Model Generalisation for Predicting the Amount of Photosynthetically Available Radiation in the Water Column from Freefall Profiler Observations

Christoph Tholen[1], Lars Nolle[1,2], Jochen Wollschläger[3] and Frederic Stahl[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@jade-hs.de

[3] Carl von Ossietzky Universität Oldenburg
School of Mathematics and Science
Institute for Chemistry and Biology of the Marine Environment (ICBM)
Ammerländer Heerstraße 114-118
26129 Oldenburg
Email: jochen.wollschlaeger@uni-oldenburg.de

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

ABSTRACT

In modern oceanography Photosynthetically Available Radiation (PAR) is used for modelling vegetation growth as it is a requirement for the process of photosynthesis. PAR as integrated value of the light spectrum between 400-700 nm can be measured directly using respective sensor systems. However, PAR can also be determined indirectly using measurements from only a small number of discrete wavelengths. In this paper, such a modelling approach is presented for predicting PAR in the water column. The approach uses spectral information within the water column and from above the sea surface. Three different modelling approaches based on artificial intelligence (AI) were used. It was shown that the artificial neural network (ANN) approach outperformed the regression tree (RT) and the linear regression (LR) approaches. It was also shown that the models generalise well, with an accuracy loss of 10 % based on the median, on data recorded in other geolocations without additional modification or re-training.

INTRODUCTION

In modern oceanography, one of the important parameters is Photosynthetically Available Radiation (PAR), which is the integrated radiation between 400-700 nm. It can be used for modelling vegetation growth due to being a requirement for the photosynthesis process (Holinde and Zielinski, 2016; Wang et al., 2013).

Therefore, measuring PAR is important. 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). 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. Due to the heavy dependency of the underwater light field on the incoming surface irradiance (E_s) (Wollschläger et al., 2020d), an alternative to incorporate pressure measurements to improve accuracy would be using these surface light field measurements. However, since Argo floats operate autonomous underwater for long time, simultaneous measurements of the surface light field is not an option.

A similar way of measuring PAR is being conducted by Freefall Profilers (Figure 1). However, different to Argo Floats, these measurements also comprise E_s. Therefore, this study tries to map the approaches from Kumm et al (2022) and Stahl et al (2022) to the freefall profiler platform. In addition, it will be investigated if incorporating E_s into the model building increases the accuracy. It will also be investigated if models trained on one set of experiments can be generalised to data from other measurements, i.e. other geolocations. If possible, it would allow marine scientists to reuse the developed models without re-training.

Figure 1 – Freefall profiler.

RADIOMETRIC PROFILING

For the data acquisition, the underwater light field was investigated using a free-falling profiling system (HyperPro II; Sea-Bird Scientific, USA, former Satlantic), which is designed to slowly sink vertically through the water column (Figure 1). The HyperPro II was equipped with two hyperspectral HyperOCR radiometers (Sea-Bird Scientific, USA, λ=350-800 nm) measuring different parts of the underwater light field: A planar cosine radiometer was mounted looking upward in order to determine the downwelling irradiance $E_d(\lambda)$, thus the overall light field propagating from the from the sea surface into the depth. Another, radiance-type radiometer with a field-of-view of 8.5° was mounted looking downward to measure the upwelling radiance $L_u(\lambda)$, thus the light field scattered back from the depth in a narrow cone in sinking direction. A third planar cosine radiometer was placed as reference in an unshaded, upright position on an elevated position on the ship in order to determine the downwelling irradiance $E_s(\lambda)$ above the seawater, thus the light field impinging on the sea surface. Its measurements allow the correction of the in-water measurements for general changes in the light field (e.g. temporary cloud coverage) during the deployment of the HyperPro II. The HyperPro II also contains sensors for additional parameters, like temperature, conductivity, depth, chlorophyll-a fluorescence, backscatter, and tilt of the instrument. All sensors on the instrument were pre-calibrated by the manufacturer, and the radiometers were checked with a reference lamp (FieldCal, TriOS GmbH, Germany) before and after the cruise, confirming that the initial calibration was still valid.

The handling of the HyperPro II followed the same protocol as in Holinde and Zielinski (2016), Mascarenhas et al. (2017), and Wollschläger et al. (2020d): Prior to the deployment of the HyperPro II at a station, the depth sensor was tared on deck of RV *Heincke* (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 2017) with the instrument in an upright position in order to adjust it to the current air pressure and ensuring correct in-water readings of the depth. Afterwards, the HyperPro II was deployed from the ship's stern, letting it drift to a distance of approx. 30 m to avoid shadow anomalies on the underwater measurements caused by the ship and its superstructures. Per station, one to three profiles were taken, depending on available time. All profiles were done as deep as possible (limited by the length of the instrument cable), but at least until the lower limit of the euphotic zone (depth in which 1% of surface PAR is available). Data were recorded using the SatView software (version 2.9.5_7). During data processing readings corresponding to an instrument tilt of >5° were discarded, as a vertical orientation of the instrument is necessary for correct measurements.

MODELLING

For modelling purposes, data from the HE533 Expedition (Voß et al., 2020e) was used after pre-processing, i.e. normalisation and removal of data records with missing values. Random sampling without replacement was applied, to split the HE533 data into a training set (70 %) and a test set (30 %). The training set was used to learn three different AI based models, i.e. a Linear Regression model (LR), an Artificial Neural Network model (ANN), and a Regression Tree model (RT).

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 where 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 2.

EXPERIMENTAL SETUP

Publicly available datasets from different ship cruises are used. All datasets can be found on the data portal Pangaea (www.pangaea.de). The data from the cruise HE533 (Voß et al., 2020e) was used to train the different models, while the data from the other cruises was used for validation (Friedrichs et al., 2020; Mascarenhas et al., 2020; Voß et al., 2020f, 2020a, 2020b, 2020c, 2020d; Wollschläger et al., 2020a, 2020b, 2020c). The HE533 dataset contains originally 9858 tuples of which

37.77 % had to be discarded because of missing values. The combined dataset for validation contains 64060 tuples of which 23.05 % for experiment 1 and 23.14 % for experiments 2 and 3 had to be discarded also because of missing values.

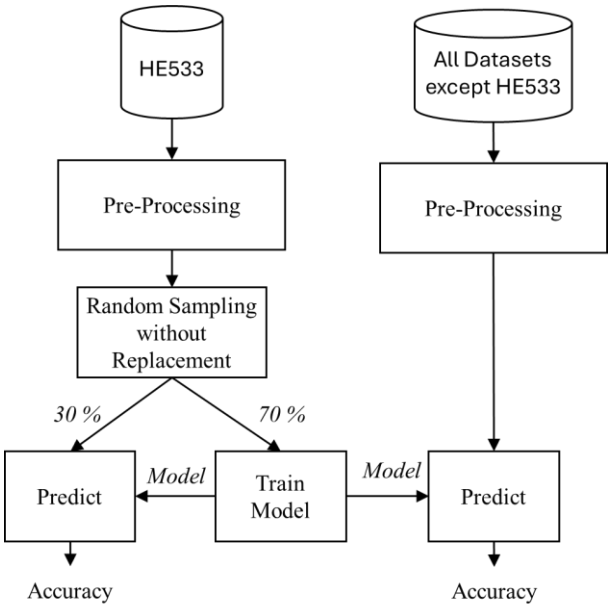

Figure 2: Modelling approach used

All models were generated using the KNIME workbench (Berthold et al., 2009). An ANN was used with one hidden layer containing 100 hidden units and trained for 1,000 epochs, using adaptive RProp (Riedmiller and Braun, 1993). For 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).

Three sets of experiments were carried out. In all the experiments, the models were trained using the HE533 dataset. In the first set of experiments, the models were trained on three wavelengths measured in the water column (E_d), 400 nm, 412 nm, and 490 nm, based on (Stahl et al., 2022). In the second set of experiments, the full spectrum of the surface light (E_s) between 400 nm and 700 nm, in 1 nm steps, was added to the inputs. In the third set of experiments, the full E_s spectrum was replaced by the same wavelengths that were used from the underwater light field. The results of the experiments are presented in the next section.

EXPERIMENTAL RESULTS AND DISCUSSION

For comparing the models, the R^2 values were calculated on the test data (see Figure 2). The R^2 value was chosen as metric to ensure comparability with previously published results (Kumm et al., 2022; Stahl et al., 2022). The results on the three experiments can be found in Table 1, where the R^2 values on HE533 correspond to the left hand side of Figure 2, whereas the R^2 for all datasets except HE533 correspond to the right hand side of Figure 2.

As can be seen in Table 1, the R^2 values on all datasets are lower compared with R^2 values on test data from HE533. This was expected since the additional data was not involved in training the models and were recorded in different geolocations with different physical properties.

Table 1: R^2 values for different models using Multiple Linear Regression (LR), Neural Network (ANN) and Regression Tree (RT).

Experiment #	Trained on	Model	R^2 on HE533	R^2 (all Datasets except HE533)
1	**HE533 E_d(400), E_d(412), E_d(490)**	LR	0.984	0.884
		ANN	0.986	0.879
		RT	0.972	0.821
2	**HE533 E_s(full spectrum) and E_d(400), E_d(412), E_d(490)**	LR	0.984	0.035
		ANN	0.989	0.899
		RT	0.977	0.795
3	**HE533 E_s(400), E_s(412), E_s(490) and E_d(400), E_d(412), E_d(490)**	LR	0.982	0.880
		ANN	0.986	0.919
		RT	0.973	0.822

When comparing Experiment 2 with Experiment 1, one can observe that the R^2 values are in the same order of magnitude for the evaluation on HE533, i.e. there was no improvement. However, when comparing results for all datasets, it can be observed that for ANNs the accuracy increases by 2.0 % whereas the performance for the regression tree deceases by 2.6 %. Noticeable, the linear regress decreases in performance by 84.9 %. It is believed that this underperformance is caused by outliers in some of the additional spectral information from the surface light. The linear regression approach will consider all spectral information including outliers. On the other hand, regression trees perform an internal selection of the best spectral information for branching and building the tree structure. Therefore, outliers may not be selected for branching. A neural network can also cope very well with outliers, since they can model non-linear dependencies.

When comparing Experiment 3 with Experiment 1 one can observe that the R^2 values are in the same order of magnitude, even for linear regression. This is in line with the observations about linear regression performance in Experiment 2, since in Experiment 3 a limited spectrum, i.e. number of input variables, was used. The datasets were normalised before training and validation took place.

Comparing results on HE533, with the results on all datasets except HE533 and for all experiments, once can see that the accuracy drops by approximately 10 % using median. The neural network-based model outperformed linear regression and regression tree-based models. This is probably because there are some non-linear factors that a neural network can compensate better. These results are in line with the findings reported by Kumm et al. (2022).

It was shown that spectral information from the surface light can be used to improve the generalisability of the models, especially of the ANN.

CONCLUSIONS AND FUTURE WORK

The paper presented a modelling approach for predicting PAR in the water column, which uses selected spectral information within the water column and additionally surface spectral information. Three different AI-based modelling approaches where used. It was shown that the ANN approach outperformed the RT and LR models. It was also shown that the models generalise well on data recorded in other geolocations without additional modification or re-training.

It should be noted that the parameter settings of the models have not been optimised yet. Therefore, further improvements are potentially possible. The selection of spectral variables was based on the literature. However, it is conceivable that different spectral information may result in more accurate models. Also, other environmental parameters such as e.g. pressure or salinity could potentially improve the models. Therefore, a more systematic variable selection process will be investigated in the future. 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 Niedersächsische Vorab (ZN3480).

REFERENCES

Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 2017. Research Vessel HEINCKE Operated by the Alfred-Wegener-Institute. Journal of large-scale research facilities JLSRF 3, A120–A120. https://doi.org/10.17815/jlsrf-3-164

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

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

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

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

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

Mascarenhas, V.J., Voß, D., Wollschläger, J., Zielinksi, O., 2017. Fjord light regime: Bio-optical variability, absorption budget, and hyperspectral light availability in Sognefjord and Trondheimsfjord, Norway. Journal of Geophysical Research: Oceans 122, 3828–3847. https://doi.org/10.1002/2016JC012610

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: IEEE International Conference on Neural Networks. Presented at the IEEE International Conference on Neural Networks, pp. 586–591 vol.1. https://doi.org/10.1109/ICNN.1993.298623

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

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

CHRISTOPH THOLEN is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), in 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.

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.

JOCHEN WOLLSCHLÄGER is a senior scientist in the working group Marine Sensor systems at the Institute for Chemistry and Biology of the Marine Environment at the Carl-von-Ossietzky-Universität Oldenburg. Being a biologist by training, he got his PhD in 2013 from the Jacobs University (now Constructor University) in Bremen and is working in the field of aquatic sensors for almost 14 years. His work focuses on the application of new and established bio-optical *in situ* instruments for the characterization of the inherent and apparent optical properties of the water. From these data, the relationship to the optically active substances (phytoplankton, CDOM, and non-algal particles) in the water are investigated to obtain ecologically relevant information.

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.