Student thesis series INES nr 576

Estimation of dissolved organic carbon from inland waters using remote sensing data and machine learning

Lasse Harkort

2022
Department of
Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden

Lasse Harkort (2022).

Estimation of dissolved organic carbon from inland waters using remote sensing data and machine learning

Master degree thesis, 30 credits in *Geomatics*

Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January* 2022 until *June* 2022

Estimation of dissolved organic carbon from inland waters using remote sensing data and machine learning

---

Lasse Harkort

Master thesis, 30 credits, in *Geomatics*

Supervisor:

Zheng Duan

Dep. of Physical Geography and Ecosystem Science, Lund University

Exam committee:

Per-Ola Olsson

Dep. of Physical Geography and Ecosystem Science, Lund University

Acknowledgements

I want to thank Zheng Duan for the great supervision, offering me useful scientific advice and guidance and organising a powerful computer without which the thesis would not have been possible to do in the way it is now. I also want to thank Renkui Gou and Salvador Hernández Malavé for the weekly meetings where we could share our worries and support each other with feedback.

I want to thank all authors that contributed to creating the openly available AquaSat dataset. Without your contribution, this thesis would not have been possible. Likewise, I want to thank all contributors of the Python scikit-learn package as well as various other contributors to open-source packages in Python and R which allowed me to write this thesis without the need for commercial licences or funding.

Abstract

This thesis presents the first attempt to estimate Dissolved Organic Carbon (DOC) in inland waters over a large-scale area using satellite data and machine learning (ML) methods. Four ML approaches, namely Random Forest Regression (RFR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), and a Multilayer Backpropagation Neural Network (MBPNN) were tested to retrieve DOC using a filtered version of the recently published open source AquaSat dataset with more than 16 thousand samples across the continental US matched with satellite data from Landsat 5, 7 and 8 missions. In this work, the AquaSat dataset was extended with environmental data from the ERA5-Land product.

Including environmental data considerably improved the prediction of DOC for all algorithms, with GPR showing the best and most robust performance results with moderate estimation errors (RMSE: 4.08 mg/L). Permutation feature importance analysis showed that from the Landsat bands, the wavelength in the visible green and for the ERA5-Land product, the monthly average air temperature were the most important variables for the machine learning approaches. The results demonstrate the predictive strength of advanced ML approaches faced with a complex learning task, such as GPR and MBPNN, and highlight the important role of considering environmental processes to explain DOC variations over large scales.

While performance evaluation showed that DOC concentrations can be retrieved with adequate accuracy, algorithm development was challenged by the heterogenous nature of large-scale open source in situ data, issues related to atmospheric correction, and the low spatial and temporal resolution of the environmental predictors. Although locally tuned models are likely to outperform the developed model in terms of accuracy, the model can address key issues of inland water remote sensing as a promising approach to overcome the lack of in-situ measurements and to map large scale trends of inland DOC dynamically over long time periods and seasons.

This research demonstrates how open source, large scale datasets like AquaSat in combination with ML and remote sensing can make research toward large scale estimations of inland water DOC more realistic while highlighting its remaining limitations and challenges.

Keywords: *dissolved organic carbon, machine learning, remote sensing, inland waters, water quality, open source data*

Table of Contents

Abbreviations

ANN	Artificial Neural Network
CDOM	Colored Dissolved Organic Matter
DOC	Dissolved Organic Carbon
GPR	Gaussian Process Regression
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
LaSRC	Landsat 8 Surface Reflectance Code
LEDAPS	Landsat Ecosystem Disturbance Adaptive Processing System
MAE	Mean Absolute Error
MBPNN	Multilayer Backpropagation Neural Network
ML	Machine Learning
MSE	Mean Square Error
R^2	Coefficient of determination
RFR	Random Forest Regression
RMSE	Root Mean Square Error
SVR	Support Vector Regression
USGS	United States Geological Survey

1 Introduction

Inland waters are aquatic ecosystems that are usually confined within land boundaries. They represent ecosystem types such as lakes, reservoirs, rivers, ponds, swamps, wetlands, and coastal areas (Ogashawara et al., 2017). Inland surface waters contain the world's main freshwater resources. They provide a broad range of ecosystem services to humankind (Millennium Ecosystem Assessment, 2005). Inland waters only occupy a small fraction of the surface of the Earth (3%), but recently it has been recognized that they play an important role in the global carbon cycle (Lauerwald et al., 2015; Raymond et al., 2013; Regnier et al., 2013).

Carbon enters a water body as carbon dioxide from the air and is fixed by photosynthesizing organisms (autochthonous carbon), from the degradation of dead terrestrial organisms (allochthonous carbon), or through ground and surface water from the catchment area (bicarbonate). This carbon is bound to living or dead organisms or is dissolved in water (dissolved organic matter, DOC) (Brönmark & Hansson, 2017). Part of the carbon that enters inland waters is emitted into the atmosphere, and another part sinks into the water column and becomes buried in the sediment. DOC has been a largely unaccounted form of carbon (Kutser et al., 2017) although its amount is typically very large in inland waters (Wetzel, 2001). To predict and better understand the dramatically changing climate on Earth in which the global carbon cycle is a major component, it is therefore highly desirable to estimate and map DOC content in inland waters. The estimation of DOC in inland waters is important not only from the global carbon cycle perspective, but also has effects on public health. DOC promotes the fouling of water and hampers the disinfection of water by chlorination. Water rich in DOC in connection with chlorination has been shown to have an increased risk of bladder and rectal cancer (Koivusalo et al., 1997) and higher frequencies of birth defects (Magnus et al., 1999). In many boreal regions (including Sweden), increased levels of DOC have led to browning of lakes that are often important sources of drinking water (Williamson et al., 2015). Therefore, routine monitoring of DOC is needed to ensure safe provide drinking water and to be able to adjust treatment processes when needed. However, in-situ measuring, and sampling of DOC concentrations is time-consuming and labor-intense.

DOC contains an optically active part, known as colored dissolved organic matter (CDOM), which can be detected by remote sensing instruments. In recent years, satellite-based techniques have evolved as an important supplement or alternative to in-situ measurements of CDOM, as they are capable of capturing the spatio-temporal dynamics of CDOM from large-

scale and historical perspectives (Brezonik et al., 2015; Zhang et al., 2021). Many studies have shown that CDOM and DOC concentrations correlate significantly, indicating that CDOM can be a reliable regional proxy for DOC (e.g. Chen et al., 2020; Erlandsson et al., 2012; Kutser et al., 2016). However, the CDOM/DOC relationship can be weak in some waterbodies or underly seasonal variations (Brezonik et al., 2015; Zhang et al., 2021). The prediction of DOC from CDOM levels is thus associated with considerable uncertainty (Brezonik et al., 2015) and thus, modelling the complex relationship of the received remote sensing signal and DOC concentration remains a big challenge.

Before reaching the sensor, the sun's light passes twice through the Earth's atmosphere, from the sun to the surface of the Earth and from the surface to the sensor. As a result, the absorption and scattering by gas molecules and particles in the environment always affects the light received at the sensor (Moses et al., 2017). Therefore, the development and research of atmospheric correction algorithms in remote sensing is a research field on its own. For open ocean water, robust atmospheric correction techniques have been developed, but those algorithms cannot simply be applied to inland waters as they violate some of the basic assumptions made for open oceans (e.g., adjacency effects from neighbouring land pixels, proximity to terrestrial sources of atmospheric pollution) (Moses et al., 2017). In the history of water quality remote sensing, inland water remote sensing has made slow progress compared to ocean remote sensing, which benefits from robust, open, and large data sets aimed at combining both in situ and radiometric observations with satellite data. Therefore, there is still a lack of universal algorithms and unified approaches to quantifying water quality parameters in inland waters (Palmer et al., 2015).

Hence, it is not surprising that machine learning (ML) has gained considerable attention in the inland water remote sensing community. ML, together with remote sensing, big data technologies, and high-performance computing, has emerged as a powerful and promising combination to quantify water quality parameters from remote sensing data (Hassan & Woo, 2021; Wagle et al., 2020). These methods have the potential to overcome the nonlinear, multicollinear, and heteroscedastic relationship of the received remote sensing signal and water quality parameters in inland water environments (Giri, 2021). Some findings even indicated that ML can achieve better predictions without atmospheric correction, as common correction algorithms are not developed for inland waters and can introduce errors (Medina-Lopez, 2020; Toming, Kutser, Laas, et al., 2016).

Given the uncertainties associated with quantifying DOC directly from remote sensing signals, DOC as a nonoptical parameter has received much less attention than other water quality parameters (e.g., chlorophyll-a or CDOM) in the inland aquatic remote sensing community. For an ML approach to successfully learn the hidden and complex patterns between DOC concentrations and the remote sensing signal a large number of DOC in-situ samples are required, which are often not available or very costly to get. Many studies only have access to in situ data from a small number of inland waters. Thus, validation studies are often biased toward certain optical water types (Palmer et al., 2015). Therefore, to date, very little is known about the potential of ML algorithms to predict DOC from remote sensing data in larger scale areas that include a diverse collection of complex optical water types.

Therefore, the first objective of this thesis is the following.

(1) Evaluate the performance of several ML algorithms that predict the concentration of DOC in inland waters from satellite data using a large-scale data set of DOC samples covering different inland water bodies.

Studies have shown that environmental variables (such as air temperature, wind speed, and precipitation) contribute to explaining the variability of in situ DOC concentrations in inland waters (Baines & Pace, 1991; Correll et al., 2001; Zhou et al., 2016). Therefore, including environmental variables has the potential to considerably improve the DOC prediction models. For the retrieval of DOC concentrations of large areas, environmental variables can play a fundamental role (Toming et al., 2020). However, only few studies assimilate in situ data within ecological and hydrodynamic models, and only little is known about the combinatorial effects of earth observation data with existing monitoring frameworks (Palmer et al., 2015). Thus, the second objective of this thesis is the following:

(2) Evaluate the effect of using several environmental variables in addition to Landsat surface reflectance to predict DOC concentration for inland waters.

To better understand the characteristics, traceability, and dynamics of DOC, it is important to map the spatio-temporal variations of the DOC content in a water body. In the last part of this work, we want to demonstrate the best-performing model in a remote sensing application by:

(3) Upscale point-based measurements to an entire waterbody using the best-performing model (based on objective 1 and objective 2) and analyze the spatiotemporal characteristics

of DOC content by mapping the DOC concentrations of a waterbody for each month of a selected year.

2 Background

2.1 Inland water remote sensing

The Earth Resources Technology Satellite (ERTS-1, later renamed Landsat-1), launched by NASA in August 1972, was the first satellite to monitor the earth surface from orbit. Researchers of the ocean remote sensing community were soon aware that chlorophyll-a and temperature could be monitored remotely. They understood that bio-optical algorithms would be needed to derive chlorophyll-a concentrations from satellite data of the ocean (Bukata, 2013). Almost at the same time, remote sensing was recognized as a possible complement to traditional lake monitoring methods. However, their application to monitor inland waters has been far less successful compared to open oceans (Ogashawara et al., 2017; Palmer et al., 2015). This is due to numerous additional challenges that remote sensing of inland waters faced and still faces when retrieving water constituents from remote sensing data. These challenges include:

(1) Challenges Related to Atmospheric Correction

Unlike open oceans, inland waters are most often located in diverse and heterogeneous environments that affect the constituents of the atmosphere. For example, they are often close to human settlements, which usually affects the atmosphere through pollution. Thereby, the atmosphere becomes optically more heterogeneous and thus more difficult to model. Inland waters are also usually confined within land boundaries, and thus scattering in the atmosphere contaminates water pixels with radiation from the surrounding land. The atmospheric correction algorithms applied by the data providers are commonly developed for terrestrial surfaces and are not suitable for inland waters, which introduces additional uncertainties for modelling atmospheric effects over inland waters (Gege, 2017). Another effect mostly unique to inland waters is an often-increased reflectance of water in the near-infrared region due to high sediment concentrations. The high sediment concentrations occur, for example, by wind-driven sediment resuspension, surface and subsurface runoff, agricultural and industrial discharge from terrestrial sources, or sediment influx from landslides and coastline erosion (Moses et al., 2017). The resulting increased reflectance of water in the near-infrared region then makes it difficult to determine and eliminate the influence of atmospheric aerosol

contribution on the signal received by the satellite sensor. In addition, inland waters can occur at very different altitudes in relation to the mean sea level. This can result in uncertainties in the estimates of aerosol content in the atmospheric column above water (Moses et al., 2017).

(2) Optical Influence of Suspended Minerals and Dissolved Organic Matter

Unlike the open ocean, where optical properties are reasonably defined by the impacts of phytoplankton and its by-products, algorithms for inland waters must account for the added optical influence of inorganic and organic suspended sediment, algae, dissolved organic matter, and other constituents (IOCCG, 2018).

(3) Lack of suitable sensors.

Satellite remote sensing for inland waters has been hampered by a lack of suitable sensors for a long time. Ocean color sensors such as the Moderate Resolution Imaging Spectroradiometer (MODIS) or the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) have a coarse spatial resolution that is not suitable for most inland waters. The rather recent launches of Landsat 8 and Sentinel 2 offer new possibilities with higher temporal, spatial, and spectral resolutions, but are still designed for terrestrial remote sensing and have only been available since a few years, making historical analysis of inland water quality difficult.

(4) Lack of funding, infrastructure, and coordination of research efforts.

Apart from the challenges arising from the optically complex properties of inland waters, Palmer et al. (2015) point out that inland water remote sensing has suffered by a lack of funding, infrastructure, and the mechanisms needed to coordinate research efforts (Palmer et al., 2015). According to Bukata (2013), oceanographers and limnologists have done their research in historical dual solitude. Although now both oceanographers and limnologist share an interest in optical complex waters, this lack of communication may have led to the loss of valuable scientific history (Bukata, 2013).

(5) Lack of in situ measurements

The lack of funding, infrastructure, and coordination of research efforts has also made it more difficult to facilitate the shared use of in situ measurements for inland waters. Consequently, there are large in situ data gaps that prevent direct calibration and validation of products derived from sensors which lead to validation studies being biased toward certain optical water types or lack standardized measurement methods (Palmer et al., 2015; Warren et al., 2021).

2.2 Retrieval algorithms for inland water constituents from satellite data

Despite the numerous challenges, many researchers have devoted their efforts to develop algorithms to retrieve water constituents in inland waters. These algorithms have been developed to translate the optical signal recorded by the remote sensor to the bio-optical and biogeochemical variables of interest and are commonly known as bio-optical algorithms or bio-optical models (IOCCG, 2018). Although there is a lack of consistency in terminology and classification of types of bio-optical algorithms, they can broadly be classified into five categories: empirical, semi-empirical, semi-analytical, quasi-analytical, and analytical (Ogashawara, 2015).

In empirical and semi-empirical algorithms, the statistical relationship between in situ measurements of water constituents and the optical signal recorded by the remote sensor is used. While semi-empirical algorithms partly rely on assumptions (e.g., reflectance signal peak near 700 nm to estimate chlorophyll-a concentration in inland and coastal waters), empirical algorithms focus only on statistical estimators (Ogashawara et al., 2017). Statistical techniques such as stepwise regression or least squares are used to find the optimal relationship between the water constituents and the radiometric data. Empirical algorithms can be implemented without the need for a priori assumptions. In contrast to this, analytical, semi-analytical, and quasi-analytical algorithms use knowledge of the underlying physics of light transfer in waters (radiative transfer theory) and analytical inversions to simultaneously estimate the variables of interest. Analytical algorithms are purely physics-based, while semi-analytical and quasi-analytical algorithms also include empirical steps.

Since empirical algorithms do not require prior knowledge of the underlying physics of light transfer in waters, they have been a popular choice for inland water constituent retrieval. However, the presence of non-linear relationships among the inland water constituents makes the application of linear regression less reliable to identify the precise relationships between the remote sensing signal and the water constituents. Another problem is that empirical relationships that are often optimized for a specific water body may not be fully transferable to other waterbodies, as they are based on the unique characteristics of the original target area (Sagan et al., 2020). Here, algorithms based on radiative transfer theory in waters may be more suitable because they are better at generalizing the relationship characteristics over different regions. However, detailed spectral information about the optically active water components in the research area is required and often cannot be obtained. Another challenge of the physics-

based approach is to select and specify the necessary parameters for modelling and solving radiation transmission equations (Sagan et al., 2020).

2.3 Machine Learning for retrieving water constituents in inland waters from satellite data

As in many other fields of research, the recent advances of ML together with big data technologies and high-performance computing have had a huge impact in the aquatic remote sensing community. As a more complex subset of empirical approaches, the application of ML to retrieve water constituents does not require prior knowledge. At the same time, ML approaches have been shown to be capable of capturing the nonuniform and complex relationships of the water constituents and the remote sensing signal (Hassan & Woo, 2021; Wagle et al., 2020) and outperform analytical approaches (Sagan et al., 2020). Recently, ML has also been shown to be able to quantify water quality parameters over large regions (Pahlevan et al., 2022; Smith et al., 2021).

Machine learning is a set of statistical algorithms that, without being explicitly programmed, can learn from data and construct a detection, estimation, or classification model that can make a prediction from data that has been previously unknown to the trained algorithm (Murphy, 2012). In the aquatic remote sensing research field this usually means that the computer learns the relationship between in-situ measurements of water constituents and spatiotemporally matched satellite data, although ML algorithms can also be used to optimize radiative transfer equations to estimate water constituents (e.g., Doerffer & Schiller, 2007). The learned relationship is then applied to previously unknown data and various evaluation metrics are used to estimate the model prediction error.

Today, there are many ML methods available and several review papers on the application of ML to retrieve inland water constituents have been published (Hassan & Woo, 2021; Sagan et al., 2020; Wagle et al., 2020). The ML algorithms used to retrieve the constituents of inland water can be categorized into several families. Due to resource and time limitations, not all ML families and algorithms can be covered in this thesis. However, from the review papers published, three common ML families can be identified: tree-based algorithms, neural networks, and kernel methods.

2.3.1 Tree-based algorithms

Tree-based machine learning algorithms use decision trees to learn from observations of a feature to predict the features target. A decision tree is a tree-like model of decisions where every question asked in the decision process is a test on one feature. Each test then results in completion or an additional test based on the current answer. The objective is to produce a tree that can generalize to predict unknown samples (Zhou, 2021a). Decision trees in which the target variable can take continuous numbers are called regression trees. Instead of the mode, the feature testing is based on the mean response of the observations falling into a certain category. A key issue of decision trees is that they are very sensitive to small variations in the training data and often do not generalize well on new data. To overcome these problems, ensemble methods are used that train and combine multiple learners to achieve more accurate and balanced models (e.g., Bagging, Boosting). The most prominent ensemble method that is based on decisions trees is called Random Forest (Breiman, 2001) (more details in Section 4.2.1).

2.3.2 Artificial neural networks

Research on artificial neural networks (ANN) already started decades ago and has become a broad and interdisciplinary research field today. ANN is a subset of ML that includes traditional and deep neural networks. Inspired by the biological nervous systems of humans and animals, ANNs consists of large interconnected hierarchical networks of simple elements (also called neurons) that combined together can solve complex learning tasks (Zhou, 2021b). In its simplest form, originally proposed by McCulloh and Pitts (1943), each neuron receives an input signal from *n* other neurons via weighted connections. The weighted sum of the received signals is compared to a threshold, and an output signal is generated by an activation function (Zhou, 2021b). The activation function defines the output of a neuron that has received an input or a set of inputs. If the output exceeds a threshold, the neuron is activated. To transform a large interval of possible values into an open unit interval between 0 and 1, a sigmoid function has been commonly used as the activation function. Today, many more types of activation functions exist, with the Rectified Linear Unit function (ReLU) being a common choice (Zhou, 2021b).

All parts of a neural network combined form a complex mathematical model with a large number of parameters that can be learned by feeding it with training and test data. At the end of the training process, the ‘knowledge’ of the neural network lies in the connection weights

and thresholds that can be used to generalize on new data (Zhou, 2021b). To learn these parameters efficiently, most neural networks use error backpropagation (Rumelhart et al., 1985). By two passes through the network (one forward, one backward), the backpropagation computes the gradients of the networks error regarding every model parameter, and thus finds out how the weighted connections and thresholds need to be tuned to reduce the error. Today, deep learning methods that contain many levels of processing layers are one of the fastest growing trends in research fields related to remote sensing and water quality monitoring (Sagan et al., 2020). In this thesis, a Multilayer Backpropagation Neural Network (MBPNN) is used to predict DOC (more details in Section 4.2.4).

2.3.3 Kernel methods

In machine learning, kernels can take care of non-linearity transformation when the feature space is high dimensional. Kernel methods avoid mapping the actual coordinates of the data in that high-dimensional feature space by computing the inner products between the images (i.e., the set of all output values a function from set x to set y may produce) of all pairs of data in the feature space. This is known as the 'kernel trick'. For example, fitting input data (in this case the spectral data matched to the DOC samples) to the output variable (DOC), the predicted DOC value ($\hat{y}$) for new input vector (x_*) (the new spectral data) can be obtained by Equation 1:

$$\hat{y} = f(x) = \sum_{i=1}^{N} a_i K_\theta(x_i, x_*) + a_o, \quad (1)$$

In this equation, $\{x_i\}\frac{N}{i=1}$ are the spectral data used during training, a_i is the weight assigned to each one of them, a_o is the bias in the regression function, and K_θ is the kernel function that measures the similarity between the new input spectrum x_* and all training spectra n in the high dimensional space (Ruescas et al., 2018). The kernel function is parametrized by a set of hyper-parameters θ. Different kernel functions exist that result in different values to the weights a_i. The most common kernel functions are linear, polynomial and radial basis functions (Malo & Camps-Valls, 2011). The ML algorithms that are based on kernel methods and used in this thesis are Support Vector Regression (SVR) (more details in 4.2.2) and Gaussian Process Regression (GPR) (Section 4.2.3).

2.4 Retrieval of the water constituent dissolved organic carbon in inland waters with remote sensing data

The retrieval of DOC from remote sensing data is mostly done by using the optically active part of DOC, CDOM, as a regional proxy (Chen et al., 2020; Erlandsson et al., 2012; Jiang et al., 2012; Kutser et al., 2016; G. Liu et al., 2021). Although CDOM is more challenging to retrieve from satellite data than other optically active substances (e.g., chlorophyll a), researchers have been able to establish empirical, semi-empirical, and semi-analytical algorithms to estimate CDOM from remote sensing signals and improve the estimation accuracy for inland waters (Zhang et al., 2021). Important wavelength regions for CDOM are in the visible range (400-700 nm). There are no wavelength bands in the visible spectrum uniquely associated with CDOM, but wavelengths in the green and red bands are generally important for estimating CDOM (Brezonik et al., 2015; Zhang et al., 2021). However, depending on the levels of spectral interference by large quantities of particulate matter in the water column, the most effective wavelength region for determining CDOM may vary (Zhu et al., 2014). In addition, previous findings from CDOM estimation that use ML indicate that a larger number of wavelength bands can further improve CDOM estimations (Ruescas et al., 2018; Sun et al., 2021).

While CDOM often correlates well to DOC, there are several processes in inland waters that can change the relationship between the optical and chemical properties of DOC (e.g. dilution, photodegradation, and flocculation) (Chen et al., 2004; Loiselle et al., 2010). Subsequently, studies have shown that in some inland waterbodies, the CDOM/DOC correlations can be weak or underly seasonal variations (Hestir et al., 2015; Massicotte et al., 2017; Toming, Kutser, Tuvikene, et al., 2016). Hence, assumptions made for CDOM retrieval may not always apply for DOC and is associated with high uncertainty (Brezonik et al., 2015; Griffin et al., 2018).

The advancement of ML algorithms offers new possibilities as the algorithms are able to learn the hidden and complex patterns between DOC and remote sensing derived reflectance. Therefore, researchers have begun to use ML to directly retrieve DOC from remote sensing signals without relying on the relationship of CDOM/DOC. For instance, Liu et al. (2021) used a Multilayer Backpropagation Neural Network (MBPNN) model to estimate DOC concentrations in the eutrophic Lake Taihu, China. They matched MODIS/Aqua data, temperature and wind speed with in-situ measurements from the lake and yielded a mean estimation error of 15.14% for the test dataset (D. Liu et al., 2021). Codden et al. (2021) applied

ML to predict DOC concentration in a dynamic salt marsh creek. They used ML coupled with low- to zero-cost predictors (e.g. local rainfall, point in year) to surmount the problem of having to use linear modelling and optical data predictors collected from high-cost, high-maintenance in situ spectrophotometer, which is often used as an alternative to expensive discrete sample collections of DOC. The ML algorithms were able to modestly improve the accuracy of linear methods while at the same time reducing the instrumentation cost by approximately 90%. They concluded that although their models were developed for a single site only, ML together with low- to zero-cost predictors is a methodology that bears high potential for others trying to model DOC dynamics and other analytes in any complex aquatic system (Codden et al., 2021). Toming et al. (2020) predicted lake DOC concentration and total lake DOC at global scale using only environmental variables and a boosted regression tree algorithm. Their analysis of variable importance showed that catchment properties and meteorological and hydrological features explained most of the variability of the lake DOC concentration (Toming et al., 2020).

These studies highlight the potential of ML-based estimations of DOC using satellite data and environmental predictors. However, while Liu et al. (2021) and Codden et. al (2021) used ML with satellite data and environmental predictors for one specific waterbody only, the model developed by Toming et al. (2020) shows global DOC estimates but cannot map spatial variability and seasonal variations within lakes. Recently, efforts have been made to use ML to retrieve water quality parameters from satellite data on a global scale (Pahlevan et al., 2022; Smith et al., 2021). However, to the best of the authors' knowledge, there is no study that tested whether ML algorithms can use satellite and environmental data to predict DOC concentration directly and dynamically for larger scale areas that include different inland water bodies, which motivated us to address this research gap. For this purpose, a newly developed data set AquaSat is used which will be described in more detail in the following chapter.

3 Data

3.1 AquaSat

For in situ DOC measurements coupled with satellite data the dataset AquaSat (Ross et al., 2019) is taken. AquaSat is a combination of existing public datasets covering the continental United States, including the Water Quality Portal (WQP) (Read et al., 2017), LAGOS NE (Soranno et al., 2017) and the Landsat archive (Wulder et al., 2016).

The samples of the dataset are matched with satellite images from the three most recent Landsat missions: Landsat 5 (Thematic Mapper, 1984–2012), Landsat 7 (Enhanced Thematic Mapper+, 1999–present), and Landsat 8 (Operational Land Imager, 2013–present). The Landsat satellites have a 16-day orbit repetition period though overlapping images at high latitudes result in shorter revisit periods. Landsat 5 and 7 onboard sensors capture imagery in seven bands, three visible wavelengths (blue, green, and red) and four infrared wavelengths (near infrared (NIR), short-wave infrared 1 (SWIR1), shortwave infrared 2 (SWIR2), and thermal band). Landsat 8 includes bands in the same spectral regions as previous sensors with enhanced signal-to-noise ratios, and an added ultrablue band. All six bands used in this study (blue, green, red, NIR, SWIR1, SWIR2) have a spatial resolution of 30m.

AquaSat includes more than 600,000 Landsat match-ups of dissolved organic carbon, chlorophyll-a, ground-based total suspended sediment, and SDD Secchi disk depth measurements matched with atmospherically corrected surface reflectance from the Landsat missions within ±1 day of each other, covering the years 1984–2019. For atmospheric correction, the United States Geological Survey (USGS) developed a surface reflectance product based on the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) for Landsat 5 and 7 (Ju et al., 2012) and the Landsat 8 Surface Reflectance Code (LaSRC) for Landsat 8 (Vermote et al., 2016). Each water quality sample in the Aquasat dataset has matched reflectance values of Landsat 5,7 or 8 from the blue, green, red, NIR, SWIR1, and SWIR2 band. According to Ross et al. (2019), AquaSat is the largest set of matchup data ever assembled and was developed to specifically meet the lack of a public remote sensing data set paired with in situ measurements of water constituents. The data set has already been used in previous studies (e.g., Pahlevan et al., 2022; Topp et al., 2021). To the best of the authors' knowledge, it has not yet been used for DOC predictions with ML.

3.2 ERA5-Land

To combine satellite data with environmental predictors, data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) Land product was used (Muñoz-Sabater et al., 2021). ERA5 is the fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from January 1950 to present. ERA5 is produced by the Copernicus Climate Change Service (C3S) at the European Centre for Medium-Range Weather Forecasts (ECMWF). The ERA5-Land product is a higher resolution version (0.1° x 0.1°, native resolution about 9km) of the ECMWF ERA5 climate reanalysis dataset. It has been

available since mid-2019 and is therefore one of the most recent products available. The data set offers a wide variety of climatic variables. Reanalysis data combine observational inputs and models to produce consistent products based on physical principles. (Copernicus Climate Change Service, 2019). Starting with a weather prediction model, ERA5 employs the Integrated Forecast System cycle 41r2 (Cy41r2), which is coupled with a large number of observations (for example, temperature observations, aircraft measurements, and satellite data) using data assimilation techniques (Hersbach et al., 2020). To achieve the final analysis as close as possible to observations and forecasts, a cost function is minimized. ERA5 has been shown to outperform gauge or satellite-based products in terms of reliability and accuracy (Zandler et al., 2020). For this thesis, the ERA5-Land monthly averaged data product from 1981 to present is used (Muñoz-Sabater, 2019).

The environmental predictors tested were selected based on a literature review and include the average monthly temperature, wind speed, leaf area index (LAI) (high vegetation), average LAI (low vegetation), evaporation over inland waters, surface net solar radiation and the monthly total precipitation. The following section shortly describes the reasoning behind choosing these variables.

- Solar radiation - surface net solar radiation has been found an important variable for global DOC estimations (Toming et al., 2020)
- Evaporation over inland waters – Anderson & Stedmon (2007) found evaporative concentration to be the first-order control on DOC concentration in lakes tested in southwest Greenland (Anderson & Stedmon, 2007).
- LAI – low vegetation, high vegetation – Yang et al (2017) studied riverine systems in the US and found that forests and shrubland are positively correlated with DOC concentration (Yang et al., 2017) which can be represented by LAI.
- Total precipitation – although the subsequent effect of rain on DOC can vary considerably its relation to DOC variation have been shown in many studies (Jennings et al., 2012; Reche & Pace, 2002; Zhou et al., 2016).
- Wind speed – wind speed has been shown to promote sediment DOC release and DOC redistribution (Zhou et al., 2016) and found to contribute to explaining DOC variations in a ML model (D. Liu et al., 2021).
- Temperature – many studies have shown temperature to be one of the main drivers of DOC variations (Freeman et al., 2001; Kellerman et al., 2014; D. Liu et al., 2021; Strock et al., 2016).

4 Methodology

The methodology in this study can be divided into three major parts: data preprocessing, model development and model evaluation (Figure 1).

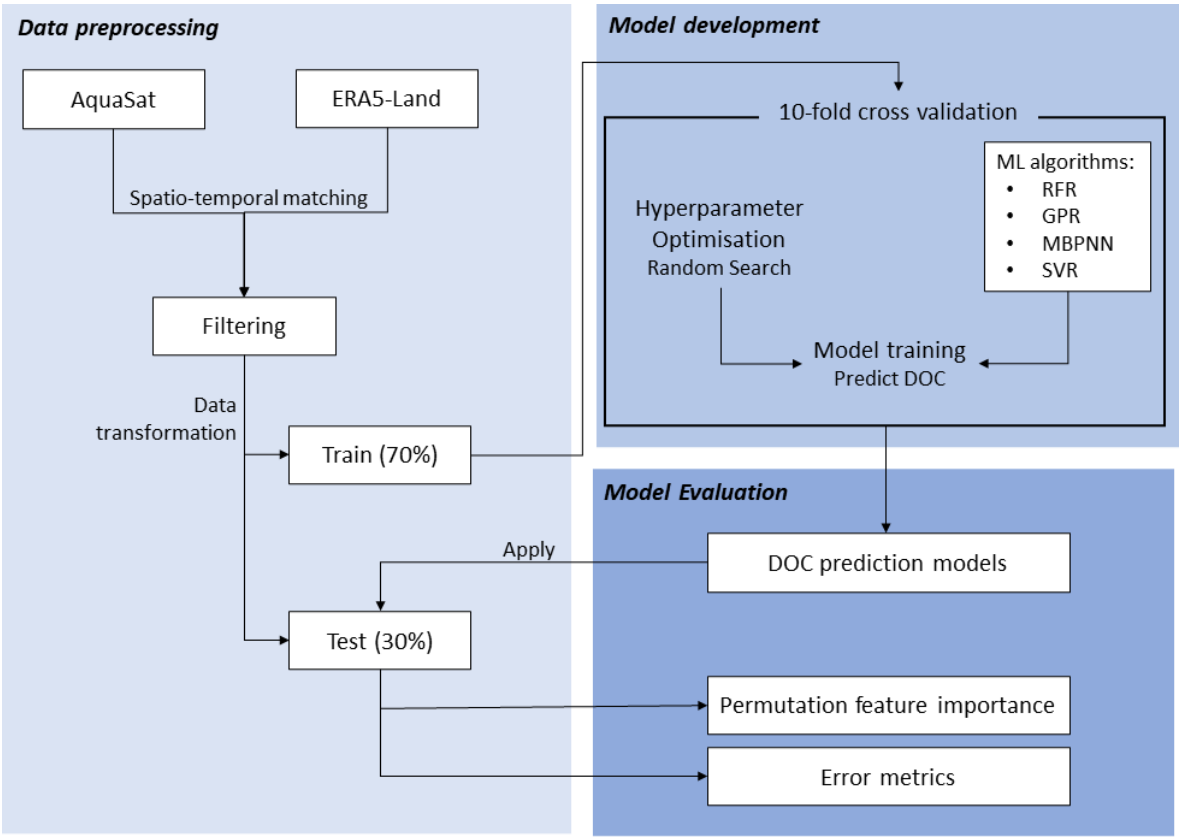

Fig. 1. Schematic workflow. After data preprocessing, the dataset is divided into training and test data sets. 10-fold cross-validation is used on the training data set for performance evaluation of the model parameters. The model development includes analysing the feature importance with Random Forest and finding the best performing ML algorithms with optimized hyperparameters. After development, each model predicts the concentration of DOC from the independent test data set and permutation feature importance and error metrics are computed for the final evaluation (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

4.1 Data Preprocessing

Although AquaSat contains an extensive number of samples, it is a dataset that is compiled from multiple sources. In such 'secondary use' data, many issues can arise through different methods used to report the same common metadata elements (Sprague et al., 2017).

Ross et al. (2019) have applied several data quality assurance procedures for data integration (Ross et al., 2019, Chapter 2.3.1) and joining Landsat data to in-situ match-ups (Ross et al.,

2019, Chapter 2.3.2). An overview of all quality assurance steps described in Ross et al. (2019) is given in the appendix (Table A2). Ross et al (2019) included as much data as possible to give future users the opportunity to set their own criteria, resulting in a dataset of purposefully lower quality (Ross et al., 2019). Although not all issues can be resolved, we intended to set a high-quality standard for the samples, and thus considerable reduction in the amount of in situ data was necessary. After extracting all DOC samples (N=42,316), they were matched with environmental predictors from the ERA5-Land monthly average data product. The year and month in which each sample was taken was matched with the nearest pixel value of the spatially and temporally corresponding year and the monthly average of the ERA 5 Land product. DOC concentrations lower than zero were removed. Samples with a water type classified as 'Facility' were removed from the data set as they indicate wastewater treatment facilities, which is not part of the purpose of the study. To ensure more robust band reflectance estimates, we only kept Landsat matchups, where at least nine water pixels were used to calculate a spatial median of reflectance in each band. Any outliers in the spectral response matched with the in situ location were investigated. We removed all samples that had saturated pixels from the bands (Morfitt et al., 2015). Such outliers in the remote sensing signal can occur e.g., by waves, sun glint, or clouds. We then converted the pixel values to surface reflectance using the Landsat-specific scale factor. The SWIR2 band was subtracted from the surface reflectance in the blue, green, red, NIR, and SWIR1 bands to remove the partial aerosol signal from the surface reflectance (Feng et al., 2018). We then joined the metadata provided to the dataset. We checked the metadata given for the analytical method and removed those where information about the analytical method used was not available or unreliable (e.g. "Historic procedure"). For DOC matchups, a ± 3 h time window for estuaries, a same-day overpass for streams, and due to the known stability of the optical active component of DOC in lakes, a ± 3 h time window for lakes was allowed (Brezonik et al., 2015; Pahlevan et al., 2022). After preprocessing, the dataset contains 16,029 samples across the continental United States (Figure 2). Previous findings from modelling CDOM with ML suggest that ML approaches can generally make use of information in a wide wavelength range (Olmanson et al., 2016; Ruescas et al., 2018). Hence, for model development, all six available Landsat bands from the AquaSat dataset were used together with the seven environmental predictors retrieved from ERA5 Land. The importance of each predictor is then assessed by permutation feature importance (Section 5.2).

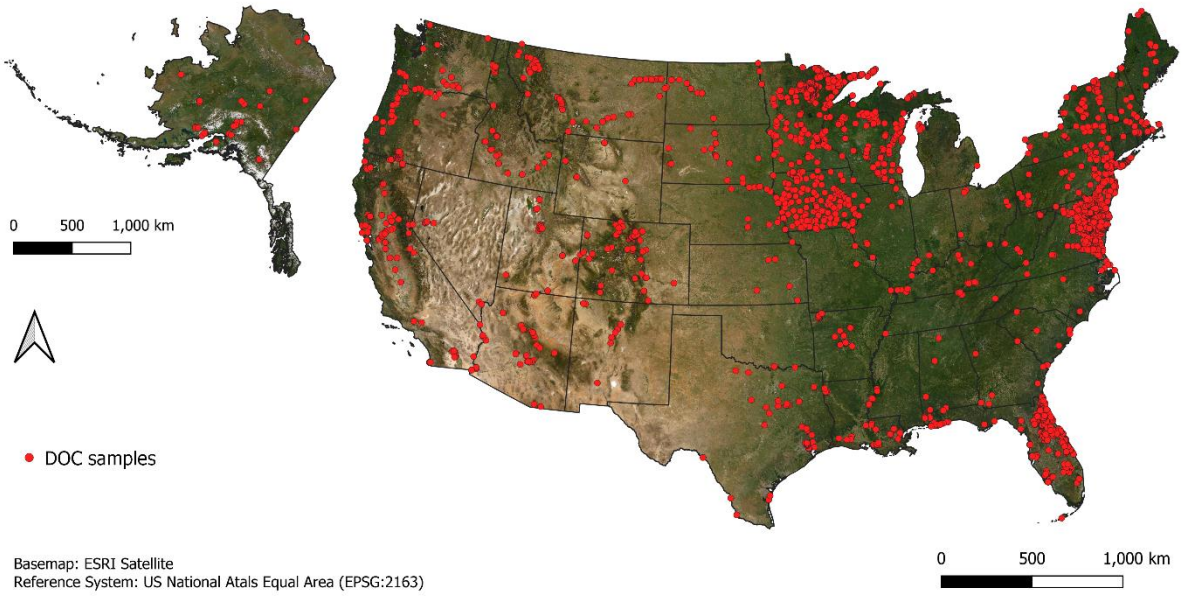

Fig. 2. Spatial distribution of DOC samples (N=16,029) in the continental United States after data filtering.

The samples were then randomly divided into training (70%) and test datasets (30%). All input features were standardized by removing the mean and scaling to unit variance (Codden et al., 2021). The output variable (i.e., DOC) was logarithmically transformed (base 10) to ensure homogeneity of variance (Figure 3) (Campbell, 1995). Transformation to logarithmic space is a common technique for data sets containing multiple inland water bodies and a wide range of water constituent's concentrations (Brezonik et al., 2015; Maier et al., 2021; Pahlevan et al., 2022). The test data was then scaled using the same function that was used to standardize the training data, resulting in input consistent with the trained algorithm's expectations.

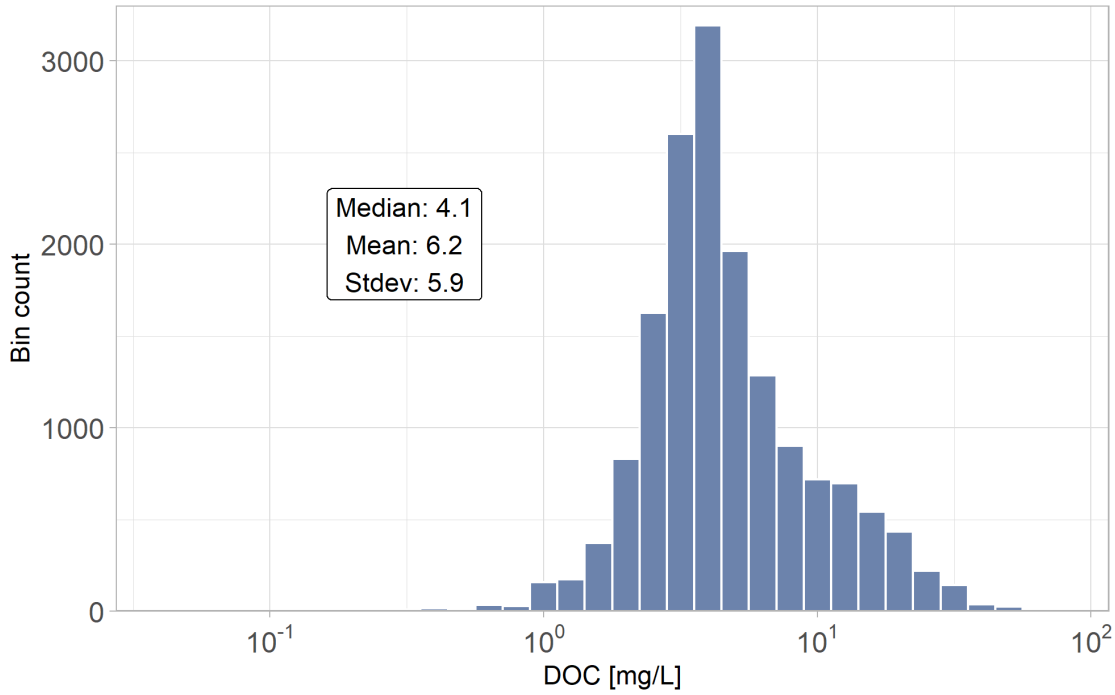

Fig. 3. Frequency distribution of DOC concentration (log10 transformed) of the samples (N=16,029).

4.2 Model development

Based on the literature review, four popular ML algorithms were selected to include at least one algorithm from each ML family described in Section 2.3. The machine learning models were implemented in the well-known Python package scikit-learn (Pedregosa et al., 2011). An exception is the ANN which core functionality was implemented in Tensorflow (Abadi et al., 2016). Data processing was done on a workstation with 64GB of RAM.

4.2.1 Cross-Validation

The samples are unevenly distributed in terms of time, location, and Landsat sensors (see Appendix, Figures A1, A2 and A3). To detect issues with overfitting, 10-fold cross-validation was implemented to test if differences in dataset partitioning influence the model performance. In 10-fold cross-validation, the training dataset, which has been extracted from the original dataset, is split into 10 groups. Each unique group is once used as a test data set, while the remaining groups are used for training the model. The trained model will then be evaluated against the test dataset and the evaluation scores are recorded. The model is then discarded, trained, and evaluated with the next train-test partition. For 10-fold cross-validation the process will therefore repeat ten times. In this thesis, the model performance is summarized by taking the average and standard deviation of the 10 evaluation score samples. A high standard

deviation indicates if the models performance depends on the training partitioning. As evaluation score, the commonly used root mean squared error (RMSE) is used (Section 4.3). The RMSE from cross-validation of the training set can then also be compared to the RMSE retrieved from the test dataset to test the robustness of the models.

4.2.2 Hyperparameter tuning

In machine learning models, two types of parameters exist: model parameters, which can be initialized and updated through the data learning process, and hyperparameters, which cannot be estimated directly from data learning and must be set before training an ML model as they define the model architecture. The process of exploring the range of possibilities and finding the optimal hyperparameter configuration is called hyperparameter tuning (Yang & Shami, 2020). A commonly used method is Grid Search, which searches the optimal configuration of hyperparameters in a fixed domain of hyperparameters (Bergstra et al., 2011). In this thesis, hyperparameter tuning was done via random search (Bergstra & Bengio, 2012). This method is similar to grid search, but instead of searching all possible parameters specified in the fixed domain, it randomly selects hyperparameter combinations and outputs the best configuration based on k-fold cross-validation. Compared to grid search, random search can find better models by effectively searching a larger, less promising configuration space (Bergstra & Bengio, 2012). This is especially useful in this thesis, as different hyperparameters may be important for the two dataset configurations (one without and one with environmental variables). In this thesis, 5-fold cross-validation was adopted with 60 trials. Around 60 trials have been shown to find better models than grid search in a larger less promising space (Bergstra & Bengio, 2012). 5-fold cross-validation was used as a compromise between robust estimates and processing time. The hyperparameter values found by random search for each of the ML algorithms are listed in the appendix (Table A1).

4.2.3 Random Forest Regression

Random Forest Regression (RFR) is an ensemble learning method that constructs many independent decision trees. Decision trees are independent because they work with random subsets of features. Many of these random trees together form a ‘Random Forest’. The average of all outputs of the independent decision trees is then taken as the model prediction result (Breiman, 2001). In this way, overfitting is avoided and the output is less sensitive to outliers and noise, which increases the generalization ability of Random Forest when applied to new data. Compared to other ML methods, RFR is easy to understand but still strong in its

performance and therefore commonly applied in water quality research (e.g., Hafeez et al., 2019; Ruescas et al., 2018; Sun et al., 2021). RFR can be tuned by many hyperparameters. Common parameters tuned are the number of decision trees in the 'forest', the maximum depth of the decision trees and the numbers of features to consider when looking for the best split. Note that when hyperparameters do not appear in Table A2 in the appendix, the default value of the Python package scikit-learn implementation is used.

4.2.4 Support Vector Regression

In traditional regression models, the loss is calculated from the difference between the model output $f_{(x)}$ and the ground-truth y. In Support Vector Regression (SVR), however, a loss only incurs if the difference between $f_{(x)}$ and y exceeds a set error margin. This error margin creates something similar to a buffer region that surrounds $f_{(x)}$. Training samples that may fall into this region are considered correctly predicted (Zhou, 2021c). In SVR, these minor inaccuracies are mostly accepted, as the main goal is to find a hyperplane in multidimensional feature space that separates the data with a maximum margin. The maximum-margin hyperplane may not always perform best on the training dataset but has the strongest generalization ability when applied to new data (Zhou, 2021c). To map samples from the original feature space to a higher dimensional space, SVR requires a kernel function that takes vectors as inputs in the original space and returns the dot product of the vectors in the feature space (2.3.3). Commonly used kernel methods include polynomial, linear, and radial basis functions. SVR has many hyperparameters to be tuned. Important hyperparameters include the C regularization parameter, which controls overfitting, the e-insensitive zone, which sets the error margin where no loss incurs, and the kernel, which defines the kernel type used in the algorithm. Compared to other ML algorithms, the performance of SVR is very dependent on the selection of these parameters (Mountrakis et al., 2011). Besides RFR, SVR can be considered a standard approach in ML algorithms and is often among the best performing algorithms in water quality applications (Guan et al., 2020; Kim et al., 2014; Tenjo et al., 2021).

4.2.5 Gaussian Process Regression

The basic idea in Gaussian Process Regression (GPR) is to give a prior probability to all functions that could explain the relationship between the input variables and the target variable, where higher probabilities are given to functions that are more likely to explain the relationship of the input variables and the target variable, for instance because they are smoother than other functions (Rasmussen & Williams, 2006). Because it is impossible to compute an infinite set

of possible functions for the relationship of x and y, a Gaussian process is used. A Gaussian process is a generalization of the Gaussian probability distribution, and instead of describing the probability distribution for random variables, it governs the properties of functions with a stochastic process. It can loosely be thought of as a very long vector where each entry in the vector specifies the function value $f_{(x)}$ at a particular input x (Rasmussen & Williams, 2006). GPR is an iterative process in which the probability of a function is estimated under the condition that a new point is observed. Therefore, the function defined by the Gaussian process is constantly adjusting as new observed points are added. Similarly to SVR, GPR uses the kernel trick to allow this process to work in a multidimensional space (2.3.3). As suggested by the findings of Zare Farjoudi & Alizadeh (2021), we used the rational quadratic kernel function for the GPR implementation (Zare Farjoudi & Alizadeh, 2021) including a white kernel for noise level estimation. Hyperparameter tuning through random search is not required for GPR as its hyperparameter are optimized by maximizing the log-marginal-likelihood in the training set (see Rasmussen & Williams, 2006). GPR has become a common ML algorithm used in water quality parameter predictions due to its excellent performance (Blix et al., 2018; Ruescas et al., 2018; Sun et al., 2021).

4.2.6 Multilayer Backpropagation Neural Network

The MBPNN used in this thesis comprises an input layer, multiple hidden layers, and an output layer. The input layer receives the external input, the hidden layers process the signals, and the output layer outputs the processed signals (Zhou, 2021b). The hidden layers consist of a large number of neurons which were optimized with the Adam optimizer (Kingma & Ba, 2014). Each neuron in each layer is connected to all neurons in the preceding and subsequent layers. Training a neural network requires selecting a structure (number of hidden layers and nodes per layer), properly initialization of the weights, shape of the nonlinearity, learning rate, and regularization parameters to prevent overfitting. To find these hyperparameters for the MBPNN, a wrapper is used to allow randomized search for hyperparameter tuning in the same way as it is done for the other ML algorithms. The implementation of the MBPNN in this thesis mainly follows the example by Géron (2019) in Chapter 10 (Géron, 2019). To prevent overfitting and reduce the amount of training time, early stopping is implemented, which interrupt training early when there is no more progress. Early stopping is defined by the patience argument and set to 10 (Géron, 2019). MPNN have shown promising results in water quality research (Chen & Hu, 2017; Sun et al., 2011) and have recently been applied for the retrieval of DOC (D. Liu et al., 2021).

4.3 Model evaluation

Guidance on commonly used error metrics (Seegers et al., 2018) and new error metrics have been proposed (Morley et al., 2018). In this thesis, a straightforward but robust combination of metrics has been chosen. The root mean squared error (RMSE, Eq. 2) is the square root of the average of the squared difference between the target value and the value predicted by the regression model. It can be thought of as the standard deviation of the residuals, which indicates how concentrated the data is around the line of best fit. The mean absolute error (MAE, Eq. 3) is simply the average difference between the measured and the estimated value. The system error (Bias, Eq. 4) is a description of the systematic direction of the error, as either overestimating or underestimating the prediction on average. The coefficient of determination (R^2, Eq. 5) indicates how much of the variation in the estimated values is explained by the measured values in the regression models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n} \tag{3}$$

$$Bias = \frac{\sum_{i=1}^{n}(x_i - y_i)}{n} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)}{\sum_{i=1}^{n}(x_i - \mu_i)} \tag{5}$$

In equations 2-5, *n* is the number of data pairs, the subscript *i* denotes individual data points, and *x* and *y* represent the measured and estimated values.

4.4 Permutation feature importance

For estimating the feature importance of the models, the inspection technique permutation feature importance was used. It is defined to be the decrease in a model score when a single feature value is randomly shuffled (Breiman, 2001). The drop in the model score then indicates how much the model depends on each of the features. Compared to Random Forest feature

importance, permutation feature importance can be used on any machine learning model. Another advantage is that it can be applied to the held-out test dataset and therefore highlight which features contribute the most to the generalization power of the model. It does, however, not reflect the intrinsic predictive value of a feature by itself, but how important this feature is for a particular model (Pedregosa et al., 2011). In this thesis, the feature importance score was based on the average increase in mean squared error (log10 transformed) when a single feature has been randomly shuffled 30 times. Multiple tests in our study and experience from other studies show that more than 30 realizations do not reduce much of the variance of the estimates (Ruescas et al., 2018).

5 Results

5.1 ML algorithm performance

Table 1 provides an overview of the metrics for all four ML methods trained by the six Landsat bands only and trained with the six Landsat bands together with the seven environmental variables. For better interpretability, predicted and measured values were transformed back from the logarithmic scale before calculating the evaluation metrics RMSE, MAE and bias. The unit for these metrics is therefore DOC [mg/L]. The cross-validation score (CV-score) is based on the training set and shows the log10 transformed RMSE with standard deviation after 10 folds (Table 2).

Overall, the ML algorithms predicted DOC with moderate to high uncertainties. Including the environmental predictors led to substantial performance improvements for all ML algorithms, with GPR and MBPNN showing the highest improvements and best overall scores. MBPNN has a slightly lower RMSE (4.02 mg/L) and Bias (-0.45) than GPR (RMSE 4.08 mg/L, Bias -0.69), but a higher MAE (MBPNN: 2.03 mg/L, GPR: 1.9 mg/L) and a lower R^2 score (MBPNN: 0.57, GPR: 0.61). Although the metric scores between both algorithms are without considerable differences, the CV-score of MBPNN (0.25 ± 0.008) is much higher than that of GPR (0.19 ± 0.008).

Table 1: Performance evaluation. For better interpretability, predicted and measured values were transformed back from the logarithmic scale before calculating RMSE, MAE and Bias. RMSE, MAE and Bias can be interpreted as DOC [mg/L]. (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

	Landsat Bands				**+ Environmental variables**			
	RMSE	**MAE**	**Bias**	**R^2**	**RMSE**	**MAE**	**Bias**	**R^2**
RFR	5.154	2.556	-1.1	0.366	4.271	1.984	-0.877	0.592
GPR	5.169	2.539	-1.092	0.364	4.078	1.902	-0.686	0.611
MPBNN	5.355	2.72	-1.508	0.273	4.048	2.029	-0.297	0.542
SVR	5.319	2.701	-1.177	0.258	4.708	2.280	-0.737	0.505

Table 2: RMSE of the test data set and 10-fold cross validation scores (average RMSE with standard deviation) of the training data set for each algorithm. Note that scores are based on logarithmic 10 transformed DOC values (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

	Test RMSE	**Train RMSE (CV-score)**
RFR	0.196	0.195 ± 0.01
GPR	0.191	0.192 ± 0.008
MBPNN	0.207	0.248 ± 0.01
SVR	0.215	0.215 ± 0.007

The performance of the ML algorithms is further illustrated in scatterplots plots (Figure 4). The reported slopes are generally closer to unity when environmental variables are included. Although a negative bias in all algorithms indicates an overall underestimation of DOC, the slopes compared to unity suggest a degraded performance of all ML algorithms at the two tails of data distribution, overestimating small DOC concentrations and underestimating larger DOC concentrations. Of all ML algorithms, the slopes of GPR and MBPNN with included environmental variables are closest to the unity line.

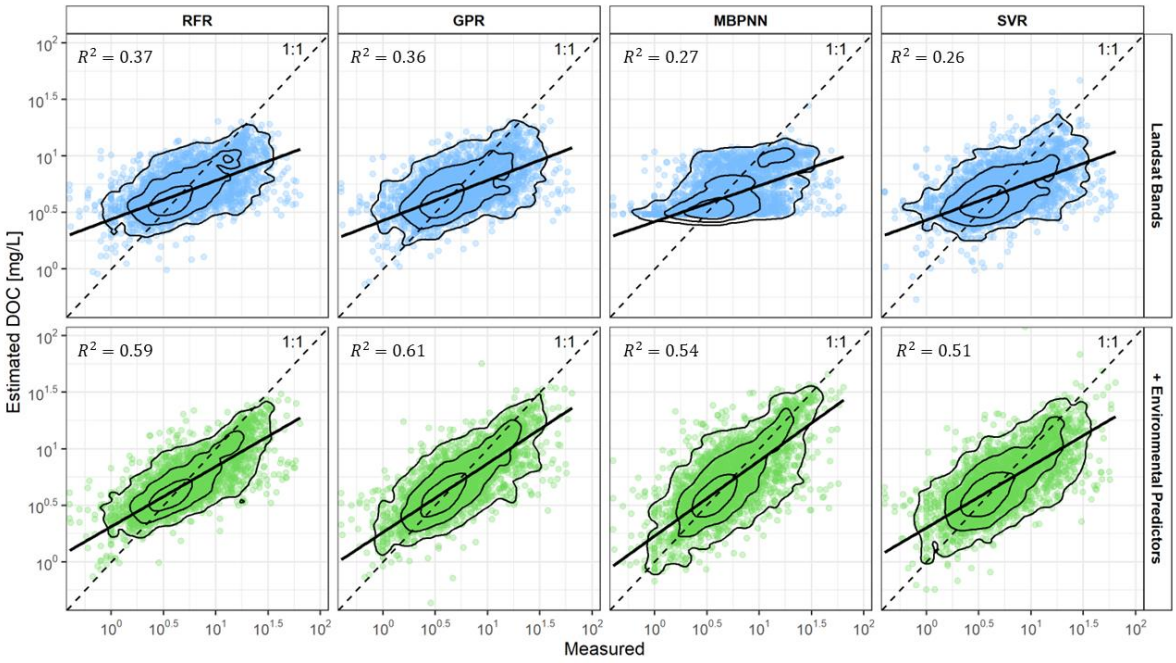

Figure 4: Scatter plots showing the performance of the four ML algorithms used in this study. The x axis shows the measured DOC from the independent test dataset (N=4809) and the y axis contains the predicted values for each of the measurements with the different ML models. Contour lines are included to better illustrate the data distributions (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MBPNN= Multilayer Backpropagation Neural Network).

The distribution of residuals further illustrates model improvements when using environmental variables (Figure 5). The interquartile range and whiskers are reduced considerably for all algorithms, most notably in MBPNN, RFR, and GPR. However, the range of outliers seems largely unchanged, indicating that including environmental predictors did not have a substantial impact when predicting extreme DOC values.

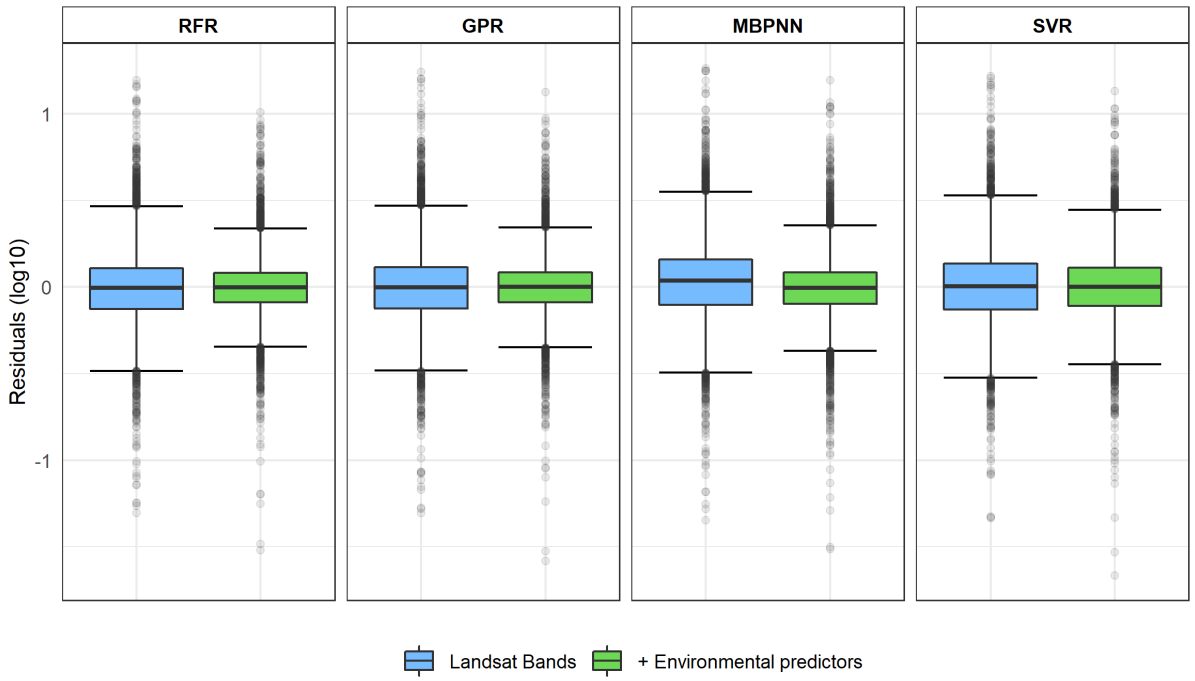

Figure 5: Log10 transformed residual distribution for each algorithm using only Landsat bands (blue) and using Landsat bands with environmental variables (green) as predictors (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

5.2 Variable importance

The analysis of feature importance is shown in Figure 6. Overall, removing the green band and air temperature leads to the highest MSE increase on average, suggesting that the two variables are most important for the models, followed by the red band and LAI - high vegetation. The general increase in MSE in RFR is low compared to SVR and MBPNN. RFR is the only ML approach where the increase in MSE for air temperature is higher than in the green band and the MSE increase for LAI - high vegetation - is higher than the red band. This indicates that the performance of RFR tends to rely more on environmental predictors compared to the other algorithms, although according to Table 1, it is not improving more than the other algorithms when environmental predictors are added. MBPNN and SVR strongly depend on the green band as large increases in MSE are observed when it is removed. Although GPR has a similar distribution of variable importance compared to SVR and MBPNN, the average MSE increase appears more balanced.

Figure 6: Permutation feature importance for each ML approach. Feature importance is quantified by the average increase in mean squared error (log10 transformed) when a single feature has been randomly shuffled 30 times. Note that the scores do not reflect the intrinsic predictive value of a feature by itself, but how important this feature is for a particular model (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer back-propagation neural network).

5.3 Visual assessments

Based on the model evaluation, the GPR approach with environmental predictors included was the best model for predicting DOC. For estimating DOC to an entire waterbody Lake Okeechobee in Florida (US) was chosen (Figure ?). Lake Okeechobee is the eights largest freshwater lake in the US and has been subject to many studies as it plays an important role for the everglades agricultural area and has experienced many cyanobacteria outbreaks in the past.

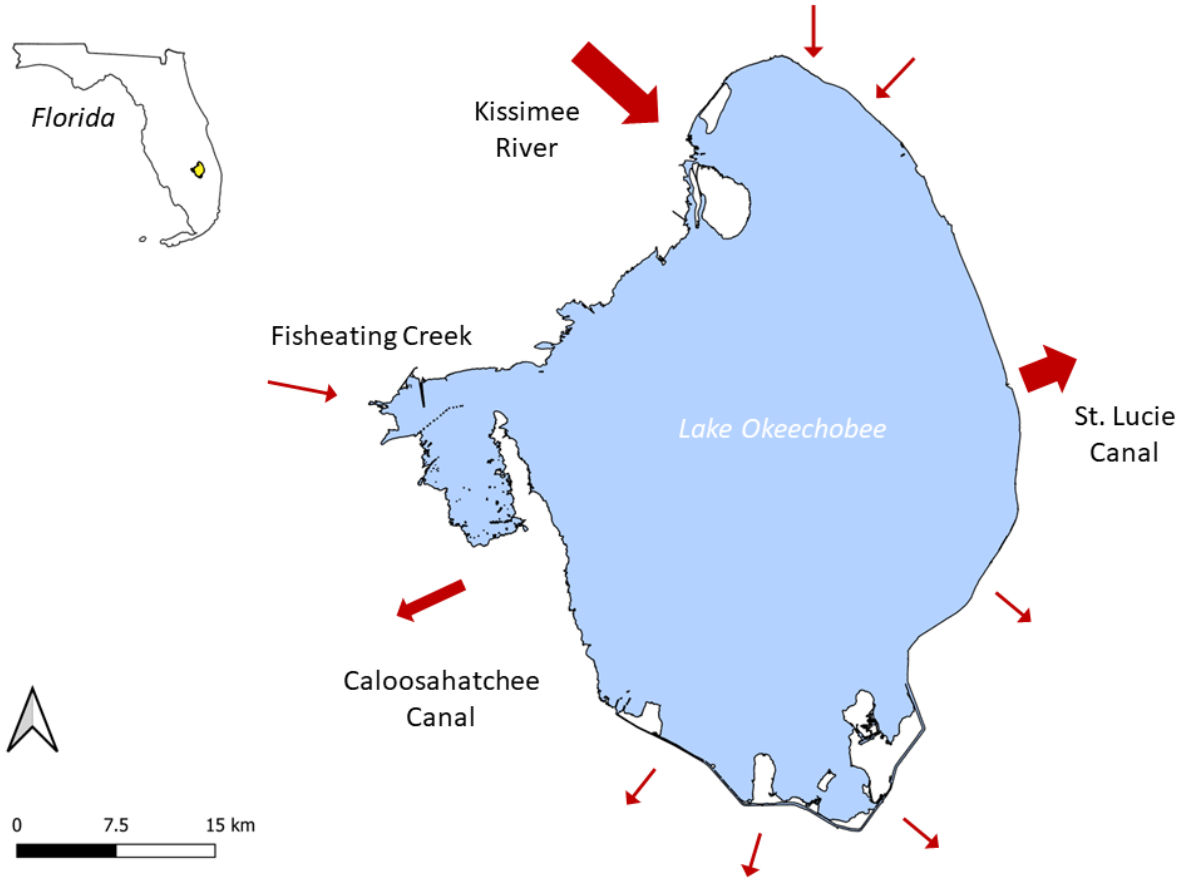

Fig. 7: Lake Okeeochobee and its main inflows and outflows (simplified). The smaller red arrows represent canals. Adopted from Havens et. al (1996) and Siders et. al (2020) (Havens et al., 1996; Siders & Havens, 2020).

The GPR model was applied to map the DOC concentration for Lake Okeechobee for each month of the year 2019 (Figure 8). DOC concentrations were mapped using Landsat 8 or Landsat 7 images. The images were visually checked for cloud cover and downloaded using the USGS earth explorer. Landsat 8 images with no or few clouds were generally preferred over Landsat 7 images which contain white stripes due to a failure of the Scan Line Corrector after 31 May 2003. As the ERA-5 Land product is provided in a reduced Gaussian grid with a quasi-uniform spacing over the globe, the layers containing the environmental information were resampled to universal transfer Mercator (UTM) and spatial resolution (30m) of the Landsat products using bilinear interpolation.

The satellite derived monthly DOC concentration in Lake Okeechobee in 2019 reveals distinct spatial and temporal patterns. High concentrations of DOC appear in the north-eastern part of the lake during the month from June to September. A high DOC concentration event was observed in June. Lower concentrations of DOC are observed from October to March. An

exception is the month of December, where higher concentrations of DOC were predicted in the south-eastern part of the lake.

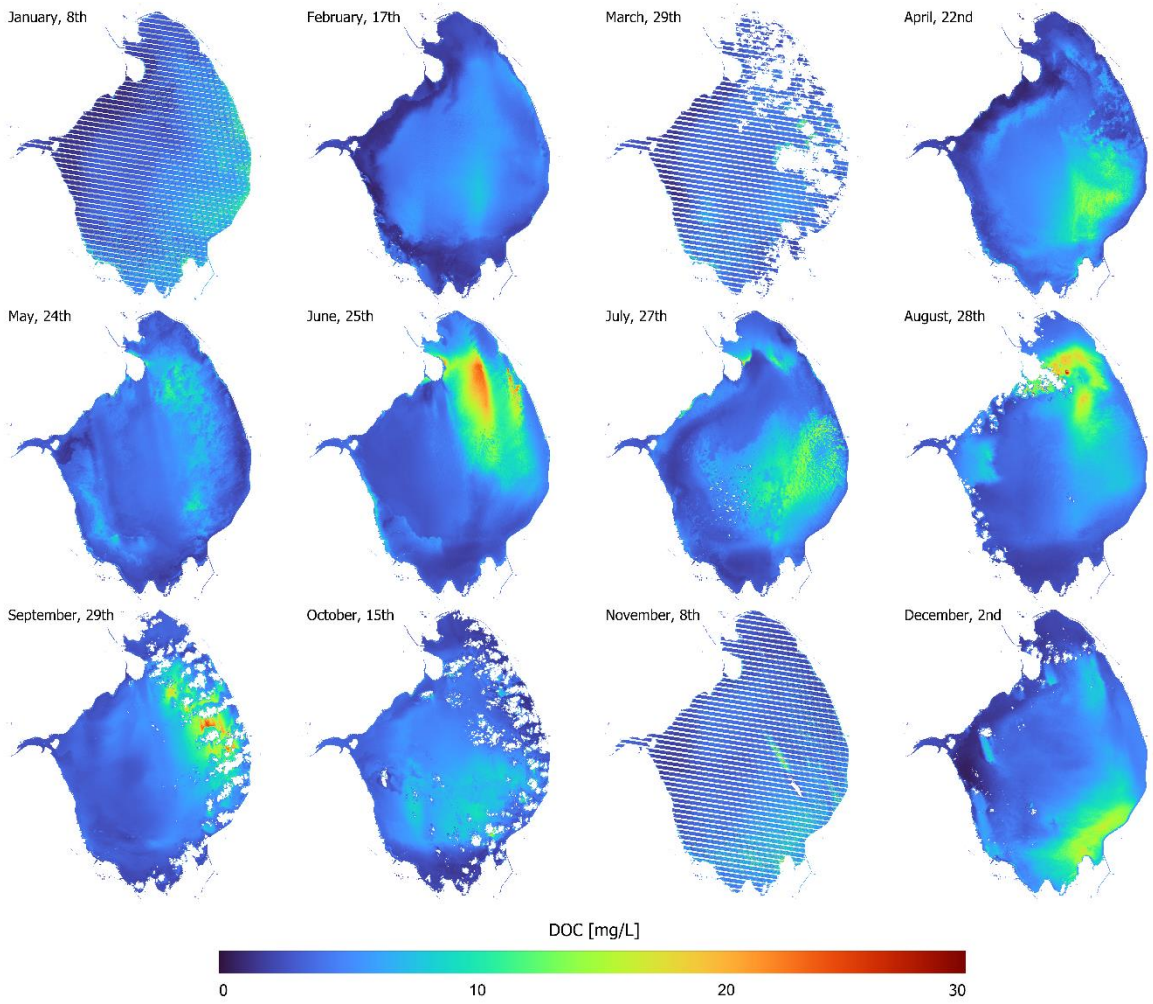

Fig. 8. DOC prediction for every month in 2019 on Lake Okeechobee, Florida based on Landsat 8 or Landsat 7 imagery using GPR with all environmental predictors. White spaces are cloud interferences. The white stripes are from failure of the Scan Line Corrector from Landsat 7 after 31 May 2003. Not all of the water body shown in Figure 7 is mapped due to low water occurrence.

6 Discussion

6.1 ML algorithm performance

The results imply the strongest performance of GPR, including environmental variables. This result is in line with several water quality studies in which GPR was found to outperform other ML approaches (Blix et al., 2018; Ruescas et al., 2018; Sun et al., 2021). Based on the model evaluation, the developed MBPNN predicted DOC concentration on the test dataset with nearly the same precision as GPR. However, cross-validation analysis showed that MBPNN was more unstable, suggesting its generalization performance varies more, according to the composition of the data it is applied to. These observations highlight the importance of cross-validation analysis in ML algorithm development and evaluation of ML algorithms.

The RFR scores showed slightly less accurate predictions on the test data set compared to GPR and MBPNN which might indicate a limitation of the traditional RFR method compared to the more advanced ML approaches when the prediction task is very difficult. Applying the developed RFR model to Lake Okeechobee showed that the model predicts DOC concentrations with a narrow range of values compared to the other approaches (Figure 9). Overall, RFR may only have moderate prediction errors, but the satellite-derived DOC concentrations suggest that it merely predicts the mean. Furthermore, the RFR prediction shows unnatural looking boundaries in the mapped DOC concentrations across the lake. This can be attributed to the influence of environmental predictors, which have a very coarse spatial resolution compared to the Landsat bands. As indicated by the permutation feature importance analysis, environmental predictors seem more important in RFR than in the other three models. Because of the coarse spatial resolution of the environmental predictors, their per-pixel variability is much lower, resulting in broad-scale structures and patterns with distinct boundaries.

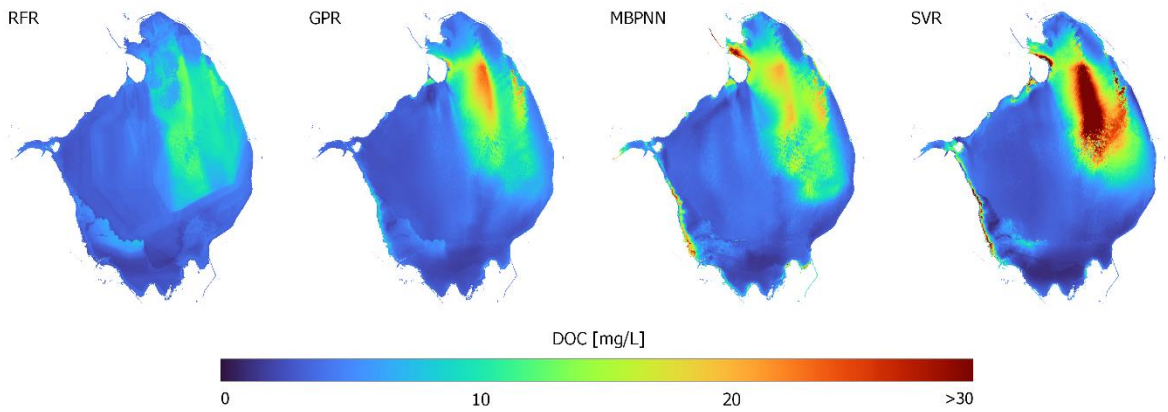

Fig 9. DOC prediction of the four ML algorithms tested for Lake Okeechobee on June 25th, 2019. (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MBPNN= Multilayer back-propagation neural network).

Among the four ML approaches tested, SVR had the poorest performance. One limitation of SVR is that it requires many hyperparameters to be tuned and its performance is very dependent on the selection of these parameters (4.2.3). In addition, the processing time SVR requires increases rapidly with large amounts of training data. Therefore, a randomized search for SVR resulted in long processing times that posed some limitations on searching the hyperparameter space. More advanced hyperparameter search methods based on Bayesian optimization could reduce such problems (Bergstra et al., 2013), but overall it highlights the limitations of SVR for such application types.

All four ML approaches tested predict very low or very high DOC values with larger errors. This is a common issue in water quality research, as there are usually fewer samples available for more extreme values. Here, further sampling efforts are necessary that focus specifically on waterbodies with very low or very high DOC concentrations, which may alleviate the degraded performance of the ML methods at tails of DOC concentration distributions. On the other hand, the Landsat data and environmental predictors might simply not be sensitive enough to capture the extreme DOC values, which could be investigated in further studies.

While cross validation analysis indicates that a robust overall performance of the ML algorithms, systematic bias might still exist on a regional level or for sites with specific characteristics. Figure 10 shows the logarithmic error of the GPR estimates compared to the measurements in the held-out test dataset based on the three different water types in which samples were taken. The plot shows that the interquartile error range of lakes expands further than estuary and streams. It suggests that the algorithm estimated DOC concentrations with larger error for lakes than for the other two water types. One reason could be the difference in the dynamics between the inland water types. While water flows in streams and estuaries are

constantly changing water constituents, water parameters in lakes have shown to be much more stable, which is also why a larger overpass time window was allowed for lakes (see section 4.1). The ML algorithm might not be able to capture these differences in dynamics, especially since there is an overrepresentation of samples for streams in the dataset (see appendix, Figure A1).

After plotting the spatial distribution of errors from the GPR model for the measurements in the held-out test dataset, no distinct spatial pattern could be observed (see appendix, Figure A4). It should however be considered that the range of samples spans over many years. It is therefore unlikely to find spatial patterns as the conditions might be considerably different even though the location of samples is in close proximity. On the other hand, if a specific time window (e.g. summer month in one year) is considered, the number of samples is too low to analyse if spatial patterns exist.

While we argue that the focus of this study is on large scale estimations of DOC, the heterogenous nature of the sample dataset limits the possibilities of more in depth analysis and consequently a better picture of how the model performs under specific conditions. It is yet another limitation that underlines the need of systematic and standardized sample efforts for inland waters.

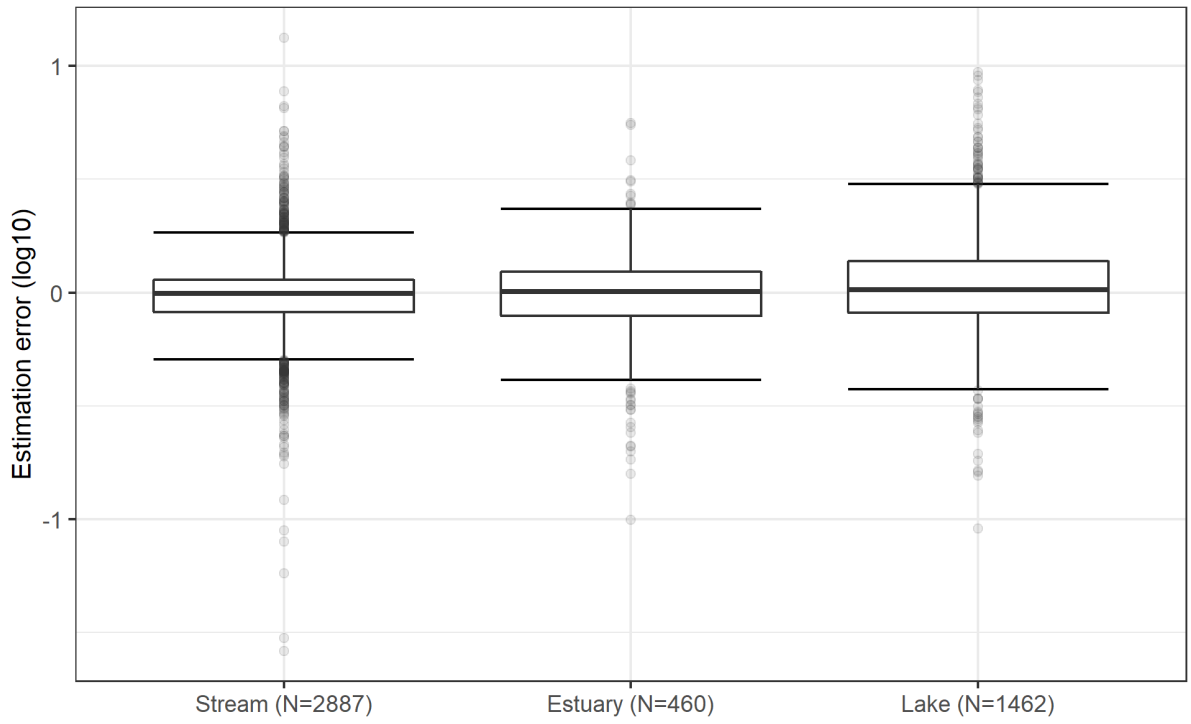

Fig. 10. Logarithmic error of the GPR estimates compared to the measurements in the hold out test dataset (N=4809) based on the three different water types.

An important aspect of ML algorithms in a practical sense is the processing times taken for training and applying the models (Table 3). The four ML algorithms differ considerably in terms of training and prediction time. The GPR model is very fast to train compared to the other methods, as hyperparameter tuning through random search is not required for GPR (4.2.5). With 60 iterations and 5-fold cross-validation, random search takes by far the most time in the other ML approaches, especially for the MBPNN. On the other hand, the GPR method takes longer to predict DOC values from satellite imagery. As with SVR, the computation complexity of GPR is $O(n^3)$, where *n* is the number of datapoints in the training set. However, unlike SVR that is a sparse solution, GPR is a dense solution that always uses all training points for prediction. As the Landsat image of Lake Okeechobee consist of nearly 1.5 Million pixels, it is likely that the difference of the dense and sparse solution reflects in the prediction time. With the trained GPR model, a workstation with 64GB RAM was close to its limits when values were predicted for the entire image of Lake Okeechobee. Hence, there may be practical issues when GPR is applied using machines with lower computational power, especially considering that predictions are usually done many times whereas the algorithm training only needs to be done once.

Table 3: Measured time (minutes) for training the algorithm with the US wide training dataset and measured time for predicting DOC concentration for the satellite image of Lake Okeechobee captured on June 25, 2019 (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

	Training (min)	**Prediction for Lake Okeechobee in June, 25th (min)**
RFR	245.2	2.77
GPR	4.57	23.2
MBPNN	558.72	0.57
SVR	308.75	9.95

Today, many different ML algorithms exist. In this thesis, we tested four popular ML approaches identified through several review papers, but other ML methods could have been of interest and possibly outperform GPR. Other advanced ML approaches such as Convolutional Neural Networks (Pu et al., 2019) or Mixture Density Networks (Pahlevan et al., 2022) have shown promising potential to retrieve water quality parameters from satellite imagery of inland waters. In addition, there usually also exist a large number of algorithmic modifications for each algorithm that make them suitable for specific settings. A time-consuming in-depth modification process for each of the ML approaches was not possible. We argue that the random search approach that is implemented here is likely to find a close to

optimal solution. However, the modification possibilities that each ML approach provides may not have been fully exploited and gaps in the optimization process of hyperparameters can exist.

6.2 Variable importance

Permutation feature importance analysis showed that the green band is the most important predictor of the six Landsat bands, followed by the red band and the blue band. This result is expected to a certain degree as wavelengths in the visible range and especially the green and red bands are commonly used to explain variations in water quality parameters, including the DOC related CDOM (Brezonik et al., 2015).

As indicated by permutation feature importance analysis, environmental predictors were not equally important for the models. Of all environmental predictors, the monthly average air temperature led to the highest increase in MSE for all models, indicating that it is the most important environmental predictor. This result is consistent with other studies that identified temperature as one of the main drivers of DOC concentrations in inland waters (Freeman et al., 2001; Kellerman et al., 2014; D. Liu et al., 2021; Strock et al., 2016). Higher average increases in MSE were also observed for LAI – high vegetation, which is an indicator of vegetation cover in the surrounding catchment. This result is in agreement with a previous US wide study, that found that forests and shrubland are positively correlated with DOC concentration in riverine systems (Yang et al., 2017). However, it should be considered that the way the information on LAI is provided for the ML approaches in this thesis does not represent the real-world conditions accurately, as it is based on a coarse resolution grid. Integrating the individual catchment of each waterbody could substantially change the importance evaluation of LAI (and related environmental predictors), but individual catchment properties are difficult to consider when the dataset consists of hundreds of waterbodies. This challenge could be addressed in a future study.

Surprisingly, total precipitation and average wind speed do not have large impacts on the model performance, which contradicts earlier findings (Jennings et al., 2012; Reche & Pace, 2002; Zhou et al., 2016). One reason might be that the low temporal and spatial resolution of environmental predictors used in this study is not capable of capturing the highly dynamic processes of precipitation and wind speed. For example, the time periods between a rainfall event and its subsequent effect on DOC concentrations in inland waters can vary considerably and depend on many factors, such as soil organic carbon of the soil, vegetation, and

hydrological conditions in the basin (Blaen et al., 2017; Wang et al., 2019). Changes in wind speed are are also highly dynamic. Thus, making use of information on wind speed and rainfall may require higher resolution monitoring (Jennings et al., 2012), implying that the ERA5 Land product based on daily averages could be more suitable for these variables. In addition, wind induced circulation and sediment distribution might be so important for streams as those are usually shielded from the wind.

LAI – low vegetation, surface net solar radiation and evaporation over inland water were not shown to have a large impact on the performance of the model. However, running the models without these three predictors still results in noticeable drops in RMSE (Table 4). Despite their seemingly marginal role in predicting DOC, the ML models seem to still gain knowledge from them. DOC variations have been related to many environmental drivers that may produce a variety of effects on DOC release that can even work in opposition to each other (Pagano et al., 2014) and vary significantly according to geographic region (Cool et al., 2014). For modelling such complex interactions in different environments, the ML approaches still seem to require as much information as possible, making use of the enhanced information due to interaction of multiple variables. This implies that using more information could still improve the model's performance. The ERA5-Land product offers many more environmental variables than those that were selected in this thesis. To reduce the complexity of the models, we selected environmental predictors which were found to be linked to DOC variations in previous studies (see section 3.2). However, other variables from the ERA5-Land product might still have improved the models even though they might not be directly associated to DOC variations.

Considering the low spatial and temporal resolution of environmental predictors in addition makes it difficult to draw final conclusions. There is still much that remains unknown about the relationships of external drivers and DOC concentration in inland waters at different spatiotemporal scales. However, the considerable improvements observed when including environmental predictors in the models underline the important role of environmental processes in modelling DOC variations in inland waters over larger scales.

Table 4: RMSE scores for ML algorithms using Landsat bands with air temperature and LAI-high vegetation as predictors and Landsat Bands with all available environmental predictors included (RFR = Random Forest Regression, SVR = Support Vector Regression, GPR = Gaussian Process Regression, MPBNN = Multilayer Backpropagation Neural Network).

	RMSE of the Landsat bands + air temperature and LAI - high vegetation	**RMSE of Landsat bands + all environmental predictors**
RFR	4.614	4.271
GPR	4.642	4.078
SVR	4.981	4.708
MBPNN	4.865	4.048

6.3 Visual assessment

Mapping the monthly DOC concentrations in Lake Okeechobee showed some distinct spatial patterns. For example, the monthly images show higher DOC in the north-eastern part of the lake from June to September. This pattern could potentially be linked to the increased cyanobacteria blooms that Lake Okeechobee has experienced lately from early May to September (Metcalf et al., 2018) as the findings of earlier studies suggest that increased DOC concentrations are linked to and partially explained by chlorophyll-a concentration (D. Liu et al., 2021; Ye et al., 2011). Furthermore, the high concentrations of DOC observed in June was in agreement with Pahlevan et al. (2020) who captured cyanobacteria blooms in Lake Okeechobee on the 5th of June 2019 using a mixture density model (Pahlevan et al., 2020, Fig. 8). The spatial pattern of DOC concentrations from June and August also suggests large input of DOC from the Kissime River where water with high levels of DOC enter the lake from a south-east direction which then distributes in the north and east part of the lake.

These observations indicate that spatial patterns of DOC concentration mapped by the GPR model in Lake Okeechobee are generally reasonable. However, since there are only very few studies that have mapped DOC concentrations for inland waters, a direct comparison of DOC estimates was not possible and thus uncertainties regarding the accuracy of mapped DOC remain. In addition, only a few images per month where available as Landsat sensors only have a revisit time of 16 days. The more recently launched Sentinel 2 satellites do not only have higher spatial and spectral resolution than Landsat 7 or 8 - one of their main advantages lies in the short revisiting times of only 5 days due to two sensors (Sentinel 2A, Sentinel 2B) in the orbit. Because of higher spatial, temporal, and spectral resolution, Sentinel 2 offers great potential in inland water remote sensing (Toming, Kutser, Laas, et al., 2016). However, images are only available from 2017 which would make the vast majority of the AquaSat dataset

unusable (AquaSat covers the period 1985 to 2019). If regular updates for the AquaSat dataset are foreseen in the future, the combination of Sentinel 2 and Landsat 8 will allow water quality mapping in a much finer spatio-temporal scale and give more robust monthly estimates (e.g., Chen et al., 2020).

A unique feature of GPR is that associated confidence intervals can be derived along with the estimations. Confidence intervals are expressed as standard deviation that accompany the mean estimation. They can be interpreted as reliability estimates and mapped for all mean predictions of the model (Verrelst et al., 2012). Figure 11 shows the standard deviations for each pixel of Lake Okeechobee on June 25th, 2019. The standard deviations clearly increase towards the shoreline of the lake. The increased uncertainties are most likely due to land adjacency effects, where scattering in the atmosphere contaminates water pixels close to the shoreline with radiation from the surrounding land (Section 2.1). Since Lake Okeechobee is a large lake with a large area of pure open water, the reduced reliability of DOC estimates is limited to the shoreline area. The proportion of uncertainties would, however, increase when inland waters are narrowly confined within land boundaries. More inland water bodies with varying morphometric characteristics could be tested in a future study to gain more insight into how land adjacency effects influence the uncertainty in DOC predictions. In the central part of the lake a random pattern can be observed which is likely an effect of atmospheric interference.

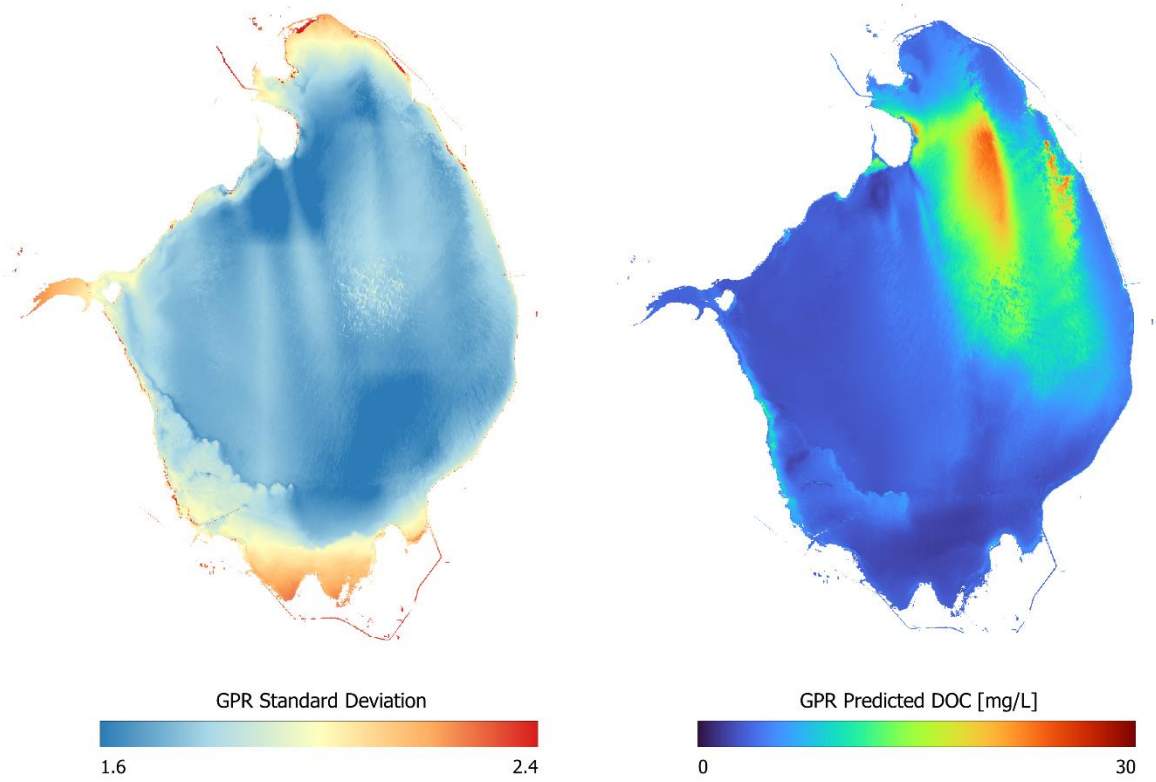

Fig. 11. GPR standard deviations of the DOC concentration estimates (left) and the mean DOC concentration prediction (right) for Lake Okeechobee on June 25th, 2019.

This problem is further illustrated in Figure 12. It shows estimated standard deviations from the GPR model in an area with cloud interference from the image taken on September 29th, 2019. GPR uncertainty estimates are generally higher around the outlines of clouds indicating that, although clouds are masked, the area surrounding the clouds is still affected by atmospheric effects and thus more difficult to model. Additionally, cloud shadows seem to influence the predictions, as their shapes are distinguishable in the predictions. Uncertainties related to atmospheric correction are one of the key challenges of water quality mapping (Section 2.1). Atmospheric correction for reflectance values in the AquaSat dataset is based on the standardized surface reflectance product developed by USGS. Although recent work found that the LaSRC model used for atmospheric correction for Landsat 8 products performs well over the Amazon River (Kuhn et al., 2019), the correction algorithms are still developed for terrestrial remote sensing. There is still big potential in optimizing atmospheric correction for inland waters. It was recently shown that different atmospheric processors can have considerable influence on the prediction outcome of water quality mapping (Pahlevan et al., 2021) or even that ML approaches can achieve better predictions without atmospheric correction, as information may be lost during the atmospheric correction progress (Medina-

Lopez, 2020). Evaluating further possibilities and estimating uncertainties in atmospheric correction processes could give important insights into how they affect satellite-derived DOC estimates.

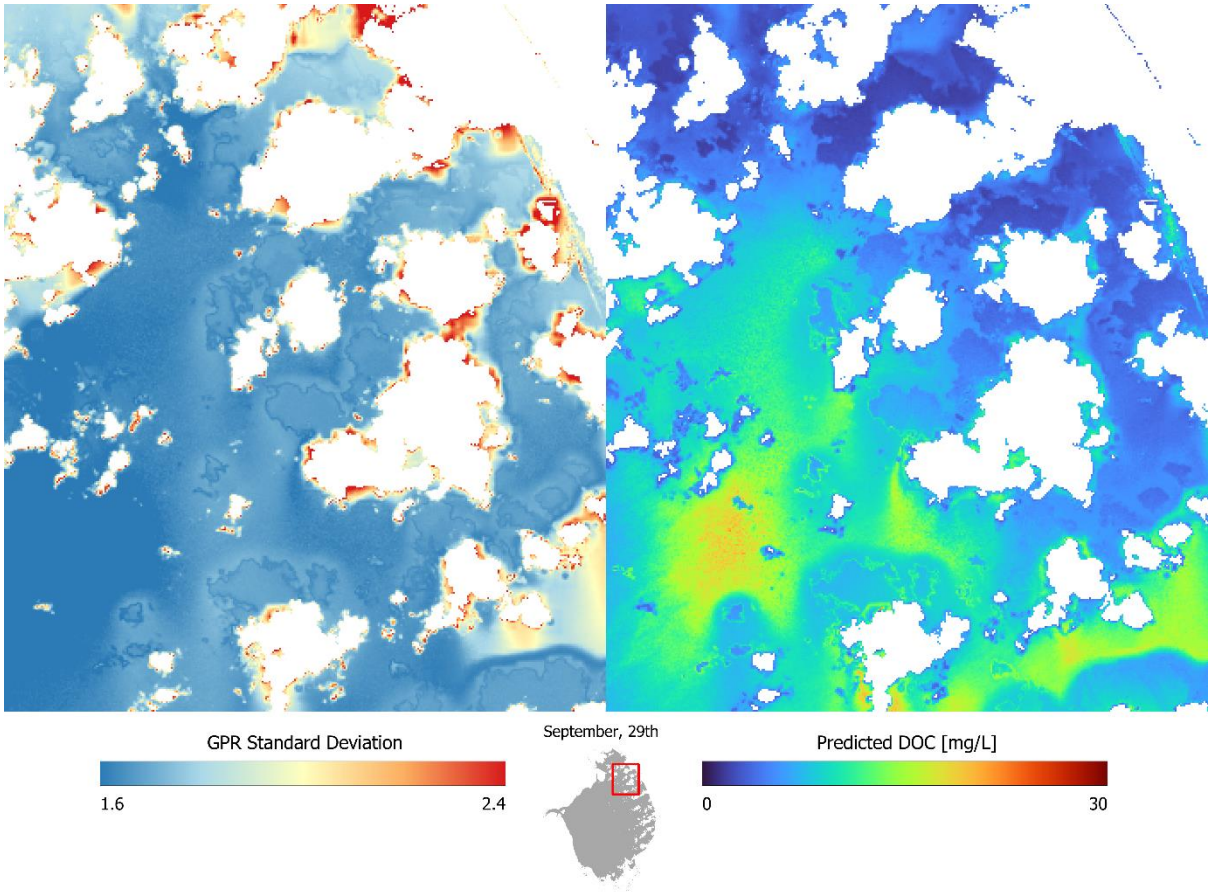

Fig. 12. GPR standard deviations of DOC concentration estimates (left) and the mean DOC concentration prediction (right) with cloud interference in the north-eastern part of Lake Okeechobee on 29th September, 2019.

6.4 Sampling

An important limitation in this project comes from the in-situ data. As described in Section 3.1, AquaSat is a combination of existing public datasets and thus the techniques, methods, and interpretations of DOC measurements may vary considerably. We have implemented several additional quality assurance steps in which the dataset was reduced notably with the intention of maintaining a high-quality standard for the remaining data (Section 4.1). Despite this, inconsistencies in the sampling technique, the instrumentation, and their corresponding calibration were not preventable. The extent to which such differences are reflected in the algorithm development and validation remains unclear. We argue that due to the large number of samples used in this thesis and the additional quality assurance steps deployed, minor

differences in some of the sampling methods will not have a significant effect on the overall model development and evaluation. However, the limitation underlines the potential value coordinated and standardized in situ data collection methods could have.

Another problem with heterogeneous datasets is the uneven distribution of features, which also exist in this dataset (see Appendix, Figure A1, A2 and A3). When the dataset is split randomly, there is an increased risk of introducing a bias in the training or test dataset. We argue that because the 10-fold cross-validation score does not differ substantially from the test score in all algorithms, except the MBPNN (Table 2), the data partitioning does not have a significant influence on the models' performance and a random split is reasonable.

Another limitation regards the temporal matching of in situ data and satellite imagery. The optimal threshold for temporal filtering is a trade-off between the accuracy of the model and its generalization capability. The high number of samples allowed us to set conservative temporal match-up limits (Section 3.1), yet any time that passes between the time of when the sample was taken and the satellite image was captured is a source of uncertainty. Also here, coordination and standardization efforts, where samples are for instance taken with consideration of satellite overpass times and clear weather conditions, bear great potential towards more robust satellite-derived water quality products. Lastly, although the dataset is not limited to a certain type of water body, all samples are located in the continental US. The models may estimate DOC reasonably well for areas in similar settings but may struggle in places that are totally different in terms of their environmental conditions. Such issues could be addressed with a larger international sample dataset.

6.5 Broader implications

This thesis presents the first attempt to predict DOC in inland waters over a large-scale area with remote sensing data. The results point towards promising potentials as well as important limitations for large-scale retrieval of DOC from inland waters. They demonstrate the predictive strength of advanced ML approaches faced with a complex learning task, such as GPR and MBPNN while at the same time highlighting limitations of the traditional RFR and SVR faced with high-dimensional data. A limitation that applied to all ML algorithms was the high prediction errors for very low and very high DOC values. Here, further sampling and research efforts are necessary that focus specifically on waterbodies with very low or very high DOC concentrations. Lastly, resource and time limitations made it impossible to cover all promising ML approaches and algorithm modifications that exist today within the scope of this

thesis. Other ML methods could have been of interest and possibly outperform the ML approaches tested.

Although there were limitations in terms of spatial and temporal resolutions, including environmental variables from the newly published ERA5 Land product resulted in considerable improvements in all ML approaches. This further highlights the potential of water quality estimations when data is assimilated within ecological and hydrodynamic models. This applies especially to DOC, which is driven by complex interactions between multiple environmental processes. Integrating more recently launched sensors like Sentinel 2 as well as environmental predictors at higher temporal and spatial resolution, such as the ERA5-Land daily average product, could further enhance the robustness of DOC inland water products on larger scales.

As confirmed in previous findings, atmospheric correction presents a major challenge in inland water remote sensing. The results of the GPR uncertainty estimates in this thesis underline the issues of land adjacency effects and cloud interference. Promising approaches for correction of atmosphere and land adjacency effects exist (e.g., Kiselev et al., 2015; Sterckx et al., 2015), and comprehensive assessments of existing atmospheric correction processors have recently been carried out with uncertainty estimates for other water quality parameters (Pahlevan et al., 2021). Further investigations could apply and test different atmospheric correction processors for DOC.

Open-source projects can accelerate scientific discovery immensely, as data is accessible to all researchers without any further restrictions (Bukata, 2013). Without funding that would allow a field sampling campaign or access to commercial data, this thesis relied entirely on open access products, most importantly, the AquaSat dataset. Thereby, this thesis demonstrates how openly available data of in situ water quality measurements can be turned into scientific knowledge. At the same time, it highlights limitations that often occur in such data. Due to the heterogeneous nature of combined data sets from many different sources, robust algorithm development was challenged by a lack of standardized collection and measurement methods, unevenly distributed features, synchronization with satellite overpass times, and limitation to a specific geographic area.

These challenges result in notable uncertainties that remain in the estimation of DOC in our model. Locally tuned ML models should usually be able to outperform our large-scale model as the prediction task is less complex due to less variations across smaller spatial scales. On the other hand, our model can estimate DOC concentrations with adequate accuracy for a much

larger variety of water bodies over much larger spatial scales. This addresses a key issue in inland water quality investigations as in-situ measurements for single waterbodies are often not available or insufficient. Furthermore, the developed model could estimate large scale trends of inland DOC dynamically over long time periods and seasons which in future could provide critical information, for example for monitoring of the global carbon cycle.

7 Conclusions

In this thesis we tested four ML approaches (RFR, GPR, MBPNN, SVR) to predict inland water DOC in the US using the novel open-source dataset AquaSat. Evaluation on the test dataset and 10-fold cross-validation scores showed that GPR using additional environmental predictors was the best performing ML algorithm with moderate prediction errors (RMSE: 4.08 mg/L). Due to fewer samples available, all models struggle to predict DOC values toward the lower and upper tails of the DOC distribution in the test dataset. Despite low spatial and temporal resolutions, environmental predictors notably improved the prediction of DOC in all algorithms highlighting the important role of environmental processes when explaining DOC variations. Permutation feature importance analysis showed that the green band, followed by the red and the blue bands were the most important predictors of the Landsat bands and that the monthly average air temperature followed by LAI -high vegetation- were the most important predictors of the environmental variables. Excluding less important environmental predictors still led to substantial drops in evaluation performance, suggesting that the ML approaches still gain knowledge due to the interaction of multiple variables.

Mapping the DOC concentrations for each month of Lake Okeechobee in 2019 showed spatial patterns with higher DOC values during the months June to September. These patterns, including an event with very high DOC concentrations captured in June, could be linked to previous findings, and indicate that DOC estimates by the GPR model are reasonable. However, the mapped DOC concentrations could not be further validated due to a lack of other mapped DOC products for inland waters. Further analysis using GPR standard deviation estimates showed high uncertainties in the predictions toward the lakes shorelines and the surrounding area of masked clouds, highlighting atmospheric correction issues for inland waters in the USGS Land surface reflectance products used in the AquaSat dataset. Uncertainties in DOC predictions related to atmospheric correction should be further investigated in the future by applying and comparing multiple atmospheric correction

processers and testing the performance with measured Landsat reflectance without atmospheric correction.

Although we note that locally tuned models are likely to outperform the developed model in terms of accuracy, it can address key issues of inland water remote sensing regarding the lack of in-situ measurements and has the potential to show large scale trends of inland DOC dynamically over long time periods and seasons. This thesis demonstrates how open source, large scale datasets like AquaSat in combination ML and remote sensing can make research toward large scale estimations of inland water DOC more realistic.

8 References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). {TensorFlow}: A System for {Large-Scale} Machine Learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16),

Anderson, N. J., & Stedmon, C. A. (2007). The effect of evapoconcentration on dissolved organic carbon concentration and quality in lakes of SW Greenland. *Freshwater Biology, 52*(2), 280-289. https://doi.org/https://doi.org/10.1111/j.1365-2427.2006.01688.x

Baines, S. B., & Pace, M. L. (1991). The production of dissolved organic matter by phytoplankton and its importance to bacteria: Patterns across marine and freshwater systems. *Limnology and Oceanography, 36*(6), 1078-1090. https://doi.org/https://doi.org/10.4319/lo.1991.36.6.1078

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems, 24*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research, 13*(2).

Bergstra, J., Yamins, D., & Cox, D. (2013). *Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms*. https://doi.org/10.25080/Majora-8b375195-003

Blaen, P. J., Khamis, K., Lloyd, C., Comer-Warner, S., Ciocca, F., Thomas, R. M., MacKenzie, A. R., & Krause, S. (2017). High-frequency monitoring of catchment nutrient exports reveals highly variable storm event responses and dynamic source zone activation. *Journal of Geophysical Research: Biogeosciences, 122*(9), 2265-2281. https://doi.org/https://doi.org/10.1002/2017JG003904

Blix, K., Pálffy, K., R. Tóth, V., & Eltoft, T. (2018). Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI. *Water, 10*(10). https://doi.org/10.3390/w10101428

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brezonik, P. L., Olmanson, L. G., Finlay, J. C., & Bauer, M. E. (2015). Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters. *Remote Sensing of Environment, 157*, 199-215. https://doi.org/https://doi.org/10.1016/j.rse.2014.04.033

Brönmark, C., & Hansson, L.-A. (2017). *The biology of lakes and ponds* [Book]. Oxford University Press. https://doi.org/10.1093/oso/9780198713593.001.0001

Bukata, R. P. (2013). Retrospection and introspection on remote sensing of inland water quality: "Like Déjà Vu All Over Again". *Journal of Great Lakes Research, 39*, 2-5. https://doi.org/https://doi.org/10.1016/j.jglr.2013.04.001

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans, 100*(C7), 13237-13254. https://doi.org/https://doi.org/10.1029/95JC00458

Chen, J., Zhu, W., Tian, Y. Q., & Yu, Q. (2020). Monitoring dissolved organic carbon by combining Landsat-8 and Sentinel-2 satellites: Case study in Saginaw River estuary, Lake Huron. *Science of The Total Environment, 718*, 137374. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.137374

Chen, S., & Hu, C. (2017). Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. *Remote Sensing of Environment, 201*, 115-132. https://doi.org/https://doi.org/10.1016/j.rse.2017.09.004

Chen, Z., Li, Y., & Pan, J. (2004). Distributions of colored dissolved organic matter and dissolved organic carbon in the Pearl River Estuary, China. *Continental Shelf Research, 24*(16), 1845-1856. https://doi.org/https://doi.org/10.1016/j.csr.2004.06.011

Codden, C. J., Snauffer, A. M., Mueller, A. V., Edwards, C. R., Thompson, M., Tait, Z., & Stubbins, A. (2021). Predicting dissolved organic carbon concentration in a dynamic salt marsh creek via

machine learning. *Limnology and Oceanography: Methods, 19*(2), 81-95. https://doi.org/https://doi.org/10.1002/lom3.10406

Cool, G., Lebel, A., Sadiq, R., & Rodriguez, M. J. (2014). Impact of catchment geophysical characteristics and climate on the regional variability of dissolved organic carbon (DOC) in surface water. *Science of The Total Environment, 490*, 947-956. https://doi.org/https://doi.org/10.1016/j.scitotenv.2014.05.091

Copernicus Climate Change Service. (2019). *C3S ERA5-Land reanalysis*. Retrieved 17/03/2022 from https://cds.climate.copernicus.eu/cdsapp#!/home

Correll, D. L., Jordan, T. E., & Weller, D. E. (2001). Effects of Precipitation, Air Temperature, and Land Use on Organic Carbon Discharges from Rhode River Watersheds. *Water, Air, and Soil Pollution, 128*(1), 139-159. https://doi.org/10.1023/A:1010337623092

Doerffer, R., & Schiller, H. (2007). The MERIS Case 2 water algorithm. *International Journal of Remote Sensing, 28*(3-4), 517-535. https://doi.org/10.1080/01431160600821127

Erlandsson, M., Futter, M. N., Kothawala, D. N., & Köhler, S. J. (2012). Variability in spectral absorbance metrics across boreal lake waters. *Journal of Environmental Monitoring, 14*(10), 2643-2652. https://doi.org/10.1039/C2EM30266G

Feng, L., Hou, X., Li, J., & Zheng, Y. (2018). Exploring the potential of Rayleigh-corrected reflectance in coastal and inland water applications: A simple aerosol correction method and its merits. *ISPRS Journal of Photogrammetry and Remote Sensing, 146*, 52-64. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2018.08.020

Freeman, C., Evans, C. D., Monteith, D. T., Reynolds, B., & Fenner, N. (2001). Export of organic carbon from peat soils. *Nature, 412*(6849), 785-785. https://doi.org/10.1038/35090628

Gege, P. (2017). Chapter 2 - Radiative Transfer Theory for Inland Waters. In D. R. Mishra, I. Ogashawara, & A. A. Gitelson (Eds.), *Bio-optical Modeling and Remote Sensing of Inland Waters* (pp. 25-67). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-804644-9.00002-1

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated. http://ebookcentral.proquest.com/lib/lund/detail.action?docID=5892320

Giri, S. (2021). Water quality prospective in Twenty First Century: Status of water quality in major river basins, contemporary strategies and impediments: A review. *Environmental Pollution, 271*, 116332. https://doi.org/https://doi.org/10.1016/j.envpol.2020.116332

Griffin, C. G., Finlay, J. C., Brezonik, P. L., Olmanson, L., & Hozalski, R. M. (2018). Limitations on using CDOM as a proxy for DOC in temperate lakes. *Water Research, 144*, 719-727. https://doi.org/https://doi.org/10.1016/j.watres.2018.08.007

Guan, Q., Feng, L., Hou, X., Schurgers, G., Zheng, Y., & Tang, J. (2020). Eutrophication changes in fifty large lakes on the Yangtze Plain of China derived from MERIS and OLCI observations. *Remote Sensing of Environment, 246*, 111890. https://doi.org/https://doi.org/10.1016/j.rse.2020.111890

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K. H., & Pun, L. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-ii waters: A case study of hong kong. *Remote Sensing, 11*(6), Article 617. https://doi.org/10.3390/rs11060617

Hassan, N., & Woo, C. S. (2021). Machine Learning Application in Water Quality Using Satellite Data. *IOP Conference Series: Earth and Environmental Science, 842*(1), 012018. https://doi.org/10.1088/1755-1315/842/1/012018

Havens, K., Aumen, N., James, R., & Smith, V. (1996). Rapid Ecological Changes in a Large Subtropical Lake Undergoing Cultural Eutrophication. *Ambio, 25*, 150-155.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M.,

Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., & Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999-2049. https://doi.org/https://doi.org/10.1002/qj.3803

Hestir, E. L., Brando, V., Campbell, G., Dekker, A., & Malthus, T. (2015). The relationship between dissolved organic matter absorption and dissolved organic carbon in reservoirs along a temperate to tropical gradient. *Remote Sensing of Environment, 156*, 395-402. https://doi.org/https://doi.org/10.1016/j.rse.2014.09.022

IOCCG. (2018). Earth observations in support of global water quality monitoring (eds. Greb, S., Dekker, A. and Binding, C.). *International Ocean-Colour Coordinating Group (IOCCG)*. https://doi.org/http://dx.doi.org/10.25607/OBP-113

Jennings, E., Jones, S., Arvola, L., Staehr, P. A., Gaiser, E., Jones, I. D., Weathers, K. C., Weyhenmeyer, G. A., Chiu, C.-Y., & De Eyto, E. (2012). Effects of weather-related episodic events in lakes: an analysis based on high-frequency data. *Freshwater Biology, 57*(3), 589-601. https://doi.org/https://doi.org/10.1111/j.1365-2427.2011.02729.x

Jiang, G., Ma, R., Loiselle, S. A., & Duan, H. (2012). Optical approaches to examining the dynamics of dissolved organic carbon in optically complex inland waters. *Environmental Research Letters, 7*(3), 034014. https://doi.org/10.1088/1748-9326/7/3/034014

Ju, J., Roy, D. P., Vermote, E., Masek, J., & Kovalskyy, V. (2012). Continental-scale validation of MODIS-based and LEDAPS Landsat ETM+ atmospheric correction methods. *Remote Sensing of Environment, 122*, 175-184. https://doi.org/https://doi.org/10.1016/j.rse.2011.12.025

Kellerman, A. M., Dittmar, T., Kothawala, D. N., & Tranvik, L. J. (2014). Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nature Communications, 5*(1), 3804. https://doi.org/10.1038/ncomms4804

Kim, Y. H., Im, J., Ha, H. K., Choi, J.-K., & Ha, S. (2014). Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing, 51*(2), 158-174. https://doi.org/10.1080/15481603.2014.900983

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

Kiselev, V., Bulgarelli, B., & Heege, T. (2015). Sensor independent adjacency correction algorithm for coastal and inland water systems. *Remote Sensing of Environment, 157*, 85-95. https://doi.org/https://doi.org/10.1016/j.rse.2014.07.025

Koivusalo, M., Pukkala, E., Vartiainen, T., Jaakkola, J. J. K., & Hakulinen, T. (1997). Drinking water chlorination and cancer - A historical cohort study in Finland. *Cancer Causes and Control, 8*(2), 192-200. https://doi.org/10.1023/A:1018420229802

Kuhn, C., de Matos Valerio, A., Ward, N., Loken, L., Sawakuchi, H. O., Kampel, M., Richey, J., Stadler, P., Crawford, J., Striegl, R., Vermote, E., Pahlevan, N., & Butman, D. (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. https://doi.org/https://doi.org/10.1016/j.rse.2019.01.023

Kutser, T., Casal Pascual, G., Barbosa, C., Paavel, B., Ferreira, R., Carvalho, L., & Toming, K. (2016). Mapping inland water carbon content with Landsat 8 data. *International Journal of Remote Sensing, 37*(13), 2950-2961. https://doi.org/10.1080/01431161.2016.1186852

Kutser, T., Koponen, S., Kallio, K. Y., Fincke, T., & Paavel, B. (2017). Chapter 4 - Bio-optical Modeling of Colored Dissolved Organic Matter. In D. R. Mishra, I. Ogashawara, & A. A. Gitelson (Eds.), *Bio-optical Modeling and Remote Sensing of Inland Waters* (pp. 101-128). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-804644-9.00004-5

Lauerwald, R., Laruelle, G. G., Hartmann, J., Ciais, P., & Regnier, P. A. G. (2015). Spatial patterns in CO2 evasion from the global river network. *Global Biogeochemical Cycles, 29*(5), 534-554. https://doi.org/https://doi.org/10.1002/2014GB004941

Liu, D., Yu, S., Xiao, Q., Qi, T., & Duan, H. (2021). Satellite estimation of dissolved organic carbon in eutrophic Lake Taihu, China. *Remote Sensing of Environment, 264*, Article 112572. https://doi.org/10.1016/j.rse.2021.112572

Liu, G., Li, S., Song, K., Wang, X., Wen, Z., Kutser, T., Jacinthe, P. A., Shang, Y., Lyu, L., Fang, C., Yang, Y., Yang, Q., Zhang, B., Cheng, S., & Hou, J. (2021). Remote sensing of CDOM and DOC in alpine lakes across the Qinghai-Tibet Plateau using Sentinel-2A imagery data. *Journal of Environmental Management, 286*, Article 112231. https://doi.org/10.1016/j.jenvman.2021.112231

Loiselle, S. A., Azza, N., Gichuki, J., Bracchini, L., Tognazzi, A., Dattilo, A. M., Rossi, C., & Cozar, A. (2010). Spatial dynamics of chromophoric dissolved organic matter in nearshore waters of Lake Victoria. *Aquatic Ecosystem Health & Management, 13*(2), 185-195. https://doi.org/10.1080/14634988.2010.481236

Magnus, P., Jaakkola, J. J. K., Skrondal, A., Alexander, J., Becher, G., Krogh, T., & Dybing, E. (1999). Water chlorination and birth defects. *Epidemiology, 10*(5), 513-517. https://doi.org/10.1097/00001648-199909000-00008

Maier, P. M., Keller, S., & Hinz, S. (2021). Deep learning with wasi simulation data for estimating chlorophyll a concentration of inland water bodies. *Remote Sensing, 13*(4), 1-27, Article 718. https://doi.org/10.3390/rs13040718

Malo, J., & Camps-Valls, G. (2011). A Review of Kernel Methods in Remote Sensing Data Analysis. In (pp. 171-206). https://doi.org/10.1007/978-3-642-14212-3_10

Massicotte, P., Asmala, E., Stedmon, C., & Markager, S. (2017). Global distribution of dissolved organic matter along the aquatic continuum: Across rivers, lakes and oceans. *Science of The Total Environment, 609*, 180-191. https://doi.org/https://doi.org/10.1016/j.scitotenv.2017.07.076

Medina-Lopez, E. (2020). Machine learning and the end of atmospheric corrections: A comparison between high-resolution sea surface salinity in coastal areas from top and bottom of atmosphere Sentinel-2 imagery. *Remote Sensing, 12*(18), Article 2924. https://doi.org/10.3390/RS12182924

Metcalf, J. S., Banack, S. A., Powell, J. T., Tymm, F. J. M., Murch, S. J., Brand, L. E., & Cox, P. A. (2018). Public health responses to toxic cyanobacterial blooms: perspectives from the 2016 Florida event. *Water Policy, 20*(5), 919-932. https://doi.org/10.2166/wp.2018.012

Millennium Ecosystem Assessment. (2005). *Ecosystems and human well-being: wetlands and water*. World resources institute. http://www.millenniumassessment.org/documents/document.358.aspx.pdf. Accessed 3 Feb 2022

Morfitt, R., Barsi, J., Levy, R., Markham, B., Micijevic, E., Ong, L., Scaramuzza, P., & Vanderwerff, K. (2015). Landsat-8 Operational Land Imager (OLI) Radiometric Performance On-Orbit. *Remote Sensing, 7*(2). https://doi.org/10.3390/rs70202208

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of Model Performance Based On the Log Accuracy Ratio. *Space Weather, 16*(1), 69-88. https://doi.org/https://doi.org/10.1002/2017SW001669

Moses, W. J., Sterckx, S., Montes, M. J., De Keukelaere, L., & Knaeps, E. (2017). Chapter 3 - Atmospheric Correction for Inland Waters. In D. R. Mishra, I. Ogashawara, & A. A. Gitelson (Eds.), *Bio-optical Modeling and Remote Sensing of Inland Waters* (pp. 69-100). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-804644-9.00003-3

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing, 66*(3), 247-259. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2010.11.001

Muñoz-Sabater, J. (2019). *ERA5-Land monthly averaged data from 1981 to present.* Retrieved 24-03-2022 from https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.68d2bb30?tab=overview

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J. N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data, 13*(9), 4349-4383. https://doi.org/10.5194/essd-13-4349-2021

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Ogashawara, I. (2015). Terminology and classification of bio-optical algorithms. *Remote Sensing Letters, 6*(8), 613-617. https://doi.org/10.1080/2150704X.2015.1066523

Ogashawara, I., Mishra, D. R., & Gitelson, A. A. (2017). Chapter 1 - Remote Sensing of Inland Waters: Background and Current State-of-the-Art. In D. R. Mishra, I. Ogashawara, & A. A. Gitelson (Eds.), *Bio-optical Modeling and Remote Sensing of Inland Waters* (pp. 1-24). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-804644-9.00001-X

Olmanson, L. G., Brezonik, P. L., Finlay, J. C., & Bauer, M. E. (2016). Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes. *Remote Sensing of Environment, 185*, 119-128. https://doi.org/https://doi.org/10.1016/j.rse.2016.01.007

Pagano, T., Bida, M., & Kenny, J. E. (2014). Trends in Levels of Allochthonous Dissolved Organic Carbon in Natural Water: A Review of Potential Mechanisms under a Changing Climate. *Water, 6*(10). https://doi.org/10.3390/w6102862

Pahlevan, N., Mangin, A., Balasubramanian, S. V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M., Giardino, C., Gurlin, D., Fan, Y., Harmel, T., Hunter, P., Ishikaza, J., Kratzer, S., Lehmann, M. K., Ligi, M., Ma, R., Martin-Lauzer, F.-R., Olmanson, L., Oppelt, N., Pan, Y., Peters, S., Reynaud, N., Sander de Carvalho, L. A., Simis, S., Spyrakos, E., Steinmetz, F., Stelzer, K., Sterckx, S., Tormos, T., Tyler, A., Vanhellemont, Q., & Warren, M. (2021). ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. *Remote Sensing of Environment, 258*, 112366. https://doi.org/https://doi.org/10.1016/j.rse.2021.112366

Pahlevan, N., Smith, B., Alikas, K., Anstee, J., Barbosa, C., Binding, C., Bresciani, M., Cremella, B., Giardino, C., Gurlin, D., Fernandez, V., Jamet, C., Kangro, K., Lehmann, M. K., Loisel, H., Matsushita, B., Hà, N., Olmanson, L., Potvin, G., Simis, S. G. H., VanderWoude, A., Vantrepotte, V., & Ruiz-Verdù, A. (2022). Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. *Remote Sensing of Environment, 270*, 112860. https://doi.org/https://doi.org/10.1016/j.rse.2021.112860

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. (2020). 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. https://doi.org/https://doi.org/10.1016/j.rse.2019.111604

Palmer, S. C. J., Kutser, T., & Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment, 157*, 1-8. https://doi.org/https://doi.org/10.1016/j.rse.2014.09.021

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Pu, F., Ding, C., Chao, Z., Yu, Y., & Xu, X. (2019). Water-Quality Classification of Inland Lakes Using Landsat8 Images by Convolutional Neural Networks. *Remote Sensing, 11*(14). https://doi.org/10.3390/rs11141674

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Raymond, P. A., Hartmann, J., Lauerwald, R., Sobek, S., McDonald, C., Hoover, M., Butman, D., Striegl, R., Mayorga, E., Humborg, C., Kortelainen, P., Dürr, H., Meybeck, M., Ciais, P., & Guth, P. (2013). Global carbon dioxide emissions from inland waters. *Nature, 503*(7476), 355-359. https://doi.org/10.1038/nature12760

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., Kreft, J., Read, J. S., & Winslow, L. A. (2017). Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research, 53*(2), 1735-1745. https://doi.org/https://doi.org/10.1002/2016WR019993

Reche, I., & Pace, M. L. (2002). Linking dynamics of dissolved organic carbon in a forested lake with environmental factors. *Biogeochemistry, 61*(1), 21-36. https://doi.org/10.1023/A:1020234900383

Regnier, P., Friedlingstein, P., Ciais, P., Mackenzie, F. T., Gruber, N., Janssens, I. A., Laruelle, G. G., Lauerwald, R., Luyssaert, S., Andersson, A. J., Arndt, S., Arnosti, C., Borges, A. V., Dale, A. W., Gallego-Sala, A., Goddéris, Y., Goossens, N., Hartmann, J., Heinze, C., Ilyina, T., Joos, F., LaRowe, D. E., Leifeld, J., Meysman, F. J. R., Munhoven, G., Raymond, P. A., Spahni, R., Suntharalingam, P., & Thullner, M. (2013). Anthropogenic perturbation of the carbon fluxes from land to ocean. *Nature Geoscience, 6*(8), 597-607. https://doi.org/10.1038/ngeo1830

Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M., & Pavelsky, T. M. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research, 55*(11), 10012-10025. https://doi.org/https://doi.org/10.1029/2019WR024883

Ruescas, A. B., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K., & Camps-Valls, G. (2018). Machine Learning Regression Approaches for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data. *Remote Sensing, 10*(5). https://doi.org/10.3390/rs10050786

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., Maalouf, S., & Adams, C. (2020). Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews, 205*, 103187. https://doi.org/https://doi.org/10.1016/j.earscirev.2020.103187

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*(6), 7404-7422. https://doi.org/10.1364/OE.26.007404

Siders, Z. A., & Havens, K. E. (2020). Revisiting the total maximum daily load total phosphorus goal in Lake Okeechobee. *Hydrobiologia, 847*(20), 4221-4232. https://doi.org/10.1007/s10750-020-04406-8

Smith, B., Pahlevan, N., Schalles, J., Ruberg, S., Errera, R., Ma, R., Giardino, C., Bresciani, M., Barbosa, C., Moore, T., Fernandez, V., Alikas, K., & Kangro, K. (2021). A Chlorophyll-a Algorithm for Landsat-8 Based on Mixture Density Networks. *Frontiers in Remote Sensing, 1*. https://doi.org/10.3389/frsen.2020.623678

Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., Boyer, M. G., Bremigan, M. T., Carpenter, S. R., Carr, J. W., Cheruvelil, K. S., Christel, S. T., Claucherty, M., Collins, S. M., Conroy, J. D., Downing, J. A., Dukett, J., Fergus, C. E., Filstrup, C. T., Funk, C., Gonzalez, M. J., Green, L. T., Gries, C., Halfman, J. D., Hamilton, S. K., Hanson, P. C., Henry, E. N., Herron, E. M., Hockings, C., Jackson, J. R., Jacobson-Hedin, K., Janus, L. L., Jones, W. W., Jones, J. R., Keson, C. M., King, K. B. S., Kishbaugh, S. A., Lapierre, J.-F., Lathrop, B., Latimore, J. A., Lee, Y., Lottig, N. R., Lynch, J. A., Matthews, L. J., McDowell, W. H., Moore, K. E. B., Neff, B. P., Nelson, S. J., Oliver, S. K., Pace, M. L., Pierson, D. C., Poisson, A. C., Pollard, A. I., Post, D. M., Reyes, P. O., Rosenberry, D. O., Roy, K. M., Rudstam, L. G., Sarnelle, O., Schuldt, N. J., Scott, C. E., Skaff, N. K., Smith, N. J., Spinelli, N. R., Stachelek, J., Stanley, E. H., Stoddard, J. L., Stopyak, S. B., Stow, C. A., Tallant, J. M., Tan, P.-N., Thorpe, A. P., Vanni, M. J., Wagner, T., Watkins, G., Weathers, K. C., Webster, K. E., White, J. D., Wilmes, M. K., & Yuan, S. (2017). LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality

for thousands of US lakes. *GigaScience, 6*(12), gix101. https://doi.org/10.1093/gigascience/gix101

Sprague, L. A., Oelsner, G. P., & Argue, D. M. (2017). Challenges with secondary use of multi-source water-quality data in the United States. *Water Research, 110*, 252-261. https://doi.org/https://doi.org/10.1016/j.watres.2016.12.024

Sterckx, S., Knaeps, S., Kratzer, S., & Ruddick, K. (2015). SIMilarity Environment Correction (SIMEC) applied to MERIS data over inland and coastal waters. *Remote Sensing of Environment, 157*, 96-110. https://doi.org/https://doi.org/10.1016/j.rse.2014.06.017

Strock, K. E., Saros, J. E., Nelson, S. J., Birkel, S. D., Kahl, J. S., & McDowell, W. H. (2016). Extreme weather years drive episodic changes in lake chemistry: implications for recovery from sulfate deposition and long-term trends in dissolved organic carbon. *Biogeochemistry, 127*(2), 353-365. https://doi.org/10.1007/s10533-016-0185-9

Sun, D. Y., Li, Y. M., Wang, Q., Lu, H., Le, C. F., Huang, C. C., & Gong, S. Q. (2011). A neural-network model to retrieve CDOM absorption from in situ measured hyperspectral data in an optically complex lake: Lake Taihu case study. *International Journal of Remote Sensing, 32*(14), 4005-4022. https://doi.org/10.1080/01431161.2010.481297

Sun, X., Zhang, Y., Zhang, Y., Shi, K., Zhou, Y., & Li, N. (2021). Machine learning algorithms for chromophoric dissolved organic matter (Cdom) estimation based on landsat 8 images. *Remote Sensing, 13*(18), Article 3560. https://doi.org/10.3390/rs13183560

Tenjo, C., Ruiz-Verdú, A., Van Wittenberghe, S., Delegido, J., & Moreno, J. (2021). A New Algorithm for the Retrieval of Sun Induced Chlorophyll Fluorescence of Water Bodies Exploiting the Detailed Spectral Shape of Water-Leaving Radiance. *Remote Sensing, 13*(2). https://doi.org/10.3390/rs13020329

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., & Tranvik, L. J. (2020). Predicting lake dissolved organic carbon at a global scale. *Scientific Reports, 10*(1), 8471. https://doi.org/10.1038/s41598-020-65010-3

Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., & Nõges, T. (2016). First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery. *Remote Sensing, 8*(8). https://doi.org/10.3390/rs8080640

Toming, K., Kutser, T., Tuvikene, L., Viik, M., & Nõges, T. (2016). Dissolved organic carbon and its potential predictors in eutrophic lakes. *Water Research, 102*, 32-40. https://doi.org/https://doi.org/10.1016/j.watres.2016.06.012

Topp, S. N., Pavelsky, T. M., Dugan, H. A., Yang, X., Gardner, J., & Ross, M. R. V. (2021). Shifting Patterns of Summer Lake Color Phenology in Over 26,000 US Lakes. *Water Resources Research, 57*(5), e2020WR029123. https://doi.org/https://doi.org/10.1029/2020WR029123

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment, 185*, 46-56. https://doi.org/https://doi.org/10.1016/j.rse.2016.04.008

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sensing of Environment, 118*, 127-139. https://doi.org/https://doi.org/10.1016/j.rse.2011.11.002

Wagle, N., Acharya, T. D., & Lee, D. H. (2020). Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data. *Sensors and Materials, 32*(11), 3879-3892. https://doi.org/10.18494/SAM.2020.2953

Wang, L., Yen, H., E, X., Chen, L., & Wang, Y. (2019). Dissolved organic carbon driven by rainfall events from a semi-arid catchment during concentrated rainfall season in the Loess Plateau, China. *Hydrol. Earth Syst. Sci., 23*(7), 3141-3153. https://doi.org/10.5194/hess-23-3141-2019

Warren, M. A., Simis, S. G. H., & Selmes, N. (2021). Complementary water quality observations from high and medium resolution Sentinel sensors by aligning chlorophyll-a and turbidity

algorithms. *Remote Sensing of Environment, 265*, 112651. https://doi.org/https://doi.org/10.1016/j.rse.2021.112651

Wetzel, R. G. (2001). 11 - The Inorganic Complex. In R. G. Wetzel (Ed.), *Limnology (Third Edition)* (pp. 187-204). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-08-057439-4.50015-0

Williamson, C. E., Overholt, E. P., Pilla, R. M., Leach, T. H., Brentrup, J. A., Knoll, L. B., Mette, E. M., & Moeller, R. E. (2015). Ecological consequences of long-term browning in lakes. *Scientific Reports, 5*, Article 18666. https://doi.org/10.1038/srep18666

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., Fosnight, E. A., Shaw, J., Masek, J. G., & Roy, D. P. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment, 185*, 271-283. https://doi.org/https://doi.org/10.1016/j.rse.2015.11.032

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295-316. https://doi.org/https://doi.org/10.1016/j.neucom.2020.07.061

Yang, Q., Zhang, X., Xu, X., & Asrar, G. R. (2017). An Analysis of Terrestrial and Aquatic Environmental Controls of Riverine Dissolved Organic Carbon in the Conterminous United States. *Water, 9*(6). https://doi.org/10.3390/w9060383

Ye, L., Shi, X., Wu, X., Zhang, M., Yu, Y., Li, D., & Kong, F. (2011). Dynamics of dissolved organic carbon after a cyanobacterial bloom in hypereutrophic Lake Taihu (China). *Limnologica, 41*(4), 382-388. https://doi.org/https://doi.org/10.1016/j.limno.2011.06.001

Zandler, H., Senftl, T., & Vanselow, K. A. (2020). Reanalysis datasets outperform other gridded climate products in vegetation change analysis in peripheral conservation areas of Central Asia. *Scientific Reports, 10*(1), 22446. https://doi.org/10.1038/s41598-020-79480-y

Zare Farjoudi, S., & Alizadeh, Z. (2021). A comparative study of total dissolved solids in water estimation models using Gaussian process regression with different kernel functions. *Environmental Earth Sciences, 80*(17), 557. https://doi.org/10.1007/s12665-021-09798-x

Zhang, Y., Zhou, L., Zhou, Y., Zhang, L., Yao, X., Shi, K., Jeppesen, E., Yu, Q., & Zhu, W. (2021). Chromophoric dissolved organic matter in inland waters: Present knowledge and future challenges. *Science of The Total Environment, 759*, 143550. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.143550

Zhou, Y., Zhou, J., Jeppesen, E., Zhang, Y., Qin, B., Shi, K., Tang, X., & Han, X. (2016). Will enhanced turbulence in inland waters result in elevated production of autochthonous dissolved organic matter? *Science of The Total Environment, 543*, 405-415. https://doi.org/https://doi.org/10.1016/j.scitotenv.2015.11.051

Zhou, Z.-H. (2021a). Decision Trees. In Z.-H. Zhou (Ed.), *Machine Learning* (pp. 79-102). Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3_4

Zhou, Z.-H. (2021b). Neural Networks. In Z.-H. Zhou (Ed.), *Machine Learning* (pp. 103-128). Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3_5

Zhou, Z.-H. (2021c). Support Vector Machine. In Z.-H. Zhou (Ed.), *Machine Learning* (pp. 129-153). Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3_6

Zhu, W., Yu, Q., Tian, Y. Q., Becker, B. L., Zheng, T., & Carrick, H. J. (2014). An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments. *Remote Sensing of Environment, 140*, 766-778. https://doi.org/https://doi.org/10.1016/j.rse.2013.10.015

9 Appendix

The appendix consists of two tables and three figures. The first table (Table A1) shows quality assurance steps when Ross et. al joined Landsat and the in-situ datasets during the development of the AquaSat dataset. The second table (Table A2) shows the tuned hyperparameter values after random search was done. The first Figure (Fig. A1) shows the number of samples per inland water type in the preprocessed dataset. The second Figure (Fig. A2) shows the number of samples per Landsat sensor type and the third Figure (Fig. A3) shows the temporal distribution of the time the samples were taken by Year, Month and Hour. The fourth Figure (Fig. A4) shows the spatial distribution of the estimation error by the GPR model for the held-out test dataset.

Table A1. Quality assurance steps when Ross et. al joined Landsat and the in-situ datasets (Ross et al., 2019).

Water Quality Data Download and Quality Control	
1	All observations were verified to have analytical methods related to parameter name; when this was not the case, samples were dropped.
2	Data across exchangeable units were harmonized (e.g. DOC in milligrams per liter)
3	Were depth information was given, all observations deeper than 100 m were removed (<1% of data). For the samples with recorded depth, more than 88% of data were sampled within 2 m of the surface of the waterbody, suggesting most samples are near surface. Given this high proportion of surface samples in the WQP data, the decision was made to keep all data with no depth information, assuming the vast majority of it was collected near the surface.
4	LAGOS-NE and WQP data were verified to have only one observation per site at a particular date and/or time. Some observations include date without timestamp; samples with one observation per date if only date information was available and one per datetime if timestamps were recorded were kept. Where the date, time, and observation value were the same for multiple observations, duplicates were converted to a single value. When the site and date or datetime were the same, but the parameter values were different, multiple observations were averaged to a single observation if the coefficient of variation (standard deviation/mean) was less than 10% and removed observations with too many simultaneous observations (five per date time combination) or too much variation with no metadata explaining the repeat observations.
5	In situ data was filtered to include only data with environmentally possible values by exploring the large data set itself and looking into the literature for reasonable values (for DOC thresholds are > 0.01 and <500 mg/L). All such sites were kept in the data set and were spatially joined to an inventory of Landsat WRS-2 paths and rows. Each site was related to its corresponding Landsat tile with sizes of about 5,000 × 5,000 pixels
Joining Landsat and In Situ Data Sets	

1	WQP and LAGOS-NE sites were preserved if they were within 200 m of at least 1 pixel with water occurrence of 80% in the Surface Water Occurrence dataset (Pekel et al., 2016)
2	Within a 200-m buffered zone any pixel not classified as water at least 80% of the time in the Landsat archive were removed
3	Landsat quality assessment bands were used to remove all pixels classified as cloud and cloud shadows, but all data classified as land, snow/ice, or water were kept, since high sediment concentrations can lead to classification as land or ice
4	Because many of the samples in the WQP are taken from or near bridges narrower than 30 m, 30-m buffer around the TIGER road data set from the U.S. Census office was created to exclude mixed water/road pixels
5	A spatial median of reflectance in each band from all remaining pixels in the 200-m buffer zone was calculated. The total pixel count and the standard deviation of reflectance values was gathered to assess uncertainty and variance. Spatial medians include a median of the quality assessment band and standard deviation of the bands.

Table A2. Tuned Hyperparameter values after random search. Note that if hyperparameters are not included the default setting in the Pyhton scikit learn package has been used.

	Hyperparameter	**Value (Landsat bands)**	**Value (Landsat bands + Environmental predictors)**
RFR	Number of trees	1354	2612
	Maximum features	2	9
	Maximum tree depth	89	538
GPR	Kernel	Rational Quadratic + White Kernel	Rational Quadratic + White Kernel
	alpha	0.00751	0.0225
	length scale	0.031	1.99
	noise level	0.00121	0.00165
MBPNN	Neurons	564	609
	Hidden Layers	3	6
	Learning rate	1.78e-5	6.22e-4
SVR	Kernel	Radial basis function	Radial basis function
	C	6	6
	epsilon	0.1	0.1
	gamma	1 / (n features * X.var())	1 / (n features * X.var())

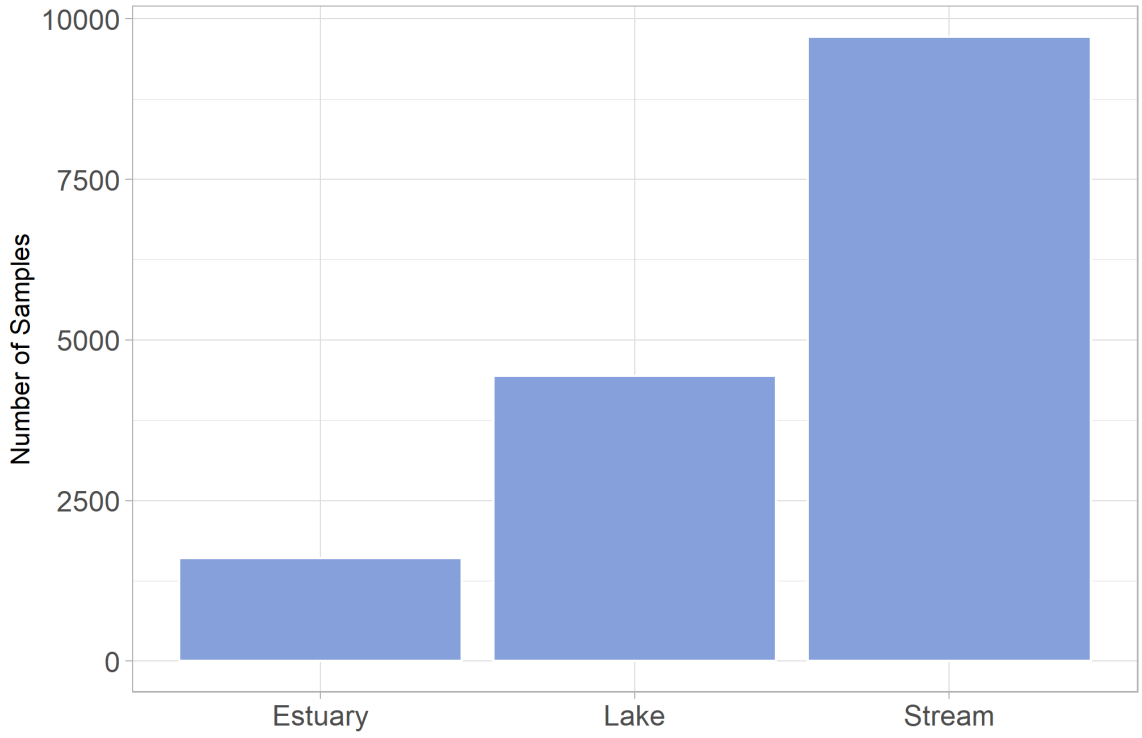

Fig. A1. Number of samples per inland water type in the preprocessed dataset.

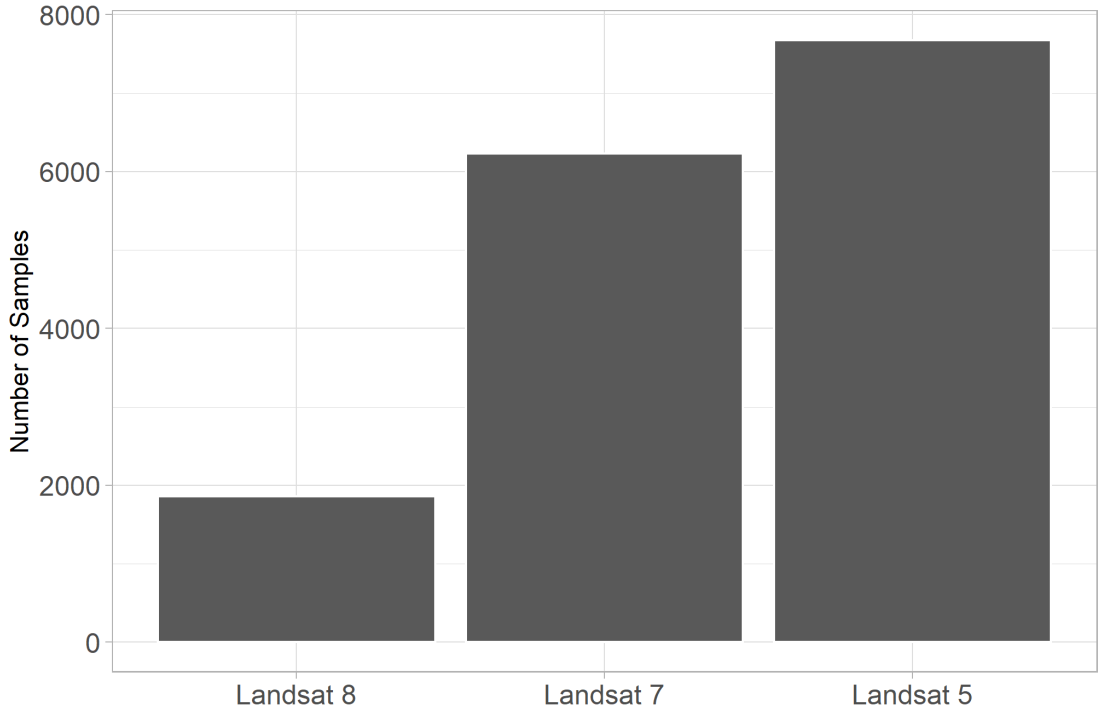

Fig. A2. Number of samples per Landsat sensor type in the preprocessed dataset.

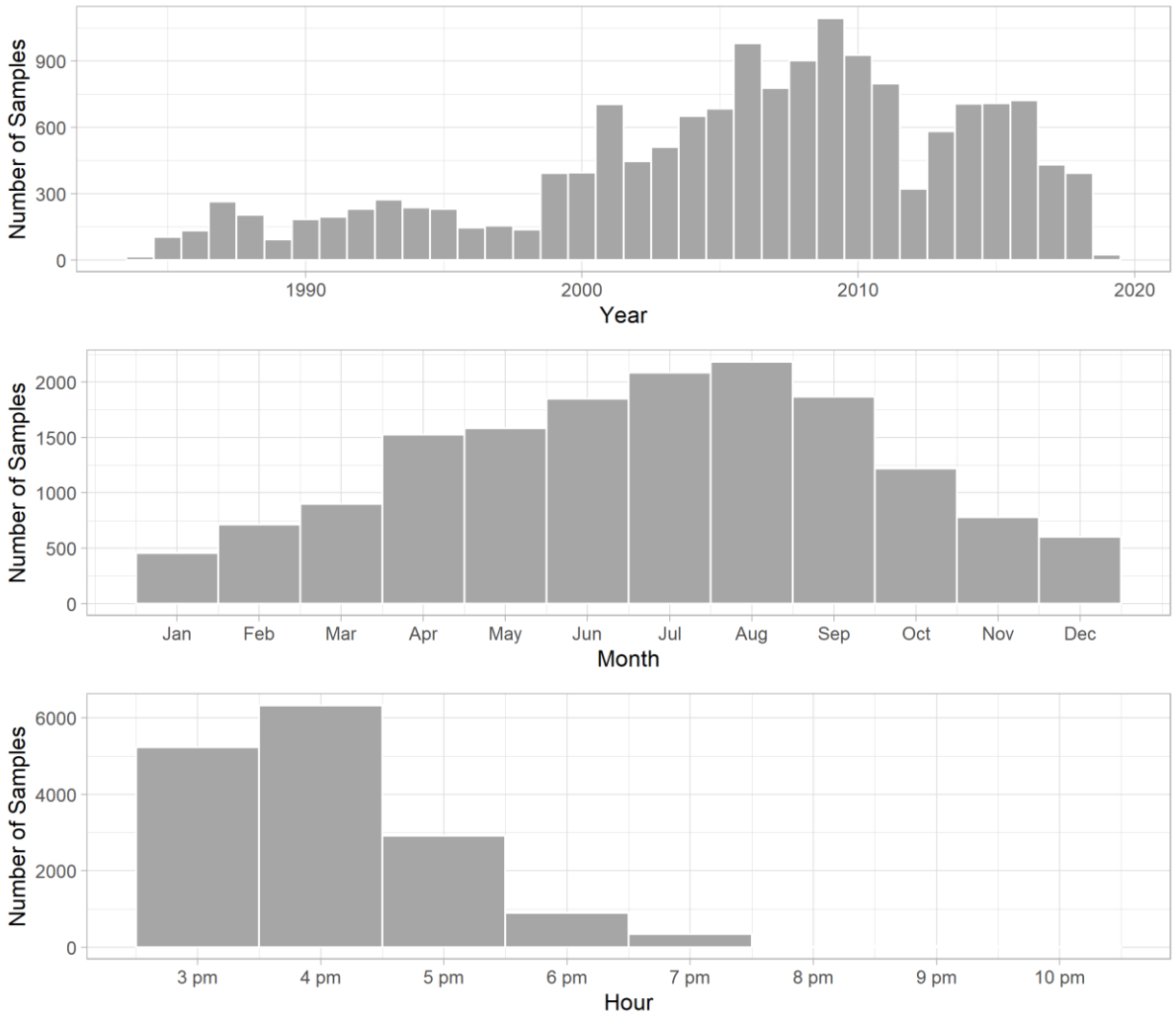

Fig. A3. Temporal distribution of the time the samples were taken by Year, Month and Hour in the preprocessed dataset.

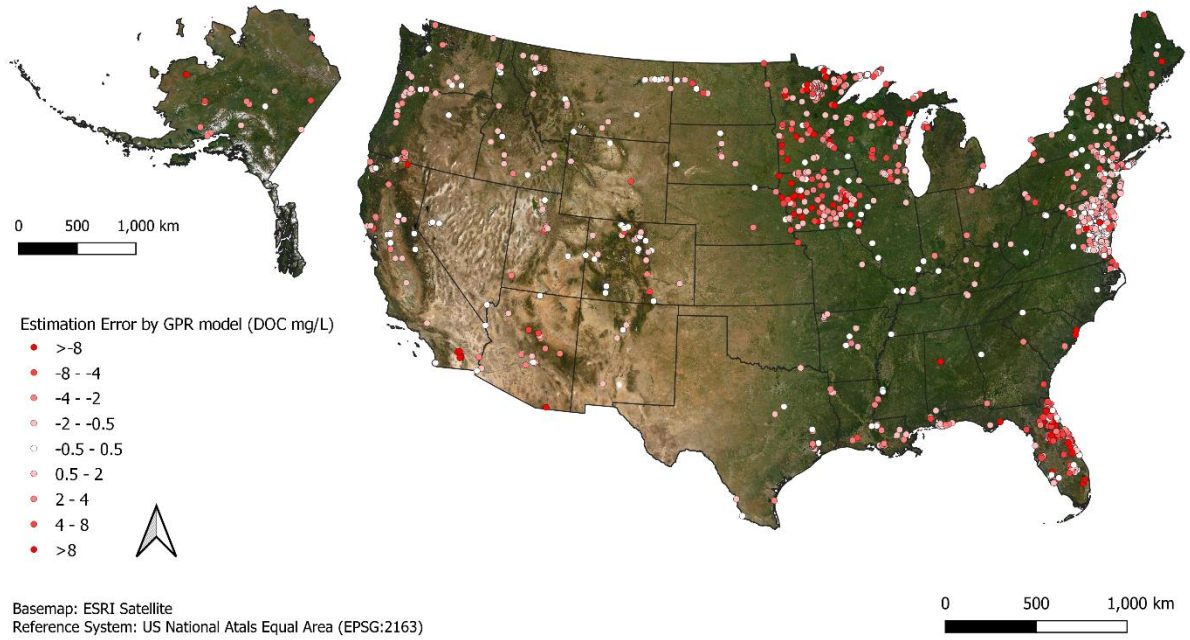

Fig. A4. Spatial distribution of the estimation error by the GPR model for the held out test dataset (N=4809).

Reference:

Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M., & Pavelsky, T. M. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research, 55*(11), 10012-10025. https://doi.org/https://doi.org/10.1029/2019WR024883