ATMOSPHERIC CORRECTION OF HYPERSPECTRAL DATA OVER COASTAL WATERS BASED ON MACHINE LEARNING MODELS

Ole Martin Borge[1], Sivert Bakken[2], Tor A. Johansen [2]

Department of Physics(1) &
Center for Autonomous Marine Operations and Systems
Department of Engineering Cybernetics(2)
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

ABSTRACT

Standard Atmospheric Correction algorithms that predict water-leaving radiance, while working well for the open-ocean using multispectral data, can be inaccurate or computationally demanding for coastal and optically-complex waters, where the phytoplankton signal might be masked or modified by the presence of other substances. Here, different Machine Learning models are presented, trained, and evaluated using simulated hyperspectral ocean color data of top-of-the-atmosphere radiance from coastal waters to predict water-leaving radiance and other ocean color variables directly, such as chlorophyll concentration. High accuracy of up to 99% for some of the variables is achieved when trained and evaluated on simulated data.

Index Terms— Atmospheric Correction, Hyperspectral Imaging, Remote Sensing, Ocean Color, Machine Learning

1. INTRODUCTION

An understanding of the biogeochemistry, ecology, and hazards of the oceans with a changing climate is critical to sustaining Earth as a habitable planet [1]. Satellite remote sensing of the ocean's spectral albedo is an effective tool to characterize and monitor the ocean environment on a global scale. Ocean Color can be used to monitor chlorophyll concentration (CHL), a common biomarker for the state of the marine ecosystems, providing aquaculture industry and government with information regarding water quality, biogeochemical cycles, and fisheries management. This requires a knowledge of the biotic signatures of the different ecosystems as well as the separation of those signals created by the atmosphere.

Here, radiative transfer (RT) models will be utilized to characterize and separate the atmospheric signals, and explore Machine Learning (ML) solutions to identify coastal ecosystems from their spectral albedo, discriminating against the atmospheric transmission scenarios typically present for Hyperspectral (HS) data [2, 3].
Originally, the analysis was performed using ML models for Neural Network (NN), Stochastic Gradient Descent Regression (SGDR), Partial Least Squares Regression (PLSR), and Support Vector Regression (SVR). All models were used to train Atmospheric Correction (AC) models on HS data, but only the results from NN and PLSR are presented in this paper, as they were the most promising. The full analysis can be found in [4]. To generalize well with ML, a lot of data representing various environmental cases would have to be obtained. Using synthetic data sets was chosen as our approach due to the difficulty of finding large amounts of HS data with corresponding metadata, such as sun-target-sensor angles alongside $in-situ$ measurements of light fields for validation. Similar approaches using simulated data of multispectral radiance has been used to verify AC algorithms in [3, 5, 6], and will form a basis for comparison.

In this paper, the RT model AccuRT [7] is utilized to simulate different HS radiance/irradiance data representative of a wide range of atmospheric and coastal oceanic environments, both strong aerosol containment and Case 2 waters. The ML models were trained on the simulated data, to predict remote sensing reflectance ($R_{rs}(\lambda)$) from different top-of-the-atmosphere (TOA) radiances. Also, water Inherent Optical Properties (IOP) retrieval algorithms based on ML were produced, aimed to predict the main IOPs of the water. The ML models will predict CHL, mineral concentration (MIN) and the absorption coefficient at $443\,\mathrm{nm}$ for coloured dissolved organic matter ($a_{cdom}(443)$) from $R_{rs}(\lambda)$, defined in Eq. (1). The trained ML models is then validated against each other concerning the accuracy, computational complexity, and interpretation capability, to study which could be suitable for on-board processing. Giving an indication as to how well the approach found in [5] for multispectral data works for HS data, specifically for optically complex waters.

2. PROBLEM FORMULATION

As most of the satellite measured TOA radiance over waters is due to atmospheric contributions, retrieving useful properties from the water-leaving radiance could only be done well if the AC algorithms are accurate. Only a relatively small portion of the incoming sunlight is backscattered from below the ocean surface in comparison with the sunlight backscattered from the atmosphere and specular reflection from the surface [8].

With $L_w(0^+, \lambda)$ as water leaving radiance just above the sea surface and F_0 as the extraterrestrial solar irradiance, the $R_{rs}(\lambda)$ can be expressed as:

$$R_{rs}(\lambda) \equiv L_w(0^+, \lambda) \,/\, \mathrm{F}_0 \cos(\theta_0)\, t_0(\lambda, \theta_0) \qquad (1)$$

With the total measured TOA radiance at a given wavelength λ, $L_t(\lambda)$, for ocean-atmosphere systems can be expressed as the partitioned linear equation [9], see Fig. 1.

$$L_t(\lambda) = L_r(\lambda) + L_a(\lambda) + t(\lambda)L_{wc}(\lambda) + t(\lambda)L_{sky}(\lambda) + T(\lambda)L_{sun}(\lambda) + t(\lambda)L_w(\lambda) \quad (2)$$

where $L_r(\lambda)$ is the radiance due to Rayleigh scattering by air molecules, $L_a(\lambda)$ is the aerosol scattering, $L_{wc}(\lambda)$ is the radiance contribution from whitecap on the sea surface, $L_{sun}(\lambda)$ is the specular reflection of direct sunlight off the sea surface, $L_{sky}(\lambda)$ is the radiance contribution from surface-reflected background atmospheric radiance and $L_w(\lambda)$ is water-leaving radiance due to photons that penetrate the sea surface and are backscattered. Diffuse and direct transmittances are given as $t(\lambda)$ and $T(\lambda)$, respectively.

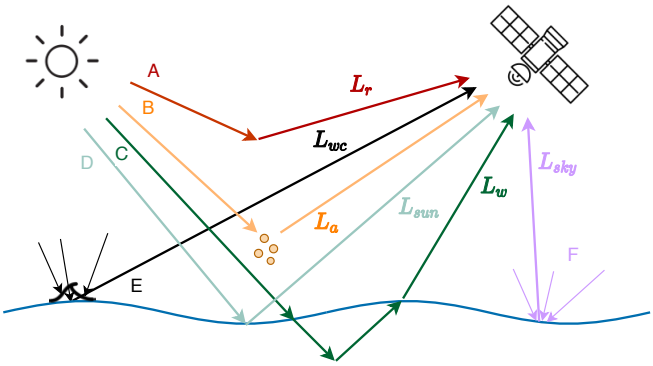

Fig. 1: Illustration of different contributions to the sensor-measured radiance.

Standard AC algorithms, e.g FLAASH, ATREM and POLDER [2, 3, 5], are based on a computationally demanding RT model, like MODTRAN, or a set of pre-calculated spectral Rayleigh scattering values, stored in Look Up Tables (LUT), to compute $L_r(\lambda)$. RT models can give an uncertainty lower than $0.5\,\%$ [5] when predicting $L_r(\lambda)$, which is also the major contribution to $L_t(\lambda)$. The algorithms could retrieve values from the LUTs matching the geometry and parameters from a scene, and use interpolation for values in between. Many AC algorithms for ocean color are based on the assumption that electromagnetic radiation in the NIR region back-propagated out of the water can be assumed to be zero, i.e. black ocean assumption. This assumption can be used to estimate aerosol contributions. While this approach works well for open oceans, it does not fit for coastal areas where the black ocean assumption tends to fail, and the aerosols can be more optically complex. One method to address this problem was studied by [5], where the combined aerosol and Rayleigh-corrected TOA radiances for multispectral data was used together as input to a NN, where $R_{rs}(\lambda)$ was predicted. A similar approach is presented here for HS data.

3. ACCURT MODEL

The coupled atmosphere-ocean RT Model AccuRT was used to simulate the interaction of solar radiation with particles and molecules in the atmosphere and ocean. AccuRT is a well-tested, user-friendly, and accurate radiative transfer model also capable of including effects from Case 2 waters, and was used to simulate different spectral radiance/irradiance data representative for strong aerosol containment and Case 2 waters [7]. AccuRT was used to generate synthetic data consisting of HS TOA radiance and corresponding $R_{rs}(\lambda)$ for a large variation of aerosol and ocean body properties for 400-$800\,\mathrm{nm}$ wavelengths with $5\,\mathrm{nm}$ spectral sampling.

3.1. Atmosphere and Aerosol

AccuRT uses a stratified vertical structure defined by the intensive properties of an atmosphere in hydrostatic balance. A 14-layer atmosphere covering the first $100\,\mathrm{km}$ was used with the predefined U.S. Standard atmospheric profile [5, 7]. The aerosols were added to the boundary level with Aerosol Optical Depths (AOD) at $869\,\mathrm{nm}$ chosen between 0.0001 and 0.4 was used. In AccuRT, it was not possible to specify AOD at any given wavelength directly, but the variation in AOD(869) could be included by varying the values of volume fraction of aerosols (f_v), the fraction of fine and coarse aerosols (f_s), and relative humidity (RH). Value ranges of these aerosol specific parameters could be chosen to get AOD(869) values between 0.0001 and 0.4, shown in Fig. 2. With θ_0, θ, $\Delta\phi$ as solar and sensor zenith, and relative sensor azimuth angle respectively, the ranges of simulated values are shown in Tab. 1.

3.2. Water IOPs

To simulate a representative synthetic dataset for the ML algorithms, water IOPs were extracted from *in-situ* field measurements. The water IOPs extracted, and needed for AccuRT data generation, were $a_{cdom}(443)$, CHL and MIN.

Data were extracted from the NASA bio-Optical Marine Algorithm Dataset (NOMAD) dataset and Coast Color Round Robin (CCRR) datasets [10]. The distribution of CHL, MIN, and $a_{cdom}(443)$ are shown in Fig. 2. The goal was to extract water IOPs representative for both Case 1 and Case 2 waters. When generating data with AccuRT, the bio-optical model CCRR was used [7].

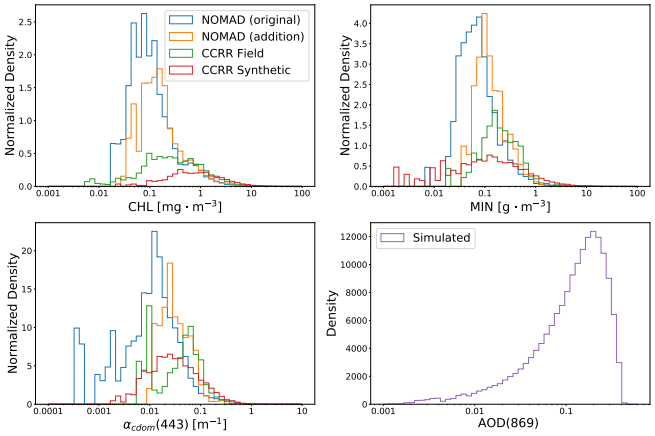

Fig. 2: Distribution of CHL, mineral concentration (MIN) and $a_{cdom}(443)$ extracted from different field datasets, together with the distribution of AOD(869) generated with AccuRT.

3.3. Data Generation with AccuRT

The main challenge for AC over coastal waters is that both the water and the aerosols are more optically complex, than for open oceans. This study will address this problem as done in [5] by looking at the simulated Rayleigh-corrected TOA radiance ($L_{rac}(\lambda)$), defined as:

$$L_{rac}(\lambda) = L_a(\lambda) + t(\lambda)L_w(\lambda) \tag{3}$$

$L_{rac}(\lambda)$ is therefore the TOA radiance corrected for atmospheric gas absorption, Rayleigh, glint and whitecaps. These effects were removed as they often are corrected for in satellite image processing [5]. The trained AC models would then predict $R_{rs}(\lambda)$ from $L_{rac}(\lambda)$,
For training and tuning/regularization 85% of the total simulated data was randomly selected, using the regularization approach described in Sec. 4.2 and 4.3. The remaining 15% was used to test the models, with the results shown in Sec. 5.

4. DATA PREPARATION & MACHINE LEARNING

4.1. Data Pre-processing

The spectral radiance input was divided by the cosine of the solar zenith angle, which is a term that can be found in Eq. (1), to get reflectance.

Table 1: Different input parameters used for the AccuRT simulations, their ranges and how the parameters were selected.

Parameter	Value range	Unit	Selection
θ_0	0-65	[°]	Uniform
θ	0-70	[°]	Uniform
$\Delta\phi$	0-180	[°]	Uniform
RH	30-95	[%]	Uniform
f_s	0-1	unitless	Uniform
f_v	1e-12 - 1e-10	unitless	Uniform
CHL	0.006 - 98	[mg/m^3]	Distribution
MIN	0.002 - 99	[g/m^3]	Distribution
$a_{cdom}(443)$	0.0004 - 5	[m^{-1}]	Distribution

The Savitzky-Golay filter (Savgol) can be applied to a set of discrete data points to smooth the spectral data without distorting the signal tendency [11]. Different types of derivatives can be applied to the spectral Savgol filtered data. This filter has long been used for spectroscopy to smooth and differentiate absorption spectra[12]. As the filter improved performance it was used as a pre-processing step.

Finally, the input data was normalized as given in Eq. (4), where $\hat{\mathrm{X}}$ is the normalized data, X is the original input data, $\overline{\mathrm{X}}$ is the mean and σ_X is the standard deviation.

$$\hat{\mathrm{X}} = \frac{\mathrm{X} - \overline{X}}{\sigma_X} \tag{4}$$

4.2. Sequential Neural Networks for Regression

It has been demonstrated that NNs with one or more hidden layers can predict non-linear relationships that could be suitable for deriving remote sensing reflectance from various TOA radiances [3, 5]. The NN presented here is a feed-forward NN, also known as the multilayer perceptron. The Python Deep Learning library Keras, with TensorFlow as a backend, was used to build the NN models, which were simple sequential models[13]. Several variations were tested, and the NN presented here used the "adam" optimizer with 2 hidden layers, 700 neurons in the first hidden layer, ReLU as activation function, and MSE as a cross-validation metric for tuning/regularization during training.

4.3. Partial Least-Squares Regression

PLSR reduce collinearity and noise within a given dataset by iteratively relating data matrices using linear multivariate models. It is a two-step algorithm that first finds uncorrelated components in the variables of a given data set and then performs the least squares regression on these components. A more in-depth description of the algorithm can be found in [14]. Several variations were tested, with 10-fold-cross-validation for tuning/regularization, and with variable selection[13].

For IOP prediction, 81 bands and 22 components gave the best results, as shown in Fig. 3. Varying the hyperparameters did not yield very different results.

5. RESULTS

The results of AC and IOP prediction using the different ML models were compared using the Pearson correlation coefficient (R), the average percentage difference (APD), and the normalized root mean squared difference (NRMSD), described in Eq. (5), (6) and (7), respectively.

$$\mathrm{R} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right)\left(\frac{\hat{Y}_i - \hat{\bar{Y}}}{\sigma_{\hat{Y}}}\right) \tag{5}$$

$$\mathrm{APD}\,[\%] = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right| \times 100 \tag{6}$$

$$\mathrm{NRMSD} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}}}{\hat{Y}_{max} - \hat{Y}_{min}} \tag{7}$$

where N is the number of samples, Y_i is the i-th predicted radiance value at a given wavelength, $\hat{Y}_i$ is the corresponding ground truth radiance value, σ_Y and $\sigma_{\hat{Y}}$ are the standard deviation of Y and $\hat{Y}$, $\bar{Y}$ and $\hat{\bar{Y}}$ are the mean values of Y and $\hat{Y}$, and $\hat{Y}_{max}$ and $\hat{Y}_{min}$ are the maximum and minimum value of $\hat{Y}$, respectively.

The ML models would give 81 predicted outputs from the wavelength bands between 400 and $800\,\mathrm{nm}$. Metric values for each predicted wavelength band were calculated and would therefore yield 81 values for each metric (R^2_{400}, R^2_{405}, ... , R^2_{800}). The optimal results of each ML model were also based on the mean of the 81 metric values, given as $\overline{\mathrm{R}^2}$, $\overline{\mathrm{APD}}$ and $\overline{\mathrm{NRMSD}}$.

5.1. Atmospheric Correction Results

In this study, the two ML models were used to predict $L_w(\lambda)$ from Rayleigh and absorption corrected radiance ($L_{rac}(\lambda)$), θ, θ_0 and $\Delta\phi$. Before finding the optimal results, a hyperparameter optimization study was done by training the ML models on a range of different Savgol filters and hyperparameters. The ML models giving the best results based on both the Pearson coefficient and NRMSD are presented in Tab. 2. The models are compared further by the aforementioned metrics and execution times.

AC of multispectral L_{rac} with MLNN done in [5] produced $\mathrm{R}^2 > 0.993$ for all 7 bands in VIS (412, 443, 488, 531, 547, 667 and $678\,\mathrm{nm}$) and $\overline{\mathrm{APD}} = 3.1\,\%$. The NN trained on HS data showed comparable results, see Tab. 2, with $\mathrm{R}^2 > 0.992$ for all 81 bands and $\overline{\mathrm{APD}} = 4.4\,\%$, and $\overline{\mathrm{R}^2}$ calculated from the NN was at 0.999, which was higher than 0.996 reported by [5]. These results imply that both models are able to predict the spectral relationship between L_{rac} and R_{rs} with similar or better accuracy.

Table 2: Optimal results when predicting $R_{rs}(\lambda)$ from $L_{rac}(\lambda)$, θ, θ_0 and $\Delta\phi$ using NN and PLSR with respect to time complexity. Computational time to fit the model ($\mathrm{T_{fit}}$), computational time to predict the output ($\mathrm{T_{pred}}$) and the number of training data ($\mathrm{N_{train}}$) are given.

Metrics	$\overline{\mathrm{R}^2}$	$\overline{\mathrm{APD}}$ [%]	$\overline{\mathrm{NRMSE}}$
NN	0.999	4.42	0.045
PLSR	0.974	34.1	0.197

Time	$\mathrm{T_{fit}}$ [s]	$\mathrm{T_{pred}}$ [s/N]	$\mathrm{N_{train}}$
NN	675	1.3e-3	91702
PLSR	166	1.0e-4	91702

5.2. IOP Prediction Results

Several ocean color algorithms [3] can predict IOPs from $R_{rs}(\lambda)$ based on empirical relationships derived from $in\text{-}situ$ measurements, like the non-linear OCx algorithm. Here, NN and PLSR models were trained to predict CHL, $a_{cdom}(443)$ and MIN from $R_{rs}(\lambda)$ from the HS $R_{rs}(\lambda)$ data. Different Savgol filters were applied to the input data for each NN model. The best results using NN to predict IOPs are shown in Tab. 3.

Table 3: Predicted chlorophyll concentration (CHL), $a_{cdom}(443)$, mineral concentration (MIN) from $R_{rs}(\lambda)$ with NN and PLSR validated with R^2, APD, and NRMSD.

	CHL		$a_{cdom}(443)$		MIN	
Metrics	NN	PLSR	NN	PLSR	NN	PLSR
$\overline{\mathrm{R}^2}$	0.999	0.987	0.996	0.961	0.998	0.994
$\overline{\mathrm{APD}}$ [%]	8.84	127.1	16.8	84.5	42.4	65.0
$\overline{\mathrm{NRMSE}}$	0.0353	0.271	0.0933	0.275	0.108	0.178

6. DISCUSSION AND CONCLUSION

AccuRT was used to simulate HS data representative for challenging coastal waters which could be used to train ML models. When predicting R_{rs} from TOA radiance corrected for Rayleigh and absorption (L_{rac}), all ML models resulted in $\mathrm{R}^2 > 0.968$, indicating that they were able to predict the spectral relationship between L_{rac} and R_{rs}. The best results were obtained with the NN algorithm (R^2=0.999), especially compared to the linear model PLSR (R^2=0.974). However, the PLSR provided interpretable coefficients, see Fig. 3, and

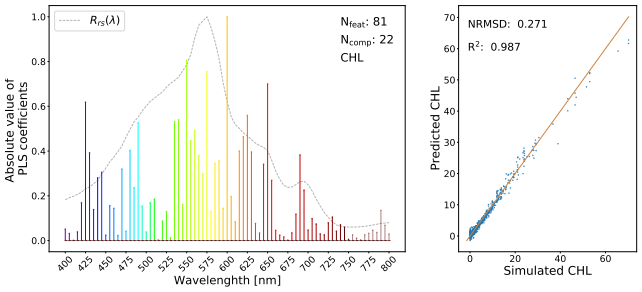

Fig. 3: Normalized absolute values of PLS coefficients as a function of wavelength bands for predicting CHL together with scatterplots of the predicted and simulated CHL. Number of features ($\mathrm{N_{feat}}$) and number of components ($\mathrm{N_{comp}}$) are highlighted in the left plot. The grey dashed curve represent one simulated $R_{rs}(\lambda)$.

shorter prediction time. A specific application or set of constraints will determine the most applicable model.

Unlike many standard AC algorithms, these models were capable of doing AC without the extra short-wave infrared (SWIR) bands, as they were trained on HS data in the wavelength region 400-800 nm. Finally, the NN approach could also be used for water IOP prediction, and provided $\mathrm{R}^2 > 0.999$ when predicting CHL from R_{rs}. The results when using synthetic data are comparable or outperforms the results reported in [5]. It should be noted that the underlying simulated data was trained for multispectral data in [5], and not HS data as in this paper.

The different AC algorithms based on ML after training are not computationally complex and, as shown in Tab. 3, can be executed quickly, and therefore could suit operational use in satellites as a part of the on-board data processing framework. For future research, the ML models should be tested on *in-situ* data and be validated against more conventional AC algorithms, such as FLAASH, ATREM or POLDER [2, 3].

7. ACKNOWLEDGEMENTS

Patrick J. Espy, Professor at NTNU Department of Physics, provided supervision, comments and guidance. This work was supported by the Norwegian Research Council through the Centre of Autonomous Marine Operations and Systems (NTNU AMOS) (grant no. 223254), the MASSIVE project (grant no. 270959).

References

[1] T. Platt *et al.*, Ed., *Why Ocean Colour? The Societal Benefits of Ocean-Colour Technology*, vol. No. 7 of *Reports of the International Ocean Colour Coordinating Group*, IOCCG, Dartmouth, Canada, 2008.

[2] M. Eismann, *Hyperspectral remote sensing*, Society of Photo-Optical Instrumentation Engineers, 2012.

[3] M. Wang, Ed., *Atmospheric Correction for Remotely-Sensed Ocean-Colour Products*, vol. No. 10 of *Reports of the International Ocean Colour Coordinating Group*, IOCCG, Dartmouth, Canada, 2010.

[4] O. M. Borge, "Atmospheric correction over coastal waters based on machine learning models," M.S. thesis, NTNU, 2020.

[5] Y. Fan *et al.*, "Atmospheric correction over coastal waters using multilayer neural networks," *Remote Sensing of Enviroment*, vol. 199, pp. 218–240, 2017.

[6] A.D. Gerace *et al.*, "Increased potential to monitor water quality in the near-shore environment with Landsat's next-generation satellite," *Journal of Applied Remote Sensing*, vol. 7, no. 1, pp. 1–19, 2013.

[7] K. Stamnes *et al.*, "Progress in Forward-Inverse Modeling Based on Radiative Transfer Tools for Coupled Atmosphere-Snow/Ice-Ocean Systems: A Review and Description of the AccuRT Model," *Applied Sciences*, vol. 8, pp. 2682, 2018.

[8] K. Moore *et al.*, "Spectral reflectance of whitecaps: Their contribution to water-leaving radiance," *Journal of Geophysical Research*, vol. 105, pp. 6493–6499, 2000.

[9] H. R. Gordon, "Atmospheric correction of ocean color imagery in the earth observing system era," *Journal of Geophysical Reseach: Atmospheres.*, , no. 102, pp. 17081–17106, 1997.

[10] K. Ruddick, "Coastcolour: Round robin protocol, version 1.2," *Brockmann Consult*, 2010.

[11] A. Savitzky and M. J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[12] "Nirpy research," `https://nirpyresearch.com/`, (Accessed on 2020-10-05).

[13] Aurélien Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, 2019.

[14] Roman Rosipal and Nicole Krämer, "Overview and recent advances in partial least squares," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer, 2005, pp. 34–51.