EVALUATING HYPERSPECTRAL SECCHI DEPTH RETRIEVAL THROUGH HYBRID MODELING AND REGRESSION

Sivert Bakken[1,2,], Kelly Luis[3], Geir Johnsen[2] and Tor Arne Johansen[2]*

[1] SINTEF Ocean, Trondheim, Norway
[2] Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[3] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, United States

ABSTRACT

This work compares different regression models combined with hybrid modeling to estimate water clarity using hyperspectral remote sensing data. The Secchi depth, a proxy of water clarity, can be modeled using first principles bio-optical modeling and other static pre-processing steps are used to generate four different feature sets. The different feature sets and regression models are evaluated using cross-validation on the recently published GLORIA dataset, representing a vast set of Secchi depth measurements from various aquatic environments ($N = 3914$). The best-performing feature generation and regression model combination can provide promising Secchi depth inference from hyperspectral data ($RMSE = 1.543$, $APD = 39.419$, $R^2 = 0.636$). The study demonstrates the potential of hyperspectral remote sensing data for monitoring and managing aquatic ecosystems.

***Index Terms*—** Water Quality, Hyperspectral, Secchi Depth, Machine Learning, hybrid modeling,

1. INTRODUCTION

Access to comprehensive information on water environments is increasingly vital for developing informed decisions regarding water resource use and development policies [1–5].

The Secchi Depth (SD) is a simple yet effective indicator of water quality, used to evaluate the long-term dynamics of water quality, and is one of the measurements that has been used since the 19^{th}-century. SD is influenced by the amount of dissolved and particulate matter in the water column, which plays a critical role in regulating various chemical, physical, and biological processes [1, 4, 5]. Furthermore, good water quality is valued for recreational activities, including boating, swimming, fishing, and sightseeing [3, 4]. Traditional observations of SD can be accurate but will be limited in terms of spatial and temporal resolutions. Fortunately, satellite remote sensing has emerged as a vital alternative tool for synoptic estimates of SD, providing large-scale observations and higher spatial and temporal resolutions [1, 4, 6]. Remote sensing can offer a broader perspective and enables us to make informed decisions that benefit our water resources and those who enjoy them [4].

The model presented in [1] provides a mechanistic model for SD retrieval, and the work is demonstrated in [7] to show how it can be applied to multispectral data from Landsat-8. Additionally, with considerable success, recent studies have developed different data-driven approaches to infer SD [2]. Results from [2] indicate that machine learning methods could have advantages over simple empirical band-ratio-based and semi-analytical methods. The models in [2] are shown to outperform the model developed in [1] when tested with multispectral data from Landsat-8. In [2], the results indicate that ensemble models, specifically Random Forrest Regression (RFR), appeared to be more reliable than single models such as Support Vector Regression (SVR). While previous work has been focused on multispectral data, with the publication of the GLORIA Data set [6], it is now possible to explore the use of hyperspectral data without being concerned with intermediate processing steps related to atmospheric compensation, data co-location, and other challenges of similar nature.

In this work, we show how the different approaches work in a hyperspectral context and show the potential benefits of feature engineering. It is shown that simple feature engineering can improve models' performance across various water bodies.

Sec. 2 provides details on the physical model, the hybrid modeling and feature engineering approach of choice, a brief introduction to the regression models, the GLORIA Data Set [6], and model evaluation metrics. In Sec. 3, details of the data handling can be found alongside some relevant considerations. Lastly, Sec. 4 provides conclusions and paths forward.

The Research Council of Norway is acknowledged for funding through AMOS (grant number 223254) and the HYPSCI project (grant number 325961. * Corresponding author, sivert.bakken@ntnu.no.

2. BACKGROUND AND THEORY

2.1. Secchi Depth Theory

The importance of the SD, z_{sd}, in water quality monitoring is detailed in Sec. 1. Several algorithms have been developed to infer the z_{sd} by remote sensing. More recently [1] provided a new mechanistic model for water visibility. For brevity, the model can be expressed as

$$z_{sd} = \frac{1}{2.5\ K_{d_{min}}} \ln \frac{|0.14\ R_{rs}(\ _k)|}{0.013} : \quad (1)$$

where $K_{d_{min}}$ is the minimum diffuse attenuation coef cient which characterizes the intensity reduction rate of the light as it passes through water. $R_{rs}(\ _k)$ refers to the remote sensing re ectance at the wavelength of $_{d_{min}}$. The parameter K_d can be expressed as

$$K_d = (1 + 0.005\ sa)\ a + (1 - 0.265\ nw)\ 4.259\ (1 - 0.52\ e^{-10.8\ a})\ bb; \quad (2)$$

with details in [8]. Here, sa denotes the solar zenith angle, a represents the total absorption, nw represents the contribution from molecular backscattering to total backscattering, and bb represents the total backscattering coef cient.

2.2. Hybrid Modeling and Regression Models

This approach combines machine learning with physical modeling to generate new models with more accurate and precise features while still being simpler to interpret. By utilizing both paradigms, hybrid models ideally ensure that mechanistic knowledge is kept in new, more data-driven techniques. Following the de nitions from [9], the rst principle model has been used to generate a suitable feature space in a serial con guration. That is, the results from the physical model described in Eq. (1) are included as an input to the data-driven model alongside the spectral re ectance. The feature engineering is tested against static spectral preprocessing steps such as log transformation and rst-order gradient. Several combinations were experimented with, and the interesting ones can be seen in Sec. 3.

SVR is a model that uses the Support Vector Machine algorithm to perform regression [10]. The model is trained using a set of training data, $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^d$ is the ith input vector of a given feature set and $y_i \in R$ is the corresponding SD for this particular problem. The model is trained by solving an optimization problem that nds a hyperplane, de ned by a kernel function, in an n-dimensional space that maps the feature vector to the target value. See [10, 11], for further details on SVR theory.

RFR is a regression model that uses the Random Forest (RF) algorithm to perform regression. This method employs several decision trees to make predictions. Each decision tree consists of nodes and branches, with each node representing a test on a feature or a group of features and each branch representing an outcome. The RFR aggregates the results from individual decision trees to make the best predictions according to a given metric or criterion,[10].

Fig. 1: Mean wavelength spectra from [6], where each water type and water body type has been plotted together.

2.3. The GLORIA Data Set

This work uses the SD reported in the GLORIA data set [6]. The instruments used for R_{rs} measurements are typically used for validating satellite-derived water re ectance with an above-surface protocol or using oating frames, with more details in [6]. In Fig. 1, the mean spectra of the data set have been plotted. The variability in the mean spectra from the different water and water body classes should be noted. The spectra have been transformed using a triangular Relative Spectral Response (RSR) function to have a Full Width at Half Maximum (FWHM) of 10 nm. This spectral resolution is more attainable with current and planned satellites with high-resolution optical imagry [12] and should improve the approach's applicability. Only the wavelengths between 443 and 700 nm as this spectral region are included in all the samples from [6] that had taken simultaneous SD measurements. The spectral resampling ensures that all SD measurements can be used as there are variations in spectral coverage of the sensors used [6]. Furthermore, by looking at this subset, it is also ensured that only the spectral region that most optical satellites can cover is used [12]

2.4. Model Evaluation

The metrics Root Mean Square Error (RMSE), Absolute Percentage Deviation (APD), and Correlation Score (R^2) are used to evaluate the different models.

RMSE can be computed as

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)^{\frac{1}{2}}. \quad (3)$$

where n is the number of observations, y_i is the SD value for the ith observation, and $\hat{y}_i$ is the model-inferred SD for the ith observation. The formula calculates the average of the squared differences between the inferred and measured values and then takes the square root of that average.

R^2 is a statistical measure representing the proportion of the variance in the dependent variable explained by the independent variables in a regression model. R^2 can be calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

A higher R^2 value indicates a better fit between the observed and the inferred data better and that the independent variables explain a more significant proportion of the variability in the dependent variable.

APD measures the relative difference between two values, expressed as a percentage. It is given here as the absolute error divided by the mean of the observed. This value can be expressed as

$$APD = \frac{|y_i - \hat{y}_i|}{\bar{y}} \cdot 100\% \quad (5)$$

3. METHODS AND RESULTS

This section describes the methods used in more detail, as well as a presentation of the results. A complete overview of the results with the chosen metrics can be found in Table 1.

This study derived various feature sets. The feature sets were labeled as

$$F0 := R_{rs}(\lambda);$$
$$F1 := \log(R_{rs}(\lambda));$$
$$F2 := \log(R_{rs}(\lambda)); \hat{y}_z; \text{ and}$$
$$F3 := \log(R_{rs}(\lambda)); r(\log(R_{rs}(\lambda))); \hat{y}_z;$$

with $R_{rs}(\lambda)$ being the set of wavelength vectors for a given data point, $\log(\cdot)$ being a logarithmic mapping of each feature, $r(\cdot)$ being the first-order gradient along the vector, and $\hat{y}_z$ being the SD derived from Eq. (1).

The correlation between radiometric intensity and the measured SD at different wavelengths can be seen for some operators in Fig. 2. A higher absolute correlation value, closer to 1, indicates that there should be a stronger relationship between the variables and the target value. The regression models are expected to perform better with a more linear relationship. The $\log(\cdot)$ compresses the dynamic range of the data, making it easier to distinguish small changes in the signal from noise and make multiplicative effects additive. The $r(\cdot)$ measures the rate of change of the signal with wavelength, which is less affected by baseline shifts than the raw spectra themselves [13].

Fig. 2: The absolute value of the correlation coefficient between each wavelength of various static preprocessing methods with the measured SD. The gradient is a first-order gradient, and for the Log Gradient, the log of the values per wavelength is derived prior to computing the gradient.

The tested regression models rely on these different feature sets as input. The SVR and RFR models are implemented using the Python library Scikit-learn [10]. The SVR model utilizes the Radial Basis Function (RBF) kernel function, the same as used in [2]. The RFR model is also configured in the same way as reported [2]. The RFR uses the absolute error to evaluate the splitting of the decision trees. See Sec. 2.3 for details on how the GLORIA data set is pre-processed. The models are evaluated using 10-fold cross-validation when the number of samples is ≥ 10. Otherwise, the number of folds is equal to the number of samples. The results reported in Table 1 are based on the predicted values on the test fold. All feature set and regression model combinations are tested and trained per Water type. For some of the values an overflow error is encountered in the exponent of Eq. (2). The associated variables have been discarded, removing in total 20 variables, less than 1 percent. In Fig. 3 three selected models are plotted for the entire data set.

4. DISCUSSION AND CONCLUSIONS

This section examines the impact of various feature sets and regression models on the estimation of SD using hyperspectral reflectance. The choice of regression models is based on [2] and alternative models that were tested through trial and error. While [2] demonstrates consistent and favorable performance with RFR, it is essential to note that [2] did not explore significant feature engineering beyond band-ratios, which corresponds to F0 here, more or less. In contrast, our feature engineering approach led to significant improvements,

Table 1: RMSE, APD and R^2 values for all subsets of the data for all the different models. An average and a weighed average, with the number of samples as weights, is also computed.

Water type	Metric	Z. Lee	SVR F0	SVR F1	SVR F2	SVR F3	RFR F0	RFR F1	RFR F2	RFR F3	N
All	RMSE	2.986	1.779	1.636	1.639	1.627	1.711	1.715	1.676	1.593	3914
Lake	RMSE	3.383	1.913	1.928	1.888	1.889	2.048	2.110	2.112	2.189	2001
Stability-1	RMSE	2.846	1.077	1.084	1.151	1.125	1.333	1.230	1.302	1.040	1410
Stability-2	RMSE	3.743	1.303	0.889	0.902	1.040	0.946	0.899	1.173	0.830	1172
Chla	RMSE	1.423	1.507	1.384	1.387	1.429	1.552	1.563	1.573	1.580	915
Stability-3	RMSE	1.620	1.920	1.654	1.612	1.569	1.713	1.564	1.521	1.401	627
TSS	RMSE	3.152	2.658	2.429	2.529	2.716	2.397	2.403	2.239	2.447	407
Clear	RMSE	7.102	2.182	2.095	2.101	2.118	2.279	2.248	2.256	2.656	327
Chla + CDOM	RMSE	1.257	1.553	1.206	1.209	1.372	1.421	1.403	1.326	1.259	315
Coastal Ocean	RMSE	0.674	1.392	1.376	1.435	1.333	1.328	1.353	1.326	1.071	63
River	RMSE	0.483	0.221	0.220	0.242	0.257	0.199	0.202	0.208	0.177	43
Estuary	RMSE	1.009	1.532	1.285	1.443	1.659	1.210	1.229	1.250	1.001	22
CDOM	RMSE	0.531	0.502	0.542	0.561	0.715	0.604	0.607	0.566	0.608	13
Turbid Coastal	RMSE	0.865	2.356	2.348	2.256	2.300	2.397	2.370	2.371	2.324	5
Other	RMSE	0.539	1.117	0.953	0.954	0.909	0.980	0.946	0.886	0.903	4
W-Avg.	RMSE	2.961	1.678	1.543	1.549	1.569	1.654	1.640	1.655	1.580	-
Avg.	RMSE	2.107	1.534	1.402	1.421	1.471	1.475	1.456	1.452	1.405	-
All	APD	47.258	41.318	34.836	35.389	35.970	39.716	40.919	40.446	37.462	3914
Lake	APD	61.991	51.184	46.292	46.142	48.723	55.586	59.410	60.490	60.781	2001
Stability-1	APD	99.937	60.467	60.881	64.788	62.892	71.339	67.623	70.943	48.865	1410
Stability-2	APD	42.417	33.698	24.629	25.256	28.088	26.725	25.922	29.814	24.682	1172
Chla	APD	46.917	43.020	33.123	33.588	38.483	46.169	48.150	48.518	49.980	915
Stability-3	APD	23.106	25.090	20.184	20.176	19.817	20.992	20.489	20.387	18.830	627
TSS	APD	86.081	73.259	61.470	68.181	77.430	66.042	66.427	62.585	65.989	407
Clear	APD	108.499	61.851	51.046	53.570	58.740	69.204	69.562	71.453	80.982	327
Chla + CDOM	APD	37.921	43.010	25.974	29.016	36.856	32.436	32.540	31.668	29.369	315
Coastal Ocean	APD	52.432	60.253	57.668	61.683	62.845	53.196	54.844	54.668	48.990	63
River	APD	50.575	22.521	22.719	23.490	25.857	18.720	18.325	20.464	16.522	43
Estuary	APD	33.682	45.470	39.745	42.882	48.837	43.519	45.589	44.381	34.770	22
CDOM	APD	30.235	26.196	29.638	32.710	41.180	31.242	30.811	29.913	33.335	13
Turbid Coastal	APD	14.128	29.306	28.597	27.836	28.765	30.190	30.946	31.145	28.368	5
Other	APD	61.832	131.938	109.163	109.196	97.342	112.803	108.941	94.579	111.236	4
W-Avg.	APD	57.523	45.770	39.419	40.614	42.451	46.258	46.978	47.743	45.021	-
Avg.	APD	53.134	49.905	43.064	44.927	47.455	47.859	48.033	47.430	46.744	-
All	R^2	0.441	0.681	0.739	0.738	0.743	0.715	0.712	0.727	0.756	3914
Lake	R^2	0.236	0.454	0.480	0.503	0.491	0.451	0.429	0.402	0.407	2001
Stability-1	R^2	0.169	0.274	0.345	0.268	0.214	0.291	0.329	0.288	0.468	1410
Stability-2	R^2	0.297	0.785	0.902	0.897	0.861	0.882	0.894	0.822	0.911	1172
Chla	R^2	0.670	0.645	0.697	0.696	0.678	0.603	0.599	0.596	0.599	915
Stability-3	R^2	0.815	0.754	0.811	0.821	0.834	0.792	0.826	0.837	0.863	627
TSS	R^2	0.334	0.274	0.460	0.380	0.218	0.484	0.488	0.561	0.475	407
Clear	R^2	-0.001	0.245	0.326	0.326	0.379	0.140	0.192	0.198	0.101	327
Chla + CDOM	R^2	0.756	0.630	0.777	0.775	0.718	0.686	0.689	0.721	0.751	315
Coastal Ocean	R^2	0.924	0.211	0.267	0.133	0.455	0.374	0.329	0.388	0.689	63
River	R^2	0.488	0.649	0.651	0.562	0.519	0.732	0.722	0.702	0.791	43
Estuary	R^2	0.837	0.463	0.758	0.629	0.494	0.667	0.652	0.638	0.813	22
CDOM	R^2	0.672	0.653	0.580	0.530	-0.151	0.454	0.452	0.524	0.417	13
Turbid Coastal	R^2	0.959	-0.654	-0.531	0.009	-0.447	-0.530	-0.384	-0.405	-0.252	5
Other	R^2	0.777	-0.331	-0.358	-0.356	-0.307	-0.303	-0.325	-0.291	-0.694	4
W-Avg.	R^2	0.391	0.568	0.636	0.626	0.609	0.598	0.602	0.595	0.636	-
Avg.	R^2	0.558	0.382	0.460	0.461	0.380	0.429	0.440	0.447	0.473	-