A Chlorophyll-a Algorithm for Landsat-8 Based on Mixture Density Networks

Brandon Smith[1,2], Nima Pahlevan[1,2], John Schalles[3], Steve Ruberg[4], Reagan Errera[4], Ronghua Ma[5], Claudia Giardino[6], Mariano Bresciani[6], Claudio Barbosa[7], Tim Moore[8], Virginia Fernandez[9], Krista Alikas[10], and Kersti Kangaro[10]

[1]NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA.
[2]Science Systems and Applications Inc. (SSAI), Lanham, MD 20706
[3]Department of Biology, Creighton University, Omaha, NE 68178 USA.
[4]Great Lakes Environmental Research Laboratory, NOAA, Ann Arbor, MI 48108 USA.
[5]Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Science, Nanjing 210008 China.
[6]National Research Council of Italy, Institute for Electromagnetic Sensing of the Environment, CNR IREA, Milan 20133 Italy.
[7]Instrumentation Lab for Aquatic Systems, National Institute for Space Research, São José dos Campos, SP, Brazil.
[8]Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL 34946 USA.
[9]Departament of Geography, Faculty of Sciences, University of the Republic, Montevideo 11400, Uruguay.
[10] Tartu Observatory of University of Tartu, Tartu, Tartumaa 61602, Estonia

Corresponding Author: nima.pahlevan@nasa.gov

Abstract— Retrieval of aquatic biogeochemical variables, such as the near-surface concentration of chlorophyll-a (Chl*a*) in inland and coastal waters via remote observations, has long been regarded as a challenging task. This manuscript applies Mixture Density Networks (MDN) that use the visible spectral bands available by the Operational Land Imager (OLI) aboard Landsat-8 to estimate Chl*a*. We utilize a database of co-located *in situ* radiometric and Chl*a* measurements (N=4354), referred to as Type A data, to train and test an MDN model (MDN_A). This algorithm's performance, having been proven for other satellite missions (Pahlevan et al. 2020b), is further evaluated against other widely used machine learning models (e.g., support vector machines), as well as other domain-specific solutions (OC3), and shown to offer significant advancements in the field. Our performance assessment using a held-out test data set suggests that a 49% (median) accuracy with near-zero bias can be achieved via the MDN_A model, offering improvements of 20 to 100% in retrievals with respect to other models. The sensitivity of the MDN_A model and benchmarking methods to uncertainties when using an atmospheric correction (AC) method, is further quantified through a semi-global matchup dataset (N=3337), referred to as Type B data. To tackle these increased uncertainties, alternative MDN models (MDN_B) are developed through

various features of the Type B data (e.g., Rayleigh-corrected reflectance spectra $\boldsymbol{\rho_s}$). Using held-out data, along with spatial and temporal analyses, we demonstrate that these alternative models show promise in enhancing the retrieval accuracy adversely influenced by the AC process. Results lend support for the adoption of $\mathrm{MDN_B}$ models for regional and potentially global processing of OLI imagery, until a more robust AC method is developed.

Index Terms—Chlorophyll-*a*, coastal water, inland water, Landsat-8, machine learning, ocean color, aquatic remote sensing.

1. Introduction

Near-surface concentration of chlorophyll-*a* (Chl*a*), a proxy for phytoplankton biomass, has been observed and quantified in aquatic ecosystems through optical remote sensing for many years (Bukata 1995; Clarke et al. 1970; Gordon et al. 1983; Smith and Baker 1982; Wezernak et al. 1976). This technique has led to the routine production of Chl*a* distributions for the global oceans for more than two decades. The heritage algorithms have used blue-green band-ratio models to estimate Chl*a* (Gordon et al. 1980; O'Reilly et al. 1998), which are realistic representations of biomass in ecosystems where other constituents, such as detritus and colored dissolved organic matter (CDOM), co-vary with Chl*a*. In inland and optically complex coastal waters however, the color of water is further modulated by the presence of organic and inorganic particles, as well as dissolved matter (Han et al. 1994; Harding et al. 1994) that do not generally co-vary with phytoplankton, rendering retrievals of Chl*a* a far more challenging task (IOCCG 2000). To improve estimates of Chl*a* in these turbid and eutrophic environments, other methods have been developed. For example, spectral bands within the red-edge (RE) region (690 – 715 nm) (Mittenzwey et al. 1992; Vos et al. 1986), and their combinations with red bands have shown to correlate well with Chl*a* in highly turbid and/or eutrophic waters (Gitelson 1992; Gitelson et al. 2007; Gower et al. 1984; Khorram et al. 1987; Munday and Zubkoff 1981; Rundquist et al. 1996).

The RE observations, however, are not available in the suite of measurements made by heritage missions – such as Landsat – which have provided the longest record of Earth observation from space (Goward et al. 2017).

The Operational Land Imager (OLI) aboard Landsat-8 was launched in February 2013 to continue Landsat's mission of monitoring Earth systems and capturing changes at relatively high spatial resolution (30 m) (Irons et al. 2012). This mission has offered significant improvements in both data quality and quantity (i.e., both spectral and spatial coverage) over previous heritage instruments (Markham et al. 2014; Markham et al. 2015; Pahlevan et al. 2014). Although several methods have been developed to retrieve Chl*a* from the four OLI visible bands (Allan et al. 2015; Concha and Schott 2016; Manuel et al. 2020; Watanabe et al. 2015), yet Chl*a* retrieval methods in inland and coastal waters using traditional approaches are challenged by optical complexity and high dynamic ranges where water types can range anywhere from very clear to highly turbid and eutrophic (Spyrakos et al. 2018). It is, therefore, critical to continue to formulate novel methodologies that enable the production of viable Chl*a* products from Landsat-8 data for global scientific studies and applications (Snyder et al. 2017). Pahlevan et al. (2020b) successfully applied Mixture Density Networks (MDNs) – a class of neural networks that estimates multimodal Gaussian distributions over a range of solutions – to Sentinel-2 and Sentinel-3 data for mapping Chl*a*. This model has further been extended to the hyperspectral domain to obtain Chl*a* and phytoplankton absorption properties from the images of the Hyperspectral Imager for the Coastal and Ocean (HICO) (Pahlevan et al. 2020a).

Our motivation for this study is to test the feasibility of using MDN algorithms extended to the Landsat-8 imagery for Chl*a* retrievals. Four different MDN models were trained, evaluated, and compared against current machine learning (ML) algorithms using the visible spectral bands. One model (MDN_A) was developed similar to that of Pahlevan et al. (2020b), using paired *in situ* Chl*a* and remote sensing reflectance (R_{rs}) (Mobley 1999), whereas three other models (MDN_B) were trained using *in situ* Chl*a* matchups and atmospherically corrected (or partially corrected) products (Cao et al. 2020). These latter models, developed to compensate for uncertainties in the atmospheric correction (AC) (Warren et al. 2019), were trained using input features comprised of: a) satellite-derived R_{rs} (hereafter referred to as R_{rs}^{Δ}); b) R_{rs}^{Δ} in combination with ancillary data; and c) intermediate Rayleigh-corrected reflectance products (ρ_s) (Wynne et al. 2013) combined with ancillary data. The manuscript follows with sensitivity analyses on: a) the contribution of different spectral bands to the outputs of the model; b) the impacts of different AC methods; and c) the implications for aquatic science and applications.

2. Chl*a* retrievals from visible bands

For satellite missions like Landsat-8 that do not support measurements in the RE region, Chl*a* algorithms tend to rely on either blue-green ratio algorithms (O'Reilly et al. 1998) or neural network (NN) models (Doerffer and Schiller 2007; Kajiyama et al. 2018) that apply all (or a subset of) bands within the visible (VIS) and near infrared (NIR) bands. Algorithms based on band ratios for Chl*a* work well in ocean environments; however, when applied to optically complex waters, such as in coastal or inland areas, performance significantly degrades (Bukata et al. 1981; Freitas and Dierssen 2019; Le et al. 2013). Most research on these environments has focused on instruments like the MEdium Resolution Imaging Spectrometer (MERIS), equipped with RE bands (Gitelson 1992; Gitelson et al. 2007; Gower et al. 2005); however, these algorithms are not viable for OLI or missions without such measurements (e.g., the Moderate Resolution Imaging Spectroradiometer (MODIS (Esaias et al. 1998), Visible Infrared Imaging Radiometer Suite (VIIRS) (Wang et al. 2014), and Geostationary Ocean Color Image (GOCI) (Ryu et al. 2012)).

Thus, the only widely used Chl*a* estimation algorithms available are those of the band-ratio Ocean Color (OC) family (e.g., OC3), a combination of those (Neil et al. 2019), or ML models. Regional and local algorithms specific to OLI imagery have also been attempted with some success in lakes and reservoirs (Allan et al. 2015; Watanabe et al. 2015).

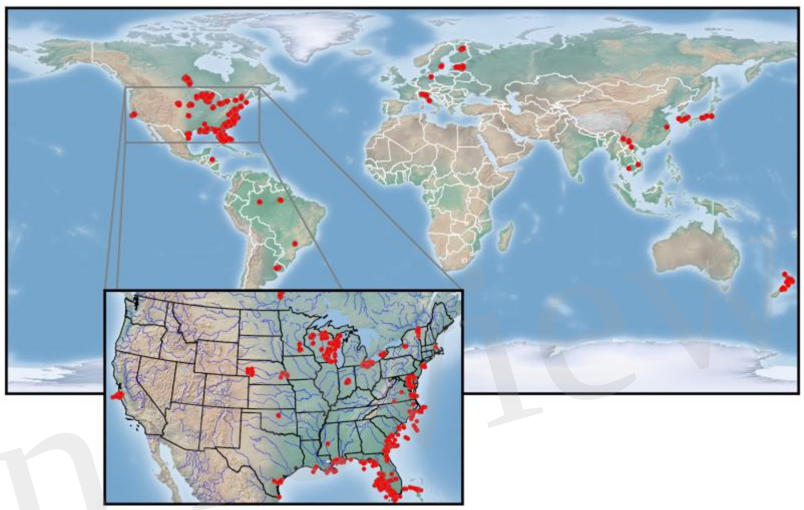

Fig. 1. Geographic distribution of in situ measurements (Type A) (N=4354)

Over the years, several generic ML methods have been utilized in the OC or aquatic remote sensing domain. Among them, Multilayer Perceptrons (MLP), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGB) have shown promise in retrieving Chl*a*. MLPs are NNs with feed-forward connections arranged in a series of layers, which perform regression by learning a set of weights that are used in dot products through sequential layers (Hinton 1990). This type of model has been employed in past research to obtain various bio-optical parameters as well as Chl*a* (Chen et al. 2014; Gross et al. 2000; Hieronymi et al. 2017; Ioannou et al. 2011; Jamet et al. 2012; Schiller and Doerffer 1999; Vilas et al. 2011). SVMs on the other hand, perform regression by finding a maximal separation hyperplane, fitting the training samples within some margin of tolerance. This margin is tunable, and influences over- and under-fitting (Chang and Lin 2011).

Similarly, this method has previously been utilized for Chl*a* estimations in open ocean environments (Kwiatkowska and Fargion 2003; Zhan et al. 2003). Lastly, XGB is a highly optimized tree-based method which fits a series of models to the training data, incrementally reducing the error through gradient boosting – a specific type of ensembling focusing on the error gradient as the target (Chen and Guestrin 2016). This approach has been proven to improve Chl*a* retrievals from OLI ρ_s products in highly turbid or eutrophic lakes in China (Cao et al. 2020). The MDNs utilized in this research is a variation of MLPs that learn a probability distribution over the output space to allow for multimodal target distributions (Section 4.1). This multimodality is a fundamental characteristic of inverse problems, owing to the non-unique relationships between input and output features (Sydor et al. 2004).

Table 1. Data distribution

	North America	South America	Europe	Asia	Australia/Oceania
Type A	2836	121	794	552	51
Type B	2287	46	568	1	420

3. Datasets

Two datasets are utilized in this study: paired *in situ* Chl*a* - R_{rs} measurements (Type A); and near-simultaneous Chl*a* - R_{rs}^{Δ} satellite matchups (Type B). Using both Type A and Type B datasets provides significant benefits in understanding an algorithm's performance. Co-located R_{rs} - Chla measurement pairs (Type A) provide the theoretically ideal environment, as performance on this dataset quantifies the quality of estimates when applied to sample spectra with minimal noise. Near-simultaneous satellite matchups (Type B), on the other hand, provide a practical demonstration of an algorithm's capability when applied with significant noise. Both sets have

Chl*a* ranging from 0.1 to > 1000 mg/m^3 (as shown in Figs. 2 and A1). Continental distributions of the datasets are provided in Table 1.

3.1. Type A: *In situ* data

Type A data consists of radiometric and biogeochemical parameters that have been collected and assembled from various lakes, bays, estuaries, coast lines, and rivers from around the world (Fig. 1), covering a wide range of trophic states and geographic locations (Pahlevan et al. 2020b). The frequency distribution of Chl*a*, total suspended solids (TSS), and the absorption by CDOM at 443 nm (a_{CDOM} (440)) is shown in Fig. 2. Although our *in situ* measurements are not void of uncertainties, this dataset has proven useful for model development and validation (Pahlevan et al. 2020b), representing the closest to ideal while still considering instrument and human errors.

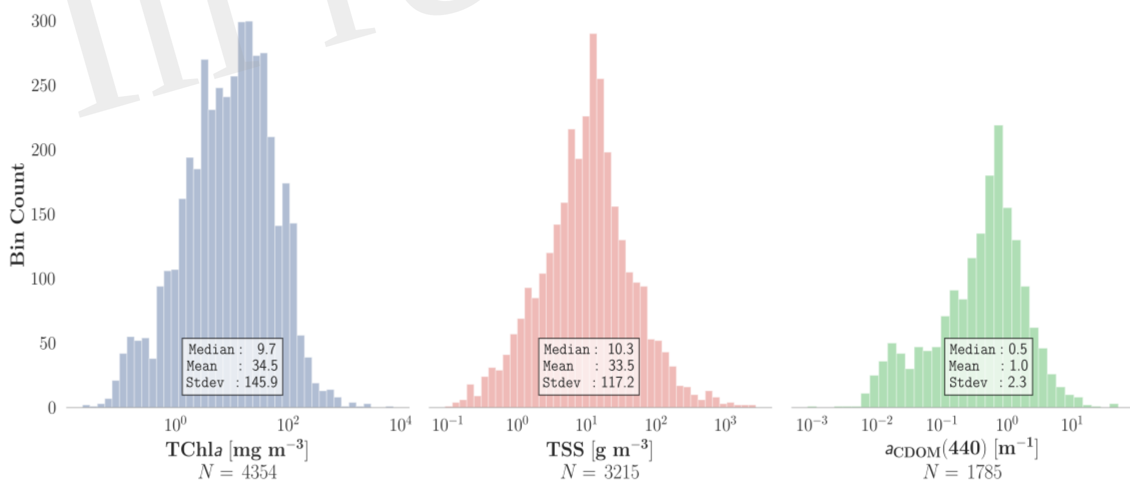

Fig. 2. Distribution of Chla, TSS, and a_{CDOM}(440) data (Type A)

The radiometric quantity primarily used for model developments in this study is the remote sensing reflectance (R_{rs}):

$$R_{rs}(\lambda) = L_w(\lambda) \,/\, E_d(\lambda) \qquad (1)$$

R_{rs} (sr^{-1})_is determined using the water-leaving radiance L_w and downwelling irradiance E_d in the air, just above the water surface. Hyperspectral R_{rs} spectra were resampled according to the OLI's relative spectral response function. Furthermore, the data was preprocessed before being used as input into any machine learning models, with R_{rs} data transformed according to a robust median-centering interquartile range (IQR) scaling process (fit to the training data); and the Chl*a* values being log-scaled, and transformed to be within the interval [-1, 1]. Type A data were used for training and validation of the first MDN model (MDN_A; Section 4.3), and for performance assessment against that of other Chl*a* algorithms (Section 5.1).

3.2. Type B: Satellite matchups

The satellite matchup dataset (Type B) is composed of two sources: Level-1TP OLI scenes, and *in situ* Chl*a* measurements made during cruises, at buoys, or via site visits carried out through routine monitoring activities. The *in situ* data were obtained via a search of national/international water quality databases, such as the Water Quality Portal (WQP), NOAA's Chesapeake Bay Interpretive Buoy System (CB), Environment and Climate Change Canada (ECCC), the World Ocean Database (WOD), and others. The data acquired from CB are near-surface calibrated fluorometry data, while remaining sites represent near-surface or depth-integrated measurements. The application of Type B data is twofold: a) to evaluate the performance of the MDN model applied to atmospherically corrected products (R_{rs}^{Δ}) (Section 4.2); and b) to explore alternative inputs for MDN models (MDN_B types) using R_{rs}^{Δ} and ρ_s products, combined with ancillary data (Section 4.3). The former analysis enables gauging the performance of MDN models in practice, comparing its efficacy with our previous studies (Pahlevan et al. 2020a; Pahlevan et al. 2020b) and quantifying how uncertainties in R_{rs}^{Δ} propagate to Chl*a* products. The motivation behind focusing on R_{rs}^{Δ} and ρ_s products (the latter analysis) is to identify alternative approaches (e.g., Cao et al. (2020)) to use the provisional, readily accessible, United States Geological Survey aquatic reflectance products (Franz et al. 2015; Pahlevan et al. 2017b). The locations of all successful OLI matchups are mapped in Fig. 3 (Section 4.2), and the histograms of relevant parameters are shown in Appendix A.

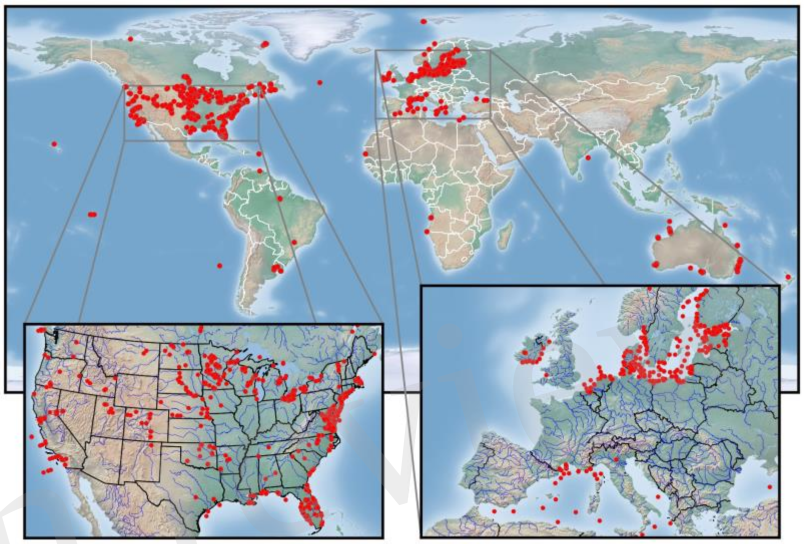

Fig. 3. Geographic distribution of satellite matchups (Type B) (N = 3371)

4. Method

4.1. Mixture Density Network (MDN)

Radiative transfer theory details a series of equations that are concerned with the forward problem: given a set of parameters which describe the inherent optical properties (IOPs) of the water, concentrations of water constituents (e.g., Chl*a*), and a set of boundary conditions, which limit the environment itself, how the relevant apparent optical properties (AOPs) can be discerned (Mobley 1994)**Error! Reference source not found.**. In the same manner, the standard target of a model in machine learning takes the form of a forward problem: given a set of independent variables, the goal is to find a function which approximates the relationship between these and the dependent variables. In particular, the relationship must be right-unique, which guarantees there is a single set of true outputs (y) for any given set of input (x) variables in a dataset D:

$$\forall(\mathrm{x},\mathrm{y}) \in D \land (x, y') \in D \Rightarrow y = y' \qquad (2)$$

Plainly, for any input-output pair in a dataset, any samples with the same input must also maintain the same output (conditioned on noise). Inverse problems reverse the relationship however, i.e. switching x and y, which leads to violations of this core assumption. In natural environments, bio-optically active constituents and illumination conditions *cause* observed R_{rs}; the same set of input parameters, with perfect knowledge, should always lead to the same R_{rs}. In the inverse formulation, we attempt to instead determine bio-optical parameters and biogeochemical properties from the R_{rs} observations, and thus have the possibility of a single R_{rs} spectrum leading to multiple sets of valid parametric solutions (and so, multiple valid environments in which the spectrum might have been observed) (Pahlevan et al. 2020a; Pahlevan et al. 2020b).

Mixture Density Networks (MDN) (Bishop 1994) are a class of neural networks which attempt to address this one-to-many mapping (Defoin-Platel and Chami 2007; Sydor et al. 2004). Where a standard neural network (e.g. MLP) directly models the R_{rs} => Chl$a$ relationship, MDNs model a conditional probability distribution, i.e., $p(\mathrm{Chl}a|\, R_{\mathrm{rs}})$, over the R_{rs} => Chl$a$ mapping as a mixture of multiple (c) Gaussian functions:

$$p(\mathrm{Chl}a|\, R_{\mathrm{rs}}) = \sum_{i=1}^{c} \pi_i(R_{\mathrm{rs}})\, \phi_i(\mathrm{Chl}a|\, R_{\mathrm{rs}}) \qquad (3)$$

$$\phi_i(\mathrm{Chl}a|\, R_{\mathrm{rs}}) = \frac{exp(-\frac{1}{2}(\mathrm{Chl}a - \mu_i)^T {\Sigma_i}^{-1}(\mathrm{Chl}a - \mu_i))}{\sqrt{(2\pi)^d|\Sigma_i|}} \qquad (4)$$

With c mixture components and dimensionality d, μ being the mean vector, and ϕ denoting a

Gaussian distribution. A valid Gaussian mixture requires that the mixing coefficients π and the covariance matrix Σ adhere to the constraints explained in detail in Bishop (1994). The final model estimate is then taken to be the maximum likelihood, which represents the area of highest probability mass. In our formulation, Bootstrap Aggregation (bagging) (Breiman 1996) is also applied to the model to improve the quality and consistency of estimates. The practice of bagging is an ensemble technique which is intended to reduce variance and improve generalization. In short, the idea is to repeatedly resample the available training set into a smaller subset (in practice, 50-75% of the original size) and train the model on this new, randomly sampled training subset. After some number of models is added to the ensemble, the median of all model estimates is taken as the final output.

4.2. Atmospheric correction and matchup selection

There are several viable AC methods suitable for OLI data processing; nonetheless we focused only on one processing chain, i.e., the SeaWiFS Data Analysis System (SeaDAS), the heritage ocean color AC processing scheme adopted for OLI (Franz et al. 2015). This processing approach is also adopted by the USGS Earth Resource Operation and Science center (EROS) to produce aquatic reflectance products, which are equivalent to R_{rs} products normalized by π. In this study, OLI images were not only fully processed to R_{rs} but also partially processed to output intermediate ρ_{s} that is corrected for atmospheric gaseous absorption, molecular scattering effects, and air-water interface multiple scattering phenomena (Gordon 1997). To compare the effects of AC schemes on Chl*a* retrievals, a single OLI image was processed by three other methods: Polynomial based algorithm applied to MERIS (POLYMER) (Steinmetz et al. 2011), Atmospheric Correction for OLI lite (ACOLITE) (Vanhellemont and Ruddick 2018), and Case-2 Extreme Waters (C2X)

(Brockmann et al. 2016) (Section 6.2).

To create satellite matchup datasets (Type B), SeaDAS-processed OLI scenes were paired with *in situ* measurements on same-day overpasses and the matchup criteria proposed in Bailey and Werdell (2006) was followed using strict spatiotemporal filters to remove matchups with questionable quality. A 3x3-element box centered on the closest geographic coordinates of the *in situ* measurement was used to select potential satellite observations (Fig. 3), and any matchup was discarded if four or more pixels were flagged as invalid. Further, any cruise samples < 500 m apart were considered duplicates and removed. Temporal mismatch criteria were further tightened for dynamic aquatic ecosystems (e.g., Chesapeake Bay, riverine systems) to 30 minutes to minimize the associated uncertainties. The median value of valid pixels within the 3x3-element box was then derived for each parameter and preprocessed in the same manner as the Type A data. All relevant reflectance products, including top-of-atmosphere reflectance (ρ_t), ρ_s, and R_{rs}^{Δ}, as well as ancillary (anc) parameters (e.g., sensor and solar viewing geometries, water vapor content, etc.), were simultaneously extracted, filtered, and stored.

4.3. MDN Model Types

The naming convention used for MDN model developments will follow the format listed in Table 2. Each of the MDN models was trained with 50% of the total available samples within the respective dataset, chosen uniformly at random (with the same set of samples used to train all ML models; Section 4.4). The remaining, held-out portion of the dataset was then used to test the models. The "bagging" scheme was also applied to all ML models – in order to ensure a fair comparison between algorithms – with 75% of the training data used per bagging estimator (without replacement), and an ensemble size of 10 estimators. All hyperparameters of the benchmark models were chosen via a 5-fold cross validation grid search on the training data. For a detailed discussion on MDN hyperparameters, see Appendix C.

In those MDN models which utilize ancillary data, these features are added alongside their respective R_{rs}^{Δ} (or ρ_s) spectra as inputs. Bagging was also applied to these ancillary features when they are included (i.e. keeping only a random 75% of the ancillary features for each estimator in the bagging ensemble); the full VIS R_{rs}^{Δ} (i.e. 443, 482, 561, 665nm) or ρ_s (443, 482, 561, 665, 865, 1609nm) spectrum for a sample was always included as input. Ancillary data (Appendix B) included per pixel imaging geometry, coarse scale wind parameter estimations, and other general atmospheric condition variables which were available from SeaDAS (e.g. NO2, O3, water vapor). We hypothesize that these additional features help the model to learn the biases and uncertainties specific to the AC method, which uses these features in their derivations of R_{rs}^{Δ}. For instance, wind parameters are known to correlate well with sunglint signal (Kay 2009; Wang and Bailey 2001), and are not utilized by default in SeaDAS. Unaccounted water vapor absorption, which affects the OLI's red and SWIR bands, can also introduce additional uncertainties in R_{rs}^{Δ}. On the other hand, ρ_s products contain aerosol scattering and absorption governed by imaging geometry parameters; hence, the sun-sensor geometries must be evaluated when retrieving Chl*a* through $MDN_B^{\rho_s,anc}$, a point not taken into consideration in other studies which utilize similar features (Cao et al. 2020).

Table 2. MDN model types. VIS indicates OLI's visible bands.

Model	Input feature	Data Type	Number of Samples		
			Training	*Testing*	*Total*
MDN_A	R_{rs}(VIS)	A	2177	2177	4354
MDN_B	R_{rs}^{Δ}(VIS), Δt	B	1686	1685	3371
MDN_B^{anc}	R_{rs}^{Δ} (VIS), anc, Δt				
$MDN_B^{\rho_s,anc}$	$\rho_s(443 \leq \lambda \leq 1609), anc, \Delta t$				

In order to help prevent the models from learning spurious relationships due to temporal misalignment, we added one additional feature to all MDN_B models, which represents the number of minutes between the satellite overpass and the *in situ* measurement, i.e., Δt. This number is negative if the *in situ* measurement was taken prior to the overpass, and positive if after. When applying the model to a scene to generate Chl*a* maps (see Section 5.2), this feature is simply set to 0 for all pixels for the exact time of the overpass.

4.4. Benchmarking

Given their previous application in the aquatic remote sensing area, MLP, SVM, and XGB were the main ML models identically trained and tested with the MDN_A model. Due to its simple implementation and successful application in classification problems, K Nearest Neighbor (KNN) was also added as another benchmark (Altman 1992). In spite of its expected performance loss in waterbodies rich in organic or inorganic material, the OC3 model was also used as another benchmark (Franz et al. 2015), owing to its global utility. The MDN_B models were further benchmarked against another XGB model, hereafter referred to by its name in original publication (BST), developed and tested by Cao et al. (2020).

To quantify performance, we primarily examined three metrics: Median Symmetric Accuracy (Morley et al. 2018), referred to as “Error” in all plots and tables; Symmetric Signed Percentage Bias (Morley et al. 2018), referred to as “Bias” in all plots and tables; and the slope of the least-squares linear regression line on the log-transformed data (Campbell 1995). All three have straightforward interpretations, though to clarify the first two:

- Median Symmetric Accuracy (“Error”) can be interpreted as a symmetric percentage error, equally penalizing over- and under-estimation. Lower values indicate better performance, with perfect accuracy being assigned a value of 0%.

$$Error = 100 \times (e^{median(\,|\,log(\,\widehat{\text{Chl}a}\,/\,\text{Chl}a)\,|\,)} - 1) \tag{5}$$

- Symmetric Signed Percentage Bias ("Bias"), as with the former metric, is interpretable as a percentage bias that maintains symmetry between over- and under-estimation. Values closer to zero indicate better performance, with positive values indicating over-estimation and negative indicating under-estimation.

$$MdLQ = median(\,log(\,\widehat{\text{Chl}a}\,/\,\text{Chl}a\,)\,) \tag{6}$$

$$Bias = 100 \times sign(MdLQ) \times (e^{\,|\,MdLQ\,|} - 1) \tag{7}$$

In the equations above (5-7), Chl$a$ denotes the *in situ* value and $\widehat{\text{Chl}a}$ is the estimated value. Both metrics are designed to address the widely documented drawbacks (Hyndman and Koehler 2006; Makridakis 1993; Tofallis 2015) of other commonly used statistical measures (e.g. MAPE). For a thorough discussion on the advantages of these methods over others commonly found in literature, we direct the reader to (Morley et al. 2018). One should also note that the above metrics are zero-centered and symmetric compared to recently proposed metrics in Seegers et al. (2018).

5. Results

5.1. Performance assessment

The performance of MDN_A, the other ML models, and OC3 and the corresponding statistical metrics (N=2177) are provided in Fig. 4. The MDN model outperforms other ML models with improvements in error ranging from 30 to 60%, and MLP ranking as the second-best performer. In fact, given the coarse hyperparameter grid search performed for other ML methods, and the lack of equivalent optimization for MDN parameters, the improvement in performance is even more significant than shown here (Appendix C). The MDN model, as well as other ML models, remarkably outperform OC3, which overestimates Chl*a* in the 1-10 mg/m^3 range and underestimates in the higher range. Yet, compared to Pahlevan et al. (2020b)[1], the performance is worse since R_{rs} simulated for the MultiSpectral Instrument (MSI) and Ocean Land Color Instrument (OLCI) contain the RE bands (Gitelson et al. 2007).

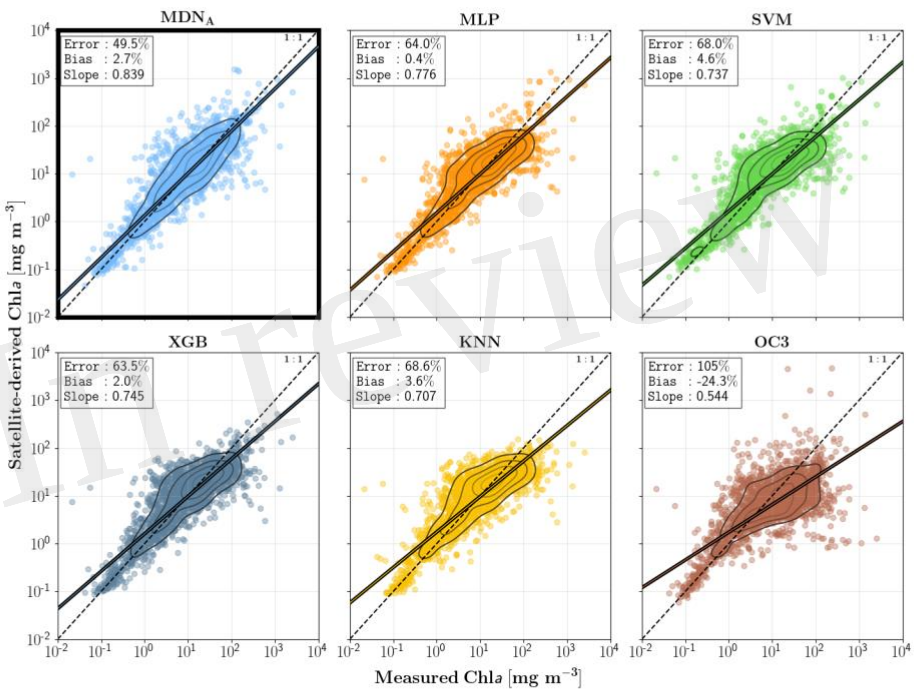

Fig. 4. Performance of all algorithms on the in situ dataset (Type A; N=2177). MDN$_A$, despite using suboptimal hyperparameters, still outperforms all other surveyed algorithms. The contour lines correspond with density estimates.

To gauge the level of noise introduced through the AC (SeaDAS), we also produced Chl*a* scatter plots via MDN$_A$ and other benchmark algorithms applied to the Type B dataset (i.e., satellite matchups) (Fig. 5). The model performances degrade considerably when applied to SeaDAS-derived R_{rs}^{A}. Error levels exceeding 100% render the utility of OLI-derived Chl*a* somewhat impractical for robust scientific applications (Bresciani et al. 2018; Trinh et al. 2017). Further, all

[1]Comparison with Pahlevan et. al. 2020b is possible through the relationship: Error = 100 x (MAE-1)

the models tend to overestimate Chl*a* > 1 [mg m^{-3}], suggesting primarily underestimated R_{rs}^{Δ}. This is in agreement with other studies in which the confounding effects of AC have been highlighted; OLI-retrieved R_{rs}^{Δ} over coastal and inland waters are known to carry these major biases and uncertainties due to imperfections in the AC process (Ilori et al. 2019; Kuhn et al. 2019; Pahlevan et al. 2017a; Werdell et al. 2009; Zibordi et al. 2009). The AC processor introduces varying degrees of error in Chl*a* estimates, due to issues in the shape and magnitude of R_{rs}^{Δ} spectra estimated. Regardless, while evaluating Chl*a* retrievals against the satellite matchups is informative, the differences between the Type A and Type B sets are too large to draw any firm conclusions about the underlying models (see Figs. 2 and A1). In Type B data, error is introduced by spatial differences, temporal differences, adjacency effects, variability in atmospheric conditions, and image artifacts, to name a few. Performance metrics will therefore reflect these errors as much as any error inherent to the models themselves.

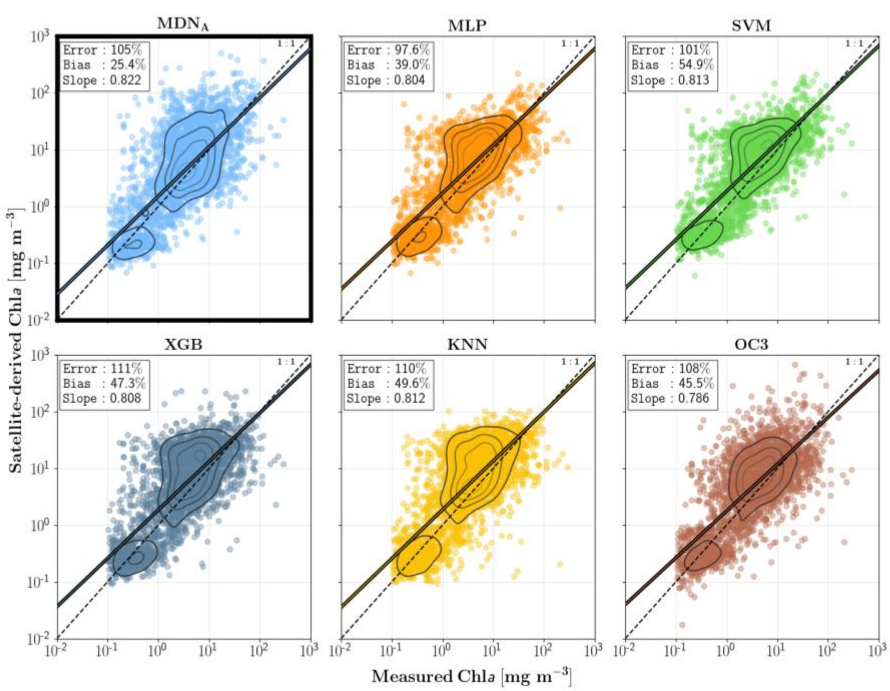

Fig. 5. OLI-retrieved Chla evaluated using satellite matchups, i.e., Type B dataset (N=3371). SeaDAS was used for the atmospheric correction in all cases.

In spite of these issues, Landsat-matchup trained MDN model (MDN_B) is to some extent capable of improving the accuracy as compared to that of the original MDN_A model (Fig. 6). In essence, the model accounts for some uncertainties inherent to the R_{rs}^{Δ} products – though still exhibiting a fair amount of error due to the limited information contained in the four available R_{rs}^{Δ} bands. This error is reduced further for MDN_B^{anc}. By including ancillary data, this MDN model compensates for uncertainties from AC sources in SeaDAS retrievals of R_{rs}^{Δ}. With some of the uncertainties in AC addressed (e.g., imperfect accounting of sun-sensor geometry; (Gilerson et al. 2018; Pahlevan et al. 2017b)), the model can, in general, make more accurate estimates. The comparable performance of $MDN_B^{\rho_s,anc}$ demonstrates viable retrievals in areas with both highly eutrophic and/or turbid waters (Bailey et al. 2010) and increased water-surface signal induced by residual and/or moderate sunglint (Harmel et al. 2018). The recently published model in Cao et al. (2020) (BST) is also added as another benchmark, which poorly predicts Chl*a* – likely because of the drastic differences in the distribution of training data, i.e., the median Chl*a* in Cao et al. (2020) was > 10 [mg m^{-3}].

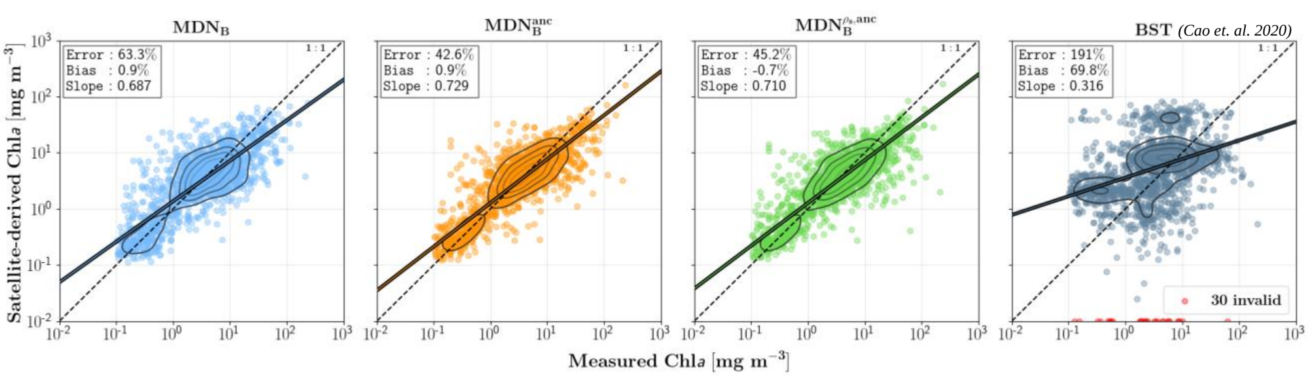

Fig. 6. Performance assessment of MDN$_B$ models using half of the Type B dataset (N=1685). The performance of the original model by Cao et al. (2020) is included. Red dots indicate negative estimates or failure.

5.2. Spatial analysis

The different models examined in Section 4 are retrained using the full datasets, rather than splitting into training and/or testing sets. This allows for the model development to have the widest range of data available. Two scenarios are demonstrated here.

As a first example, a natural color image of Lake Erie and the derived products during a harmful algal bloom event on Sept. 14th, 2015 was used (Fig. 7). The black "x" markers indicate the positions of the three monitoring stations visited by the Great Lakes Environmental Research Laboratory (GLERL) within ± 1 hr of Landsat-8 overpass. High concentrations of Chl*a* in the southwestern section of the lake are evident in the natural color image generated from ρ_s products. The elevated backscatter evident in in the natural color image of the northern and central Lake St. Claire discharging into Lake Erie is commonly attributed to suspended sediments and resuspension events (Avouris and Ortiz 2019; Bukata et al. 1988; Czuba et al. 2011; Hawley and Lesht 1992). Table 3 contains the extracted Chl*a* measurements and the Chl*a* estimates from the MDN models. Although there is a slight temporal mismatch between Landsat-8 overpass and *in situ* measurements, there is a clear pattern in the results which quantitatively supports the accuracy improvement made by the MDN models.

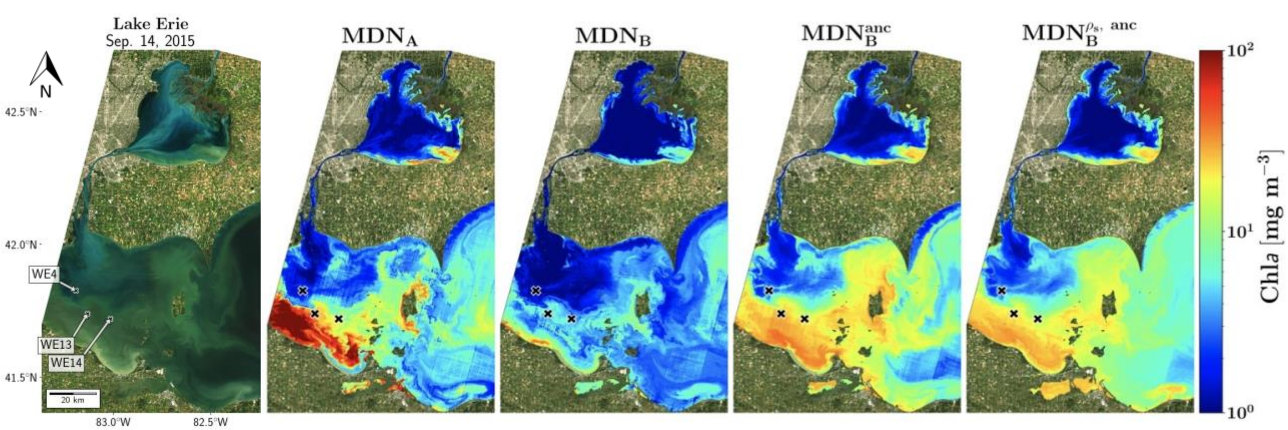

Fig. 7. Algorithm estimates for Lake Erie (Sept. 14th, 2015), with a bloom event occurring in the south-western portion of the lake apparent in the natural color image (left column).

One point to note is the apparent underestimation of the MDN Type B models. This can be at least partially attributed to the AC frequently failing in highly eutrophic waters (Wang et al. 2019): since the MDN_B models are only trained on samples (Fig. 3 and A1) for which SeaDAS gives a valid result, there will necessarily be a bias towards lower concentration samples within the training data. This bias does not appear to be present in the MDN_A model, as it would not have such a selection bias in its training set. Taking WE13 station as an example, we note that in spite of the reported concentrations > 50 mg m^{-3}, OC3 estimates a maximum of around 10.1 mg m^{-3}, with the majority of examined areas below 11 mg m^{-3} (Table 3). $\mathrm{MDN_B}$ exhibits similar behavior, possibly due to the previously discussed selection bias in the training data. However, when ancillary features are added to the model inputs (as is the case with $\mathrm{MDN_B^{anc}}$), the estimated Chl*a* concentrations become far more plausible. Furthermore, there is a notable consistency between the

Table 3. In situ Chl*a* data collected near-coincident (±1 hr) with Landsat-8 overpasses. The estimated Chl*a* from various MDN models and OC3 are also tabulated.

Site	Station	*In situ*	MDN_A	MDN_B	MDN_B^{anc}	$MDN_B^{\rho_s,anc}$	OC3
Lake Erie	WE14	45.6	18.2	5.7	**21.3**	16.7	9.4
	WE13	53.9	**45.9**	6.2	28.7	22.2	10.1
	WE4	6.5	1.4	1.1	2.3	**2.5**	**2.5**
San Francisco	33	8.8	15.6	6.6	5.4	**6.8**	5.57
	32	7.9	**8.9**	5.8	5.3	5.5	5.10
	31	6.2	5.9	**6.0**	5.2	5.6	5.12
	30	6.0	5.6	3.9	5.6	**5.7**	5.17
	29.5	7.2	**8.3**	3.6	5.6	5.9	5.15
Error [%]		NA	15.0	83.7	55.9	42.3	56.4

maps predicted from MDN_B^{anc} and $MDN_B^{\rho_s,anc}$, implying ρ_s has the potential to be used as a substitute for R_{rs}^{Δ}. The benefit of this substitution is demonstrated in the southern area of the image: for a large proportion of Sandusky Bay, R_{rs}^{Δ} through SeaDAS is unavailable and thus missing within the MDN_B^{anc} map.

Our second example is comprised of a scene over the San Francisco Bay (SFB), imaged on April 27th, 2017 at 18:43 GMT (Fig. 8), for which there were a few near-simultaneous *in situ* Chl*a* measurements (Table 3), provided by SFB monthly cruises. In contrast to MDN_B-derived maps, the map obtained from MDN_A shows highly trophic areas (> 20 mg m^{-3}) in the south bay region. Retrievals from MDN_A were similar to *in situ* measurements (in the lower bay) except for the estimate at station 33 – which is far greater than the measured concentrations (Table 3). This might be caused by any number of factors, not least of which being those biases inherent to the AC process. An interesting feature in Fig. 8 is the Chl*a* field outside the bay (or in San Pablo Bay) that has been merely predicted by $MDN_B^{\rho_s,anc}$, suggesting the advantage of this model over other models limited by the failure in the AC (e.g., due to haze or over highly turbid/eutrophic waters).

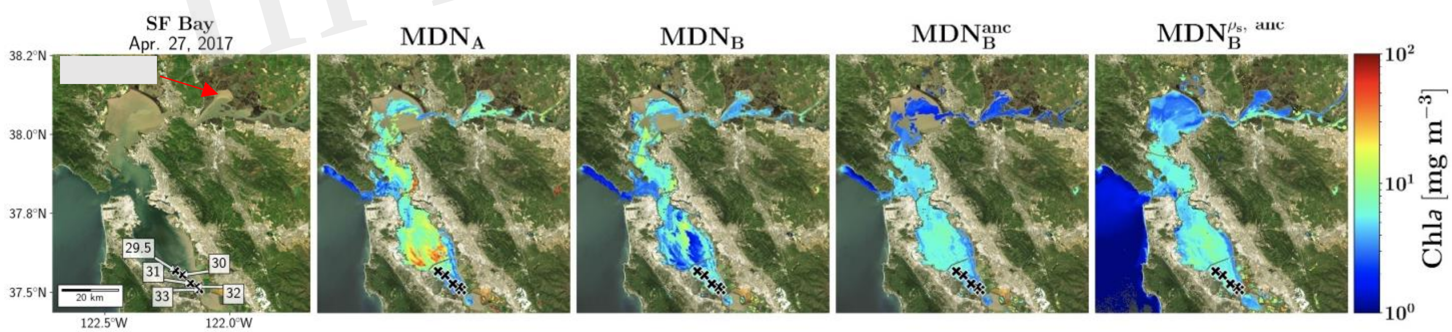

Fig. 8. Algorithm estimates for San Francisco (SF) Bay (April 27th, 2017) shown along the natural color image (left column). Note the successful retrievals via $MDN_B^{\rho_s,anc}$ in the Pacific coastal waters where SeaDAS did not return valid R_{rs}^{Δ}. In situ Chla measurements for the stations are provided in Table 3. The Grizzly Bay station whose time-series Chla data are shown in Fig. 9.

5.3. Temporal analysis

The time-series of the MDN_B^{anc} and $MDN_B^{\rho_s,anc}$ performances for Chl*a* were compared with *in situ* measurements (Fig. 9). In this case, the *in situ* data (N ~ 30) were measured via calibrated autonomous fluorometers deployed near-surface in Grizzly Bay, the northern section of the SFB region (Fig. 8). The errors (Eq. 5) for the different models amounted to 54%, 41%, 120%, and

241% for $\mathrm{MDN}_{\mathrm{B}}^{\mathrm{anc}}$, $\mathrm{MDN}_{\mathrm{B}}^{\rho_s,\mathrm{anc}}$, $\mathrm{MDN}_{\mathrm{A}}$, and OC3, respectively. The $\mathrm{MDN}_{\mathrm{A}}$ predictions, although exhibiting enhancements over OC3, contain outliers, but the $\mathrm{MDN}_{\mathrm{B}}^{\rho_s,\mathrm{anc}}$ and $\mathrm{MDN}_{\mathrm{B}}^{\mathrm{anc}}$ predictions resemble the *in situ* measurements without considerable noise, suggesting their viability if adequate training data are supplied. This time-series analysis highlights the primary role of AC in achieving high-quality Chl*a* retrievals and the need for improvement.

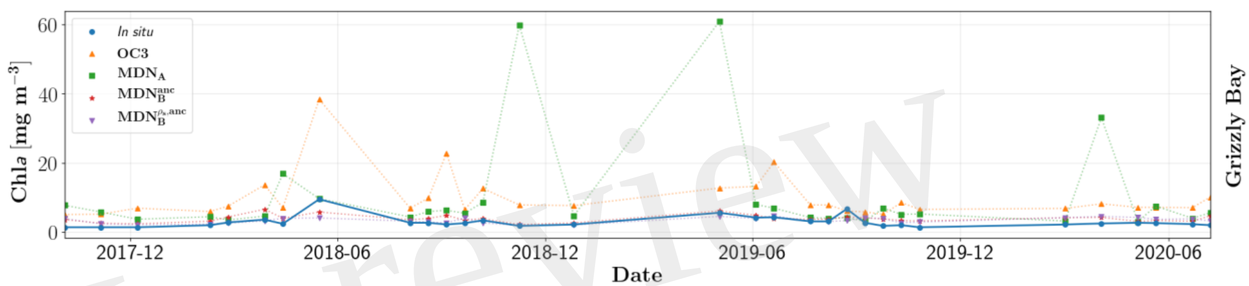

Fig. 9. Time-series data extracted from in situ fluorometric Chla measurements along with estimates derived from different versions of MDN models and OC3.

6. Discussion

Based on these results, one infers that the MDN is a promising model in retrieving Chl*a* from OLI, offering improvements in accuracy over other current models. Here, we further address why this model is a viable choice for retrievals and demonstrate its strength in suppressing noise in R_{rs} compared to other ML models. This is followed by a discussion on the impacts of varying AC methods on the performance of $\mathrm{MDN}_{\mathrm{A}}$ and the implications of this research for studying and monitoring global waterbodies using Landsat-8 and other missions.

6.1. Model Validity

Neural network models have long been regarded as black-box models, with their complexity being a double-edge: providing more accurate solutions than have been previously available, at the cost of understanding the rationale in their estimations. This loss of explicability is of great concern for those involved in critical applications, due to the costs incurred in the event of failures. Without a source to identify as the cause – as is often the case in these models – the trust placed in the application is eroded. Recent research has led to a number of methods which allow for better model transparency, however. For instance, with many models it can be helpful to visualize the effect a given input feature has on the output of the model; in this case the effect of a certain band on the Chl*a* estimation.

One such method to do this is called an Accumulated Local Effects (ALE) plot (Apley and Zhu 2016). The interested reader may also examine the literature for the related Partial Dependence (PD) plot – though these have the disadvantage of assuming independence between input features, and so are not the best choice in this case. ALE plots, on the other hand, calculate the effect of a feature *conditional* upon the other input features. Another way to explain this is, they examine the

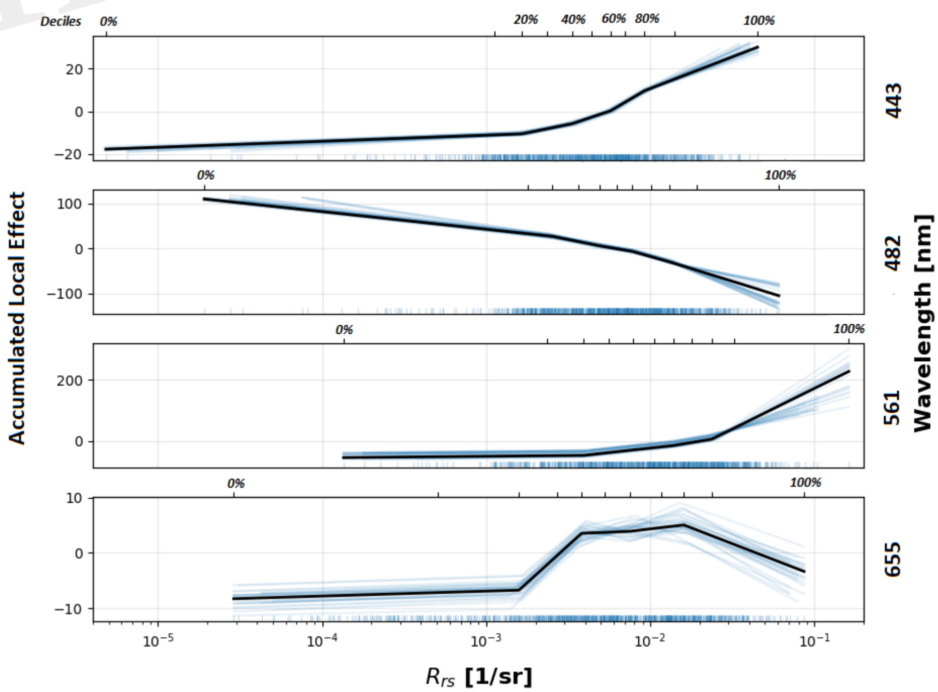

Fig. 10. ALE plots showing the sensitivity of MDN Chla estimates, with respect to changes in the input bands. Y-axes measure the accumulated local effect, which can be thought of as "change in estimated Chla."

average change in prediction over a window around an input feature's values, only conditioning upon other features in areas for which values exist in the data set. Fig. 10 shows the ALE plots for OLI bands, generated via the Type A data set. Note that the y-axis values correspond to the accumulated local effect, which can be thought of as "change in estimated Chl*a"*. These plots indicate that 561nm, when observed with a large magnitude, has the greatest (positive) effect on chlorophyll-a estimation. Not surprisingly, 482nm also appears to significantly impact chlorophyll estimates with an inverse relationship: low magnitude reflectances indicating a higher than average chlorophyll-a value, and high magnitude reflectances indicating a lower than average chlorophyll-s value. On the other hand, since the 655-nm band does not fully cover Chl*a* absorption at 676 nm, it appears to contain limited spectral information pertaining to Chl*a* (see Helder et al. (2018)).

6.2. Impacts of atmospheric correction

Although the primary processing scheme used here was SeaDAS, here, we underscore the importance and challenges associated with the AC. To that end, C2X, ACOLITE, and POLYMER, in addition to SeaDAS, were implemented to a sample OLI scene over Lake Peipsi, June 14th, 2016, followed by applying the MDN_A model to all the derived R_{rs} products. Fig. 11 shows the corresponding Chl*a* map products. The inconsistency in the relative distribution and the overall magnitude of products is largely noticeable. For example, C2X appears to predict a large bloom in the center of the lake whereas other schemes provide values closer to the lake-wide average estimate. Moreover, SeaDAS and ACOLITE tend to estimate relatively high Chl*a* in the southern and eastern basins while POLYMER retrieves only slightly higher-than-average estimates. Same-day *in situ* Chl*a* measurements and the estimated Chl*a* from MDN_A and OC3 from the four processors are included in Table 4 (Alikas et al. 2015). Despite that the *in situ* dataset does not represent the entirety of the ecosystem, it allows to better comprehend the complexity induced through the AC process and how confusing the output products may be. Given the statistics in Table 4, there is no single processor that distinctly outperforms the rest for this instance of OLI image and/or lake. It is worth noting that while SeaDAS, ACOLITE, and POLYMER statistically yield better Chl*a* estimates via MDN$_A$, retrieved Chl*a* values from C2X through OC3 resembles *in situ* samples more closely. This observation and the discrepancies in the performances further corroborates the need for an improved AC method for the OLI data processing to achieve the theoretical limit shown in Fig. 4.

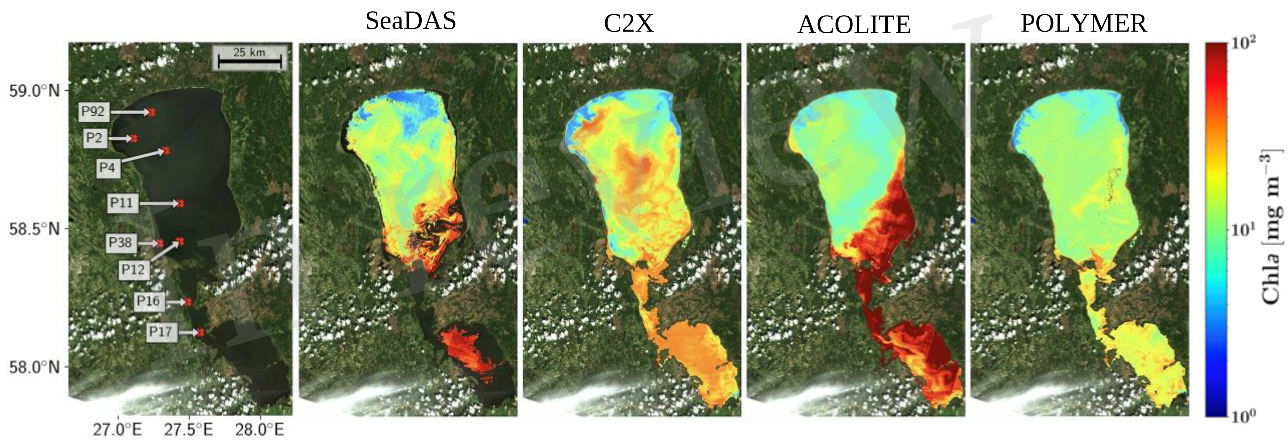

Fig. 11. An OLI scene acquired on June 14th, 2016 processed via four different atmospheric correction methods. The MDN$_A$ model was used to generate Chla from the output of the processors.

6.3. Implications for aquatic studies

Landsat-8 data, when combined with the data from Sentinel-2 and -3, are expected to allow for near-daily global observations of inland and coastal waters. Irrespective of differences in their observation modalities, creating consistent Chl*a* products is key for successful assessment and monitoring of these ecosystems. Considering the missing RE measurements in the OLI suite of observations however, retrieving Chl*a* as accurately as that with MSI and OLCI appears challenging. Although the number of matchups assessed is different, comparing to our previous results (Pahlevan et al. 2020b), it can be inferred that R_{rs}, or their equivalent R_{rs}^{Δ} and ρ_s, within

Table 4. *In situ* Chl*a* and retrieved Chl*a* derived from four different AC processors by applying MDN_A and OC3. The boldfaced values correspond to best estimates at each station.

		SeaDAS		C2X		ACOLITE		POLYMER	
Station	***In Situ***	**MDN_A**	**OC3**	**MDN_A**	**OC3**	**MDN_A**	**OC3**	**MDN_A**	**OC3**
P17	36.3	NA	NA	**32.5**	32.3	168.5	7.39	23.6	10.0
P16	37.2	NA	NA	3.9	**25.6**	73.8	5.29	12.2	9.5
P12	15.7	48.3	51.08	25.8	**18.5**	54.7	4.64	11.0	8.8
P38	10.3	8.7	119.62	4.4	12.3	**10.2**	3.28	8.8	5.7
P11	13.8	**10.6**	15.60	26.9	18.2	5.9	3.26	17.3	9.0
P4	9.2	14.4	15.90	27.1	21.3	12.0	3.70	11.7	**8.4**
P92	14.0	5.7	23.00	27.5	21.6	**14.4**	3.47	10.3	8.7
P2	10.6	13.9	22.72	21.0	19.0	**10.6**	3.55	11.3	8.4
	Error [%]	43.3	92.5	97.3	38.4	61.0	269.5	**31.5**	69.4
	Bias [%]	**5.2**	92.5	78.9	25.5	15.8	-269.5	-26.1	-69.4

the RE region contain significant information related to Chl*a* in eutrophic waters. These channels also help constrain the solution space in less eutrophic waters with higher turbidity, which may be mislabeled as having high Chl*a* with OLI. That said, the addition of spectral information within the 865 nm band may prove valuable under such circumstances (Cao et al. 2020), which has been the case for our $\mathrm{MDN}_{\mathrm{B}}^{\rho_s,\mathrm{anc}}$ model.

In addition to SeaDAS, we also implemented and tested ACOLITE and POLYMER to assess the performance of MDN_A models. Our analyses showed that these alternative models yield Chl*a* as inaccurate as that from SeaDAS (Fig. 5). The performance of MDN_B and MDN_B^{anc} using R_{rs}^{Δ} derived from ACOLITE and POLYMER showed consistent performances however, similar to those illustrated in Fig. 6. Therefore, it is surmised that until major improvements in the state-of-the-art AC methods are achieved, an alternative approach to obtaining improved Chl*a* is through $\mathrm{MDN_B}$ models supplied with R_{rs}^{Δ}, ρ_s, or a combination of both. Assuming adequate matchups spanning a wide array of trophic states and aerosol conditions are incorporated in training (beyond what was used in our Type B dataset; Fig. A1), such a model should provide global retrievals nearly as robust as those obtained by MSI and OLCI. This is likely the path forward, given the extent to which AC degrades the performance of MDN_A (rendering it virtually equivalent to OC3). This approach is, in particular, applicable to regional monitoring sites (e.g., western Lake Erie, Lake Taihu) where ample high-quality, historic *in situ* datasets are available for model training. In other words, a successful compilation of high-quality discrete samples of Chl*a* at global scales is a challenging task and may take several years to achieve.

Spatial and temporal mismatches inherent to satellite matchups introduce further uncertainties in our assessments. Of concern is, in particular, our same-day criteria. We made an attempt to diminish the impact of this noise source by supplying Δt to the MDN_B models, in the hopes that the model could adjust for the importance of temporally distant samples during the learning process. Overall, choosing an optimal threshold for the temporal filtering is a trade-off between the accuracy of the model, and the generalization capability conferred by a wider range of samples and environments.

The inclusion of ancillary data, such as the solar angles, sensor viewing angle, wind data, water vapor, and others, enhanced the model performance noticeably when added to model input features, regardless of AC processor. The improvements stem from how SeaDAS utilizes the ancillary information itself (Mobley et al. 2016) while some of the parameters are not often used, e.g., wind speed, wind angle. In some cases, the algorithm may apply simplifications, for example to reduce computational burden, that may preclude a rigorous integration of ancillary data in the process. In particular, we found that our model is very sensitive to sensor azimuth angles, which change sign for the two adjacent focal plane modules (Markham et al. 2014). Our spatial analysis suggested that including these angles yield alternate low-high Chl*a* for the odd and even focal plane modules (Pahlevan et al. 2017b); hence, we decided to discard this information, which led to more spatially uniform maps. Further, it is worth pointing out that most ancillary variables (e.g., water vapor concentration) are coarse-resolution features with little to no per-pixel variability. Therefore, any fine structures or patterns seen in the estimates can only be influenced by the spectral information itself.

6.4. Future work

This work introduces a great number of potential directions for exploration. From the perspective of the aquatic remote sensing field as a whole, it is yet to be determined how the MDN model fares when applied to other missions. In the future, the performance evaluation is expected to be carried out for other ocean color missions, such as MODIS, which do not measure in the RE region but provide relevant spectral content in the vicinity of 750-nm region. Similarly, the MDN developed for OLI's visible bands might be further extended by including the panchromatic band to further constrain the solution space (Castagna et al. 2020). As the atmospheric correction process has been shown to introduce significant errors in downstream products of such missions, and ρ_s being a feasible substitute to bypass portions of this process, the question becomes whether it is possible to bypass AC completely and allow for direct retrieval of the relevant biogeochemical properties. Alternatively: whether there are certain AC-specific parameters which might be tuned, in order to provide a more amenable input for learning the product inverse function.

Other directions include those focused more on ML, and the MDN itself. For instance, the MDN model also has the capability to simultaneously estimate multiple products; to what extent do the inclusion of additional variables in the model output (e.g. TSS) affect performance? Intuitively these additions should serve only to improve accuracy overall, given the additional information of target covariances – but which products might be estimated synergistically is yet to be explored.

More theoretically, we might ask if there are non-Gaussian distributions (e.g. Laplace, which may better represent the data); or, whether the learned mixture components might relate to the physical environments of the samples assigned. Further exploration is required in regard to the model hyperparameters, and the mixture components especially. There are very likely advancements in the field of machine learning (e.g. activation functions, batch normalization procedures, convolutional / temporal architectures, etc.) which could also be applied to enhance retrievals – though potentially requiring alternative data formulations, such as incorporating spatial or temporal information.

7. Conclusion

In this work, we have gathered a global dataset of both R_{rs} – Chl*a* and R_{rs}^{Δ} – Chl*a* matchups, compiled from a variety of different sources. These two datasets are used to train several machine learning (ML) models, and in particular, used to train the Mixture Density Network (MDN) – an algorithm which we introduce as a theoretically plausible model for biogeochemical variable retrieval via remotely sensed radiometric data. These ML algorithms are benchmarked against each other, in order to provide empirical justification alongside the theory of MDN superiority on inverse problems; as well as against OC3 to demonstrate accuracy on the described task. Furthermore, we show that instead of using the standard R_{rs}^{Δ} values as input, it is feasible to instead use ρ_s to achieve similar performance in Chl*a* estimation. The benefits of this are briefly touched upon, where ρ_s-trained models are shown to seamlessly retrieve Chl*a* where previously unavailable due to AC failure.

The MDN algorithm represents a promising step towards the goal of global simultaneous biophysical and biogeochemical variable retrieval, in the context of aquatic remote sensing. While results are promising, much work is left to be done in both data acquisition and model validation. To truly design a global-scale model, capable of approximating an inverse solution to the radiative transfer equations, significantly more data is required. Simultaneously retrieving all parameters of interest to the community requires the potential dataset to have the necessary information to learn relevant covariances in all atmospheric conditions. Just as important, the various sources of uncertainty and misalignment must also be minimized in order for the model to accurately learn these relationships.

We conclude with broad discussions of other justifications and benefits, analyses on the hyperparameters, implications of the model within the broader community, and potential directions for further experimentation. These discussions are far from exhaustive, but we hope they will provide the seed for future advancements in remote sensing.

Acknowledgement

We would like to recognize all the individuals and entities that acquired, processed, and prepared in situ data that are central to the development of global algorithms. The principal investigators providing the data include Caren Binding, Daniela Gurlin, Steve Greb, Bunkei Matsushita, Anatoly Gitelson, Wesley Moses, Moritz Lehman, and Michael Ondrusek. We also acknowledge NASA's support in creating and maintaining SeaBASS. We acknowledge the European Union's Horizon 2020 research and innovation program (grant agreement no. 730066, EOMORES) to support *in situ* data collection in Estonian inland waters. Nima Pahlevan is funded under NASA ROSES contract # 80HQTR19C0015, Remote Sensing of Water Quality element, and the USGS Landsat Science Team Award # 140G0118C0011.

Appendix A

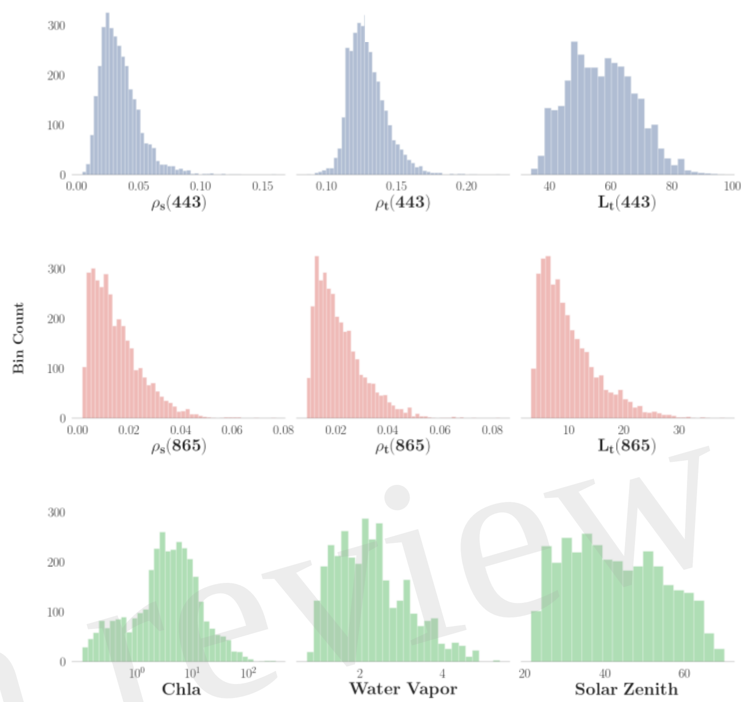

Fig. A1. Satellite matchup (Type B) distributions for select parameters (N = 3337). Wide ranges in spectral band parameters indicate the representation of many atmospheric and imaging conditions within the dataset. ρ_s, ρ_t, and L_t correspond to Rayleigh-corrected reflectance, top-of-atmosphere (TOA) reflectance, and TOA radiance, respectively. L_t is in units of W m^{-2} sr^{-1} μm^{-1}.

Appendix B

Table B1. Ancillary parameters

Name	Description	Mean	Min	Max
humidity	Relative humidity (%)	70	35	91
no2_frac	Fraction of tropospheric NO2 above 200m	0.741	0.379	0.977
no2_strat	Stratospheric NO2 (molecules/cm^2)	3.2e+15	1.8e+15	4.9e+15
no2_tropo	Tropospheric NO2 (molecules/cm^2)	3.0e+15	1.2e+14	1.9e+16
ozone	Ozone concentration (cm)	0.305	0.246	0.401
pressure	Surface pressure (millibars)	1010	804	1024
mwind	Meridional wind speed at 10m (m/s)	0.302	-5.324	5.887
zwind	Zonal wind speed at 10m (m/s)	0.795	-7.121	4.354
windangle	Wind direction at 10m (degree)	-61	-179	180
windspeed	Wind speed at 10m (m/s)	1.709	0.009	9.071
scattang	Scattering angle (degree)	138	107	166
senz	Sensor zenith angle (degree)	4.405	0.236	8.603
sola	Solar azimuth angle (degree)	136	33	177
solz	Solar zenith angle (degree)	43	21	70
water_vapor	Precipitable water vapor (g/cm^2)	2.307	0.544	5.389
Δt	Difference in time between satellite overpass and *in situ* measurement (minutes)	8.881	-1420	1365

Of the ancillary features, only windangle is treated as periodic – having an end point equal to its start point (0° = 360°). In order for a machine learning algorithm to understand this periodicity, we separate the variable into two separate features:

$$feature_1 = sin(\ 2\pi \ \times \ feature \ / \ 360 \ )$$

$$feature_2 = cos(\ 2\pi\ \times\ feature\ /\ 360\ )$$

These two features are included in the full input set, and any bagging applied includes these as well. Other features which have degree units do not span the entire range of values (e.g. only allowing 20° to 70°), and so are left as their original value. These may also benefit from a conversion however, which will be explored further in future work where ancillary feature contributions are examined in more detail.

Appendix C

While no hyperparameter optimization has taken place over the course of this work, it is an important step to ensure optimal performance and generality. All models examined in this study have simply used 'reasonable' default values for their hyperparameters (Glorot and Bengio 2010; Hinton 1990; Kingma and Ba 2014) namely: a five layer neural network with 100 nodes per layer, learning a mixture of five gaussians; a learning rate, L2 normalization rate, and ϵ value all set to 0.001; and with training performed via Adam optimization over a set of 10,000 stochastic mini-batch iterations, using a mini-batch size of 128 samples.

These choices are mostly arbitrary, as the goal of this study was to present the feasibility and theoretical backing of the MDN model. A full optimality demonstration, on the other hand, would require a significantly longer study than already presented. Nevertheless, we would be remiss to exclude any discussion which examines the hyperparameter choices, and so what follows is a (very) brief look at how select parameters affect performance. Note that for the sake of brevity, all data used within this section is from the Type A dataset; the noise inherent to the Type B set would likely overshadow any actual information about the algorithm that might be gleaned from such an experiment. Nevertheless, results should be generalizable between Type A and Type B datasets.

First, some terminology must be defined in order to make it clear what is being examined. Normally in hyperparameter optimization, and machine learning in general, the dataset is split into three parts: training, validation, and testing. The training set is of course used to train the model on; the validation set is used to optimize the model using data unseen during training; and the testing set is used only at the end, in order to benchmark performance.

As mentioned, no explicit hyperparameter optimizations have taken place thus far in the study. Default values were chosen based on those commonly used in literature and in available libraries (e.g. scikit-learn), and as will be seen, do not represent the optimal values for our specific data. As such, no separate validation set was set aside for the purposes of such an exercise.

Lest this call into question any results in readers' minds, it should be pointed out that we have also shown similar model performance on a completely independent dataset in this study (Type B), using the exact same default parameter choices. This set has, in effect, functioned as the true testing set, isolated from even any unintended optimization choices made over the course of algorithm implementation.

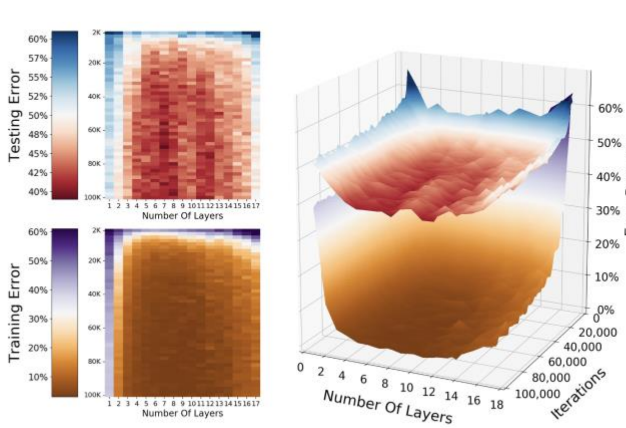

Fig. C1. Error surfaces that examine the training and testing performance over a wide variety of training iterations, and number of network layers. Heatmaps correspond to a flattened version of the three-dimensional surfaces.

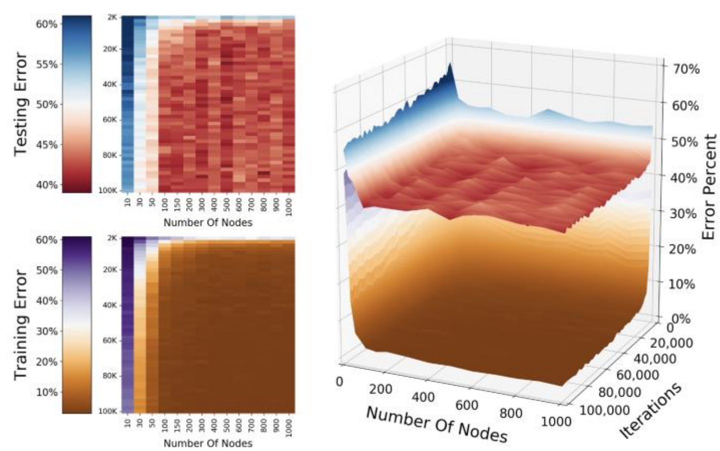

Fig. C2. Error surfaces that examine the training and testing performance over a wide variety of training iterations, and number of nodes within each network layer. Heatmaps correspond to a flattened version of the three-dimensional surfaces.

With this stated, subsequent plots will retain the “Training” and “Testing” labels as indicating those samples from the Type A set which are seen by the model when learning weights, and unseen until final estimations take place, respectively. While Section **Error! Reference source not found.**1 (Network Size) maintains the original 50%-50% training-testing split as used elsewhere in this study, the reader should note that Section **Error! Reference source not found.**2 (Learning Curves) switches to a 75%-25% split in order to more thoroughly show the extent to which training set size impacts performance.

1) Network Size

One of the main questions any reader may have is, “how large is the model?”. Naturally, the size of the underlying network determines a great deal about learning capacity, and similarly, the propensity to overfit. Shown in Fig. C1 and Fig. C2 are two components of network size – number of layers, and number of nodes per layer – plotted against the number of iterations for which the model is trained. In each image, two different perspectives of the same information is shown: the heatmap of the error simply shows a flattened version of the 3D perspective.

From Fig. C1, we see that essentially any number of layers between four and twelve performs well, with the optimal number being seven or eight. The slight increase in error for layer counts of nine and ten, given we do not see such behavior in the training set performance, is likely explained by randomness in both the train/test data split, and in the initialization weights of the networks. It is also likely that performance could be maintained beyond the limit we currently see at twelve layers, by including additional regularization techniques (e.g. dropout) in the model implementation.

Fig. C2 shows an even greater range of viable options for the number of nodes per layer, with all choices larger than 50 nodes providing a network which performs well. Ordinarily we would see wider networks begin to exhibit overfitting with further training iterations; i.e. where testing error begins to increase as more iterations are performed, even as training error continues to decline. However, we speculate that due to regularization the bagging procedure provides, along with a fairly aggressive L2 parameter setting, 'widening' the networks never seem to reach the point where overfitting becomes apparent.

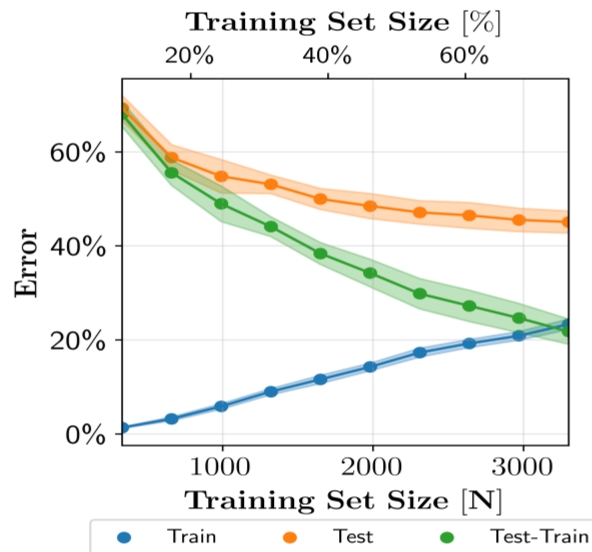

Fig. C3. Learning curves showing how the number of training samples affects performance. "Test-Train" refers to the distance between the testing and training curves. Testing set size is fixed at 1100 samples (25%), while an increasing number of samples are added to the training set. The characteristic pattern of increasing training error, and decreasing testing error, is seen.

Note that the performance in both figures appears to reach ~40% error or lower, which improves upon the 50% error shown earlier in the paper. This can likely be reduced even further, through additional hyperparameter optimizations. Again however, the focus of this manuscript is not to thoroughly optimize the model, but rather to provide demonstration and documentation of the viability.

2) Learning Curves

Though not an explicit hyperparameter *per se*, the training set size is an important factor in model performance. Much can be learned by viewing how training and testing error rates change as the amount of data provided to the model increases (Amari et al. 1992; Thoma 2017). In particular, these learning curve plots can help determine how much impact additional data may have in improving model accuracy – and in turn, how much effort ought to be expended upon collecting such data.

Furthermore, learning curves can help diagnose a model's bias and variance: how well the model is fit to the training data, and how well it generalizes that knowledge outside of training. There is a tradeoff between the two characteristics, which is inherent to any algorithm: a model with lower bias will have a higher variance due to overfitting the training set (and thus generalizing poorly); a model with higher bias has underfit the training data, and so will have high error but low variance when predicting unseen data (Thoma 2017). Optimally, we want to find some middle ground where both bias and variance are relatively low.

In relation to learning curves this is shown by the height of the training and testing curves, as well as the space between them. For instance, Fig. C3 shows the learning curves for our default model parameters. We can see that while the test set error is high, the space between the two lines is converging fairly rapidly as more data is added. This tells us two things: that the model has a relatively high bias (shown via high training error, which has not yet plateaued); and that adding more data will very likely improve performance further. Observe that the training set size also reaches up to 75% (3300 samples) of the overall available data – meaning 25% (1100 samples) of the data is always held-out as the testing set.

Due to the overall complexity of the MDN model, we can assume that, in general, it is a fairly low bias algorithm. A number of factors are likely affecting the high bias / low variance we see in Fig. C3, such as the learning rate, the L2 normalization rate, and the number of iterations for which we train. As we have already seen in the prior section, the hyperparameter choices are not optimal for this model. Perhaps most importantly, the number of training iterations (10K) are too few – thus the model is currently underfitting the training set.

To examine this relationship between dataset size and training iterations, we can follow the same procedure as in the prior section. Shown in Fig. C4 are the same training, testing, and difference curves, expanded into a third dimension (training iterations). Smoothing has been applied to the 3D surfaces to better show the overall trends, so care should be taken when analyzing any minor features, as these may simply be artifacts of the smoothing process. Though it may be more difficult overall to interpret the three-dimensional error surfaces compared to previous figures, the heatmaps show clearly a number of important patterns, and retain the data in its original form without any smoothing applied.

First, looking at the testing error heatmap, we see that along the y-dimension (iterations), performance appears to eventually plateau at a given training set size. While further training iterations may slightly improve performance, there are significantly decreasing returns beyond a certain point. It is interesting to note however, that the model never appears to overfit the training set – which would be indicated by a subsequent increase in testing error with further iterations (especially for small training set sizes). This can likely be attributed to a fairly high regularization pressure through L2 normalization and bagging.

In the same subplot, we also see that a given level of performance requires fewer iterations as more data is made available to train on. Overall, we can conclude that further additions to training data is imperative for improving test set performance, with marginal increases to model complexity likely required to reap the full benefits.

Second, in examining the training error heatmap, we see virtually the opposite pattern as before: increasing the amount of training data available decreases the performance on that set. This is exactly as we expect, as it becomes harder to perfectly fit a set of data as more complex relationships need to be modeled. Similarly, increasing the number of training iterations allows

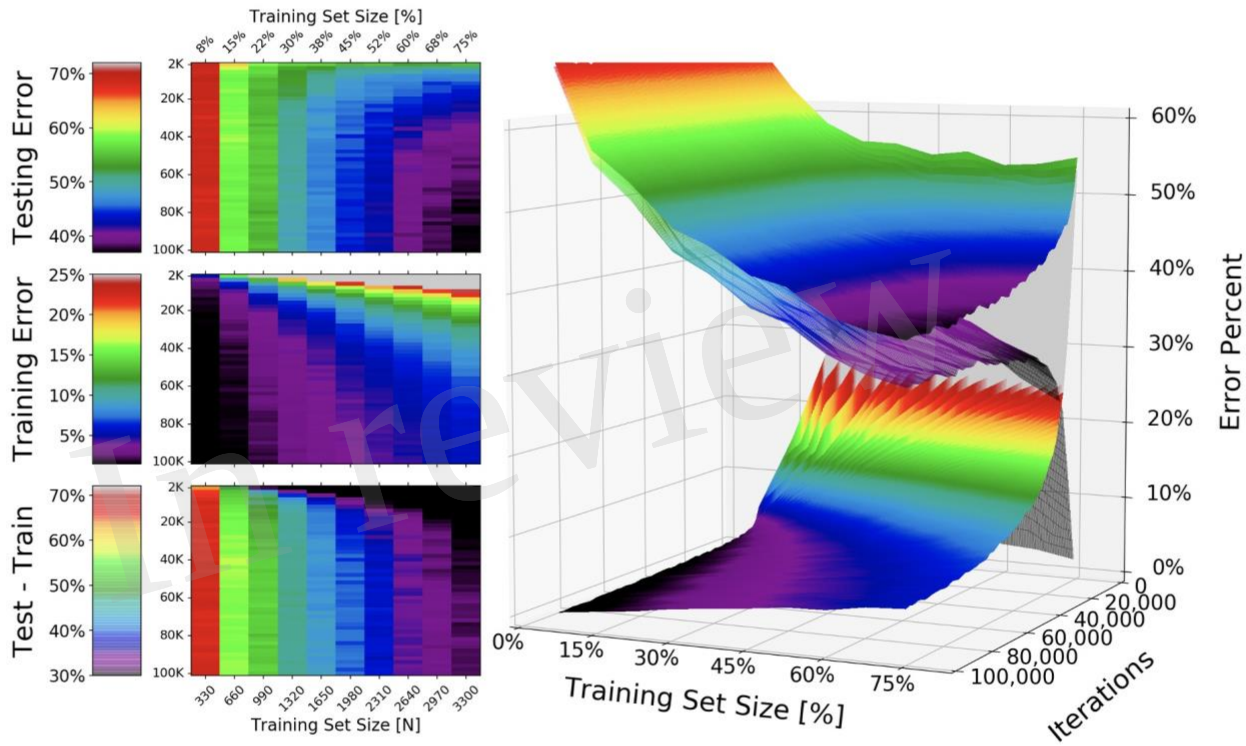

Fig. C4. Learning curves extended into a third dimension, showing how the curves evolve with respect to number of training iterations. Most importantly, it can be inferred that significant improvement can still be gained through the addition of further training samples.

the MDN to retain the same level of performance as the training data is expanded (in particular, observe the ~5% dark blue band). Again, we can conclude that increasing model complexity, either through number of training iterations or some other hyperparameter, is necessary to fully take advantage of the information contained within additional training data.

Finally, the heatmap showing the space between the two error surfaces (“Test - Train”) summarizes the relationship between the training and testing error. From this subplot, we see that increasing the number of training iterations also increases the distance between surfaces, i.e.

increases variance. Again, this is as expected, given higher complexity models exhibit lower bias and higher variance. What is interesting about this subplot, however, is how much impact additional data has on the metric: as further samples are included in the training set, the distance between the surfaces quickly closes without any sign of a plateau. In contrast, the increase in distance with further training iterations *does* appear to plateau. These observations reinforce the idea that the current bottleneck to model performance is in data availability, rather than the current model's complexity.

References

Alikas, K., Kangro, K., Randoja, R., Philipson, P., Asuküll, E., Pisek, J., & Reinart, A. (2015). Satellite-based products for monitoring optically complex inland waters in support of EU Water Framework Directive. *International Journal of Remote Sensing, 36*, 4446-4468

Allan, M.G., Hamilton, D.P., Hicks, B., & Brabyn, L. (2015). Empirical and semi-analytical chlorophyll a algorithms for multi-temporal monitoring of New Zealand lakes using Landsat. *Environmental monitoring and assessment, 187*, 364

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*, 175-185

Amari, S.-i., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation, 4*, 605-618

Apley, D.W., & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*

Avouris, D.M., & Ortiz, J.D. (2019). Validation of 2015 Lake Erie MODIS image spectral decomposition using visible derivative spectroscopy and field campaign data. *Journal of Great Lakes Research, 45*, 466-479

Bailey, S.W., Franz, B.A., & Werdell, P.J. (2010). Estimation of near-infrared water-leaving reflectance for satellite ocean color data processing. *Optics express, 18*, 7521-7527

Bailey, S.W., & Werdell, P.J. (2006). A multi-sensor approach for the on-orbit validation of ocean color satellite data products. *Remote Sensing of Environment, 102*, 12-23

Bishop, C.M. (1994). Mixture Density Networks. NCRG/94/004. Aston University Birmingham http://www.ncrg.aston.ac.uk.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123-140

Bresciani, M., Cazzaniga, I., Austoni, M., Sforzi, T., Buzzi, F., Morabito, G., & Giardino, C. (2018). Mapping phytoplankton blooms in deep subalpine lakes from Sentinel-2A and Landsat-8. *Hydrobiologia, 824*, 197-214

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S., & Ruescas, A. (2016). Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. *ESASP, 740*, 54

Bukata, R., Jerome, J., Bruton, J., Jain, S., & Zwick, H. (1981). Optical water quality model of Lake Ontario. 1: Determination of the optical cross sections of organic and inorganic particulates in Lake Ontario. *Applied Optics, 20*, 1696-1703

Bukata, R.P., Jerome, J.H., & Bruton, J.E. (1988). Particulate concentrations in Lake St. Clair as recorded by a shipborne multispectral optical monitoring system. *Remote Sensing of Environment, 25*, 201-229

Bukata, R.P., Jerome, J.H., Kondratyev, K.Y, Pozdnyakox, D.V. (1995). *Optical Properties and Remote Sensing of Inland and Coastal Waters*. New York: CRC Press

Campbell, J.W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans, 100*, 13237-13254

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sensing of Environment, 248*, 111974

Castagna, A., Simis, S., Dierssen, H., Vanhellemont, Q., Sabbe, K., & Vyverman, W. (2020). Extending Landsat 8: Retrieval of an orange contra-band for inland water quality applications. *Remote Sensing, 12*, 637

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST), 2*, 1-27

Chen, J., Cui, T., Ishizaka, J., & Lin, C. (2014). A neural network model for remote sensing of diffuse attenuation coefficient in global oceanic and coastal waters: Exemplifying the applicability of the model to the coastal regions in Eastern China Seas. *Remote Sensing of Environment, 148*, 168-177

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794)

Clarke, G.L., Ewing, G.C., & Lorenzen, C.J. (1970). Spectra of backscattered light from the sea obtained from aircraft as a measure of chlorophyll concentration. *Science, 167*, 1119-1121

Concha, J.A., & Schott, J.R. (2016). Retrieval of color producing agents in Case 2 waters using Landsat 8. *Remote Sensing of Environment, 185*, 95-107

Czuba, J.A., Best, J.L., Oberg, K.A., Parsons, D.R., Jackson, P.R., Garcia, M.H., & Ashmore, P. (2011). Bed morphology, flow structure, and sediment transport at the outlet of Lake Huron and in the upper St. Clair River. *Journal of Great Lakes Research, 37*, 480-493

Defoin-Platel, M., & Chami, M. (2007). How ambiguous is the inverse problem of ocean color in coastal waters? *Journal of Geophysical Research: Oceans, 112*

Doerffer, R., & Schiller, H. (2007). The MERIS Case 2 water algorithm. *International Journal of Remote Sensing, 28*, 517-535

Esaias, W.E., Abbott, M.R., Barton, I., Brown, O.B., Campbell, J.W., Carder, K.L., Clark, D.K., Evans, R.H., Hoge, F.E., & Gordon, H.R. (1998). An overview of MODIS capabilities for ocean science observations. *Ieee Transactions on Geoscience and Remote Sensing, 36*, 1250-1265

Franz, B.A., Bailey, S.W., Kuring, N., & Werdell, P.J. (2015). Ocean color measurements with the Operational Land Imager on Landsat-8: implementation and evaluation in SeaDAS. *Journal of Applied Remote Sensing, 9*, 096070-096070

Freitas, F.H., & Dierssen, H.M. (2019). Evaluating the seasonal and decadal performance of red band difference algorithms for chlorophyll in an optically complex estuary with winter and summer blooms. *Remote Sensing of Environment, 231*, 111228

Gilerson, A., Carrizo, C., Foster, R., & Harmel, T. (2018). Variability of the reflectance coefficient of skylight from the ocean surface and its implications to ocean color. *Optics express, 26*, 9615-9633

Gitelson, A. (1992). The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing, 13*, 3367-3373

Gitelson, A.A., Schalles, J.F., & Hladik, C.M. (2007). Remote chlorophyll-a retrieval in turbid, productive estuaries: Chesapeake Bay case study. *Remote Sensing of Environment, 109*, 464-472

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In, *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256)

Gordon, H.R. (1997). Atmospheric correction of ocean color imagery in the Earth Observing System era. *Journal of Geophysical Research: Atmospheres, 102*, 17081-17106

Gordon, H.R., Clark, D.K., Brown, J.W., Brown, O.B., Evans, R.H., & Broenkow, W.W. (1983). Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates. *Applied Optics, 22*, 20-36

Gordon, H.R., Clark, D.K., Mueller, J.L., & Hovis, W.A. (1980). Phytoplankton pigments from the Nimbus-7 Coastal Zone Color Scanner: comparisons with surface measurements. *Science, 210*, 63-66

Goward, S.N., Williams, D.L., Arvidson, T., Rocchio, L.E., Irons, J.R., Russell, C.A., & Johnston, S.S. (2017). Landsat's Enduring Legacy: Pioneering Global Land Observations from Space

Gower, J., King, S., Borstad, G., & Brown, L. (2005). Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. *International Journal of Remote Sensing, 26*, 2005-2012

Gower, J., Lin, S., & Borstad, G. (1984). The information content of different optical spectral ranges for remote chlorophyll estimation in coastal waters. *Remote Sensing, 5*, 349-364

Gross, L., Thiria, S., Frouin, R., & Mitchell, B.G. (2000). Artificial neural networks for modeling the transfer function between marine reflectance and phytoplankton pigment concentration. *Journal of Geophysical Research: Oceans, 105*, 3483-3495

Han, L., Rundquist, D., Liu, L., Fraser, R., & Schalles, J. (1994). The spectral responses of algal chlorophyll in water with varying levels of suspended sediment. *International Journal of Remote Sensing, 15*, 3707-3718

Harding, L.W., Itsweire, E.C., & Esaias, W.E. (1994). Estimates of phytoplankton biomass in the Chesapeake Bay from aircraft remote sensing of chlorophyll concentrations, 1989–92. *Remote Sensing of Environment, 49*, 41-56

Harmel, T., Chami, M., Tormos, T., Reynaud, N., & Danis, P.-A. (2018). Sunglint correction of the Multi-Spectral Instrument (MSI)-SENTINEL-2 imagery over inland and sea waters from SWIR bands. *Remote Sensing of Environment, 204*, 308-321

Hawley, N., & Lesht, B.M. (1992). Sediment resuspension in lake St. Clair. *Limnology and Oceanography, 37*, 1720-1737

Helder, D., Markham, B., Morfitt, R., Storey, J., Barsi, J., Gascon, F., Clerc, S., LaFrance, B., Masek, J., & Roy, D. (2018). Observations and Recommendations for the Calibration of Landsat 8 OLI and Sentinel 2 MSI for improved data interoperability. *Remote Sensing, 10*, 1340

Hieronymi, M., Müller, D., & Doerffer, R. (2017). The OLCI Neural Network Swarm (ONNS): a bio-geo-optical algorithm for open ocean and coastal waters. *Frontiers in Marine Science, 4*, 140

Hinton, G.E. (1990). Connectionist learning procedures. *Machine learning* (pp. 555-610): Elsevier

Hyndman, R.J., & Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting, 22*, 679-688

Ilori, C.O., Pahlevan, N., & Knudby, A. (2019). Analyzing Performances of Different Atmospheric Correction Techniques for Landsat 8: Application for Coastal Remote Sensing. *Remote Sensing, 11*, 469

Ioannou, I., Gilerson, A., Gross, B., Moshary, F., & Ahmed, S. (2011). Neural network approach to retrieve the inherent optical properties of the ocean from observations of MODIS. *Applied Optics, 50*, 3168-3186

IOCCG (2000). Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters. S. Sathyendranath Reports of the International Ocean-Colour Coordinating Group.

Irons, J.R., Dwyer, J.L., & Barsi, J.A. (2012). The next Landsat satellite: The Landsat data continuity mission. *Remote Sensing of Environment, 122*, 11-21

Jamet, C., Loisel, H., & Dessailly, D. (2012). Retrieval of the spectral diffuse attenuation coefficient Kd (λ) in open and coastal ocean waters using a neural network inversion. *Journal of Geophysical Research: Oceans, 117*

Kajiyama, T., D'Alimonte, D., & Zibordi, G. (2018). Algorithms Merging for the Determination of Chlorophyll-${a} $ Concentration in the Black Sea. *IEEE Geoscience and Remote Sensing Letters, 16*, 677-681

Kay, S., Hedley, J.D., Lavender, S. (2009). Sun Glint Correction of High and Low Spatial Resolution Images of Aquatic Scenes: a Review of Methods for Visible and Near Infrared Wavelengths. *Remote Sensing, 1*, 33

Khorram, S., Catts, G.P., Cloern, J.E., & Knight, A.W. (1987). Modeling of Estuarne Chlorophyll a from an Airborne Scanner. *Ieee Transactions on Geoscience and Remote Sensing*, 662-669

Kingma, D.P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Kuhn, C., de Matos Valerio, A., Ward, N., Loken, L., Sawakuchi, H.O., Kampel, M., Richey, J., Stadler, P., Crawford, J., & Striegl, R. (2019). Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. *Remote Sensing of Environment, 224*, 104-118

Kwiatkowska, E.J., & Fargion, G.S. (2003). Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data. *Ieee Transactions on Geoscience and Remote Sensing, 41*, 2844-2860

Le, C., Hu, C., English, D., Cannizzaro, J., Chen, Z., Feng, L., Boler, R., & Kovach, C. (2013). Towards a long-term chlorophyll-a data record in a turbid estuary using MODIS observations. *Progress in Oceanography, 109*, 90-103

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International journal of forecasting, 9*, 527-529

Manuel, A., Blanco, A., Tamondong, A., Jalbuena, R., Cabrera, O., & Gege, P. (2020). Optmization of Bio-Optical Model Parameters for Turbid Lake Water Quality Estimation Using Landsat 8 and WASI-2D. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, 67-72

Markham, B., Barsi, J., Kvaran, G., Ong, L., Kaita, E., Biggar, S., Czapla-Myers, J., Mishra, N., & Helder, D. (2014). Landsat-8 operational land imager radiometric calibration and stability. *Remote Sensing, 6*, 12275-12308

Markham, B.L., Barsi, J.A., Morfitt, R., Choate, M., Montanaro, M., Arvidson, T., & Irons, J.R. (2015). Landsat 8: status and on-orbit performance. In, *SPIE Remote Sensing* (pp. 963908-963908-963908): International Society for Optics and Photonics

Mittenzwey, K.H., Ullrich, S., Gitelson, A., & Kondratiev, K. (1992). Determination of chlorophyll a of inland waters on the basis of spectral reflectance. *Limnology and Oceanography, 37*, 147-149

Mobley, C.D. (1994). *Light and Water: Radiative transfer in natural waters*. Academic Press, Inc.

Mobley, C.D. (1999). Estimation of the remote-sensing reflectance from above-surface measurements. *Appl Opt, 38*, 7442-7455

Mobley, C.D., Werdell, J., Franz, B., Ahmad, Z., & Bailey, S. (2016). Atmospheric correction for satellite ocean color radiometry. NASA/TM-2016-217551, GSFC-E-DAA-TN35509

Morley, S.K., Brito, T.V., & Welling, D.T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather, 16*, 69-88

Munday, J., & Zubkoff, P.L. (1981). Remote sensing of dinoflagellate blooms in a turbid estuary. *Photogrammetric Engineering and Remote Sensing, 47*, 523-531

Neil, C., Spyrakos, E., Hunter, P.D., & Tyler, A.N. (2019). A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. *Remote Sensing of Environment, 229*, 159-178

O'Reilly, J.E., Maritorena, S., Mitchell, B.G., Siegel, D.A., Carder, K.L., Garver, S.A., Kahru, M., & McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans, 103*, 24937-24953

Pahlevan, N., Lee, Z., Wei, J., Schaff, C., Schott, J., & Berk, A. (2014). On-orbit radiometric characterization of OLI (Landsat-8) for applications in aquatic remote sensing. *Remote Sensing of Environment, 154*, 272–284

Pahlevan, N., Roger, J.-C., & Ahmad, Z. (2017a). Revisiting short-wave-infrared (SWIR) bands for atmospheric correction in coastal waters. *Optics express, 25*, 6015-6035

Pahlevan, N., Schott, J.R., Franz, B.A., Zibordi, G., Markham, B., Bailey, S., Schaaf, C.B., Ondrusek, M., Greb, S., & Strait, C.M. (2017b). Landsat 8 remote sensing reflectance (R rs) products: Evaluations, intercomparisons, and enhancements. *Remote Sensing of Environment, 190*, 289-301

Pahlevan, N., Smith, B., Binding, C., Gurlin, D., Li, L., Bresciani, M., & Giardino, C. (2020a). Hyperspectral retrievals of phytoplankton absorption and chlorophyll-a in inland and nearshore coastal waters. *Remote Sensing of Environment*, 112200

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., Matsushita, B., Moses, W., Greb, S., Lehmann, M.K., Ondrusek, M., Oppelt, N., & Stumpf, R. (2020b). Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment, 240*, 111604

Rundquist, D.C., Han, L., Schalles, J.F., & Peake, J.S. (1996). Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm. *Photogrammetric Engineering and Remote Sensing, 62*, 195-200

Ryu, J.-H., Han, H.-J., Cho, S., Park, Y.-J., & Ahn, Y.-H. (2012). Overview of geostationary ocean color imager (GOCI) and GOCI data processing system (GDPS). *Ocean Science Journal, 47*, 223-233

Schiller, H., & Doerffer, R. (1999). Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data. *International Journal of Remote Sensing, 20*, 1735-1746

Seegers, B.N., Stumpf, R.P., Schaeffer, B.A., Loftin, K.A., & Werdell, P.J. (2018). Performance metrics for the assessment of satellite data products: an ocean color case study. *Optics express, 26*, 7404-7422

Smith, R., & Baker, K. (1982). Oceanic chlorophyll concentrations as determined by satellite (Nimbus-7 Coastal Zone Color Scanner). *Marine biology, 66*, 269-279

Snyder, J., Boss, E., Weatherbee, R., Thomas, A.C., Brady, D., & Newell, C. (2017). Oyster aquaculture site selection using Landsat 8-derived sea surface temperature, turbidity, and chlorophyll a. *Frontiers in Marine Science, 4*, 190

Spyrakos, E., O'Donnell, R., Hunter, P.D., Miller, C., Scott, M., Simis, S.G., Neil, C., Barbosa, C.C., Binding, C.E., & Bradt, S. (2018). Optical types of inland and coastal waters. *Limnology and Oceanography, 63*, 846-870

Steinmetz, F., Deschamps, P.-Y., & Ramon, D. (2011). Atmospheric correction in presence of sun glint: application to MERIS. *Optics express, 19*, 9783-9800

Sydor, M., Gould, R.W., Arnone, R.A., Haltrin, V.I., & Goode, W. (2004). Uniqueness in remote sensing of the inherent optical properties of ocean water. *Applied Optics, 43*, 2156-2162

Thoma, M. (2017). Analysis and optimization of convolutional neural network architectures. *arXiv preprint arXiv:1707.09725*

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society, 66*, 1352-1362

Trinh, R.C., Fichot, C.G., Gierach, M.M., Holt, B., Malakar, N.K., Hulley, G., & Smith, J. (2017). Application of Landsat 8 for monitoring impacts of wastewater discharge on coastal water quality. *Frontiers in Marine Science, 4*, 329

Vanhellemont, Q., & Ruddick, K. (2018). Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. *Remote Sensing of Environment, 216*, 586-597

Vilas, L.G., Spyrakos, E., & Palenzuela, J.M.T. (2011). Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment, 115*, 524-535

Vos, W., Donze, M., & Buiteveld, H. (1986). *On the Reflectance Spectrum of Algae in Water: The Nature of the Peak at 700nm and Its Shift with Varying Algal Concentration*. Delft University of Technology, Faculty of Civil Engineering

Wang, D., Ma, R., Xue, K., & Loiselle, S.A. (2019). The assessment of Landsat-8 OLI atmospheric correction algorithms for inland waters. *Remote Sensing, 11*, 169

Wang, M., & Bailey, S.W. (2001). Correction of Sun glint Contamination on the SeaWiFS Ocean and Atmosphere Products. *Applied Optics, 40*, 4790-4798

Wang, M., Liu, X., Jiang, L., Son, S., Sun, J., Shi, W., Tan, L., Naik, P., Mikelsons, K., & Wang, X. (2014). Evaluation of VIIRS ocean color products. In, *Ocean Remote Sensing and Monitoring from Space* (p. 92610E): International Society for Optics and Photonics

Warren, M.A., Simis, S.G., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C., & Ansper, A. (2019). Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters. *Remote Sensing of Environment, 225*, 267-289

Watanabe, F.S.Y., Alcântara, E., Rodrigues, T.W.P., Imai, N.N., Barbosa, C.C.F., & Rotta, L.H.d.S. (2015). Estimation of chlorophyll-a concentration and the trophic state of the Barra Bonita hydroelectric reservoir using OLI/Landsat-8 images. *International journal of environmental research and public health, 12*, 10391-10417

Werdell, P.J., Bailey, S.W., Franz, B.A., Harding Jr, L.W., Feldman, G.C., & McClain, C.R. (2009). Regional and seasonal variability of chlorophyll-a in Chesapeake Bay as observed by SeaWiFS and MODIS-Aqua. *Remote Sensing of Environment, 113*, 1319-1330

Wezernak, C., Tanis, F., & Bajza, C. (1976). Trophic state analysis of inland lakes. *Remote Sensing of Environment, 5*, 147-164

Wynne, T., Stumpf, R., & Briggs, T. (2013). Comparing MODIS and MERIS spectral shapes for cyanobacterial bloom detection. *International Journal of Remote Sensing, 34*, 6668-6678

Zhan, H., Shi, P., & Chen, C. (2003). Retrieval of oceanic chlorophyll concentration using support vector machines. *Ieee Transactions on Geoscience and Remote Sensing, 41*, 2947-2951

Zibordi, G., Berthon, J.-F., Mélin, F., D'Alimonte, D., & Kaitala, S. (2009). Validation of satellite ocean color primary products at optically complex coastal sites: Northern Adriatic Sea, Northern Baltic Proper and Gulf of Finland. *Remote Sensing of Environment, 113*, 2574-2591