



Monitoring of Coastal Water Turbidity in the Presence of Cloud Cover Using Sentinel 2 Imagery with a Machine Learning Approach

A Thesis presented to
the Department of Geoinformation in Environmental Management
CIHEAM-Mediterranean Agronomic Institute of Chania (MAICh)

In Partial Fulfillment of the Requirements
for the Degree Master of Science

By

Lokmen Farhat

April 2022

Supervisor: Ioannis Manakos

Co-supervisor: Georgios Sylaios

M.Sc. Coordinator: Chariton Kalaitzidis

Acknowledgements

I love writing, but writing a thesis is an endeavour that should not be done alone. Mentors, faculty, family and friends are essential for completing this work. First and foremost, I would like to thank Dr. Ioannis Manakos for encouraging me to try new ideas and the hours he put into reading, editing, and discussing how best I can evolve the quality of my work. This has certainly helped me develop good scientific habits. I would also like to thank Prof. Georgios Sylaios who has always showed enthusiasm for my project and who has kept providing me with crucial advices in all our meetings. I will aim to always maintain the same excitement and curiosity towards the ideas of other researchers or colleagues. I am, as well, very grateful towards Dr. Chariton Kalaitzidis who has first introduced me to the world of remote sensing and for his thoughtful comments and support whenever I ran into problems. That certainly helped in improving this work. I would be remiss if I did not acknowledge CIHEAM MAICh and its Director Dr. George Baourakis for giving me access to the necessary space and computers to speed up my research and complete my work in a timely manner.

Many thanks go as well towards my friends and colleagues whose positivity, optimism, and good vibes were essential during my stay in MAICh. I really wish to thank Tahar Benamira, Fethi Morsli, Imen Brini, Imen Bouhamed, and Georgios Xaralampa for being a source of laughter, learning and support and I really think that meeting them has helped me grow as a person. I hope that life treats them well and that they succeed in whatever they decide to pursue. Special thanks go, as well, towards my friend Radhouane Arfaoui, from my home-country,

who has always cheered for me and supported me when things seemed too difficult. His friendship has always been a blessing for me and I really wish him and his family the best of things.

Moreover, I would like to thank my own family. My kind-hearted parents Mohammed and Nabiha, as well as, my amazing sisters Binen and Rabeb who have kept me motivated and grounded not only in the last two years, but for my whole life. Although, my family does not understand much of my research, they sacrificed a lot for me and immensely supported me in many ways. I really appreciate their unwavering love and support and I hope that I will be able to return their kindness in the near future.

Table of Contents

Acknowledgements	i
Table of Contents	iii
List of Figures	v
List of Tables.....	vi
List of Abbreviations	vii
Abstract	viii
Chapter 1: Introduction	1
1.1 Background and Rationale	1
1.2 Research Objectives	3
1.3 Thesis Outline	4
Chapter 2: Literature Review	5
2.1 Mediterranean Coastal Waters	5
2.2 Monitoring Coastal Water Quality	6
2.2.1 Traditional approaches.....	7
2.2.2 Optical remote sensing approaches.....	7
2.2.3 Water quality indicators measurable from satellite images	9
2.2.3.1 Turbidity.....	9
2.2.3.2 Secchi Disk Depth	10
2.2.3.3 Chlorophyll-a.....	10
2.2.3.4 Coloured dissolved organic material	11
2.3 Advances in Water Quality Mapping	11
2.3.1 Water quality retrieval algorithms	11
2.3.1.1 Empirical approaches.....	11
2.3.1.2 Analytical approaches	12
2.3.1.3 Hybrid approaches	12
2.3.2 Gap-Filling algorithms	13
2.3.2.1 Temporal approaches	13
2.3.2.2 Spatial approaches	14
2.3.2.3 Machine Learning approaches	14
Chapter 3: Materials and Methods.....	15
3.1 Study Area.....	15
3.2 Earth Observation Datasets.....	16
3.3 Applied Methodology	18
3.3.1 Pre-processing	20
3.3.1.1 Experimental setting.....	21
3.3.1.2 Performance assessment	26
3.3.2 Development of machine learning models.....	27
3.3.2.1 Experimental setting.....	28
3.3.2.2 Model selection and interpretation.....	30
Chapter 4: Results	31
4.1 Performance Evaluation of the Modified Gap Filling Algorithm	31

4.2	Performance Evaluation of the Adopted Machine Learning Algorithms	33
4.3	Spatial and Temporal Variation of Turbidity	35
Chapter 5:	Discussion	39
5.1	Gap Filling Algorithm.....	39
5.2	Machine Learning Algorithms	39
5.3	Spatial and Temporal Characteristics	40
Chapter 6:	Conclusions and Recommendations	41
Bibliography		43
Appendices		55

List of Figures

Figure 2.1 Mediterranean Sea.	5
Figure 2.2. Expected minimum and maximum changes in sea surface temperature for the 2070–2099 period (vs. 1961–1990) (Kim et al., 2019).	6
Figure 2.3. Sediment plume discharged by the Rhone River in southern France (Credit: adopted image, EU, Copernicus Sentinel-2 imagery).	8
Figure 2.4. Components of radiance (R) reaching a sensor: (a) R from the water, (b) R reflected from the bottom, (c) R reflected by the air-water interface, (d) R scattered by the atmosphere, (e) R reflected from land and scattered to the sensor by the atmosphere (Credit: logos downloaded from logomakr.com).....	9
Figure 3.1. Study area location.....	15
Figure 3.2. Distribution of S2 according to designated classes of cloud cover.	18
Figure 3.3. Workflow of applied methodology.	19
Figure 3.4. Adopted land mask (a) before and (b) after applying a buffer.	20
Figure 3.5. Workflow of the gap-filling algorithm (UP: Unknown Pixel, KP: Known Pixel).	22
Figure 3.6. Proposed new filling path method for unknown pixels.	23
Figure 3.7. Example of a targeted search operation.	24
Figure 3.8. Preparation of data to be fed into ML models (TP: turbidity pixel; KP: known pixel.	29
Figure 3.9. Example of a 5-folds cross validation approach.....	30
Figure 4.1. Boxplot representing the R2 metric obtained on the cross-validation (colored dots correspond to outliers); the metric on the test set is marked with a red star.....	35
Figure 4.2. Turbidity maps corresponding to two months (a: original Copernicus image; b: simulated image without GF; c: simulated image with GF; Red circles indicate gaps due to cloud cover.....	36
Figure 4.3. Temporal graphs of average, interquartile range (IQR), and standard deviation (SD) turbidity values in the study area derived from actual and predicted observations, with and without GF.	38

List of Tables

Table 1: Acquisition dates of turbidity maps (spatial resolution of 100 m)	17
Table 2: General characteristics of S2 bands	17
Table 3: Parameter values that were investigated for achieving the best gap filling results.....	25
Table 4: Acquisition dates and cloud percentage of S2 images that were used in testing the modified DS algorithm	26
Table 5: Number of S2 images used in ML.....	28
Table 6: The average number of extracted pixel values per case of selected best Sentinel2 images.....	29
Table 7: Error metrics values for gap filled images (Blue Band).....	31
Table 8: Statistical indices of gap filled images (Blue Band)	32
Table 9: Comparison of processing time between UL and BL versions (Blue Band)	32
Table 10: The processing time of a UL algorithm based on the selected number of sub-groups (Blue Band)	33
Table 11: Mean error metric values derived from the cross-validation step	34
Table 12: Change in the number of unknown pixels and error percentages corresponding to simulated turbidity maps.....	37

List of Abbreviations

CC	Climate Change;
IPCC	Intergovernmental Panel for Climate Change;
EU	European Union;
ML	Machine Learning;
EO	Earth Observation;
UNESCO	United Nations Educational, Scientific and Cultural Organization;
GMRT	Global Multi-Resolution Topography;
S2	Sentinel-2;
NTU	Nephelometric Turbidity Units;
L2A	Level-2A;
UTM	Universal Transverse Mercator;
WGS84	World Geodetic System 1984;
SCL	Scene Classification Layer;
GF	Gap-Filling;
DS	Direct Sampling;
UP	Unknown Pixel;
KP	Known Pixel;
Nx	Neighboring pixels of an UP value at location 'x' in target image;
Ny	Neighborhood pixels of a replicate at location 'y' in training image;
TI	Training Image;
UL	uni-layer;
BL	bi-layer;
SD	Standard Deviation;
MdAPE	Median Absolute Percentage Error
RMSE	Root Mean Square Error;
RMSLE	Root Mean Square Logarithmic Error;
MAE	Mean Absolute Error;
R ²	Coefficient of Determination;
MSE	Mean Squared Error;
RF	Random Forests;
ANN	Artificial Neural Networks;
KNN	K-Nearest Neighbors regressor;
CV	Cross Validation;
R	Radiance;
Chl-a	Chlorophyll-a;
CDOM	Coloured Dissolved Organic Material;
SDD	Secchi Disk Depth;
SPM	Suspended Particulate Matter.

Abstract

The scarcity of field data measurements in water quality has led researchers to focus on developing empirical algorithms, based on remote sensing data, for the monitoring of coastal water quality indicators. Empirical algorithms rely, primarily, on the use of Earth Observation data to predict the selected water quality indicators. Coastal water turbidity is one such indicator that is considered essential for the study of water pollution in the Mediterranean Sea. This research work aims to: (i) modify an existing gap-filling method to produce satellite images without gaps, and (ii) build three machine learning models that can reliably predict coastal water turbidity, by using Sentinel-2 imagery. The validation of the modified gap-filling approach was performed by imposing artificial clouds on a selected image (containing several bands) and comparing the simulated images with the original one, based on five error metrics. The validation of the machine learning models was implemented by sampling more than 4700 reference values per image from existing Copernicus turbidity maps and assessing the performance based on two error metrics. These existing maps were derived using a semi-analytical approach with a 100 m resolution and for a limited period of 15 months. The gap-filling results reveal that acceptable performance can only be achieved when considering parametrization for three S2 bands (blue, green, and red). The average R^2 accuracy value for both versions of the gap-filling algorithm is around 83%. The machine learning results indicate that the random forests model tends to outperform both the artificial neural network and the K-nearest neighbors' models. When using only two S2 images to feed the best learner model, the average R^2 value derived with or without gap-filling is about 90%.

Keywords: coastal water turbidity, Earth Observation, Sentinel-2, gap filling, empirical algorithms, machine learning models, Mediterranean Sea.

Chapter 1: Introduction

1.1 Background and Rationale

Coastal waterbodies contribute significantly to the daily wellbeing of people all over the world (Neumann et al., 2017). Indeed, human communities tend to thrive more when they are located near coastal areas as the latter can provide a boost to the prosperity of populations. Coastal ecosystem services vary from addressing the basic human need for food to supporting the economy of a country through fishing campaigns, and/or promoting recreational and tourism-related activities (Lau et al., 2019; Mehvar et al., 2018; Rao et al., 2015). However, coastal waterbodies, particularly in the Mediterranean Basin, continue to be affected by several types of pressures such as coastal runoff, the overuse of plastics, and the illegal dumping of waste and wastewater near the shore (Micheli et al., 2013; Prévost and Robert, 2016; Robert et al., 2019). Professionals and researchers expect the situation to be further exacerbated by global challenges related to the rising urbanization and climate change (CC) (Calvo et al., 2011; Lejeusne et al., 2010). If land management practices do not change, it is anticipated that an overall decrease of coastal water quality may occur; resulting in the deterioration of the habitats of marine species and the widening of socio-economic inequalities in the future.

Coastal areas are highly vulnerable to many pollution types. The latter can be a slow or rapid process and may be attributed to both anthropogenic and natural causes. For example, runoff water that end up in the sea may be rich in dissolved and suspended pollutants, namely nutrients, toxic metals, organic matter, pesticides and fertilizers that have been transported, over time, from nearby watersheds (Oelsner and Stets, 2019). In addition, it is well known that the impacts of water pollution are not limited to the simple loss in water clarity, but may, also, be associated with the decline of aquatic life and several human health issues (El-Said, 2013). In fact, the increase of suspended sediments and nutrient concentrations in coastal areas coupled with warming weather conditions will, most likely, result in a rapid expansion of algae and the eutrophication of the

ecosystem as a whole (de Paul Obade et al., 2014; Paerl, 2017). When left undetected for a long period of time, the spreading pollution may endanger the community at large or cause an irreversible loss in suitable habitats for some marine species.

Spurred by this threat, the European Union (EU) has established several guidelines regarding the required indicators for monitoring water quality (i.e. Marine Strategy Framework Directive and the Water Framework Directive) (Law et al., 2010; Regional Environmental Center, 2008). Turbidity is one such indicator that is widely adopted by local managers working in coastal ecosystems (Gholizadeh et al., 2016). Turbidity is a reflection of the total amount of different components that exist in water at the time of measurement. The recent advances in satellite technology have made it possible to complement traditional measurement approaches (Wang and Yang, 2019). In addition, the consistency and long data continuity of some satellite missions such as Landsat and Sentinel series have helped in better tracking the variability of different water quality parameters over space and time. For example, Caballero et al. (Caballero et al., 2018), as well as, Dorji and Fearn (Dorji and Fearn, 2017) have successfully evaluated the use of different satellite imagery (e.g. Sentinel, Landsat, MODIS, etc.) for estimating turbidity in coastal waters, respectively, in Spain and Australia.

The use of satellite imagery is based on the principle that a change in the physical and chemical properties of water will result in a proportional change in its inherent optical properties (Bresciani et al., 2018; Lins et al., 2017). In this context, many models were developed for the purpose of facilitating more proactive strategies when dealing with coastal water quality issues. These models can be purely empirical/analytical, or semi-empirical/analytical (known as hybrids) (Chang et al., 2015). As a general rule, purely empirical models are based on describing the relationships between field and remote sensing data from the perspective of a statistician (Ramadas and Samantaray, 2018), while purely analytical models refer to the characterization of light interaction with the water surface from the perspective of a physicist (Dekker et al., 1996). Whereas, hybrid models entail combining the best of both worlds, namely, light theory with empirical formulas, which makes them more generalizable (Ma et al., 2006).

The recent advances in technology (i.e. the introduction of cloud platforms and the substantial increase in the computational power of computers), have boosted the popularity of different machine learning (ML) techniques (considered to be part of the category of empirical methods) among professionals and researchers (Lary et al., 2016; Olden et al., 2016). For instance, Hafeez et al. (Hafeez et al., 2019) have developed several ML models (e.g. artificial neural network, random forests, etc.) to estimate water quality using a handheld multispectral radiometer and Landsat imagery as input layers. In-situ physical data, corresponding to three parameters (suspended solids, chlorophyll-a, and turbidity), were collected from 76 local stations. Results indicate that the neural network model outperforms the other models, with overall accuracy results being 89%-Chl-a, 93%-SS, and 82%-turbidity (handheld radiometer), and 91%-Chl-a, 92%-SS, and 85%-turbidity (satellite data). Accuracy was measured using the Root Mean Square Error (RMSE) formula.

1.2 Research Objectives

In the context of the above, the present thesis proposes a framework where Earth Observation (EO) data can be used, in combination with ML techniques, to retrieve information on turbidity levels over space and time in a Mediterranean coastal area.

The primary objectives of this thesis will be about:

- 1) Modifying a gap-filling algorithm to improve its processing speed and evaluating its performance when applied on Sentinel 2 bands;
- 2) Juxtaposing the ML derived maps against the Copernicus ones to evaluate their validity in being used as surrogates;
- 3) Testing the sensitivity of various spectral bands and band ratio combinations to achieve the best performance in adopted ML models;
- 4) Examining whether the use of ML coupled with the latest EO products may expand the ability to more regularly produce turbidity maps across wider coastal areas.

1.3 Thesis Outline

The thesis is organized in the following manner:

- Chapter 1 is an introduction presenting the scope of the study, the major objectives, and the structure of the thesis;
- Chapter 2 provides a review of the literature for water quality monitoring using EO data and gap-filling methods;
- Chapter 3 describes in detail the input data and methodology used for data preparation as well as data analysis;
- Chapter 4 is dedicated for the presentation of results;
- Chapter 5 provides a discussion of the results;
- Chapter 6 is a conclusion summarizing the main findings and presenting potential improvements upon the developed methodology as well as prospective future research directions.

Chapter 2: Literature Review

2.1 Mediterranean Coastal Waters

The Mediterranean region depicts in a way a crossroads of civilizations and cultures and has, therefore, been characterized by constant land use changes (Figure 2.1). The coastal areas represent 12% of the total surface areas corresponding to all Mediterranean countries and they are, in general, where most human development and population growth tend to occur (BluePlan, 2005). The Mediterranean Sea covers about 26 million Km² (0.8% of total surface area of oceans) and is characterized by richness in biodiversity and natural habitats. For instance, it is estimated that between 4 and 18% of all identified marine species live in the Mediterranean Sea (Coll et al., 2010). About a fifth of these species are actually considered endemic (BluePlan, 2005). Moreover, it is to be noted that most of these species tend to settle near the shore, which is why they have always been very vulnerable to human activities.

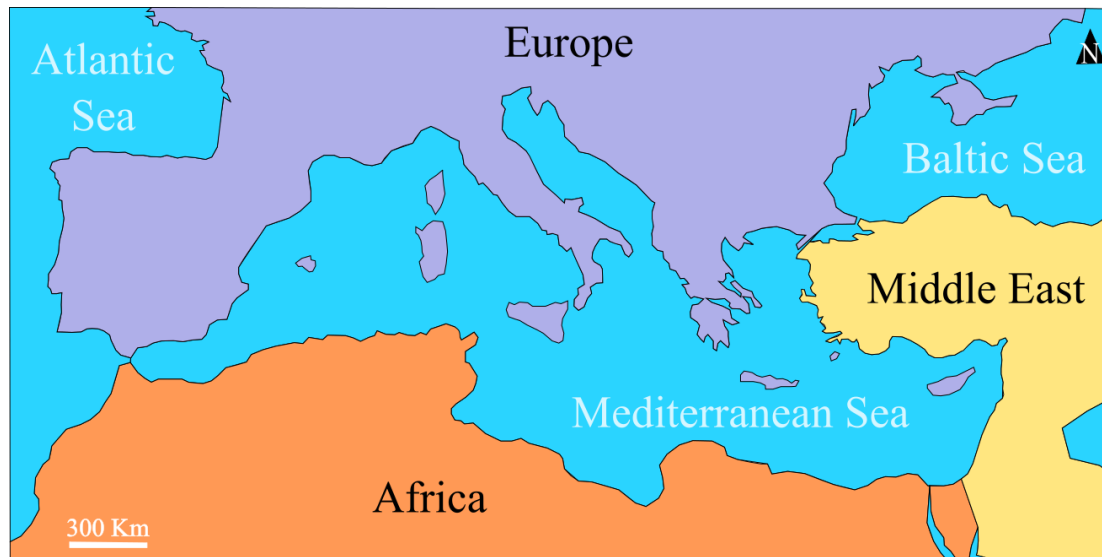


Figure 2.1 Mediterranean Sea.

The Mediterranean region has always been shaped by trade, industrial, and touristic human needs. All of which have made the region a “hotspot” in terms of areas most affected by Climate Change (CC) (Figure 2.2). The reports published by the Intergovernmental Panel for Climate Change (IPCC) show that the synergistic impacts resulting from both climate change, and human development may cause

an irreversible loss in biodiversity and the degradation of natural habitats (Boero, 2015; Frölicher and Laufkötter, 2018; Giorgi, 2006; Mancinelli and Vizzini, 2015; Pachauri et al., 2014). For example, it has been reported that biological traits related to population growth, survival, and migration of fish are closely tied to the warming of the Mediterranean Sea (Marbà et al., 2015). The latter has been estimated to be at 0.4°C per decade for the period between 1985 and 2006. The projections for 2100 vary between +1.8°C and +3.5°C in average compared to the period between 1961 and 1990 (Adloff et al., 2015).

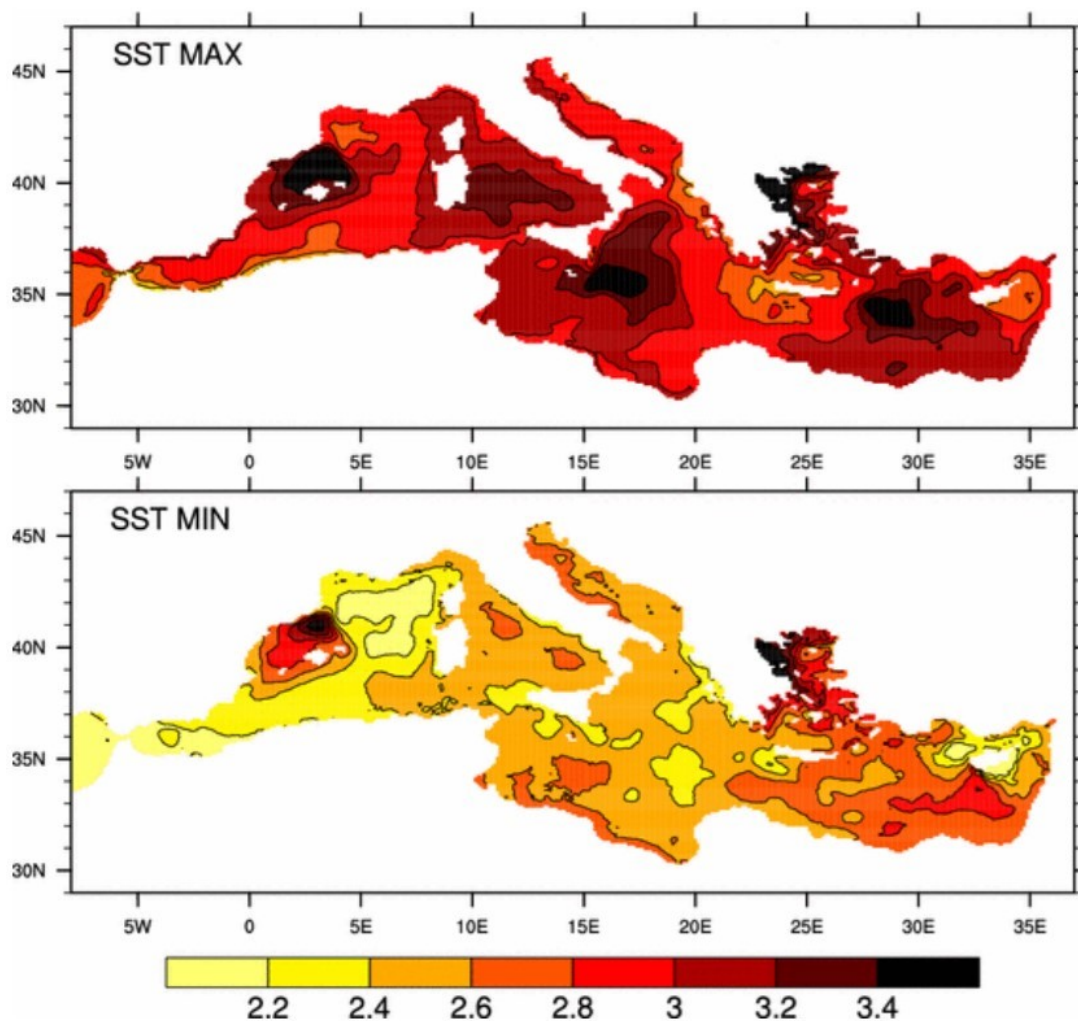


Figure 2.2. Expected minimum and maximum changes in sea surface temperature for the 2070–2099 period (vs. 1961–1990) (Kim et al., 2019).

2.2 Monitoring Coastal Water Quality

It is expected that the degradation of Mediterranean marine ecosystems will not only have environmental consequences (e.g., loss in biodiversity) but it will

also result in further widening the gap between the South and the North and among the communities within the same country as well. This is attributed to the fact that the livelihoods of millions of people depend on fishing, recreational, and cultural activities associated with coastal areas. Indeed, such issues will most likely lead to political, economic and social unrest in the long term (BluePlan, 2005). Hence, the importance of establishing solid policy responses to the complex problem of coastal water pollution that are based on the clear and effective communication of scientific evidence. To achieve this aim, researchers would first need to adopt systematic approaches for recording and monitoring of various water quality indicators. In this regard, the combination of both traditional and modern technological methods is of paramount importance for better understanding the key drivers of pollution in coastal waters.

2.2.1 Traditional approaches

A traditional approach for water quality assessment is based on either performing on-site measurements (e.g., Temperature and Secchi Depth) or collecting water samples for analysis in a laboratory (e.g., Turbidity and Chlorophyll-a). The inherent limitations of these approaches include: (i) the discrete nature of collected data, (ii) the required intensive field/laboratory preparation, and (iii) the high financial costs. All of these reasons explain how most field-collected data tend to suffer from discontinuity or an uneven distribution of observations over space and time (de Paul Obade and Moore, 2018). In many cases, environmental experts and researchers find themselves in need of applying interpolation methods to make extrapolations of their collected datasets to wider areas.

2.2.2 Optical remote sensing approaches

The use of optical remote sensing to assess water quality is founded on the principle that the amount of reflected light from water surface can vary based on the types of components existing in the area of interest at the time. For instance, an increase in the flux of suspended sediments floating in water due to intense precipitation will make coastal waters appear more brown in color and the reflected light detected by the satellite sensor will mirror that change (Figure 2.3). In a way, any change in the structural, physical, and chemical properties of coastal

waters will result in a change of their spectral signature as well (Lillesand et al., 2015). There are four main advantages for using optical remote sensing to monitor water quality: (i) the gain in spatial resolution and coverage, (ii) the ability to capture a continuous series of images over a long time period (e.g., there exist about 49 years of Landsat images and seven years of Sentinel 2 images), (iii) the ability to study the variability of multiple water quality indicators using the same type of images, and (iv) the existence of many open-source satellite imagery repositories. In most cases, environmental experts and researchers opt for the use of optical remote sensing in combination with in-situ observations for a better tracking of water quality indicators, establishing baselines, and determining trends.

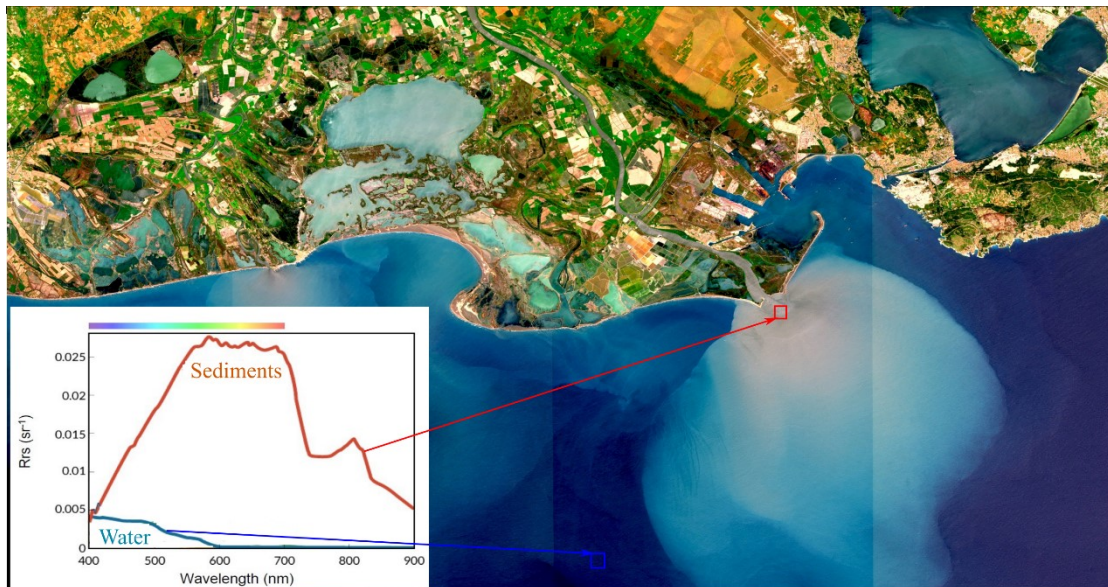


Figure 2.3. Sediment plume discharged by the Rhone River in southern France (Credit: adopted image, EU, Copernicus Sentinel-2 imagery).

Optical remote sensing corresponds to measuring the sunlight reflected by coastal waters in the direction of the passive sensor. The incident solar radiation represents a mixture of five components illustrated in Figure 2.4. The overall reflectance value recorded by the sensor, and the contribution of each of the five components will differ, largely, based on the following factors: (i) the considered wavelength, (ii) the adjacency of water pixels to coastline pixels, and (iii) the depth of sea water (Asrar, 1989; Hellweger et al., 2004). The target properties can then be approximately predicted using several types of algorithms (e.g., machine learning algorithms). In the context of water quality monitoring, the most used

wavelength varies typically between 400 and 1020 nm (visible to infra-red) (Jenni, 2019). Optical remote sensing can either be performed through airborne or satellite sensors. The acquired images are, mainly, characterized by their spatial, temporal, and spectral resolutions.

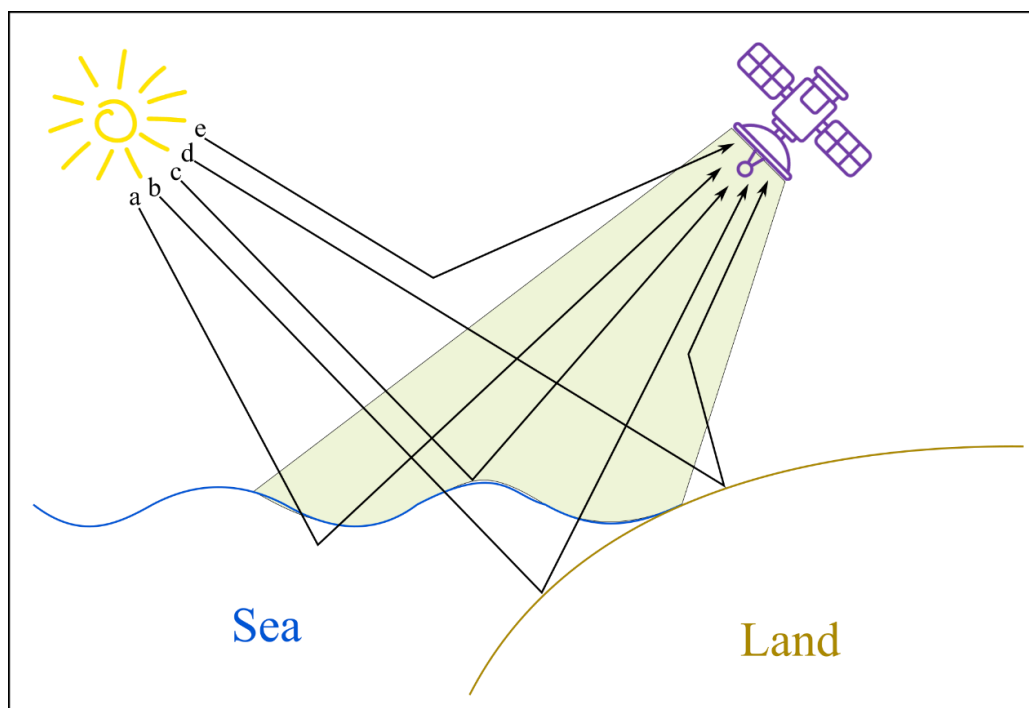


Figure 2.4. Components of radiance (R) reaching a sensor: (a) R from the water, (b) R reflected from the bottom, (c) R reflected by the air-water interface, (d) R scattered by the atmosphere, (e) R reflected from land and scattered to the sensor by the atmosphere (Credit: logos downloaded from logomakr.com)

2.2.3 Water quality indicators measurable from satellite images

The deterioration of the quality of coastal waters may be attributed to several causes, such as: illegal waste discharges, transported sediments, nutrients, and pesticides (following runoff events). The study of the extent and intensity of water pollution is, primarily, based on measuring several water quality indicators, such as: turbidity, secchi disk depth, chlorophyll-a, and colored dissolved organic matter (Chang et al., 2015; de Paul Obade and Moore, 2018; Gholizadeh et al., 2016; Wang and Yang, 2019).

2.2.3.1 Turbidity

Turbidity represents an indicator for describing the difficulty at which the incoming radiation travels through water. It is regarded as the opposite of water

clarity and is based on measuring optical scattering in water. The latter is closely related to the concentration of suspended particulate matter (SPM) present at the time (Myint and Walker, 2002). SPM refers to suspended particles on which substances like organic matter, metals, algae, etc. are adsorbed. Typically, turbidity is determined with adequate accuracy using bands that range between 400 and 800 nm (Ritchie et al., 1976). Several studies indicate that the use of either a single band or a band ratio can help in monitoring turbidity in water (Doxaran et al., 2002; Nechad et al., 2010; Usali and Ismail, 2010).

2.2.3.2 Secchi Disk Depth

Secchi disk depth (SDD) represents an indicator for describing water clarity. SDD is inversely correlated with turbidity and varies in relation to the concentration of SPM present in water. The measurement of SDD is based on lowering a white disk with a diameter ranging, usually, between 12 to 20 cm into the water and determining the depth at which the disk can no longer be visually distinguished from the surrounding water (Preisendorfer, 1986). The determined depth indicates the maximum depth at which the water layer receives sunlight and photosynthesis is possible. Although, this indicator is easy to measure, it should be noted that the latter is heavily dependent on illumination conditions and the eyesight of the person in question. Remote sensing products have been used by several researchers for keeping track of SDD (Allan et al., 2011; Olmanson et al., 2008).

2.2.3.3 Chlorophyll-a

Chlorophyll-a (Chl-a) represents an indicator for describing algae growth and the existence of phytoplankton biomass. Chl-a can help us to indirectly quantify the nutrients present in water and assess its trophic state (Huang et al., 2010). Ensuring the continuous measurement of Chl-a helps in detecting algae blooms at an early stage, and therefore in avoiding the eutrophication of the area of interest. The eutrophication phenomenon is attributed to the sharp decrease in dissolved oxygen that is usually associated with a sudden rise in phytoplankton species. According to Cannizzarro and Carder (Cannizzarro and Carder, 2006) the variation in Chl-a is optimally detected using a wavelength that ranges between 440 and 550 nm.

2.2.3.4 Coloured dissolved organic material

Coloured dissolved organic material (CDOM) represents an indicator for describing the part of the humic substances that subsisted in its dissolved state in water. The humic substances refer to those resulting from the decomposing of plant tissue, other than transforming into carbon dioxide, inorganic nitrogen, sulphur and phosphorus (Kirk, 1994). CDOM can color water in brown and yellowish brown for high concentrations. It is to be noted that CDOM co-varies with Chl-a and is usually detected using a wavelength at 440 nm (Nelson et al., 2007).

2.3 Advances in Water Quality Mapping

In recent years, the number of available satellite images has been increasing rapidly. This is particularly the case for the open-access imagery repositories, such as those acquired from Sentinel, Landsat, Terra and Aqua satellites. These repositories have made a difference in the fight against environmental pollution. For instance, they have helped in studying the seasonal characteristics of water quality, and by extension in identifying anomalies at an early stage (Caballero et al., 2018; Kupssinskü et al., 2020). They have, also, provided decision-makers with critical information regarding the impact of urban expansion and the human footprint on coastal ecosystems (Chen et al., 2005; Li et al., 2014). With this abundance in remote sensing datasets, the aim of researchers has shifted towards developing better tools for data analysis in terms of improving the prediction accuracy, reducing the computational time, and filling gaps in existing datasets.

2.3.1 Water quality retrieval algorithms

One of the primary points of focus for researchers is the refinement of existing water quality features retrieval algorithms and the development of new ones. In general, this type of algorithms can be divided into three categories that are, namely: (i) empirical approaches, (ii) analytical approaches, and (iii) hybrid approaches (i.e., semi-empirical/analytical) (Chang et al., 2015).

2.3.1.1 Empirical approaches

Empirical approaches correspond to algorithms that use linear and multi-linear statistical equations, as well as, non-linear machine learning techniques to

predict the variable(s) of interest (Chang et al., 2009; Dekker et al., 1996; Ting, 2005). These approaches are based on establishing a statistical relationship between one or many bands/band ratios and the selected water quality variable. In general, empirical models have two main advantages over analytical models: (i) they are easier to implement, and (ii) they tend to be more accurate when applied to the same area from which in-situ data were collected (Liu et al., 2003). However, the major drawback of empirical models lies in their low performance when trying to apply them to other locations. Therefore, these models have more site-specific capacities.

2.3.1.2 Analytical approaches

Analytical approaches correspond to algorithms that use the inherent and apparent optical characteristics of water to predict the variable(s) of interest (Chang et al., 2009; Gholizadeh et al., 2016). These optical properties include the absorption coefficient, the scattering coefficient, the diffuse attenuation coefficient, etc. (Dekker et al., 1996; Ma et al., 2006). These approaches are based on the principle that the optical properties of water vary in correlation with water quality. For instance, an increase in sediments in water will be associated with a higher scattering coefficient. Therefore, understanding how these optical properties vary in space and time would help in describing the changes in water quality in an area of interest. In general, analytical models have two main advantages over empirical models: (i) they are more generalizable and can be applied across large areas, and (ii) they are independent of in-situ water quality data. However, the major drawback of analytical models is in relation to the availability of a complete and accurate dataset of optical properties. In addition, it should be highlighted that the accuracy of such a dataset is highly dependent on using an efficient atmospheric correction model (Fan et al., 2017).

2.3.1.3 Hybrid approaches

Hybrid approaches correspond to algorithms that combine optical theory with statistical techniques to predict the variable(s) of interest (Ma and Dai, 2005). They are based on expressing radiative transfer theory in an empirical way. Doing so helps in improving the accuracy or/and coverage of predictions (Chang et al., 2015). In general, hybrid models are better at distinguishing the different

constituents existing in water. However, there are two major drawbacks of hybrid models. They are highly reliant on applying an accurate atmospheric correction model, and they require a more sophisticated training and preparation (more data-driven) compared to purely empirical models (Matthews et al., 2010).

2.3.2 Gap-Filling algorithms

Gaps resulting from cloud cover represent one of the major limitations of optical remote sensing. Based on the time of image acquisition and the area of investigation, they can be large and contribute to omitting critical information. Typically, researchers tend to use interpolation methods for filling gaps in satellite imagery (El Fellah et al., 2017; Holloway et al., 2019; Wang et al., 2012). In general, these interpolation methods are divided into three categories: (i) temporal methods that are based on using a time series of images, (ii) spatial methods that are based on using spatial information, and (iii) machine learning methods.

2.3.2.1 Temporal approaches

Temporal approaches are based on looking for the time series corresponding to a selected pixel in a satellite image, and using this information to estimate the unknown pixel value (Pede and Mountrakis, 2018). An example of a temporal approach is the “Linear Interpolation Method”, where the algorithm fills the gaps by considering a pre-defined number of nearest neighbours (in time), assigning them weights based on time distance, and then computing the predicted value (Metz et al., 2014). Other methods that are categorized as temporal approaches for filling gaps are “Harmonic Analysis” and “Temporal Fourier Analysis” methods. Both of which are based on the following three steps: (i) selecting a time-series corresponding to an unknown pixel, (ii) investigating the seasonality characteristics of this time-series, and (iii) using the latter to predict the unknown pixel value (Scharlemann et al., 2008; Xu and Shen, 2013). A major drawback of temporal approaches is in regards to its low performance when applied in locations where cloud percentage tends to be high for an extended period of time (i.e., winter season). Consequently, if for a selected pixel, the extracted time-series consists primarily of missing values, the overall accuracy of the predicted pixel values will drop drastically (Metz et al., 2014; Scharlemann et al., 2008).

2.3.2.2 Spatial approaches

Spatial approaches refer to the algorithms that look for the nearest known neighbours (in space) to a selected unknown pixel and using them to predict the value of the latter. They are based on the assumption that neighbouring pixels are highly correlated (Yu et al., 2015). Hence, it is not recommended to use these approaches where this assumption is false. For example, using spatial methods to fill gaps in satellite images along a narrow coastline where ship traffic is high may yield poor results. This is because the variation among neighboring pixels of the unknown pixel would be very high (Ke et al., 2012). Moreover, it is recommended to use auxiliary data to improve the performance of these methods. For example, using kriging interpolation on land surface temperature data with elevation as an auxiliary layer would improve the accuracy of results by a great margin (Metz et al., 2014; Neteler, 2010).

2.3.2.3 Machine Learning approaches

Machine learning approaches are based on feeding a selected model (e.g., convolutional neural networks) with enough temporal, spatial, and reflectance information to fill the gaps in an image (Zhang et al., 2018). Machine learning methods allow the user to change the hyper-parameters as necessary. In some cases, it is possible to apply the linear interpolation method using two different types of images. For example, Cresson et al. (Cresson et al., 2019) used Sentinel-1 SAR images alongside Sentinel-2 images to fill gaps in the latter. However, a major drawback of these methods is in regards to their requirement in large amounts of data to produce reliable estimations. In addition, based on which machine learning model is used, the processing time may be long.

Chapter 3: Materials and Methods

3.1 Study Area

The study area is located in the southern coast of France, where the Rhone River flows out to the Sea (Figure 3.1). According to Antonelli et al. (Antonelli et al., 2008), the Rhone river is considered to be a major contributor of sediments to the Mediterranean Sea. The Rhone river stems from Geneva Lake, located in the Alps, and flows through many highly populated areas (~ 180 inhabitants/km²) until reaching the sea, near the city of Arles, with an overall length of 512 Km (Delile et al., 2020). In addition, it is characterized by a catchment area of 98,000 km² (Comby et al., 2014), and with an average amount of precipitations being 843 mm (Ludwig and Meybeck, 2003). The outlet of the Rhone River appears to be in the form of a delta that contains a protected area known as “Parc Naturel Régional de Camargue”. This wetland site has been declared as a “Biosphere Reserve” by the United Nations Educational, Scientific and Cultural Organization (UNESCO) (<https://en.unesco.org/biosphere/eu-na/camargue>, accessed online on the 26th January 2022).

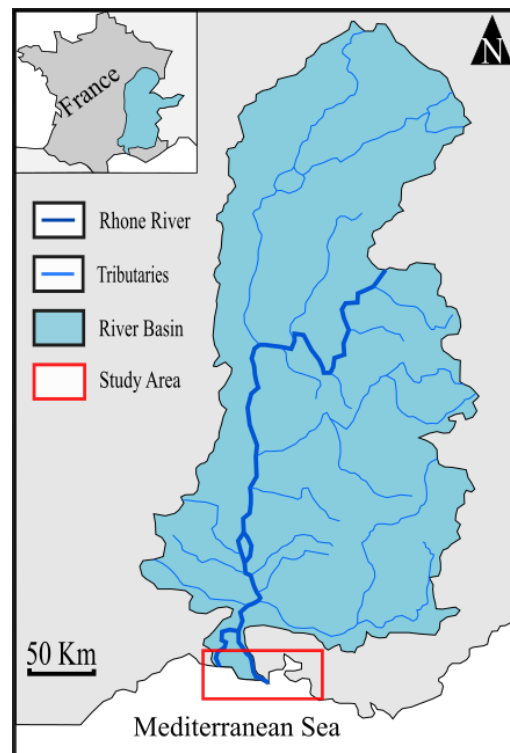


Figure 3.1. Study area location.

3.2 Earth Observation Datasets

In the present work, three types of Earth Observation (EO) data have been acquired, namely: (i) Turbidity maps that are available through the Copernicus global land service program (land.copernicus.eu/global/products/lwq, accessed online on the 26th January 2022) (ii), Bathymetry maps from the Global Multi-Resolution Topography (GMRT) product (Ryan et al., 2009), and (iii) Sentinel-2 (S2) imagery that are available through the Copernicus Open Access Hub. The acquired bathymetry data represent a synthesized product from different sources (e.g., sonar survey missions, international bathymetric charts, elevation data, etc.) [7]. In case of the present study area, the retrieved bathymetry map is characterized by a spatial resolution of 45 m.

The Copernicus global land service program provides ready to use turbidity maps that were derived from S2 using the semi-analytical approach described by Nechad et al. [8]. These turbidity maps have an approximate spatial resolution of 100 m and cover, only, the period between January 2019 and March 2020. In addition, it is to be noted that the publicly available maps represent a 10-days aggregated product that is expressed in Nephelometric Turbidity Units (NTU). In this study, one turbidity map per month was downloaded for a total of 15. The selected turbidity maps represent the period between the 11th and 20th day of every month. However, when a map contained very few valid pixels (or none at all), the previous or next map was used instead (based on which one has the highest number of known pixels). This was the case for three dates: January 2019, December 2019, and January 2020 (**Erreur ! Source du renvoi introuvable.**).

S2 imageries were acquired from the twin polar-orbiting satellites known as S2 A/B (both of which are equipped with a multispectral imager). S2A was launched on 23 June 2015 and S2B followed on 7 March 2017 (Agency, 2015). The general characteristics of available spectral bands are described in Table 2. The downloaded S2 scenes correspond to Level-2A (L2A) products. L2A products consist of 100 km² tiles that are orthorectified and spatially registered images in Universal Transverse Mercator (UTM) / World Geodetic System 1984 (WGS84) projection. In case of the considered study area, the number of available scenes during the 10-days period is four. Moreover, for the purpose of performing the

gap-filling process, S2 images that are available within a three days' difference from the 10-days period were downloaded.

Table 1: Acquisition dates of turbidity maps (spatial resolution of 100 m)

Image N°	Date	N° Valid Pixels
1	2019/01/21	91373
2	2019/02/11	47096
3	2019/03/11	102902
4	2019/04/11	102530
5	2019/05/11	102437
6	2019/06/11	102235
7	2019/07/11	103057
8	2019/08/11	102814
9	2019/09/11	103058
10	2019/10/11	7015
11	2019/11/11	67187
12	2019/12/01	102968
13	2020/01/01	103019
14	2020/02/11	100043
15	2020/03/11	103058

Table 2: General characteristics of S2 bands

Spectral Band	Central Wavelength (nm)	Band Width (nm)	Spatial Resolution (m)
B01	443	20	60
B02	490	65	10
B03	560	35	10
B04	665	30	10
B05	705	15	20
B06	740	15	20
B07	783	20	20
B08	842	115	10
B8a	865	20	20
B09	945	20	60
B10	1380	30	60
B11	1610	90	20
B12	2190	180	20

The average cloud cover percentage of S2 images (corresponding to the 10-days period) is 36.41%. Cloud percentage values were calculated after applying masks on S2 images. Figure 3.2 indicates that the cloud cover percentage in most S2 images is below 25%. However, it appears that for every 10 day-period of every month, there is at least one S2 image that is severely affected by clouds (cloud percentage above 75%). S2 images acquired during March 2019 seem to contain the least number of missing data, whilst S2 images acquired during November 2019 seem to contain the highest number of missing data.

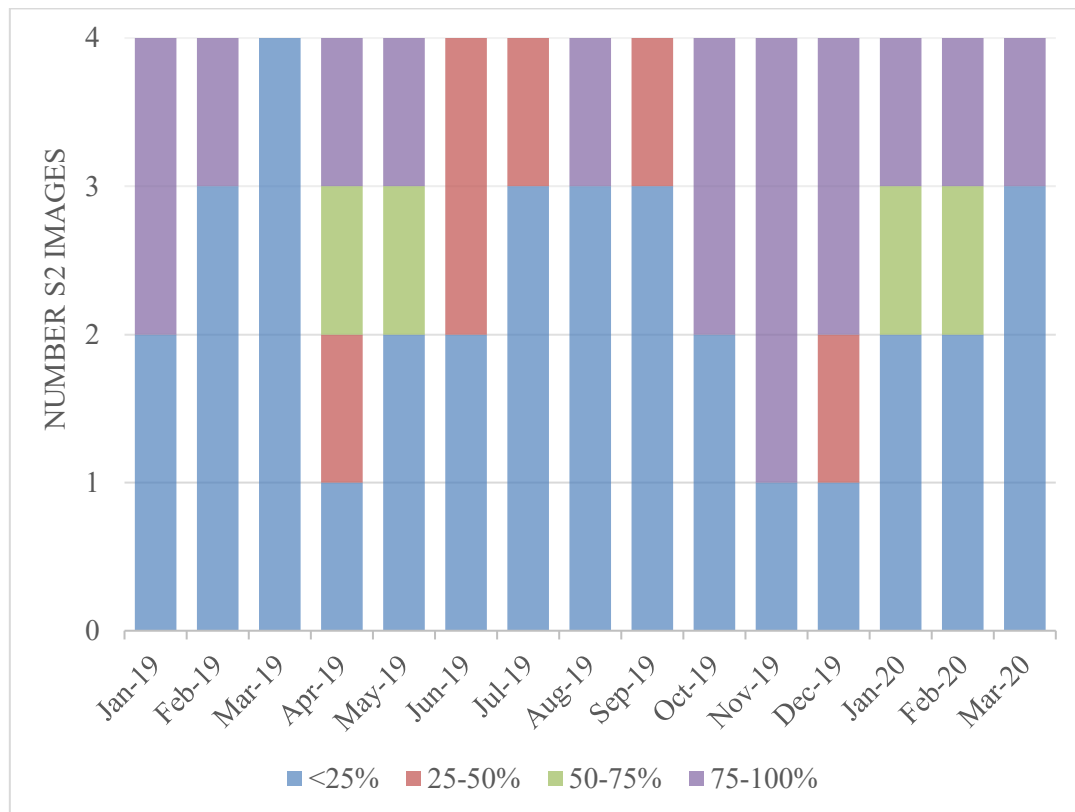


Figure 3.2. Distribution of S2 according to designated classes of cloud cover.

3.3 Applied Methodology

The existing turbidity maps are provided in an aggregated format and tend to contain varying amounts of gaps, with the latter being mainly attributed to clouds and shadows. In addition, they are only available between January 2019 and March 2020. To address these inherent limitations (i.e., limited spatial and temporal coverage), the present study suggests the methodology illustrated in Figure 3.3, where the existing Copernicus maps will be used for validation purposes. This methodology represents an empirical approach and is composed

of four main steps, namely: (i) Data acquisition, (ii) Data pre-processing, (iii) Data processing, and (iv) Data analysis and interpretation.

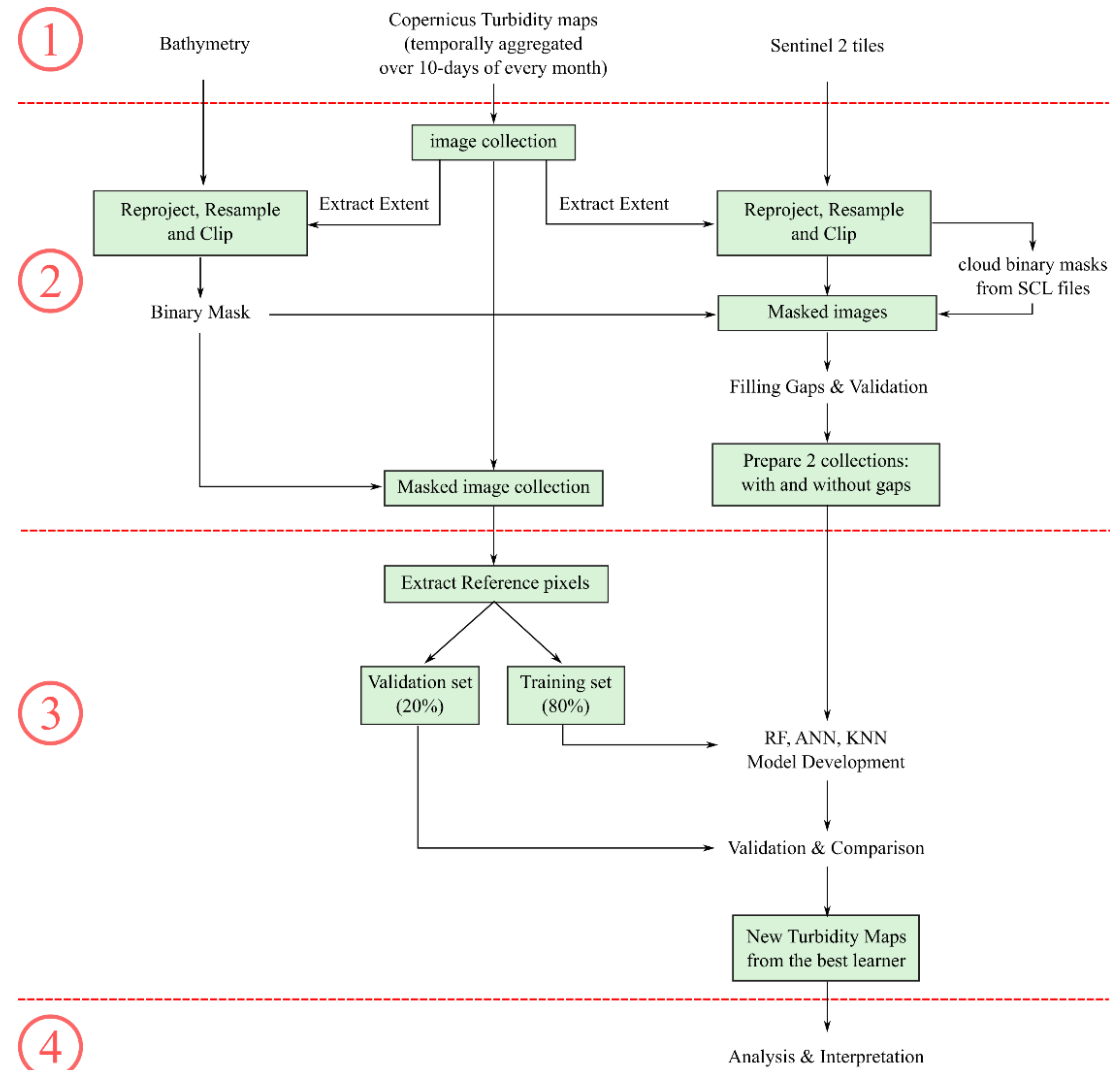


Figure 3.3. Workflow of applied methodology.

This methodology has been implemented using Python programming language through the Jupyter Notebook interactive web tool. The main open-source packages used in this work are the following: SentinelSat (pypi.org/project/sentinelSAT/, accessed online on the 26th January 2022), pandas (pandas.pydata.org, accessed online on the 26th January 2022), numpy (numpy.org/, accessed online on the 26th January 2022), gdal (gdal.org/, accessed online on the 26th January 2022), rasterio (rasterio.readthedocs.io/, accessed online on the 26th January 2022), sklearn (scikit-learn.org, accessed online on the 26th January 2022), keras (keras.io, accessed online on the 26th January 2022), and joblib (joblib.readthedocs.io, accessed online on the 26th January 2022). In

addition, it is to be noted that simulations were run using a personal laptop (Windows system on an Intel Core i7 2.6 Ghz processor with eight cores and 16 GB of RAM) and 12 desktop laboratory computers (Windows system on an Intel Core i7 3.7 Ghz processor with four cores and 16 GB of RAM).

3.3.1 Pre-processing

The pre-processing step refers to preparing the EO data for use as input data in the ML models. This step is composed of two main tasks: (i) initial transformation of data so that they share the same geographic extent, coordinate system, and grid cell size ($\sim 100\text{m}$), and (ii) gap-filling (GF) to be applied on images that suffer from cloud cover. Initial pre-processing includes constructing two binary masks from: (i) the S2 Scene Classification Layer (SCL) image, and (ii) the bathymetry image. The first mask will help in excluding pixels that represent cloud shadows, clouds with either a medium or high probability, and cirrus. Whereas, the second mask will help in excluding pixels with a depth less than 1m. It is to be noted though, that the binary mask derived from the bathymetry product was not very accurate. Thus, after testing, the latter has been modified so as to exclude the shoreline, the river mouth, shipyard, and a small section of the wetland (Figure 3.4). Experimentation proved that doing so will help in: (i) reducing the number of erroneous estimations attributed to adjacency from land and bottom reflectance, as well as (ii) speeding the convergence of the algorithm (Appendix E).

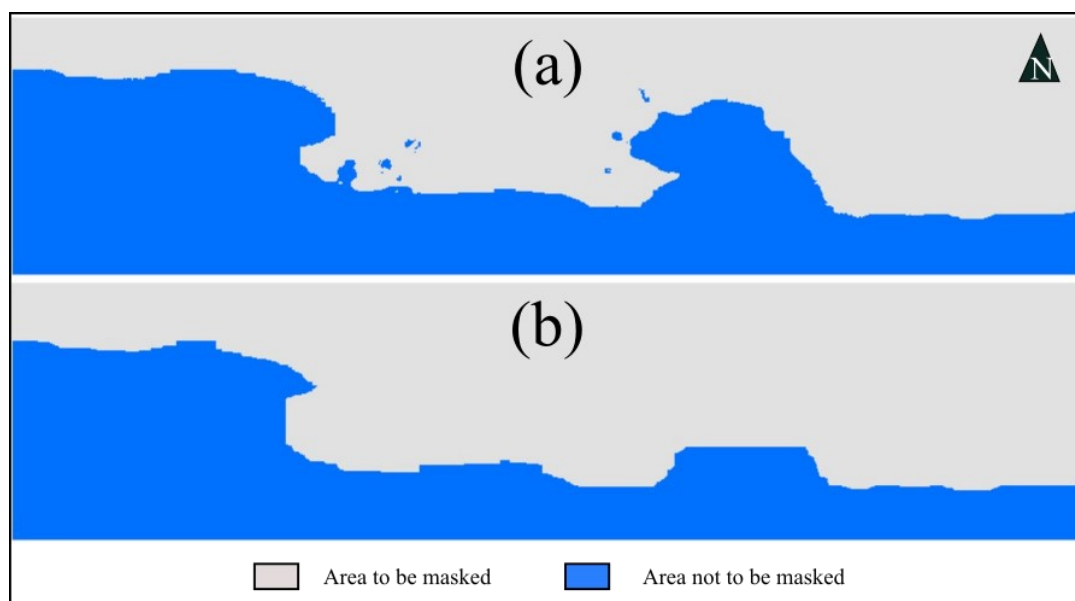


Figure 3.4. Adopted land mask (a) before and (b) after applying a buffer.

The gap filling step is based on applying a variation of the Direct Sampling (DS) approach on downloaded satellite images. The DS approach represents a multiple-point statistical simulation technique that has been first introduced by Mariethoz et al. (Mariethoz et al., 2010) as a way to reconstruct images that suffer from data gaps. The core concept of this approach is about identifying the neighborhood pixels (N_x) of an unknown pixel value at location 'x' (in the target image) and then searching for a replicate with a neighborhood (N_y) that closely matches N_x (in the training image). The filling path of unknown pixels (UP) can be random or unilateral and the selected training image (TI) may represent a different measurement acquisition date. The potential of the DS approach has been examined by experts working in different research fields ranging from mapping soil patterns (E. Meerschman et al., 2013) to improving remote sensing products (i.e. Landsat-7 imagery that were affected by systematic gaps due to an instrument failure) (Yin et al., 2017). For a detailed description of the original DS method, readers are referred to the papers written by Mariethoz and Renard (Mariethoz and Renard, 2010) and Meerschman et al. (Eef Meerschman et al., 2013). Though, it is to be noted that over the study period, there were two instances where the selected S2 image was 100% covered by clouds. In these two cases, the simulated cloud free image was derived from the mean of S2 images acquired within a three days period.

3.3.1.1 Experimental setting

The implementation of the DS method adopted in this study proposes two changes to the original algorithm. These changes have been tested only for coastal waterbodies. The main aim of these changes is in decreasing the randomness factor associated with the DS method, while improving the overall processing speed. Figure 3.5 shows the different steps associated to applying the modified GF algorithm for a selected image band.

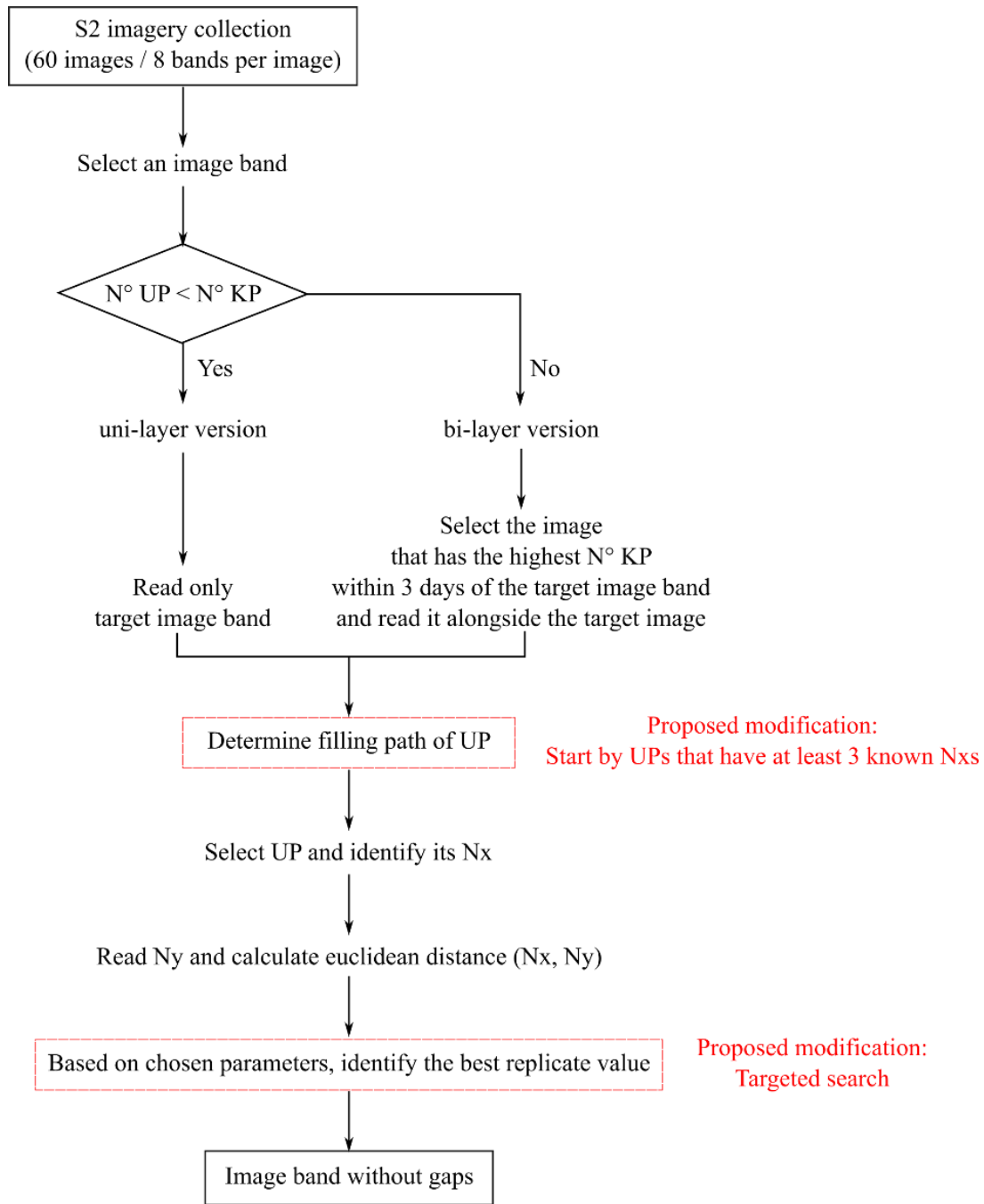


Figure 3.5. Workflow of the gap-filling algorithm (UP: Unknown Pixel, KP: Known Pixel).

In the DS method, the user can define the filling path of UPs to be either random or unilateral. Most often, experts tend to opt for a random approach so as to reduce the burden in processing time. However, even when the user chooses the unilateral approach, there is a good chance that the first selected UP does not have any known neighbors, and therefore it will be assigned a random value. If this value is too different from the actual observation, it will negatively affect most

of the subsequent iterations. To avoid such a situation, the present study suggests that giving priority to filling UPs that have at least three known neighbors would give more robust results (less random) without slowing the algorithm too much. The part of the algorithm that will search for a replicate will then be run using the selected subset of UP (with the order of items in the subset being random). At the end, the algorithm will update the target image with the new KP values and check again for UP that have at least three KP in N_x . This operation will be repeated until no UP is identified in the target image (Figure 3.6). It is to be noted that in case the algorithm scans the whole image and does not find any UP with at least 3 known neighbors, it will instead randomly select the coordinates of five UP (If the number of UP is below five, it will select one UP. Though, such a situation did not occur in the present study).

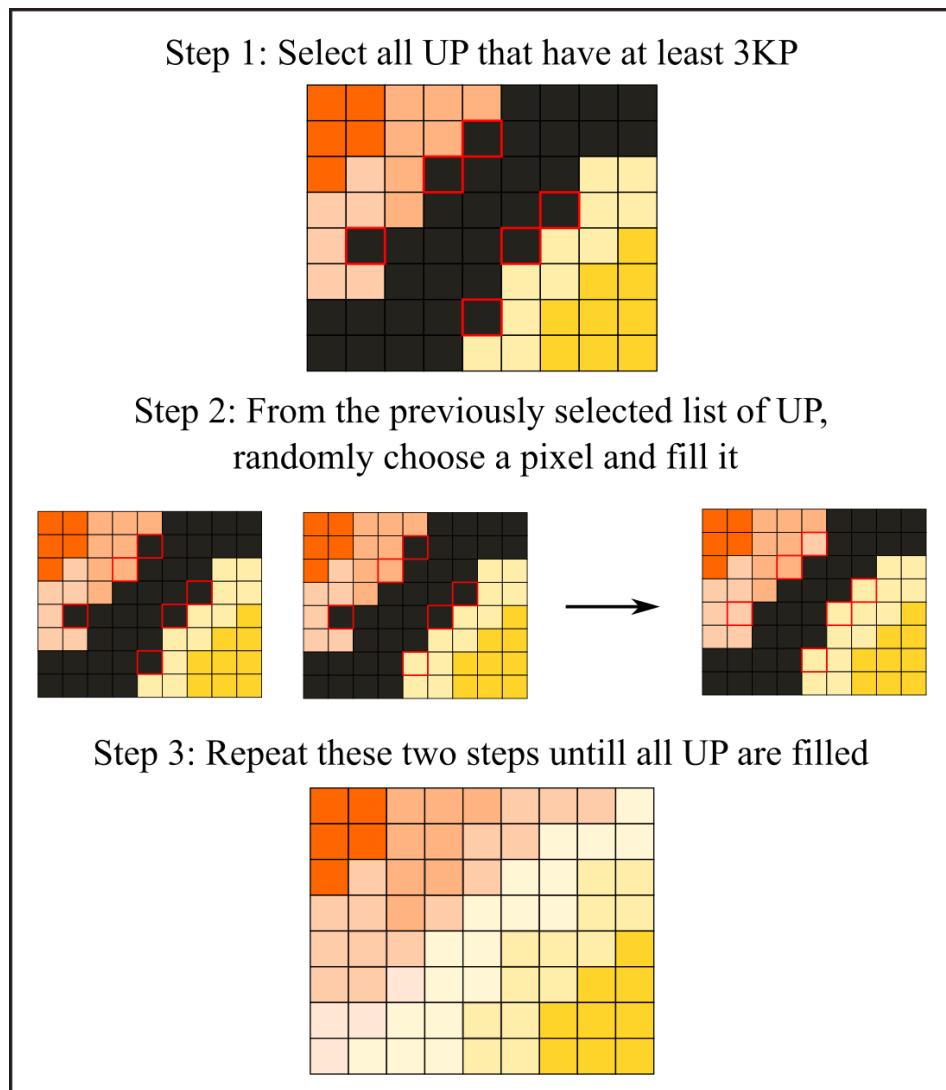


Figure 3.6. Proposed new filling path method for unknown pixels.

With regards to the extreme cost in time of this method, a possible solution is to run the algorithm using all computer processing cores in parallel. In addition to that, experimentation indicates that applying a more targeted search (for replicate values) will allow the algorithm to converge faster (Figure 3.7). Basically, instead of exhaustively searching the whole TI or a random portion of it (if the “fraction” parameter is set to be less than 1), the following steps are proposed: (i) Construct multiple sub-lists from the original list of KP coordinates in the TI. The number of sub-lists is specified by the user based on a new parameter called “n2”. For the example presented in Figure 3.7, if KP values are between 100 and 300 and n2 is set to be two. The algorithm will divide the KP coordinates into two sub-lists, with the first one representing KP values between 100 and 149 and the second sub-list representing KP values between 150 and 300. It should be noted that the length of sub-lists will almost always differ from each other; (ii) For each element in N_x , identify the sub-list it corresponds to, and then join them all into one; and (iii) Search for a replicate in the latter. It should be noted that the user still maintains the ability to apply the “fraction” parameter on the newly constructed list of KP coordinates. Following these steps will reduce the excessive computational burden of DS without compromising the overall accuracy. This idea is based on the assumptions that, in general: (i) N_x tends to share similar spectral characteristics with the selected UP and (ii) that high/low pixel values tend to be clustered together. It is to be noted that assigning a value of one to the parameter “n2” means that the algorithm will behave as defined by the original DS (no construction of new sub-lists).

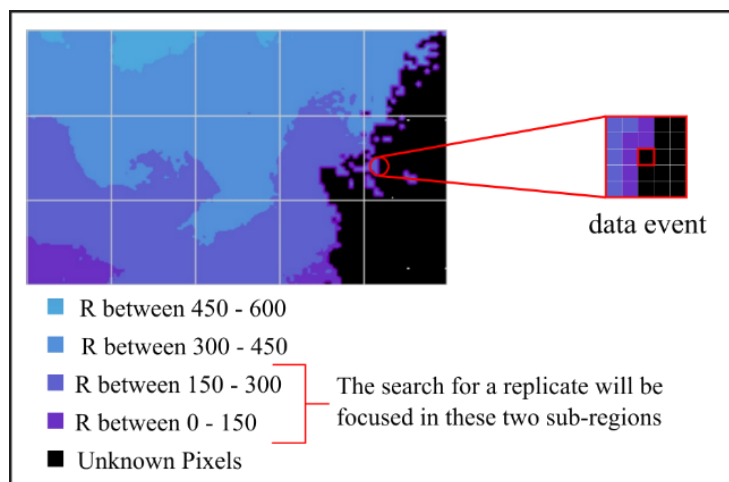


Figure 3.7. Example of a targeted search operation.

For the present implementation of the DS method, two versions of the algorithm were prepared: one for a uni-layer (UL) simulation and the other for a bi-layer (BL) simulation. In case of a UL simulation, the algorithm will consider the target image as the TI if the number of pixels covered with clouds does not exceed the number of existing KP. Otherwise, it will read the S2 images acquired within three days from the acquisition date of the target image and select the one with the highest number of KP as the TI. In case of a BL simulation, the construction of sub-lists from KP values will only be applied to the target image, while the selection of an auxiliary image will be based on which image (within 3-days period) has the highest number of KP. The main parameters adopted in these two versions are summarized in **Erreur ! Source du renvoi introuvable.** The search for the best parameter combination was carried out using the following approach: (i) select an image and the parameter to be investigated, (ii) choose fixed values for the other parameters, and (iii) run the GF algorithms for all possible cases indicated for the selected parameter. Repeat the previous steps for other parameters. After finishing the first iteration of tests, the same process will be done for other images corresponding to different cloud percentages. At the end, a unified parameter combination will be adopted irrespective of cloud percentage. Similar to the work of Yin et al. [10], the pattern matching between N_x and N_y was assessed based on calculating the Euclidean distance equation described in their paper.

Table 3: Parameter values that were investigated for achieving the best gap filling results

Parameter	Cases
n1	8 – 14 – 20 – 24
t	0 – 0.005 – 0.008 – 0.01
f	0.6 – 0.8 – 1
n2	1 – 25 – 50 – 75 – 100
(w1; w2) (Their sum should be equal to 1)	(0.5; 0.5) – (0.7; 0.3) – (0.8; 0.2) – (0.9; 0.1)

n1: Number of neighbors to be checked; t: Distance threshold; f: Search fraction; n2: Number of sub-groups; w1: Weight value corresponding to target image (Bi-layer algorithm); w2: Weight value corresponding to auxiliary image (Bi-layer algorithm).

For the purpose of testing the developed algorithm, five realistic simulations of clouded images were used. These artificially clouded images were generated by:

(i) preparing five cloud masks from Scene Classification Layers (SCL) corresponding to satellite images that were acquired on five different dates, and (ii) applying them on an originally S2 cloud free image. The selection of all these images was random (Table 4). Testing the different parameter combinations has been done on the blue band in a way to balance the prediction accuracy and the cost in time. A unified best parameter configuration was selected that is the following: $n1=20$; $t=0.008$; $f=0.8$; $n2=100$ and $(w1; w2) = (0.5; 0.5)$. After selecting the best parameter configuration, the latter was applied on three S2 bands that are namely: B02, B03, and B04 (Table 2).

These S2 bands are characterized by a good radiometric quality (12 bits), a high revisit frequency (2-3 days at the study area), and a high spatial resolution (10 m), all of which contribute in better capturing of differences in water reflectance (Gholizadeh et al., 2016; Kupssinskü et al., 2020). Moreover, it should be noted that the GF algorithms were tested on other S2 bands (i.e., bands 5 to 8a). However, as the computed error metrics were very low or in the negative, they were not reported nor included in further analyses. The unsuitability of using the modified DS method to fill the gaps in the aforementioned bands (i.e., bands 5 to 8a) is attributed to the high variation among neighboring reflectance values. The use of another gap filling method that is based on the analysis of time-series data will most likely allow the user to overcome this issue.

Table 4: Acquisition dates and cloud percentage of S2 images that were used in testing the modified DS algorithm

Used Image	Date	Cloud (%)
Original cloud free image (target image)	2019-09-12	0.00
Cloud mask 1	2019-10-17	11.57
Cloud mask 2	2019-03-14	23.35
Cloud mask 3	2019-06-19	38.89
Cloud mask 4	2020-02-19	57.30
Cloud Mask 5	2019-11-19	79.30
Image acquired after target image (within 3 days)	2019-09-10	100.00
Image acquired before target image (within 3 days)	2019-09-15	38.82

3.3.1.2 Performance assessment

The evaluation of gap-filling results has been performed five times for each artificially clouded image. This is to confirm that the randomness factor in the GF algorithm does not have a large effect on overall accuracy. Performance was assessed in two ways: (i) based on comparing the mean reflectance, standard deviation (SD) and skewness coefficient of an image before and after applying the GF algorithm, and (ii) using the mean values of each of the following five error metrics: (i) median absolute percentage error (MdAPE), (ii) root mean square logarithmic error (RMSLE), (iii) root mean square error (RMSE), (iv) mean absolute error (MAE), and (v) coefficient of determination (R^2). Their formulas are expressed below:

$$MdAPE (\%) = \text{median}_{i=1,N} \left(100 * \frac{|Ei - Mi|}{Mi} \right)$$

Where: E: estimated value, M: actual value, i: selected UP, and N: total number of UP.

MdAPE is the middle value of the absolute percentage errors after ordering them ascendingly (or the average of the two middle values if N is an even number). The closer the MdAPE value is to zero the more unbiased the estimated value is.

$$RMSLE = \sqrt{\frac{\sum_{i=1}^N (\log_{10}(Ei) - \log_{10}(Mi))^2}{N}}$$

RMSLE is a modified version of the RMSE that reflects the relative error between the estimated and actual values without giving importance to the error magnitude. It is usually adopted when the user does not want to penalize large individual errors that may occur.

$$RMSE = \sqrt{MSE} \quad \text{with} \quad MSE = \frac{\sum_{i=1}^N (Ei - Mi)^2}{N}$$

RMSE is one of the most commonly used error metrics for quantifying loss. However, it can be significantly affected by outliers.

$$MAE = \sum_{i=1}^N \frac{|Ei - Mi|}{N}$$

MAE is usually used to give an average estimate of errors. MAE is more robust to the presence of a few error outliers than RMSE and R^2 .

$$R^2 = \left(\frac{\sum_{i=1}^N (Ei - \bar{Ei})(Mi - \bar{Mi})}{\sqrt{\sum_{i=1}^N (Ei - \bar{Ei})^2 \sum_{i=1}^N (Mi - \bar{Mi})^2}} \right)^2$$

R^2 is a widely adopted metric for assessing the consistency between estimated and actual observations. However, a few extreme errors can have a severe impact on this metric.

3.3.2 Development of machine learning models

The processing of the organized image collections refers to developing ML models that can reliably predict the turbidity levels in the study area using the selected bands and/or band ratios. For this purpose, three ML techniques that are: random forests (RF), artificial neural networks (ANN), and K-nearest neighbors regressor (KNN) were examined. The simulated results were evaluated against the turbidity maps available through the Copernicus Global Land Service Program. In total, 15 turbidity maps were used for this purpose with each one representing a specific month.

3.3.2.1 Experimental setting

In the present study, two collections of S2 images were prepared (with and without gap filling). Each one of these images are composed of nine layers (i.e., green, red and blue bands; as well as their corresponding six ratios), which were used to feed the aforementioned ML models. Pixel values were extracted from the selected input layers according to the three cases described in Table 5.

Table 5: Number of S2 images used in ML

	N° of S2 images used within those available in the 10-days period	Total n° of input layers (bands and band ratios)
Case 1	Use only the best image	9
Case 2	Use only the best two images	18
Case 3	Use only the best three images	27

The best image refers to the image with the least cloud cover%.

Erreur ! Source du renvoi introuvable. illustrates the necessary steps for preparing data to be fed into a selected ML model. The extraction of reference turbidity and reflectance pixel values was done while ensuring a representative selection of the latter. The necessary steps to perform the extraction task are the following: (i) Merge the selected turbidity map with the corresponding reflectance layers into one stacked image, (ii) Exclude any reflectance pixel value, if it corresponds to an unknown turbidity pixel value, (iii) Classify the turbidity map into several groups based on an equal distance of 20 NTU, (iv) Determine the number of pixel coordinates to be retained per identified turbidity group (n_2), (v) Randomly select n_2 pixel coordinates per identified turbidity group, and (vi) Extract corresponding turbidity and reflectance values. It is to be noted that the

number of coordinates of pixel values from images with and without gap filling is close to each other (**Erreur ! Source du renvoi introuvable.**). Overall, the average number of extracted pixels per image is 4713 (without GF), and 5093 (with GF). In addition, the extracted pixel values will be randomly split into a training and testing datasets based on an 8:2 ratio. The training dataset will be used for tuning the ML models, whilst the testing dataset was used for ensuring that the models can be generalized on unseen data and for comparing the RF, ANN, and KNN models. The same partitioning of training/testing datasets was maintained across the different ML models (Fernández-Delgado et al., 2014).

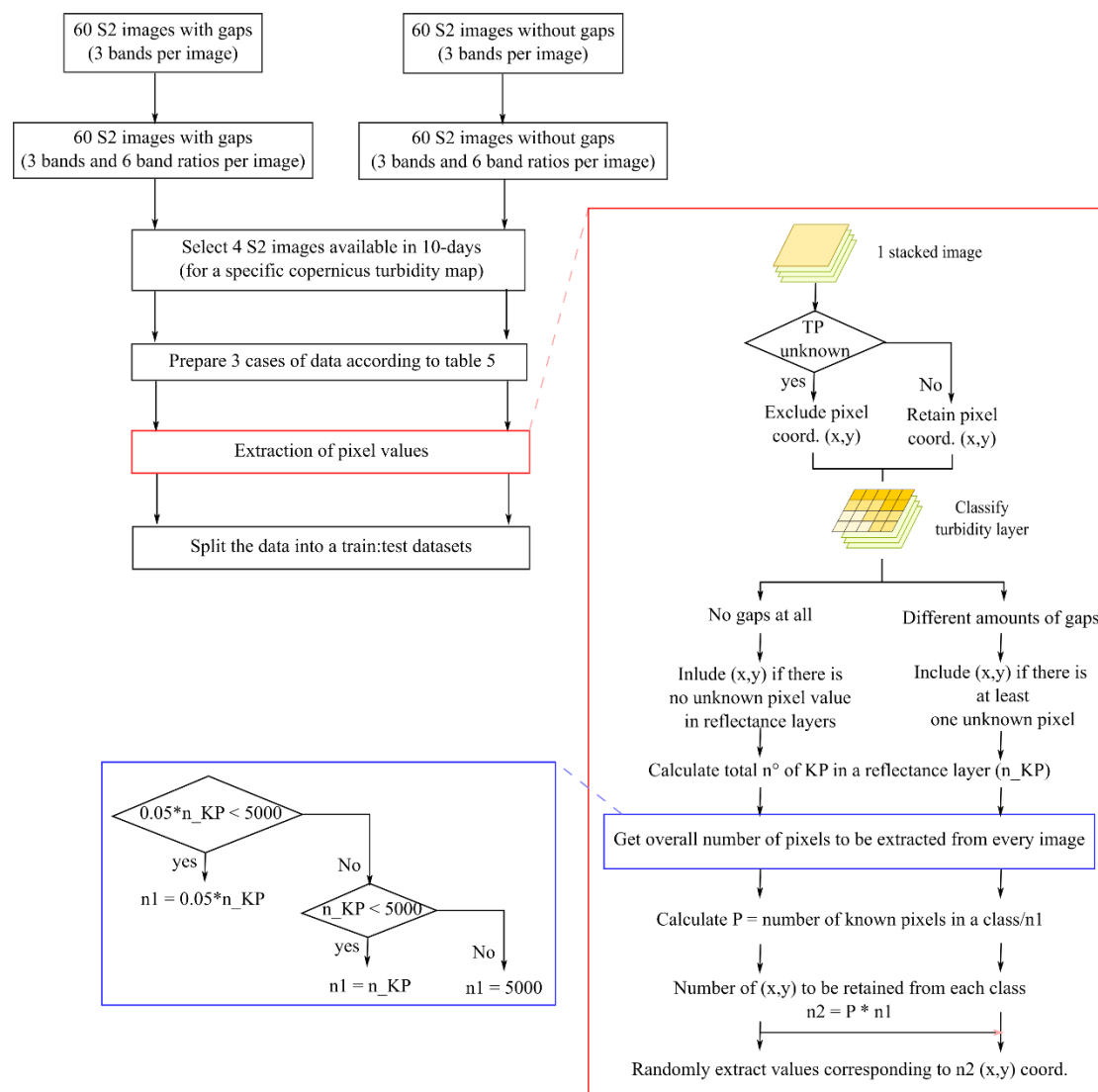


Figure 3.8. Preparation of data to be fed into ML models (TP: turbidity pixel; KP: known pixel).

Table 6: The average number of extracted pixel values per case of selected best Sentinel2 images.

	Best_1	Best_2	Best_3
Without GF	5079	5049	4011
With GF	5093	5093	5093

The ML models were trained and tuned using a double cross validation (CV) approach. The K-folds CV approach is based on dividing a dataset into K non-overlapping folds. Each of the latter were used for testing the model that has been trained using the other folds (Figure 3.9). The first CV (k=5) was applied to find the optimal hyper-parameter configuration for all 15 turbidity maps. The second CV (k=10) was applied after adopting the unified hyper-parameter configuration. Its results will be used for assessing overall model performance and double-checking that the model is not over-fitting. The evaluation of all simulations was carried out by reporting the average values of two performance metrics that are: (i) the mean squared error (MSE) and (ii) the coefficient of determination (R^2).

Validate	Train	Train	Train	Train
Train	Validate	Train	Train	Train
Train	Train	Validate	Train	Train
Train	Train	Train	Validate	Train
Train	Train	Train	Train	Validate

Figure 3.9. Example of a 5-folds cross validation approach.

3.3.2.2 Model selection and interpretation

To further ensure that the three ML models do not exhibit overfitting, the latter was run a third time using all training data without CV (Kupssinskü et al., 2020). After verifying that the resulting average error metrics are within the range of MSE and R^2 values computed with CV, the best learner was selected. This selection is based on which one has the highest accuracy value. Following the selection of the best learner, the latter was used for deriving new turbidity maps.

Afterwards, these maps were analyzed and interpreted in terms of their spatial and temporal variation.

Chapter 4: Results

4.1 Performance Evaluation of the Modified Gap Filling Algorithm

As was described before, two versions of the GF algorithm (UL and BL) were implemented on five artificially clouded images (with every image containing three bands). The performance of the aforementioned algorithms was, mainly, evaluated based on: (i) five error metrics (Table 8), and (ii) three statistical metrics (Table 7). Measurements of the processing time were taken for each experiment. However, there is some bias associated with these measurements as the performance of used computers vary between each other. In fact, as they lack proper maintenance, they tend to lag at different times during the processing (especially when the latter takes more than one day). To simplify the presentation of results, the blue band was chosen as a representative example to compare the two versions of the GF algorithm. The full set of results is summarised in appendices A, B and C. The results presented in tables Table 8 to Table 10 correspond to the mean realization of five repetitions.

The examination of the gap filled images through error metrics indicates that, when filling images with cloud% below 50, both the UL and BL algorithms generate similar good results. In case of images with cloud percentage above 50%, only the BL algorithm is capable of generating good results (Table 7). In addition, the quantitative assessment based on the statistical metrics indicates that the reconstructed images are, in general, similar to the original ones. For instance, in case of the blue band, the mean reflectance of reconstructed images ranges between 305.22 and 308.20, which is very similar to the mean reflectance corresponding to the original cloud-free image (estimated to be 308.30) (Table 8).

Table 7: Error metrics values for gap filled images (Blue Band)

	Cloud%	RMSLE	MdAPE	MAE	RMSE	R2
UL	11.57%	0.06	5.38	31.45	45.20	77.75
	23.35%	0.03	3.83	16.11	27.64	93.26
	38.89%	0.05	5.09	24.38	36.78	86.40
	57.30%	0.17	35.38	107.39	124.3	-21.68
	79.30%	0.21	53.74	130.44	147.18	-88.69

BL	11.57%	0.06	5.20	31.27	45.34	77.61
	23.35%	0.03	3.88	16.12	27.41	93.37
	38.89%	0.05	5.14	24.80	37.54	85.83
	57.30%	0.07	10.52	42.53	58.31	73.23
	79.30%	0.06	5.49	27.87	46.87	80.87

Table 8: Statistical indices of gap filled images (Blue Band)

	Cloud%	Mean	SD	Skew
Original Image	0%	308.30	109.23	0.90
Gap Filled Image (UL)*	11.57%	306.33	107.86	0.94
	23.35%	307.47	109.30	0.88
	38.89%	308.20	108.35	0.87
	11.57%	306.62	108.11	0.94
Gap Filled Image (BL)	23.35%	307.60	109.28	0.88
	38.89%	307.63	108.24	0.87
	57.30%	305.22	92.03	0.99
	79.30%	308.20	102.09	0.76

*: results corresponding to negative R^2 values were not reported.

In terms of processing time, it appears that when the cloud percentage is small, the BL algorithm tends to converge faster than the UL algorithm, and vice versa (Table 9). However, it seems that the difference in processing time between the UL and BL algorithms is, particularly, significant for the red and green bands and in case of a cloud percentage $>35\%$ (Appendix C). For the mass-production of gap-filled images, as long as the cloud percentage is below 50%, the UL version was chosen instead of the BL. This choice is attributed to how the UL version will require way less imagery data, and by extension the costs in downloading, storing, and pre-processing the latter will be more manageable.

Table 9: Comparison of processing time between UL and BL versions (Blue Band)

Algorithm	Cloud%	Time (min)
UL	11.57%	214.0
	23.35%	293.2
	38.89%	318.8
BL	11.57%	150.2
	23.35%	274.4
	38.89%	1649.8

An analysis was also undertaken to assess the gain in processing time after modifying the original Direct Sampling method. As was described before, the latter has been modified in a way to allow for a more targeted search for the best replicate values. This modification was carried out by introducing another parameter “n2” that will classify pixel values into sub-groups. Assigning a value of 1 to n2 means that the user will not adopt this modification when running the code. The results revealed in Table 10 indicate: (i) some slight changes in overall performance, and (ii) a clear gain in processing time. For instance, in case of an image with a cloud percentage of 38.89, the cost in time was reduced from 20 hours and 32 minutes to 2 hours and 48 minutes (UL version).

Table 10: The processing time of a UL algorithm based on the selected number of sub-groups (Blue Band)

Cloud%	n1	t	f	n2	R ² *	Time	Gain in Time (%)
11.57%	14	0.01	0.6	Benchmark	73.02	18h 20m	-
				25	72.6	3h 3m	83.36
				50	73.28	1h 59m	89.18
				75	76.07	1h 35m	91.36
				100	77.09	1h 15m	93.18
23.35%	14	0.01	0.6	Benchmark	89.17	3d 1h	-
				25	92.62	5h 45m	68.64
				50	92.23	10h 12m	44.36
				75	92.92	2h 1m	89.00
				100	92.88	2h 2m	88.91
38.89%	14	0.01	0.6	Benchmark	85.84	20h 32m	-
				25	83.32	3h 15m	82.27
				50	84.89	2h 39m	85.55
				75	85.11	2h 46m	84.91
				100	85.88	2h 48m	84.73

*: these values were derived from the UL algorithm; h: hours; m: minutes; Benchmark: n2=1.

4.2 Performance Evaluation of the Adopted Machine Learning Algorithms

As was described before, three ML algorithms (RF, ANN, and KNN) were implemented on eight image collections. The latter were prepared based on the number of used S2 images and whether or not the GF algorithm was applied. The

assessment of the performance of these algorithms was carried out while adopting the same train-test split of the datasets. From the train split, a K-fold cross-validation was performed, and the mean values for MSE and R^2 were generated and are reported in Table 11. Generally speaking, all three ML models gave similar results with the RF model always reaching the highest R^2 values. In addition, results indicate that the higher the number of S2 images that are fed to the models, the better their performance will be. However, at the same time, it is worth mentioning that using only two S2 images to feed the ML models still gave very acceptable results. For instance, in case of the RF model, R^2 values are around 90% when using the best two S2 images. Table 11 reveals, as well, that applying the GF algorithm on S2 images does not negatively affect the overall performance of ML models. Moreover, the inspection of results indicates that only during the summer (June, July, and August) do R^2 values fall below 80%.

Table 11: Mean error metric values derived from the cross-validation step

	N° images	Algorithm	MSE	R^2
With GF	1	ANN	0.0027	76.61
		KNN	0.0026	78.02
		RF	0.0025	78.60
	2	ANN	0.0013	88.88
		KNN	0.0011	89.87
		RF	0.0010	90.93
	3	ANN	0.0010	91.44
		KNN	0.0010	91.16
		RF	0.0008	93.02
Without GF	1	ANN	0.0021	76.92
		KNN	0.0020	78.70
		RF	0.0019	79.46
	2	ANN	0.0008	88.12
		KNN	0.0007	88.93
		RF	0.0007	89.66
	3	ANN	0.0016	88.94
		KNN	0.0018	87.76
		RF	0.0015	89.82

Following the cross-validation step, the number of used S2 images was fixed at two. Subsequently, the three ML models were run on the whole training set and assessed based on the testing set. Figure 4.1 shows that the results derived without CV are within the range of results derived with CV. Thus, it is unlikely that any of

the three ML models exhibits over-fitting. The RF model was selected as the best learner. The analysis of its results indicates that the most important band or band ratio layers are the following: 'B/R', 'G/B', 'G' (without GF) and 'B/G', 'G/B', 'G' (with GF).

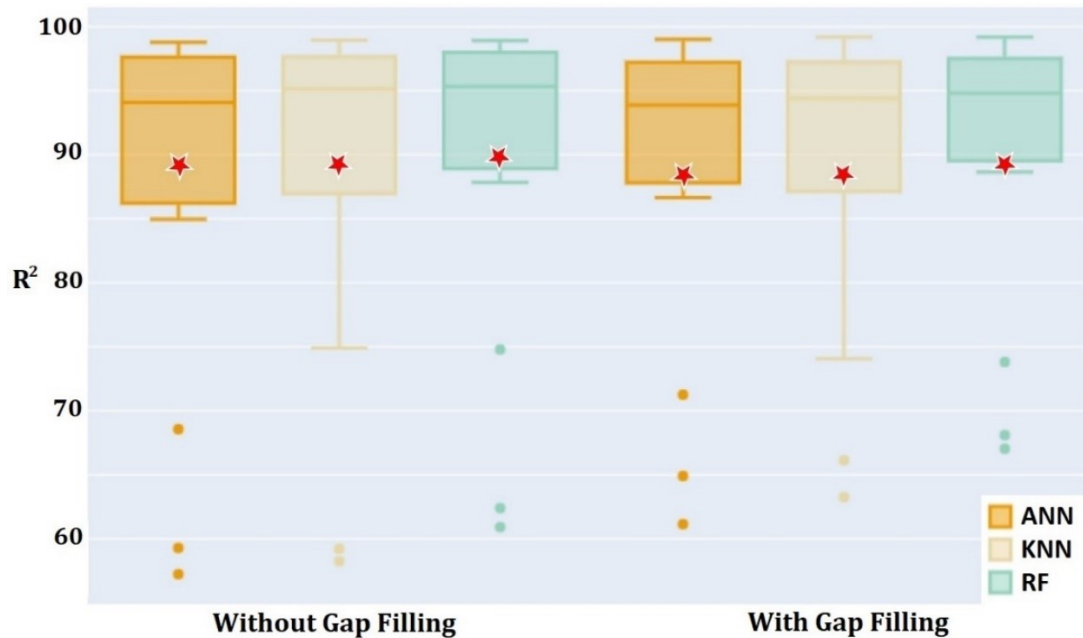


Figure 4.1. Boxplot representing the R2 metric obtained on the cross-validation (colored dots correspond to outliers); the metric on the test set is marked with a red star.

4.3 Spatial and Temporal Variation of Turbidity

The RF model was applied on all processed images in the study area. Figure 4.2 compares actual turbidity maps with simulated ones for a selected two representative dates (21 January 2019, and 11 March 2019). The visual inspection of the maps reveals that the spatial distribution of turbidity values in images derived with and without GF is close to that of actual observations. For these two selected dates, the overall average error percentage corresponding to simulated turbidity maps is about 13.7% (without GF) and 12.2% (with GF). Indeed, **Erreur ! Source du renvoi introuvable.** further confirms that turbidity maps simulated with GF: (i) slightly outperform those without GF in terms of error metrics, and (ii) have the clear advantage of being always gap-free. However, **Erreur ! Source du renvoi introuvable.** reveals, as well, that using the best two S2 images (without GF) can in most cases reduce the number of unknown pixels to a large extent or even totally. Although, it should be noted that in one case the number of

unknown pixels rather increased by a small margin, following the use of S2 images without GF (2019/08/11). Moreover, it should be highlighted that even though the general pattern of original turbidity maps is well maintained in the simulated ones, the latter appears to suffer from some per-pixel inaccuracies in case of outlier turbidity values. For instance, the maximum turbidity value in the example simulated maps does not exceed 20 NTU, whilst in the original images, the maximum value does reach 45 NTU (January 2019) and 215 NTU (March 2019).

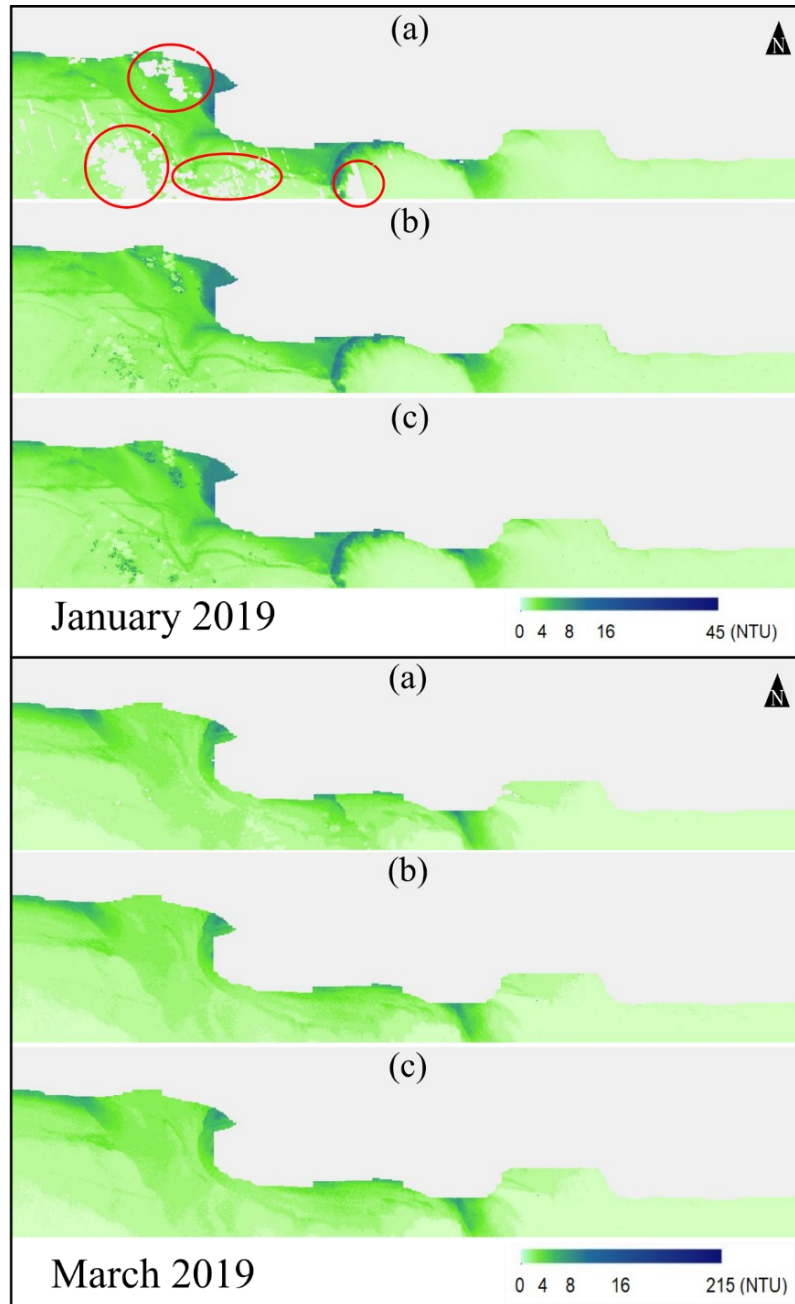


Figure 4.2. Turbidity maps corresponding to two months (a: original Copernicus image; b: simulated image without GF; c: simulated image with GF; Red circles indicate gaps due to cloud cover.

Table 12. Change in the number of unknown pixels and error percentages corresponding to simulated turbidity maps.

	SI Without GF				SI With GF		
	CUP (%)	MdAPE	AEP		CUP (%)	MdAPE	AEP
2019/01/21	-100	8.76	-3.57	-100		8.76	-3.27
2019/02/11	-98	45.28	-62.45			41.47	-52.18
2019/03/11	-100	9.9	-4.19			9.74	-4.46
2019/04/11	-100	11.2	-3.37			9.59	-4.52
2019/05/11	-100	8.26	-3.24			8.22	-3
2019/06/11	-100	19.14	-19.19			19.25	-20.12
2019/07/11	-100	19.4	-18.09			19	-16.68
2019/08/11	+4	14.37	-14.74			11.86	-8.18
2019/09/11	---	7.74	-2.83			7.69	-1.7
2019/10/11	-100	16.7	-17.55			16.7	-17.55
2019/11/11	-65	11.61	-77.46			7.74	-7.79
2019/12/01	-100	14.98	-66.22			5.94	-2.84
2020/01/01	-97	7.6	0.93			7.5	-2.19
2020/02/11	-100	6.28	-1.28			6.05	-1.52
2020/03/11	---	4.38	-0.9			4.38	-0.9
Average Value		-19.61	13.71			-9.79	12.26

AEP: Average Error Percentage (%) = $\text{mean}[100 * (\text{actual} - \text{predicted}) / \text{actual}]$; SI: Simulated Image; CUP: Change in the number of Unknown Pixels compared to original turbidity maps (%) / a positive sign indicates an increase, whilst a negative sign indicates a decrease; ---: both original and simulated turbidity maps did not contain any gaps.

Figure 4.3 describes the temporal variability of turbidity over the study period, in terms of average, interquartile range (IQR), and standard deviation (SD) values. It is clear that the highest turbidity values, as well as, the highest fluctuation tend to occur in the winter season. This is, most likely, attributed to the increase in the frequency and intensity of runoff and soil erosion phenomena in that period of the year. In addition, it appears that the turbidity maps that were simulated with GF are the most coherent with the variability trend of actual observations. Generally speaking, the average of actual turbidity observations ranges between 0.32 and 7.49 NTU. In comparison, the average of simulated turbidity observations: ranges between 0.33 and 10.66 NTU (without GF), and ranges between 0.32 and 7.45 NTU (with GF).

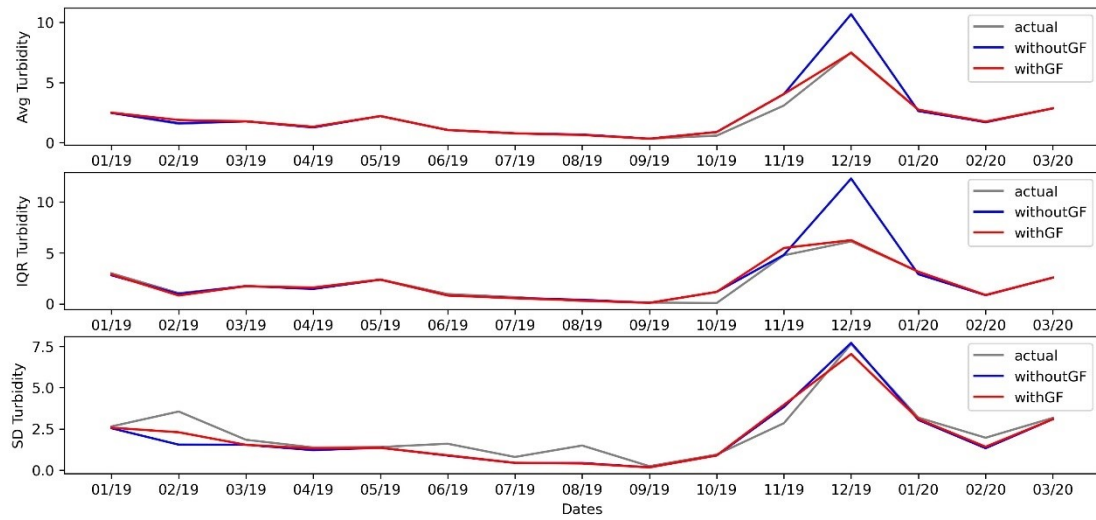


Figure 4.3. Temporal graphs of average, interquartile range (IQR), and standard deviation (SD) turbidity values in the study area derived from actual and predicted observations, with and without GF.

Chapter 5: Discussion

5.1 Gap Filling Algorithm

The results derived from the modified GF algorithms reveal a satisfactory performance when filling in the blue, green, and red bands of the S2 satellite images for all five considered cases of artificially clouded images. Overall accuracy is largely dependent on two factors: (i) the percentage of cloud cover, and (ii) the number of similar patterns an identified N_x has. This means that an increase in cloud percentage does not always equate to a decrease in overall performance of the used GF algorithm. In fact, as long as all N_x s have enough similar N_y s, the adopted GF algorithm will achieve acceptable results. In addition, it should be mentioned that even though the Direct Sampling method is able to preserve the general patterns in a satisfactory manner, it tends to be less sensitive of individual pixel values. This has been highlighted by Yin et al. (Yin et al., 2017) as an inherent disadvantage of this method, which can be mitigated by opting for the multi-variate version of the algorithm (e.g., bi-layer). However, in case of the present study, when filling gaps in all collected S2 images, the BL version was implemented only when the cloud percentage of an image exceeded 50%. This is mainly due to the technical difficulties associated with downloading, storing and pre-processing a larger number of satellite images. Furthermore, it is to be noted that the Direct Sampling method can be applied to images corresponding to several types of environments (e.g., grassland, forests, farms, etc.).

5.2 Machine Learning Algorithms

The results achieved by the ML algorithms indicate that all developed models can reach a satisfactory performance, with the RF model being consistently the best learner. These findings are similar to those published by Kupssinskü et al. (Kupssinskü et al., 2020) in their study of the behaviour of suspended sediments in two lakes located in Brazil. It is worth mentioning as well that providing the RF model with more input variables such as pH, water temperature, and sea surface salinity may further increase its performance (Alizadeh et al., 2018; Kupssinskü et al., 2020; Sagan et al., 2020). A second

possible solution when reaching the limit of what hyper-parameter tuning can offer is to divide the dataset into subsets, build models that are tailored for each of the latter, and then use ensemble learning to combine the results. However, doing so will increase the complexity of the algorithm and will come with its cost in processing time. The construction of these subsets can be based on a time criterion (e.g., different seasons) or a space criterion (e.g., different water depths). Moreover, it should be noted that using only three bands (blue, green and red) to predict turbidity represents a major advantage of the developed ML model. In fact, this choice allows for faster computations and the possibility to adopt other types of imagery such as those derived from Landsat series or unmanned aerial vehicles (UAVs).

5.3 Spatial and Temporal Characteristics

In regards to the spatial and temporal characteristics of the simulated turbidity maps, the latter seems to suffer from an issue known as “Model Saturation”. In fact, the high turbidity values exceeding 20 NTU tend to be underestimated. This issue has been highlighted in the works of Pereira et al. (Pereira et al., 2018), and Long and Pavelsky (Long and Pavelsky, 2013) and has been attributed to the fact that pixels with high turbidity values tend to be few in number. On another note, it should be highlighted that the generated predictions of turbidity levels may not necessarily follow a realistic seasonal behaviour. This is mainly attributed to the short study period (15 months) that does not allow the model to incorporate seasonality when predicting turbidity levels for a specific location. Using more data with better spatial resolution and for a longer period of time, as well as, validating the model with ground-truth data instead of the 10-days Copernicus product will probably allow the user to overcome both of these issues and improve overall quality.

Chapter 6: Conclusions and Recommendations

The main conclusions of this study can be summarized as follows: (i) The modified GF approach demonstrated a clear reduction in processing time, with a slight improvement in performance compared to the original algorithm, (ii) The performance of the GF algorithm was acceptable only for three bands that are B02, B03, and B04 of S2 images, (iii) The developed GF algorithm has the potential of being applied on imagery taken from other types of environments, (iv) The ML derived turbidity maps have a satisfactory quality, (v) Using the best two S2 images in 10-days period helped, in many cases, to produce gap-free turbidity maps even without GF, (vi) The use of the GF algorithm has contributed to providing more accurate turbidity maps, and the advantage of this algorithm will be more evident in locations and during seasons where rainfall events are more frequent, (vii) The most sensitive bands and band ratios for predicting turbidity are: 'B/R', 'B/G', 'G/B', and 'G', (viii) It was confirmed that the use of ML in combination with remote sensing products can help in generating turbidity maps with a higher spatial resolution (from 100 to 10 m), across a wider area (gaps were filled), for a longer period of time (not only those 15 months), and in a more regular manner (if in-situ data can be provided).

The main recommendations for further improving the performance of the developed algorithms can be summarized as follows: (i) Limiting the use of GF to cases where the best two S2 images are still considered very cloudy (to decrease the overall cost in processing time), (ii) Investigating whether using other auxiliary data, such as pH, water surface temperature and sea salinity can contribute to more robust ML models, (ii) Writing the code in another programming language that supports both fast computations and ML model development. For instance, Julia Language is, recently, gaining a lot of momentum and popularity due to the way it is focused on ensuring a good compilation speed

Bibliography

- Adloff, F., Somot, S., Sevault, F., Jordà, G., Aznar, R., Déqué, M., Herrmann, M., Marcos, M., Dubois, C., Padorno, E., Alvarez-Fanjul, E., Gomis, D., **2015**. Mediterranean Sea response to climate change in an ensemble of twenty first century scenarios. *Climate Dynamics* 45, 2775–2802.
<https://doi.org/10.1007/s00382-015-2507-3>
- Agency, E.S., **2015**. Sentinel-2 User Handbook, *Industrial and Engineering Chemistry*. Paris. <https://doi.org/10.1021/ie51400a018>
- Alizadeh, M.J., Kavianpour, M.R., Danesh, M., Adolf, J., Shamshirband, S., Chau, K.W., **2018**. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Engineering Applications of Computational Fluid Mechanics* 12, 810–823.
<https://doi.org/10.1080/19942060.2018.1528480>
- Allan, M.G., Hamilton, D.P., Hicks, B.J., Brabyn, L., **2011**. Landsat remote sensing of chlorophyll a concentrations in central North Island lakes of New Zealand. *International Journal of Remote Sensing* 32, 2037–2055.
<https://doi.org/10.1080/01431161003645840>
- Antonelli, C., Eyrolle, F., Rolland, B., Provansal, M., Sabatier, F., **2008**. Suspended sediment and ¹³⁷Cs fluxes during the exceptional December 2003 flood in the Rhone River, southeast France. *Geomorphology* 95, 350–360.
<https://doi.org/10.1016/j.geomorph.2007.06.007>
- Asrar, G., **1989**. Theory and Applications of Optical Remote Sensing. John Wiley & Sons, Chichester.
- BluePlan, **2005**. A sustainable future for the Mediterranean, *Rangeland Journal*.
<https://doi.org/10.1071/RJ11056>
- Boero, F., **2015**. The future of the Mediterranean Sea Ecosystem: towards a different tomorrow. *Rendiconti Lincei* 26, 3–12.
<https://doi.org/10.1007/s12210-014-0340-y>
- Bresciani, M., Cazzaniga, I., Austoni, M., Sforzi, T., Buzzi, F., Morabito, G., Giardino,

- C., **2018**. Mapping phytoplankton blooms in deep subalpine lakes from Sentinel-2A and Landsat-8. *Hydrobiologia* 824, 197–214.
<https://doi.org/10.1007/s10750-017-3462-2>
- Caballero, I., Steinmetz, F., Navarro, G., **2018**. Evaluation of the first year of operational Sentinel-2A data for retrieval of suspended solids in medium- to high-turbidity waters. *Remote Sensing* 10.
<https://doi.org/10.3390/rs10070982>
- Calvo, E., Simó, R., Coma, R., Ribes, M., Pascual, J., Sabatés, A., Gili, J.M., Pelejero, C., **2011**. Effects of climate change on Mediterranean marine ecosystems: The case of the Catalan Sea. *Climate Research* 50, 1–29.
<https://doi.org/10.3354/cr01040>
- Cannizzaro, J.P., Carder, K.L., **2006**. Estimating chlorophyll a concentrations from remote-sensing reflectance in optically shallow waters. *Remote Sensing of Environment* 101, 13–24. <https://doi.org/10.1016/j.rse.2005.12.002>
- Chang, N. Bin, Imen, S., Vannah, B., **2015**. Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: A 40-year perspective. *Critical Reviews in Environmental Science and Technology* 45, 101–166. <https://doi.org/10.1080/10643389.2013.829981>
- Chang, N.-B., Daranpob, A., Yang, Y.J., Jin, K.-R., **2009**. Comparative data mining analysis for information retrieval of MODIS images: monitoring lake turbidity changes at Lake Okeechobee, Florida. *Journal of Applied Remote Sensing* 3, 033549. <https://doi.org/10.1117/1.3244644>
- Chen, S. Sen, Chen, L.F., Liu, Q.H., Li, X., Tan, Q., **2005**. Remote sensing and GIS-based integrated analysis of coastal changes and their environmental impacts in Lingding Bay, Pearl River Estuary, *South China*. *Ocean and Coastal Management* 48, 65–83. <https://doi.org/10.1016/j.ocecoaman.2004.11.004>
- Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F.B.R., Aguzzi, J., Ballesteros, E., Bianchi, C.N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Frogli, C., Galil, B.S., Gasol, J.M., Gertwage, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., Kitsos, M.S., Koukouras, A., Lampadariou, N., Laxamana, E., de la Cuadra, C.M.L.F., Lotze, H.K., Martin, D., Mouillot, D., Oro, D., Raicevich,

- S., Rius-Barile, J., Saiz-Salinas, J.I., Vicente, C.S., Somot, S., Templado, J., Turon, X., Vafidis, D., Villanueva, R., Voultsiadou, E., **2010**. The biodiversity of the Mediterranean Sea: Estimates, patterns, and threats. *PLoS ONE* 5.
<https://doi.org/10.1371/journal.pone.0011842>
- Comby, E., Le Lay, Y.F., Piégay, H., **2014**. How chemical pollution becomes a social problem. Risk communication and assessment through regional newspapers during the management of PCB pollutions of the Rhône River (France). *Science of the Total Environment* 100–115, 100–115.
<https://doi.org/10.1016/j.scitotenv.2014.02.137>
- Cresson, R., Ienco, D., Gaetano, R., Ose, K., Tong Minh, D.H., **2019**. Optical image gap filling using deep convolutional autoencoder from optical and radar images. *IEEE International Geoscience and Remote Sensing Symposium* 218–221. <https://doi.org/10.1109/igarss.2019.8900353>
- de Paul Obade, V., Lal, R., Moore, R., **2014**. Assessing the Accuracy of Soil and Water Quality Characterization Using Remote Sensing. *Water Resources Management* 28, 5091–5109. <https://doi.org/10.1007/s11269-014-0796-7>
- de Paul Obade, V., Moore, R., **2018**. Synthesizing water quality indicators from standardized geospatial information to remedy water security challenges: A review. *Environment International* 119, 220–231.
<https://doi.org/10.1016/j.envint.2018.06.026>
- Dekker, A.G., Zamurović-Nenad, Hoogenboom, H.J., Peters, S.W.M., **1996**. Remote sensing, ecological water quality modelling and in situ measurements: a case study in shallow lakes. *Hydrological Sciences Journal* 41, 531–547.
<https://doi.org/10.1080/02626669609491524>
- Delile, H., Masson, M., Miège, C., Le Coz, J., Poulier, G., Le Bescond, C., Radakovitch, O., Coquery, M., **2020**. Hydro-climatic drivers of land-based organic and inorganic particulate micropollutant fluxes: The regime of the largest river water inflow of the Mediterranean Sea. *Water Research* 185.
<https://doi.org/10.1016/j.watres.2020.116067>
- Dorji, P., Fearn, P., **2017**. Impact of the spatial resolution of satellite remote sensing sensors in the quantification of total suspended sediment

- concentration: A case study in turbid waters of Northern Western Australia. *PLoS ONE* 12, 1–24. <https://doi.org/10.1371/journal.pone.0175042>
- Doxaran, D., Froidefond, J.-M., Lavender, S., Castaing, P., **2002**. Spectral signature of highly turbid waters. *Remote Sensing of Environment* 81, 149–161. [https://doi.org/10.1016/s0034-4257\(01\)00341-8](https://doi.org/10.1016/s0034-4257(01)00341-8)
- El-Said, G.F., **2013**. Bioaccumulation of Key Metals and Other Contaminants by Seaweeds from the Egyptian Mediterranean Sea Coast in Relation to Human Health Risk. *Human and Ecological Risk Assessment* 19, 1285–1305. <https://doi.org/10.1080/10807039.2012.708253>
- El Fellah, S., Rziza, M., El Haziti, M., **2017**. An efficient approach for filling gaps in Landsat 7 satellite images. *IEEE Geoscience and Remote Sensing Letters* 14, 62–66. <https://doi.org/10.1109/LGRS.2016.2626138>
- Fan, Y., Li, W., Gatebe, C.K., Jamet, C., Zibordi, G., Schroeder, T., Stamnes, K., **2017**. Atmospheric correction over coastal waters using multilayer neural networks. *Remote Sensing of Environment* 199, 218–240. <https://doi.org/10.1016/j.rse.2017.07.016>
- Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., **2014**. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15, 3133–3181. <https://doi.org/10.1117/1.JRS.11.015020>
- Frölicher, T.L., Laufkötter, C., **2018**. Emerging risks from marine heat waves. *Nature Communications* 9, 2015–2018. <https://doi.org/10.1038/s41467-018-03163-6>
- Gholizadeh, M.H., Melesse, A.M., Reddi, L., **2016**. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors* 16. <https://doi.org/10.3390/s16081298>
- Giorgi, F., **2006**. Climate change hot-spots. *Geophysical Research Letters* 33, 1–4. <https://doi.org/10.1029/2006GL025734>
- Hafeez, S., Wong, M., Ho, H., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K., Pun, L., **2019**. Comparison of Machine Learning Algorithms for Retrieval of Water

- Quality Indicators in Case-II Waters: A Case Study of Hong Kong. *Remote Sensing* 11, 617. <https://doi.org/10.3390/rs11060617>
- Hellweger, F.L., Schlosser, P., Lall, U., Weissel, J.K., **2004**. Use of satellite imagery for water quality studies in New York Harbor. *Estuarine, Coastal and Shelf Science* 61, 437–448. <https://doi.org/10.1016/j.ecss.2004.06.019>
- Holloway, J., Helmstedt, K.J., Mengersen, K., Schmidt, M., **2019**. A decision tree approach for spatially interpolating missing land cover data and classifying satellite images. *Remote Sensing* 11. <https://doi.org/10.3390/rs11151796>
- Huang, Y., Jiang, D., Zhuang, D., Fu, J., **2010**. Evaluation of hyperspectral indices for chlorophyll-a concentration estimation in Tangxun lake (Wuhan, China). *International Journal of Environmental Research and Public Health* 7, 2437–2451. <https://doi.org/10.3390/ijerph7062437>
- Jenni, A., **2019**. Water quality monitoring and assessment of the Northern Baltic Sea using Earth Observation. *Aalto University*, 84.
- Ke, L., Song, C., Ding, X., **2012**. Reconstructing complete MODIS LST based on temperature gradients in northeastern Qinghai-Tibet Plateau. *International Geoscience and Remote Sensing Symposium (IGARSS)* 3505–3508. <https://doi.org/10.1109/IGARSS.2012.6350664>
- Kim, G.U., Seo, K.H., Chen, D., **2019**. Climate change over the Mediterranean and current destruction of marine ecosystem. *Scientific Reports* 9, 1–9. <https://doi.org/10.1038/s41598-019-55303-7>
- Kirk, J., **1994**. Light and photosynthesis in aquatic ecosystems, Second Edi. ed. *Cambridge University Press*, 13.
- Kupssinskü, L.S., Guimarães, T.T., De Souza, E.M., Zanotta, D.C., Veronez, M.R., Gonzaga, L., Mauad, F.F., **2020**. A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. *Sensors* 20. <https://doi.org/10.3390/s20072125>
- Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., **2016**. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7, 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>

- Lau, J.D., Hicks, C.C., Gurney, G.G., Cinner, J.E., **2019**. What matters to whom and why? Understanding the importance of coastal ecosystem services in developing coastal communities. *Ecosystem Services* 35, 219–230.
<https://doi.org/10.1016/j.ecoser.2018.12.012>
- Law, R., Hanke, G., Angelidis, M., Batty, J., Bignert, A., Dachs, J., Davies, I., Denga, Y., Duffek, A., Herut, B., Hylland, K., Lepom, P., Leonards, P., Mehtonen, J., Piha, H., Roose, P., Tronczynski, J., Velikova, V., Vethaak, D., **2010**. Marine Strategy Framework Directive: Task Group 8 Report Contaminants and pollution effects. *Joint Research Centre, European Commission and International Council for the Exploration of the Sea*. <https://doi.org/10.2788/85887>
- Lejeusne, C., Chevaldonné, P., Pergent-Martini, C., Boudouresque, C.F., Pérez, T., **2010**. Climate change effects on a miniature ocean: the highly diverse, highly impacted Mediterranean Sea. *Trends in Ecology and Evolution* 25, 250–260. <https://doi.org/10.1016/j.tree.2009.10.009>
- Li, Y., Shi, Y., Zhu, X., Cao, H., Yu, T., **2014**. Coastal wetland loss and environmental change due to rapid urban expansion in Lianyungang, Jiangsu, China. *Regional Environmental Change* 14, 1175–1188.
<https://doi.org/10.1007/s10113-013-0552-1>
- Lillesand, T.M., Kiefer, R.W., Chipman, J.W., **2015**. Remote Sensing and Image Interpretation, 7th editio. ed, *Photogrammetric Engineering & Remote Sensing*. Wiley. <https://doi.org/10.14358/pers.81.8.615>
- Lins, R.C., Martinez, J.M., Marques, D. da M., Cirilo, J.A., Fragoso, C.R., **2017**. Assessment of chlorophyll-a remote sensing algorithms in a productive tropical estuarine-lagoon system. *Remote Sensing* 9, 1–19.
<https://doi.org/10.3390/rs9060516>
- Liu, Y., Islam, A., Gao, J., **2003**. Quantification of shallow water quality parameters by means of remote sensing. *Progress in Physical Geography* 27, 24–43.
<https://doi.org/10.1191/0309133303pp357ra>
- Long, C.M., Pavelsky, T.M., **2013**. Remote sensing of suspended sediment concentration and hydrologic connectivity in a complex wetland environment. *Remote Sensing of Environment* 129, 197–209.

<https://doi.org/10.1016/j.rse.2012.10.019>

Ludwig, W., Meybeck, M., **2003**. Riverine Transport of water, sediments and pollutants to the Mediterranean Sea, *UNEP/MAP/MED POL, MAP Technical Reports Series*, 50.

Ma, R., Dai, J., **2005**. Investigation of chlorophyll-a and total suspended matter concentrations using landsat ETM and field spectral measurement in Taihu Lake, China. *International Journal of Remote Sensing* 26, 2779–2795.
<https://doi.org/10.1080/01431160512331326648>

Ma, R., Tang, J., Dai, J., **2006**. Bio-optical model with optimal parameter suitable for Taihu Lake in water colour remote sensing. *International Journal of Remote Sensing* 27, 4305–4328.
<https://doi.org/10.1080/01431160600857428>

Mancinelli, G., Vizzini, S., **2015**. Assessing anthropogenic pressures on coastal marine ecosystems using stable CNS isotopes: State of the art, knowledge gaps, and community-scale perspectives. *Estuarine, Coastal and Shelf Science* 156, 195–204. <https://doi.org/10.1016/j.ecss.2014.11.030>

Marbà, N., Jordà, G., Agustí, S., Girard, C., Duarte, C.M., **2015**. Footprints of climate change on Mediterranean Sea biota. *Frontiers in Marine Science* 2, 1–11.
<https://doi.org/10.3389/fmars.2015.00056>

Mariethoz, G., Renard, P., **2010**. Reconstruction of Incomplete Data Sets or Images Using Direct Sampling. *Mathematical Geosciences* 42, 245–268.
<https://doi.org/10.1007/s11004-010-9270-0>

Mariethoz, G., Renard, P., Straubhaar, J., **2010**. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research* 46, 1–14. <https://doi.org/10.1029/2008WR007621>

Matthews, M.W., Bernard, S., Winter, K., **2010**. Remote sensing of cyanobacteria-dominant algal blooms and water quality parameters in Zeekoevlei, a small hypertrophic lake, using MERIS. *Remote Sensing of Environment* 114, 2070–2087. <https://doi.org/10.1016/j.rse.2010.04.013>

Meerschman, Eef, Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M.,

- Renard, P., **2013**. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. *Computers and Geosciences* 52, 307–324. <https://doi.org/10.1016/j.cageo.2012.09.019>
- Meerschman, E., Van Meirvenne, M., Van De Vijver, E., De Smedt, P., Islam, M.M., Saey, T., **2013**. Mapping complex soil patterns with multiple-point geostatistics. *European Journal of Soil Science* 64, 183–191. <https://doi.org/10.1111/ejss.12033>
- Mehvar, S., Filatova, T., Dastgheib, A., de Ruyter van Steveninck, E., Ranasinghe, R., **2018**. Quantifying economic value of coastal ecosystem services: A review. *Journal of Marine Science and Engineering* 6. <https://doi.org/10.3390/jmse6010005>
- Metz, M., Rocchini, D., Neteler, M., **2014**. Surface temperatures at the continental scale: Tracking changes with remote sensing at unprecedented detail. *Remote Sensing* 6, 3822–3840. <https://doi.org/10.3390/rs6053822>
- Micheli, F., Halpern, B.S., Walbridge, S., Ciriaco, S., Ferretti, F., Frascchetti, S., Lewison, R., Nykjaer, L., Rosenberg, A.A., **2013**. Cumulative human impacts on Mediterranean and Black Sea marine ecosystems: Assessing current pressures and opportunities. *PLoS ONE* 8. <https://doi.org/10.1371/journal.pone.0079889>
- Myint, S.W., Walker, N.D., **2002**. Quantification of surface suspended sediments along a river dominated coast with NOAA AVHRR and Sea WiFS measurements: Louisiana, USA. *International Journal of Remote Sensing* 23, 3229–3249. <https://doi.org/10.1080/01431160110104700>
- Nechad, B., Ruddick, K.G., Park, Y., **2010**. Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters. *Remote Sensing of Environment* 114, 854–866. <https://doi.org/10.1016/j.rse.2009.11.022>
- Nelson, N.B., Siegel, D.A., Carlson, C.A., Swan, C., Smethie, W.M., Khatiwala, S., **2007**. Hydrography of chromophoric dissolved organic matter in the North Atlantic. *Deep-Sea Research Part I: Oceanographic Research Papers* 54, 710–731. <https://doi.org/10.1016/j.dsr.2007.02.006>

- Neteler, M., **2010**. Estimating daily land surface temperatures in mountainous environments by reconstructed MODIS LST data. *Remote Sensing* 2, 333–351. <https://doi.org/10.3390/rs1020333>
- Neumann, B., Ott, K., Kenchington, R., **2017**. Strong sustainability in coastal areas: a conceptual interpretation of SDG 14. *Sustainability Science* 12, 1019–1035. <https://doi.org/10.1007/s11625-017-0472-y>
- Oelsner, G.P., Stets, E.G., **2019**. Recent trends in nutrient and sediment loading to coastal areas of the conterminous U.S.: Insights and global context. *Science of the Total Environment* 654, 1225–1240. <https://doi.org/10.1016/j.scitotenv.2018.10.437>
- Olden, J.D., Lawler, J.J., Poff, N.L., **2016**. Machine Learning Methods Without Tears : A Primer for Ecologists. *The Quarterly Review of Biology* 83, 171–193.
- Olmanson, L.G., Bauer, M.E., Brezonik, P.L., **2008**. A 20-year Landsat water clarity census of Minnesota's 10,000 lakes. *Remote Sensing of Environment* 112, 4086–4097. <https://doi.org/10.1016/j.rse.2007.12.013>
- Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P., Dubash, N.K., Edenhofer, O., Elgizouli, I., Field, C.B., Forster, P., Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., Hegerl, G., Howden, M., Jiang, K., Jimenez Cisneroz, B., Kattsov, V., Lee, H., Mach, K.J., Marotzke, J., Mastrandrea, M.D., Meyer, L., Minx, J., Mulugetta, Y., O'Brien, K., Oppenheimer, M., Pereira, J.J., Pichs-Madruga, R., Plattner, G.K., Pörtner, H.O., Power, S.B., Preston, B., Ravindranath, N.H., Reisinger, A., Riahi, K., Rusticucci, M., Scholes, R., Seyboth, K., Sokona, Y., Stavins, R., Stocker, T.F., Tschakert, P., van Vuuren, D., van Ypserle, J.P., **2014**. Climate Change **2014**: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC*, Geneva, Switzerland.
- Paerl, H.W., **2017**. Controlling cyanobacterial harmful blooms in freshwater ecosystems. *Microbial Biotechnology* 10, 1106–1110. <https://doi.org/10.1111/1751-7915.12725>
- Pede, T., Mountrakis, G., **2018**. An empirical comparison of interpolation

- methods for MODIS 8-day land surface temperature composites across the conterminous United States. *ISPRS Journal of Photogrammetry and Remote Sensing* 142, 137–150. <https://doi.org/10.1016/j.isprsjprs.2018.06.003>
- Pereira, L.S.F., Andes, L.C., Cox, A.L., Ghulam, A., **2018**. Measuring Suspended-Sediment Concentration and Turbidity in the Middle Mississippi and Lower Missouri Rivers Using Landsat Data. *Journal of the American Water Resources Association* 54, 440–450. <https://doi.org/10.1111/1752-1688.12616>
- Preisendorfer, R.W., **1986**. Secchi disk science: Visual optics of natural waters. *Limnology and Oceanography* 31, 909–926. <https://doi.org/10.4319/lo.1986.31.5.0909>
- Prévost, A., Robert, S., **2016**. Local spatial planning practices in four French Mediterranean coastal territories under pressure. *Land Use Policy* 56, 68–80. <https://doi.org/10.1016/j.landusepol.2016.04.034>
- Ramadas, M., Samantaray, A.K., **2018**. Applications of Remote Sensing and GIS in Water Quality Monitoring and Remediation: A State-of-the-Art Review, in: *Water Remediation, Energy, Environment, and Sustainability*. pp. 225–246. https://doi.org/10.1007/978-981-10-7551-3_13
- Rao, N.S., Ghermandi, A., Portela, R., Wang, X., **2015**. Global values of coastal ecosystem services: A spatial economic analysis of shoreline protection values. *Ecosystem Services* 11, 95–105. <https://doi.org/10.1016/j.ecoser.2014.11.011>
- Regional Environmental Center, **2008**. Section 5 Water Protection Legislation, in: *Handbook on the Implementation of EC Environmental Legislation*. pp. 613–733.
- Ritchie, J.C., Schiebe, F.R., McHenry, J.R., **1976**. Remote sensing of suspended sediments in surface waters. *Photogrammetric Engineering and Remote Sensing* 42, 1539–1545.
- Robert, S., Fox, D., Boulay, G., Grandclément, A., Garrido, M., Pasqualini, V., Prévost, A., Schleyer-Lindenmann, A., Trémélo, M.L., **2019**. A framework to analyse urban sprawl in the French Mediterranean coastal zone. *Regional*

- Environmental Change* 19, 559–572. <https://doi.org/10.1007/s10113-018-1425-4>
- Ryan, W.B.F., Carbotte, S.M., Coplan, J.O., O'Hara, S., Melkonian, A., Arko, R., Weissel, R.A., Ferrini, V., Goodwillie, A., Nitsche, F., Bonczkowski, J., Zemsky, R., **2009**. Global multi-resolution topography synthesis. *Geochemistry, Geophysics, Geosystems* 10. <https://doi.org/10.1029/2008GC002332>
- Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S., Adams, C., **2020**. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing, *Earth-Science Reviews*. <https://doi.org/10.1016/j.earscirev.2020.103187>
- Scharlemann, J.P.W., Benz, D., Hay, S.I., Purse, B. V., Tatem, A.J., Wint, G.R.W., Rogers, D.J., **2008**. Global data for ecology and epidemiology: A novel algorithm for temporal fourier processing MODIS data. *PLoS ONE* 3. <https://doi.org/10.1371/journal.pone.0001408>
- Ting, Y., **2005**. Utility of remote sensing data in retrieval of water quality constituents concentrations in coastal water of New Jersey. *New Jersey Institute of Technology*, 14.
- Usali, N., Ismail, M.H., **2010**. Use of Remote Sensing and GIS in Monitoring Water Quality. *Journal of Sustainable Development* 3, 228–238. <https://doi.org/10.5539/jsd.v3n3p228>
- Wang, G., Garcia, D., Liu, Y., de Jeu, R., Johannes Dolman, A., **2012**. A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations. *Environmental Modelling and Software* 30, 139–142. <https://doi.org/10.1016/j.envsoft.2011.10.015>
- Wang, X., Yang, W., **2019**. Water quality monitoring and evaluation using remote-sensing techniques in China: A systematic review. *Ecosystem Health and Sustainability* 5, 47–56. <https://doi.org/10.1080/20964129.2019.1571443>
- Xu, Y., Shen, Y., **2013**. Reconstruction of the land surface temperature time series using harmonic analysis. *Computers and Geosciences* 61, 126–132. <https://doi.org/10.1016/j.cageo.2013.08.009>

- Yin, G., Mariethoz, G., McCabe, M.F., **2017**. Gap-filling of landsat 7 imagery using the direct sampling method. *Remote Sensing* 9, 1–20.
<https://doi.org/10.3390/rs9010012>
- Yu, W., Nan, Z., Wang, Z., Chen, H., Wu, T., Zhao, L., **2015**. An Effective Interpolation Method for MODIS Land Surface Temperature on the Qinghai-Tibet Plateau. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, 4539–4550.
<https://doi.org/10.1109/JSTARS.2015.2464094>
- Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., **2018**. Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing* 56, 4274–4288. <https://doi.org/10.1109/TGRS.2018.2810208>

Appendices

Appendix A

Statistical indices of gap filled images

	Cloud%	Band	Mean	SD	Skew
Original Image	No Clouds	B02	308.30	109.23	0.90
		B03	186.25	125.49	1.16
		B04	24.99	30.82	3.89
Gap Filled Image (UL)	11.57 %	B02	306.33	107.86	0.94
		B03	183.36	123.79	1.23
		B04	23.89	29.60	4.20
	23.35 %	B02	307.47	109.30	0.88
		B03	185.35	124.73	1.14
		B04	24.63	29.45	1.57
	38.89 %	B02	308.20	108.35	0.87
		B03	184.79	122.42	1.09
		B04	24.39	29.03	2.57
Gap Filled Image (BL)	11.57 %	B02	306.62	108.11	0.94
		B03	183.33	123.62	1.22
		B04	23.48	29.31	4.32
	23.35 %	B02	307.60	109.28	0.88
		B03	185.37	124.83	1.14
		B04	24.63	29.41	1.57
	38.89 %	B02	307.63	108.24	0.87
		B03	184.91	123.17	1.10
		B04	23.91	28.80	2.65
	57.30 %	B02	305.22	92.03	0.99
		B03	174.87	97.57	1.39
		B04	20.16	24.23	6.47
	79.30 %	B02	308.20	102.09	0.76
		B03	184.92	114.73	1.04
		B04	22.69	25.97	1.46

B02: Blue band; B03: Green band; and B04: Red band.

Appendix B

Error metrics values for gap filled images

Algorithm	Cloud %	Band	RMSLE	MdAPE	MAE	RMSE	R2
UL	11.57%	B02	0.06	5.38	31.45	45.20	77.75
		B03	0.11	9.16	39.31	60.75	68.69
		B04	0.42	20.66	13.71	21.27	49.34
	23.35%	B02	0.03	3.83	16.11	27.64	93.26
		B03	0.06	6.39	16.97	34.36	92.30
		B04	0.36	21.33	6.59	19.56	68.85
	38.89%	B02	0.05	5.09	24.38	36.78	86.40
		B03	0.09	9.83	29.93	49.95	83.73
		B04	0.42	36.46	8.98	18.09	70.71
BL	11.57%	B02	0.06	5.20	31.27	45.34	77.61
		B03	0.11	9.32	39.52	60.59	68.84
		B04	0.45	23.16	14.89	22.70	42.32
	23.35%	B02	0.03	3.88	16.12	27.41	93.37
		B03	0.06	6.35	17.12	34.84	92.08
		B04	0.37	21.96	6.67	19.64	68.60
	38.89%	B02	0.05	5.14	24.80	37.54	85.83
		B03	0.09	9.79	28.60	47.56	85.25
		B04	0.42	34.32	9.27	18.91	68.00
	57.30%	B02	0.07	10.52	42.53	58.31	73.23
		B03	0.16	23.48	53.85	80.12	65.25
		B04	0.52	53.87	13.63	23.34	48.73
	79.30%	B02	0.06	5.49	27.87	46.87	80.87
		B03	0.11	10.34	31.61	57.69	78.28
		B04	0.42	34.28	9.19	19.29	59.87

B02: Blue band; B03: Green band; and B04: Red band.

Appendix C

Processing time needed for the GF algorithm to find best replicates for all unknown pixels in artificially clouded images

Algorithm	Cloud%	Band	Time (min)
UL	11.57%	B02	214.00
		B03	177.40
		B04	191.00
	23.35%	B02	293.20
		B03	268.80
		B04	580.60
	38.89%	B02	318.80
		B03	270.80
		B04	283.00
BL	11.57%	B02	150.20
		B03	138.20
		B04	259.40
	23.35%	B02	274.40
		B03	305.80
		B04	864.60
	38.89%	B02	1649.80
		B03	2140.40
		B04	8082.00
	57.30%	B02	1508.60
		B03	1966.20
		B04	11615.80
	79.30%	B02	399.00
		B03	451.20
		B04	1780.60

B02: Blue band; B03: Green band; and B04: Red band; Time represents the time spent to finish the search for replicate values and not other tasks.

Appendix D

**Table of cloud percentages corresponding to Sentinel-2 images acquired
within the 10-days period from January 2019 to March 2020**

Number	Date of acquired S2 image	Cloud Percentage
1	2019-01-23	99.7
2	2019-01-25	0
3	2019-01-28	99.13
4	2019-01-30	0.73
5	2019-02-12	8.54
6	2019-02-24	10.13
7	2019-02-17	24.01
8	2019-02-19	87.67
9	2019-03-11	0
10	2019-03-14	23.35
11	2019-03-16	0
12	2019-03-19	19.81
13	2019-04-13	1.7
14	2019-04-15	37.73
15	2019-04-18	98.9
16	2019-04-20	52.39
17	2019-05-13	0.02
18	2019-05-15	0.08
19	2019-05-18	64.07
20	2019-05-20	83.12
21	2019-06-12	0.01
22	2019-06-14	46.75
23	2019-06-17	0.02
24	2019-06-19	38.89
25	2019-07-12	2.88
26	2019-07-14	1.33
27	2019-07-17	43.6
28	2019-07-19	0
29	2019-08-11	14.23
30	2019-08-13	97.63
31	2019-08-16	4.02
32	2019-08-18	7.44
33	2019-09-12	0
34	2019-09-15	38.82
35	2019-09-17	5.52
36	2019-09-20	0.33
37	2019-10-12	99.84
38	2019-10-15	0.01
39	2019-10-17	11.57

40	2019-10-20	85.67
41	2019-11-11	76.64
42	2019-11-14	100
43	2019-11-16	13.69
44	2019-11-19	79.42
45	2019-12-01	99
46	2019-12-04	100
47	2019-12-06	0
48	2019-12-09	49.45
49	2020-01-03	99.05
50	2020-01-05	0.01
51	2020-01-08	11.46
52	2020-01-10	58.93
53	2020-02-12	10.2
54	2020-02-14	0.49
55	2020-02-17	99.77
56	2020-02-19	57.3
57	2020-03-13	24.25
58	2020-03-15	0
59	2020-03-18	0
60	2020-03-20	95.59

Appendix E

Results from testing the adopted land mask before and after applying a buffer

Model	Buffer	Band	RMSLE	MdAPE	MAE	RMSE	R2	Gain in Time (%)
UV	No	B02	0.11	6.9	73.5	185.2	23.56	-
		B03	0.17	11.9	112.3	273.1	12.64	-
		B04	0.70	31.8	130.6	392.9	-29.52	-
	Yes	B02	0.05	5.2	31.7	46.1	76.87	33.64
		B03	0.11	9.5	39.8	61.27	68.17	27.17
		B04	0.37	21.4	13.7	20.97	50.81	48.95
BV	No	B02	0.11	6.6	72.0	182.2	26.04	-
		B03	0.16	11.2	104.0	256.8	22.79	-
		B04	0.63	42.8	126.6	382.9	-23.06	-
	Yes	B02	0.05	4.8	29.4	43.1	79.74	51.11
		B03	0.105	8.93	37.5	58.2	71.26	44.90
		B04	0.525	20.9	14.5	22.6	42.68	27.10