade University of Applied Sciences
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
Email: {lars.nolle|martin.kumm}@jade-hs.de

KEYWORDS
Machine Learning, Underwater Light Field, Photosynthetic Active Radiation, Freefall Profiler, KNIME

ABSTRACT

Photosynthetically Available Radiation (PAR) is a crucial parameter in oceanography. This study explores the optimization of machine learning models to predict PAR in the water column using selected wavelengths of downwelling irradiance. By leveraging Genetic Algorithms (GA), optimal wavelength combinations were identified for two machine learning models: Linear Regression (LR) and Regression Trees (RT). The models were trained on data from the HE533 expedition and validated using datasets from multiple ship expeditions across different geolocations. Experimental results indicate that the LR model, with an optimal wavelength combination of $E_d(469)$, $E_d(501)$, and $E_d(600)$, achieved the highest prediction accuracy (R^2 = 0.9992, MAE = 5.78). The RT model, using $E_d(433)$, $E_d(586)$, and $E_d(687)$, performed slightly worse (R^2 = 0.9954, MAE = 16.37). While both models generalised well to unseen datasets, significant prediction errors were observed for small PAR values at lower water depths.

INTRODUCTION

Photosynthetically Available Radiation (PAR) is an important parameter in modern oceanography. PAR is defined as the integrated radiation between 400-700 nm. One potential application is modelling vegetation growth, because the radiation in this wavelength is a requirement for the photosynthesis process (Holinde and Zielinski, 2016; Wang et al., 2013).

Modelling vegetation growth in the water column is crucial for understanding ecosystem dynamics (Krause-Jensen and Duarte, 2016), predicting climate change impacts (Dutkiewicz et al., 2019), managing water resources (Glibert, 2020) or modelling the oxygen production (Field et al., 1998). Therefore, providing reliable PAR values for different water bodies is an important task.

Figure 1 – Freefall profiler

As proven in previous work, the PAR values can be re-constructed using only discrete wavelengths from the underwater light field and, if necessary, additional environmental parameters (Stahl et al., 2022; Kumm et al., 2022; Tholen et al., 2024). Predicting PAR has been explored in the context of autonomous Argo Float devices (Sloyan et al., 2018) in (Stahl et al., 2022) using multiple linear regression and regression trees. Kumm et. al. (2022) showed that these results can be improved by using artificial neural networks-based models and further improved by incorporating additional environmental parameters, i.e. pressure. Tholen et al. (2024) used data from Freefall Profilers (Figure 1) to

improve the prediction accuracy of PAR by incorporating incoming surface irradiance (E_s) (Wollschläger et al., 2020d). Most recently Pitarch et al. (2025) investigated the accurate estimation of PAR based on the different radiance wavelength available on different Argo Float configurations. The long term goal of this series of research is to make the PAR sensor, mounted on the Argo Floats obsolete, to allow the replacement with another sensor (Stahl et al., 2022). However, even if the models show a good performance on the data gathered by Argo Floats (Kumm et al., 2022; Stahl et al., 2022). However, most recent research has shown a unsatisfactory accuracy of the models relaying on the three wavelengths $E_d(400)$, $E_d(412)$, $E_d(490)$ for the Freefall Profiler dataset, which covers more geolocations and is therefore more complex (Tholen et al., 2024).

In Figure 2 two spectral profiles for a water depth of o1 m and 195 m are shown. It can be observed that the radiation profile depends on the water depth. Furthermore, one can observe that the wavelength measured by the Argo Floats might best reflect the characteristics of the full spectra. Therefore, in this paper Genetic Algorithm (GA) (Holland, 1975) will be used to optimise the selection of suitable wavelength for PAR predictions on Argo Floats.

DATASETS

In this research, data from different ship expeditions are used. The different models are trained on data from the HE533 Expedition (Voß et al., 2020e). As shown in Figure 3, this expedition was undertaken near the northern coast of Norway. The HE533 dataset contains originally 9858 tuples of which 37.77 % had to be discarded because of missing values.

To validate the generalisability of the models developed, data from ten other cruises was used in the validation process (Friedrichs et al., 2020; Mascarenhas et al., 2020; Voß et al., 2020f, 2020a, 2020b, 2020c, 2020d; Wollschläger et al., 2020a, 2020b, 2020c). As shown in Figure 3 these cruises cover different geolocations all over the world. The combined dataset for validation contains 64,060 tuples. However, only 10,954 tuples can be used due to missing values. All datasets used in this research are publicly available on the data portal Pangaea[1].

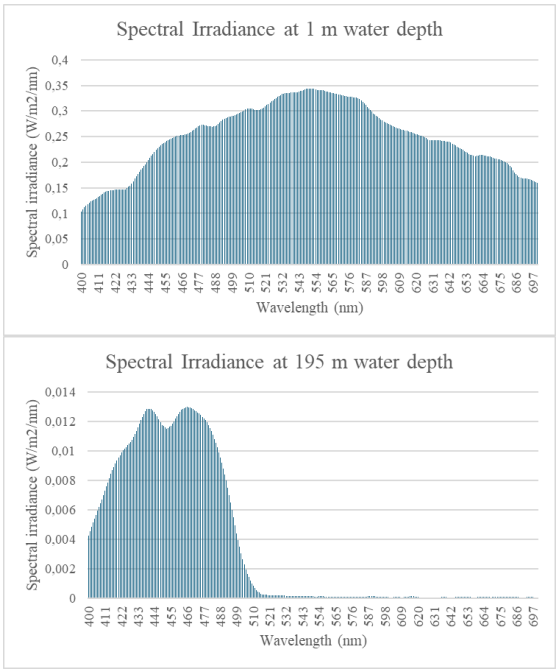

Figure 2 Example Spectral Irradiance for 1m and 195 water depth

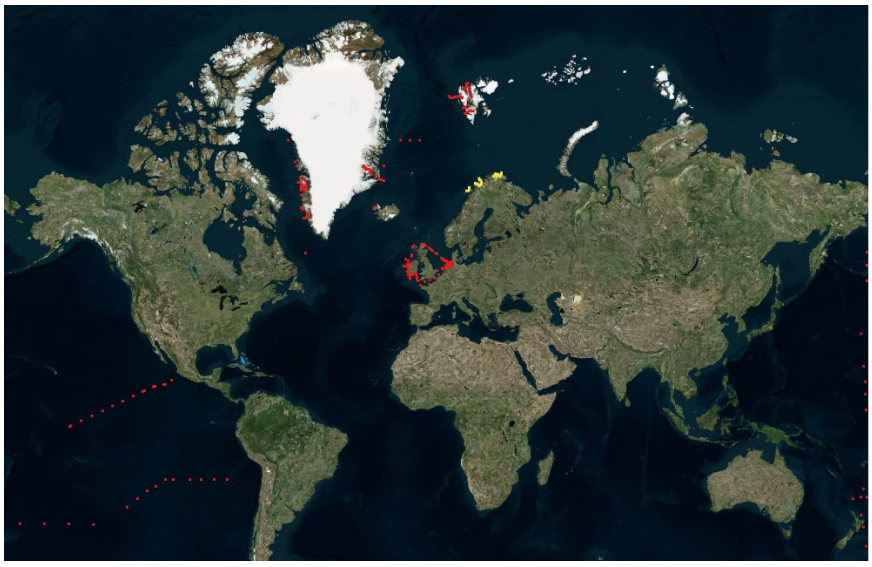

Figure 3 Locations of stations from cruise HE 533 (yellow) used for training and all other cruises (red) used for validation

[1] www.pangea.de

MODELLING

For modelling purposes, data from the HE533 dataset was used after pre-processing, i.e. normalisation and removal of data records with missing values. Random sampling without replacement was applied, to split this data into a training set (70 %) and a test set (30 %). The training set was used to find optimal wavelength for PAR prediction for two different AI based models, i.e. a Linear Regression model (LR), and a Regression Tree model (RT) utilising GA.

The test set was then used to validate the models generated in terms of accuracy. The outcome of this validation serves as a baseline to investigate the generalisability of the different models to measurements in other geolocations.

The models generated on HE533 were then applied on the other datasets available and evaluated in terms of accuracy. This accuracy was then compared with the baseline accuracy calculated from the HE533 test data. The modelling approach described is visualised in Figure 4.

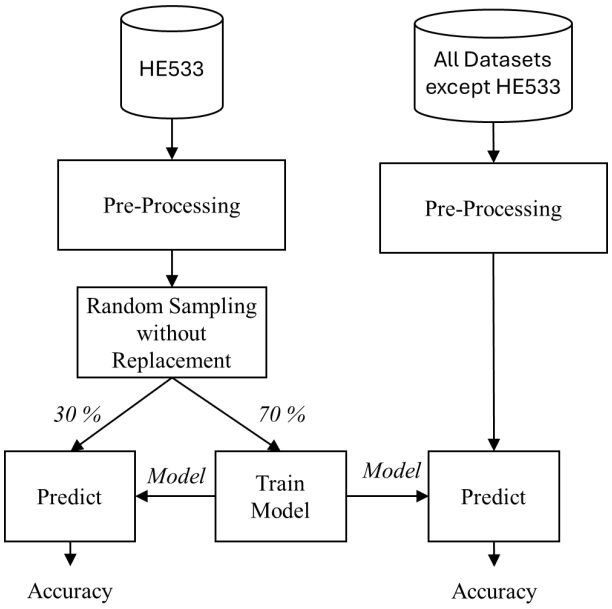

Figure 4: Modelling approach used

EXPERIMENTAL SETUP

All experiments were conducted using the KNIME workbench (Berthold et al., 2009). For training the RT the procedure described by Breiman et al. (1984) is applied with a couple of simplification, for instance no pruning, not necessarily binary trees. LR model uses standard multiple linear regression (Freedman, 2009).

During the experiments, GA is used to find an optimal subset of wavelengths to model the PAR value. Due to the limitations of the Argo Float platforms, a maximum number of three wavelengths is chosen during the optimization. The optimization was done utilizing the KNIME *Feature Selection Node* applying Genetic Algorithm as feature selection strategy. The population size was set to 20, while the maximum number of iterations was set to 10. The optimal parametrisation of the GA is not discussed in this research.

RESULTS

For the LR the highest accuracy was achieved by the combination of $E_d(469)$, $E_d(501)$, and $E_d(600)$, achieving an R^2 value of 0.9992. For this configuration, the accuracy on the validation dataset was 0.9982. A scatter plot showing the relation between predictions based on the LR-model and the true PAR values is given in Figure 5. It can be observed that the model tends to underestimate the PAR values, especially for higher values.

Figure 5 Scatter Plot for Linear Regression Predictions of PAR for the validation Datasets

For RT the highest accuracy was achieved by the combination of $E_d(433)$, $E_d(586)$, and $E_d(687)$, achieving an R^2 value of 0.9954. For this configuration, the accuracy on the validation dataset was 0.9603. Figure 6 shows the relation between PAR predictions and true PAR values utilising the RT model as scatter plot. It can be observed that the model performance decreases for higher PAR values.

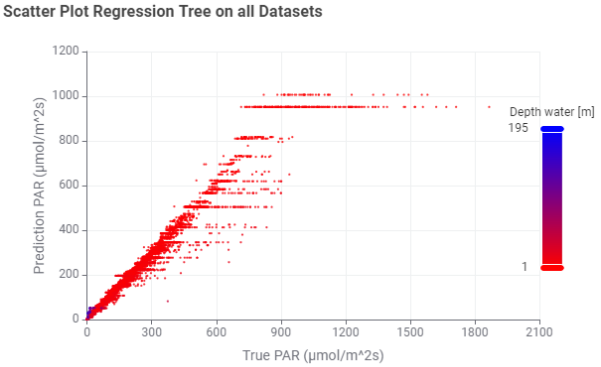

Figure 6 Scatter Plot for Regression Tree Predictions of PAR for the validation dataset

Statistical results of both best found ML models are given in Table 1. The mean absolute error (MAE) of the

LR is approximately three times smaller than the MAE of the RT model.

Table 1 Statistical values for both ML models on the validation set

Statistical Value	LR	RT
R^2	0.9969	0.9564
Mean Absolute Error	5.780	16.37
Mean Squared Error	141.8	1979
Mean Absolute Percentage Error	0.1160	0.2208

DISCUSSION

Models trained in this research, i.e. based on the optimised subset of wavelength, clearly outperform the models trained in Tholen et al. (2024). The best performing LR model found in previous work achieved an R^2 of 0.884, trained on $E_d(400)$, $E_d(412)$, $E_d(490)$. The best performing RT model achieved an R^2 of 0.822 on six wavelengths, incorporating three wavelengths of the surface radiation.

In Figure 6, depicting the results of the regression tree, it can be seen that groups the plotted data points are aligned horizontally. This is because a regression tree partitions the feature space into distinct regions based on decision rules. This is also known as local discretisation (Bramer, 2020). Within each region, the model assigns a constant predicted value to that region. If predicted values (i.e. the predicted PAR values), are plotted on the vertical axis against an independent variable on the horizontal axis (i.e. the true PAR values), all data points within the same leaf node will have the same predicted value. This results in horizontally aligned points because multiple input values share the same output value.

Wollschläger et al (2020d) introduced a trimodal approach to model the underwater light field, splitting the spectral information into three different bands. Within the optimisation GA automated chooses one wavelength from each of the bands for the LR model, while for the RT-model two wavelengths from the third band are chosen. For the RT model, two wavelength from the third band are chosen, while none of the wavelength from the second band is chosen. The trimodal approach was motivated by the absorption characteristics of the dominant substances affecting the underwater light field (Wollschläger et al., 2020d). Thus, the wavelength chosen for the RT might not be optimal for generalising PAR predictions.

As summarised in Table 1, the LR model outperforms the RT-model. Therefore, further discussions will focus on the LR model. To validate the suitability of the ML model to replace the PAR sensor, the relative error of predictions compared to true values is calculated as follows:

$$E_{relative} = \frac{PAR_T - PAR_P}{PAR_T} \cdot 100 \qquad (1)$$

Where PAR_T denotes the true PAR value, while PAR_P is the predicted PAR value. In Figure 7 the relative error is shown in dependence of the water depth for the LR model. One can observe measurements with high relative errors for low water depths.

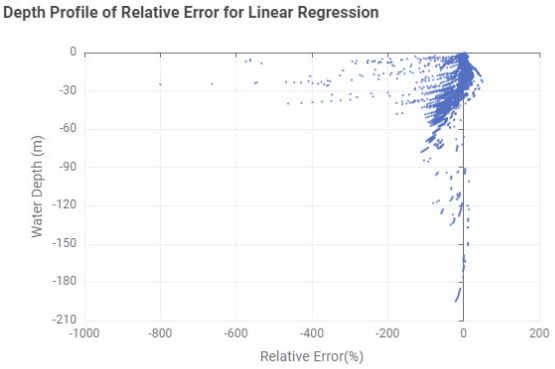

Figure 7 Depth profile of relative prediction error for the linear regression model

In Table 2 the ten highest relative error values are given together with the true PAR values and the PAR predictions. It can be observed that in all cases the true PAR value was smaller than 0.2197.

Table 2 Summary of data tuples with the highest relative error values of the validation set for linear regression

Relative Error (%)	True PAR	Prediction LR
-800.08	0.1328	1.195
-663.74	0.1591	1.215
-573.67	0.1867	1.258
-562.72	0.1917	1.270
-562.05	0.1888	1.250
-549.95	0.1903	1.237
-545.66	0.1895	1.224
-533.03	0.1974	1.250
-468.04	0.2197	1.248
-462.45	0.2175	1.223

As shown in Table 2 in all cases of high relative errors the true PAR values are small. For further investigation, Table 3 summarises statistics of the true PAR values of the subset of the validation set with higher relative errors than the given threshold, while the relative error over the true PAR values is shown in Figure 8. It can be observed that high relative errors occur for small true PAR values. If the prediction is limited to PAR values greater than one the maximum relative error is reduced to 50.04 %. Higher thresholds for the maximal relative errors result in smaller mean values for true PAR.

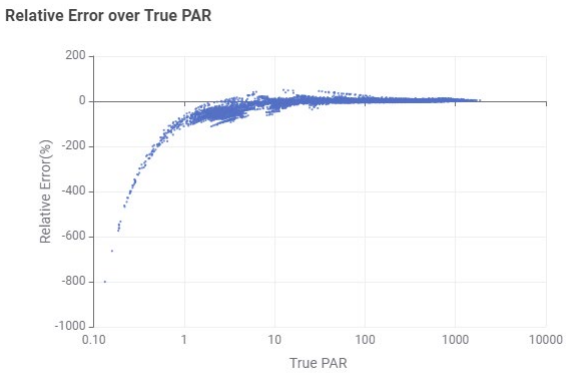

Figure 8 Relative error (%) of the linear regression model over the true PAR value

Table 3 Summary of statistics of true PAR values for tuples of the validation set with specific max. relative error values higher than the threshold

	Max. relative error (threshold)			
Statistics true PAR	**10 %**	**50 %**	**100 %**	**200 %**
Samples	1930	705	134	57
Mean	31.745	2.092	0.617	0.322
Std. dev.	110.7	1.258	0.42	0.096
Minimum	0.133	0.133	0.133	0.133
Maximum	900.44	12.538	2.353	0.508

CONCLUSION AND FUTURE WORK

The paper presented a modelling approach for predicting PAR in the water column, which uses downwelling irradiance in selected wavelengths. Two different AI-based modelling approaches, i.e. linear regression and regression tree, were used. The wavelength selection was optimised utilising a Genetic Algorithm. All experiments were conducted using the KNIME workbench.

It was shown that the linear regression model outperforms the regression tree model in terms of R^2 and mean absolute error. It was also shown that the models generalise well on data recorded in other geolocations without additional modification or re-training.

However, further analysis revealed that for small true PAR values high relative prediction errors occur, especially in lower water depths. It was already shown that incorporating additional environmental parameters such as e.g. pressure or salinity enhance the accuracy of the regression (Kumm et al., 2022). Therefore, the incorporation of additional parameters can also make the regression more stable and prevent high relative errors. This will be investigated in future research. In addition, domain experts will be incorporated to decide whether the performance of the ML models is high enough to replace the PAR sensor on the Argo Floats for future missions.

In addition, methods to improve linear regression models, such as regression splines (Friedman, 1991) or generalised additive models (Wood et al., 2015), will be investigated.

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Science and Culture, Lower Saxony, Germany, through funds from the zukunft.niedersachsen (ZN3480).

REFERENCES

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. KNIME - the Konstanz information miner: version 2.0 and beyond. SIGKDD Explor. Newsl. 11, 26–31. https://doi.org/10.1145/1656274.1656280

Bramer, M., 2020. Principles of Data Mining, Undergraduate Topics in Computer Science. Springer London, London. https://doi.org/10.1007/978-1-4471-7493-6

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Routledge, New York. https://doi.org/10.1201/9781315139470

Dutkiewicz, S., Hickman, A.E., Jahn, O., Henson, S., Beaulieu, C., Monier, E., 2019. Ocean colour signature of climate change. Nat Commun 10, 578. https://doi.org/10.1038/s41467-019-08457-x

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science 281, 237–240. https://doi.org/10.1126/science.281.5374.237

Freedman, D.A., 2009. Statistical Models: Theory and Practice, 2nd ed. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511815867

Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics 19, 1–67. https://doi.org/10.1214/aos/1176347963

Friedrichs, A., Schwalfenberg, K., Voß, D., Wollschläger, J., Zielinski, O., 2020. Hyperspectral underwater light field measured during the cruise MSM56 with RV MARIA S. MERIAN. https://doi.org/10.1594/PANGAEA.917534

Glibert, P.M., 2020. Harmful algae at the complex nexus of eutrophication and climate change. Harmful Algae, Climate change and harmful algal blooms 91, 101583. https://doi.org/10.1016/j.hal.2019.03.001

Holinde, L., Zielinski, O., 2016. Bio-optical characterization and light availability parameterization in Uummannaq Fjord and Vaigat–Disko Bay (West Greenland). Ocean Science 12, 117–128. https://doi.org/10.5194/os-12-117-2016

Holland, J., 1975. Adaptation in Natural and Artificial Systems.

Krause-Jensen, D., Duarte, C.M., 2016. Substantial role of macroalgae in marine carbon sequestration. Nature Geosci 9, 737–742. https://doi.org/10.1038/ngeo2790

Kumm, M.M., Nolle, L., Stahl, F., Jemai, A., Zielinski, O., 2022. On an Artificial Neural Network Approach for Predicting Photosynthetically Active Radiation in the Water Column, in: Bramer, M., Stahl, F. (Eds.), Artificial Intelligence XXXIX, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 112–123. https://doi.org/10.1007/978-3-031-21441-7_8

Mascarenhas, V.J., Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020. Hyperspectral underwater light field measured during the cruise MSM65 with RV MARIA S. MERIAN. https://doi.org/10.1594/PANGAEA.917564

Pitarch, J., Leymarie, E., Vellucci, V., Massi, L., Claustre, H., Poteau, A., Antoine, D., Organelli, E., 2025. Accurate estimation of photosynthetic available radiation from multispectral downwelling irradiance profiles. Limnology & Ocean Methods lom3.10673. https://doi.org/10.1002/lom3.10673

Sloyan, B., Roughan, M., Hill, K., 2018. Global Ocean Observing System.

Stahl, F., Nolle, L., Zielinski, O., Jemai, A., 2022. A Model for Predicting the Amount of Photosynthetically Available Radiation from BGC-ARGO Float Observations in the Water Column, in: ECMS 2022 Proceedings Edited by Ibrahim A. Hameed, Agus Hasan, Saleh Abdel-Afou Alaliyat. Presented at the 36th ECMS International Conference on Modelling and Simulation, ECMS, pp. 174–180. https://doi.org/10.7148/2022-0174

Tholen, C., Nolle, L., Wollschlaeger, J., Stahl, F., 2024. Model generalisation for predicting the amount of photosynthetically available radiation in the water column from freefall profiler observations, in: ECMS 2024 Proceedings Edited by Daniel Grzonka, Natalia Rylko, Grazyna Suchacka, Vladimir Mityushev. Presented at the 38th ECMS International Conference on Modelling and Simulation, ECMS, pp. 381–386. https://doi.org/10.7148/2024-0381

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020a. Hyperspectral underwater light field measured during the cruise SO248 with RV SONNE. https://doi.org/10.1594/PANGAEA.911988

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020b. Hyperspectral underwater light field measured during the cruise SO267/2 with RV SONNE. https://doi.org/10.1594/PANGAEA.912028

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020c. Hyperspectral underwater light field measured during the cruise SO245 with RV SONNE. https://doi.org/10.1594/PANGAEA.911558

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020d. Hyperspectral underwater light field measured during the cruise SO254 with RV SONNE. https://doi.org/10.1594/PANGAEA.912001

Voß, D., Wollschläger, J., Henkel, R., Zielinski, O., 2020e. Hyperspectral underwater light field measured during the cruise HE533 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.918041

Voß, D., Wollschläger, J., Henkel, R., Zielinski, O., 2020f. Hyperspectral underwater light field measured during the cruise HE492 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.918047

Wang, L., Gong, W., Li, C., Lin, A., Hu, B., Ma, Y., 2013. Measurement and estimation of photosynthetically active radiation from 1961 to 2011 in Central China. Applied Energy 111, 1010–1017. https://doi.org/10.1016/j.apenergy.2013.07.001

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020a. Hyperspectral underwater light field measured during the cruise HE503 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912073

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020b. Hyperspectral underwater light field measured during the cruise HE516 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912033

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020c. Hyperspectral underwater light field measured during the cruise HE527 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912054

Wollschläger, J., Tietjen, B., Voß, D., Zielinski, O., 2020d. An Empirically Derived Trimodal Parameterization of Underwater Light in Complex Coastal Waters – A Case Study in the North Sea. Frontiers in Marine Science 7.

Wood, S.N., Goude, Y., Shaw, S., 2015. Generalized Additive Models for Large Data Sets. Journal of the Royal Statistical Society Series C: Applied Statistics 64, 139–155. https://doi.org/10.1111/rssc.12068

AUTHOR BIOGRAPHY

FREDERIC STAHL is Principal Researcher at the German Research Center for Artificial Intelligence (DFKI), where he is heading the Marine Perception research department. He has been working in the field of Data Mining for more than 17 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 85 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 also is affiliated with the Marine Perception research department at the German Research Center for Artificial Intelligence (DFKI). His main research interests are computational optimisation methods for real-world scientific and engineering applications.

MARTIN MAXIMILIAN KUMM graduated from Jade University of Applied Sciences in Wilhelmshaven, Germany, with a Master degree in Mechanical Engineering in 2022. Since 2020 he is a research fellow at the Jade University of Applied Sciences responsible for the research aircraft. He also works in a joint project between Jade University of Applied Sciences, the German Research Center for Artificial Intelligence (DFKI) and marinom GmbH for the development of explainable artificial intelligence decision support systems for nautical officers.

CHRISTOPH THOLEN is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), where he is Deputy Head of the Marine Perception research department. His current research interests including the application of Artificial Intelligence applied to the maritime context, with a special focus on the identification and quantification of plastic litter using remote sensing. He received his doctoral degree in 2022 from the Carl von Ossietzky University of Oldenburg. From 2016 to 2022, he worked on a joint project between the Jade University of Applied Science and the Institute for Chemistry and Biology of the Marine Environment (ICBM), at the Carl von Ossietzky University of Oldenburg for the development of a low cost and intelligent environmental observatory.