Towards a Generalized Machine Learning Approach for Estimating Chlorophyll Values in Inland Waters with Spectral Data

---

Philipp M. Maier

Karlsruhe, 2021

Towards a Generalized Machine Learning Approach for Estimating Chlorophyll Values in Inland Waters with Spectral Data

Zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS (Dr.-Ing.)

von der Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Philipp M. Maier

geboren in Göppingen

Tag der mündlichen Prüfung 28.10.2021

Referent Prof. Dr.-Ing. habil. Stefan Hinz
Institut für Photogrammetrie und Fernerkundung (IPF)
Karlsruher Institut für Technologie (KIT)

Korreferent Prof. Dr. Stefan Norra
Institut für Angewandte Geowissenschaften
Karlsruher Institut für Technologie (KIT)

Karlsruhe (2021)

Philipp M. Maier

Towards a Generalized Machine Learning Approach for Estimating Chlorophyll Values in Inland Waters with Spectral Data

Doctoral Thesis
Date of examination: 28.10.2021
Reviewers: Prof. Dr.-Ing. habil. Stefan Hinz and Prof. Dr. Stefan Norra

Karlsruhe Institute of Technology (KIT)
Department of Civil Engineering, Geo and Environmental Sciences
Institute of Photogrammetry and Remote Sensing (IPF)
Kaiserstr. 12
76131 Karlsruhe

Abstract

Water is an essential element of life. However, its quality is threatened for example by harmful algal blooms or man-made pollution. Monitoring enables the detection of changes in inland water quality. Conventional monitoring of water quality is mainly conducted with in situ sampling, an expensive and labor-intensive technique. Spectral remote sensing can be an alternative to in situ monitoring. The visible and near-infrared radiation recorded by a sensor has interacted with the water body and its constituents. Hence, the radiation contains information related to absorption and scattering processes in the water column. One parameter that strongly interacts with the light is the herbal pigment chlorophyll *a*. Since chlorophyll *a* is a proxy for phytoplankton abundance, it can be related to water quality.

One challenge in retrieving chlorophyll *a* based on spectral information concerns the spectral overlapping with other water parameters in the water column. Therefore, a reliable modeling approach is needed that can solve the non-linear regression task to retrieve continuous chlorophyll *a* values from spectral data. An additional requirement for such a model is the applicability to most of the global inland water bodies since the lack of reference data does not allow building specialized models for every single lake. This generalization requirement perfectly matches supervised machine learning approaches.

Hence, the main investigation of this thesis is to train and evaluate supervised machine learning approaches that can estimate continuous chlorophyll *a* values of multiple water bodies. Therefore, the examined studies rely entirely on spectral in situ measurements. This setup allows a more detailed analysis of the relations between spectral data and water parameters. Besides, the influence of the atmosphere is reduced. Three different datasets have been collected in the scope of this thesis to investigate the generalization process. The variability of the datasets increases consecutively. Therefore, three study setups were designed for these datasets, which consecutively increase the models' challenge to generalize. In the first setup, only models relying on a single water body are investigated. In contrast, the last setup relies on an entirely simulated dataset for the models' training process, whereas their evaluation is conducted on a completely unknown dataset containing eleven different inland water bodies. The idea of this concept is if models can estimate the chlorophyll *a* values of the completely independent eleven water bodies, they will likely perform well on most of the global water bodies with similar conditions. As a result of the conducted study setups, an One-dimensional CNN as a deep learning model succeed the task of the final setup and proves itself as a generalized model with suitable performance.

Further attention is given to the spectral resolution. A decrease in the spectral resolution from hyperspectral to multispectral is accompanied by a loss of information. The estimation results from the One-dimensional CNN reveal that hyperspectral resolution is necessary for an entirely generalized model. However, multispectral resolution is sufficient for less generalized models. These findings are important considering an upscaling approach to actual satellite data to fulfill the monitoring approach area-wide in future research.

Zusammenfassung

Wasser ist ein wesentliches Element des Lebens. Seine Qualität ist jedoch bedroht, zum Beispiel durch schädliche Algenblüten oder anthropogene Verschmutzungen. Regelmäßige Kontrollen ermöglichen das Erkennen von Veränderungen der Wasserqualität von Binnengewässern. Konventionelle Wasserqualitätskontrollen werden hauptsächlich mittels In-situ-Probenahmen durchgeführt, eine teure und arbeitsintensive Vorgehensweise. Spektrale Fernerkundung kann eine Alternative zu In-situ-Beprobungen sein. Die sichtbare und nahinfrarote Strahlung, die von einem Sensor aufgenommen wird, hat mit dem Wasserkörper und dessen Inhaltsstoffen interagiert. Dadurch enthält die Strahlung Informationen über Absorptions- und Streuprozesse in der Wassersäule. Ein Parameter, der stark mit der Strahlung wechselwirkt, ist das pflanzliche Pigment Chlorophyll *a*. Chlorophyll *a* ist ein Proxy für die Phytoplanktonabundanz und kann daher mit der Wasserqualität in Verbindung gebracht werden.

Die spektrale Überlappung mit anderen Wasserinhaltsstoffen erschwert die Bestimmung des Chlorophyll *a*-Gehalts mit spektralen Daten in der Wassersäule. Daher ist ein zuverlässiger Modellierungsansatz erforderlich, um diese nicht-lineare Regressionsaufgabe zu lösen und damit kontinuierliche Chlorophyll *a*-Werte aus Spektraldaten zu gewinnen. Eine zusätzliche Anforderung an einen solchen Ansatz ist die Anwendbarkeit auf die meisten der weltweiten Binnengewässer, da der Mangel an Referenzdaten nicht erlaubt, spezialisierte Modelle für jeden einzelnen See zu generieren. Diese Generalisierungsanforderung passt perfekt zu Ansätzen des überwachten maschinellen Lernens.

Ein Hauptziel dieser Arbeit ist daher das Trainieren und Evaluieren von überwachten maschinellen Lernverfahren zum Schätzen kontinuierlicher Chlorophyll *a*-Werte von mehreren Binnengewässern. Die untersuchten Studien stützen sich dabei vollständig auf spektrale In-situ-Messungen. Dieser Aufbau erlaubt eine detailliertere Analyse der Beziehungen zwischen spektralen Daten und Wasserparametern. Außerdem wird der Einfluss der Atmosphäre verringert. Drei verschiedene Datensätze wurden im Rahmen dieser Arbeit aufgenommen, um den Generalisierungsprozess der generierten Modelle zu untersuchen. Die Variabilität der Datensätze nimmt dabei sukzessive zu. Daher wurden für diese Datensätze drei Studienkonfigurationen entworfen, die sukzessive die Anforderung zur Generalisierung der Modelle erhöhen. In der ersten Konfiguration werden lediglich Modelle untersucht, die sich auf ein einzelnes Gewässer beziehen. Im Gegensatz dazu stützt sich die letzte Konfiguration auf einen vollständig simulierten Datensatz für den Trainingsprozess der Modelle, während deren Evaluierung auf einem völlig unabhängigen Datensatz mit elf verschiedenen Binnengewässern erfolgt. Die Idee hinter diesem Konzept ist, wenn die Modelle die Chlorophyll *a*-Werte

der elf völlig unbekannten Binnengewässer schätzen können, werden sie vermutlich auch weltweit, die Werte ähnlicher Binnengewässer schätzen können. Ein eindimensionales CNN als Vertreter der Deep-Learning-Verfahren hat sich dabei als das Modell mit den besten Generaliserungseigenschaften bei zufriedenstellender Schätzgenauigkeit erwiesen.

Ein weiteres Augenmerk wird auf die spektrale Auflösung gelegt. Eine Verringerung der spektralen Auflösung von hyperspektral auf multispektral ist mit einem Informationsverlust verbunden. Die Schätzungsergebnisse aus dem eindimensionalen CNN zeigen, dass eine hyperspektrale Auflösung für ein vollständig generalisierendes Modell notwendig ist. Eine multispektrale Auflösung ist jedoch ausreichend für weniger generalisierende Modelle. Diese Erkenntnisse sind wichtig um im Hinblick auf ein zukünftiges Forschungsvorhaben den Upscaling-Ansatz auf reale Satellitendaten zu realisieren und damit eine flächendeckende Überwachung der Wasserqualität zu verwirklichen.

Contents

1 Introduction

1.1 Motivation

Water is a vital component of life. It is crucial for humans but also for ecosystems. Its importance is therefore stipulated in two out of the 17 Sustainable Development Goals (SDG) of the United Nations (UN) [1]. These two goals concern the provisioning of drinking water (6) and the protection of marine ecosystems (14). The first affects humans, whereas the latter concerns environmental systems. The conservation of inland waters is highlighted by the European Water Framework Directive (WFD) of the European Union (EU) [2] and the American counterpart, the US Clean Water Act (CWA) [3]. Both guidelines aim for a good status of inland waters concerning chemical, structural and ecological manner. To recognize such changes, consequent monitoring of water bodies and their quality is necessary.

For consequent monitoring, parameters that can be related to the water body's quality need to be frequently available. Chlorophyll *a* is such a parameter. It is a herbal pigment that appears in phototrophic organisms such as phytoplankton. Hence, its concentration is a proxy for the primary production in a water body, the basis of the food web, and thus, crucial for the ecosystem. Additionally, the chlorophyll *a* concentration correlates with the nutrition and the oxygen supply [4, 5, 6]. Besides seasonal variations of the phytoplankton occurrence, long-term trends can exist, forced by human impacts. Phytoplankton growth is boosted by two human-derived nutrients; phosphorus and nitrogen that can lead to eutrophication of a water body [7]. The first is mainly the limiting factor in freshwater, whereas the latter is limiting in ocean water [8]. Phosphorous input can originate, e.g., from sewage and animal wastes, atmospheric deposition, agricultural fertilizer runoff, or groundwater inflow [9]. During the photosynthesis process, phytoplankton produces oxygen. However, there is a demand for oxygen in the decomposition phase. In combination with high water temperatures and their related low oxygen solubility, mass death of fish can appear [10, 9]. Hence, chlorophyll *a* is a crucial parameter for understanding and evaluating changes in water ecosystems [11].

Diatoms are highly sensitive to environmental conditions, respond quickly to chemical, biological, and physical changes, and grow rapidly [12]. Therefore, the class of diatoms are bioindicators and directly related to water quality in terms of the Trophic Diatom Index [13]. However, not only do humans affect the water quality by nutrient enrichment, the other way around is also valid. Algae abundance is relevant for drinking water quality. In water bodies, different phytoplankton species can occur, varying in their proportions and over seasons [9].

Some of these species, especially the ones belonging to the class cyanobacteria, can cause harmful effects on human and animal life. Harmful algal blooms of cyanobacteria in drinking water can lead to intoxications of humans and animals and even further to death [14]. Hence, monitoring is essential not only for the environment but also for human health.

Conventional monitoring of water bodies is conducted either by water samples evaluated in the laboratory or by in situ sampling with a probe device. Both techniques have similar disadvantages. The measurements are based on a point sample taken as representative for the whole water body. For small water bodies, this generalization is acceptable, but for larger water bodies, this is hardly possible. Wind, for example, influences phytoplankton aggregation, which can lead to strong over- or underestimations [15, 16]. Imagine a river flowing into a lake and bringing nutrients that are accumulated. This accumulation leads to algae growth, especially in that area, and a single point sample would not represent the whole water body, irrespective of where it is taken from. Another disadvantage of current monitoring approaches is the cost. The in situ monitoring is labor-intensive, so the number of controls during a year is often deficient [15, 16]. Since phytoplankton grows exponentially, especially in spring, it is crucial to have a high temporal resolution of the sampling if measures by the local authorities are necessary.

Remote sensing is a technique that can support conventional measurements (see, for example, [17, 16]). Working with current satellite data allows consequent environmental monitoring and is an opportunity to fill the gap in between infrequent in situ measurements. The satellites cover a huge area with a single image and return in a certain amount of days on the same spot at the same daytime. So, extracting information about the condition of a water body from satellite images would provide a vast benefit for local authorities.

To retrieve information from a satellite image, a model must link the spectral information to continuous values or discrete quality classes of a parameter. Chlorophyll *a* is one parameter with characteristic optical properties for the estimation of water quality with spectral data. This is due to the pigment's specific absorption features and scattering. These features can be extracted by models and related to the chlorophyll *a* concentration. Many other pigments or particles show such characteristic features. Unfortunately, most of these features overlap with others, or their concentration is low, so the impact on the spectrum is negligible. In total, the entire spectral signature of a water body is a mixture composition between different pigments, molecules, and other parameters. Hence it is a challenging task for a model to extract a reliable chlorophyll *a* value. Such spectral characteristics do not exist for nutrients (e.g., phosphorus, nitrogen) or heavy metals (e.g., lead and cadmium) in the visible spectrum. Thus, these chemical parameters are not retrievable with spectral remote sensing techniques.

Currently, empirical band ratio (BR) models combined with a linear regression are the most common approaches to estimate the chlorophyll *a* concentration with spectral data. The idea of a band ratio model is to select influential bands for the respective water parameter,

e.g., a feature influenced by pigment absorption and another one without absorption. Then, its ratio is the input for a regression model, which is able to estimate the parameter's concentration. Such models work remarkably well and noise resistant on satellite images of individual water bodies. A disadvantage of such a model is its specialization on one particular water body. Hence, for a monitoring approach, such a model needs to be calibrated with reference data for every single water body. Unfortunately, reference data is sparse for most of the 117 million lakes globally with a surface area greater than $0.002\,\mathrm{km}^2$ [18]. On the contrary, a generalized model would be able to estimate the water parameter values with a suitable performance independent of the underlying water body. Hower, such models rarely exist or show poor performance.

For a generalized approach, a model is necessary that can derive the relevant information from the spectrum consisting of overlying spectral signatures. Supervised machine learning (ML) models, such as neural networks may present approaches to meet this challenge. These models take the whole spectrum in the visible range as input data to extract information about the water ingredients and to derive its chlorophyll *a* concentration. However, for such an approach, a vast amount of data from different water bodies is needed. In addition, ML models are often considered as a black box.

One disadvantage of estimating water parameters with satellite data is the trade-off between spectral and spatial resolution. As a technical limitation, increasing spatial resolution leads to decreased spectral resolution. This trade-off means that a reduced spatial resolution leads to a decrease in the number of water bodies that can be monitored due to their low surface area. However, a decrease in spectral resolution is followed by a loss of information. This decrease in spectral resolution impedes especially the small-scale inland water monitoring. Another constraint is the atmosphere that strongly influences the retrieved satellite image. One additional limitation in building retrieval models is the relation between the water reference samples and the satellite image. The experience of measuring in the field shows that, e.g., the chlorophyll *a* concentration of water samples varies about several percent in short time intervals or several meters. Thus, relating the values to the exact time of a satellite image or to pixel sizes of, e.g., 300 m (Sentinel-3) leads to unwanted miscalculations. Therefore, building models based on in situ spectrometer data may be a valid option to investigate the potential of supervised learning models.

1.2 Main Objective and Research Goals

This thesis aims to build and evaluate ML models for estimating quality parameters of inland waters, especially chlorophyll *a*, with spectral data. Therefore, several data have been collected on various water bodies. It is important to highlight that the thesis is completely built on in situ measured spectral data. When working on actual satellite data,

the influence of the atmosphere on the spectral signal offers much more degrees of freedom than working only on in situ remote sensing data. Hence, the demand for data would increase significantly, which would go far beyond the scope of this thesis. One advantage of in situ data is a finer coupling between spectral data and water parameters, that allows a more exact model training. Besides, the spectrometer data can be downsampled to various satellite resolutions, to compare their impact on the models on the same dataset. A later modular upscaling approach may allow closing the loop to the monitoring approach with real satellite data. Concluding the challenges in remote sensing of inland water bodies, several Research Goal (RG) are identified. Thus, the research structure of this thesis is set into five RG with Research Question (RQ), presented in the following:

RG 1: Models and Parameters The quantitative retrieval of inland water parameters, mainly chlorophyll *a* with spectral data, is an ill-posed problem [19] due to the overlay of the spectral signature of several water constituents. Hence, it is a non-linear regression challenge. Contrary to most of the literature, the idea of this thesis is to face the problem solely with data-driven supervised ML models. This RG is determined as **RG Models and Parameters**. Thus, one main objective of this thesis is framed by a fundamental double **RQ 1**:

- **Can supervised ML models provide a suitable estimation performance of water parameters, especially chlorophyll *a*?**
- **Which of the applied ML models is the most promising to retrieve water parameters with spectral data?**

RG 2: Feature Importance The second RG concerns the accusation of ML models to be a black box. Thus, the idea is to regard the variable importance ranking of models and compare the most important ones with the spectral features of the parameter. It is determined as **RG Feature Importance**. Thus, **RQ 2** means:

- **Do the applied ML models rely on similar features as the empirical retrieval algorithms in the literature?**

RG 3: Generalization The third objective of this thesis concerns the generalization regarding several water bodies. It is determined as **RG Generalization**. Most of the studies in the literature show that their models are working fine for the dataset they were built on. However, their generalization on unknown datasets is poor [17], which means they need to be adapted to local conditions. Therefore, **RQ 3** is:

- **Can the applied ML models generalize on multiple water bodies? Is it even possible to build a model that is able to estimate the chlorophyll *a* concentration of completely unknown water bodies?**

RG 4: Spectral Resolution The background of the fourth RG is a later application on satellite data, where both the spectral and the spatial resolution are limited. Hence, it is essential to solve the trade-off and fulfill the parameter estimation with a low as possible spectral resolution accompanied by a still acceptable performance. A decrease in spectral resolution results in less input features for the models. The RG is determined as **RG Spectral Resolution**. This leads to **RQ 4**:

- **How much can the spectral resolution decrease to still get a suitable estimation performance by the models?**

RG 5: Transferability The final RG focuses on the limitations of ML models. One disadvantage of data-driven ML models is the vast demand for training data, especially for Deep Learning (DL) techniques. However, the amount of data that can be collected during a specific period is limited. The RG is determined as **RG Transferability**. To solve the sparse data challenge, the fifth and final **RQ 5** is:

- **Can a model trained on simulated data be able to estimate the chlorophyll *a* concentrations of real-world water bodies?**
- **Is the transferability between simulated and measured data given?**

1.3 Study Design

To answer the posted **RQs**, the thesis relies on different **Study Setups** (see Figure 1.1) comprising three datasets, conducted within its scope. Each dataset consists of spectral data and simultaneously measured water parameters. Generally, in each study setup, various ML models are built to estimate the underlying water parameters with spectral data.

Study Setup I is a preliminary study relying on a dataset conducted on the River Elbe. The investigations concern **RQ 1** and **RQ 2** (see Figure 1.1). The study's design focuses on the general applicability of various ML models in estimating different water parameters (**RQ 1**). Besides, the accessibility of frequent measurements of various water parameters may allow conclusions about the important features selected by the applied models (**RQ 2**). Within the scope of this thesis, three studies have been published relying on these investigations on **Study Setup I** ([20, 21, 22]).

Study Setup II relies on the SpecWa dataset, consisting of spectral measurements combined with chlorophyll *a* concentrations of eleven different small-scale inland water bodies. One focus of this study is a higher degree of generalization since the models need to generalize on multiple water bodies (**RQ 1** + **RQ 3**). Additionally, the influence of different spectral resolutions on the model performance is investigated (**RQ 4**). Eventually, the investigated

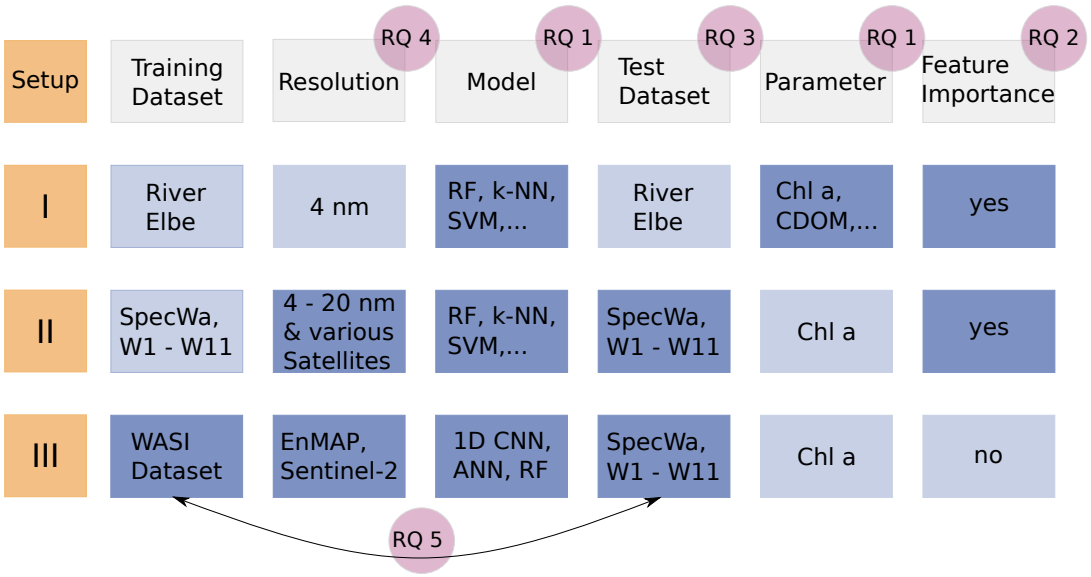

Figure 1.1: The investigations of the RQ is comprised in **Study Setups I-III** (in orange), focusing on different Datasets. The grey rectangles represent the characteristics of each **Study Setup** and can be related to **RQs**. The dark blue rectangles reflect the focus of the respective study and indicates an overlap with RQs. **RQ 5** represents the transferability between datasets, solely investigated in **Study Setup III**.

models allow conclusions on the highly-rated input features for specialized (selected lakes) and for generalized models (the whole dataset) (**RQ 2**). Most of these investigations are content of the published studies ([23, 24]). The dataset itself is published in [25].

The final **Study Setup III** relies on two datasets. First, an entirely simulated one, that is used to train the ML models. Second, the SpecWa dataset mentioned above is applied as the test dataset. The study focuses on **RQ 1** and **RQ 3-5**. Hereby, the best performing ML models from previous studies complemented with a 1D CNN, representing a DL model (**RQ 1**), are tested. The approach allows conclusions about the generalization (**RQ 3**) since the datasets are entirely independent. Then, the transferability between the simulated training and the real-world test dataset is examined (**RQ 5**). Besides, the impact of the spectral resolution on the estimation performance is investigated (**RQ 3**). The investigations of **Study Setup III** are published in [26].

1.4 Thesis Outline and Contributions

The thesis is organized into seven chapters. Chapter 2 describes the relevant parameters inside and outside a water body, influencing its spectral signal. It amplifies the physical background of remote sensing of water constituents. The section should clarify

the demand for answering the RQs posted in Section 1.2. The state of the art concerning water parameter retrieval algorithms is presented in Chapter 3. Only a few studies exist, which follow the supervised ML approach, whereas most of the literature focuses on empirical band arithmetic approaches.

For remote sensing of inland waters, benchmark datasets rarely exist. Thus, three datasets are collected within the scope of this thesis, as basis of the investigations summarized in the RQs presented in Section 1.2. The description of the datasets and their measuring setups are presented in Chapter 4. Chapter 5 provides the theoretical background of supervised ML approaches and briefly introduces several selected algorithms. Its second part is the applied methodology chapter. It contains data processing and the ML pipeline that is employed in the studies.

For the evaluations of the five research goals, three study setups exist, relying all on different datasets. The challenge for the models to retrieve the respective water parameters is increasing in each study due to a rising heterogeneity in each subsequent dataset. After each presented study results, a discussion concerning the affected RQs is conducted in Chapter 6. Finally, the discussions and the following conclusions focusing on each RQ are presented in Chapter 7. Subsequently, the main conclusions are made and an outlook to further research is given.

The thesis contributes to monitoring inland water quality with spectral remote sensing data. Modern DL techniques combined with a continuously growing amount of data and computational feasibility may significantly improve the performance of water parameter retrieval models. Generalized models are substantial, following the aspiration of monitoring most of the global inland waters. For such an approach, suitable satellite sensors need to deliver satisfying image quality, which is rarely given for inland waters. Therefore the investigation of the spectral resolution's impact on the retrieval model performance is crucial.

2 Physical Background

The spectral radiation coming from a water body to a remote sensing sensor is an interaction of multiple physical parameters. These physical parameters are essential to understand the complexity and the challenges of retrieving water parameter values from the spectral composition. Water ingredients, such as chlorophyll *a* show spectral absorption features. However, such features must be detectable by spectral bands to derive information about a water body's chlorophyll *a* concentration. As a result, there is a demand for a fine spectral resolution.

A sensor's spectral resolution determines how accurate the radiation coming to the sensor can be disaggregated in spectral channels. Of course, a finer spectral resolution may contain more relevant information. However, the increase in spectral resolution often results in a decrease in spatial resolution. This trade-off is relevant for a later monitoring application of water parameters on real satellite data.

The following Section 2.1 will inform about the physical relations between water constituents and the spectral signal conducted with a remote sensing sensor. Then, the focus is on the trade-off between a sensor's spectral and spatial resolution (see Section 2.2).

2.1 Light Behavior and Interactions in the Water Body

This section focuses on the composition of the spectral signature measured with a spectrometer such as the RoX, later described in Chapter 4, placed slightly above the water surface. It is to mention that the measured radiance reflectance above the water surface has only the intensity of about 1 % of the radiation that is measured before the interaction with the water body. The rest of the visible light is absorbed by the water and its ingredients. Different water compositions result in various spectral signals, which in turn may allow to deducing the concentration of its components.

Since the sensor signal measured above the water surface does contain not only the information of the water column but also artifacts from the atmosphere mirrored at the surface, and ground reflectance in shallow waters, it is crucial to understand the parameters influencing those processes, not only the ones in the water column. In general, there are several parameters in and above the water that influences its spectral signature. The most important ones are described in the following.

Figure 2.1 shows the most influential parameters for a natural water body's spectrum schematically. The scheme refers to a normalizing sensor, such as the RoX spectrometer. This means that the spectral signature is expressed as radiance reflectance, the ratio between the incoming irradiance and the reflected radiance. In the application on a water surface, unfortunately, the latter is set together from the water-leaving radiance, containing the information of the water column and the surface reflectance.

The wanted signal for most applications is the remote sensing reflectance, which is defined as the ratio between the downwelling irradiance and the water-leaving radiance [27]. Thus, to obtain remote sensing reflectance, the surface reflectance needs to be subtracted [28]. This can be fulfilled by different correction models [29] to get the comparability, e.g., between satellite images. For inland waters, specific correction models are necessary since in opposite to the land surface, a water surface reflection is more challenging due to its specular effect. The surface reflectance is an unwanted signal, whereas the water-leaving radiance carries the information for parameter estimation.

Nevertheless, in this thesis, the focus is on radiance reflectance due to three reasons. First, existing models for surface (and atmosphere) correction are complex and need additional measuring information, which is not given for the underlying datasets. Then, several models exist, but they do not work suitably for every water body [29, 30, 31]. Finally, from its magnitude, the surface reflectance can be much higher than the water-leaving radiance. Hence, miscalculations in pre-processing may lead to even higher errors in the final parameter estimations by error propagation. Since data-driven ML models are the main subject of this thesis, surface reflectance is an additional task to be solved within one model. The approach is then to train the model on varying surface reflectance conditions.

For the following explanation, the interaction areas of the visible light with a water body are divided into three levels. First, the interactions that occur above the water, including the water surface reflectance. Then, the interactions in the water column and, finally, the interactions with the ground.

2.1.1 Water Surface and Above

When evaluating satellite images, atmospheric correction is significant to get comparability between the different images. The light comes from the sun passing the whole atmosphere to the water body and back through the atmosphere to the sensor on the satellite. On this path, many interactions are possible that do not concern the water body itself but show a vast influence on the spectral signature. Here comes the advantage when measuring with a spectrometer, such as the RoX. The RoX measures not only the reflectance that comes from the water, but also the irradiance that comes directly from the sun and diffusely from the sky to the upwards-looking sensor. Its ratio results in a normalized value that

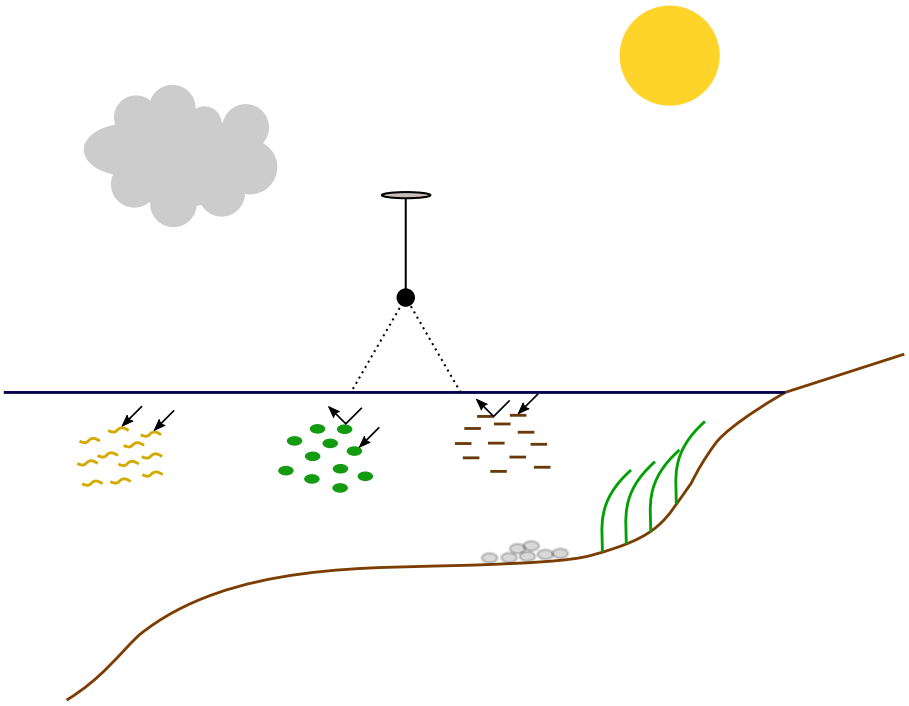

Figure 2.1: Composition of the reflected radiance to a sensor measuring above the water level. The illustration refers to a normalizing sensor, such as the RoX. The measured signal is composed by the water-leaving radiance and the surface reflectance. The latter depends on the condition of the atmosphere and the altitude of the sun. The water-leaving radiance depends mainly on absorption and scattering processes of water ingredients, represented by three groups. First, absorption on CDOM (yellow), then absorption and scattering on phytoplankton (green) and finally, absorption and scattering on NAP (brown). Additionally, for shallow waters, the benthic substrate's signal is added.

allows a comparison between any measured spectrum. Only in the approximate one meter between the two sensors of the RoX occur unsupervised interactions. For sensors carried by planes or UAVs, this distance is much larger. Hence, different sun angles and slight cloud occurrence still lead to stable measurements of the RoX sensor and their effects on the spectrum are minimized. However, there is also the effect of surface reflectance that overlays the water-leaving radiance of a water body. The surface reflectance is related to sun glint and sky glint and influenced by, e.g., the zenith angle of the sun as well as scattering in the atmosphere [31]. The latter consists of two components, an aerosol scattering component and a Rayleigh scattering component [28]. In general, there is a trend: the more overcast a sky, the higher the surface reflectance at a radiance sensor. Then, it is also possible that the surface reflectance signal exceeds the water-leaving radiance [31]. The surface reflectance signal of the overcast sky is nearly wavelength-independent, but in turn, it is wavelength-dependent on clear-sky conditions [32, 31]. This means it does not result in a simple y-offset in the spectrum. To sum up, sun altitude and the condition of the atmosphere mainly determine the spectral signal that is reflected from the water surface. The surface

reflectance does not carry information concerning the water ingredients, but its magnitude needs to be understood to determine the signal from the water column.

2.1.2 Water Column

In the water column itself, the water-leaving radiance, and hence the spectrum's shape, is determined. It carries the information to derive water parameters. The attenuation describes the decrease of natural illumination in a water column interval dependent on the solar zenith angle [28]. It is driven by backscattering and absorption processes in the water column. The more particles in the water that absorb light, the less radiation is measured in the radiance sensor, whereas particles that scatter visible light enforce the signal. These spectral signals are wavelengths dependent.

Morel and Prieur [33] distinguished the global water bodies in CASE I and CASE II waters regarding the optical complexity. The CASE I waters comprise the seas and the oceans. Their color is dominated by phytoplankton occurrence with a rare impact of Colored Dissolved Organic Matter (CDOM) and non-algal particles (NAP) [34]. The CASE II waters encompass the coastal waters and the inland waters. Contrary to CASE 1 waters, CASE 2 waters are optically more complex since they are affected by terrestrial substances, like CDOM or NAP, especially inorganic matter. Both groups are optically active, and hence their absorption and scattering processes overlap each other.

Absorption appears on different water content materials, such as CDOM, chlorophyll, or other pigments occurring in different phytoplankton classes, and water itself. In modeling, the various absorption coefficients can be summarized as a single absorption coefficient. The higher the absorption, the more the attenuation in the water.

Scattering of light in water occurs on particles such as clay minerals but also on organic particles such as phytoplankton [35]. In modeling, same as for the absorption, the underlying backscattering coefficients of a water body can be summarized to a single one. The higher the backscattering coefficient, the more light is scattered back to the water surface, resulting in a higher magnitude of the water spectrum. More detailed information about absorption and scattering behavior on different materials is given in the following paragraphs.

Chlorophyll and Phytoplankton

The spectral signature of phytoplankton is related to its pigments. The most prominent pigment is chlorophyll *a*, which appears in every phytoplankton species. There are other chlorophyll derivatives, but their concentration is lower than the *a* modification, and their absorption features are overlying with other substances, so they are hardly recognizable in

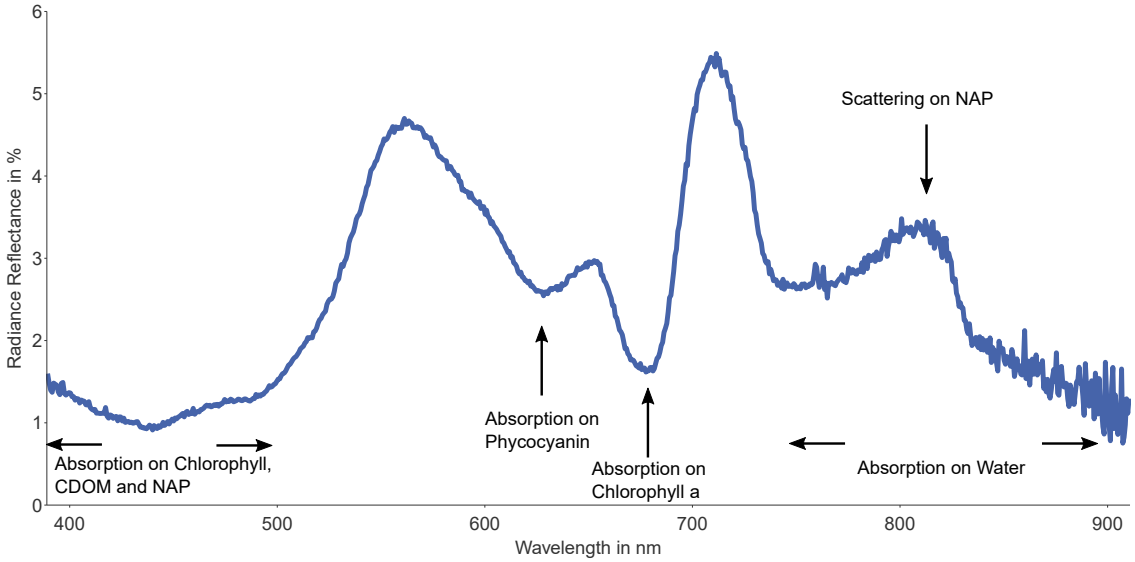

Figure 2.2: An exemplarily measured spectrum of the SpecWa dataset. The spectral signature is influenced by high chlorophyll *a* concentration of $190\,\mu g\,L^{-1}$ whereof $50\,\mu g\,L^{-1}$ is related to cyanobacteria. Hence two distinct absorption minima occur at 620 nm related to phycocyanin and at 670 nm related to chlorophyll *a*. Due to high turbidity which is influenced by living and degraded phytoplankton (detritus), a scattering peak at 810 nm occurs.

the spectral signal by remote sensing. Besides chlorophyll, phytoplankton species consist of various other pigments, such as carotenes, xanthophylls, and phycobiliproteins [36]. Since phytoplankton species, such as green algae, diatoms, or cyanobacteria, consist of different pigments, a distinction between the species may be allowed by different spectral signals.

Chlorophyll *a* shows two distinct absorption features, which result in a minimum reflectance in the spectral signature (see Figure 2.2). The most prominent one is placed at about 675 nm [37, 33, 38]. One advantage of this feature is its distinctness since only a little overlap with other substances occurs at these wavelengths. Hence, this feature is mostly used to derive chlorophyll *a* concentrations of inland water by remote sensing. A second absorption feature lies between 400 nm to 500 nm. Unfortunately, overlapping occurs with strong absorption on CDOM [38], which hinders deviating chlorophyll models on this feature. At around 560 nm, a peak in the spectrum occurs related to the low absorption of algae material. As a result, the global reflectance maximum of a water spectrum is placed on this wavelengths area [38].

Another characteristic of the reflectance spectrum of natural water bodies is their reflectance peak with a maximum at 685 nm to 715 nm [38]. This peak can be related to the low absorption of chlorophyll and water [39, 40, 41]. With wavelengths longer than 700 nm absorption by water increases strongly [42, 43]. Others additionally relate the peak to scattering on phytoplankton. It shifts towards longer wavelengths with higher chlorophyll concentrations [38].

In general, backscattering on phytoplankton occurs, but its wavelength-dependent curve is similar to the one by non-algal particles. According to Gege [28], most frequently,

no distinction is made between both particles for backscattering since scattering on these particles depends more on the size and the distribution than on their chemical composition [28]. The shorter the wavelengths, the higher the backscattering.

A distinction between phytoplankton species is a difficult task since their pigment features often overlap. In the literature, the focus on species distinction is mostly on cyanobacteria due to their toxicity. Then, in phytoplankton mixture communities, the cyanobacteria's concentration can not be extracted by the prominent chlorophyll feature [44]. Fortunately, this species is the most feasible to distinguish because of the absorption feature on its unique pigment phycocyanin at 615 nm that is visible as a local minimum in the spectral signature [45, 44] (see Figure 2.2).

To compare green algae with diatoms, the first contains more carotenes than the latter, while both have a significant amount of various xanthophylls [36]. A study with different mixtures of green algae and diatoms have shown, that the global reflectance peak at around 560 nm shifts towards 520 nm the more green algae is apparent. The other way around, the peak shifts towards 570 nm [36]. However, the nutrition supply and the physiological condition of phytoplankton may influence the pigment composition in natural waters [36]. Unfortunately, high concentrations of CDOM disable the differentiation between pigments in wavelengths shorter than 600 nm due to their high absorption [36].

Fluorescence is an effect that allows algae to release light at the characteristic wavelength of 685 nm. Since the fluorescence peak lies between the minimum and maximum of the absorption (see Figure 2.2), it is hardly recognizable with remote sensing techniques. Thus, fluorescence is not assessed in this thesis.

CDOM

CDOM is a group of water-soluble organic substances that can originate in an autochthonous or allochthonous way [46]. Autochthonous CDOM mainly consists of degraded algae material, whereas allochthonous CDOM comes from outside the water body, such as humic plant- or soil materials. Hence, CDOM is a parameter that varies slightly in its properties [47].

CDOM is not only determined of environmental concerns but also important to understand the spectral behavior of a water body. It mainly absorbs light, whereas its scattering is negligible. Its absorption follows the principle that the shorter the wavelength, the higher the impact. Hence, there is an overlap in the spectral signature between phytoplankton pigments and CDOM, especially for higher concentrations of the latter [36]. This fact often results in a poor estimation performance of models in water bodies with high concentrations of fulvic and humic substances [48].

The CDOM concentration's influence on the spectral signature is more simple than the one of chlorophyll *a*. It can be expressed by exponential functions [49, 50], which are validated for a vast variety of natural water bodies [50, 51]. Since only absorption occurs on CDOM, but no scattering, the reflectance of water bodies with a high CDOM concentration is very low compared to others [52]. As a consequence, CDOM-rich water bodies containing a low amount of backscattering substances are difficult to investigate with remote sensing since atmospheric effects overtop the signal from the water body[53, 48].

Non-Algal Particles

Besides CDOM and phytoplankton, a third group of parameters exists relevant to the light behavior in the water column of a natural water body, the NAP [28]. This group includes suspended materials but excludes living algae particles, whereas both together are called Total Suspended Matter (TSM) or Total Suspended Solids (TSS). Contrary to TSS, the term turbidity means the proportion between light absorption and scattering [54]. Another related parameter is the Secchi Depth (SD), which is often applied as a remote sensing target parameter.

NAP can have an organic or inorganic origin. The part of degraded algae material that does not belong to CDOM belongs to the organic NAP (other denomination: tripton, detritus, bleached particles). Suspended Particular Inorganic Matter (SPIM) also belongs to the NAP. They include, e.g., soil materials, such as iron oxides or clay minerals, and are hence related to the geogenic origin.

NAP influence the spectrum of a natural water body due to scattering and absorption processes. Water bodies with high concentrations of NAP show higher reflectance due to scattering effects on the particles [55]. The higher the concentration of suspended materials, the higher the reflectance, especially in the longer wavelengths of visible light [55]. A peak occurring at about 810 nm can be related to tripton [56] (see Figure 2.2). Backscattering on NAP is more or less constant over the visible spectrum but tends to have more impact on the shorter wavelengths [28]. In general, it is similar to backscattering on phytoplankton [28]. This sounds contradictory to [55], but with increasing NAP concentration, the absorption increases as well, which is then again stronger on the shorter wavelengths [56]. The absorption curve of NAP can be approximated with exponential function similar to CDOM [57, 28]. However, absorption on NAP varies strongly due to its variable constitution and depending on the geogenic background [57]. Especially, iron oxides have an impact [57].

2.1.3 Bottom Reflectance

The attenuation of the water column in combination with the depth of a water body determines whether the bottom affects the spectrum or not. If yes, the reflected radiance

is attenuated again through the water column and adds its signal to the sensor. Different benthic substrates, such as sand, silt, clay, or water plants, possess different reflection spectra. In real-world conditions, it is often a mixture between various benthic substrates. The overlay with spectral absorption features and scattering on particles with bottom reflectance leads to an even more mixed spectral signal, which is again more difficult to interpret for parameter retrieval models. This mixture culminates when plant material, such as submerged plants, occurs and overlays with the chlorophyll *a* feature of phytoplankton.

2.1.4 The WASI tool as a Physical Model to Simulate Water Spectra

To sum up the previous subchapters, the spectral signature of a water body is a complex composition of various interactions of multiple parameters in and above the water body. Each of the processes can be modeled, and hence, software exists that allows simulations of water spectra according to selected concentrations based on that models. WASI (Water Color Simulation) is one of the software tools that allow such modeling. It can either generate a spectrum according to given input parameters or estimate water parameters with, e.g., least-square fits of a given spectrum [58]. For a simulation of a dataset with WASI, a setup of different selected parameters and their respective range will be described later in section Section 4.3. Other simulation tools similar to WASI are BOMBER (Bio-Optical Model Based tool for Estimating water quality and bottom properties from Remote sensing images) [59] and HydroLight [60]. However, WASI is the most suitable to be applied in this thesis since BOMBER requires remote sensing reflectance, which is not given for the employed spectrometer, and HydroLight is only available with a commercial license.

2.2 Spectral vs. Spatial Resolution for Inland Waters

Regular recordings with satellites to derive water parameters would benefit the work of the authorities in environmental concerns. For example, monitoring with ESA's satellite systems Sentinel-2 [61] and Sentinel-3 [62] would provide data every second to fifth day. However, satellite measurements also have their limitations.

When it comes to remote sensing with satellite images, spatial resolution is often crucial, depending on the application. This also concerns remote sensing of water parameters. The relative amount of mixed pixels between, e.g., the shoreline and pure water area increase with smaller lake size and bigger pixel size. Since it is hardly possible to retrieve water parameters from mixed pixels, they are mostly discarded. Additionally, water parameters, such as chlorophyll, can vary highly over a lake, e.g., forced by wind shift of phytoplankton. Then, in the case of small lakes, trends may be no more recognizable with a pixel size of,

e.g., 300 m for the OLCI instrument on Sentinel-3 [62]. Often, only a single pixel remains after discarding all mixed pixels. If small water bodies shall be supervised with satellite remote sensing, the spatial resolution needs to be far better than the 300 m of Sentinel-3.

Unfortunately, a small pixel size comes along with a worse spectral resolution. This is due to energy effects on a sensor. The smaller the pixel size on the water surface that reflects or emits light, the less energy comes to the sensor, decreasing the signal-to-noise ratio. For a fine spectral resolution, the spectral bands are narrow, resulting in a similar effect as for the small pixel size. Hence, there is always a trade-off between spectral and spatial resolution. Its optimum depends on the application.

The spectral resolution is crucial since it determines how much information is lost on a satellite resolution compared to a full spectrometer resolution. Regarding the shape of the spectrum in Figure 2.2, it becomes evident that the placement of satellite bands is essential. To retrieve, e.g., the phycocyanin concentration, a band placement at 615 nm is vital, which is given for, e.g., Sentinel-3. Chlorophyll *a* is mostly retrieved by the ratio between the minimum at 670 nm and the maximum at 700 nm. Such a ratio model is possible for both Sentinel satellites but not for the Landsat satellites since they lack bands in the specific spectral region entirely [63]. For the latter, a loss of information due to its spectral resolution obviously occurs. Investigations on the performance of retrieval algorithms concerning the spectral resolution were made by, e.g., [63, 64]. Of course, hyperspectral satellites with a constant bandwidth of approximately 10 nm would be the best solution concerning the spectral resolution. However, they come along with either a poor spatial resolution or are not even in orbit yet (see Environmental Mapping and Analysis Program (EnMAP) (Environmental Mapping Analysis Program) [65]).

The optical complexity of CASE II waters is another issue that clarifies the need for a good spectral resolution. CASE I waters have a more constant level of CDOM and only a limited amount of NAP. Hence, the spectral shape is determined mainly by phytoplankton. In addition, the spatial resolution for CASE I waters is of minor concern since they are more homogeneous over broad areas. Even for the Landsat-5 satellite, models exist to retrieve suitable chlorophyll *a* values with the first two bands due to the lack of disturbing particles. This is not possible for CASE II waters, since chlorophyll *a* absorption overlies with CDOM and NAP absorption on those bands. CDOM varies seasonally in structure and amount, whereas the inorganic part of NAP often depends on rainfall, erosion, and the quantity of influx. The shallow littoral zones of CASE II waters are often influenced by their benthic substrate, which impedes the retrieval of water parameters. On the other side, high concentrations of suspended materials increase the signal-to-noise ratio [36]. To conclude this paragraph, CASE II waters have two main disadvantages over CASE I waters: they need an acceptable spatial resolution and a fine spectral resolution due to their optical complexity [16]. As a result, the retrieval of water parameters from CASE I waters with remote sensing techniques

is far more successful than for CASE II waters. **RG 4: Spectral Resolution** deals exactly with the trade-off between varying spectral resolution and the respective model performance.

3 State of the Art – Water Parameter Retrieval Models

The spectral remote sensing signal consists of reflectance values at each spectral band of the respective pixel. Models are the key component to retrieve continuous water parameter values from the respective reflectance values. Three different categories of such models exist: analytical models, empirical models, and data-driven ML models [17]. These model types are briefly explained in the following subsections Section 3.1 and Section 3.2. Data-driven ML models can be seen as a category of empirical models. Since they are crucial in the scope of this thesis, they are treated separately in Section 3.3. However, the fundamentals of ML and some of its algorithms are explained later in Chapter 5. A comparison between the three model types is presented in Section 3.4.

3.1 Analytical Models

Analytical approaches base upon the inherent optical properties (IOPs) [66], describing the underwater light field. Absorption, scattering, and backscattering coefficients of the water constituents reveal the IOPs of a water column [67]. Hence, a physical relation between the IOPs, the subsurface irradiance reflectance, atmospheric conditions, and the sun exists. They can be approximated or calculated with radiative transfer models [68, 69, 70, 71, 72]. As a result, the IOPs are quasi-independent of changing illumination conditions [73]. Thus, theoretically, the analytical models work for most of the water bodies. Unfortunately, exact information and an appropriate parametrization of local IOPs are required for a robust model [74, 17].

Analytical approaches to retrieve water parameters consist of two models, a forward model, and an inversion model. The bio-optical forward model relates the optically active constituents, such as CDOM, NAP, and phytoplankton, with the IOPs. Then, according to the IOPs, radiative transfer models calculate, e.g., the remote sensing reflectance as a representative of the (AOPs). They can be performed by analytical relationships (e.g. [75, 67, 76]) or by numerical radiative transfer (e.g. [77, 78]). The inverse model matches the measured AOP (e.g., remote sensing reflectance or radiance reflectance) with the simulated ones to retrieve the water parameters. Neural networks frequently represent these kinds of models (e.g., [79, 80]. For example, WASI is such an analytical model [58].

Its forward mode allows simulations of water spectra according to selected parameters, whereas for the inversion, it applies, e.g., a least square fit [73].

One advantage of the analytical approach is the simultaneous retrieval of the optical active water parameters [76]. In turn, they need measurements at the local water bodies for a suitable approximation of the local IOPs. Compared to empirical models, analytical models are generally complex and sensitive to errors concerning atmospheric correction [76, 81]. Another issue is the overlay of IOPs, such as scattering and absorption on NAP and chlorophyll *a*, that does not allow a unique solution [82].

3.2 Empirical Models

The most prominent approaches to retrieve water parameters are the empirical models. The idea of empirical models is to select spectral bands that are usually physically related to the target parameter (see Figure 2.2). Frequently, band arithmetic is employed to create newly features, such as two band ratios [44, 83, 84, 85, 86, 87, 88, 89] or normalized band ratios [90]. Then, a regression function is calculated with newly obtained features for the available water samples. Thus, they are also called feature engineering approaches.

Commonly, the regression is calculated as a linear function (see, e.g., [91, 92, 83, 93, 94, 85, 86]). But there are also other possibilities like: polynomial with second degree (e.g. [95, 84]) or third degree (e.g., [88]), logarithmic functions (e.g. [96, 89]), or generally spoken non-linear functions (e.g. [90, 97, 86, 87]).

Empirical models are not only built with two spectral channels, but also single-band approaches [95, 86] and ratio approaches with up to four bands exist [90, 93, 94]. Others use the area under the peak at 700 nm [98] or directly the peak height [38, 99]. Some multi-band ratio approaches calculate the logarithm (e.g., [100, 96]) of the band's reflectance. Another possibility is multiple linear regression models without band arithmetic but band selection. Therefore several bands are selected, and the regression function is calculated, e.g., a linear regression [101, 100, 102].

In general, empirical approaches exist for every water parameter mentioned in Section 2.1.2. The easiest parameter to retrieve with such models are TSS, investigated by (e.g., [90, 95, 83, 87, 88]). Closely related to TSS is the Secchi Depth (e.g., [86, 100]). For chlorophyll *a* hundreds of empirical algorithms exist, most of them are related to the 670 nm absorption feature, if it is available for the applied sensor (e.g. [90, 97, 91, 92, 83, 93, 84, 94, 85, 101, 100, 102, 96, 98]). Turbidity is related to chlorophyll and investigated by (e.g., [95, 86, 100, 102]). Phycocyanin retrieval is related to its distinct absorption feature at 620 nm, which is always considered for feature engineering approaches (e.g., [44, 89]). CDOM estimations often deliver unsatisfying results [47, 103], since it has not directly a spectral

feature and overlaps with chlorophyll and other pigment features. Hence, it is the most challenging parameter to retrieve [47]. Especially, the performance of the models decreases in water bodies with moderate to high chlorophyll concentrations [48]. Mainly bands with the shortest wavelengths are applied in the empirical algorithms due to the stronger absorption in the blue region of the visible spectrum [90, 83, 86, 100].

The Advantage of empirical ratio models comes to forth for satellite bands. Setting the spectral bands into a ratio neglects the absolute value. Hence, the surface reflectance induced by sky glint, which is rarely wavelengths dependent Section 2.1.1, is automatically corrected, which leads to stable results. Calculating derivatives lead to a similar result in neglecting the absolute values. Several authors (e.g., [104, 105, 106]) investigated derivative-based models up to the fourth grade. The main disadvantage of the empirical models is the transferability to other water bodies. Mostly, it is not the selected bands that need to be adapted, but the model parameters. In the case of a linear regression model, a different slope and intercept are optimal for every water body. To obtain both model parameters, reference data of the specific lake is needed with various values of the target parameter. In sum, such models are simple in their construction but generalize poorly, and in turn, they need data for every water body.

3.3 ML Models

Data-driven ML models can be seen as a category of empirical models. ML approaches have been rarely seen in the related work of water parameter retrieval. This may be due to the demand for a vast amount of data for its training process [17]. Especially when working with satellite data, the atmosphere has an additional impact compared to spectrometer data (cf. Section 2.1.1). Either its impact has to be corrected, or the model itself needs to find solutions for varying atmosphere conditions, resulting in an even higher demand for data. Contrary, spectrometer data may allow such approaches since the impact of the atmosphere is lower.

The general functioning of ML models is explained more intensively in Chapter 4. In this thesis, the following ML algorithms are investigated as candidates for water parameter retrieval models: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARS), k-Nearest Neighbors (k-NN), Artificial Neural Network (ANN), and 1-dimensional (1D) Convolutional Neural Network (CNN).

In the related literature, MARS has not been applied once, yet in the context of retrieving water parameters. Slightly more applications concern the k-NN and the GB. The latter was applied to retrieve CDOM with Sentinel-2 images in a reservoir [107], whereas the k-NN has been used for chlorophyll *a* and TSS estimations with Sentinel-2 in two different lakes [108]. SVM and RF show significantly more applications than the algorithms before. For example, [109] used a RF and a SVM applied on satellite data to estimate chlorophyll *a* and suspended

particulate matter in coastal water. Whereas the ANN-based models are clearly the most picked supervised ML models to retrieve water parameters (e.g. [110, 111, 112, 113, 19]). Recently, even CNNs have been trained for water monitoring. [114, 115] applied CNNs on Landsat 8 and Sentinel-3 images to quantify water quality. Contrary to the CNN applied in this thesis, Pu et al. [114] focused on a classification with different quality levels. However, the WaterNet by Syariz et al. [115] estimates the chlorophyll *a* concentration, also considering spatial features. Silveira Kupssinskü et al. [108] conducted a comparative study with Sentinel-2 data on two different lakes to retrieve TSS and chlorophyll *a*. In this study, the RF was the superior model in estimating both parameters for both lakes over the ANN, SVM, and k-NN. ANN and k-NN achieved a similar performance but slightly worse than RF. The SVM clearly underperformed compared to the other three models. However, in a comparative study with satellite data to retrieve chlorophyll *a* and TSS, a SVM outperformed the RF in costal waters [109]. In a comparison study with Landsat images by [116] to retrieve TSS and chlorophyll *a*, an ANN apparently performed best, whereas the SVM slightly outperformed the RF. Hence, it is not generally possible to determine the best algorithm in before. Pahlevan et al. [19] use a mixture density model as a kind of neural network approach on in situ water samples of several water bodies of various types. After scaling the spectrometer data to satellite resolution, they applied the model to atmospherically corrected satellite images.

3.4 Model Synopsis

A vast amount of water parameter retrieval algorithms have been published in recent decades. The models presented in the sections above work exceptionally well, independent of their type, whether it is an analytical, an empirical, or a machine learning model. Indeed, they work suitable in their studies, but mostly, they are only valid for the water body they were designed for [17, 81, 16]. The main issue of such models is the degree of generalization, which is always a trade-off with the performance. Depending on the application, different degrees of generalization are suitable. For water bodies with a long available data history, a specialized model may be suitable. Then, for every possible and realistic parameter combination of the specific water body, data exists, and the model can be tuned precisely for every case. Unfortunately, such a data history is rarely available, and most likely, such a specialized model will work poorly on other water bodies [17].

All over the world, about 117 million lakes exist with a surface greater than $0.002\,\mathrm{km}^2$ [18]. Most of them have not even been monitored once [16]. Hence, it is not feasible to build specialized models for most of them. This is where the advantage of a generalized model comes to forth. Those models can be built and validated on multiple water bodies with different characteristics. Naturally, their performance will not achieve the quality of a specialized model made only for a single water body. But it is possible to retrieve suitable values for many

lakes without any prior knowledge. Unfortunately, it is difficult to evaluate the degree of generalization since no unbiased dataset represents the 117 million lakes properly yet [16].

Pahlevan et al. [19] presented one approach towards generalized models. They scaled in situ water spectra of various lakes to the Sentinel-2 and Sentinel-3 resolution and applied the NN-based model on real satellite images. A comparison with an empirical two band and three-band ratio approach reveal an outperformance of the neural network model. Spyrakos et al. [117] suggested 13 optical water types. For each of the types, semi-generalized models can be trained. In a first step, a water body is assigned to one of the 13 types. Then, the respective semi-generalized model estimates the water parameter values [118]. A similar approach was also followed by [19] and revealed improvements compared to a complete generalized model.

To sum up, there is a demand for a higher degree of generalization of the models. Especially the water bodies that have not frequently been monitored would benefit from such generalized models. But for such models, a vast demand for data exists.

4 Datasets and Data Preparation

Datasets are necessary to investigate the RGs, presented in Section 1.2. Since the focus is on data-driven ML approaches to estimate various water parameters, especially chlorophyll *a*, a vast demand for data exist. There are two reasons due to the thesis relies on in situ measurements. First, spectral in situ data allows a finer coupling with reference data than satellite images since the distance between both samplings is optimized. Hence training the models with more exact datapoints may provide additional potential. Second, the atmospheric correction of satellite images above inland waters is a challenge that is not well understood yet [16, 119]. Following the ML approach on satellite data would increase the demand for data, including reference data, significantly due to even more degrees of freedom for the models considering the atmosphere. Hence the ML approach for estimating water parameters with spectral data was investigated on in situ measured data.

During this thesis, the data have been conducted on many recording days and different field campaigns. The data applied in this thesis can be sorted into three categories. First, there was the participation at the "Elbschwimmstaffel" organized by the Federal Ministry of Education and Research in June 2017. It consists of about 1000 datapoints, with reference values of chlorophyll *a*, Green Algae, Diatoms, CDOM, and turbidity linked to hyperspectral data along the River Elbe. A detailed description of the genesis of the dataset is given in Section 4.1. The general applicability of ML models in the retrieval of different water parameters and the comparison between suitable models will be investigated on this dataset.

Secondly, the SpecWa dataset was conducted on eleven different water bodies in the surrounding region of Karlsruhe during the summer half of 2018 and 2019. The dataset consists of 3645 datapoints, measured with a spectrometer, whereas the reference data is measured by either the AlgaeTorch or the Phycolab. Since the instruments measure different water parameters, the focus is only on the overlapping chlorophyll *a* concentration. This dataset represents the primary investigations and encloses all RGs of this thesis. The dataset is described intensively in Section 6.2.

Finally, a third dataset exists, the WASI dataset, which is entirely simulated with the WASI tool. It provides 528 000 datapoints to face the sparse data challenge and allows the application of DL techniques. The trained models shall show the generalization ability of such models and, thus, the transferability between water bodies and different datasets. The simulation of the data is specified in Section 4.3.

4.1 The River Elbe Dataset

This section includes material from the publication

Sina Keller, Philipp M Maier, Felix M Riese, Stefan Norra, Andreas Holbach, Nicolas Börsig, et al. "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity". In: *International journal of environmental research and public health* 15.9 (2018), p. 1881. It is cited as Keller et al. [22] and marked with a may green line.

and from the publication

Philipp M Maier, Stefan Hinz, and Sina Keller. "Estimation of Chlorophyll A, Diatoms and Green Algae Based on Hyperspectral Data with Machine Learning Approaches". In: *Tagungsband der 37. Wissenschaftlich-Technische Jahrestagung der DGPF e.V.*. Vol. 27. Munich, Germany: Deutsche Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation, 2018, pp. 49–57. It is cited as Maier et al. [20] and marked with a pink line.

The following section describes the collection and the components of the Elbe dataset, the first dataset that was conducted and evaluated within the scope of this thesis. The dataset was measured during the Elbe field campaign[1] along a 575 km stretch from Bad Schandau downstream to Geesthacht from June 24 to July 12, 2017. The study area and the location of the probe measurements are visualized in Figure 4.1. The dataset is composed of three different sensors: The bbe fluorometer PhycoSens, the Biofish multi-sensor system, and a visible and near-infrared (VNIR) hyperspectral sensor (Cubert UHD 285). All sensors were carried on the research vessel *Elbegrund* of the German Federal Waterways and Shipping Administration of Germany. The measured water parameters include the concentrations of CDOM, chlorophyll *a*, green algae, diatoms, and turbidity that are timely connected to the hyperspectral images. It is the data basis for the estimation of several water parameters with ML models on a single water body, which was published in [21, 20, 22]. In the following, the measurement setup with its three different sensor systems is described.

4.1.1 Sampling Chlorophyll *a*, Green Algae, and Diatoms

The PhycoSens fluorometer is mounted in front of the research vessel. This instrument enables in-situ measurements of water quality parameters without additional sample preparation by filtration or with solvent. It simultaneously determines chlorophyll *a* concentrations, transmission of light, and optional photosynthetic activity. This sensor also measures

[1] The Elbe field campaign was funded by the German Federal Ministry of Education and Research.

Figure 4.1: Map of the study area along the River Elbe with the probe measurements. Taken from [22].

the amount of unbound phycocyanin, which mirrors the release of blue-green algae contents. Chlorophyll *a* and phycocyanin content is excited by seven LEDs at frequencies of 370 nm, 430 nm, 470 nm, 525 nm, 590 nm, and 610 nm to obtain a meaningful fluorescence excitation spectrum. The fluorescence emission is measured as an answer to the excitation and allocated to the different algae classes such as green algae, cyanobacteria, cryptomonads, or the class of diatoms. In this context, the green algae and diatom concentrations are expressed as the chlorophyll *a* equivalent concentrations derived from specific fluorescence signatures of green algae. The sampling interval of the PhycoSens was set to every 5 min over the whole field campaign.

4.1.2 Sampling CDOM and Turbidity

The multi-sensor system Biofish monitors eight relevant water quality parameters: temperature, electrical conductivity, oxygen concentration and saturation, pH, CDOM, chlorophyll *a*, turbidity, and photosynthetic active radiation. During the Elbe field campaign, the Biofish sensor system is installed at a fixed depth of around 0.5 m underneath a floating cylinder and is mounted on a crane in front of the research vessel (cf. Figure 4.2). All data is sampled online at a 4 Hz frequency and is tagged immediately with GPS measurements. For a detailed view of the sensor specifications, see Holbach et al. [120]. To get more stability in

Figure 4.2: Application of the three sensor systems on the research vessel *Elbegrund*. The cranes on the front carries the Biofish sensor system (white floating cylinder, front right) and the PhycoSens collects water (front left) in the river Elbe. The hyperspectral sensor is mounted close to the railing in between the cranes. Taken from [22].

the data, the median values for each parameter are calculated for every minute. Although the Biofish monitors the chlorophyll *a* concentrations additional to the PhycoSens, the dataset relies only on the PhycoSens values since they allow a more detailed view of the different algae classes. Solely the CDOM and turbidity values of the Biofish were added to the dataset.

4.1.3 Recording Hyperspectral Images

The hyperspectral snapshot sensor Cubert UHD 285 records high-dimensional images non-invasively every 0.5 min to 1 min. It is mounted on a tripod next to the Biofish sensor system at the front of the research vessel. The sensor was calibrated every 20 min with a white reference to compensate for the varying sun altitude. Every minute we captured a hyperspectral snapshot within a 70° angle towards the water surface. The spectralon was placed on the railing so that a part of it is visible in every snapshot to control the reflectance. In addition, we equalized minor radiative fluctuations such as slight cloud occurrences. Based on the spectralon reflectance values of the calibration, we removed outlying images whose reflectance on the spectralon differed around the factor of 0.3. The outlying behavior was mainly affected by shadow occurrence caused by ship turnings, bridge crossings, or sudden cloud coverages.

Each hyperspectral image is characterized by 50×50 pixels and 125 spectral channels, each with a spectral resolution of 4 nm. The spectrum ranges from 450 nm to 950 nm. We select an area in each image which is free of bubble formations, shadows, or waves to calculate

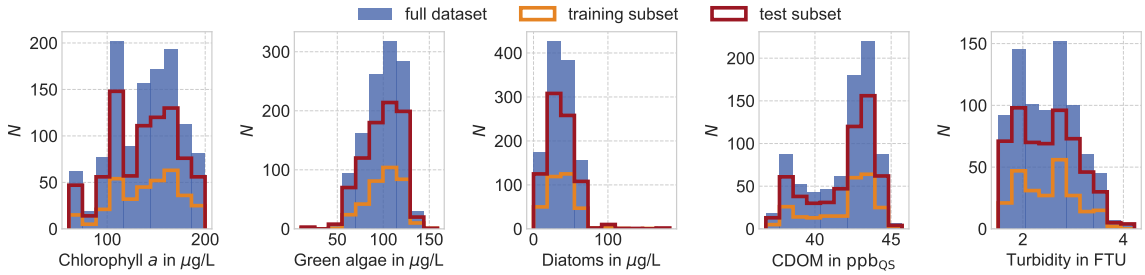

Figure 4.3: Distributions of the water quality parameter values. Each full dataset (blue bars) is split randomly into a training (orange) and a test (red) subset. The number of datapoints is symbolized as N. Taken from [21].

a mean spectra per image manually. To exclude sensor errors, we apply a feature band selection resulting in a range of wavelengths between 470 nm to 910 nm.

4.1.4 Elbe Field Campaign Datasets

For a temporal matching of the sampled PhycoSens data to the hyperspectral data, we need to extend the former by linear interpolation. This is possible due to the continuous change of the sampled chlorophyll *a*, green algae, and diatoms concentrations. The data sampled by the Biofish sensor system can be matched directly to the hyperspectral data due to its continuously high temporal resolution.

The Elbe field campaign results in five datasets, one dataset for each of the five water quality parameters. A dataset contains datapoints, whereby one datapoint is defined by 111 selected hyperspectral bands and one value of a respective water quality parameter. The distribution of the water quality parameters is visualized in Figure 4.3. Datapoints with a chlorophyll *a* concentration higher than 200 µg/L are dismissed since they exceed the measurement range of the PhycoSens.

4.2 The SpecWa Dataset

This section includes material from the publication

Philipp M Maier and Sina Keller. “SpecWa: Spectral remote sensing data and chlorophyll a values of inland waters”. In: *GFZ Data Services* (2020). It is cited as Maier and Keller [25] and marked with an orange line.

The following section describes the SpecWa datset and its collection setup during the years 2018 and 2019. It is the second dataset that was measured and evaluated within the scope

of this thesis. Contrary to the Elbe dataset, the SpecWa dataset was collected at 11 inland water bodies and with different instruments. Since each measurement period lasts for four to five months, seasonal effects occur additionally, resulting in different concentrations of water parameters and spectral variability. This dataset was published separately as [25]. The SpecWa dataset is part of each published study (see Maier and Keller [23, 24] and Maier et al. [26]) succeeding the River Elbe evaluations. However, the first two only rely on the 2018 part of the dataset due to their earlier release.

4.2.1 Description of the Equipment

To measure the spectral data of the water bodies, we applied a RoX spectrometer (Reflectance Box, see JB Hyperspectral Devices) covering a spectral range from 341 nm to 1014 nm with an intermediate sampling interval of 0.65 nm. It consists of two fiber-optic cables. One fiber-optic cable has a cosine receptor at its top with a field of view of 180°, while the other one has a field of view of 25°. The integration time of the sensor is determined by the incoming radiation on the cosine receptor. This is very useful for measuring under various angles of the sun or changing cloud conditions. The final spectral result is the reflectance, the ratio of the reflected radiance to the incoming irradiance on the cosine receptor for each wavelength. The RoX spectrometer was calibrated by the company in the laboratory against an Ulbricht sphere.

The water parameters were monitored by the AlgaeLabAnalyser (bbe moldaenke) in 2018 and by the AlgaeTorch (bbe moldaenke) in 2019. The AlgaeLabAnalyser is a spectral fluorometer. It measures chlorophyll *a* in a range from $0.1\,\mu g\,L^{-1}$ to $200\,\mu g\,L^{-1}$ with a resolution of $0.01\,\mu g\,L^{-1}$. In addition, it can distinguish the following algae classes: green algae, diatoms, cryptophytaceae, and cyanobacteria. Moreover, it measures CDOM (bbe moldaenke). The AlgaeLabAnalyser is a device for measuring in the laboratory. Therefore, we collected samples at the water bodies. In the laboratory, each of the samples was analyzed in the device, where they are stimulated by various LEDs. The AlgaeTorch, however, uses a measurement technique to monitor chlorophyll *a* directly in the water. It measures the chlorophyll *a* concentration in the range from $0.1\,\mu g\,L^{-1}$ to $200\,\mu g\,L^{-1}$ with a resolution of $0.01\,\mu g\,L^{-1}$. The device is placed in the water, and after a measurement time of about 15 s, the values of the concentrations are stored and displayed. It measures the intensity of fluorescence by the stimulation of the algae with different LEDs. In addition, it measures turbidity and the proportion of chlorophyll *a* related to cyanobacteria (see bbe moldaenke).

4.2.2 Experimental Setup

The measurements intended to collect on-site hyperspectral data of different water bodies and the corresponding near-time water parameters on the local scale. For the recording

Figure 4.4: Measurement setup of the RoX spectrometer at a natural water body (left) and artificial ponds (right). Taken from [23]

of the hyperspectral data, we installed the RoX spectrometer on a tripod. The fiber cable with the cosine receptor was directed perpendicular to the sky while the bare fiber cable pointed to the water surface, hence in the opposite direction (see Figure 4.4). Depending on the water body, we applied two kinds of installation for the tripod and the spectrometer: when monitoring a natural water body with a flat littoral zone, we installed the tripod as far inside the water as it was possible with respect to the water level (see Figure 4.4, left), otherwise when monitoring an artificial water body without a flat littoral zone, we installed the tripod outside the water (see Figure 4.4, right). In addition, we aimed to minimize the influence of the ground during the hyperspectral data acquisition as well as other influencing factors such as shadows of surrounding trees or buildings.

The sampling frequency of the RoX was set to 15 s in case of the year 2019 measurements and 20 s in case of the year 2018 measurements. That sampling frequency resulted in three to four spectra per minute. During the measurements in the year of 2018, we took a reference water sample to measure the water parameters every fifth minute, which we later analyzed with the AlgaeLabAnalyser in the laboratory. During the measurements in the year of 2019, we sampled in-situ measurements of the water parameters approximately every second minute with the AlgaeTorch. The water samples were collected in a depth of 20 cm below the water surface and as close as possible to the spectrometer measurements. Thus, we ensured to record the same water conditions with the remote sensing sensor and the in situ probes. At each water body, we switched the devices to different spots in the water. Furthermore, we aimed to conduct our measurements under optimal conditions with clear sky and without clouds, nevertheless during a few measurements, slight cloud cover occurred.

4.2.3 Measured Water Quality Parameters

Table 4.1 summarizes the measured water parameters. Besides the chlorophyll *a* concentration, represented by the different algae classes, we monitored turbidity and CDOM. Since

Table 4.1: The water parameters retrieved by the reference measurements with the respective device in the years 2018 and 2019. Taken and adapted from [25].

Parameter	Unit	Description	Year
Chlorophyll *a*	$\mu g\,L^{-1}$	Measured concentration of chlorophyll *a*	2018 and 2019
Cyanobacteria	$\mu g\,L^{-1}$	Measured concentration of cyanobacteria	2018 and 2019
Green algae	$\mu g\,L^{-1}$	Measured concentration of green algae	only 2018
Diatoms	$\mu g\,L^{-1}$	Measured concentration of diatoms	only 2018
Cryptophytaceae	$\mu g\,L^{-1}$	Measured concentration of cryptophytaceae	only 2018
Turbidity	FTU	Measured turbidity	only 2019
CDOM	$\mu g\,L^{-1}$	Measured concentration of CDOM	only 2018

the calibration range for the chlorophyll *a* concentration is from $0\,\mu g\,L^{-1}$ to $200\,\mu g\,L^{-1}$, we removed the chlorophyll *a* values above $200\,\mu g\,L^{-1}$. Note that we monitored different water parameters during the measurements in the years of 2018 and 2019 due to the characteristics of the two different in situ measurement devices.

4.2.4 Pre-Processing of the Spectrometer Data

As a first pre-processing step, we dropped the spectral data at the edge of the spectral range of the RoX spectrometer due to e.g., sensor noise and atmospheric disturbances. This results in spectral data in the range of 389 nm to 910 nm. Furthermore, we excluded spectra, which we identified as outliers, such as spectra including waves at the water surface or spectra measured during cloudy conditions. To identify outliers, we first calculated the median spectrum for each water body at each specific measurement date. Then, we compared the respective spectrum to the median spectrum and excluded it if the specific spectrum was higher than twice the median spectrum or lower than half of the median spectrum. Since the measurement conditions were carefully chosen, the number of outliers is low. In the last pre-processing step, the spectra were assigned to the measured values of the water parameters using the respective timestamp. Since the sampling frequency of the spectral RoX sensor was higher than the sampling frequency of the probe measurements, some spectra were aligned to the same probe measurements. In total, the dataset for the year 2018 consists of 1305 datapoints and the dataset for 2019 consists of 2380 datapoints. We define one datapoint as the spectral data in the range of 389 nm to 910 nm and all respective reference values such as the chlorophyll *a* concentration and turbidity.

4.2.5 Site Properties

The measurements in the year 2018 took place at eleven different inland waters, whereas for the year 2019, the focus was on only six specific water bodies. However, we increased

the frequency on those water bodies compared to the measurements in 2018. The locations of samplings are illustrated in Table 4.2 and Figure 4.5.

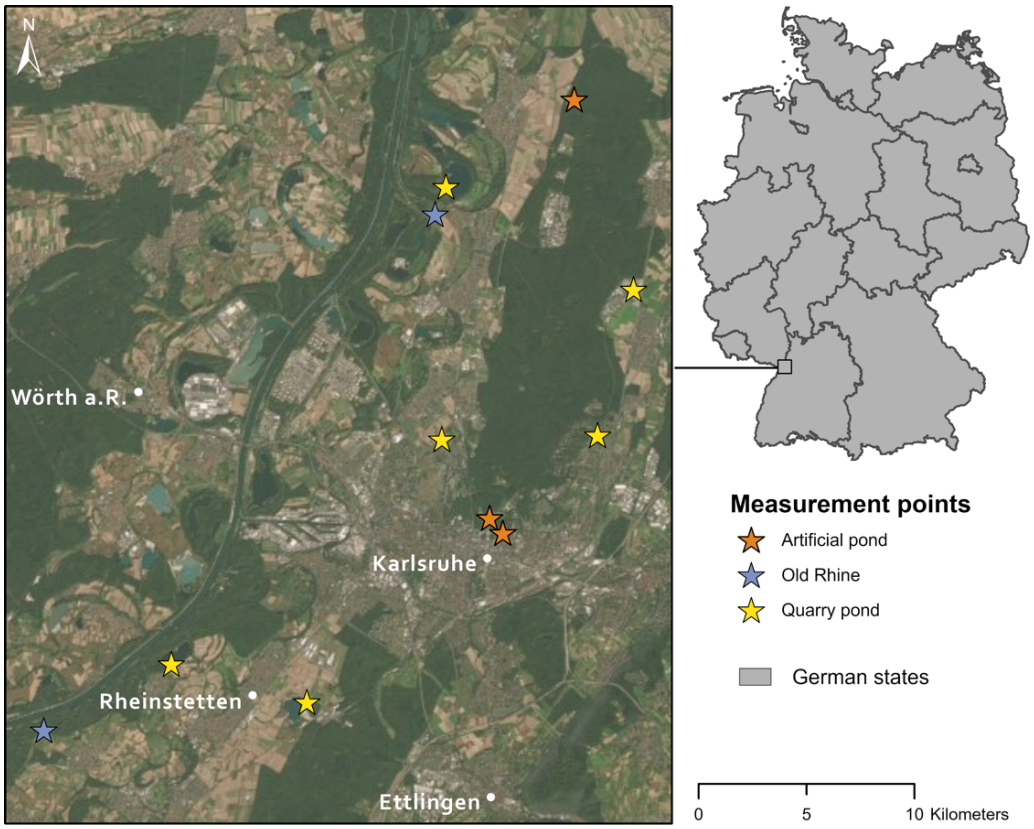

Figure 4.5: Map of the study area including the investigated water bodies of the SpecWa dataset. Map projection: WSG 1984. Taken from [25].

Table 4.2: Brief summary of the investigated water bodies. The number of datapoints refers to the number of spectral RoX measurements assigned to the respective values of the water parameters. The term ap refers to *artificial pond* while qp means *quarry pond*. A water ID is given for each water body for more simple overview in figures. A means artificial, whereas W means a more naturally water body. The water depth refers to the measuring point under the RoX. Taken and adapted from [25].

Water body	Water ID	Lon	Lat	Number of datapoints	Water depth in m	Chl *a* range in $\mu g\,L^{-1}$
ap castle garden	A1	8.4049	49.0170	1322	1.0 to 2.0	6.3 to 171.1
ap KIT	A2	8.4104	49.0129	907	0.5 to 1.0	22.2 to 199.6
ap TMB	A3	8.4401	49.1312	60	2.0 to 3.0	61.0 to 100.5
old rhine au	W1	8.2193	48.9590	21	2.0 to 3.0	4.7 to 9.1
old rhine leopoldshafen	W2	8.3822	49.0999	8	0.5 to 1.0	9.8 to 11.0
qp blankenloch	W3	8.4648	49.0794	494	0.5 to 3.0	2.4 to 21.0
qp epple	W4	8.3288	48.9667	42	1.0 to 3.0	1.6 to 13.0
qp ferma	W5	8.2724	48.9771	20	1.0 to 3.0	3.3 to 6.4
qp heide	W6	8.3850	49.0385	221	1.0 to 3.0	1.7 to 16.5
qp leopoldshafen	W7	8.3867	49.1074	105	1.5 to 3.0	0.0 to 8.7
qp waldstadt	W8	8.4498	49.0396	485	1.5 to 3.0	0.0 to 17.0

In the entire dataset, we identified the three artificial ponds (A1-A3) as eutrophic water bodies with the highest concentrations of chlorophyll *a*. On the contrary, the quarry pond in Leopoldshafen (W7) is characterized by a very low chlorophyll *a* concentration. The chlorophyll *a* concentrations of the other water bodies lay in between. Most of the measurements took place at the artificial ponds in the castle garden (A1) and the KIT (A2) due to an enormous change in chlorophyll *a* concentrations over the seasons. Additionally, both water bodies are the only ones in the SpecWa dataset with cyanobacteria occurrence during the summer period. Both old Rhine branches in Au (W1) and Leopoldshafen (W2) belong to the natural water bodies of this dataset. A low water depth characterizes the measurement side of the latter. The quarry ponds in Waldstadt (W8) and Blankenloch (W3) seem to be similar to each other concerning the chlorophyll *a* concentrations at a medium level. Both are the most frequently investigated natural water bodies in the SpecWa dataset. Finally, the quarry pond Epple (W4) differs strongly caused by its high content of suspended materials, leading to high reflectance values. This is due to the still active gravel depletion.

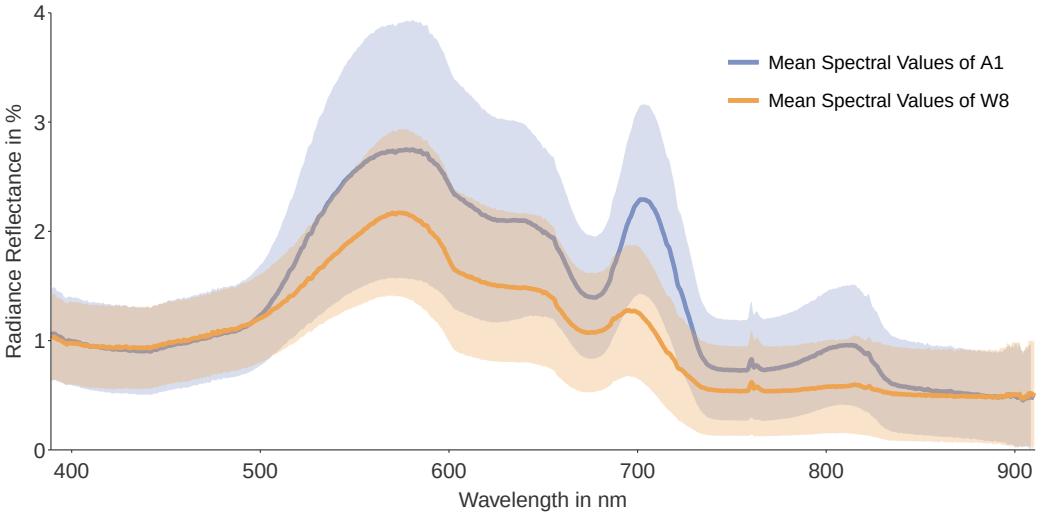

Figure 4.6: Visualization of the spectral data of the two SpecWa water bodies. The solid lines refer to the mean of the waterbodies A1 (blue) and W8 (orange). The brighter area represents the respective standard deviations. In the SpecWa dataset, the radiance reflectance is the normalized ratio between the water leaving radiance combined with the surface reflectance and the incoming irradiance. The corresponding chlorophyll *a* concentrations and information are given in Table 4.2. Taken from [26].

Figure 4.6 shows the radiance reflectance of the water bodies A1 and W8 with the mean and the respective standard deviations to get a first impression of spectral signatures of the dataset. Both spectra are similar to each other in the short wavelengths of 400 nm to 500 nm and the longer wavelengths. Between 500 nm to 800 nm, the reflectance of A1 is significantly higher than of W8. Likely, this is due to scattering effects related to high chlorophyll *a* concentrations of A1 (see Table 4.2). For both spectra, the absorption minimum at 670 nm is visible but more distinct for A1.

Figure 4.7 is a visualization approach to cluster the datapoints of SpecWa by considering the first four principal components. Some clusters are recognizable that are mainly forced by the water bodies. As the most important finding, there is no trend recognizable towards a specific water parameter. Regarding the datapoints of A1 – A3, representing the high chlorophyll *a* concentrations, they are distributed all over the plot. This finding may be influenced by a substantial contribution of surface reflectance and bottom reflectance to the measured radiance reflectance. Therefore, a Principal Component Analysis (PCA) [121] may not be suitable for clustering water bodies, showing the spectral composition's complexity.

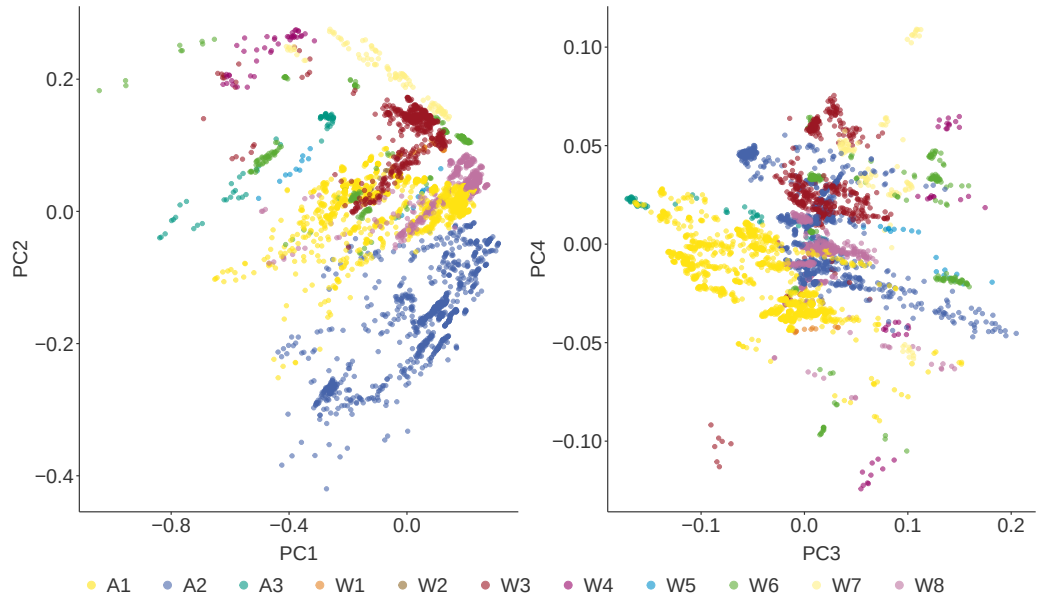

Figure 4.7: A visualization of the SpecWa dataset with the first four principal components. The first two components are plotted on the left, whereas the third and the fourth component are plotted on the right. The different colors represent the IDs of the eleven different inland water bodies described in Table 4.2. As input for the Principal Component Analysis, the reflectance values of the wavelengths 389 nm to 910 nm have been selected for every datapoint of the SpecWa dataset.

4.2.6 Data Application in Published Studies

The SpecWa dataset is applied in three published studies that are subject to this thesis. In each of the three studies, the SpecWa dataset was modified compared to its published version [25]. For the final study in this thesis [26], the datapoints of Table 4.2 are reduced according to a threshold for the concentration of chlorophyll *a* and cyanobacteria. This threshold was necessary due to the combination of the dataset with the WASI tool and its limitations. For the Studies [24, 23] only the 2018 part of the SpecWa dataset was applied since they have been published before the data collection of the 2019 part of the dataset. Additionally, they include datapoints, sorted out for the publication of the SpecWa dataset [25]. This removal concerned two water bodies with a meager amount of datapoints,

that have been measured from a bridge. Thus, the measurement setup was problematic since shadows and reflections on the water surface occurred due to the higher measurement level compared to the other water bodies. Another difference between the two studies and the published SpecWa dataset was the conception of the datapoints. In the studies relying solely on the 2018 part of the dataset [24, 23], a different data fusion approach was applied compared to the published version [25]. In the SpecWa dataset, a reference value was matched with multiple measured spectra. However, for [24, 23], a reference value was matched with only a single spectrum. This limitation is necessary due to the conception of the studies. The spectra that are sampled during the minute in which the reference sample is collected are expected to be similar to each other. When applying a random split, as it was done in both studies [24, 23], it will be problematic if one of the spectra is used in the training and the other one in the test dataset. Such a setup would facilitate the estimation task significantly. Therefore, only one spectrum was combined with a reference value [24, 23]. Contrary, the WASI study [26] is based on entirely different datasets. Hence a linkage between a reference value with multiple spectra is not problematic.

4.3 The WASI Dataset

This section includes material from the publication

Philipp M Maier, Sina Keller, and Stefan Hinz. "Deep Learning with WASI Simulation Data for Estimating Chlorophyll a Concentration of Inland Water Bodies". In: *Remote Sensing* 13.4 (2021), p. 718. It is cited as [26] and marked with a purple line.

The following paragraphs describe the last dataset involved in the scope of this thesis. Contrary to the River Elbe and the SpecWa dataset, it is entirely simulated with the WASI tool and therefore named WASI dataset. The WASI tool was briefly described in Section 2.1.4. It bases mainly on the physical relationships presented in Section 2.1. The dataset's genesis is forced by the need for a vast amount of data, allowing the application of deep learning techniques to estimate the chlorophyll *a* values of the entirely independent SpecWa dataset. This approach presumes transferability between the datasets in terms of the spectral signals and the environmental parameters. A successful application on the SpecWa dataset would confirm a high degree of the model's generalization. Therefore, the WASI dataset must comprise the possible combinations of water ingredients and atmospheric variations underlying in the SpecWa dataset to enable transferability. Thus, the final study (Maier et al. [26]) in the scope of this thesis faces the challenges above, relying on the WASI and the SpecWa dataset.

4.3.1 Fundamentals and Genesis

Since models developed on the WASI dataset should show the transferability to datapoints of another dataset, the input data structure of the WASI-generated simulation data needs to be similar to the SpecWa data structure. In our case, the spectral data of the SpecWa dataset is given in radiance reflectance (see [25]). The radiance reflectance is the sum of the water leaving radiance and the reflectance on the water surface on a radiance sensor in proportion to the incoming irradiance [31, 28]. Consequently, we simulate data with the WASI tool as radiance reflectance data by varying the input parameters. For our setup, the WASI tool provides 33 parameters that can be adapted for radiance reflectance, but only three parameters can be varied simultaneously. Besides, we select twelve out of the possible 33 WASI parameters which affect the water-leaving radiance and the surface reflectance. The remaining WASI parameters are set to constant default values. Table 4.3 summarizes the twelve selected WASI parameters.

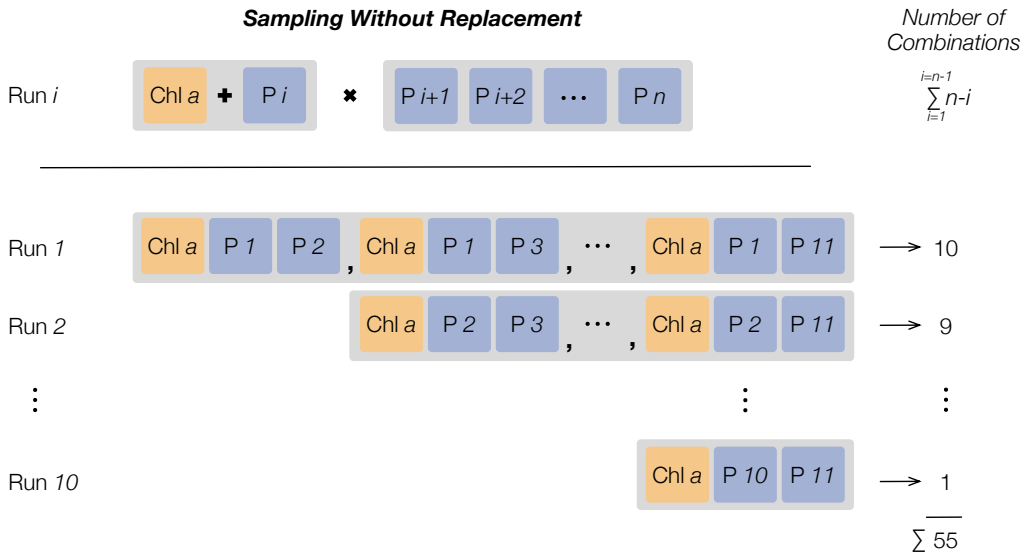

Figure 4.8: Sampling schema of the selected WASI parameters. Chl *a* refers to the Chlorophyll *a*, while P I are the remaining n = 11 parameters given in Table 4.3. I describes the control variable and I = 1 ... n – 1. Taken from [26].

We define a sampling schema to handle the parameter combination, which is visualized in Figure 4.8. We include the chlorophyll *a* concentration as a parameter in every run and select two additional out of the remaining 11 parameters (see Table 4.3) in an iterative process as variable settings. Since we aim to cover every combination of the parameters and the chlorophyll *a* concentration, we receive 55 WASI-parameter combinations in ten runs, as shown in Figure 4.8.

In each run, we consider the selected three parameters' value range and their frequencies' distribution (see Table 4.3, *range* and *steps* column). Both, the range and the frequency,

are selected according to the following two criteria: (i) We use the range of the respective parameter so that a wide variety of possible inland water bodies is represented. (ii) In addition, we simulate the respective parameter value's frequency so that it is nearly equally or logarithmically equally distributed. This distribution is crucial to ensure that the DL approaches are provided with the full range of the data and not only the majority.

The remaining WASI parameters, which are not among the selected three parameters, are set to a constant value according to Table 4.4. These constant values are also given in Table 4.3 at column *Standard*.

4.3.2 Parameter Selection

Among the twelve parameters varied in the simulation grid, there are not only water parameters, but also atmospheric parameters and different types of benthic substrate. Those twelve parameters that we consider essential are explained in the following passages and summarized in Table 4.3.

Table 4.3: Summary of the relevant WASI simulation parameters with their respective range. The sampling schema is described in Figure 4.8. The range and the respective steps define the possible occurring parameter values. For chlorophyll *a* and the concentration of NAP, a logarithmic scale is chosen. W.P. means WASI parameter. Taken from [26].

W.P.	Range	Standard	Steps	Log scale	Description
Chl *a*	$1\,\mu g\,L^{-1}$ to $100\,\mu g\,L^{-1}$	-	30	yes	concentration of chlorophyll *a*
C_X	$0.1\,mg\,L^{-1}$ to $100\,mg\,L^{-1}$	1	20	yes	concentration of non-algal particles type I
C_{Mie}	$1\,mg\,L^{-1}$ to $20\,mg\,L^{-1}$	0	20	no	concentration of non-algal particles type II
C_Y	$0.1\,m^{-1}$ to $5\,m^{-1}$	0.1	20	no	CDOM absorption
zB	1 m to 5 m	2	10	no	water depth
Sun	35° to 65°	50	10	no	sun position
FA1	0.1 to 5	0	10	no	background type sand
FA2	0.1 to 5	0	10	no	background type silt
FA5	0.1 to 3	0	10	no	background type macrophyte
g_{dd}	$0\,Sr^{-1}$ to $0.5\,Sr^{-1}$	0.02	10	no	fraction of sky radiance due to direct solar radiation
g_{dsr}	$0\,Sr^{-1}$ to $1\,Sr^{-1}$	0.318	10	no	fraction of sky radiance due to molecule scattering
g_{dsa}	$0\,Sr^{-1}$ to $1\,Sr^{-1}$	0.318	10	no	fraction of sky radiance due to aerosol scattering

Regarding the SpecWa dataset, green algae and diatoms mainly occurred as phytoplankton classes. The wavelength-dependent absorption for both algae classes is read in with the default file provided by the WASI tool. We simulated the whole sampling schema twice,

once for each algae class. The first time, green algae represent the varying chlorophyll *a* concentration, whereas the diatom concentration is excluded, while the second time, it is vice versa. The WASI-generated chlorophyll *a* data covers the range from $0.1\,\mu g\,L^{-1}$ to $100\,\mu g\,L^{-1}$ in 30 steps in each run. Regarding the chlorophyll *a* distribution in the SpecWa dataset and the distributions in other studies (e.g., Pahlevan et al. [19]), we decided on a logarithmic distribution, resulting in fewer datapoints with high concentrations.

For the simulation of NAP, we selected both different types implemented in WASI: C_X and C_{Mie}. We use the standard model settings of WASI for both parameters and change only the concentration according to Table 4.3. For both parameters, we selected different simulation ranges: C_X is simulated from $0.1\,mg\,L^{-1}$ to $100\,mg\,L^{-1}$ with 20 steps in a logarithmic step width, and C_{Mie} is simulated from $0\,mg\,L^{-1}$ to $20\,mg\,L^{-1}$ with a linear step width. C_X is varied with a logarithmic stepsize, according to less frequency of high concentrations in natural water bodies, especially lakes. The two NAP parameters differ in their scattering model.

The effect of high CDOM absorption in waters is intensively explained in Section 2.1.2. For the simulation, the standard settings of WASI for the absorption model are applied. The parameter is varied within 20 steps in between the $0.1\,m^{-1}$ to $5\,m^{-1}$. The range mainly comprises the CDOM values of natural water bodies (e.g., Pahlevan et al. [19]). However, we cut the highest values because SpecWa's water bodies do not belong to the humic water bodies and show low CDOM concentrations.

The water depth is another parameter that affects the water spectrum. It determines the attenuation of the light in the water column and hence functions as a weight for the reflectance on the bottom constituents. The water depth for the simulation is selected in between 1 m to 5 m, which approximately fits SpecWa's water bodies (see Table 4.2). Since the SpecWa dataset was measured on shallow water bodies, we decided that a maximum of 5 m is enough for the simulation.

As background type of the water bodies, we simulated sand, silt, and macrophytes that are read in by the default WASI file. The more shallow a water body and the less suspended materials occur in the water, the higher the benthic substrate's impact on the spectrum. We decided to use the three different constituents since we think one or a mixture represents the water bodies' benthic substrate in the best manner.

Another parameter with a crucial impact on the spectrum is the solar zenith angle. Since the SpecWa dataset is measured in the season between June and October with varying daytime, many different solar zenith angles occur. It influences the water body's illumination conditions, the sun glint on the water surface, and the attenuation in the water column. The sun zenith angle is varied in the simulation from 35° to 65°, representing the conditions during the measurements.

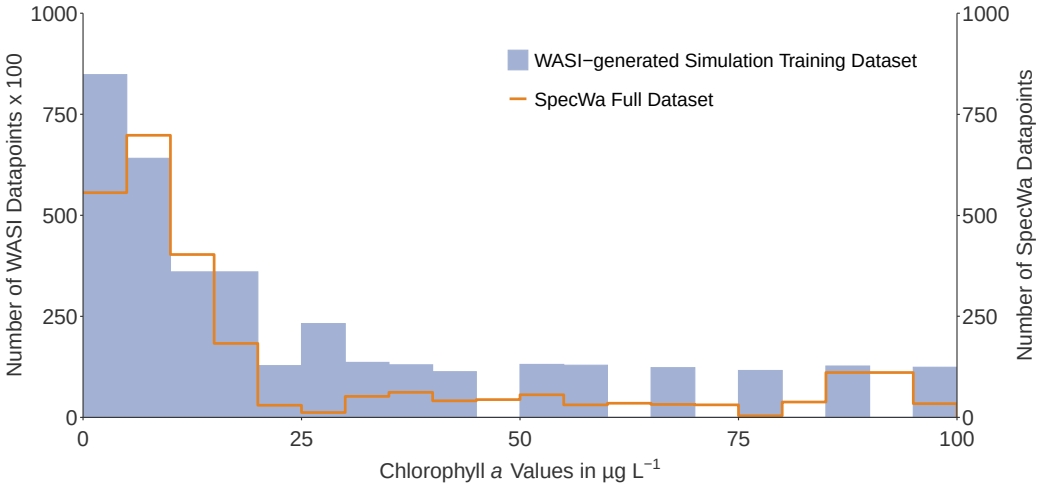

Figure 4.9: Distributions of of the chlorophyll *a* values between the WASI training subset and the SpecWa dataset. The WASI training set refers to the left y-axis and has about 100 times more datapoints than the SpecWa dataset which is referred to the right y-axis. The training set represents 70 % of the complete WASI dataset. Taken from [26].

To finally get a partly independent dataset of different illumination conditions, three atmospheric parameters (g_{dd}, g_{dsr}, and g_{dsa}) are varied during the simulation. The factor g_{dd} is a weighting factor for the direct solar radiation, whereas g_{dsr} and g_{dsa} are factors for diffuse radiation. These parameters heavily influence the surface reflectance due to sun and sky glint, which can be stronger than the water leaving radiance itself. The selected values for g_{dd} in the simulation can double the reflectance with the same setup for the other water parameters. g_{dd} and g_{dsa} result approximately in a simple y-offset of the radiance reflectance, whereas the effect of g_{dsr} depends more on the wavelength.

The part of the 33 WASI parameters that remains constant over the whole simulation process is summarized in Table 4.4. Finally, we receive a number of 528 000 WASI-generated datapoints containing the radiance reflectance values in the range of 400 nm to 900 nm with a spectral resolution of 1 nm and varying values of the twelve selected parameters. The distribution of the chlorophyll *a* concentration of the WASI dataset and the SpecWa dataset is visualized in Figure 4.9.

4.4 Downsampling of the Spectral Data

The influence of the spectral resolution on the estimation performance of a ML model is one topic covering this thesis. The measurements that have been conducted during the field campaigns in the scope of this thesis have been made with two different spectral sensors. First, the Cubert UHD 285 was applied in the Elbe field campaign, and second, the

Table 4.4: Summary of the default WASI simulation parameters. Taken from [26], (cf. [28]).

Parameter	Standard Value	Unit	Description
C[0]	0	$\mu g\,L^{-1}$	Concentration of phytoplankton class 0
C[1]	0	$\mu g\,L^{-1}$	Concentration of phytoplankton class 1
C[2]	0	$\mu g\,L^{-1}$	Concentration of phytoplankton class 2
C[4]	0	$\mu g\,L^{-1}$	Concentration of phytoplankton class 4
fluo	0		chlorophyll a fluorescence quantum yield
S	0.014	nm^{-1}	Exponent of CDOM absorption
n	−1	-	Angström exponent of particle scattering
T_W	25	°C	Water temperature
f	0.033	-	f-factor of R
Q	5	Sr^{-1}	Anisotropie factor of upwelling radiation
z	0	m	Sensor depth
view	0	°	Viewing angle
bbs_phy	0.001	$m^2\,mg^{-1}$	Specific backscattering coefficient of phytoplankton
f_nw	0	-	Fraction of non-water area
fA[0]	0	-	fraction of bottom type #0 (constant)
fA[3]	0	-	fraction of bottom type #3 (seagrass)
fA[4]	0	-	fraction of bottom type #4 (mussel)
f_dd	1	-	Fraction of direct downwelling irradiance
f_ds	1	-	Fraction of diffuse downwelling irradiance
H_oz	0.38	cm	Scale height of ozone
alpha	1.3170	-	Angström exponent of aerosols
beta	0.2606	-	Turbidity coefficient
WV	2.500	cm	Scale height of precipitable water in the atmosphere
rho_L	0.020 06	-	Fresnel reflecance of downwelling radiance
rho_dd	0.033 25	-	Reflection factor of E_{dd}
rho_ds	0.0889	-	Reflection factor of E_{ds}

RoX spectrometer was employed for the SpecWa dataset. Additionally, a third imaginary sensor exists, that produced the simulated WASI dataset.

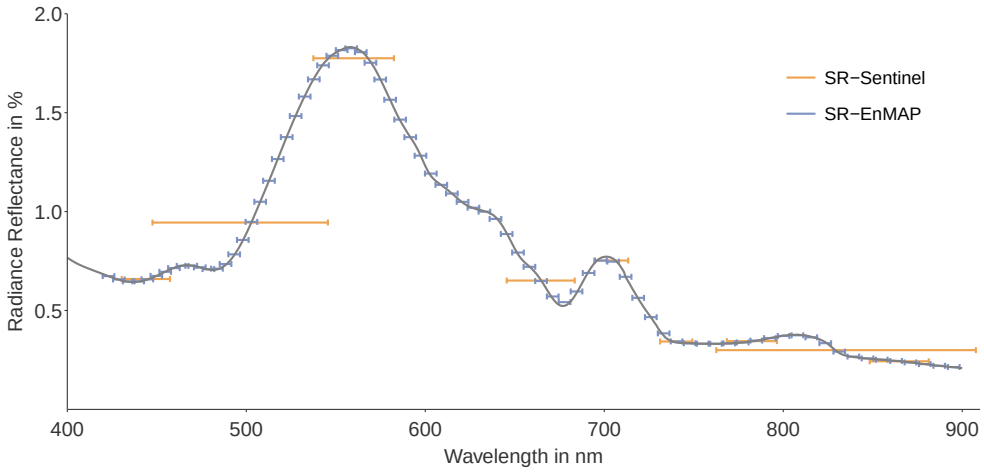

Figure 4.10: Visualization of the two different downsampled spectral resolutions. The spectral resolution of the Sentinel-2 mission (orange) is referred to as `SR-Sentinel`, and the spectral resolution of the EnMAP mission (blue) as `SR-EnMAP`. The grey line represents a selected WASI-generated simulation spectrum with a chlorophyll a value of $51\,\mu g\,L^{-1}$, a concentration of suspended materials type I of $7.8\,mg\,L^{-1}$, and a sandy bottom substrate. The additonal WASI parameters are set to the (default) values according to Table 4.3 and Table 4.4. Taken from [26].

For the River Elbe dataset, representing the study setup I, the spectral resolution of the Cubert UHD 285 has not been mutated. Hence the data with a spectral resolution of 4 nm has been used directly as input data for the ML models. However, the spectral data of the RoX sensor shows an averaged spectral resolution of 0.65 nm. Fitting ML models with such kind of resolution would result in strong overfitting of the models since the neighboring spectral bands are highly correlated. In addition, such a spectral resolution is not feasible when concerning a later application with satellite data. The same applies to the WASI dataset, which is originally available with a spectral resolution of 1 nm.

Thus, for the investigation of the SpecWa dataset, two approaches were followed concerning study setup II. First, a scaling to spectral bands with continuous values (published as Maier and Keller [23]). Therein, the 0.65 nm resolution is scaled to hyperspectral bands with a resolution of: 4 nm, 8 nm, 12 nm and 20 nm. Therefore, the input of the overlying raw spectral bands are averaged. In a follow-up study (published as Maier and Keller [24]), the resolution is scaled to the ones of real satellites. Therefore, six satellite missions are selected: the multispectral Landsat 5 and Landsat 8 mission, the multispectral Sentinel-2 and Sentinel-3 mission and the hyperspectral Hyperion and EnMAP mission. The key characteristics of the mentioned satellite missions are summarized in Table 4.5. For most of the satellite missions, the actual spectral response function exists. If the spectral response function is available, it will be used to scale the data to the spectral bands of the concerning satellite mission. This is not the case for the hyperspectral EnMAP mission since

EnMAP has not been launched yet, and so no spectral response function exists. Hence, the EnMAP bands have been calculated with a Gaussian function over each band's central wavelengths. This applies also to the Hyperion and the Sentinel 3 mission. All these scaling operations are implemented and performed with the `hsdar` package [122] in R. The scaling approach of the spectral data was inspired by, e.g., [64].

Table 4.5: Summary of some characteristics of the different satellite systems used for the data simulation covering the spectral range between 400 nm to 900 nm. The hyperspectral satellite missions are highlighted by $*$. SRF means spectral response function, GF means Gaussian function. Taken and adapted from [24].

Satellite mission	Number of bands	Bandwidth in nm	Spectral range in nm	Spatial resolution in m	Approach for the simulation
Sentinel 2	9	18 to 145	443 to 865	10 to 60	SRF
Sentinel 3	19	2.5 to 75	400 to 900	300 to 1000	GF
Landsat 8	5	16 to 60	443 to 865	30	SRF
Landsat 5	4	60 to 140	485 to 840	30	SRF
Hyperion*	54	10	406 to 895	30	GF
EnMAP*	77	6.5	423 to 895	30	GF

The downsampling to another resolution is especially necessary for the combined application of different spectral data. This is intended for the last study setup III, within the scope of this thesis combining the SpecWa dataset and the WASI dataset. Though, the 1 nm resolution of the WASI dataset and the 0.65 nm of the SpecWa dataset are downsampled to the same spectral satellite resolution to get comparable input features for the ML approach. Therefore, a common denominator is found in the spectral bands of the multispectral Sentinel-2 and the hyperspectral EnMAP resolution. The resulting spectral bands are illustrated in Figure 4.10.

5 Fundamentals and Training Process of Applied Machine Learning Approaches

ML approaches have been rarely seen in the related work of water parameter retrieval. This might be caused by multiple reasons, such as, e.g., computational power for earlier studies in previous decades or a vast demand for data. Especially the latter is often the case impeding ML applications for inland water remote sensing. However, the composition of inland waters' spectral signal is a mixture of many water ingredients. Their spectral features mostly overlap themselves (see Section 2.1.2). Hence, it is not feasible to retrieve a single water parameter without understanding the others. This might be only acceptable for models designed only for a specific water body, in which only the target parameter varies, and the others remain more or less constant. However, to get a more generalized model applicable for multiple water bodies, which is one goal in this thesis, they need to understand the relevant features throughout the spectrum. ML models might derive such information out of the complex composition of the spectral signal. Therefore, it might be suitable to rely on ML approaches for generalized models. In the following, an introduction to ML models and their training process is given.

5.1 Machine Learning in General

ML is an umbrella term for different kinds of learning approaches. In the scope of this thesis, ML mainly refers to supervised learning. Rarely, unsupervised learning techniques are subject to this thesis. The central point of a learning approach is the dataset. The datasets presented in Chapter 4) follow all the same schemes. They consist of several datapoints, whereas a datapoint is set together of a spectrum represented by the reflectance values at each band and a reference value, e.g., the chlorophyll *a* concentration at the same time. In the context of ML, the reference values refer to target values of a target variable or outputs, and the reflectance values refer to input features or inputs [123].

One difference between supervised and unsupervised learning techniques is the reference to a target variable. Supervised learning techniques require target variables. For supervised learning, each spectrum (input values) is related to a reference value (output value) at

the same time (see Chapter 4). The supervised model learns the linkage between the reflectance value for each band (input) and the chlorophyll *a* concentration (output) during a training process [124, 123, 125]. Minimization of the error guides improvements of the learning calculated between the model output and the available reference values [126]. The learning process is described in more detail in Section 5.3.

Contrary to supervised learning, unsupervised learning techniques are independent of target parameters [123]. These are often not available, such as satellite images without reference values. Therefore, unsupervised learning is frequently applied for clustering approaches and dimensionality reduction [126, 125]. Dimensionality reduction is valuable for some tasks, especially for highly correlated input data combined with a limited amount of datapoints [125]. Besides, information instead of noise can be lost due to the reduction. Clustering approaches with unsupervised learning techniques can be suitable, such as pre-studies to find, e.g., similarities between datapoints or identifying trends in the dataset. In the case of a Principal Component Analysis (PCA) [121], e.g., correlations between principal components and environmental variables may be found. Since target values, such as the chlorophyll *a* concentration, exist for every datapoint investigated in this thesis, the focus is on supervised learning techniques. Mainly, a PCA is applied as pre-processing to reduce the dimensionality of the input data for some supervised learning models.

Depending on the target variable, whether it is continuous or discrete, a regression or a classification task is underlying. For continuous water parameters such as chlorophyll *a* or CDOM, like every parameter in this thesis, regression models are selected. If the water quality is determined to different quality levels, for instance, the different trophic states, it would have been a classification task. Since the presented datasets all obtain continuous target values, only regression models are trained for their estimation in this thesis.

In general, every regression task is different. There is no unique approach that adapts well on all datasets [127]. Hence it is a good approach to select multiple ML algorithms finding the most suitable one for the challenge to be solved. Therefore, during this thesis, a framework of different supervised ML approaches is always applied and evaluated on different datasets described in Chapter 4. The different selected supervised learning models, their possible pre-processing possibilities, and their training process are subject to the following subsections.

Supervised ML models have rarely been applied to estimate water parameters with remote sensing data yet. Hence, it is one focus in this thesis. Therefore, the following sections concerning supervised ML are introduced in more detail. First, data preparation, including necessary dataset splits and pre-processing of the data, are described in Section 5.2. Then, the training and evaluation process is in the focus Section 5.3. Eventually, the distinct learning algorithms and their application in the related work are given in Section 5.4. However, the exact setup that is applied to train the later models is presented in Chapter 4.

5.2 Data Preparation

Before training a supervised learning model, a split of the complete dataset needs to be carried out. Therefore, the whole dataset is divided into different subsets of variable sizes. There are two options for the split: first, into a training and a test set, or second, into three sets, a training, a validation, and a test set [125]. The first option is schematically illustrated in Figure 5.1. In general, it is advantageous when the distribution of the target values is considered for the split so that the shape of their histogram is similar to each other set. If extreme values only occur in the test set, it is less likely that the model estimates them well. The ratio between the training and the test set is often a trade-off. The bigger the test set, the better to show a model's generalization ability. Otherwise, the training set must be representative of the whole dataset. A model needs enough datapoints to learn the linkage between input features and target parameters. In addition, the more input features exist, the more training data is necessary to avoid overfitting. After the split, storing the datasets is essential to obtain reproducible and comparable results for each applied model.

Pre-processing of the data may provide improvements in the model's performance. Within the scope of this thesis, three different kinds of pre-processing methods will be applied: Scaling, a PCA, and calculating derivatives of the spectrum. Scaling is essential when the magnitude of the input features differs strongly, which may influence the model performance. In this context, an option is min-max scaling, which is defined as:

$$\Delta_{\text{scaled}}(x) = \frac{x - \min(X_{\text{train,i}})}{\max(X_{\text{train,i}}) - \min(X_{\text{train,i}})}, \tag{5.1}$$

with a training dataset $X_{\text{train,i}}$ of the hyperspectral band i. This results in normalized input features with a range from 0 and 1. With such a scaling approach, no feature is favored according to its magnitude. A PCA [121] is an unsupervised technique but can be applied as another pre-processing possibility. PCA is an orthogonal transformation that finds new axes, the so-called Principal Components, according to the variance of the input data. With ascending Principal Components, the variance decreases. Hence, most of the variance and its related information are summarized in the first few components. For instance, the first four Principal Components of Figure 4.7 contain 98.9 % of the available variance of the SpecWa dataset. Using only a small number of the Principal Components to fit the model may reduce overfitting, especially for small datasets with high dimensions.

Eventually, derivatives of a spectrum can be applied as pre-processing steps. Only the first derivative is calculated in this thesis, whereas derivatives up to the fourth grade have been

applied in the literature [105, 106]. An advantage of derivatives is to reduce noise, e.g., y-offsets, influenced by, e.g., sky glint, mostly disappear, and mainly the shape of the spectrum is investigated. However, these offsets can also contain information, which is lost in turn.

In the case pre-processing is applied, it must be conducted after the split is made, especially for scaling and a PCA. Then, the transformation is made on the training set and finally applied to the test set and the validation set (see Figure 5.1 dashed arrows). This is important to get an honest model evaluation since scaling the data and a PCA before the splitting leads to an interaction between training and test set, which are no longer independent in turn. Following this procedure, the final test set remains unseen for the model till its evaluation [123, 125]. Contrary to the other pre-processing methods, calculating derivatives is an invariable function that is not affected by each other's datapoints.

5.3 Training and Evaluation Process of Supervised ML Approaches

Figure 5.1 visualizes the possible steps that are conducted during the supervised learning process, from the pre-processing of the data to the evaluation of the final model.

The training process is the centerpiece of supervised ML. It takes place solely on the training set. During the training phase, the supervised models are optimized to minimize the training error, e.g., by a loss function [126]. This optimization concerns the so-called model parameters that depend on the training data. Model parameters are, e.g., the weights of a neural network or the value of a split on a decision tree. Frequently applied training error measurements are the Mean Squared Error (MSE) or the Root Mean Squared Error (RMSE). Different metrics lead to different effects on the penalization of errors. For example, the weight of outliers has a more substantial impact on the MSE than on the RMSE. The selection of the metric depends on the model's application.

Optimizing the model parameters is a task that is repeated very often during the training phase because, first, the hyperparameter tuning is conducted. Hyperparameters are the parameters that define the configuration of a model, e.g., the number of layers in a neural network or its amount of neurons. Each ML algorithm has individual hyperparameters that need to be adapted for the regression task. The suitable set of hyperparameters is adapted before training the final model parameters. One option to get the best fitting set of hyperparameters is a grid search. If a grid search is applied, possible values for each hyperparameter are defined and combined with each possible value of the other tunable hyperparameters. Then, models are trained and evaluated for every hyperparameter combination, whereas the best combination is the hyperparameter setting for the final model. One established method is the k-fold cross-validation (CV), which is schematically

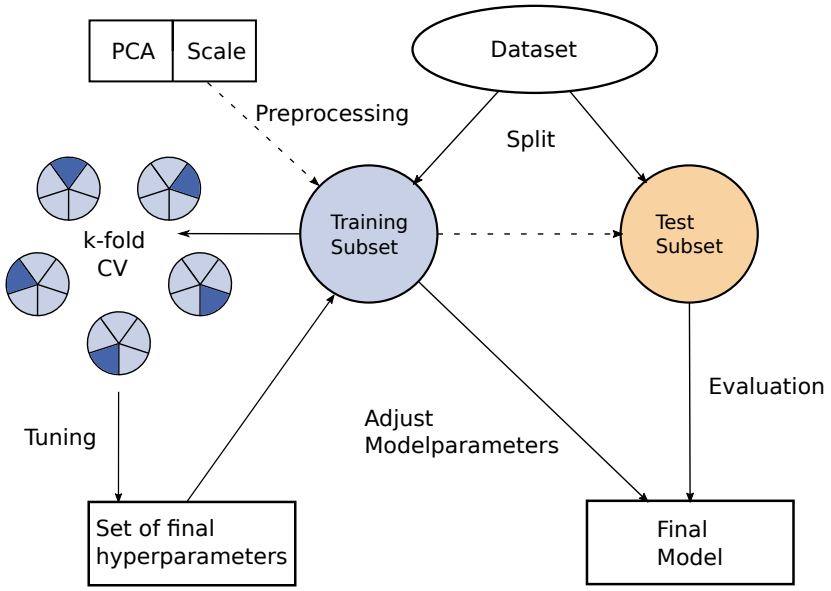

Figure 5.1: Flowchart of the training process of a supervised learning model with its possible steps on two distinct datasets. The k-fold CV can be replaced by bagging. Pre-processing is an optional step. This transformation is always calculated on the training subset and then applied to the test subset. The hyperparameters are adjusted with a grid search by training each combination with CV. Model parameters are trained on the whole training subset with the before-adjusted hyperparameter combination. The final evaluation of the model is conducted on the test subset.

implemented in the training process in Figure 5.1. Therefore, the training set is split into k parts without replacement. Then, on k-1 parts, the model is trained, whereas the model is evaluated on the remaining part. This process is conducted to choose every of the k parts for the evaluation. The hyperparameter setup that performs best in average over the k-folds is set to be the combination for the final model. Additionally, the CV can be repeated with a different random composition of the k-folds to get a more stable model. Multiplying the number of folds in the CV and its repetitions with the possible combinations of the grid search result in a vast amount of trained models to solely find the perfect set of hyperparameters. Ultimately, the training of the final model takes place on the whole training set without partitions. For its evaluations, the final model is applied to the hold-out test set. An alternative to a full grid search may be a predefined number of random selections in the grid that shortens the hyperparameter tuning process [125]. Moreover, trying different hyperparameter values may also be an option [125].

Another option to avoid the long-lasting CV process is a split of the whole dataset into three subsets. Then, the hyperparameter tuning can occur on the whole training set, whereas the validation set is used to measure the performance of each hyperparameter combination. Then again, the evaluation of the final model is conducted on the test sub-

set. Therefore, much data is required since, for each set, the distribution of the reference values should be considered. Such splitting is often conducted for neural networks to obtain a completely independent model fit.

In general, the aim of a ML model is to learn the linkage between the inputs and the output of the training dataset and generalize them to unknown datasets. Nevertheless, achieving both simultaneously is nearly impossible. Hence, it is always a trade-off between optimization and generalization of a model [128, 123, 129, 125]. If a model is strongly optimized, it adapts well to the training data. Since the model learns not only the connections between input features and target parameter but also outliers and the overlying noise of the data, this results in a model that shows a tendency towards overfitting. Otherwise, underfitting can occur when the model does not respond to the characteristics of a dataset. Then it may be able to ignore outliers and does not respond to noise, but it still results in a weak performance.

With the application on the unknown dataset, the model's performance is finally proven. In terms of regression, it can be computed as, e.g., the coefficient of determination (R^2) or the MAE and the RMSE. The advantage of MAE and RMSE is that they are noted in the same physical unit and are therefore simple to interpret. A considerable difference in the performance metrics between the test and the training set may indicate a weak trade-off between optimization and generalization. Otherwise, dataset shifts may occur, mainly when datasets collected with different sensors are applied and are hard-split into training and test set.

5.4 Applied ML Algorithms

In the following subsections, the functioning and the respective characteristics of supervised learning algorithms that are implemented in the scope of this thesis is briefly explained. The algorithms include Random Forest (RF), Gradient Boosting (GB), Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS), k-Nearest-Neighbor (k-NN), Artificial Neural Network (ANN), and one-Dimensional Convolutional Neural Network (1D CNN).

5.4.1 Random Forest

The Random Forest (RF) [130] algorithm is a basic supervised learning model and the one that is applied in every published study within the scope of this thesis. RF is an ensemble method of decision trees. A decision tree consists of root nodes and leaves connected by branches. At each node, a split occurs according to an input variable's threshold value that is adapted during the training phase. Splits occur at each following node, after which the importance of the variable used for the split is descending. The importance of a feature is determined by e.g., permutation. In total, these splitting procedures lead to a tree-like structure.

The cutting process for each tree stops when the maximum number of nodes, depth, or purity is achieved, which was determined before. Finally, every datapoint is assigned to a leaf. In the case of a regression, the average of the target parameter value is calculated at each leaf [125].

To get RF algorithm from a decision trees, bagging is applied. For bagging, a certain amount of datapoints is selected randomly with replacement from the training set, which is repeated several times. As a result, slightly different datasets occur. On each of the selected parts, the decision trees are trained, whereas on the other part, the estimation performance is tested, the so-called out-of-bag error. The advantage of such a procedure is to reduce the variance. The number of trees is determined before training the model. For a regression model, the estimated parameter value of a datapoint is the average estimation of each tree.

Another strength that improves the stability and the performance of a RF model is the so-called feature bagging. Feature bagging means that at each tree split, the number of possible features (variables) is randomly selected. Then again, the most important one is selected as a split variable. Feature bagging leads to less correlation between the trees [123]. The number of features that are considered for feature bagging at each split can be tuned as a hyperparameter in the grid search.

The minimum number of observations in a leaf is another hyperparameter. It is an essential parameter to get influence on the generalization ability of a model. Small numbers lead to good training performance, whereas higher numbers result in more generalization. Thus, it allows control of not creating a new leafs for a single datapoint. Limiting the tree depth would be another option to concern this issue leading to similar effects.

The split rule is crucial for the RF. It is determinable in a grid search [131]. At this point, the focus is on extratree since this option always outperformed the others when applied in a grid search. In the literature, extratree stands for extremely randomized trees (ET) [132], and it is often seen as an own algorithm. In this thesis, it is treated as an adaption of the RF. Hence, in this thesis and the publications in its context, it always means extratree when RF is mentioned. For extratree, the splitting threshold at each node is set randomly for each feature and random subset. In the end, the best threshold is selected [123, 125].

The RF algorithm is characterized by a simple training concept with a low amount of hyperparameters in the grid search. Thus, it provides acceptable first results makes with low effort. In addition, slight variations in the hyperparameters do not influence the result strongly. The availability of a feature importance function also allows conclusions to physical interactions at the different wavelengths, which is calculated by, e.g., permutation.

5.4.2 Gradient Boosting

The gradient boosting (GB) algorithm is another ensemble method based upon decision trees [133]. In contrary to bagging on RF, boosting is applied. In gradient boosting, shallow trees are fitted incrementally and optimized alongside gradient descent. Since gradient boosting includes many more degrees of freedom in the grid search with its multiple adjustable parameters Table 5.1, a RF model is easier to optimize.

5.4.3 Support Vector Machine

Support Vector Machines (SVM) [134] are algorithms that try to find decision boundaries in high dimensional feature space to separate datapoints in terms of a classification task. In the learning process, the SVM tries to maximize the margin between the datapoints and their classes [123]. The closest points to the hyperplane are the so-called support vectors. Since finding linear boundaries are impossible in high-dimensional data, a so-called kernel trick is applied. This kernel can have various characteristics, such as linear, sigmoidal, or polynomial constitution. With the kernel trick, the dimensionality can be increased, in which the datapoints are then again separable with a linear boundary. The decision boundary is a soft margin with a distance ϵ to the hyperplane [123].

Contrary to common regressions, the regression version of the SVM tries to fit the boundary lines so that they include a maximum amount of datapoints within a specific ϵ value. This means that the error is not penalized as long as it is less than ϵ [135, 123]. Else, the C (cost) parameter is implemented as a trade-off that deals with the values that exceed ϵ [135]. This is necessary so the model does not adapt to each outlier. C is a tunable parameter that is adapted within a grid search.

One of the most common kernels for SVM is the radial basis kernel, which is the only one applied within this thesis. γ is the tunable kernel parameter that determines the shape of the kernel. Increasing γ leads to more complex separation structures, which may lead to overfitting. On the contrary, if γ is too low, the model is too shallow and unable to learn the data's complexity. Here again, the trade-off between optimization and generalization appears.

An advantage of SVM is its good performance on small datasets. Contrary to the RF, the hyperparameters need to be tuned more precisely within the grid search. For big datasets, the optimization of a SVM is a time-consuming process [125]. Several pre-processing steps are possible for the SVM. The input data can be scaled, or a PCA can be applied. Relying on derivatives of the spectrum can also be meaningful.

5.4.4 Multivariate Adaptive Regression Spline

Multivariate Adaptive Regression Spline (MARS) is an adaption of regression splines. First of all, a regression spline is an expansion of a linear stepwise regression, which consists of linear regression models for different intervals. Contrary to a linear regression line, a spline consists of polynomials of the grade N. Hence, it is some kind of function of piecewise polynomials. The respective smoothing parameters are the number and placement of knots that define the intervals of a spline and its degree [123].

Since selecting the number and the placement of the knots is a combinatorically complex task, simplifications are made, which results in the MARS algorithm. MARS can be seen as a generalization of the linear stepwise regression [123].

Two different passes are run during the model's training process: a forward and a backward pass [136]. The MARS algorithm does not use the original input variables in the later phase. Instead, it generates pairwise functions. The second part of the pair is always a mirrored function. These functions are the so-called candidate functions. The best-suited candidate function is selected according to minimizing the loss in the forward pass. Since this reduction still results in a huge model, pruning is applied, where the functions with less impact on the final model are dismissed. Pruning is conducted stepwise on the backward pass. The number of allowed functions used at the maximum and the grade of the polynomial can be determined within a grid search.

5.4.5 K-Nearest Neighbor

In the k-Nearest Neighbor method (k-NN), the k datapoints with the closest distance to the observation in the feature space determine its output value [123]. For the regression case, the output is calculated by the average value of the k closest datapoints weighted by the inverse distance to the observation [125]. So the closer a datapoint, the higher its impact. An advantage of the k-NN algorithm is the simple training process. K is the only hyperparameter that needs to be tuned. Its disadvantage comes with high-dimensional data. The difference between close and far away decreases with rising dimensions [125].

5.4.6 Artificial Neural Networks

One of the most powerful supervised learning approaches is an artificial neural network (ANN). The basic idea of an ANN is to imitate biological neurons. Each network consists of a certain amount of neurons, which have an input and an output. The input to a neuron is the weighted output of a predecessor neuron combined with an activation function. Its output is then again calculated and passed through the next neuron. Each connection

between the neurons depends on the weights adapted during the training phase through backpropagation [126]. If neurons are arranged in different layers, connected to each other, the network is called fully-connected (FC). Such networks are the most common ones [125]. Different activation functions exist for the activation of the neurons in each layer. The shape of the function determines how a neuron is activated. Recently, the Rectified Linear Unit (ReLU) function [137] is the most commonly used activation between the layers. Each ANN consists of an input layer and an output layer. The output layer is commonly activated with a linear activation function for regression tasks. It consists of only a single neuron. Optional layers in between the input and the output layer are called hidden layers. With an increasing number of hidden layers, the ANN shifts towards a deep neural network. A deep neural network with its vast number of layers and neurons is able to learn more complex tasks. But in turn, the adjustable parameters and hence, the degrees of freedom increase strongly. Therefore, a vast amount of datapoints is necessary. Overfitting may occur. Regularization techniques can help to prevent overfitting. Such techniques are L2 regularization contributing to the loss function, dropout layers, or random noise. A dropout layer randomly disables neurons in the training phase [126]. Random noise layers add some kind of noise to the data, most commonly Gaussian noise. Both dropout and noise layers are only implemented in the training phase and not in the final model.

One option to fit neural networks (either an ANN or a CNN) is the `Adam` optimizer [138]. Adam optimization is an extension of a stochastic gradient descent, which updates the network's weights more efficiently by using momentum and adaptive learning rates for faster convergence. Training of neural networks takes place in epochs. The process ends when the maximum number of epochs is accomplished or a stopping criterion is achieved [126].

5.4.7 One Dimensional Convolutional Neural Network

Convolutional neural networks (CNNs) [140] belong to artificial neural networks. They consist of an input layer, several convolutional layers, and an output layer. Fully connected layers are often applied between the convolutional layers and the output layer, as schematically illustrated in Figure 5.2. CNNs extract features in different directions that is why they are popular in image understanding. Therefore, several filter layers are applied with different filters and filter sizes, convolved with the input data. Since this thesis's datasets only exist of spectral data, without spatial information, a one-dimensional CNN (1D CNN) is the suitable structure for the application. Therefore, one-dimensional filters are applied along the spectral curve. Pooling layers reduce the output of a filter layer. CNNs have been recently implemented in remote sensing in terms of classification tasks with spectral images (e.g., [141, 142]) or as 1D CNN (e.g.,[139]).

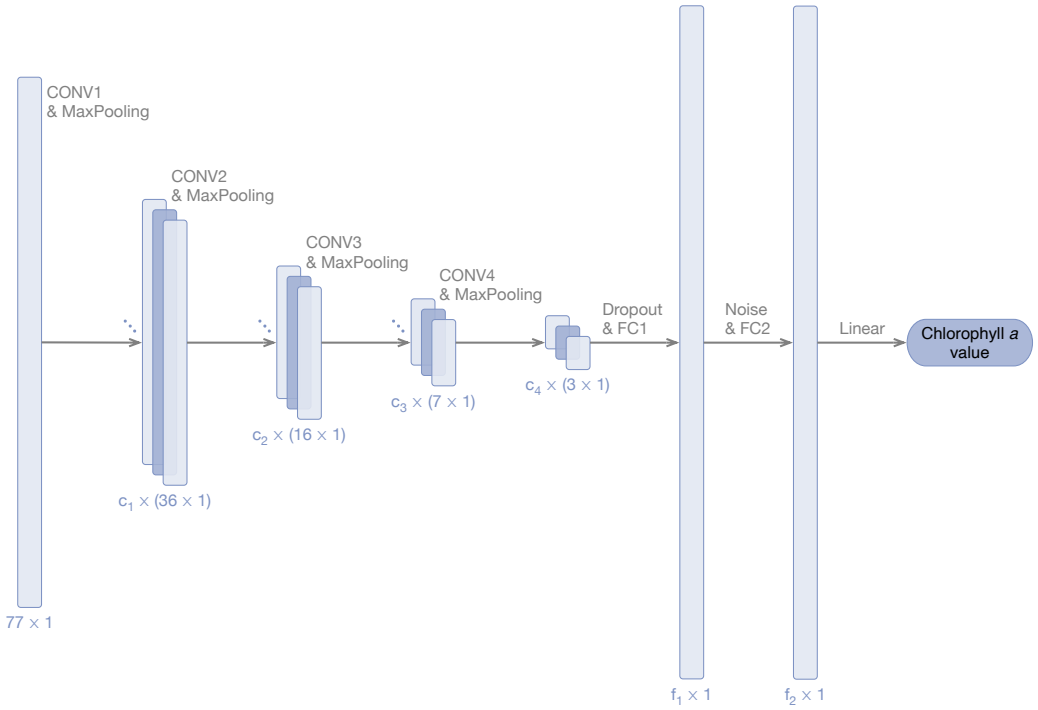

Figure 5.2: Flowchart of the 1D CNN for the SR-EnMAP spectral input features during the training process. The network includes convolutional (CONV), fully-connected (FC), and max-pooling (MaxPooling) layers. Besides, a layer with Gaussian noise (Noise) and a dropout layer (dropout) exist. The I-th CONV layer contains c_i filters and the j-th FC layer contains f_j units. At the end of the network, a linear activation function is applied. Taken from [26], adopted from [139]

5.5 Supervised Learning - Application in the Study Setups

The following section describes the different proceedings in the applied supervised learning approaches for the respective study setups. Besides, technical aspects such as the relying software packages are denoted.

In every study setup, a framework of various supervised learning models is conducted. That means a model is trained and tested for each applied algorithm. To produce comparable results, each of the framework's models follows the same scheme, presented in Figure 5.1. First, the dataset split is executed. This is executed individually for every study setup due to different requirements and will be introduced in the study setup itself. Hereby, splits into two or three subsets are made, and the split ratio varies, depending on the setup and the underlying dataset. It is essential to mention that the training subset is always the same for the framework's models. As a second step, pre-processing is applied to the training dataset and executed on the test and the potential validation dataset. Scaling, a PCA or derivatives are the applied options for pre-processing. If pre-processing is not a part of a study's investigations, the best-performing option is always chosen.

Table 5.1: Summary of the applied models in studies accompanying this thesis, their adjustable hyperparameters and possible pre-processing options. The listed models are the ones trained within the caret environment. N stands for number, min for minimum, max for maximum.

Model	Study Setup	Pre-Processing	Package	Hyperparameter	Description
RF	I, II, III	PCA, Der	`ranger` [131]	mtry min node size	N of variables randomly selected at each node min number of observations at each node
GB	I	PCA, Der	`xgboost` [144]	nrounds max depth eta gamma colsample bytree min child weight subsample	N of boosting iterations max tree depth shrinkage min loss reduction subsample ratio of columns min sum of instance weights subsample percentage
k-NN	I	PCA, Der, Scale	-	k	N of considered neighbors
MARS	I, II	PCA, Der, Scale	`earth` [136]	degree nprune	degree of the polynomial max number of terms
SVM	I, II	PCA, Der, Scale	`kernlab` [145]	gamma cost	kernel parameter penalty factor
ANN	II	Scale	`nnet` [146]	size decay	N of hidden units in the layer weighting factor of the decay

For the training process itself, a technical turning point occurs between **Study Setup I + II** and **Study Setup III**. The trained models in the first two setups do all belong to the shallow learning models and are hence conducted with the `caret` package [143] in `R`. The package then again relies on other packages that provide the functions for the different algorithms. Table 5.1 summarizes the applied shallow learning algorithms, including their possible hyperparameters, their optional pre-processing steps, the belonging to the **Study Setup**, and the package they rely on. The optimal set of hyperparameters of these models are retrieved by a grid search while executing a k-fold CV. To complement the content of Table 5.1, the RF model is always conducted with the splitrule extratree. Besides, the SVM is always executed with a radial-basis kernel.

Contrary to **Study Setup II**, the applied neural networks of **Study Setup III** rely on `R's` version of `tensorflow` [147] and the `keras` [148]. The same applies to the 1D CNN. Switching to this setup is possible due to the vast amount of training data generated by the WASI tool. The 1D CNN architecture that is finally implemented in this thesis is inspired by the LeNet5 network [149] and the LucasCNN [139]. Since estimating chlorophyll *a* is a regression task, especially the output layer and its activation function are adapted as a

difference to the LucasCNN [139]. For the regression, the final layer consists of a single neuron that is activated by a linear activation function. The architecture of the 1D CNN that is applied on the EnMAP resolution in **Study Setup III** is visualized in Figure 5.2 schematically. It consists of four convolutional layers (Conv) with different filters and filter sizes (see Table 5.2). After each Conv layer, a max-pooling layer (MaxPooling) follows, whereas a flatten layer follows the last max-pooling layer. A Gaussian noise (Noise) layer and a dropout layer with a dropout rate of 0.2 are employed to counter overfitting in the training process. The latter follows the flatten layer, whereas the Noise layer is placed between two fully-connected (FC) layers. Both FC layers are located at the end of the network. Except for the last layer, all layers are activated with the commonly applied ReLU function [137]. The `Adam` optimizer [138] is applied to fit the network. During the training process, the MSE represents the loss function. The training process is conducted on a batch size of 256 datapoints in 100 epochs. However, the dropout and the noise layers are only present during the training phase of the 1D CNN. The final adjustment of the hyperparameters for 1D CNN and ANN is illustrated in Table 5.2. A different hyperparameter setup is chosen for each resolution because of the varying amount of input features for the different resolutions.

Table 5.2: Hyperparameters of the 1D CNN (here: CNN) and the ANN with their respective simulated spectral resolutions. The number of filters in the I-th CONV layer is defined as c_i and the number of units in the I-th FC layer is defined as f_j. Taken from [26].

Hyperparameters	CNN + SR-EnMAP	ANN + SP-EnMAP	CNN + SR-Sentinel	ANN + SR-Sentinel
Number of epochs	50	50	100	100
Batch size	256	256	256	256
Kernel size 1	5	-	3	-
Kernel size 2	4	-	2	-
Kernel size 3	3	-	-	-
Kernel size 4	2	-	-	-
Pooling size	2	-	2	-
Activations	ReLU	ReLU	ReLU	ReLU
c_1	128	-	128	-
c_2	128	-	128	-
c_3	256	-	-	-
c_4	256	-	-	-
f_1	200	100	100	100
f_2	200	100	100	100
Dropout	0.2	0.2	0.2	0.2
Loss	Mean squared error			
Optimizer	Adam			

6 Towards Generalized ML Approaches - Study Setups and Evaluation

The primary idea of this thesis is to build generalized ML models for retrieving water parameters with spectral data. Moreover, to identify the performance of various potential approaches for such a model. Generalized models are needed since, on the contrary, specialized models need to be trained explicitly for every water body, which mostly fails due to missing data and considering the enormous number of lakes. Since a later application of such models will be on real satellite data, which may allow consequent monitoring of water ingredients, investigations on the spectral resolution are conducted. The spectral resolution is often limited since its increase comes with a decrease in the spatial resolution (see Section 2.2). A fine spectral resolution for satellite images would be accompanied by a better estimation performance of the retrieval models. However, it is essential to find models that perform suitably on a moderate spectral resolution.

The following chapter represents the research conducted within the scope of this thesis. It is organized in three **Study Setups** (cf. Figure 1.1), which consecutively increases the models' challenges.

The **Study Setups** follow a similar structure. One dataset, described in Chapter 4, consisting of spectral input data and output data, mainly chlorophyll *a*, is evaluated with different kinds of supervised ML approaches. This means a ML regression model is trained on one part of the dataset to estimate the water parameter values, e.g., the chlorophyll *a* concentration. In contrast, its estimation performance is evaluated on the other part (see Section 5.2). Since various ML algorithms perform heterogeneous on different datasets, always multiple approaches are applied in a framework for each setup. The selected algorithms in the framework are presented in Section 5.4. With every additional study, the requirements for the models increase since they more and more need to generalize their learnings according to more challenging datasets. Hence, the studies are ordered in three setups, represented by the applied datasets and the reasonable degree of generalization.

In **Study Setup I**, the relying dataset consists of water samples of only a single water body (see Section 4.1). Most likely, the models trained on that dataset will adapt well to the local constituents. Thus they are specialized to exactly this water body. However, they will likely not be able to perform well on multiple water bodies.

Study Setup II relies on the SpecWa dataset, presented in Section 6.2. Thus, models build on that dataset need to generalize on multiple water bodies and variable conditions. Contrary to the first study, the task for the models in the framework is more challenging since the water parameters described in Section 2.1 vary strongly among the water bodies. However, the models still see some datapoints from each water body in the training process. Thus, the generalization in this setup can only be confirmed for the water bodies in that dataset.

However, **Study Setup III** focuses on exactly the separation between the datasets. In this approach, the datasets are strictly separated. This means the framework's models are trained on the entirely simulated WASI dataset (see Section 4.3), whereas their performance is evaluated on the SpecWa dataset (see Section 6.2). Additionally, a deep learning technique is introduced with the application of a 1D CNN (cf. Section 5.4.7). The idea of such a setup is to find a model that is capable of learning and generalizing the relevant information of the spectrum, completely independent of the evaluation data. Since the measurement setup of the training and the test set differs, slight dataset shifts in terms of the spectral data and the reference data are expected. Hence, the performance of this generalized approach is not expected to be as good as the specialized approaches in the first study.

6.1 Setup I – Models Trained and Applied on the River Elbe Dataset

This chapter includes material from the publication

Philipp M Maier and Sina Keller. "Machine learning regression on hyperspectral data to estimate multiple water parameters". In: *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2018, pp. 1–5. It is cited as Maier and Keller [21] and marked with a red line.

and from the publication

Sina Keller, Philipp M Maier, Felix M Riese, Stefan Norra, Andreas Holbach, Nicolas Börsig, et al. "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity". In: *International journal of environmental research and public health* 15.9 (2018), p. 1881. It is cited as Keller et al. [22] and marked with a may green line.

The following section describes **Study Setup I**. The study's setup is introduced in the beginning, containing the relying dataset, the applied models, and the investigation targets. Then, the results are presented and discussed, and finally set in the context of the RGs. **Study Setup I** affects **RG 1: Models and Parameters** and **RG 2: Feature Importance**.

6.1.1 Setup I – Data Preparation and Study Design

The idea of **Study Setup I** is to evaluate the performance of various ML models in estimating several water parameters on a single water body. The underlying data is the River Elbe dataset, presented in Section 4.1. Thus, the determinable water parameters are chlorophyll *a*, green algae, diatoms, CDOM, and turbidity. The underlying spectral resolution is 4 nm in the boundaries between 470 nm to 910 nm. Hence, it is denominated as a hyperspectral dataset.

Five ML models, k-NN, RF, SVM, MARS, and XGB (see Section 5.4) are selected in the regression framework. They perform the ML regression with the hyperspectral data as input vector and the respective water parameter as target value. Models are built for each parameter combined with every ML algorithm in the framework to estimate the continuous parameter values. The River Elbe dataset, comprising about 1.000 datapoints, is split into a training and a test subset. The training set contains 30 % of the available datapoints, whereas the remaining 70 % is used for the evaluation. All models of the framework are trained on the respective training subset and are evaluated on the respective test subset. For a reasonable estimation of each water parameter, the distributions of the subsets have to be representative of the specific water parameter's distribution. The split ratio, as well as the distinction between training and test subset, decreases overfitting. Figure 4.3 provides histograms of the distribution of each water parameter.

To reduce the dimensionality of the hyperspectral input data, we apply a principal component analysis (PCA) as a pre-processing step. We choose the first eight principal components to fit the models since they cover 99.9 % of the overall variability. Alternatively, the regression framework performs the estimation of the water parameters without PCA, to which we refer as raw bands. With respect to the k-NN, SVM, and MARS regression models, the input data always is scaled [150, 151, 152].

A comparison between the models' performance regarding the underlying algorithm and the respective water parameter is conducted to face **RQ 1**. Finally, the wavelengths considered essential for the RF model are discussed in context with the literature to face **RQ 2**. The **Study Setup I** has to be seen as a first investigation of the potential of different supervised learning models and their ability to estimate various water parameters.

6.1.2 Setup I – Estimation Results of the ML Framework

The regression results of the framework for the estimation of the five water parameters are summarized in Figure 6.1. The regression performance is expressed in terms of the coefficient of determination (R^2) and the root mean squared error (RMSE). The results reveal that the five regression models estimate all water parameters with solely hyperspectral input data

remarkably well. Tables 6.1 to 6.5 contain the regression results sorted by water parameters. In total, RF, SVM, and k-NN models perform better than the MARS and XGB models.

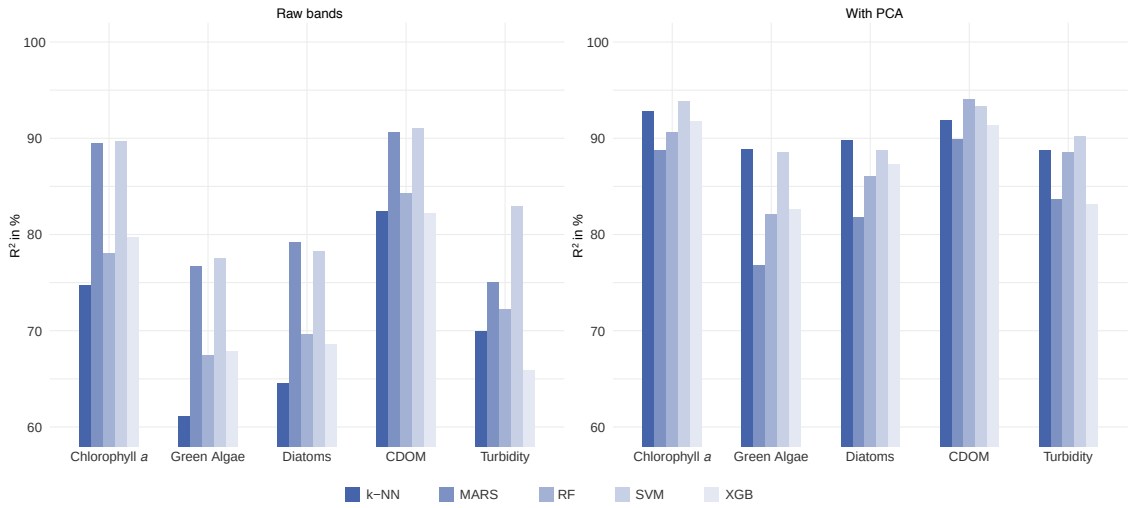

Figure 6.1: Regression results for the estimation of the water parameters expressed as R^2 without PCA (left) and with PCA (right). Taken from [21].

The best estimation performance of each model is obtained by applying the PCA as preprocessing step. The chlorophyll *a* and CDOM concentrations are estimated well by the regression framework (cf. Tables 6.1 and 6.4). Each model reaches R^2 of approximately 90 % with PCA. The estimation performances of diatoms, green algae, and turbidity concentrations achieve $R^2 > 80\,\%$ with PCA (cf. Tables 6.2, 6.3 and 6.5).

The MARS regression model performs worst compared to the others. RF provides the best performances of estimating CDOM while the best estimation of the turbidity is achieved by the SVM regression. The SVM and the k-NN model outperform the other models when estimating the concentrations of chlorophyll *a*, green algae, and diatoms. We point out that outliers in the distribution of the reference data (cf. Figure 4.3) might influence any regression results.

Table 6.1: Regression results of the framework for chlorophyll *a* estimation. Taken from [21].

Model	raw bands		with PCA	
	R^2 in %	RMSE in $\mu g\,L^{-1}$	R^2 in %	RMSE in $\mu g\,L^{-1}$
k-NN	74.7	17.2	92.8	9.1
RF	78.1	16.1	90.7	11.5
SVM	89.7	10.9	93.9	8.3
MARS	89.5	11.2	88.8	11.4
XGB	79.7	15.3	91.8	9.7

Table 6.2: Regression results of the framework for green algae estimation. Taken from [21].

Model	raw bands		with PCA	
	R^2 in %	RMSE in $\mu g\,L^{-1}$	R^2 in %	RMSE in $\mu g\,L^{-1}$
k-NN	61.1	13.2	88.9	7.1
RF	67.5	12.2	82.1	9.4
SVM	77.6	10.0	88.6	7.2
MARS	76.7	10.2	76.8	10.2
XGB	67.9	12.0	82.6	8.8

Table 6.3: Regression results of the framework for diatoms estimation. Taken from [21].

Model	raw bands		with PCA	
	R^2 in %	RMSE in $\mu g\,L^{-1}$	R^2 in %	RMSE in $\mu g\,L^{-1}$
k-NN	64.6	11.5	89.8	6.2
RF	69.7	10.9	86.1	8.1
SVM	78.3	9.0	88.8	6.5
MARS	79.2	8.9	81.8	8.2
XGB	68.6	10.8	87.3	7.0

Table 6.4: Regression results of the framework for CDOM estimation. CDOM is measured in parts per billion, $ppb_{QS} = 10^{-9}$ and is calibrated against Quinine Sulfate (QS). Taken from [21].

Model	raw bands		with PCA	
	R^2 in %	RMSE in ppb_{QS}	R^2 in %	RMSE in ppb_{QS}
k-NN	82.4	0.9	91.9	0.6
RF	84.3	0.8	94.1	0.5
SVM	91.1	0.6	93.3	0.5
MARS	90.7	0.6	89.9	0.7
XGB	82.2	0.9	91.4	0.6

Figure 6.2 exemplifies the regression results of the ET model compared to the real probe measurements matched with their respective recorded GPS data along the Elbe. In addition, plots in the right columns represent the min-max scaled deviation Δ_{scaled} between the measured (meas) and the estimated (est) values. The scaled deviation Δ_{scaled} allows the comparison of the estimation performance of all water quality parameters. We define Δ_{scaled} according to Equation (6.1). Eventually, this results in a measure independent of the unit and the range of the target variable.

Table 6.5: Regression results of the framework for turbidity estimation. The turbidity is measured in Formazin Turbidity Unit (FTU). Taken from [21].

Model	raw bands		with PCA	
	R^2 in %	RMSE in FTU	R^2 in %	RMSE in FTU
k-NN	70.0	0.3	88.8	0.2
RF	72.3	0.2	88.6	0.2
SVM	83.0	0.3	90.2	0.2
MARS	75.1	0.3	83.7	0.2
XGB	65.9	0.3	83.2	0.2

$$\Delta_{\text{scaled}}(x) = \frac{X_{\text{est}} - \min(X_{\text{meas}})}{\max(X_{\text{meas}}) - \min(X_{\text{meas}})} \tag{6.1}$$

Regarding the estimation of chlorophyll *a*, the ET model underestimates the chlorophyll *a* concentration at the beginning of the Elbe field campaign (cf. Figure 6.2, second row). Over the central profile and the end of the field campaign, an overestimation occurs. With respect to the green algae and diatoms concentrations (cf. Figure 6.2, third and fourth row), similar conditions in the course of the Elbe field campaign can be deduced regarding the over- and underestimation of the regressor. As for the turbidity (cf. Figure 6.2, last row), the models underestimated this parameter at the beginning of the field campaign. Later along the track of the Elbe, it changes to an overestimation. For CDOM (cf. Figure 6.2, first row), the deviation is generally less distinct, but some outliers, especially in the middle and the end of the track, occur.

6.1.3 Setup I – Feature Importance of the ET Model – Results

In Figure 6.3 we show the feature importance distributions for the hyperspectral input data of all water quality parameters generated by ET without pre-processing. Additionally, the hyperspectral mean spectrum with standard deviation is included. To derive statements of the hyperspectral input data, we need a feature importance distribution of the raw bands without pre-processing. Furthermore, we choose the ET model due to its technically integrated variable importance function combined with a suitable performance on the raw bands. Except for the green algae distribution, the four other distributions are characterized by similar peaks: one more significant peak at around 735 nm and one smaller peak at around 680 nm. The first peak at around 680 nm is only weakly pronounced for the green algae, whereas the second peak shifts towards the longer wavelengths at around 820 nm. Besides the two peaks in the middle of the spectrum, CDOM shows one distinct

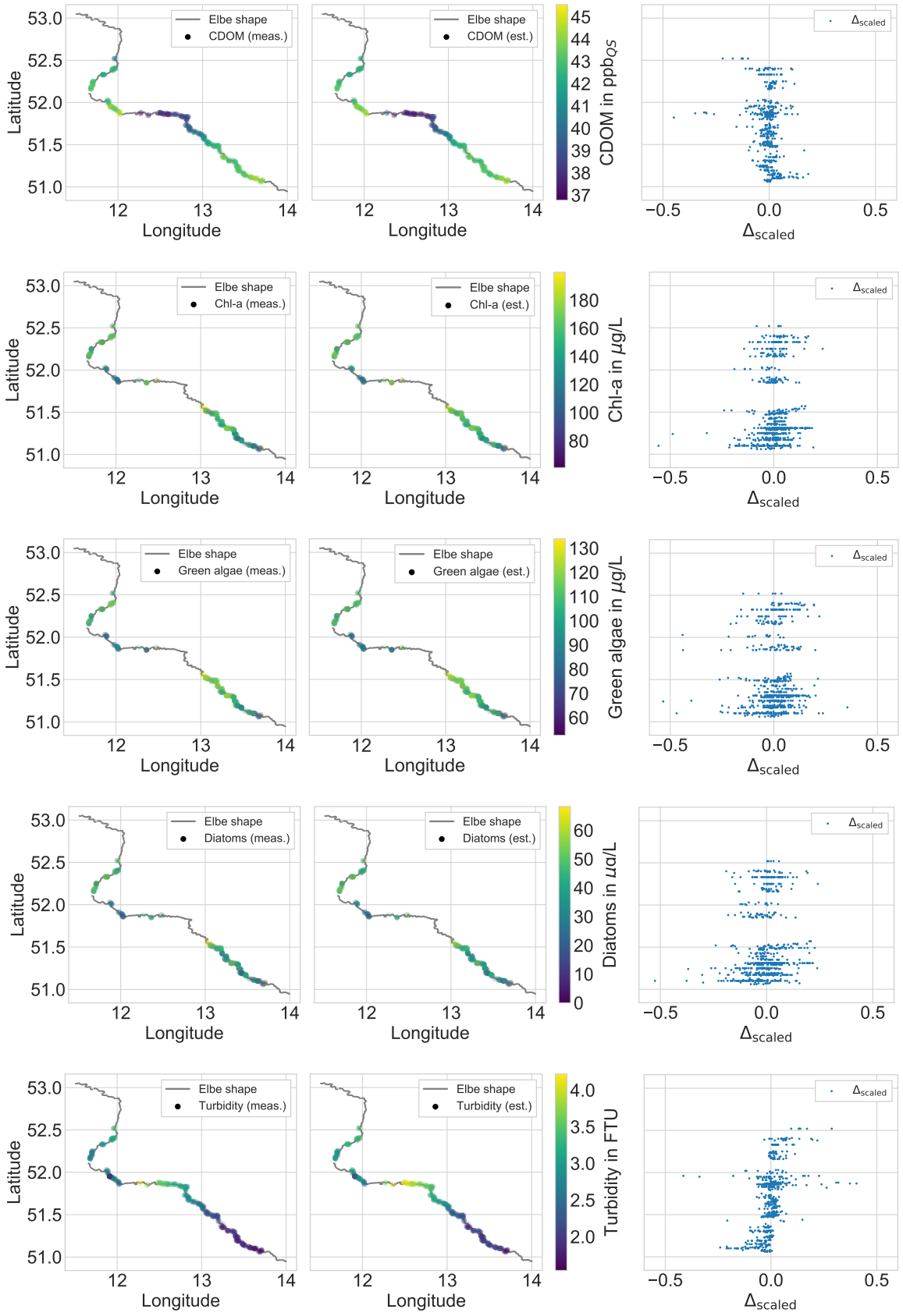

Figure 6.2: Visualization of the regression results generated by the ET model (central columns) compared to the real probe measurements (left columns) matched with their respective recorded GPS data along the river Elbe. The min-max scaled deviations Δ_{scaled} between the measured and the estimated values of the water quality parameters are illustrated in the right columns. We refer to the chlorophyll *a* concentration in this plot as Chl-a. In this context, ET and RF are synonymous. Taken from [22].

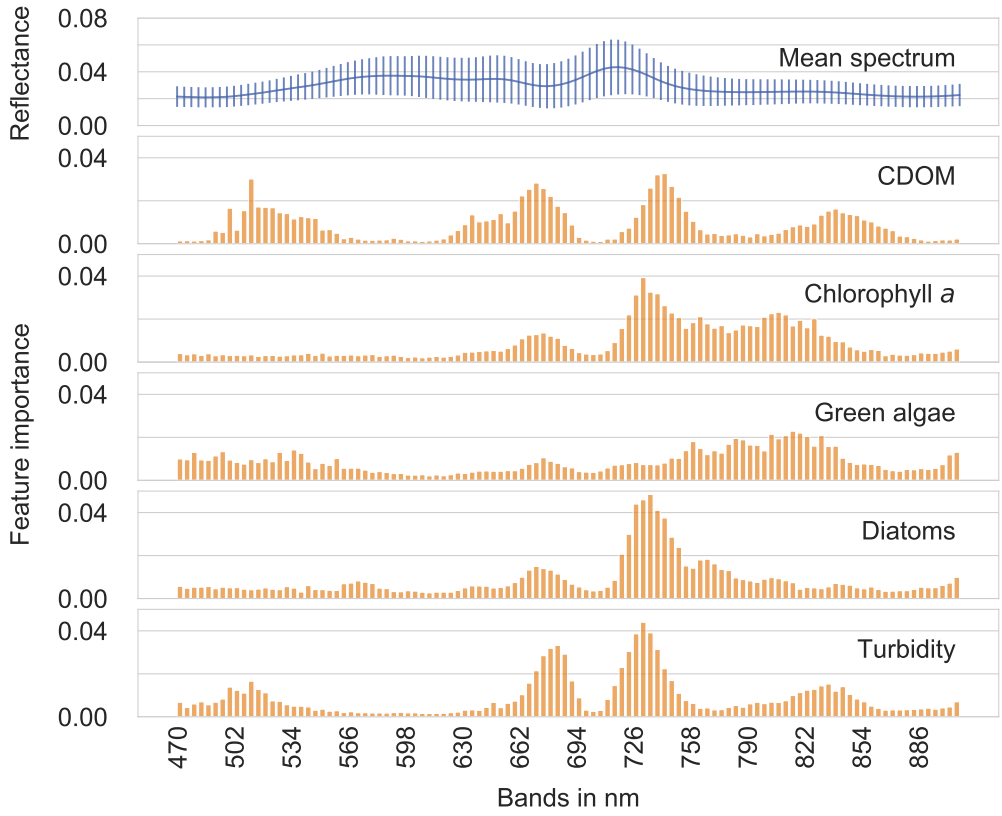

Figure 6.3: Feature importance of the ET regressor without pre-processing (baseline). The upper plot represents the mean spectrum of the hyperspectral data. Taken from [22].

peak in the shorter wavelength ranges and one in the longer wavelengths. This is similar to the shape of turbidity's feature importance distribution.

6.1.4 Setup I – Feature Importance of the ET Model – Discussion

To discuss and understand the features selected by the ET models for the respective parameter, we relate them with the physical properties of the parameters. First of all, from the physical point of view, we would expect similar selected features for chlorophyll *a*, green algae, and diatoms. They should occur at the specific absorption feature at around 670 nm and the global maximum at 710 nm, where only little absorption on the water takes place (see, e.g., [37] and Section 2.1.2). Additionally, most of the empirical models for retrieving chlorophyll *a* also rely on both features. Alternatively, another important feature would be the local maximum at around 560 nm related to low absorption by algal materials.

To distinguish green algae and diatoms, we would expect to see important features between 520 nm to 570 nm due to the findings by Hunter et al. [36] related to different pigments (see Section 2.1.2).

Since turbidity is highly related to chlorophyll *a*, especially for water bodies with high chlorophyll *a* concentrations, such as the River Elbe, we would also assume multiple correlated features. For all measured datapoints of the River Elbe, the linear correlation between chlorophyll *a* and turbidity is $r^2 = 48\,\%$. Additionally, another essential feature for turbidity is expected to be at around 810 nm due to scattering on detritus (see [56], Section 2.1.2).

However, for the CDOM estimation, we expect entirely different importance features compared to chlorophyll *a* due to its absorption at the shorter wavelengths.

When comparing the expectations from the physical point of view with the obtained results by the ET model's feature importance, we derive different findings. Firstly, two parameters perfectly match our expectations, whereas secondly, we are surprised by the model's choice for three parameters. Chlorophyll *a* and turbidity belong to the parameters that match our expectations. Both parameter rely on the relevant features for chlorophyll *a* that are also used for the band ratio approaches, presented in Section 3.2. Besides the important features, the non-important features for chlorophyll *a* also match the physical expectations. Only the peak at about 810 nm does not match our expectations. One possible explanation might be the relation to degraded algal materials, which might correlate with high chlorophyll *a* concentrations. The peak is also pronounced for the turbidity model that fits our expectations related to the scattering peak at about 810 nm, but with a slight shift towards the longer wavelengths.

The expected distinguishing of the two phytoplankton classes in the green wavelengths between 520 nm to 570 nm related to their pigments was missed. For green algae, the ET model shows higher importance for this spectral region. Contrary to diatoms, merely, a slight increase at around 570 nm exists, but it seems to be insignificant compared to the importance of the other features. Besides, the feature importance for diatoms follows more or less the ones of chlorophyll *a*, except for the last peak at 810 nm. This peak is followed in turn by green algae. It seems reasonable that green algae do not have the same importance features as chlorophyll *a* at a second glance. The diurnal variations during the measurement campaign reveal increasing chlorophyll *a* concentrations till afternoon, but the ratio between green algae and diatoms decreases in that period. This relation may indicate that applying similar features to retrieve chlorophyll *a* seems possible for diatoms but not for green algae.

For the retrieval of CDOM in the River Elbe dataset, the ET model's selection of the important features seems to be astonishing at first glance. The most important features are the ones affected by chlorophyll *a*. Therefore, no physical relation exists. However, one explanation for that finding might be that the models need to understand the varying chlorophyll *a* concentration to conclude CDOM values. The third peak in the ET model's feature importance can be directly related to CDOM's high absorption in the shorter blue wavelength.

The surface reflectance impacts the spectrum, and in turn, it might be essential for the feature importance. Imagine the atmospheric conditions described in Section 2.1.1 influencing the surface reflectance. This often occurs as a simple y-offset. Thus, it would be

possible that some models use irrelevant features for the water ingredients to, e.g., normalize the spectrum and hence, decrease the impact of surface reflectance. However, regarding this study setup, this cannot be proven.

6.1.5 Setup I – Parameter and Model Performance – Discussion

In this section, we discuss the results from two different points of view. First, we focus on the performance from the view of the water parameters. Second, regarding the algorithms and the performance difference between the PCA-based and the raw bands-based models.

Regarding the water parameters, the models for CDOM and chlorophyll *a* perform best, whereas the models for both phytoplankton species perform a bit worse, as well as the models for turbidity. One explanation for this finding might be that the chlorophyll *a* concentration is a composition of both algae classes. Additionally, it is a pigment with distinct absorption features. Thus it seems to be easier for the models to retrieve only chlorophyll *a*. Most likely, it is more difficult to relate the chlorophyll *a* concentration to the related algae classes. In the literature, retrieving CDOM is the most challenging task in remote sensing of inland waters [48]. Therefore, the model's excellent performance is a bit surprising. Nevertheless, the models seem to adapt very well to the local data.

Focusing on the models, we find the apparent trend that dimensionality reduction to only eight principal components leads to a substantial performance increase. This finding might be related to overfitting of the models relying on the raw bands. Thus most of the models cannot handle the high dimensional input data. From the raw band models, two specific outperform the other three. On the one hand, this concerns the MARS model and, on the other hand, the SVM.

The MARS model's outperformance on the raw bands might be due to its conceptions described in Section 5.4.4, containing a forward and a backward path, as well as pruning. It does not use the actual input features but creates mirrored functions. Thus, it better addresses overfitting-susceptible datasets. In turn, MARS cannot learn the whole complexity of the data, resulting in an underperformance on the PCA-based dataset.

The SVM shows the best performance on the raw dataset. This might be related to its tuning parameters that can be better adjusted to prevent overfitting than the other models Section 5.4.3. Thus, the SVM might generalize better and still achieve R^2 values of up to 90 % on the raw band dataset.

The poor performance of the k-NN on the raw dataset seems to be also an overfitting problem. It is the most simple algorithm that relies on similarities of the datapoints. Therefore, outliers can have a vast impact on the performance.

Regarding the estimations of the parameters along the course of the River Elbe, there seems to be no section where the ET models thoroughly perform badly. Thus, we assume that the model's generalization on the dataset is suitable, and overfitting does not dominate the process.

6.1.6 Setup I – Research Goal 1: Models and Parameters

The conducted study focused on the general applicability and the retrievable water parameters with multiple supervised ML models. All five applied ML models achieve a satisfactory estimation performance on every investigated water parameter. However, the dimensionality reduction on this dataset is crucial for good and stable estimation results. This finding may indicate, that the models tend toward overfitting. However, they can learn the data of the underlying dataset. In general, the models seem to be highly specialized. Even the shallow k-NN model achieved a good performance.

Regarding the estimation performance from the water parameters' point of view, CDOM and chlorophyll *a* are the best-performing ones. The models can even distinguish between phytoplankton species. However, it does not mean that it is possible in general on multiple water bodies.

Our regression framework serves as starting point for further investigations like adapting the framework to estimate water parameters of different types of inland waters. As a consequence of a similar performance between the GB models and the RF models, despite more effort in the training phase with multiple hyperparameters (see Table 5.1), we dismiss the GB models from the framework.

To answer the **RQ 1**, the models provide a suitable estimation performance on every applied parameter. Concerning this study, the SVM produces the most promising models. Additionally, regarding the ET models' important features, there seems to be too much information in the spectrum left, to only rely on several selected bands as proposed by the empirical models in the literature. This fortifies our ML approach in working with all available input features.

6.1.7 Setup I – Research Goal 2: Feature Importance

Regarding the selected features by the models concerning its respective water parameter, the physical relations are not always apparent. Especially the differentiation between both two algae classes is not comprehensible with a physical background. However, for chlorophyll *a* and turbidity, the selected features of the models seem reasonable with the physical background. Additionally, the selected features for CDOM can be explained, but only at a second glance. To sum up, for chlorophyll *a*, the most important parameter,

the model clearly selects spectral features that can be physically related to the optical properties of the parameter. Thus, for this study setup, **RQ 2** can be answered with yes. The models mainly rely on physically relevant spectral features. However, we have to keep in mind that these models are specialized on the River Elbe dataset. Thus it is possible that some overfitting may occur, so in turn, the models might apply features that are not transferable to other datasets. Hence, we expect to find more distinguishable features in more generalized models, relying on multiple datasets.

6.2 Setup II – Models Trained and Applied on the SpecWa Dataset

This section includes material from the publication

Philipp M Maier and Sina Keller. "Estimating chlorophyll a concentrations of several inland waters with hyperspectral data and machine learning models". In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4 (2019). It is cited as Maier and Keller [23] and marked with a blue line.

and from the publication

Philipp M Maier and Sina Keller. "Application of different simulated spectral data and machine learning to estimate the chlorophyll a concentration of several inland waters". In: *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2019, pp. 1–5. It is cited as Maier and Keller [24] and marked with a green line.

The general idea of **Study Setup II** is to estimate the chlorophyll *a* concentrations of multiple inland water bodies with the ML framework. Therefore, we rely on the 2018 part of the SpecWa dataset, presented in Section 6.2. Concerning the RGs, we meet the same as in the previous **Study Setup I** (see Section 6.1), but the focus is solely on chlorophyll *a*. In addition, we approach the generalization with the inclusion of multiple water bodies provided by the SpecWa dataset (**RG 3: Generalization**). Contrary to **Study Setup I**, the models must show that they can transfer the learnings from the training dataset to multiple water bodies' datapoints to face the generalization challenge. Moreover, we investigate the impact of the spectral resolution on the estimation performance of the models (**RG 4: Spectral Resolution**).

Study Setup II is organized into three sub-setups, relying on the SpecWa dataset presented in Section 6.2. The first two sub-setups are similar to each other. We investigate both approaches in an integrated manner. Their primary investigation is the combined analysis of the spectral resolution and the applied ML models in the framework. The

first relies on different continuous hyperspectral resolutions, whereas the second uses the spectral resolution of actual satellite missions.

The third sub-setup bases on the full SpecWa dataset. It focuses on the feature importance of the RF model concerning a generalized model for the whole dataset. Besides, we investigate models that represent only selected individual water bodies, representing more specialized models (**RG 2: Feature Importance**). This approach is similar to the feature importance investigation in **Study Setup I** (see Section 6.1.3).

6.2.1 Setup II/1 – Comparison of Different Continuous Hyperspectral Resolutions

For the investigations of **Study Setup II/1** and later **Study Setup II/2**, we rely on the 2018 part of the SpecWa dataset. This is because the respective studies are published before the collection of the 2019 part (see [23, 24]). In the 2018 part of the SpecWa dataset, a small number of datapoints from two additional water bodies accrue that have not later been published in the SpecWa dataset (see [25]) due to inconsistent measurements.

For the first investigation, we split the dataset into two equal-sized halves while considering the distribution of the chlorophyll *a* concentration in both sets. The first part is used to train the models, whereas the second part is held out for their evaluation. We illustrate the dataset according to the chlorophyll *a* concentrations with its respective split in Figure 6.4. The dataset consists of measurements at 13 water bodies with a varying sampling frequency, visualized in Figure 6.5. Regarding both distributions in Figure 6.4 and Figure 6.5, it will be a challenging task for the ML models since they usually prefer an equally distributed and represented data basis. Artificial ponds occur more frequently than natural water bodies. Since we rely on real-world data, a uniform distribution is not feasible without removing too many datapoints.

In **Study Setup II/1**, the spectrometer data of the SpecWa dataset is aggregated to continuous intervals of 4 nm, 8 nm, 12 nm, and 20 nm, representing the new spectral features in the range from 400 nm to 900 nm (see Section 4.4). On every resolution, each of the framework's models is trained to investigate the impact of the different resolutions on the models' performance (cf. Section 5.5). The investigation on the resolution is conducted for practical reasons since sensors for monitoring water bodies area-wide obtain a less spectral resolution than the spectrometer. Additionally, we analyze the effect of applying the first derivative of the spectrum. Thus, we compare the performance of the models relying on the first derivative bands to the models relying on the raw bands. We no longer apply the PCA as a pre-processing step since models based on the principal components underperformed the others in a preliminary investigation. For the investigation of **Study Setup II/1**, the

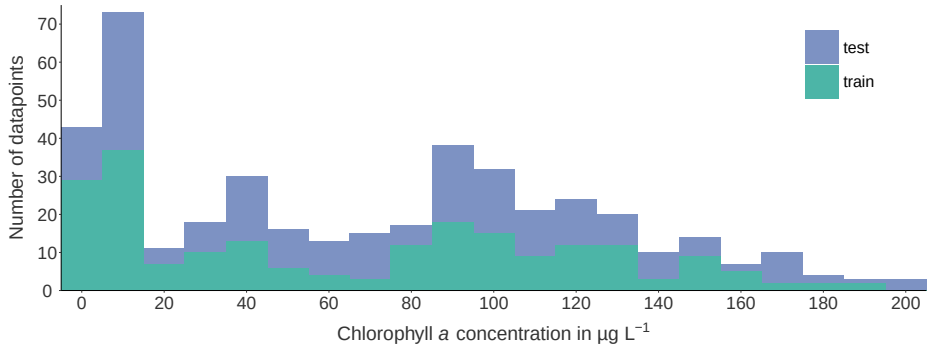

Figure 6.4: Histogram of the chlorophyll *a* concentration of all datapoints of the 2018 part of the SpecWa dataset. The different colors indicate the split in the training and the test set. For this split, the concentration is considered, leading to a similar shape of each set's histogram. Taken and adapted from [23].

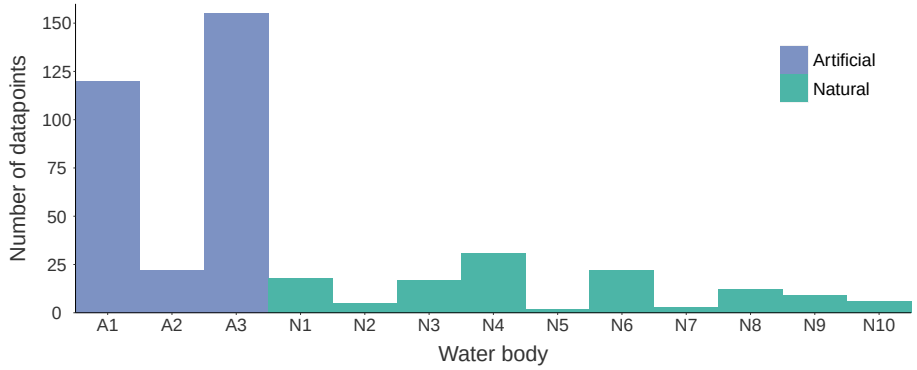

Figure 6.5: The number of datapoints per water body in the 2018 part of the SpecWa dataset. The blue color indicates an artificial water body with high chlorophyll *a* concentrations, whereas the green color represents more natural water bodies with lower chlorophyll *a* concentration. Taken and adapted from [23].

supervised ML framework consists of RF, SVM and ANN. The models should link the spectral information of unknown water samples to continuous chlorophyll *a* values.

6.2.2 Setup II/1 – Continuous Hyperspectral Resolutions – Results

Tables 6.6 to 6.9 present the estimation results of the regression models with different spectral resolutions on both datasets, the raw dataset (raw) and the dataset with the derivatives (der). We expected to obtain a better estimation performance with a higher spectral resolution. This hypothesis could not be proved. Even though we achieved better results on both datasets with higher resolution than on the datasets with the lower resolution, it could not be generalized that a higher resolution must lead to a better estimation performance. For the best performing regression model on the 20 nm-dataset, the ANN model, the R^2 score

of 87.1 % is only 2 % lower than on the 4 nm-dataset. The results for the 4 nm-dataset and the 8 nm-dataset seem to be more or less the same. Whereas the regression performance is better on the 20 nm-dataset than on the 12 nm-dataset.

Table 6.6: Regression results for chlorophyll *a* estimation with 4 nm spectral resolution for the raw dataset (raw) and the dataset with derivatives (der). Taken from [23].

	Model	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
raw	RF	81.7	22.5	15.9
	SVM	86.8	19.1	13.5
	ANN	89.0	17.7	12.0
der	RF	87.1	19.0	12.9
	SVM	85.2	20.5	14.9
	ANN	85.2	20.7	14.8

Table 6.7: Regression results for chlorophyll *a* estimation with 8 nm spectral resolution for the raw dataset (raw) and the dataset with derivatives (der). Taken from [23].

	Model	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
raw	RF	77.5	24.3	16.5
	SVM	88.1	17.8	12.3
	ANN	88.3	17.5	12.1
der	RF	87.8	18.1	12.0
	SVM	88.6	17.5	12.5
	ANN	89.2	17.0	11.9

Table 6.8: Regression results for chlorophyll *a* estimation with 12 nm spectral resolution for the raw dataset (raw) and the dataset with derivatives (der). Taken from [23].

	Model	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
raw	RF	76.4	25.5	17.2
	SVM	82.4	22.2	14.5
	ANN	82.6	22.1	14.6
der	RF	83.3	21.4	13.3
	SVM	83.5	22.0	14.6
	ANN	81.5	22.9	13.9

Regarding the effects of derivatives by comparing the upper and the lower half of the Tables 6.6 to 6.9, we can observe that the RF model experiences the strongest influ-

Table 6.9: Regression results for chlorophyll *a* estimation with 20 nm spectral resolution for the raw dataset (raw) and the dataset with derivatives (der). Taken from [23].

	Model	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
raw	RF	79.4	24.4	16.9
	SVM	81.4	23.2	15.2
	ANN	87.1	19.4	13.2
der	RF	84.7	20.9	12.5
	SVM	85.2	20.7	14.3
	ANN	87.1	19.5	13.0

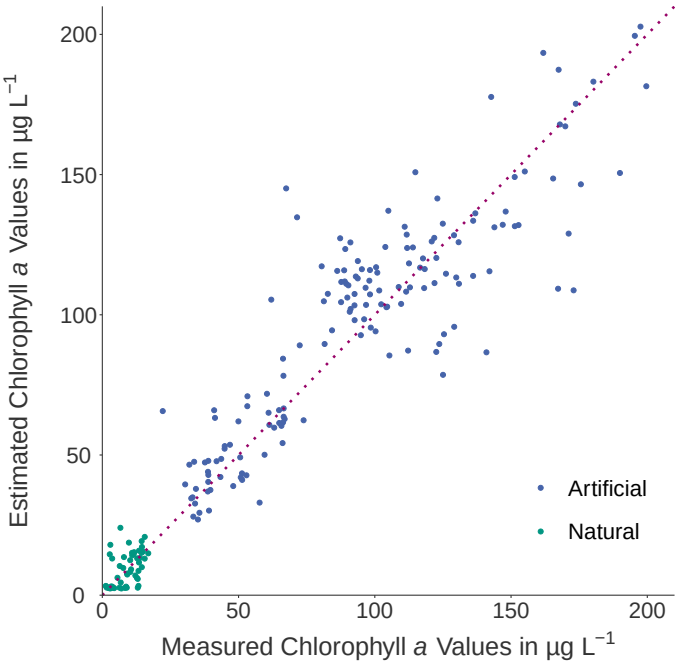

Figure 6.6: Example of a scatterplot of the ANN regressor showing the estimated vs. the measured chlorophyll *a* concentration for the 4 nm spectral resolution. The color defines the status of the water body: natural waters (green) and artificial ponds (blue). The red dotted line indicates the perfect regression line. Taken and adapted from [23].

ence by derivatives. Improvements between 5 % and 10 % in the R^2 score for all resolutions are reached. For the SVM, we notice slight improvements three times by different resolutions with the derivatives, but the effect is not convincing. Applying the ANN model on the dataset with the derivatives, the effect is reversed compared to the SVM due to the specific characteristics of an ANN.

Comparing the estimation results between the three machine learning models, ANN shows the best performance for the raw bands and overall. The best performance is achieved for

the 8 nm-dataset with an R^2 of 89.2 %. RF and SVM demonstrate a slightly worse estimation performance. Both models are in a range of 1 % R^2 to each other for every spectral resolution.

Figure 6.6 visualizes the estimation result from the ANN model for every datapoint of the test dataset with the 4 nm spectral resolution. In general, we can recognize the regression line distinctly. However, some points exist, which are estimated poorly.

6.2.3 Setup II/1 – Conclusions and Adaptions for Study Setup II/2

To sum up the results of **Study Setup II/1**, we are satisfied by the performance of all applied models in the framework. The performance metrics for chlorophyll *a* are slightly worse than in **Study Setup I**, but the models' task in estimating the values of different water bodies was more challenging. Moreover, we recognized that the resolution for all four hyperspectral datasets is not as important as we would have thought for the estimation performance when designing this setup. The RF model improved the estimation performance noticeably when relying on the first derivative of the spectrum.

Encouraged by the findings in **Study Setup II/1**, we decided to investigate further the impact of the spectral resolution on the models' estimation performances. Therefore, **Study Setup II/2** has to be seen as an add-on to the previous setup. It is based on the same dataset with the same 50 : 50 training and test split (see Figure 6.4) and the same model types. Instead of using continuous hyperspectral features, the focus of **Study Setup II/2** is to even more widen the bandwidths concerning actual satellite resolutions. Hence, the spectrometer resolution is downsampled to different satellite missions, including hyperspectral resolutions, as well as multispectral resolutions with broad bands as, e.g., for Landsat 5. The simulated satellite resolutions comprise the two hyperspectral sensors on EnMAP and Hyperion, the multispectral ESA missions Sentinel-2 and Sentinel-3, and the multispectral NASA missions Landsat 5 and Landsat 8. Their characteristics are summarized in Table 4.5. The scaling approach for the satellite missions was explained in detail in Section 4.4. Since for most of the simulated satellite resolutions, the amount of bands decreases and the bandwidth increases compared to **Study Setup II/1**, a decrease in the estimation performance of the models is expected. The ML framework is the same as in **Study Setup II/1**, but we added the MARS model from **Study Setup I**. In this setup, we no longer investigate the first derivative's effect on the performance. However, for the hyperspectral datasets, we use the first derivative for training the RF models.

6.2.4 Setup II/2 – Comparison of Different Satellite Resolutions – Results

Figure 6.7 and Table 6.10 present the regression performance of estimating the chlorophyll *a* concentration with respect to the applied ML models as well as the different

simulated satellite data. Regarding Figure 6.7, the regression performances of the four ML models are all in the same range.

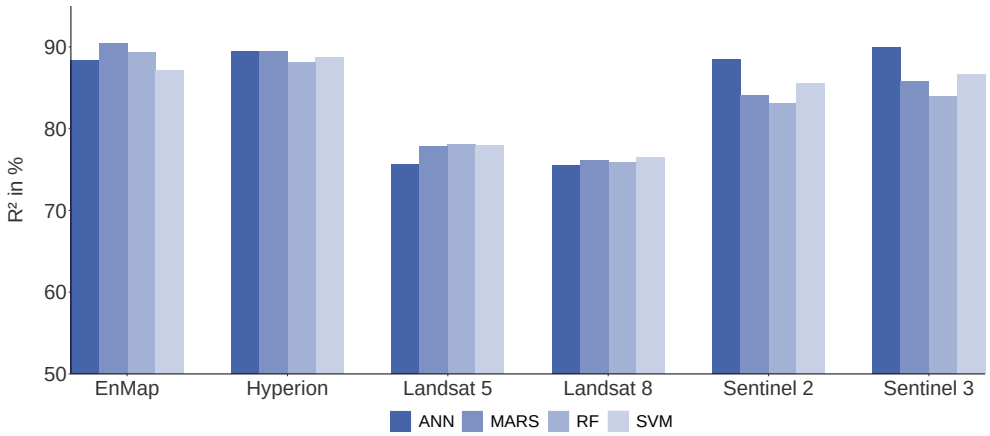

Figure 6.7: Regression results (R^2 in %) of the four ML models in combination with the different simulated satellite data. Taken and adapted from [24].

When considering the simulated satellite input data for estimating the chlorophyll *a* concentration, the regression results expressed as R^2 are distinguishable. For the simulated hyperspectral satellite data (EnMAP and Hyperion), the coefficient of determination (R^2) is quite similar. In the case of the simulated Landsat data, the regression results are closely related. In detail, the ANN model performs worse than the other three models on these two simulated datasets. However, for the simulated Sentinel data, the ANN model provides the best regression results.

Table 6.10: Performance of the regression models expressed by MAE in $\mu g\,L^{-1}$. Taken from [24].

Simulated satellite data	RF	SVM	ANN	MARS
EnMAP	10.9	12.6	11.7	10.1
Hyperion	11.3	12.2	11.3	10.5
Landsat 5	17.8	18.5	19.6	19.0
Landsat 8	18.8	18.8	20.0	20.5
Sentinel 2	14.8	13.2	11.5	14.2
Sentinel 3	14.3	14.1	10.9	13.0

Considering the different simulated satellite data, the regression with the simulated hyperspectral data based on the EnMAP and Hyperion mission achieves the best results. The corresponding MAE values range between $10.1\,\mu g\,L^{-1}$ to $12.6\,\mu g\,L^{-1}$ (see Creftab:MAE). The MAE values of the models with simulated multispectral data according to the Sentinel missions are in the range between $10.9\,\mu g\,L^{-1}$ to $14.8\,\mu g\,L^{-1}$. The estimation of the chlorophyll *a* concentration of all regression models with simulated Landsat data performs the worst compared to the other simulated satellite data. Thereby, the MAE ranges between $17.8\,\mu g\,L^{-1}$ to $20.5\,\mu g\,L^{-1}$.

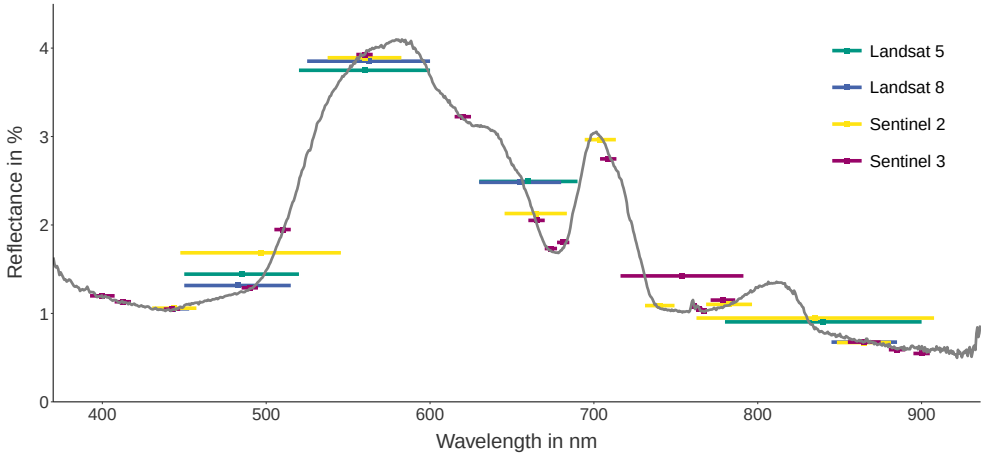

Figure 6.8: The figure shows a continuous spectrum (grey) of the SpecWa dataset. Contrary, the colored lines represent the bandwidth of four multispectral satellite missions. The colored point represents the actual reflectance value as the central wavelengths of the spectral band. The figure explains the challenges for some satellite missions in representing the continuous spectrum. Taken and adapted from [24].

6.2.5 Setup II/2 – Comparison of Different Satellite Resolutions – Discussion

Analyzing the bandwidth, the number of bands, the spectral range, and the resolution of the simulated satellite data, Figure 6.8 shows that Landsat 5 (green) and Landsat 8 (blue) have similar bands with a similar band positioning. The three bands between 450 nm to 700 nm are nearly the same. In the spectral range of 800 nm to 900 nm, Landsat 8 provides a narrower band than Landsat 5, and it has an additional fifth narrow band near 430 nm. With respect to the estimation of the chlorophyll *a* concentration, this additional band has no further impact on the regression task. This is remarkable since we thought that with only four bands, every additional information provided by a fifth band would significantly increase the models' performance.

Similar to both simulated Landsat data, the simulated multispectral Sentinel-3 data provides a better spectral resolution and accounts for more bands with narrower bandwidths than Sentinel-2. Especially in the essential range for chlorophyll *a* retrieval between 660 nm to 710 nm, Sentinel-3 provides more spectral bands.

However, the regression performance of the ML models on simulated Sentinel-3 data is not clearly better than the regression performance of the models with simulated Sentinel-2 data. When comparing the estimation performance with either simulated Sentinel data or simulated Landsat data, the outperformance of the models using the simulated Sentinel data can be well explained. First, the simulated Sentinel data is characterized by more bands. And second, these bands are well positioned within the spectral range of 400 nm to 900 nm. For example, the simulated Sentinel data includes the extremes in

the range of 660 nm to 710 nm, which are related to chlorophyll *a* (see, e.g., [37]). The mentioned spectral range is not included in the two Landsat missions and explains the poor chlorophyll *a* estimation of all models (see, e.g., [63]).

The simulated hyperspectral data (EnMAP and Hyperion) with a nearly constant spectral resolution of 6.5 nm and 10 nm are not shown in Figure 6.8 due to reasons of transparency. Comparing the regression results with the simulated hyperspectral and the simulated Sentinel data, the models relying on the hyperspectral datasets perform only slightly better. This finding indicates that the band positioning of the Sentinel missions is good for the estimation of chlorophyll *a* concentrations.

Regarding the applicability of the simulated satellite data for a general monitoring approach in the context of inland waters, the Sentinel-2 data serves its purpose. It provides data with appealing spectral resolution and a sufficient spatial resolution characterized by a high temporal frequency. Hyperspectral data with a better spectral resolution leads to a satisfying chlorophyll *a* estimation by applying the same ML models. However, their temporal resolution stays behind the temporal resolution of the Sentinel missions referring to two satellite systems. Differentiating between the two Sentinel missions, the application of the Sentinel-3 satellites is limited to large inland water surfaces due to their poor spatial resolution of 300 m to 1000 m. In addition, the Landsat satellite missions provide an attractive spatial and temporal resolution as well. However, the regression results of the models are the worst with this data since the Landsat missions are characterized by the lowest spectral resolution of all simulated satellite missions.

6.2.6 Setup II/3 – Feature Importance for the EnMAP Resolution – Description

To investigate the feature importance concerning multiple water bodies, we rely on the originally published SpecWa dataset containing eleven water bodies and 3.685 datapoints showing a chlorophyll *a* range from $0\,\mu g\,L^{-1}$ to $200\,\mu g\,L^{-1}$. From experience in **Study Setup II/2**, we select the EnMAP resolution consisting of about 6.5 nm broad channels for this investigation. To gain the feature importance, we rely again on the RF models. Therefore, five RF models are trained. First, one model representing the whole SpecWa dataset as a generalized model. Then, models for the water bodies with the most datapoints: ap castle garden Karlsruhe, qp Blankenloch, qp Waldstadt, and ap Kit. These models are trained on all respectively available datapoints. The idea of such an approach is again to find relations between the physical properties of the water ingredients and the models' feature choice. We expect a similar feature distribution as in **Study Setup I - RQ 2** for the specialized models only trained on a single water body. However, we expect more distinct features for the general model trained on all of the eleven water bodies.

6.2.7 Setup II/3 – Feature Importance for the EnMAP Resolution – Results

Figure 6.9 illustrates the selected important features of the five RF models. The spectrum visualizes the mean and the shape represents the standard deviation for the respective subset. The orange bars are normalized between 0 and 1 and stand for the models' important features. Regarding the model based on the whole dataset, nearly every selected feature lies between 700 nm to 800 nm. For the models build on single water bodies, the distribution of the selected features is more widespread. The model for the artificial pond at the castle garden in Karlsruhe selects important features in between 500 nm to 750 nm. The one for the quarry pond in Blankenloch selects important features at around 670 nm and 720 nm but also around 800 nm. The quarry pond Waldstadt's model relies on features between 500 nm to 720 nm. However, the model for the artificial pond at KIT shows distinct features at 720 nm and around 800 nm.

6.2.8 Setup II/3 – Feature Importance for the EnMAP Resolution – Discussion

Regarding the selected features for the generalized model covering eleven water bodies, most of them can be related to the physical parameters. The peak at 710 nm is related to scattering on phytoplankton materials. Moreover, this peak occurs due to high absorption on water in the longer wavelengths and on chlorophyll a on the shorter wavelengths. Thus, it is related to chlorophyll *a* absorption itself. Besides, regarding the chlorophyll *a* absorption at 670 nm, the mean value for all five spectra is approximately the same. This might be the reason why the models do not pick the absorption feature to determine the chlorophyll *a* concentration. Additionally, regarding the standard deviation, the magnitude at the scattering peak is stronger than on the absorption feature.

The indicator for the reflectance peak at 810 nm is scattering on detritus materials [56] that also contains degraded algal material. Especially in the later season, a high concentration of detritus might also indicate a high concentration of living phytoplankton. Since the SpecWa dataset contains many datapoints with high chlorophyll *a* concentrations up to $200\,\mu g\,L^{-1}$, a high detritus concentration might also be suitable. Thus it seems, that the general RF model selects the chlorophyll *a* and detritus-related features for the estimation task. The same applies to the model of the artificial pond at KIT. Especially on its datapoints, we see a distinct peak at 810 nm and with minor characteristic on the artificial pond at the castle garden. Both water bodies represent the high chlorophyll *a* values in the SpecWa dataset. In turn, the model for the pond in the castle garden does not rely on the 810 nm feature. However, besides the peak at 710 nm, this model relies on the broad peak between 500 nm to 600 nm related to low chlorophyll *a* absorption.

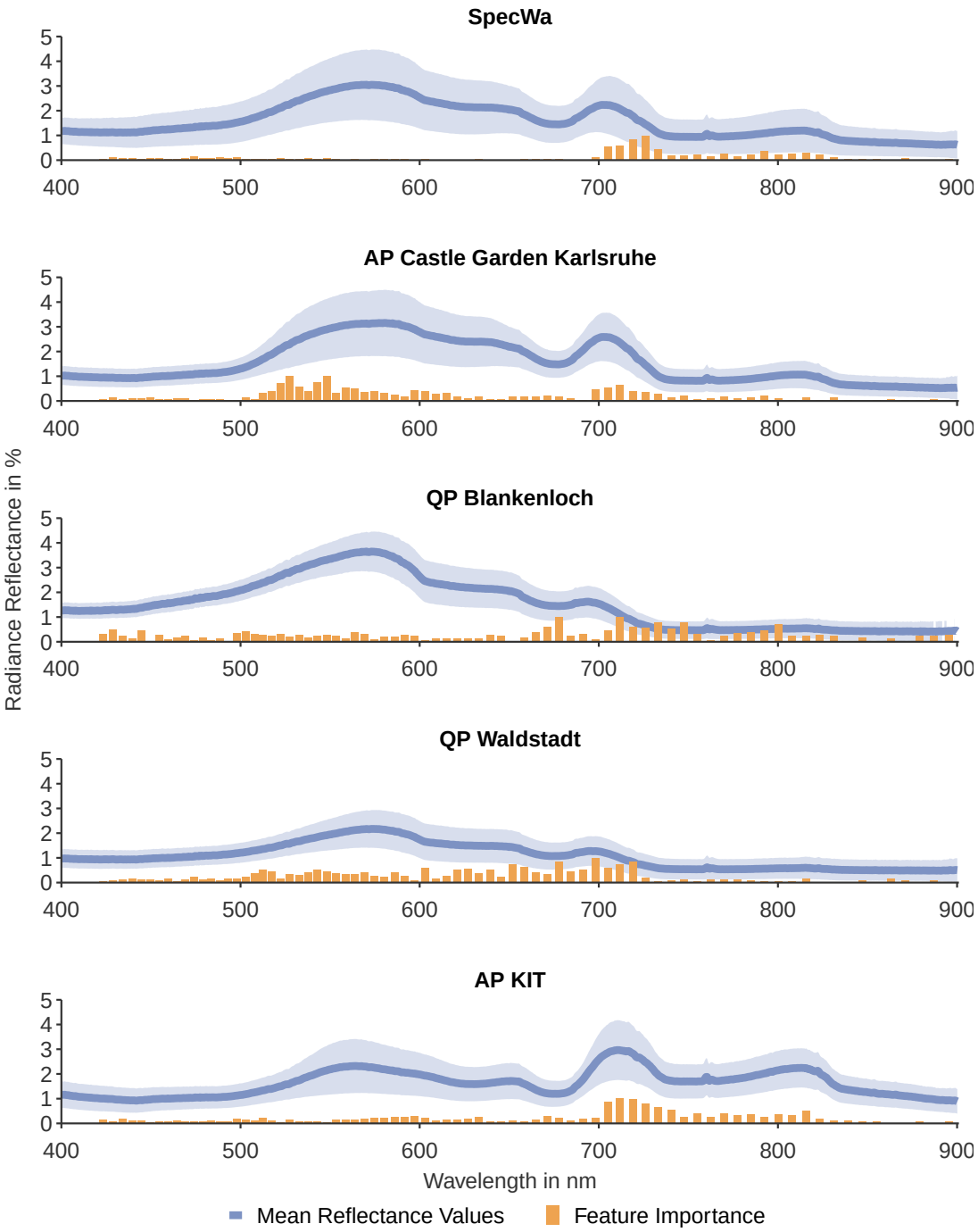

Figure 6.9: The figure illustrates the mean spectrum (dark blue) with the standard deviation (blue shade) of various subsets of the SpecWa dataset and the feature importance of the respective RF model trained on that subset. The feature importance is generated by permutation in the RF model. The feature importance values are normalized between 0 and 1. The first spectrum represents the whole SpecWa dataset, whereas the following represent some specific water bodies.

The important features of the two remaining quarry ponds are more spread over the spectrum, but for Waldstadt it is more narrow than for Blankenloch. The Waldstadt model emphasizes the features directly related to high or low chlorophyll *a* between 500 nm to 720 nm. For the

Blankenloch model, we see the features distinctly at 670 nm and 710 nm but also features in the longer wavelengths. This might be because the lake's spectrum is highly influenced by its benthic substrate that experiences more impact on the longer wavelengths features. However, both models show a less extent in the chlorophyll *a* concentration and a less amount of datapoints. Thus, it might be more difficult for the models to find distinct important features compared to the other three models.

Comparing the selected features for chlorophyll *a* in **Study Setup II** with the ones in **Study Setup II**, for most of the models, the features are more distinct in this setup. This concerns especially the generalized model comprising all water bodies, but also most of the more specialized models. A higher degree of the models' generalization might be an indicator for these findings, even for the models of a single water body. One reason for a higher degree of generalization might be the longer period of the data collection with varying seasonality, sun altitude and a higher chlorophyll *a* concentration range.

6.2.9 Setup II – Research Goal 1: Models and Parameters

Regarding the conducted **Study Setup II**, the focus was solely on chlorophyll *a* as a retrieval parameter. According to all obtained estimation results, the performance of models was satisfying. Of course, the performance depends on the spectral resolution, but the benefit of a narrow resolution is not as substantial as we thought before conducting the study. Applying a shallow ANN, for the first time in the scope of this thesis, lead to good estimation results, but no clear outperformance to the RF and the SVM was shown. Thus, the applied models in both frameworks show only marginally differences in their performance for this study design. One interesting finding is the positive influence of the first derivative as input features for the RF models. Calculating derivatives result in a loss of the absolute reflectance value. It seems that this circumstance does not strongly influence the estimation performance of the models in general. To answer **RQ 1** for **Study Setup II**, the models achieve suitable estimation results but there is no final model choice.

6.2.10 Setup II – Research Goal 2: Feature Importance

Regarding the feature selection of the RF models, we conclude that with increasing generalization, meaning, in this case, eleven water bodies, a chlorophyll *a* range of $0\,\mu g\,L^{-1}$ to $200\,\mu g\,L^{-1}$, and varying seasonality, the model picks mainly the spectral features physically related to chlorophyll *a*. The most frequently selected features by the models are not the ones related to the chlorophyll *a* absorption feature at 670 nm but the slope of the following-up reflectance peak at about 720 nm. One reason for this finding might be that the absorption feature itself does not contain information since its reflectance value is

more or less similar at most of the investigated mean spectra throughout all water bodies. For water bodies containing high chlorophyll *a* concentrations, we also see important features around 820 nm likely related to degenerating phytoplankton.

One additional finding is the shifting reflectance peak at around 710 nm that was also published by, e.g., [38]. Such a shifting peak would not allow a simple band ratio model relying on that feature to achieve a good estimation performance on multiple water bodies, especially for hyperspectral features. For such an approach with hyperspectral features, different bands need to be applied for varying chlorophyll *a* concentrations. The shifting mainly occurs by more chlorophyll *a* absorption at 670 nm, which widens the absorption feature and in turn shifts the peak. To answer **RQ 2** for **Study Setup II**, yes, especially the more generalized models mainly rely on the relevant features for chlorophyll *a*, applied in empirical models in the literature.

6.2.11 Setup II – Research Goal 3: Generalization

Focusing on the generalization ability of the models, we conclude they generalize well on the given dataset since the models show a suitable performance on multiple water bodies. However, the models have seen datapoints from every water body in the training phase. Thus, we cannot assume that those models would perform well on a completely unknown water body. Additionally, the distribution of the water bodies and their chlorophyll *a* concentrations is not equally weighted. Therefore, the models might get suitable results if they only perform well on the most frequently occurring water bodies.

It is not excluded that the models on **Study Setup II/1** suffer from some overfitting. For example, the models based on the 20 nm resolution rely on only a fifth of the 4 nm resolution's features. So fewer features may lead to less overfitting but simultaneously less performance by a lower spectral resolution.

In the context of **Study Setup II**, we can answer **RQ 3** with yes. The selected ML models can generalize on multiple water bodies. However, it cannot be proven that they can estimate chlorophyll *a* concentrations of completely unknown water bodies.

6.2.12 Setup II – Research Goal 4: Spectral Resolution

The impact of the spectral resolution on the models' performance was one primary investigation in **Study Setup II**, representing **RQ 4**. To generalize our findings, we have seen an impact of the spectral resolution on the performance, but it was less pronounced than expected. Models based on hyperspectral resolutions reveal satisfying performance. With different hyperspectral resolutions, there is nearly no difference in the models' performance.

Even some models based on multispectral resolution show a good estimation performance as well. This concerns both Sentinel resolutions. However, for the multispectral resolutions, there is no continuous finding concerning the bandwidth. Applying an ANN on the Sentinel-2 resolution lead to nearly the same performance as for the Sentinel-3, or both hyperspectral resolutions. Nevertheless, there is a vast decrease in the estimation performance for the models based on the two Landsat resolutions. The difference between Landsat and Sentinel-2 are mainly Sentinel-2's two additional bands between 670 nm to 720 nm. The relevance of the spectral range around 720 nm can be concluded from the variable importance analysis, especially for the generalized model. Thus, it might not be directly the bandwidth that causes the performance decrease, but also the positioning. Additional bands must not lead to better performance when they are placed on insignificant positions for water parameters. This is shown, for example, by the Landsat 8 and Landsat 5 resolution. Landsat 8 has an additional band, but this has no impact on the models' performance, though it has only five bands.

However, the findings between Sentinel-2 and the hyperspectral resolutions are only valid for the observation on this dataset. We cannot prove if the models generalized only on SpecWa's water bodies or if they are applicable with suitable performance on unseen water bodies.

Additionally, we identify the peak shift at 700 nm for the different water bodies (cf. [38]). This might be crucial for generalized models. Such a shift is hardly recognizable for, e.g., Sentinel-2-based models. Thus we think such a shift provides additional information, which cannot be used when the resolution is too low.

Answering **RQ 4** on **Study Setup II**, we can say that multispectral resolution similar to Sentinel-2 is still suitable for a good estimation performance on chlorophyll *a*. However, it might be possible that this is not valid for every water body. Additionally, it must not be practicable on actual satellite data due to atmospheric correction noise.

6.3 Setup III – Models Trained on WASI Data and Applied on the SpecWa Dataset

This section includes material from the publication

Philipp M Maier, Sina Keller, and Stefan Hinz. "Deep Learning with WASI Simulation Data for Estimating Chlorophyll a Concentration of Inland Water Bodies". In: *Remote Sensing* 13.4 (2021), p. 718. It is cited as [26] and marked with a purple line.

Study Setup III is the last conducted study within the scope of this thesis. Its main target is to fulfill the generalization approach by a complete separation of the datasets. Therefore,

we train the models on the simulated WASI dataset, presented in Section 4.3, and evaluate their performance on the SpecWa dataset. Chlorophyll *a* is the investigated water parameter. Hereby, different datasets mean not only unknown datapoints, as in **Study Setup II**, but also entirely unknown water bodies. One consequence of such an approach might be that models performing well on the unknown SpecWa dataset, would likely also perform well on other water bodies with the same characteristics as the SpecWa dataset. In the context of inland water monitoring with spectral data, such an approach has not been applied yet. This approach meets **RG 3** (**Generalization**) in terms of generalization as well as **RG 5** (**Transferability**) in terms of model training on simulated datasets and their transferability to real water samples. The transferability between the spectral data is one challenging task since both datasets contain surface reflectance, overlying the water leaving radiance signal. Therefore, we simulated various kinds of surface reflectance with multiple parameters in the WASI dataset (see Section 4.3.2). Hence, it is an additional task for the ML models to find the important features in the spectral data. Contrary, approaches in the literature try to remove the surface reflectance by applying correction models.

To face this challenging approach, we employ a 1D CNN as a representative for DL techniques. Moreover, we use a deeper ANN as in **Study Setup II** in the framework. Contrary to the other study setups, employing such techniques is possible in this setup due to the vast amount of data provided by the WASI tool. Besides the neural networks, we use a RF as a shallow learning model and the BR approach by Moses et al. [91] in combination with a linear regression model to compare the estimation performance with a standard empirical model as a baseline. Those four models represent **RG 1** (**Models and Parameter**) in estimating the chlorophyll *a* concentration. For **Study Setup III**, **RG 1** highly depends on **RG 3** and **RG 5**.

RG 4 (**Spectral Resolution**) is the last RG we face in this setup. We investigate the influence of the spectral resolution again. In **Study Setup III**, we focus on the hyperspectral EnMAP resolution as `SR-EnMAP` and the multispectral Sentinel-2 resolution as `SR-Sentinel`.

6.3.1 Setup III – Data and Study Preparation

In this section, we briefly describe the study and data preparation. The applied methodology bases on the explanations in Section 4.4 and Section 5.5.

Study Setup III relies on the WASI dataset (see Section 4.3) and the SpecWa dataset (see Section 6.2). Since both datasets originally have two different spectral resolutions, the one simulated with WASI with a respective 1 nm resolution and the SpecWa dataset with a respective 0.65 nm resolution, we scale the data to a common one. The common downampled resolutions are represented by the hyperspectral `SR-EnMAP` containing 77 spectral features and the multispectral `SR-Sentinel` with 9 spectral features. The spectral

downsampling follows the same scheme as in **Study Setup II**, described in Section 4.4). All obtained spectral bands and their central wavelengths are visualized in Figure 4.10.

Concerning the compatibility between the WASI dataset and the SpecWa dataset, we have to make some constraints for the latter because the WASI tool offers some limitations. First, we need to cut the maximum chlorophyll *a* concentration to $100\,\mu g\,L^{-1}$. Such a limitation is necessary because WASI is mostly designed for natural water bodies that normally do not exceed this value. Second, we must exclude datapoints with a cyanobacteria concentration higher than $5\,\mu g\,L^{-1}$ since cyanobacteria are not implemented in WASI. However, we think that values up to $5\,\mu g\,L^{-1}$ are still feasible for the models since we do not want to remove too many datapoints. As a consequence, 2.617 datapoints remain for the evaluation of the models.

For the training of the models, we conducted a split of the WASI data into a training, a test, and a validation set, summarized in Table 6.11. The training is conducted on the training set, whereas the validation set is employed for the model validation during the training phase. In this setup, we employ two test sets, the WASI test set and the SpecWa datset. The whole data processing and preparation is presented as an overview in Figure 6.11.

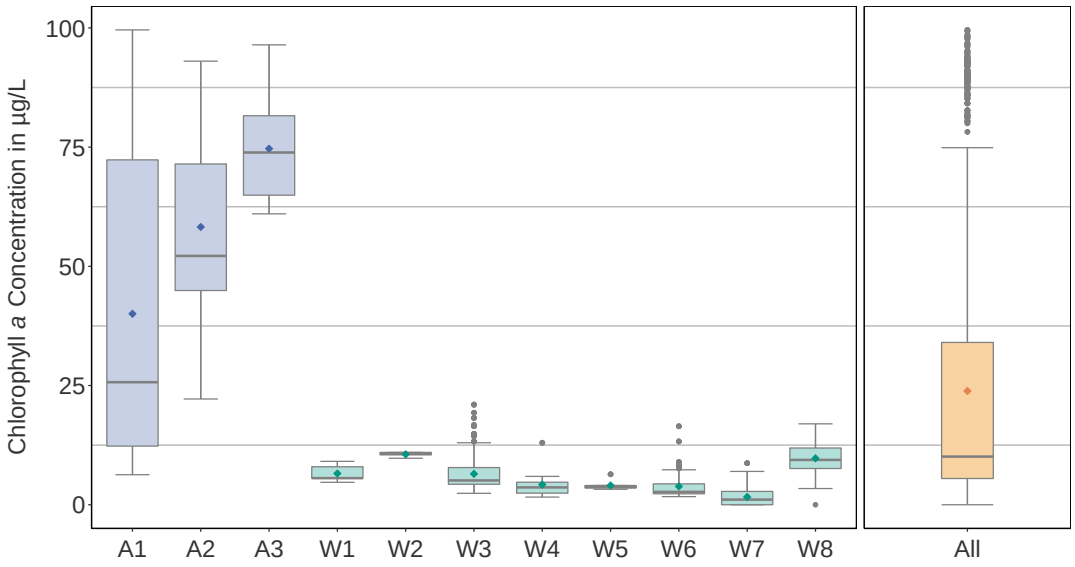

Figure 6.10: Boxplots of the chlorophyll *a* concentration range for each inland water body individually, and for all water bodies of the SpecWa dataset. The water ID is given in Table 4.2. The diamonds in the boxes symbolize the respective mean, the lines the respective median value. The lower limit of each box is the 25^{th} percentile (Q1), the upper limit the 75^{th} percentile (Q3) so that the difference builds the interquartile range (IQR). Whiskers extend to Q1 – 1.5 ∗ IQR and Q3 + 1.5 ∗ IQR. Any points beyond the whiskers are outliers and are plotted as points. Taken from [26].

The distribution of the chlorophyll *a* values between the WASI-training set and the SpecWa-test set was visualized in Section 4.3 and Figure 4.9. Besides, we illustrate the ap-

plied water samples of the SpecWa dataset concerning their chlorophyll *a* range and the respective water bodies in Figure 6.10.

Table 6.11: Number of datapoints of the WASI-generated simulation dataset for each of the three subsets. Taken from [26].

Dataset	%	Number of datapoints
Training	70	369 600
Validation	15	79 200
Test	15	79 200

We explained the training process of the 1D CNN and the ANN in Section 5.5. The 1D CNN's final structure concerning its layers and the respective amount of neurons was visualized in Figure 5.2. Table 5.2 summarizes the final hyperparameter set for the 1D CNN and the ANN for both applied spectral resolutions. As for the minor dimensioned `SR-Sentinel` input data, the 1D CNN architecture differs. Since the empirical three-band approach by Moses et al. [91] (BR) is related to the MODIS and the MERIS resolution, we need to slightly adapt the bands to the setups' `SR-EnMAP` and `SR-Sentinel`. In the case of the `SR-EnMAP` resolution, we select bands at 664.7 nm, 711.9 nm, and 755 nm, while in the case of `SR-Sentinel` resolution bands at 665 nm, 705 nm, and 740 nm are chosen. We parametrize the BR approach on the WASI-generated simulation training dataset with linear regression.

We do not analyze only the models' performance for the entire dataset, as we have done in the previous setups. We additionally evaluate the models for each of the investigated water bodies. This allows a better view on the models' generalization ability.

6.3.2 Setup III – Results

In the following, we present the chlorophyll *a* estimation results of the different ML approaches combined with the two downsampled resolutions, `SR-EnMAP` and `SR-Sentinel`. The 1D CNN, the ANN, and the RF performances on the WASI test dataset are greater than $R^2 = 99\,\%$ for the `SR-EnMAP` and $R^2 = 98\,\%$ for the `SR-Sentinel`, as expected since the WASI-generated simulation data is relatively homogeneous. Therefore, we focus on the estimation performance with the independent, real-world SpecWa dataset, representing the study's primary objective. The results are structured in three parts: (1) We describe the applied models' overall estimation results on the complete SpecWa dataset. (2) Subsequently, the best ML model is selected to investigate the estimation performance on the respective dataset in detail. (3) The specific estimations for the eleven water bodies of the SpecWa dataset are described.

Table 6.12 shows the overall estimation performance of the 1D CNN, the ANN, and the two baseline models, RF and BR, in terms of the three metrics R^2, RMSE, and MAE, for the

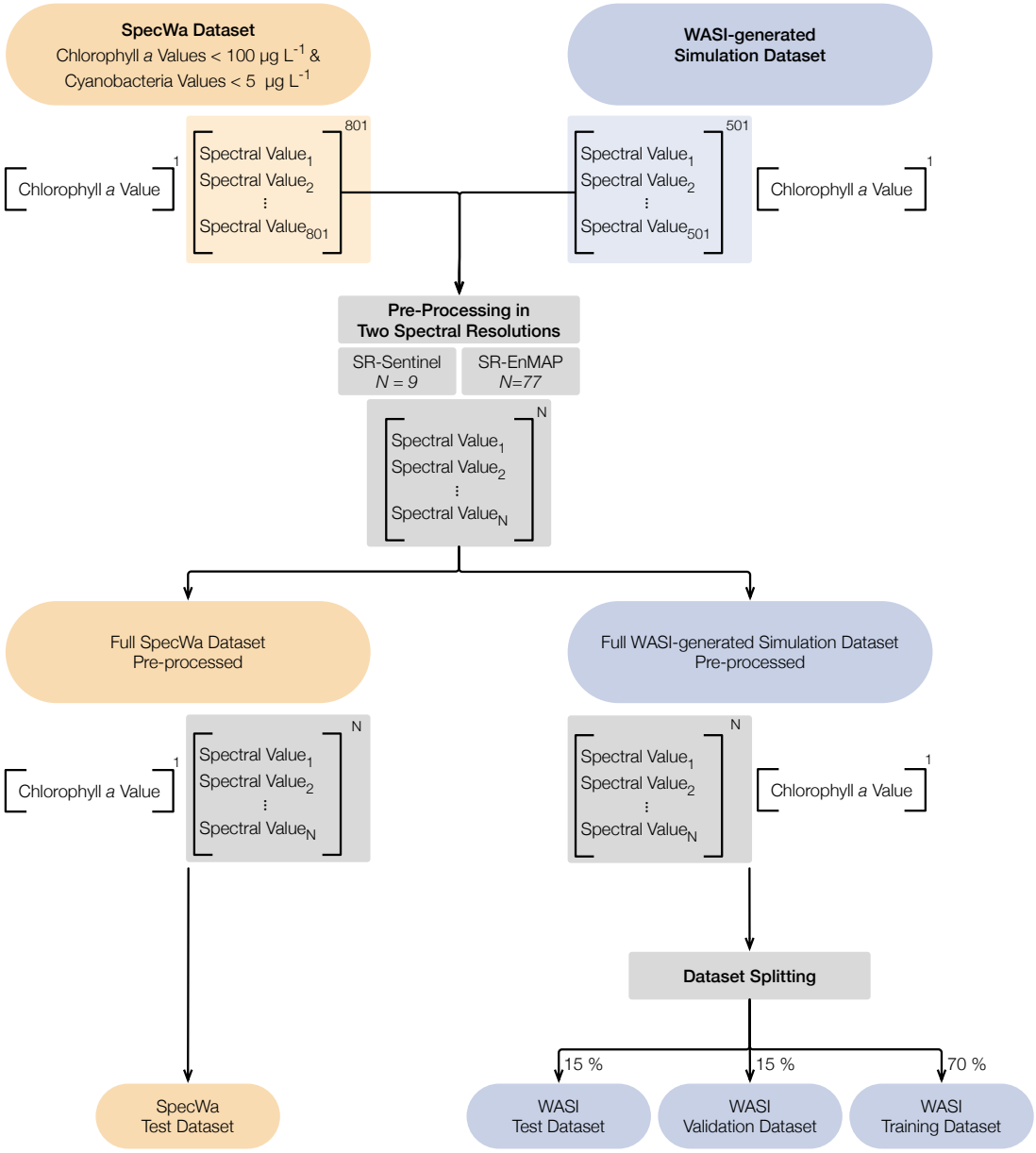

Figure 6.11: Pre-processing schema for SpecWa dataset (orange) and the WASI-generated simulation dataset (blue). N refers to either 9 spectral input features in the case of the `SR-Sentinel` resolution or 77 in the case of the `SR-EnMAP` resolution. Taken from [26].

SpecWa (test) dataset. Concerning the two spectral resolutions, the 1D CNN and the ANN achieve better estimation results on the finely resolved `SR-EnMAP` data. At the same time, RF and BR perform better on the `SR-Sentinel` data for all performance metrics. Overall, the 1D CNN represents the best estimation model with $R^2 = 81.9\,\%$, RMSE $= 12.4\,\mu g\,L^{-1}$, and MAE $= 6.7\,\mu g\,L^{-1}$ on the `SR-EnMAP` data. The ANN performs as the second-best model on the `SR-EnMAP` data, but it is significantly worse than the 1D CNN. On the `SR-Sentinel` data, the 1D CNN is also the best model. However, in this case, the 1D CNN underperforms with an $R^2 = 62.4\,\%$, RMSE $= 19.3\,\mu g\,L^{-1}$, and MAE $= 14.6\,\mu g\,L^{-1}$ compared to its performance

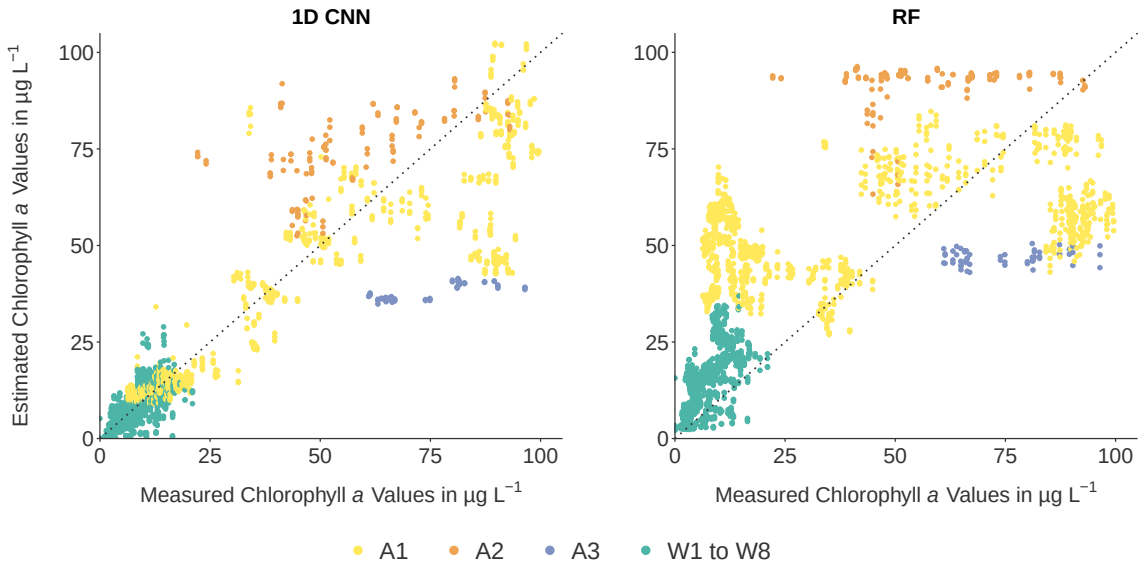

Figure 6.12: Visualization of the estimation results (y-axes) generated by the 1D CNN and the baseline RF model compared to the measured chlorophyll *a* values (x-axes) on the SpecWa dataset. The natural water bodies W1 to W8 are colored in green, while the artificial water bodies are characterized by three different colors: A1 in yellow, A2 in orange, and A3 in blue. Taken from [26].

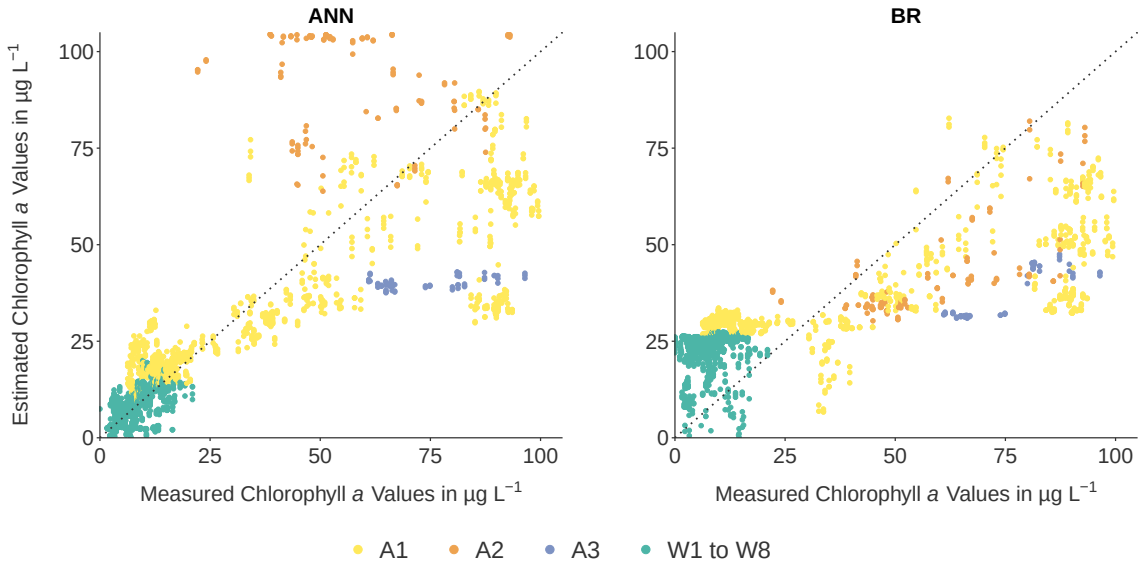

Figure 6.13: Visualization of the estimation results (y-axes) generated by the ANN and the baseline BR model compared to the measured chlorophyll *a* values (x-axes) on the SpecWa dataset. The natural water bodies W1 to W8 are colored in green, while the artificial water bodies are characterized by three different colors: A1 in yellow, A2 in orange, and A3 in blue. Taken from [26].

on the finely resolved `SR-EnMAP` data. Regarding the three estimation metrics and the `SR-Sentinel`, the 1D CNN's performance is only better in the case of the R^2-score compared to the RF. Otherwise, the range of the models' performance metrics is smaller on the `SR-Sentinel` data. For example, the R^2-score of all models ranges from 37.9 % to 81.9 % on the `SR-EnMAP` data, while R^2-score varies from 51.5 % to 62.4 % on the `SR-Sentinel`.

In order to compare the estimation results with the measured chlorophyll *a* values in the subsequent Section 6.3.3, we provide additional information about the SpecWa chlorophyll a values (see also Figure 6.10). The chlorophyll *a* ranges from $0\,\mu g\,L^{-1}$ to $99.6\,\mu g\,L^{-1}$ for all SpecWa inland water bodies with a mean value of $23.85\,\mu g\,L^{-1}$ and a median value of $10.1\,\mu g\,L^{-1}$. Based on this information, in sum, the results on the `SR-Sentinel` data remain unconvincing (see Table 6.12, right part) for all selected models.

Table 6.12: Results for the chlorophyll *a* estimation of all SpecWa inland water bodies. Taken from [26].

Model	SR-EnMAP			SR-Sentinel		
	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$	R^2 in %	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
1D CNN	**81.9**	**12.4**	**6.7**	**62.4**	**19.3**	**14.6**
ANN	66.6	16.6	9.3	54.8	23.4	17.1
RF	51.1	22.7	17.0	51.1	20.2	14.7
BR	37.9	23.0	19.3	51.5	22.3	17.8

Since the 1D CNN represents the best model on the complete SpecWa dataset, especially for the `SR-EnMAP` data, we focus on its estimation performance in detail. Figure 6.12 (left) shows the results generated by the 1D CNN model compared to the measured chlorophyll *a* values on the SpecWa dataset. On the right of Figure 6.12, we provide the estimation results generated by the RF baseline compared to the measured chlorophyll *a* values on the SpecWa dataset. As for the visualized 1D CNN-generated distribution of the estimated and measured chlorophyll *a* values, we notice that most of the low values of the natural water bodies W1 to W8 are estimated correctly (low bias). The 1D CNN over- and underestimates a limited amount of datapoints. This finding is primarily related to higher chlorophyll *a* values. The chlorophyll *a* values of the water body A2 are consequently overestimated whereas the water body A3 values are underestimated. In contrast, the RF (Figure 6.12, right) shows a significantly worse distribution of the estimated and measured chlorophyll *a* values (high bias). The RF overestimates most of the SpecWa chlorophyll *a* values; solely the values of the water body A3 are underestimated. In addition, Figure 6.13 visualizes the estimation results generated by the ANN and the baseline BR.

With respect to a detailed analysis of the individual water bodies, we summarize the estimation performance of all selected ML models on each SpecWa water body in Figure 6.14 in terms of the RMSE, accompanied by the specific values for the MAE and RMSE in Table 6.13. Note that the R^2-score is not provided. In the case of the individual inland water bodies, the number of datapoints is too low and the range of the chlorophyll *a* concentration is partly too small, resulting in inconclusive R^2-values. The estimation results for the different water bodies can be sorted into three parts. The first part includes water bodies whose chlorophyll *a* values are generally well-estimated by several models independently of the two spectral resolutions. Secondly, water bodies exist whose chlorophyll *a* values are only well-estimated

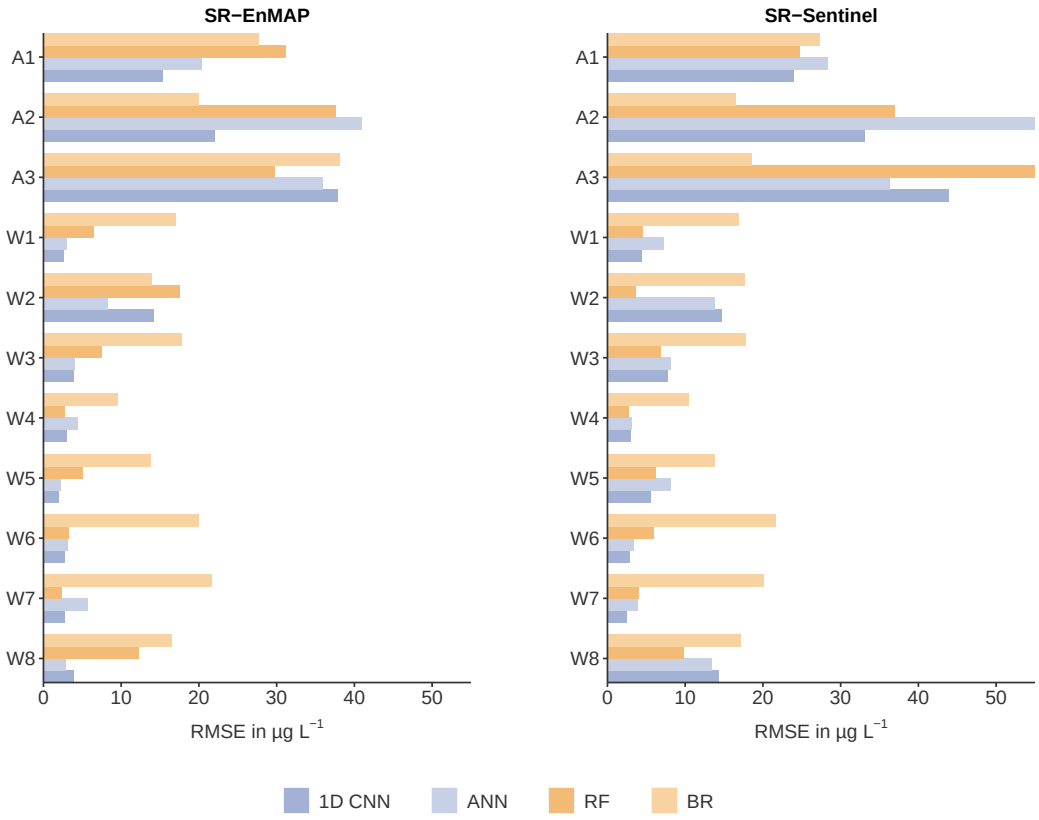

Figure 6.14: Visualization of the models' estimation results based on the RMSE-scores on the different water bodies with the two different downsampled resolutions. Taken from [26].

by a few models or only on one of the downsampled spectral resolutions. Furthermore, water bodies whose chlorophyll *a* values are generally hard to estimate by all ML models.

As shown in Figure 6.14 and Table 6.13, all ML models can predict the chlorophyll *a* values of nearly all natural water bodies W1 to W8 satisfyingly. Besides, Figure 6.15 exemplifies the scaled deviation between the estimation results of the 1D CNN and the RF model and the measured chlorophyll *a* values of the SpecWa dataset as violin plots. The deviation Δ_{scaled} emphasizes the comparison of the estimation performance concerning all water bodies since we apply a min-max-scaling for each water body individually. This scaling is performed for the measured and estimated chlorophyll *a* values individually. Eventually, the resulting values are normalized in the range of 0–1 and are independent of the target variable's unit. Figure 6.15 provides information about the median of the deviations between the scaled estimated and scaled measured chlorophyll *a* values (white dot) and the entire distribution of these deviations. The natural water body W2 is an exception concerning the satisfied estimation since only the RF on the `SR-Sentinel` performs well on this waterbody's datapoints. However, W8 is an example that is estimated well by both neural networks but only on the `SR-EnMAP` (see Figure 6.14 and Figure 6.15). In addition to W2, the datapoints of

A1 are also estimated appropriately by one model, the 1D CNN. This finding refers especially to the SR-EnMAP data (see, for example, Figure 6.15). A2 and A3 represent the third part of inland water bodies since no ML model can estimate the measured chlorophyll *a* values satisfyingly. This finding is revealed for the 1D CNN and the RF model when focusing on Figure 6.15. To sum up, both neural networks and the RF perform well, but their performance depends highly on the individual water bodies and their chlorophyll *a* concentrations.

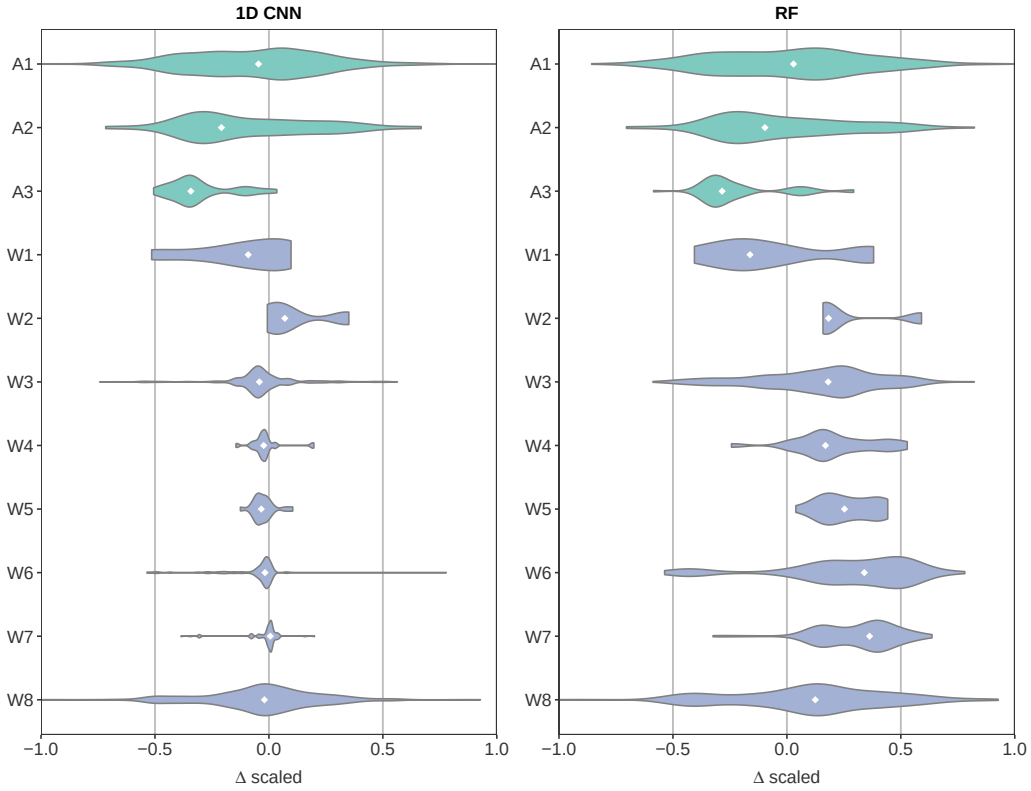

Figure 6.15: Visualization of the estimation results on the SR-EnMAP resolved data generated by the 1D CNN (left) and the RF model (right) as the min-max scaled deviation Δ_{scaled} between the estimated and measured chlorophyll *a* values. The violin plots reveal the median of the deviations between the scaled estimated and scaled measured chlorophyll *a* values (white dot) and the entire distribution of these deviations. Taken from [26].

6.3.3 Setup III – Estimation Performance Concerning the two Downsampled Spectral Data and the Different ML Models – Discussion

As for the downsampled spectral data shown in Figure 4.10, the models' estimation performance on the total SpecWa dataset varies between the well-performing neural networks and the two baseline models on the SR-EnMAP data (see Table 6.12, Figure 6.12 and Figure 6.13).

Table 6.13: Results for the chlorophyll *a* estimation of the SpecWa inland water bodies. Taken from [26].

Water ID	Model	SR-EnMAP		SR-Sentinel	
		RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$	RMSE in $\mu g\,L^{-1}$	MAE in $\mu g\,L^{-1}$
A1	**1D CNN**	**15.4**	**9.1**	24.0	20.9
	ANN	20.4	13.8	**21.8**	**14.1**
	RF	31.2	27.7	24.7	21.1
	BR	27.8	22.0	27.3	19.7
A2	1D CNN	22.0	18.2	33.1	27.5
	ANN	40.9	34.6	56.7	53.3
	RF	37.6	33.4	37.0	32.6
	BR	**20.0**	**17.2**	**16.6**	**13.7**
A3	1D CNN	37.9	36.8	43.9	42.5
	ANN	35.9	34.5	33.2	32.0
	RF	**29.7**	**27.8**	57.8	57.0
	BR	38.1	37.6	**18.6**	**16.3**
W1	**1D-CNN**	**2.6**	**2.4**	4.5	**3.2**
	ANN	3.1	3.0	**3.9**	3.6
	RF	6.5	6.2	4.6	4.0
	BR	17.0	16.8	16.9	16.8
W2	1D CNN	14.2	14.1	14.9	14.9
	ANN	**8.4**	**8.3**	14.6	14.5
	RF	17.5	17.4	**3.6**	**3.3**
	BR	13.9	13.9	17.7	17.7
W3	**1D CNN**	**3.9**	**3.0**	7.8	6.4
	ANN	4.0	**3.0**	**4.5**	**3.4**
	RF	7.5	6.6	6.8	6.0
	BR	17.8	16.8	17.8	16.9
W4	1D CNN	3.0	2.0	3.0	2.0
	ANN	4.4	3.8	3.3	2.5
	RF	**2.8**	**1.7**	**2.8**	**1.8**
	BR	9.6	8.3	10.5	7.6
W5	**1D CNN**	**2.0**	**1.4**	5.5	4.0
	ANN	2.3	2.1	**2.3**	**2.0**
	RF	5.1	4.4	6.3	6.0
	BR	13.8	12.3	13.8	12.5
W6	**1D CNN**	**2.8**	**1.6**	2.9	2.2
	ANN	3.2	2.0	**2.7**	**1.6**
	RF	3.3	2.7	6.0	5.6
	BR	20.0	17.2	21.7	17.2
W7	**1D CNN**	2.8	**1.8**	**2.5**	**2.1**
	ANN	5.7	4.8	3.3	2.9
	RF	**2.4**	2.0	4.1	3.7
	BR	21.7	21.2	20.1	19.5
W8	1D CNN	3.9	2.9	14.3	13.2
	ANN	**2.9**	**2.3**	**5.9**	**4.8**
	RF	12.3	11.2	9.8	9.1
	BR	16.6	16.2	17.2	16.9

Most models perform similarly and worse on the `SR-Sentinel` data, as shown in Table 6.12.

Except, the BR performance increases on the SR-Sentinel data since the selected bands are optimized for a multispectral resolution. These worse models' performances on the SR-Sentinel data can be explained by an information loss due to the coarser spectral downsampling. In addition, the models vary highly in their ability to handle multiple, multicollinear input data and select important features from these data. The information loss is caused by the downsampling of the 1 nm resolution original data to the SR-Sentinel, as shown in Figure 6.10. In contrast, the SR-EnMAP with 6.5 nm-bandwidth characterizes approximately the spectrum's original distribution, whereas the SR-Sentinel data no longer presents this characteristic. The SR-EnMAP data includes, for example, the typical chlorophyll *a* absorption at 685 nm [33] and the following scattering peak 705 nm, which are covered by several bands (see Figure 4.10). This information is provided by only two broad bands for the SR-Sentinel data. Moreover, the potential peak shift in the respective spectral range, presented in Section 6.2.8 is not detectable for the SR-Sentinel-based models. Besides, the chlorophyll *a*'s origin depending on either green algae or diatoms is only recognizable in the finer resolved SR-EnMAP data. Information about, for example, the water composition of different parameters is missing in the SR-Sentinel. The loss of information impedes the estimation for any data-driven ML model.

Regarding the different models' performances on the finer resolved SR-EnMAP data, the two baseline models cannot exploit the detailed spectral features (see Table 6.12, and Figure 6.12 and Figure 6.13, right). The RF model, for example, cannot use the high-dimensional WASI-generated simulation data to transfer the linkages of the training process on the unknown SpecWa test dataset as visualized, for example, in Figure 6.12. This finding is sustained since the RF's performance on the SR-EnMAP data is even slightly worse than on the more downsampled SR-Sentinel data. In contrast, both neural networks, especially the 1D CNN, benefit strongly from the finer resolution of the SR-EnMAP data, resulting in more spectral input features. This finding is clearly revealed in Figure 6.12 (left). Therefore, the 1D CNN can learn the linkage between the WASI-generated simulation data and the chlorophyll *a* values. This DL approach can generalize its learnings on the WASI-generated simulation test dataset, particularly on the entirely unknown SpecWa dataset.

When comparing the two best performing models, the ANN and the 1D CNN, the estimation results reveal that the 1D CNN estimates the chlorophyll *a* concentrations on the SpecWa dataset better than the ANN in terms of 15.3 p.p for the R^2-score. This outperforming can be explained by the characteristic ability of the 1D CNN to exploit the information implicitly contained in the large number of spectral features of SR-EnMAP better than a conventional ANN. As the 1D CNN architecture includes different filters and kernels, the spectral features' information is perfectly processed, which seems similar to applying different spectral derivatives [23].

Besides, when focusing on the noise in the real-world dataset, such as the SpecWa dataset, the sun's spectral signal and the sky glint can be higher than the water-leaving radiance [31].

This noisy part of the signal is reflected on the water surface and is included in the SpecWa test dataset, which affects the estimation performances in Table 6.12. We have varied the parameters influencing the surface reflection to cover a broad range of occurring real-world conditions during the WASI simulation process. Generally, the ML models need to handle these different illumination conditions as noise. Returning to the comparison of both neural network approaches, the estimation results reveal that the 1D CNN can minimize occurring noise, such as the effect of the absolute reflectance values in the input data. Therefore, the generalization abilities of the 1D CNN ensure the transfer to further spectral data provided by different sensors, illumination conditions, and eventually, distinct water bodies that are prevailed by the underlying estimation task in our study (see, for example, Figure 6.12, right). The weaker performance of the ANN (see Figure 6.13, right) can be explained by strongly responding to the noise in the input data.

Finally, we analyze our estimation results and the estimation results provided by a previously conducted study on one part of the SpecWa dataset by Maier and Keller [24]. They have investigated the performance of several shallow ML models to estimate the chlorophyll *a* concentration with different spectral resolutions. Note that a detailed comparison of the estimation performances is infeasible since the training dataset are entirely different. In the previous study, all ML models have been trained on datapoints of the SpecWa dataset [24]. The SpecWa dataset has been known to the models, which is not the case for our study. Therefore, the overall models' estimation performance in the previous study is better when regarding the absolute figures. A finding of Maier and Keller [24] is that the models' performances are only slightly better on the finer resolved `SR-EnMAP` data. For example, the ANN model performs slightly worse on the `SR-Sentinel` data than on the `SR-EnMAP` data (see [24], Table 6.13). In our study, the ML models show larger differences in their estimation performance between the `SR-Sentinel` and `SR-EnMAP`. As for the arising questions addressing the generalization abilities of the models in Maier and Keller [24], they have not been applied to unknown water bodies. Therefore, we assume that these models would perform significantly worse than the models in the underlying study due to the lack of noise and variability generated in our case by the WASI tool.

6.3.4 Setup III – Estimation Performance Concerning the Individual Water Bodies – Discussion

Since the SpecWa dataset consists of chlorophyll *a* concentrations of eleven different inland water bodies (see, for example, Figure 6.10), we discuss the applied ML models' estimation performance concerning these water bodies. In contrast to the commonly applied BR approaches (see, for example, [91, 153]), and the more advanced BR approaches for specific water types [118, 119], the proposed data-driven ML approaches are not trained on datapoints of the individual water bodies. Our objective is to provide a more generalized

approach trained and parametrized on a dataset not necessarily including chlorophyll *a* values of the target water bodies. This approach differs also from the study of Pahlevan et al. [19], relying on a mixture density model with in-situ data of several inland water bodies. A large amount of training data is required concerning the applicability of a DL approach, which motivates employing simulation data such as the used WASI-generated simulation data. One advantage of data simulation is, for example, the possibility to generate a broader range of chlorophyll *a* values than appearing in natural water bodies.

Our models' performance results are given in Section 6.3.2, Table 6.13, Figure 6.14 and Figure 6.15. Most of the natural water bodies (W1 to W8) are well estimated by the majority of ML models (see Figure 6.14) and even on the more downsampled `SR-Sentinel` data. When focusing on the CNN as the best performing approach (see Table 6.12) on the `En-MAP` resolution, we retrieve this assessment based on the deviation distributions of W1-W8 in Figure 6.15 (left). As shown, the median of the deviations between the scaled estimated and measured chlorophyll *a* values are allocated around 0. A good estimation of the natural water bodies' chlorophyll *a* values implies that the proposed approach to train the ML models on an entirely different and simulated dataset and, subsequently, apply them to an unknown real-world dataset, works well for the entire SpecWa dataset and concerning most of the individual water bodies (see Figure 6.14). The natural water body W2 is an exception as for the good models' estimation performance with 6.4 p.p. difference of the best individual RMSE-score. This exception is also illustrated by the severe deviations and filling shapes of the CNN's and RF's distributions in Figure 6.15. In addition to the lowest number of datapoints (8, see Table 4.2), this exception seems understandable when investigating the water body's composition. The water level ranges between 0.5 m to 1 m and is low (see Table 4.2). Therefore, the benthic substrate strongly influences the spectral signature. Besides, small water plants might float on the water surface during the spectral measurements. The latter might cause an overestimation of the chlorophyll *a* concentration of this specific water body. We have to consider that such effects of polluted water surfaces always appear when measuring real-world water bodies with, for example, dust or pollen cover.

Regarding the artificial water bodies, it can be seen in Figure 6.14 as well as Table 6.13 that the ML models' estimation performances are worse compared to the natural water bodies. This finding can also be retrieved when focusing on the deviation distributions of the artificial water bodies compared to the natural ones in Figure 6.15. One reason for the worse estimation results can arise since the WASI tool has been originally developed for water bodies with generally lower chlorophyll *a* concentrations, such as the lake Constance in Germany [154].

Therefore, the WASI-generated simulation training dataset contains only a few higher chlorophyll *a* values resembling artificial water bodies. Another reason might be that water bodies characterized by broader chlorophyll *a* concentration ranges are generally more complex to estimate as they contain a variety of substances. These substances overlap in the measured spectral signatures and cannot be deducted easily. Increasing chlorophyll *a* concentrations

evoke a diversity in the inland water body concerning the horizontal distribution and vertical distribution of the water composition. From this perspective, the estimation task is heavily ill-posed for such water bodies, since the in-situ chlorophyll *a* measurement is taken at a rather specific depth range while the spectral data capture reflectances over the whole water column. Especially for water bodies with higher concentrations, this effect can lead to a higher estimation bias. Against the background of the mentioned aspects, the 1D CNN's performance on the water body A1 can be slightly revised since most of the datapoints are estimated satisfactorily for the chlorophyll *a* range of $16.3\,\mu g\,L^{-1}$ to $99.6\,\mu g\,L^{-1}$ (see Table 6.13).

In sum, it is obvious that a stepwise simulation characterizes the WASI-generated data cannot directly correspond to a natural water body. This is the reason why a particular estimation bias is expected (see, for example, Figure 6.15). Such a bias can be originated due to make assumptions before the simulation process. For example, these assumptions are:

- We simulated the WASI-generated dataset with three different benthic substrates: sand, silt, and a macrophyte species. Natural water bodies have additional materials such as gravel, leaves, or other organic materials that are not covered in the WASI tool.
- In different geogenic regions, a diversity of minerals occur, resulting in distinct reflective properties and colors for, e.g., suspended materials.
- Besides, several phytoplankton species exist, while the WASI-generated simulation data consist only of two species.

Concerning these mentioned aspects and many more, the ML model's performance, especially the 1D CNN's, estimate the chlorophyll *a* of the entire SpecWa dataset successfully and satisfactory for most individual water bodies.

6.3.5 Setup III – Research Goal 4: Spectral Resolution

Regarding the impact of the spectral resolution on the models' performance, the hyperspectral EnMAP resolution significantly increases the performance for both neural networks compared to multispectral Sentinel-2 resolution. Merely, for some water bodies, the Sentinel-2-based models achieve a similar performance to the EnMAP-based models. Regarding the overall metrics for the multispectral resolution, even the 1D CNN model could rarely transfer the learnings from the WASI data to the SpecWa test set. This underperformance might be due to the low number of spectral bands that impede the 1D CNN's advantage in creating new features. So the models can still adapt to the training data but cannot perform the transfer task. To sum up, the impact of the spectral resolution on the models' estimation performance in textbfStudy Setup III was stronger than on **Study Setup II**. Thus, the Sentinel-2 resolution is not suitable for such a generalized approach.

6.3.6 Setup III – Research Goals 1, 3, 5: Models and Parameters, Generalization and Transferability

To analyze **Study Setup III** in the context of the RGs, we discuss **RG 1**, **RG 3**, and **RG 5** at once since all three are related to each other in this setup. For this discussion, we rely solely on the results of the EnMAP resolution-based models. First of all, we observe an equal performance of the 1D CNN, the ANN, and the RF on the WASI-test set. This finding means that the models learned the linkages between the input features and the output parameter. However, observing only the models' performance on the WASI-test set would be similar to the approach in **Study Setup II**. Therefore, we applied the models on the independent SpecWa dataset as the main target of **Study Setup III** to demonstrate the generalization task. As a result, on the SpecWa dataset, we observe an outperformance of the 1D CNN over the other models concerning all metrics. With respect to the previous studies, the performance difference between the applied models in **Study Setup III** is enormous. Comparing the models' performance between **Study Setup I + II** and **Study Setup III** also reveals a significant decrease. The 1D CNN's R^2 score is lower by about 10 % to the best models in **Study Setup II**, but this is acceptable concerning the more challenging task in the current setup. With respect to the RF model's R^2 score, the difference of about 40 % emphasizes the performance of the 1D CNN. Although the numbers between the studies are not directly comparable, considering the distance between both is still valuable. Besides, we are dissatisfied with the overall performance of the ANN. We have not expected to see the ANN's performance that far below the CNN's one. However, the ANN performed well on multiple water bodies.

Trying to explain the performance difference between 1D CNN and the ANN, two reasons might appear. One problem for the ANN might be difficulties in its training process. So, the model was able to perform well on the WASI-test set, similar to the 1D CNN. Nevertheless, it was not able to bring the transfer to the real world data. This finding might indicate that the WASI data was too simple to learn for the ANN. So the ANN learned relations that were sufficient to estimate the WASI test set but insufficient for the SpecWa dataset. The other explanation might be the surface reflectance, which varies highly in the WASI data and also in the SpecWa data and strongly affects the spectra's magnitude. But the extent might differ in the datasets as well. Here comes the advantage of the 1D CNN. Its filters create various features that might neglect most of the surface reflectance effect. These features are kind of similar to the derivatives used as input for the RF, leading to more noise resistance. Noise can be seen as, e.g., changing weather conditions during the measurements or variation in the calibration techniques. As a result, the overfitting effect on the WASI data was rather low for the 1D CNN, which might explain its better performance on the SpecWa dataset. Nevertheless, the performance of both neural networks exceeds the RF's, which most likely suffered the same issues as the ANN. However, all three ML models significantly

outperformed the empirical BR model. To finally answer **RQ 1**, the 1D CNN is the only model that shows promising results in the context of **Study Setup III**.

Focusing on the generalization (**RQ 3**), we presented a 1D CNN that can estimate most of the unknown inland water bodies with suitable performance. Since we clearly separated the training and the test dataset, we can answer the question distinctly with yes. Additionally, we think that the model can also estimate most of the global water bodies, represented by the value ranges of the WASI-generated dataset.

Regarding the transferability in **RQ 5**, we have to differentiate between the applied models. The WASI-simulated dataset was easy to learn for both neural networks and the RF, but only the 1D CNN learned the features that allow a transfer to the SpecWa test set. However, for the ANN and the RF the difference between the WASI-test set and the SpecWa-test set is too much. Hence further studies beyond the scope of this thesis need to find a way for a good trade-off between adaption on the simulated training set and generalization. Additionally, we have to point out that the WASI tool is a physical model and always a simplification of the complex reality of a water body. Thus, the transferability might be improved by a more complex simulation concerning different influence parameters, additional algae classes, or benthic substrates. Nevertheless, we can answer **RQ 5** with yes. Without simulated data, such a generalized model as the presented 1D CNN, accompanied with a suitable estimation performance on chlorophyll *a*, would not have been possible. In sum, we are satisfied by its performance concerning the challenging task.

7 Conclusion and Outlook

The thesis aimed to build and apply ML models that can estimate water parameters of multiple inland water bodies based on spectral data. These models were trained and evaluated in three study setups, with rising complexity in Chapter 6. Five RGs with the respective RQs have been posted alongside these study setups. The conclusion of these RGs and answering the RQs are the focus of this chapter.

The chapter is organized into four sections. The first section is a short repetition of the studies conducted during the thesis and their embedding in the context of the literature (see section 7.1). In Section 7.2, the focus is on the conclusions concerning the RQs across the different study setups and their embedding in the related work. Subsequently, a summarized conclusion of the conducted research in this thesis is presented in Section 7.3. Eventually, an outlook is given for the future work containing an upscaling approach for the application on real satellite data, relying on the findings of this thesis (see Section 7.4).

7.1 Summary and Embedding of the Thesis' Content

Chlorophyll *a* is an essential parameter to understand the condition of an inland water body. Since their concentration varies over a certain period driven by natural or anthropogenic influences, frequent monitoring of the water bodies is advised to cover dangerous increases with actions. One cost-efficient monitoring option is the chlorophyll *a* retrieval with spectral remote sensing data [16]. However, this is a challenging task. There are two major challenges for parameter retrieval of inland water bodies.

The first one is to find a reliable model that can link spectral data to the chlorophyll *a* concentration. Well-performing models exist, unfortunately, only for few lakes with an extended data history. Besides, these models only achieve good estimation results for exactly the respective water body they are specialized in [17]. For most of the global water bodies, such a data history does not exist. Therefore, there is a vast demand for generalized models that can estimate the chlorophyll *a* concentration for most of the water bodies without adaptions. These kind of models were intensively investigated in this thesis. The other challenge concerns the satellite images in terms of atmospheric correction over inland water bodies, which will be investigated in future research.

In general, three different approaches exist for water parameter retrieval from spectral data: analytical approaches, empirical approaches, and data-driven ML approaches. Analytical approaches have the claim to be good in the generalization task since they rely on physical modeling. However, they need to be adapted to the local optical properties [74, 17] and sensitively depend on the atmospheric correction models [76, 81]. On the contrary, empirical models achieve good estimation results on distinct water bodies with less influence by the atmosphere. However, they can be rarely employed for unknown water bodies [17].

A water body's spectrum mainly depends on three occurring parameters in the water column: chlorophyll, NAP, and CDOM [28]. The parameters' influence on the spectrum was exemplified in Section 2.1.2). Inland water bodies mainly contain all three of them, so they are called optically complex with multiple overlying spectral features [33]. Therefore, continuous water parameter retrieval is a nonlinear regression task. ML approaches are designed for such complex estimation tasks. Hence, this thesis aimed to present a generalized, data-driven ML approach that can be applied to multiple water bodies. Hereby, several water parameters were investigated to retrieve, such as chlorophyll *a*, CDOM, turbidity, or differentiation between algae classes. However, the main focus was on estimating the chlorophyll *a* concentration. The ML techniques and the applied algorithms were exemplified in Chapter 5. Several ML approaches have been tested on different in situ-sampled datasets to find the most suitable approach for the generalization task.

For the examination of the ML approaches, three ground-based datasets were collected within the thesis' scope to build and evaluate these models. Their characteristics were intensively described in Chapter 4. Hereby, a dataset is set together of multiple datapoints, containing the reflectance values and water parameter values. For each of the subsequent datasets, the heterogeneity of the water constituents increases, so the applied models' challenge increases as well. The first dataset consists of datapoints solely conducted on the River Elbe (see Section 4.1). Subsequently, the second dataset contains datapoints of eleven different inland water bodies with various chlorophyll *a* concentrations (see Section 6.2). A third dataset is entirely simulated with the WASI tool, an analytical model [58] (see Section 4.3).

Three different study setups were designed to train and test various ML models on these datasets. The spectral data was employed as input, whereas the water parameters were estimated as output. With each subsequent setup, the demand for the models' generalization ability increased, which represented the main target of the thesis. Besides the generalization, multiple research goals were investigated in the study setups.

In **Study Setup I** (see Section 6.1), the general functioning of ML models relying on hyperspectral data of the River Elbe were examined. The selected ML algorithms, such as RF, SVM, and k-NN, were conducted in a framework, relying on the same training and test sets. Besides, the feature importance of the RF models was employed to retrieve information

about which spectral features the final model relies on for the respective water parameters. So, possible physical relations between the features and the water parameter can be derived.

The subsequent **Study Setup II** (see Section 6.2) followed the same approach but for multiple water bodies, relying on the second dataset. Thus, the models in the framework needed to show more generalization.

In the final **Study Setup III** (see Section 6.3), the demand for generalization of the models increased even further. The idea was to train the models on the simulated WASI dataset and evaluate their performance on an entirely unknown real-world test dataset. The study's background was that if the models can estimate the chlorophyll *a* concentration of the entirely unknown water bodies, they are also able to perform on most of the global water bodies. DL techniques were applied for such an approach, which was enabled by the vast amount of simulated data provided by the WASI tool.

The investigations in this thesis aimed for a later upscaling approach to real satellite data. In turn, a succeeding upscaling approach with generalized models would allow a global monitoring approach of inland waters.

Since the spectral resolution of the currently available satellites is often limited, a side investigation was conducted concerning its impact on the models' estimation performance in **Study Setup II + III**. Therefore, the spectrometer data were downsampled to resolutions of promising satellite missions. However, a well-performing generalized model on the spectrally downsampled in situ data might not allow the conclusion that the approach will also perform on real satellite data. Nevertheless, the approach indicates the quality that is achievable from the spectral point of view.

7.2 Conclusion - Research Goals

The thesis' findings and conclusions are presented in the following subsections alongside the research goals and their respective questions posted in Section 1.2. Since the study setups were consecutively ordered in terms of complexity, accompanied by rising demand for the models' generalization, the **RG Generalization** is concluded integrated with the other research goals.

7.2.1 Research Goal 1: Models and Parameter

The following subsection focuses on answering **RQ 1**. It relies on all conducted study setups.

- **Can supervised ML models provide a suitable estimation performance of water parameters, especially chlorophyll *a*?**
- **Which of the applied ML models is the most promising to retrieve water parameters with spectral data?**

Regarding the different water parameters estimated in the study setups by multiple models, only the **Study Setup I** on the River Elbe dataset allowed conclusions about different water parameters. On the respective dataset, the best-estimated parameters were CDOM, chlorophyll *a*, and turbidity, with an R^2 score of about 90 %. This finding was surprising since CDOM is seen as the most challenging water parameter to be retrieved with remote sensing data [48]. Besides, the models for estimating the concentration of green algae and diatoms also achieved good estimation results with an R^2 score of slightly below 90 %. Thus, it seems more challenging to relate the chlorophyll *a* concentration to the related algae classes. Nevertheless, estimating diatoms and green algae and differentiating their concentrations with ML models was an entirely new approach. However, the models applied on the River Elbe dataset must be seen as highly specialized. Hence, they will most likely perform well only on the respective water body and likely for the specific period and parameter range. In context with the literature, the estimation results have shown satisfying results regarding the respective parameters. Nevertheless, specialized BR approaches from the literature often provide good estimation results for the respective parameters as well.

More generalized models exist only for the chlorophyll *a* estimation on the subsequent study setups. Regarding the chlorophyll *a* estimations in the other two setups, at least one model exists that provided a good estimation quality. In **Study Setup II**, a higher degree of generalization is assumed since the models were trained on multiple water bodies. In spite of this setup, the R^2 score remained high for most of the models with about 85 %. With respect to **Study Setup III**, the highest degree of generalization for the chlorophyll *a* retrieval model was assumed. Despite the restrictions for the models, the best-performing one still achieved a satisfying R^2 score of 81.9 %. To sum up and answer the first part of the research question: **Can supervised ML models provide a suitable estimation performance of water parameters, especially chlorophyll *a*?**, yes, the ML models provided promising estimation results for all applied parameters. However, only for chlorophyll *a*, the question can be answered with yes in the context of generalization. For the other parameters, it could not be proven in the applied setups.

Focusing on the ML approaches in general, throughout the different study setups, different observations can be found. In the first two study setups, most of the applied ML approaches

performed well independent of the target parameter, whereas their performance difference was relatively low. However, in the third setup, the 1D CNN outperformed the others significantly. Thus, one conclusion is that the models' performance strongly depends on the dataset's demand for generalization. For a single water body, the estimation results were rarely influenced by the chosen ML approach. However, for the challenging approach in **Study Setup III**, only the 1D CNN could have handled the appearing difficulties. This task was not feasible with acceptable performance for the RF-based model and the ANN. Compared with the selected empirical approach by Moses et al. [91], the 1D CNN showed a significant outperformance. The good performance concerning the generalization is most likely related to the typical properties of a CNN in relying on the features provided by its filters. Thus, the model deeply investigated the spectrum's shape without being influenced by noise occurring as an offset. This property makes the 1D CNN resistant against overfitting.

Answering the research question: **Which of the applied ML models is the most promising to retrieve water parameters with spectral data?**, is difficult since the best model highly depends on the degree of generalization. For estimating distinct or multiple water bodies' chlorophyll *a* concentration, models provided by shallow learning approaches such as a RF, a SVM, and a shallow ANN were sufficient if they knew the water bodies. However, if they were applied on entirely unknown water bodies, the 1D CNN was the most promising approach.

7.2.2 Research Goal 2: Feature Importance

The following subsection focuses on answering **RQ 2**.

- **Do the applied ML models rely on similar features as the empirical retrieval algorithms in the literature?**

The findings concerning the feature importance rely only on the RF algorithm and refer to **Study Setup I + II**. The idea of this research goal was to supervise the models in terms of their feature selection. ML approaches are often seen as a black box. Explaining their primary features might give the ML approaches more confidence.

Regarding the selected important features of the specialized RF models in **Study Setup I**, mainly the physically related features to the respective parameter are picked by the models (cf. Figure 6.3, Section 6.1.3). Nevertheless, the models also chose some features that cannot be distinctly explained. However, these models have shown some performance increase when building them on their principal components. This increase indicates that these models suffer some overfitting that might be related to the pretended selection of important features. These models relied on a 4 nm resolution, so the data is high-dimensional. Thus, they are susceptible to overfitting, especially such specialized ones.

In **Study Setup II**, the resolution is a bit coarser, and simultaneously, the generalization demand is higher due to the application on multiple water bodies. Thus more distinguishable features are seen for chlorophyll *a* estimation by the RF models on **Study Setup II** (cf. Figure 6.9, Section 6.2.8). Even for more specialized models trained on a single water body, the features are more distinctive than for the models trained on the River Elbe. This finding might be explained by a longer measurement period containing several months over two years, with changing conditions of the water constituents.

Comparing the selected features for chlorophyll *a* between the River Elbe from **Study Setup I**, the selected water bodies of **Study Setup II**, and the model representing all water bodies, it is obvious that the RF model on all water bodies relies on less selected features. This finding might emphasize the generalization ability of the RF model on all investigated water bodies.

One interesting finding is that the models rather pick the slope of the peak at around 720 nm than the peak itself or the chlorophyll *a* absorption feature at 670 nm. A possible explanation might be that this peak shifts towards longer wavelengths with higher chlorophyll *a* concentrations [38].

Additionally, for water bodies with high chlorophyll *a* concentrations, the models have also chosen features from the peak at 810 nm related to detritus [56]. This selection can make sense when the detritus mainly consists of degenerating phytoplankton materials. Then again, it might be related to the chlorophyll *a* concentration.

The research question: **Do the applied ML models rely on similar features as the empirical retrieval algorithms in the literature?**, can finally be answered with yes, at least for the RF model. The selected important features by the RF models are known to be physically and logically related to the target parameters and mainly similar to the ones applied in the BR approaches in the literature. Supplementary, the RF models extend the features, especially for the shifting peak and its slope at around 720 nm concerning chlorophyll *a*. Besides, it is also possible to learn from the models' feature selection for the empirical approaches. Regarding the multiple selected features strengthens the approach of applying ML models. Much more information in the whole spectrum can be employed to retrieve a target parameter than selecting just some bands like in the empirical approaches.

7.2.3 Research Goal 3: Generalization

The following subsection focuses on the generalization to answer **RQ 3**.

- **Can the applied ML models generalize on multiple water bodies? Is it even possible to build a model that is able to estimate the chlorophyll *a* concentration of completely unknown water bodies?**

Models that can estimate water quality parameters for most of the possible water bodies are essential for a global monitoring approach. Although the estimation performance of generalized models is lower than that of specialized models, the latter is not feasible since they rely on a long data history for their calibration, which is not given for most of the water bodies. However, a universal, generalized model with an acceptable estimation performance on multiple water bodies does not exist yet. Pahlevan et al. [19] presented one approach for such a generalized, neural network-based model, relying on Sentinel-2 or Sentinel-3 data. Their approach outperformed the common BR approaches. Nevertheless, their performance for the high and the low chlorophyll *a* concentrations was not convincing. Ansper and Alikas [119] and Neil et al. [118] suggest semi-generalized models that achieve good estimation results on specific optical water types.

Regarding the estimation performance of the models in **Study Setup II**, it can be concluded that the models generalized well on the applied water bodies. However, they already saw datapoints of those water bodies during their training process. Thus, the maximum degree of generalization cannot be assumed. Nevertheless, the research question: **Can the applied ML models generalize on multiple water bodies?**, can be answered with yes.

Focusing on **Study Setup III**, the 1D CNN trained on the simulated data clearly generalized well since it estimated most of the unknown water bodies with reliable performance. In combination with the well-performing 1D CNN, this study setup allows answering the research question: **Is it even possible to build a model that is able to estimate the chlorophyll *a* concentration of completely unknown water bodies?**, with yes. However, this is only valid for the hyperspectral EnMAP resolution. With the focus on the Sentinel-2 resolution, this was not feasible. One possible option to achieve good estimation results with the Sentinel-2 resolution might be the semi-generalized models, suggested by Neil et al. [118]. Therefore, the task might be less challenging and hence, more suitable for the multispectral resolution.

7.2.4 Research Goal 4: Spectral Resolution

The following subsection focuses on answering **RQ 4**.

- **How much can the spectral resolution decrease to still get a suitable estimation performance by the models?**

The RQ's background refers to a later application of the thesis' approach on real satellite data. It focuses on the trade-off between spectral and spatial resolution. Sentinel-2, for example, provides a fine spatial resolution that would allow monitoring of small-sized water bodies. Additionally, it has a suitable temporal resolution. However, it is questionable if its spectral resolution is sufficient for a generalized approach. Nevertheless, the models relying on the

spectrally fine resolved hyperspectral satellite resolutions, such as EnMAP, would provide better estimation results for water parameters. Unfortunately, they suffer under a high temporal repetition time that impedes the monitoring approach. Besides, EnMAP is not in orbit yet.

The analysis of **RQ 4** concerning the impact of the spectral resolution on the models' estimation performance relied on the **Study Setups II + III**. In the **Study Setup II**, different continuously downsampled hyperspectral resolutions and various satellite resolutions were investigated to retrieve the chlorophyll *a* concentration. The first investigation revealed that each applied hyperspectral resolution was sufficient for the respective estimation task. The performance differences between a constant 4 nm resolution and a constant 20 nm were nearly negligible. Therefore, the spectral data were downsampled to different satellite resolutions to investigate their potential for a later application on real satellite data. The evaluation of the models on both multispectral Sentinel satellite resolutions revealed that, chlorophyll *a* monitoring would be feasible from the spectral view. Their models' performance was only slightly worse than the one of the hyperspectral missions Hyperion and EnMAP. However, the performance of the Landsat-based models decreased compared to the others. Therefore, it is to conclude that band positioning is nearly as important as the general bandwidth itself. Especially bands around the chlorophyll *a* absorption feature and the follow-up reflectance peak are crucial for its estimation. The poor estimation performance of the models based on the Landsat resolutions corresponds to the findings in analyzing the feature importance. However, for the generalizing approach to entirely unknown water bodies in **Study Setups III**, the performance distance between EnMAP and Sentinel-2 increased significantly. One explanation for this finding might be that the only model that can handle this challenge is the 1D CNN. The strength of that model is on evaluating the spectrum intensively with its various filters. However, this is limited for the Sentinel-2 resolution, relying on only nine spectral bands.

The research question: **How much can the spectral resolution decrease to still get a suitable estimation performance by the models?**, can only be answered accompanied by the demand for generalization. The spectral resolution of Sentinel-2 is sufficient, but only when little generalization is required. Thus, for the Sentinel-2 resolution, semi-generalized models that perform well on a certain water type might be suitable. On the contrary, the hyperspectral EnMAP resolution is sufficient to train entirely generalized models for the application on entirely unknown water bodies.

7.2.5 Research Goal 5: Transferability

The following subsection focuses on the last **RQ 5**.

- **Can a model trained on simulated data be able to estimate the chlorophyll *a* concentrations of real-world water bodies?**

- **Is the transferability between simulated and measured data given?**

The need for simulated data comes with the approach to train a DL model to solve the generalization task for retrieving the chlorophyll *a* concentration of multiple entirely unknown water bodies. Unfortunately, for the training of the DL model, the amount of measured data in the field campaigns is not sufficient. This vast demand for data was covered by simulations with the WASI tool. The approach for the simulation was to vary the relevant parameters intensively to comprise most of the global water bodies. Moreover, not only the concentration of the water ingredients has been varied but also atmospheric parameters that influence the surface reflectance. Most of the approaches in the literature rely on data with corrected surface reflectance. This correction was not feasible for the collected dataset due to its measuring setup. Therefore, various surface reflectance in the simulated dataset was added.

This research question solely affects the **Study Setups III**. Regarding the estimation performance of the 1D CNN on the entirely unknown test set, with an R^2 score of 81.9 % and a MAE of 6.4 $\mu g\,L^{-1}$, clearly reveals the feasibility of this approach. However, only the 1D CNN performed well on that task. Even though the other models could learn the simulated data, they were not capable of performing the transfer and the generalization task. One important finding of this approach is that surface reflectance must not be corrected to achieve a good estimation result with the 1D CNN. This finding is essential since surface reflectance is a very challenging task for inland waters [32, 28].

To focus on the first part of the research question: **Can a model trained on simulated data be able to estimate the chlorophyll *a* concentrations of real-world water bodies?** Yes, but only the 1D CNN combined with the hyperspectral EnMAP resolution could succeed in that task. The second part of the research question: **Is the transferability between simulated and measured data given?**, is difficult to answer. This transfer was only feasible for the 1D CNN, relying on its features that neglect the absolute reflectance values. However, for the other models, the transfer was not feasible. Moreover, it was not possible for the models trained on the Sentinel-2 resolution.

7.3 General Conclusion

The three study setups conducted within the scope of this thesis provided many innovative findings for the inland water remote sensing community. ML models based on spectral data have shown an excellent estimation performance for multiple water parameters. They are indeed suitable to face the non-linear regression challenge in estimating chlorophyll *a* concentrations in multiple water bodies solely with spectral data. Depending on the challenge of the underlying dataset, multiple supervised learning approaches are suitable. Even a differentiation between algae classes was feasible, however only shown for specialized

models. Unfortunately, cyanobacteria could not be investigated with a ML approach due to the lack of the respective data. However, a succeeding differentiation between the phytoplankton classes may also encompass the cyanobacteria since they show at 620 nm the most distinct spectral feature among all phytoplankton classes.

The feature importance provided by the RF models revealed essential spectral bands, such as the 810 nm feature for chlorophyll *a*, that have been rarely employed in the empirical approaches in the related literature. An additional finding concerning the important features provided by the RF models is their distribution. The more generalized a model must be to solve the estimation task, the more narrow the selected features. However, this finding concerns only the investigated SpecWa dataset that does not represent, e.g., humic water bodies that might rely on other spectral features. Besides, it is not determined which are the most important features for the 1D CNN. In general, it still seems crucial to provide the models the full range of the spectral data to definitively do not lose information.

Regarding a potential satellite designed only for monitoring inland water bodies, a model's feature importance analysis of multiple representatively selected water bodies can provide valuable information about the essential bands. For example, the shifting peak at 700 nm provides crucial information about a water body's chlorophyll *a* concentration. However, this information is rarely detectable with the respective two broad bands on Sentinel-2. One option for a potential satellite would be to increase the spectral resolution towards a hyperspectral one at the respective area, even though the spatial resolution would decrease.

Focusing on the generalization, from the spectral point of view, an entirely generalized model that can estimate unknown water bodies' chlorophyll *a* concentration is feasible. This generalization was shown with the 1D CNN, at least for the evaluated SpecWa dataset, containing eleven inland water bodies. However, applied to real satellite data, the atmospheric correction might impede this approach. Besides, this generalization was only shown for hyperspectral data, for which only a few satellites exist yet. Regarding multispectral satellite resolutions, such as Sentinel-2, an entirely generalized model was not feasible, even without the impeding impact of the atmosphere. However, semi-generalized models, inspired by the research of Ansper and Alikas [119] and Neil et al. [118], might be feasible.

Regarding the spectral resolution, not only the bandwidth but also the positioning of the bands is essential for good model performance. This finding was even shown on the medium generalized models relying on the SpecWa dataset. Without spectral bands at the spectral range around the local maximum at 700 nm, a reliable estimation performance for chlorophyll *a* is not feasible for multiple water bodies.

Referring to the simulated data, this approach clearly revealed their potential for building or improving models. The resulting 1D CNN performed well on completely unknown data. However, the models' training process on the simulated data can still

be improved. One option might be to enhance the data simulation by multiple additional phytoplankton classes, benthic substrates, various suspended solids, or any kind of noise that makes the data more natural.

One important finding is the 1D CNN's well-performing on radiance reflectance data. Thus, the presented approach makes a surface reflectance correction unnecessary unless the models are trained for various surface reflectance conditions. This outcome is essential since surface reflectance correction is a huge challenge, especially for satellite images over inland water bodies [32, 28]. The 1D CNN might even further show its strength on real hyperspectral satellite data for a later upscaling approach due to its noise-resistant feature exploitation.

With the launch of EnMAP in the near future and a respectively available atmospheric correction model, the proposed monitoring approach may be fulfilled. Then the presented 1D CNN can be coupled with the atmospherically corrected satellite images to retrieve the chlorophyll *a* concentration of affected water bodies. If a surface reflectance model is applied, the 1D CNN needs to be slightly adapted to remote sensing reflectance as input data. Such an application would close the loop, beginning with spectrometer measurements to the final application on satellite images.

7.4 Outlook

The multiple findings discovered in the scope of this thesis motivates further research. Filazzola et al. [155] recently published a huge dataset covering chlorophyll measurements of 11,959 lakes across 72 countries. A coupling of the in situ measured chlorophyll data with satellite images, provided by, e.g., Sentinel-2 or Sentinel-3, would provide sufficient data to employ a DL approach under real-world conditions. Multiple options for such an upscaling approach exist. One possible option is employing two neural network-based models. The first model is trained for correcting the atmosphere, including the surface reflectance, whereas the second model estimates the chlorophyll values based on the corrected remote sensing signal. Such an approach was presented by [19]. The second model for estimating the water parameters is similar to the 1D CNN presented in the third study setup. However, the 1D CNN has shown that it does not need remote sensing reflectance for a good estimation quality. It also performs on radiance reflectance that includes the surface reflectance. If the atmospheric correction model provides better performance without correcting the surface reflectance of the water bodies, the resulting radiance reflectance will be exploited by a 1D CNN. Hereby, also simulated data, as provided by the WASI tool might help to train the second model. A comparison between both modular approaches will be interesting since a successful conduction would provide a basis for a consequent monitoring approach for the global inland water bodies.

An alternative option to avoid the coupling between two models might be an application within a single model. Then, the approach relies on uncorrected top of the atmosphere data. So the model directly learns the linkage between the measured spectrum and the water parameters, including the atmospheric noise. However, there are many degrees of freedom in such an approach, so there is again a vast demand for data.

Since most of the available satellites provide multispectral data, the performance of the models relying on that data must be enhanced to retrieve reliable chlorophyll values. One option for improving the neural networks' performance might be working with pre-trained models. Hereby the models can be pre-trained on WASI data and in a second training round adapted to real-world data, similar to [156]. This approach would be especially promising for the Sentinel-2-based models. In the thesis was shown that even these models were able to learn the WASI-generated data. However they could not perform the transfer to the real-world data.

Alternatively, another promising option for the Sentinel-2 based approaches are the already-mentioned semi-generalized approach, focusing on optical water types, inspired by [117, 118]. To conclude the given outlook, there are multiple options for an upscaling approach that may allow consequent monitoring of inland water bodies.

Bibliography

[1] Frank Biermann, Norichika Kanie, and Rakhyun E Kim. “Global governance by goal-setting: the novel approach of the UN Sustainable Development Goals”. In: *Current Opinion in Environmental Sustainability* 26 (2017), pp. 26–31 (cit. on p. 1).

[2] Water Framework Directive. “Water Framework Directive”. In: *Journal reference OJL* 327 (2000), pp. 1–73 (cit. on p. 1).

[3] Claudia Copeland. “Clean Water Act: a summary of the law”. In: Congressional Research Service, Library of Congress Washington, DC. 1999 (cit. on p. 1).

[4] Robert E Carlson. “A trophic state index for lakes”. In: *Limnology and oceanography* 22.2 (1977), pp. 361–369 (cit. on p. 1).

[5] Roland I Hall, Peter R Leavitt, Roberto Quinlan, Aruna S Dixit, and John P Smol. “Effects of agriculture, urbanization, and climate on water quality in the northern Great Plains”. In: *Limnology and Oceanography* 44.3part2 (1999), pp. 739–756 (cit. on p. 1).

[6] Miles J Furnas. “In situ growth rates of marine phytoplankton: approaches to measurement, community and species growth rates”. In: *Journal of Plankton Research* 12.6 (1990), pp. 1117–1151 (cit. on p. 1).

[7] JoAnn M Burkholder, David A Dickey, Carol A Kinder, et al. “Comprehensive trend analysis of nutrients and related variables in a large eutrophic estuary: a decadal study of anthropogenic and climatic influences”. In: *Limnology and Oceanography* 51.1part2 (2006), pp. 463–487 (cit. on p. 1).

[8] David W Schindler. “Evolution of phosphorus limitation in lakes”. In: *Science* 195.4275 (1977), pp. 260–262 (cit. on p. 1).

[9] Donald M Anderson, Patricia M Glibert, and Joann M Burkholder. “Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences”. In: *Estuaries* 25.4 (2002), pp. 704–726 (cit. on p. 1).

[10] Michael A Mallin, Hans W Paerl, Joseph Rudek, and Paul W Bates. “Regulation of estuarine primary production by watershed rainfall and river flow”. In: *Marine Ecology-Progress Series* 93 (1993), pp. 199–199 (cit. on p. 1).

[11] József Kovács, Péter Tanos, Gábor Várbıró, et al. “The role of annual periodic behavior of water quality parameters in primary production–Chlorophyll-a estimation”. In: *Ecological Indicators* 78 (2017), pp. 311–321 (cit. on p. 1).

[12] Joana Amorim Visco, Laure Apothéloz-Perret-Gentil, Arielle Cordonier, et al. “Environmental monitoring: inferring the diatom index from next-generation sequencing data”. In: *Environmental science & technology* 49.13 (2015), pp. 7597–7605 (cit. on p. 1).

[13] M G Kelly and Brian A Whitton. “The trophic diatom index: a new index for monitoring eutrophication in rivers”. In: *Journal of applied phycology* 7.4 (1995), pp. 433–444 (cit. on p. 1).

[14] Ingrid Chorus and Martin Welker. *Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management*. Taylor & Francis, 2021 (cit. on p. 2).

[15] Blake A Schaeffer, Kelly G Schaeffer, Darryl Keith, et al. “Barriers to adopting satellite remote sensing for water quality management”. In: *International Journal of Remote Sensing* 34.21 (2013), pp. 7534–7544 (cit. on p. 2).

[16] Stephanie CJ Palmer, Tiit Kutser, and Peter D Hunter. *Remote sensing of inland waters: Challenges, progress and future directions*. 2015 (cit. on pp. 2, 17, 22, 23, 25, 99).

[17] Mark W Matthews. “A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters”. In: *International Journal of Remote Sensing* 32.21 (2011), pp. 6855–6899 (cit. on pp. 2, 4, 19, 21, 22, 99, 100).

[18] Charles Verpoorter, Tiit Kutser, David A Seekell, and Lars J Tranvik. “A global inventory of lakes based on high-resolution satellite imagery”. In: *Geophysical Research Letters* 41.18 (2014), pp. 6396–6402 (cit. on pp. 3, 22).

[19] Nima Pahlevan, Brandon Smith, John Schalles, et al. “Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach”. In: *Remote Sensing of Environment* 240 (2020), p. 111604 (cit. on pp. 4, 22, 23, 39, 95, 105, 109).

[20] Philipp M Maier, Stefan Hinz, and Sina Keller. “Estimation of Chlorophyll A, Diatoms and Green Algae Based on Hyperspectral Data with Machine Learning Approaches”. In: *Tagungsband der 37. Wissenschaftlich-Technische Jahrestagung der DGPF e.V.* Vol. 27. Munich, Germany: Deutsche Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation, 2018, pp. 49–57 (cit. on pp. 5, 26, 129, 130).

[21] Philipp M Maier and Sina Keller. “Machine learning regression on hyperspectral data to estimate multiple water parameters”. In: *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2018, pp. 1–5 (cit. on pp. 5, 26, 29, 60, 62–64, 129).

[22] Sina Keller, Philipp M Maier, Felix M Riese, et al. “Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity”. In: *International journal of environmental research and public health* 15.9 (2018), p. 1881 (cit. on pp. 5, 26–28, 60, 65, 66, 129).

[23] Philipp M Maier and Sina Keller. “Estimating chlorophyll a concentrations of several inland waters with hyperspectral data and machine learning models”. In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4 (2019) (cit. on pp. 6, 30, 31, 35, 36, 42, 70–74, 93, 129).

[24] Philipp M Maier and Sina Keller. “Application of different simulated spectral data and machine learning to estimate the chlorophyll a concentration of several inland waters”. In: *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2019, pp. 1–5 (cit. on pp. 6, 30, 35, 36, 42, 43, 70, 71, 76, 77, 94, 129).

[25] Philipp M Maier and Sina Keller. “SpecWa: Spectral remote sensing data and chlorophyll a values of inland waters”. In: *GFZ Data Services* (2020) (cit. on pp. 6, 29, 30, 32, 33, 35–37, 71, 129).

[26] Philipp M Maier, Sina Keller, and Stefan Hinz. “Deep Learning with WASI Simulation Data for Estimating Chlorophyll a Concentration of Inland Water Bodies”. In: *Remote Sensing* 13.4 (2021), p. 718 (cit. on pp. 6, 30, 34–38, 40–42, 55, 57, 83, 85–92, 129).

[27] Curtis D Mobley and Charles D Mobley. *Light and water: radiative transfer in natural waters*. Academic press, 1994 (cit. on p. 10).

[28] Peter Gege. “WASI5 Manual”. In: (2019) (cit. on pp. 10–15, 37, 41, 100, 107, 109).

[29] Susan Kay, John D Hedley, and Samantha Lavender. “Sun glint correction of high and low spatial resolution images of aquatic scenes: a review of methods for visible and near-infrared wavelengths”. In: *Remote sensing* 1.4 (2009), pp. 697–730 (cit. on p. 10).

[30] Victor Martinez-Vicente, Stefang Simis, R Alegre, P E Land, and S B Groom. “Above-water reflectance for the evaluation of adjacency effects in Earth observation data: initial results and methods comparison for near-coastal waters in the Western Channel, UK”. In: *Journal of the European Optical Society-Rapid publications* 8 (2013) (cit. on p. 10).

[31] Peter Gege and Philipp Grötsch. “A spectral model for correcting sunglint and skyglint”. In: *Proceedings of Ocean Optics XXIII 2016* (2016), pp. 1–10 (cit. on pp. 10, 11, 37, 93).

[32] Dierdre A Toole, David A Siegel, David W Menzies, Michael J Neumann, and Raymond C Smith. “Remote-sensing reflectance determinations in the coastal ocean environment: impact of instrumental characteristics and environmental variability”. In: *Applied Optics* 39.3 (2000), pp. 456–469 (cit. on pp. 11, 107, 109).

[33] André Morel and Louis Prieur. “Analysis of variations in ocean color 1”. In: *Limnology and oceanography* 22.4 (1977), pp. 709–722 (cit. on pp. 12, 13, 93, 100).

[34] Andre Morel, Bernard Gentili, Malik Chami, and Joséphine Ras. “Bio-optical properties of high chlorophyll Case 1 waters and of yellow-substance-dominated Case 2 waters”. In: *Deep Sea Research Part I: Oceanographic Research Papers* 53.9 (2006), pp. 1439–1459 (cit. on p. 12).

[35] G Quibell. “The effect of suspended sediment on reflectance from freshwater algae”. In: *International Journal of Remote Sensing* 12.1 (1991), pp. 177–182 (cit. on p. 12).

[36] Peter D Hunter, Andrew N Tyler, Mátyás Présing, Attila W Kovács, and Tom Preston. “Spectral discrimination of phytoplankton colour groups: The effect of suspended particulate matter and sensor spectral resolution”. In: *Remote Sensing of Environment* 112.4 (2008), pp. 1527–1544 (cit. on pp. 13, 14, 17, 66).

[37] R A Neville and J F R Gower. “Passive remote sensing of phytoplankton via chlorophyll α fluorescence”. In: *Journal of Geophysical Research* 82.24 (1977), pp. 3487–3493 (cit. on pp. 13, 66, 78).

[38] Anatoly Gitelson. “The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration”. In: *International Journal of Remote Sensing* 13.17 (1992), pp. 3367–3373 (cit. on pp. 13, 20, 82, 83, 104).

[39] M Kishino, S Sugihara, and N Okami. “Theoretical analysis of the in-situ fluorescence of chlorophyll-a on the underwater spectral irradiance”. In: *Bulletein de la Societe Franco-Japanaise d’Oceanographie* 24 (1986), pp. 130–138 (cit. on p. 13).

[40] Willem L Vos, Marcel Donze, and H Buiteveld. *On the Reflectance Spectrum of Algae in Water: The Nature of the Peak at 700nm and Its Shift with Varying Algal Concentration*. Delft University of Technology, Faculty of Civil Engineering, 1986 (cit. on p. 13).

[41] Anatoly Gitelson and KY Kondratyev. “On the mechanism of formation of maximum in the reflectance spectra near 700 nm and its application for remote monitoring of water quality”. In: *Transactions Doklady of the USSR Academy of Sciences: Earth Science Sections*. Vol. 306. 1991, pp. 1–4 (cit. on p. 13).

[42] Linhong Kou, Daniel Labrie, and Petr Chylek. “Refractive indices of water and ice in the 0.65-to 2.5-μm spectral range”. In: *Applied optics* 32.19 (1993), pp. 3531–3540 (cit. on p. 13).

[43] Robin M Pope and Edward S Fry. “Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements”. In: *Applied optics* 36.33 (1997), pp. 8710–8723 (cit. on p. 13).

[44] Stefan GH Simis, Antonio Ruiz-Verdú, Jose Antonio Domınguez-Gómez, et al. “Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass”. In: *Remote Sensing of Environment* 106.4 (2007), pp. 414–427 (cit. on pp. 14, 20).

[45] Donald A Bryant. “The Photoregulated Expression of Multiple Phycocyanin Species: A General Mechanism for the Control of Phycocyanin Synthesis is Chromatically Adapting Cyanobacteria”. In: *European Journal of Biochemistry* 119.2 (1981), pp. 425–429 (cit. on p. 14).

[46] George R Aiken, Diane M McKnight, Patrick MacCarthy, and RL Wershaw. *Humic substances in soil, sediment, and water: geochemistry, isolation, and characterization*. Vol. 1. Wiley-Interscience, 1985 (cit. on p. 14).

[47] Patrick L Brezonik, Leif G Olmanson, Jacques C Finlay, and Marvin E Bauer. “Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters”. In: *Remote Sensing of Environment* 157 (2015), pp. 199–215 (cit. on pp. 14, 20, 21).

[48] Kevin D Menken, Patrick L Brezonik, and Marvin E Bauer. “Influence of chlorophyll and colored dissolved organic matter (CDOM) on lake reflectance spectra: Implications for measuring lake properties by remote sensing”. In: *Lake and Reservoir Management* 22.3 (2006), pp. 179–190 (cit. on pp. 14, 15, 21, 68, 102).

[49] Gunnar Nyquist. *Investigation of some optical properties of seawater with special reference to lignin sulfonates and humic substances [Swedish coastal waters, Baltic Sea]*. 1979 (cit. on p. 15).

[50] Annick Bricaud, Andre Morel, and Louis Prieur. “Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domains 1”. In: *Limnology and oceanography* 26.1 (1981), pp. 43–53 (cit. on p. 15).

[51] Kendall L Carder, Robert G Steward, George R Harvey, and Peter B Ortner. “Marine humic and fulvic acids: Their effects on remote sensing of ocean chlorophyll”. In: *Limnology and oceanography* 34.1 (1989), pp. 68–81 (cit. on p. 15).

[52] Kari Kallio, Tiit Kutser, Tuula Hannonen, et al. “Retrieval of water quality from airborne imaging spectrometry of various lake types in different seasons”. In: *Science of the Total Environment* 268.1-3 (2001), pp. 59–77 (cit. on p. 15).

[53] Tiit Kutser, Antti Herlevi, Kari Kallio, and Helgi Arst. “A hyperspectral model for interpretation of passive optical remote sensing data from turbid lakes”. In: *Science of the Total Environment* 268.1-3 (2001), pp. 47–58 (cit. on p. 15).

[54] Andrew D Eaton, Lenore S Clesceri, Eugene W Rice, Arnold E Greenberg, Mary Ann H Franson, et al. “Standard methods for the examination of water and wastewater”. In: *American public health association* 1015 (2005) (cit. on p. 15).

[55] David Doxaran, Jean-Marie Froidefond, Samantha Lavender, and Patrice Castaing. “Spectral signature of highly turbid waters: Application with SPOT data to quantify suspended particulate matter concentrations”. In: *Remote sensing of Environment* 81.1 (2002), pp. 149–161 (cit. on p. 15).

[56] RF Arenz Jr, William Lewis Jr, and JF SAUNDERS III. “Determination of chlorophyll and dissolved organic carbon from reflectance data for Colorado reservoirs”. In: *International Journal of Remote Sensing* 17.8 (1996), pp. 1547–1565 (cit. on pp. 15, 67, 79, 104).

[57] Marcel Babin, Dariusz Stramski, Giovanni M Ferrari, et al. “Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe”. In: *Journal of Geophysical Research: Oceans* 108.C7 (2003) (cit. on p. 15).

[58] Peter Gege and Andreas Albert. “A tool for inverse modeling of spectral measurements in deep and shallow waters”. In: *Remote sensing of aquatic coastal ecosystem processes*. Springer, 2006, pp. 81–109 (cit. on pp. 16, 19, 100).

[59] Claudia Giardino, Gabriele Candiani, Mariano Bresciani, et al. “BOMBER: A tool for estimating water quality and bottom properties from remote sensing images”. In: *Computers & Geosciences* 45 (2012), pp. 313–318 (cit. on p. 16).

[60] Curtis D Mobley and Lydia K Sundman. “HYDROLIGHT 5 ECOLIGHT 5”. In: *Sequoia Scientific Inc* (2008) (cit. on p. 16).

[61] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, et al. “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services”. In: *Remote sensing of Environment* 120 (2012), pp. 25–36 (cit. on p. 16).

[62] Karen Fletcher. *Sentinel-3: ESA’s global land and ocean mission for GMES operational services*. ESA Communications, 2012 (cit. on pp. 16, 17).

[63] Arnold G Decker, Tim Malthus, MM Wijnen, and E Seyhan. “The effect of spectral bandwidth and positioning on the spectral signature analysis of inland waters”. In: *Remote Sensing of Environment* 41.2-3 (1992), pp. 211–225 (cit. on pp. 17, 78).

[64] Richard Beck, Shengan Zhan, Hongxing Liu, et al. “Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations”. In: *Remote Sensing of Environment* 178 (2016), pp. 15–30 (cit. on pp. 17, 43).

[65] Luis Guanter, Hermann Kaufmann, Karl Segl, et al. “The EnMAP spaceborne imaging spectroscopy mission for earth observation”. In: *Remote Sensing* 7.7 (2015), pp. 8830–8857 (cit. on p. 17).

[66] Rudolph W Preisendorfer. “Application of radiative transfer theory to light measurements in the sea”. In: *Union Geod. Geophys. Inst. Monogr.* 10 (1961), pp. 11–30 (cit. on p. 19).

[67] Howard R Gordon, Otis B Brown, and Michael M Jacobs. “Computed relationships between the inherent and apparent optical properties of a flat homogeneous ocean”. In: *Applied optics* 14.2 (1975), pp. 417–427 (cit. on p. 19).

[68] Kendall L Carder, FR Chen, ZP Lee, SK Hawes, and D Kamykowski. “Semianalytic Moderate-Resolution Imaging Spectrometer algorithms for chlorophyll a and absorption with bio-optical domains based on nitrate-depletion temperatures”. In: *Journal of Geophysical Research: Oceans* 104.C3 (1999), pp. 5403–5421 (cit. on p. 19).

[69] Arnold G Dekker, HJ Hoogenboom, LM Goddijn, and TJM Malthus. “The relation between inherent optical properties and reflectance spectra in turbid inland waters”. In: *Remote Sensing Reviews* 15.1-4 (1997), pp. 59–74 (cit. on p. 19).

[70] Arnold G Dekker, RJ Vos, and Steef WM Peters. “Comparison of remote sensing data, model results and in situ data for total suspended matter (TSM) in the southern Frisian lakes”. In: *Science of the Total Environment* 268.1-3 (2001), pp. 197–214 (cit. on p. 19).

[71] Daniel Odermatt, Thomas Heege, Jens Nieke, Mathias Kneubühler, and Klaus Itten. “Water Quality Monitoring for Lake Constance with a Physically Based Algorithm for MERIS Data”. In: *Sensors (Basel, Switzerland)* 8.8 (2008), pp. 4582–4599 (cit. on p. 19).

[72] Helmut Schiller and Roland Doerffer. “Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data”. In: *International journal of remote sensing* 20.9 (1999), pp. 1735–1746 (cit. on p. 19).

[73] Peter Gege and Arnold G Dekker. “Spectral and radiometric measurement requirements for inland, coastal and reef waters”. In: *Remote Sensing* 12.14 (2020), p. 2247 (cit. on pp. 19, 20).

[74] Zhong-Ping Lee et al. *Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications.* International Ocean Colour Coordinating Group (IOCCG), 2006 (cit. on pp. 19, 100).

[75] Andreas Albert and Curtis D Mobley. “An analytical model for subsurface irradiance and remote sensing reflectance in deep and shallow case-2 waters”. In: *Optics Express* 11.22 (2003), pp. 2873–2890 (cit. on p. 19).

[76] ZhongPing Lee, Kendall L Carder, and Robert A Arnone. “Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters”. In: *Applied optics* 41.27 (2002), pp. 5755–5772 (cit. on pp. 19, 20, 100).

[77] Curtis D Mobley. “A numerical model for the computation of radiance distributions in natural waters with wind-roughened surfaces”. In: *Limnology and oceanography* 34.8 (1989), pp. 1473–1483 (cit. on p. 19).

[78] Peng-Wang Zhai, Yongxiang Hu, Jacek Chowdhary, et al. “A vector radiative transfer model for coupled atmosphere and ocean systems with a rough interface”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 111.7-8 (2010), pp. 1025–1040 (cit. on p. 19).

[79] Roland Doerffer and Helmut Schiller. “The MERIS Case 2 water algorithm”. In: *International Journal of Remote Sensing* 28.3-4 (2007), pp. 517–535 (cit. on p. 19).

[80] Thomas Schroeder, M Schaale, and J Fischer. “Retrieval of atmospheric and oceanic properties from MERIS measurements: A new Case-2 water processor for BEAM”. In: *International Journal of Remote Sensing* 28.24 (2007), pp. 5627–5632 (cit. on p. 19).

[81] Daniel Odermatt, Anatoly Gitelson, Vittorio Ernesto Brando, and Michael Schaepman. “Review of constituent retrieval in optically deep and complex waters from satellite imagery”. In: *Remote sensing of environment* 118 (2012), pp. 116–126 (cit. on pp. 20, 22, 100).

[82] Michael Defoin-Platel and Malik Chami. “How ambiguous is the inverse problem of ocean color in coastal waters?” In: *Journal of Geophysical Research: Oceans* 112.C3 (2007) (cit. on p. 20).

[83] Sampsa Koponen, Jenni Attila, Jouni Pulliainen, et al. “A case study of airborne and satellite remote sensing of a spring bloom event in the Gulf of Finland”. In: *Continental Shelf Research* 27.2 (2007), pp. 228–244 (cit. on pp. 20, 21).

[84] Niklas Strömbeck, Gabriele Candiani, Claudia Giardino, and Eugenio Zilioli. “Water quality monitoring of Lake Garda using multi-temporal MERIS data”. In: *Proc. of MERIS User Workshop*. 2003 (cit. on p. 20).

[85] Herman J Gons, Machteld Rijkeboer, and Kevin G Ruddick. “Effect of a waveband shift on chlorophyll retrieval from MERIS imagery of inland and coastal waters”. In: *Journal of Plankton research* 27.1 (2005), pp. 125–127 (cit. on p. 20).

[86] Kari Kallio, Jenni Attila, Pekka Härmä, et al. “Landsat ETM+ images in the estimation of seasonal lake water quality in boreal river basins”. In: *Environmental management* 42.3 (2008), pp. 511–522 (cit. on pp. 20, 21).

[87] David Doxaran, P Castaing, and SJ Lavender. “Monitoring the maximum turbidity zone and detecting fine-scale turbidity features in the Gironde estuary using high spatial resolution satellite sensor (SPOT HRV, Landsat ETM+) data”. In: *International Journal of Remote Sensing* 27.11 (2006), pp. 2303–2321 (cit. on p. 20).

[88] David Doxaran, J-M Froidefond, and P Castaing. “A reflectance band ratio used to estimate suspended matter concentrations in sediment-dominated coastal waters”. In: *International Journal of Remote Sensing* 23.23 (2002), pp. 5079–5085 (cit. on p. 20).

[89] Peter D Hunter, Andrew N Tyler, Nigel J Willby, and DJ Gilvear. “The spatial dynamics of vertical migration by Microcystis aeruginosa in a eutrophic shallow lake: A case study using high spatial resolution time-series airborne remote sensing”. In: *Limnology and Oceanography* 53.6 (2008), pp. 2391–2406 (cit. on p. 20).

[90] Anatoly Gitelson, G Garbuzov, Ferenc Szilagyi, et al. “Quantitative remote sensing methods for real-time monitoring of inland waters quality”. In: *International Journal of Remote Sensing* 14.7 (1993), pp. 1269–1295 (cit. on pp. 20, 21).

[91] Wesley J Moses, Anatoly A Gitelson, Sergey Berdnikov, and Vasiliy Povazhnyy. “Satellite estimation of chlorophyll-a concentration using the red and NIR bands of MERIS—the Azov sea case study”. In: *IEEE Geoscience and Remote Sensing Letters* 6.4 (2009), pp. 845–849 (cit. on pp. 20, 84, 86, 94, 103).

[92] Wesley J Moses, Anatoly A Gitelson, Sergey Berdnikov, and V Povazhnyy. “Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—successes and challenges”. In: *Environmental research letters* 4.4 (2009), p. 045005 (cit. on p. 20).

[93] Claudia Giardino, Gabriele Candiani, and Eugenio Zilioli. “Detecting chlorophyll-a in Lake Garda using TOA MERIS radiances”. In: *Photogrammetric Engineering & Remote Sensing* 71.9 (2005), pp. 1045–1051 (cit. on p. 20).

[94] Dana Floricioiu, C Riedl, Helmut Rott, and Eugen Rott. “Envisat MERIS capabilities for monitoring the water quality of perialpine lakes”. In: *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*. Vol. 3. IEEE. 2003, pp. 2134–2136 (cit. on p. 20).

[95] Caroline Petus, Guillem Chust, Francis Gohin, et al. “Estimating turbidity and total suspended matter in the Adour River plume (South Bay of Biscay) using MODIS 250-m imagery”. In: *Continental Shelf Research* 30.5 (2010), pp. 379–392 (cit. on p. 20).

[96] Anatoly A Gitelson, Yosef Z Yacobi, Arnon Karnieli, and Nurit Kress. “Reflectance spectra of polluted marine waters in Haifa Bay, Southeastern Mediterranean: Features and application for remote estimation of chlorophyll concentration”. In: *Israel Journal of Earth Sciences* 45.3 (1996), pp. 127–136 (cit. on p. 20).

[97] Anatoly A Gitelson, Daniela Gurlin, Wesley J Moses, and Tadd Barrow. “A bio-optical algorithm for the remote estimation of the chlorophyll-a concentration in case 2 waters”. In: *Environmental Research Letters* 4.4 (2009), p. 045003 (cit. on p. 20).

[98] Sandra Mannheim, Karl Segl, Birgit Heim, and Hermann Kaufmann. “Monitoring of lake water quality using hyperspectral CHRIS-PROBA data”. In: *Proc. of the 2nd CHRIS/PROBA Workshop*. 2004, pp. 28–30 (cit. on p. 20).

[99] Kaire Toming, Tiit Kutser, Alo Laas, et al. “First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery”. In: *Remote Sensing* 8.8 (2016), p. 640 (cit. on p. 20).

[100] Patrick Brezonik, Kevin D Menken, and Marvin Bauer. “Landsat-based remote sensing of lake water quality characteristics, including chlorophyll and colored dissolved organic matter (CDOM)”. In: *Lake and Reservoir Management* 21.4 (2005), pp. 373–382 (cit. on pp. 20, 21).

[101] Cankut Ormeci, Elif Sertel, and O Sarikaya. “Determination of chlorophyll-a amount in Golden Horn, Istanbul, Turkey using IKONOS and in situ data”. In: *Environmental monitoring and assessment* 155.1 (2009), pp. 83–90 (cit. on p. 20).

[102] RN Fraser. “Hyperspectral remote sensing of turbidity and chlorophyll a among Nebraska Sand Hills lakes”. In: *International journal of remote sensing* 19.8 (1998), pp. 1579–1589 (cit. on p. 20).

[103] Weining Zhu, Qian Yu, Yong Q Tian, et al. “An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments”. In: *Remote Sensing of Environment* 140 (2014), pp. 766–778 (cit. on p. 20).

[104] Donald C Rundquist, Luoheng Han, John F Schalles, and Jeffrey S Peake. “Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm”. In: *Photogrammetric Engineering and Remote Sensing* 62.2 (1996), pp. 195–200 (cit. on p. 21).

[105] Susanne E Craig, Steven E Lohrenz, Zhongping Lee, et al. “Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, Karenia brevis”. In: *Applied Optics* 45.21 (2006), pp. 5414–5425 (cit. on pp. 21, 48).

[106] Mak Kisevic, Mira Morovic, and Roko Andircevic. “The use of hyperspectral data for evaluation of water quality parameters in the River Sava”. In: *Fresenius environmental bulletin* 25.11 (2016), pp. 4814–4822 (cit. on pp. 21, 48).

[107] Zeliang Zhang, Weining Zhu, Jiang Chen, and Qian Cheng. “Remotely observed variations of reservoir low concentration chromophoric dissolved organic matter and its response to upstream hydrological and meteorological conditions using Sentinel-2 imagery and Gradient Boosting Regression Tree”. In: *Water Supply* 21.2 (2021), pp. 668–682 (cit. on p. 21).

[108] Lucas Silveira Kupssinskü, Tainá Thomassim Guimarães, Eniuce Menezes de Souza, et al. “A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning”. In: *Sensors* 20.7 (2020), p. 2125 (cit. on pp. 21, 22).

[109] Yong Hoon Kim, Jungho Im, Ho Kyung Ha, Jong-Kuk Choi, and Sunghyun Ha. “Machine learning approaches to coastal water quality monitoring using GOCI satellite data”. In: *GIScience & Remote Sensing* 51.2 (2014), pp. 158–174 (cit. on pp. 21, 22).

[110] Yuanzhi Zhang, Jouni Pulliainen, Sampsa Koponen, and Martti Hallikainen. “Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data”. In: *Remote sensing of environment* 81.2-3 (2002), pp. 327–336 (cit. on p. 22).

[111] KP Sudheer, Indrajeet Chaubey, and Vijay Garg. “Lake water quality assessment from landsat thematic mapper data using neural network: an approach to optimal band combination selection1”. In: *JAWRA Journal of the American Water Resources Association* 42.6 (2006), pp. 1683–1695 (cit. on p. 22).

[112] Luis González Vilas, Evangelos Spyrakos, and Jesus M. Torres Palenzuela. “Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain)”. In: *Remote Sensing of Environment* 115.2 (2011), pp. 524–535 (cit. on p. 22).

[113] Yirgalem Chebud, Ghinwa M Naja, Rosanna G Rivero, and Assefa M Melesse. “Water quality monitoring using remote sensing and an artificial neural network”. In: *Water, Air, & Soil Pollution* 223.8 (2012), pp. 4875–4887 (cit. on p. 22).

[114] Fangling Pu, Chujiang Ding, Zeyi Chao, Yue Yu, and Xin Xu. “Water-quality classification of inland lakes using landsat8 images by convolutional neural networks”. In: *Remote Sensing* 11.14 (2019), p. 1674 (cit. on p. 22).

[115] Muhammad Aldila Syariz, Chao-Hung Lin, Manh Van Nguyen, Lalu Muhamad Jaelani, and Ariel C Blanco. “WaterNet: A convolutional neural network for chlorophyll-a concentration retrieval”. In: *Remote Sensing* 12.12 (2020), p. 1966 (cit. on p. 22).

[116] Sidrah Hafeez, Man Sing Wong, Hung Chak Ho, et al. “Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong”. In: *Remote sensing* 11.6 (2019), p. 617 (cit. on p. 22).

[117] Evangelos Spyrakos, Ruth O’Donnell, Peter D Hunter, et al. “Optical types of inland and coastal waters”. In: *Limnology and Oceanography* 63.2 (2018), pp. 846–870 (cit. on pp. 23, 110).

[118] Claire Neil, Evangelos Spyrakos, Peter D Hunter, and Andrew N Tyler. “A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types”. In: *Remote Sensing of Environment* 229 (2019), pp. 159–178 (cit. on pp. 23, 94, 105, 108, 110).

[119] Ave Ansper and Krista Alikas. “Retrieval of chlorophyll a from Sentinel-2 MSI data for the European Union water framework directive reporting purposes”. In: *Remote Sensing* 11.1 (2019), p. 64 (cit. on pp. 25, 94, 105, 108).

[120] Andreas Holbach, Stefan Norra, Lijing Wang, et al. “Three Gorges Reservoir: density pump amplification of pollutant transport into tributaries”. In: *Environmental science & technology* 48.14 (2014), pp. 7798–7806 (cit. on p. 27).

[121] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572 (cit. on pp. 35, 46, 47).

[122] Lukas W Lehnert, Hanna Meyer, Wolfgang A Obermeier, et al. “Hyperspectral data analysis in R: The hsdar package”. In: *arXiv preprint arXiv:1805.05090* (2018) (cit. on p. 43).

[123] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009 (cit. on pp. 45, 46, 48, 50–53).

[124] Tom M Mitchell. *Machine learning*. 1997 (cit. on p. 46).

[125] Felix M. Riese and Sina Keller. “Supervised, Semi-supervised, and Unsupervised Learning for Hyperspectral Regression”. In: *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing*. Ed. by Saurabh Prasad and Jocelyn Chanussot. Cham: Springer International Publishing, 2020, pp. 187–232 (cit. on pp. 46–54).

[126] Ke-Lin Du and Madisetti NS Swamy. *Neural networks and statistical learning*. Springer Science & Business Media, 2013 (cit. on pp. 46, 48, 54).

[127] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014 (cit. on p. 46).

[128] Stuart Geman, Elie Bienenstock, and René Doursat. “Neural networks and the bias/variance dilemma”. In: *Neural computation* 4.1 (1992), pp. 1–58 (cit. on p. 50).

[129] Andreas Merentitis, Christian Debes, and Roel Heremans. “Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.4 (2014), pp. 1089–1102 (cit. on p. 50).

[130] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984 (cit. on p. 50).

[131] Marvin N Wright and Andreas Ziegler. “ranger: A fast implementation of random forests for high dimensional data in C++ and R”. In: *arXiv preprint arXiv:1508.04409* (2015) (cit. on pp. 51, 56).

[132] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning* 63.1 (2006), pp. 3–42 (cit. on p. 51).

[133] Jerome H Friedman. “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378 (cit. on p. 52).

[134] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995 (cit. on p. 52).

[135] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14.3 (2004), pp. 199–222 (cit. on p. 52).

[136] Stephen Milborrow. "Notes on the earth package". In: *URL https://CRAN. R-project* (2014) (cit. on pp. 53, 56).

[137] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016 (cit. on pp. 54, 57).

[138] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 54, 57).

[139] Felix M. Riese and Sina Keller. "Soil Texture Classification with 1D Convolutional Neural Networks based on Hyperspectral Data". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W5 (2019), pp. 615–621 (cit. on pp. 54–57).

[140] Yann LeCun, Bernhard Boser, John S Denker, et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551 (cit. on p. 54).

[141] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. "Deep convolutional neural networks for hyperspectral image classification". In: *Journal of Sensors* 2015 (2015) (cit. on p. 54).

[142] Lanfa Liu, Min Ji, and Manfred Buchroithner. "Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery". In: *Sensors* 18.9 (2018), p. 3169 (cit. on p. 54).

[143] Max Kuhn et al. "Building predictive models in R using the caret package". In: *J Stat Softw* 28.5 (2008), pp. 1–26 (cit. on p. 56).

[144] Tianqi Chen, Tong He, Michael Benesty, et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015) (cit. on p. 56).

[145] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. "kernlab-an S4 package for kernel methods in R". In: *Journal of statistical software* 11.9 (2004), pp. 1–20 (cit. on p. 56).

[146] Brian Ripley, William Venables, and Maintainer Brian Ripley. "Package 'nnet'". In: *R package version* 7 (2016), pp. 3–12 (cit. on p. 56).

[147] Martın Abadi, Paul Barham, Jianmin Chen, et al. "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 2016, pp. 265–283 (cit. on p. 56).

[148] Taylor B Arnold. "kerasR: R interface to the keras deep learning library". In: *Journal of Open Source Software* 2.14 (2017), p. 296 (cit. on p. 56).

[149] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 56).

[150] N S Altman. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". In: *The American Statistician* 46.3 (1992), p. 175 (cit. on p. 61).

[151] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. 1995 (cit. on p. 61).

[152] Jerome H. Friedman. "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1 (1991), pp. 1–67 (cit. on p. 61).

[153] Kari Kallio, Tiit Kutser, Tuula Hannonen, et al. "Retrieval of water quality from airborne imaging spectrometry of various lake types in different seasons". In: *Science of the Total Environment* 268.1-3 (2001), pp. 59–77 (cit. on p. 94).

[154] Peter Gege. "The water color simulator WASI: An integrating software tool for analysis and simulation of optical in situ spectra". In: *Computers & Geosciences* 30.5 (2004), pp. 523–532 (cit. on p. 95).

[155] Alessandro Filazzola, Octavia Mahdiyan, Arnab Shuvo, et al. "A database of chlorophyll and water chemistry in freshwater lakes". In: *Scientific Data* 7.1 (2020), pp. 1–10 (cit. on p. 109).

[156] Jordan S Read, Xiaowei Jia, Jared Willard, et al. "Process-guided deep learning predictions of lake water temperature". In: *Water Resources Research* 55.11 (2019), pp. 9173–9190 (cit. on p. 110).

List of Abbreviations

1D	1-dimensional
ANN	Artificial Neural Network
AOP	Apparent Optical Properties
ap	Artificial Pond
BR	Band Ratio
CDOM	Colored Dissolved Organic Matter
CNN	Convolutional Neural Network
Conv	Convolutional
CV	Cross-Validation
CWA	US Clean Water Act
der	Derivatives
DL	Deep Learning
EnMAP	Environmental Mapping and Analysis Program
ESA	European Space Agency
ET	Extremely Randomized Trees
EU	European Union
FC	Fully-Connected
FTU	Formazin Turbidity Unit
GB	Gradient Boosting
IOP	Inherent Optical Properties
k-NN	k-Nearest Neighbors
MARS	Multivariate Adaptive Regression Splines
ML	Machine Learning
MSE	Mean Squared Error
NAP	Non-Algal Particles
NASA	National Aeronautics and Space Administration
PCA	Principal Component Analysis
qp	Quarry Pond
QS	Quinine Sulfate
R^2	Coefficient of Determination
ReLU	Rectified Linear Unit
RF	Random Forest
RG	Research Goal
RQ	Research Question

RMSE	Root Mean Squared Error
RoX	Reflectance Box
SD	Secchi Depth
SDG	Sustainable Development Goals
SPIM	Suspended Particular Inorganic Matter
SVM	Support Vector Machine
TSM	Total Suspended Matter
TSS	Total Suspended Solids
UN	United Nations
VNIR	Visible and Near-Infrared
WASI	Water Color Simulation
WFD	Water Framework Directive

List of Figures

List of Tables

List of Publications

A

In the following, the publications are listed, which have been published by or with the author of this thesis, Philipp Maier, within the time period 2018-2021. The ideas and developments presented in this thesis have partly been published in these publications and are clearly marked in this thesis.

Publications:

- Philipp M Maier, Sina Keller, and Stefan Hinz. "Deep Learning with WASI Simulation Data for Estimating Chlorophyll a Concentration of Inland Water Bodies". In: *Remote Sensing* 13.4 (2021), p. 718. **Peer-reviewed**, cited as [26] and marked in purple.
- Philipp M Maier and Sina Keller. "SpecWa: Spectral remote sensing data and chlorophyll a values of inland waters". In: *GFZ Data Services* (2020). Cited as [25] and marked in orange.
- Philipp M Maier and Sina Keller. "Application of different simulated spectral data and machine learning to estimate the chlorophyll a concentration of several inland waters". In: *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2019, pp. 1–5. **Peer-reviewed**, cited as [24] and marked in green.
- Philipp M Maier and Sina Keller. "Estimating chlorophyll a concentrations of several inland waters with hyperspectral data and machine learning models". In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4 (2019). **Peer-reviewed**, cited as [23] and marked in blue.
- Sina Keller, Philipp M Maier, Felix M Riese, Stefan Norra, Andreas Holbach, Nicolas Börsig, et al. "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity". In: *International journal of environmental research and public health* 15.9 (2018), p. 1881. **Peer-reviewed**, cited as [22] and marked in may green.
- Philipp M Maier and Sina Keller. "Machine learning regression on hyperspectral data to estimate multiple water parameters". In: *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2018, pp. 1–5. **Peer-reviewed**, cited as [21] and marked in red.
- Philipp M Maier, Stefan Hinz, and Sina Keller. "Estimation of Chlorophyll A, Diatoms and Green Algae Based on Hyperspectral Data with Machine Learning Approaches". In:

Tagungsband der 37. Wissenschaftlich-Technische Jahrestagung der DGPF e.V.. Vol. 27. Munich, Germany: Deutsche Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation, 2018, pp. 49–57. **Peer-reviewed**, cited as [20] and marked in pink.

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.