UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Ocean surface visibility prediction

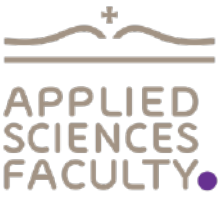
Author:
Volodymyr PRYPESHNIUK

Supervisor:
Dmytro KARAMSHUK

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences



Lviv 2023

Declaration of Authorship

I, Volodymyr PRYPESHNIUK, declare that this thesis titled, "Ocean surface visibility prediction" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Yesterday is history, tomorrow is a mystery, but today is a gift, that is why it is called present."

Master Oogway

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Ocean surface visibility prediction

by Volodymyr PRYPESHNIUK

Abstract

Seawater transparency is an indispensable ecological parameter with substantial impacts on the health and productivity of aquatic ecosystems. Its significance spans across various industries, including environment protection, fishing and tourism. The fluctuating nature of aquatic systems and their intricate interplay with human activities often induce substantial variability in seawater transparency. This underlines the pressing necessity for effective predictive tools in the stewardship and preservation of our invaluable water resources.

Despite the clear importance of water transparency, ocean forecasting remains a considerably understudied field, some work has been done on using satellite for monitoring, but literature is scarce for forecasting with only few simple models explored. There is an evident gap in research and tools focused on predicting changes in this crucial ecosystem, underlining the novelty and urgency of our work.

In this research, we aim is to develop a forecasting model that not only excels in precision and speed, but is also flexible enough to encompass a vast array of potential future scenarios. We primarily employed SimVP, a spatio-temporal convolutional neural network, for ocean forecasting purposes. This model was trained using the earth observation data from the Copernicus Marine Service. This data were collected for 20 years of daily observation of water transparency in the marine environment surrounding the UK, with a spatial resolution of 4km x 4km.

Our findings showed that SimVP substantially outperformed the baseline models (AutoRegressive Integrated Moving Average (ARIMA) and Simple Exponential Smoothing (SES)) in predicting the next day seawater transparency, demonstrating an improvement of 17.4%, and a notable reduction in the Root Mean Square Error (RMSE) from 2.63 to 2.24, and improvement in inference time efficiency in 66.3 times (334.6 -> 5.04 seconds). We show that this method better performs better on regions with minor variation like Irish Sea or English Channel, and performs worse on regions with high variations like Atlantic Ocean or North Sea.

Our study demonstrates the advantage of adopting the spatio-temporal neural network architectures for ocean monitoring and paves the way for future research in adopting advanced machine learning techniques in this field.

Contents

List of Figures

List of Tables

List of Abbreviations

RMSE	**R**oot **M**ean **S**quare **E**rror
RRMSE	**R**elative **R**oot **M**ean **S**quare **E**rror
SES	**S**imple **E**exponential **S**smoothing
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **Average**
CPU	**C**entral **P**rocessing **U**nit
GPU	**G**raphics **P**rocessing **U**nit

List of Symbols

t	time
f	some function
ϕ	model parameters
$\hat{Y}$	forecasted value
ϵ	error term
x	historical observation

Chapter 1

Introduction

1.1 Motivation

The omnipresence and vitality of oceans to life on earth cannot be overstated. Oceans serve as the planet's largest habitat, are central to the climate system, and provide resources such as food and energy that society depends upon (NOAA, 2023). In that regard, ocean surface visibility, often measured through the proxy of Secchi depth, plays a significant role in oceanographic studies as it gives insights into several biogeochemical properties of the water column, as well as the health and productivity of marine ecosystems. Consequently, the ability to predict ocean surface visibility is crucial for numerous reasons that span both ecological and societal realms.

From an ecological perspective, the visibility of the ocean surface directly affects the photosynthesis process, thus regulating primary productivity. Primary productivity serves as the baseline for the marine food web, impacting the availability of resources for higher trophic levels. Platt et al., 2017 Predicting visibility allows for forecasting of marine productivity changes, providing valuable insights for fishery management, conservation strategies, and the evaluation of potential impacts of environmental changes.

Furthermore, water transparency, has significant implications for the propagation of light and heat in the water column. Increased turbidity, leading to lower visibility, can alter the thermal stratification of water bodies, thus potentially influencing patterns of oceanic circulation. Such changes can in turn impact the distribution and behavior of marine species Li, 2020. Consequently, a reliable prediction model for water transparency can offer insights into these potential changes, aiding in biodiversity preservation and the understanding of ecosystem shifts.

From a societal perspective, better prediction of water transparency is a step forward in improving water quality monitoring, recreational planning, and navigation safety. For example, coastal regions with high recreational value often depend on water clarity for activities such as swimming, boating, and scuba diving. Meanwhile, in sectors like shipping and offshore wind farms, improved visibility prediction can significantly enhance operational safety and efficiency. Therefore, advances in forecasting water transparency can directly benefit economic sectors and societal activities that rely on the oceans.

The need for precise prediction models is heightened in the face of climate change. With oceans experiencing shifts in temperature, acidity, and nutrient loads due to anthropogenic activity, the clarity of ocean waters is anticipated to change, potentially disrupting the balance of marine ecosystems (Krajick, 2022). Therefore, a robust model that can predict changes in ocean surface visibility can provide an early warning system, aiding adaptive management efforts to mitigate potential impacts.

The body of research in this area, however, is sparse and mostly limited to linear models or simple neural networks (Yipeng Liao, 2021, Heddam, 2016). Although

these models have provided preliminary insights, they lack the ability to capture complex spatio-temporal patterns and non-linear interactions between influencing factors. By using modern Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), it is possible to move towards a more accurate and comprehensive model of water transparency forecasting. This advancement will enable researchers, policymakers, and industry professionals to make more informed decisions, fostering the sustainable use and conservation of our invaluable marine ecosystems.

1.2 Goals of the master thesis

Our main goal and focus of this thesis is to train a spatio temporal model, possessing the capacity to effectively predict the water transparency for the subsequent day, utilizing an extensive historical dataset. The specific objectives to accomplish this endeavor have been delineated as follows:

- **Train suitable Spatio-Temporal Neural Network:** The prime focus is to research and train a sophisticated machine learning algorithm, specifically a neural network, capable of exploiting both spatial and temporal features for water transpareceny prediction purposes.
- **Constructing a Suitable Dataset:** It is crucial to curate an appropriate dataset, amenable to training the neural network. This dataset must encapsulate a multitude of relevant features and observations that are conducive to the predictive capabilities of the model.
- **Conducting a Thorough Model Evaluation:** An in-depth evaluation of the model must be performed to discern its strengths and limitations. The goal is to create a forecasting model that's not only accurate and fast, but also versatile enough to handle a wide range of possible future scenarios.

1.3 Structure of the Thesis

This thesis is organized into six chapters, each dedicated to a distinct aspect of our research.

Chapter 2 provides an extensive review of the literature in the field of our research. This includes works related to satellite sensing, the process of forecasting, and the measurement of Secchi depth. It offers a comprehensive analysis of prior studies and methodologies, positioning our work within this larger academic discourse.

In **Chapter 3**, we delve into the specifics of the data employed in our study. This chapter includes detailed descriptions of the different features and geographic characteristics of the data. It provides a solid foundation for the understanding of our problem space and the subsequent technical formulations.

Chapter 4 is where we present the core technical problem that we aim to solve in our research. We provide a rigorous mathematical formulation of the problem and propose the method we intend to use to tackle this challenge.

Following the problem formulation, **Chapter 5** showcases the results of our experiments and provides an in-depth evaluation of the models we have trained. This includes a thorough analysis of their performance, potential limitations, and the insights they offer into our research question.

Finally, in **Chapter 6**, we offer a comprehensive summary of the work carried out in this thesis. We reflect on the findings of our study, their implications, and potential avenues for further research. This chapter not only concludes our present work but also outlines prospective directions for future improvements and investigations based on our research.

Chapter 2

Related works

2.1 Satellite sensing

Satellite remote sensing is the technique of obtaining information about the Earth's atmosphere, surface, and oceans from sensors onboard satellites. These sensors capture data in the form of images and provides measurements for analysis. Satellite remote sensing can be passive or active. Passive sensors detect natural radiation that is emitted or reflected by the object or surrounding areas. Active sensors, on the other hand, emit energy in order to scan objects and areas whereupon a sensor then detects and measures the radiation that is reflected or backscattered from the target. (Loredana Tecar, 2014)

Satellite remote sensing is impacting various aspects of human life, which includes:

Monitoring and Understanding the Environment: Satellite remote sensing is critical for studying various aspects of the Earth's environment (Parra, 2019), including weather and climate, the health of ecosystems, the status of agricultural crops, and the quality of bodies of water (Godson Ebenezer Adjovu, 2023). By observing these elements from space, scientists can track changes over time and assess the impact of natural events and human activities.

Disaster Management: Satellite remote sensing provides a critical tool for predicting, monitoring, and responding to natural disasters. For example, remote sensing data can be used to track the formation and path of hurricanes, allowing for early warnings and evacuation orders. After a disaster, such as a flood, wildfire or military conflicts, remote sensing data can be used to assess the damage and coordinate recovery efforts (Serhii Shevchuk, 2022). Additionally, monitoring earthquake-prone regions can provide data about ground deformation patterns, contributing to earthquake preparedness. (Wen Liu, 2019)

Resource Management: Remote sensing is used in the management of various natural resources. For instance, in forestry, remote sensing can be used to assess forest health (Angela Lausch, 2016), track deforestation, and plan sustainable logging. In mineral exploration, certain spectral characteristics can indicate the presence of valuable minerals (Sabins, 1999). In water management, remote sensing can track changes in river courses, monitor reservoir levels, and assess the health of watersheds.

Infrastructure and Urban Planning: Remote sensing can provide detailed, up-to-date images of cities and infrastructure. (Thilo Wellmann, 2020) These images can be used in urban planning to analyze land use patterns, assess the impact of new development, and monitor changes over time. Remote sensing can also be used to monitor large infrastructure projects, such as the construction of highways, dams, or airports, providing valuable information for project management.

Climate Change Studies: Satellite remote sensing plays a crucial role in studying climate change. (Jun Yang, 2013) For example, satellites can track changes in sea ice extent and thickness, providing critical data about the impacts of global warming. Satellites can also monitor global vegetation patterns, helping scientists understand the impacts of climate change on ecosystems. Data on ocean temperatures, atmospheric CO2 levels, and sea-level rise are also gathered through remote sensing, contributing to our understanding of climate change dynamics.

Defense and Intelligence: Satellite imagery is commonly used in defense and intelligence for a variety of purposes. (Ricky J. Lee, 2014) These can include identifying and tracking military activities, such as troop movements or construction at military sites. Remote sensing can also be used for border surveillance, contributing to border security efforts. Additionally, satellite imagery can support tactical mission planning by providing detailed terrain information.

Navigation and Communication: Global navigation systems like GPS rely on satellites to provide accurate location information. Similarly, communication satellites are essential for transmitting signals for television, telephone, and internet services, especially in remote areas. In both cases, the ability to monitor the status of these satellites and understand the space environment is critical. (Pelton, 2017)

I our case, water transparency forecasting would fall under the category of Monitoring and Understanding the Environment and Resource Management.

In the context of environmental monitoring, water transparency is a crucial indicator of water quality (Timo Toivanen, 2013) and can be influenced by factors such as algal blooms, suspended solids, and dissolved organic matter. Remote sensing can be used to monitor these factors over time and across large areas. This information can be crucial for understanding the health of aquatic ecosystems and the processes affecting them.

From a resource management perspective, water transparency can affect a range of human activities. For example, it can influence recreational uses of water bodies, such as swimming and boating. In aquaculture, water transparency can affect the health and growth of cultured species. Therefore, forecasting changes in water transparency can inform the management of these activities and help to mitigate potential impacts.

In both of these contexts, satellite remote sensing provides a valuable tool for assessing and predicting water transparency at large scales and over extended periods. It allows for continuous, objective, and consistent observations that can greatly enhance our understanding and management of water resources.

2.2 Forecasting

Forecasting is a critical discipline within a multitude of scientific fields, with applications ranging from weather prediction and environmental science to economics, epidemiology, and beyond. It is the science of making predictions about future outcomes based on historical data and statistical analysis. This discipline capitalizes on a myriad of methodologies, some of which include time series analysis, regression models, machine learning algorithms, and more recently, deep learning techniques.

The crux of forecasting lies in identifying patterns within historical data, quantifying uncertainties, and extrapolating these patterns into the future. The scientific value and societal impact of accurate forecasting are immense. It aids in decision-making, policy formulation, risk management, and strategic planning across various sectors, notably in finance, healthcare, in our case climate science.

One of the most notable advancements within this field in the recent decade has been the increased adoption of machine learning and artificial intelligence techniques. Traditional statistical methods, such as autoregressive integrated moving average (ARIMA) and exponential smoothing, while still valuable and widely used, sometimes fall short when dealing with large, complex datasets. Machine learning models, such as random forests, support vector machines, and neural networks, have demonstrated their potential to handle high-dimensional data, recognize complex patterns, and enhance predictive accuracy.

Especially notable is the emergence of deep learning algorithms in forecasting. These neural network-based models, such as long short-term memory (LSTM) units (Sepp Hochreiter, 1997) and convolutional neural networks (CNNs) (Yann LeCun, 2015), have shown exceptional promise in processing time-series data. (Aji Prasetya Wibawa, 2022). Their ability to capture temporal and spatial dependencies and handle non-linear relationships between variables can lead to superior forecast results, often outperforming traditional methods.

However, as with any discipline, forecasting is not without challenges. It faces the perennial issue of balancing model complexity and interpretability. While complex machine learning and deep learning models often yield more accurate predictions, they tend to be "black boxes," making their inner workings difficult to interpret. This lack of transparency can be a significant concern in fields where understanding the underlying causal mechanisms is as crucial as prediction accuracy itself.

Another challenge is the inherent uncertainty and volatility of some systems under study. Forecasting models assume that the future will behave similarly to the past, an assumption that might not always hold. Changes in trends, abrupt shocks, and unpredictable events pose significant challenges to accurate forecasting.

Data quality and availability are further challenges. Models are only as good as the data they are trained on. Biased, incomplete, or noisy data can lead to poor forecasts. Despite these challenges, the field of forecasting continues to innovate, with emerging techniques aiming to address these issues and improve accuracy and reliability.

In summary, forecasting is a dynamic and evolving field with substantial real-world significance. Its progression has been marked by the integration of sophisticated computational and statistical methods that have continually improved our ability to anticipate future outcomes. Despite the challenges posed by data quality, model transparency, and inherent system volatility, forecasting remains a crucial discipline, pushing the boundaries of scientific inquiry and technological innovation. The continuous advancement in this field promises an exciting future where our predictive capabilities can further contribute to societal and scientific progress.

In the context of our study, conventional forecasting algorithms would require a significant computational investment when training on each spatial point. Fortuitously, the domain of forecasting has experienced the emergence of a unique subfield, commonly known as *spatial forecasting*. This approach is particularly pertinent to the challenges presented in our research.

2.3 Spatial forecasting

Spatiotemporal forecasting is a complex and increasingly significant subset of the broader field of forecasting, which aims to predict future outcomes considering both

spatial and temporal dimensions. This field is especially relevant in areas such as climatology, epidemiology, traffic management, and energy consumption, where data exhibit dependencies across both time and space. The spatiotemporal aspect brings an additional layer of complexity to forecasting models as they must account not only for how variables change over time but also how they change across different geographical locations.

Techniques for handling spatiotemporal data have developed rapidly in recent years, leveraging advanced machine learning and statistical methodologies. For example, convolutional long short-term memory neural networks (ConvLSTMs) have been employed to capture the spatial dynamics of a region effectively. (Xingjian Shi, 2015). ConvLSTMs are a type of neural network that merges the spatial feature extraction capabilities of convolutional neural networks (CNNs) with the temporal dynamic learning capabilities of long short-term memory neural networks (LSTMs), making them ideal for handling spatiotemporal data. (Espeholt, 2022)

Another promising technique is the use of Graph Neural Networks (GNNs) which represent the spatial data as a graph and operate directly on it. (Bing Yu, 2017) This approach can capture complex spatial dependencies and, when combined with temporal models such as recurrent neural networks or transformer-based models, can effectively handle spatiotemporal data.

However, similar to traditional forecasting, spatiotemporal forecasting also faces significant challenges, such as computational demand, data scarcity or quality in certain regions, and the need to correctly model complex spatial and temporal dependencies. Despite these challenges, the growing importance of spatiotemporal forecasting in our increasingly interconnected world, along with advancements in computational capacities and methodologies, paints a promising picture for the future of this field.

I this study we will use SimVP (Cheng Tan, 2022) convolution based method, as our spatiotemporal model for forecasting water transparency. This method allows relatively easy training with *Mean Square Error* loss and fast inference time.

2.4 Secchi depth

The method of determining water clarity by using a simple tool known as the Secchi disk was devised 150 years ago by Angelo Secchi, an Italian priest, and has been used by oceanographers ever since. The Secchi depth, which represents the point at which the Secchi disk is no longer visible to the human eye, provides a reliable and robust technique for assessing water quality. (Wernand, 2010) Despite the advent of more advanced technologies, the Secchi disk remains a valuable tool due to its simplicity, cost-effectiveness, and the vast amount of data it has accumulated over the years, enabling researchers to evaluate changes in aquatic environments over a long period.

Historically, a significant discrepancy existed between the measurements of Secchi depths and its theoretical relationship, a challenge that has been addressed by ZhongPing Lee, 2015. They identified inaccuracies in the traditional assumptions about the detection of Secchi disks by human eyes and the mechanisms our eyes use to distinguish an object from its surroundings. The team developed a new theoretical model based on radiative transfer, thereby resolving the long-standing inconsistency between field measurements and theoretical predictions. Their model was validated

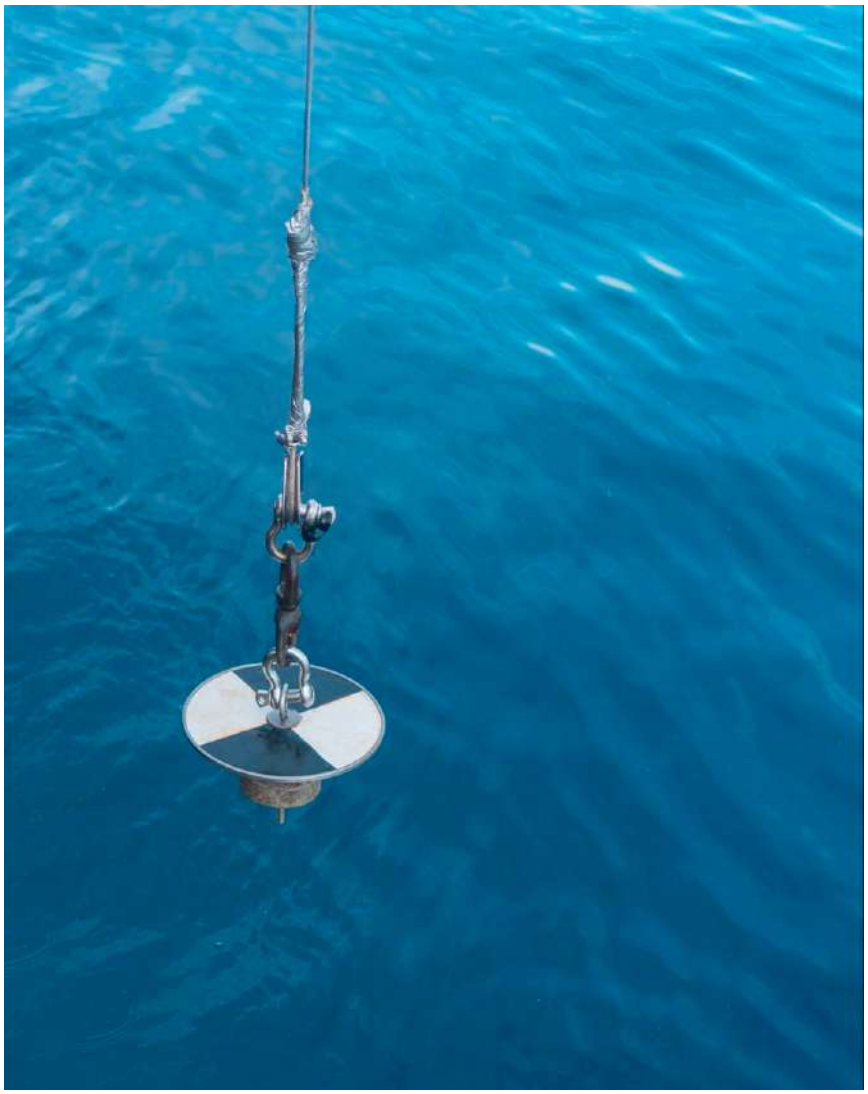


FIGURE 2.1: Modern Secchi disk for in situ measurements

using published data of Secchi depth and diffuse attenuation coefficient over a nine-decade timespan, covering a range of water bodies including oceans, coastal waters, and inland lakes.

Recently, researchers continue to advance our understanding of the Secchi depth and its application in monitoring water quality. A 2023 study by Robert J. W. Brewin, 2023 conducted on four Atlantic Meridional Transect cruises demonstrated how traditional Secchi depth measurements, alongside modern radiometric measurements of ocean clarity and color, can be used to monitor chlorophyll-a concentration and evaluate remote-sensing algorithms. This study provided several key findings that enhance our understanding of Secchi depth and its relationship with other optical and environmental variables:

1. **Correlation with Optical Variables:** The study found that the Secchi depth is closely correlated with several optical variables, including the Forel-Ule colour (a scale used to measure the color of bodies of water) and beam and diffuse attenuation (measures of how light is absorbed or scattered in the water). Specifically, the Secchi depth was inversely related to Forel-Ule colour and to beam and diffuse attenuation, and positively related to the ratio of blue to green

remote-sensing reflectance and euphotic depth (the depth to which sufficient light exists for photosynthesis to occur). These relationships validate the use of Secchi depth as a robust measure of water clarity.

2. **Correlation with Chlorophyll-a Concentration:** The study also found a tight correlation between Secchi depth and chlorophyll-a concentration, a measure often used to estimate the amount of phytoplankton (microscopic plants) in the water. This correlation ranged from 71-81

3. **Performance of Remote-Sensing Algorithms:** The research evaluated existing algorithms that predict chlorophyll-a from these variables, and found them to perform well, albeit with some systematic differences. Moreover, remote sensing algorithms of Secchi depth were in good agreement with in-situ data over the range of values collected, but with a slight positive bias.

4. **Impact of Environmental Conditions:** Importantly, the study also found that wind speed can impact the estimation of Secchi depth. The researchers suggested a path forward to include the effect of wind in the current Secchi depth theory, which could improve the accuracy of this measurement under different environmental conditions.

Their study provided a comprehensive analysis of the relationships between Secchi depth, various optical variables, chlorophyll-a concentration, and environmental conditions. These findings not only validate the use of Secchi depth as a robust measure of water clarity but also suggest ways to refine its application under varying environmental conditions (Robert J. W. Brewin, 2023.)

However, the use of Secchi depth as a measure of water clarity has not been without challenges. For many decades, researchers struggled to reconcile the discrepancy between measurements of Secchi depth and its theoretical relationship to an optical property named the 'diffuse attenuation coefficient'. Recent research by ZhongPing Lee, 2015 and his colleagues has addressed these challenges by developing a new theory based on radiative transfer, which provides a more accurate model for predicting Secchi depth in different waters.

Furthermore, modern research has found correlations between Secchi depth and other environmental variables such as chlorophyll-a concentration and wind speed. These findings underscore the value of Secchi depth as a robust measure of water clarity and suggest ways to refine its application under varying environmental conditions.

Chapter 3

Data

3.1 Data sources

In this study we use Copernicus Marine Service which is headed by European Commission, acting on behalf of the European Union. Copernicus is the world's largest and most ambitious Earth Observation system. It provides accurate, timely and easily accessible information to improve management of the environment, and to enable us to understand and mitigate the effects of climate change while ensuring civil security (Joaquim Alves Gaspar, 2019). In collaboration with European Space Agency they developed new family of satellites called Sentinels, currently they launched 3 types of satellites: Sentinel-1 which provides all-weather, day and night radar imagery for land and ocean services. Sentinel-2 provides high-resolution optical imagery for land services. It provides for example, imagery of vegetation, soil and water cover, inland waterways and coastal areas. And Sentinel-3 which provides high-accuracy optical, radar and altimetry data for marine and land services. It measures variables such as sea-surface topography, sea- and land-surface temperature, ocean colour and land colour with high-end accuracy and reliability. (Copernicus, 2023) Each of 3 Sentinel family satellites are present in 2 examples so-called twins to cover more globe surface. What is remarkable about this service is that data is available 365 days a year and is free to use for everyone. In upcoming years they planning to launch new Sentinel family satellites Figure 3.1, but we will not discuss this in this work. For now Sentinel-3 satellite is the main source of data for our study, since its provides wide range of marine characteristics.

Copernicus has several process levels of satellite data that can be delivered to end user, from Level 0 to Level 4, where Level corresponds to how much of post processing is applied on raw satellite data. Raw measurements are called Level 0 data which is the lowest level data in Copernicus products. Level 1 products additionally takes into account ancillary information including radiometric and geometric calibration coefficients and georeferencing parameters computed and added to the data. Level 2 is when products include geophysical variables such as secchi depth that we use as our target for forecasting. They achieve this with processing to remove the atmospheric component of signal and applying several algorithms to raw measurements. This products has highest spatial and temporal resolution. Level 3 and 4 products has little amount of missing areas, which is achieved by merging and interpolation of L2 products and typically has longer available time periods.

3.2 Datasets

In this study we incorporated few Copernicus Marine Service products to futher train forecasting model on them.

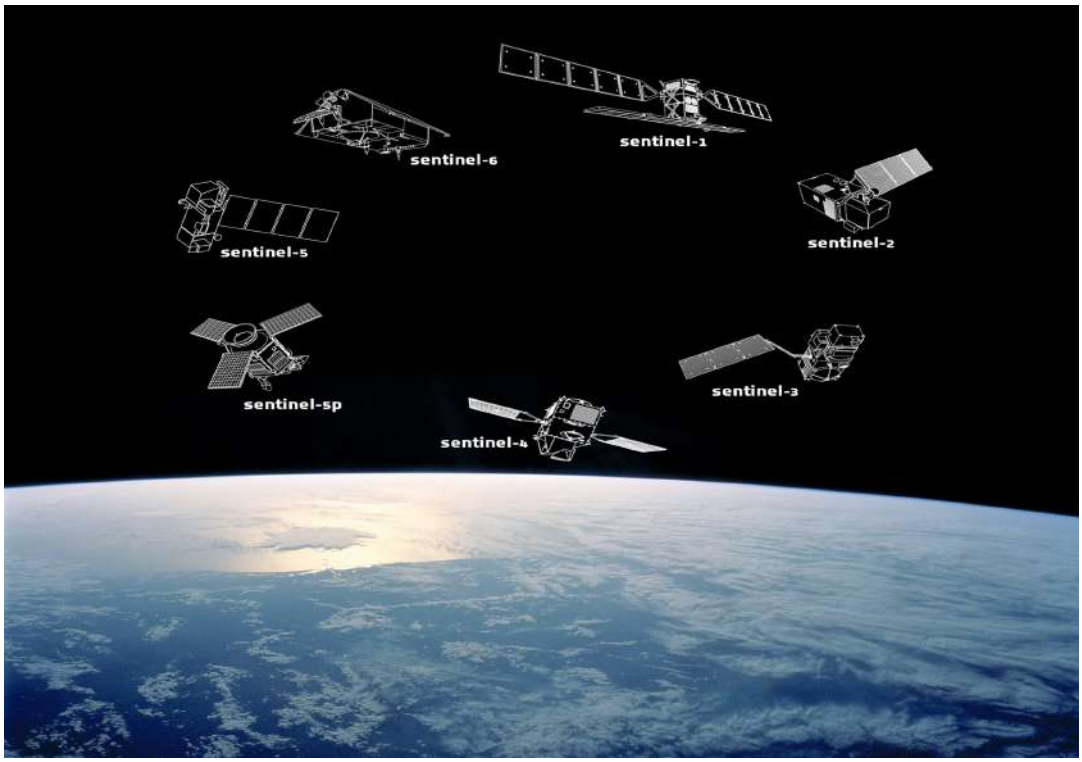


FIGURE 3.1: Sentinel family satellites

In case of forecasting we need the data to match several characteristics.
Firstly, data should be available as soon as possible, to make robust forecasts. The ideal case is when data is available for access immediately after satellite measurements, but this is not realistic for now, but we can access products that could be delivered in less than a day after satellite measurements, which is not ideal, but good enough to forecast.
And thirdly, it has to have good temporal resolution which is crucial for forecasting and enough time period length so we would be capable to train generalizable neural network without overfitting.
Secondly, it should be as much clean as it could. The raw measurements from satellite are complex and has a lot of missing values due to various climate reasons, such as cloud density, sun angle etc. For example on Figure 3.2 you can see winter period when sun angle limits the visibility of satellite in the north part of the considered region. To fill in those gaps would require a lot of affords and aggregation of different measurements from different sources.
For this reasons we used available Near real Time L4 products. Which have records from 1997 and are updated daily. L4 data has almost no missing values both spatially and temporally. We picked several biogeochemical variables which we used to forecast secchi depth. First of all we used ZSD as it is our target feature and has direct signal. For additional features we selected mass concentration of chlorophyll a in sea water (CHL).

All of data we would be talking about in next subsection could be downloaded via this link[1]

[1] https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L4_MY_009_104

FIGURE 3.2: Gaps in data due to various climate reasons

3.2.1 ZSD. Secchi depth of sea water

In this study, we propose using water transparency as a key feature for predicting future water transparency states using neural networks. Important to note that we use ZSD both as a feature and target, to forecast its future states based on previous/historical states.

The use of water transparency as a predictive feature is rooted in the inherent temporal dependencies observed in water bodies. Changes in water transparency often exhibit temporal autocorrelation, meaning that the state of water transparency at a given time is likely to be influenced by its previous states. For instance, an influx of sediment or pollution could lead to a period of decreased water transparency, while the growth of light-absorbing phytoplankton may show seasonal patterns tied to temperature and sunlight availability.

Including water transparency as an input feature allows our neural network model to capture these temporal dependencies, making it better equipped to predict future states of water transparency. By using historical water transparency data, the model can learn the underlying patterns and trends, accounting for both the short-term fluctuations and long-term changes in water transparency.

Moreover, water transparency can serve as a proxy for various environmental factors not directly included in our feature set. For instance, it can indirectly reflect factors such as nutrient concentration, sediment load, and even climatic variables, given that these factors can all affect the transparency of a water body. Hence, the use of water transparency as a feature can help in integrating the effects of these various factors into our predictive model, even if they are not explicitly included as separate features.

In summary, the inclusion of water transparency as a predictive feature in our

neural network model offers the potential for more accurate forecasting of future water transparency states. By leveraging the temporal autocorrelation in water transparency data and its indirect reflection of various environmental factors.

3.2.2 CHL. Mass concentration of chlorophyll a in sea water

Mass concentration of chlorophyll a in sea water(CHL) can directly influence sea water transparency. Chlorophyll a is a primary pigment found in cyanobacteria, algae, and plants, is essential for photosynthesis, the process by which sunlight is converted into chemical energy. This pigment plays a pivotal role in marine ecosystems, driving primary production and forming the basis of the marine food web (NOAA, 2020).

Moreover, the concentration of chlorophyll a in seawater significantly influences water transparency. Phytoplankton, microscopic marine plants that utilize chlorophyll a for photosynthesis, are directly responsible for this phenomenon. When chlorophyll a concentrations are high, indicating a dense population of phytoplankton, water transparency decreases (Jun Song Kim, 2019). Phytoplankton scatter and absorb light, thus reducing the depth to which sunlight can penetrate, known as the euphotic zone. This scattering and absorption of light also impart a greenish hue to the water, due to chlorophyll a's characteristic absorption and reflection patterns across the light spectrum.

Therefore, chlorophyll a could be very helpful signal for water transparency forecasting.

3.2.3 KD. Volume attenuation coefficient of downwelling radiative flux in sea water

In a natural setting, the volume attenuation coefficient represents the cumulative effect of three principal components: absorption, scattering, and reflection. Absorption, primarily caused by water molecules, colored dissolved organic matter (CDOM), and particles, converts light energy into other forms of energy. Scattering and reflection, on the other hand, alter the direction of light propagation without necessarily changing its intensity.

KD is inversely proportional to water transparency meaning a higher KD value suggests lower water transparency and vice versa. When KD increases, the available light at a given depth decreases more rapidly, limiting the depth of the euphotic zone where photosynthesis can occur. Thus KD can give us a direct signal for our forecasting model.

3.2.4 Study area and training dataset structure

Marine area around Great Britain coastline were chosen, which include parts of North Sea, Celtic Sea, Irish Sea, English Channel and part of Atlantic Ocean. Precise location is shown in Figure 3.3. Other large water bodies in the region - such as lakes and rivers - are excluded for the purposes of this project. 26.11% of the study area is ether surface or surface water, which are excluded from training. So 74.89% of the area shown in Figure 3.3 represents our study area. To describe the state of the marine environment in this area, we use various bio-geo-chemical and physical characteristics of the water aggregated in a grid with 4x4 kilometers cells.

While training deep neural networks it's crucial to have a large enough dataset to train robust model. Thus we choose to use historical data of selected region starting

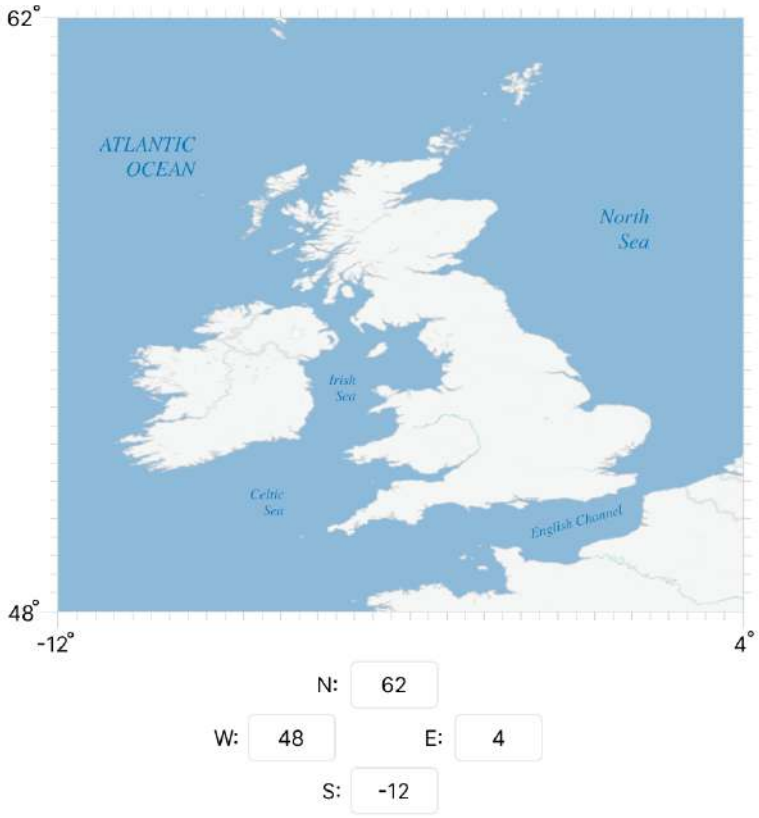


FIGURE 3.3: Location of study area, including parts of North Sea, Celtic Sea, Irish Sea, English Channel and part of Atlantic Ocean. Each pixel represents 4x4 kilometers area with

from 01.01.2000 and up to our days. So we could have around 8200 snapshots of this region, which is enough for our purposes.

The process of preparing data for training deep neural networks involves structuring the collected features in a manner conducive to the training mechanism. Initially, we divide our dataset into two: a training dataset and a validation dataset. The training dataset encompasses data from January 1, 2000, to January 1, 2022. The validation dataset consists of more recent data, specifically from January 1, 2022, to January 1, 2023. This bifurcation aims to provide a more accurate evaluation of our model's performance by testing it on unseen data.

Our training dataset, however, only contains 8030 images, which may not be sufficient to effectively train a deep neural network. Thus, following the WeatherBench experiment as outlined in Cheng Tan, 2022, we divided each image into smaller patches with a resolution of 32x64 to effectively increase the size of our dataset. This division inflates our dataset size by approximately 45 times, producing around 361350 data samples, as well as allow us to train smaller model because of decreased spatial resolution.

Subsequently, we conduct per-channel normalization on the data. The formula for this operation is as follows:

$$nD_{\text{train}} = \frac{D_{\text{train}} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{3.1}$$

Where nD_{train} represents the normalized training imagery data, D_{train} signifies the original training imagery data, μ_{train} is the mean of all points(pixels) in train dataset, and σ_{train} is the standard deviation of all points in train dataset.

This normalization process is crucial as it prevents domination of the model by features with different scales, accelerates the convergence speed of the neural network, and generally augments its accuracy. We store the values of σ and μ for future use in denormalizing the model's output.

Our dataset samples, denoted as x_i, adopt the shape (1, C, 10, 32, 64). Here, C refers to the number of channels or features, and 10 stands for a sequence of days utilized to forecast the succeeding day. The target secchi depth y_i embodies a shape of (1, 32, 64).

In the construction of the training batch, we exclude all patches inclusive of surface waters, such as lakes or rivers, which are not specifically a part of our study. Since our original dataset lacks a feature to differentiate between ocean/sea water and surface water, we implement a tree traversal algorithm to identify these regions, save them as a mask, and apply this mask post data loading. This ensures that our model trains only on pertinent data. You can see estimated mask in Figure 3.4.

(A) Surface area

(B) Surface area after tree traversal algorithm

FIGURE 3.4: Surface region mask

If you look closer on Figure 3.4a you will notice white dots on British and Irish islands, those are lakes, rivers and other surface waters, each with its own biochemical characteristics, which will add additional noise to our training data. After applying tree traversal algorithm we exclude them all from dataset, the resulted mask with solid surface is shown in Figure 3.4b

Chapter 4

Forecasting model

4.1 Problem formulation

The problem we formulate in this work centers around the prediction of ocean transparency within a designated rectangular region on Earth, specified by given longitudinal, latitudinal coordinates and time. This is essentially a *spatiotemporal forecasting problem*, that aims to predict the condition of the marine environment, specifically its transparency, in both spatial and temporal dimensions. The challenge of this problem is the multi-faceted nature of the ocean transparency that is influenced by an array of factors including, but not limited to, biological productivity, water temperature, salinity, and human activity. Simplistically we can describe it with this math definition:

$$\hat{Y}_t^l = f(x_{t-1}) \tag{4.1}$$

where:

- $\hat{Y}_t^l$ is value we want to model with timestamp t and longitude,latitude as l,
- f is some function (neural network, linear model, etc.),
- x_{t-1} represents the state of environment prior to time t,

The problem necessitates the use of historical data that encapsulates these aspects and potentially other relevant factors. This data, after a rigorous preprocessing and exploratory analysis, will serve as the foundation for the construction of the forecasting model. The spatiotemporal nature of the data introduces complexities like spatial autocorrelation, temporal autocorrelation, and potentially intricate interactions between spatial and temporal elements that must be carefully considered during model building and validation processes.

One way to solve this problem is to use statistical time series models. Statistical time series models, like *ARIMA* or *SARIMA*, could be extended into the spatial dimension, which we will use as our baselines. From a machine learning standpoint, cutting-edge models such as *convolutional neural networks* (CNNs) for spatial data, *recurrent neural networks* (RNNs) for temporal data, or a combination of both, might be more suitable for this task due to their proficiency in handling high dimensional data and capturing intricate patterns.

Recurrent architectures, including LSTM (Sepp Hochreiter, 1997) and its convolutional variant (ConvLSTM) (Xingjian Shi, 2015), have been favored in past spatiotemporal predictive learning work. Other significant recurrent architectures include Spatiotemporal LSTM (ST-LSTM) units (Qicheng Tang, 2019), which model spatial appearances and temporal variations in a unified memory pool, and PhyDNet's (Vincent Le Guen, 2020) two-branch architecture, which involves physical-based PhyCells and ConvLSTMs. Architectures based on flow, such as the invertible two-way autoencoder proposed by CrevNet, are also take place.

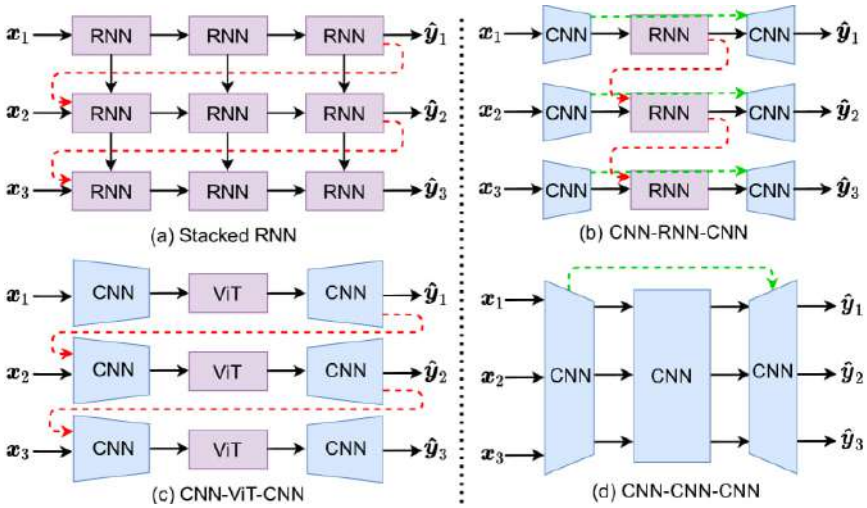


FIGURE 4.1: Major categories of the architectures for spatiotemporal predictive learning according to SimVP work. The red and blue dotted line are available to learn the temporal evolution and spatial dependency.

Authors of SimVP (Cheng Tan, 2022) categorize the prevalent methods for spatiotemporal predictive learning into four categories 4.1, based on the types of layers utilized in the architecture: (a) RNN-RNN-RNN, (b) CNN-RNN-CNN, (c) CNN-ViT-CNN, and (d) CNN-CNN-CNN. For categories (a)-(c), models generate predictions frame by frame using the previous output to capture temporal evolution. On the other hand, in category (d), models generate predictions in a one-shot manner and may employ Unet connections between the convolutional layers.

4.2 Baselines

For establishing benchmarks, we have selected several elementary yet resilient statistical methodologies. The establishment of these baselines is a pivotal step in the research process, as they allow for the comparison of results derived from the application of neural network models against those obtained from methods that have already demonstrated effectiveness and validity. This comparison is integral to determining the relative value and improvement the neural network models offer in this context. In this task we choose 2 methods, *Simple Exponential Smoothing* and *Autoregressive Integrated Moving Average*.

4.2.1 Simple Exponential Smoothing

Simple Exponential Smoothing (SES), also known as Exponential Weighted Moving Average (EWMA), is a time series forecasting method for univariate data without clear trend or seasonal components. It forms part of the broader class of exponential smoothing techniques that include methods capable of handling trends and seasonality.

The principle behind SES is to assign exponentially decreasing weights over time. This makes the method suitable for forecasting data with no clear trend or seasonal patterns, as it essentially uses a weighted average of past observations for its forecasts.

In its simplest form, the SES model can be stated with the following formula:

$$\hat{Y}_t^l = \alpha x_{t-1} + (1-\alpha)\hat{Y}_{t-1}^l \tag{4.2}$$

where:

- $\hat{Y}_t^l$ is the value we want to model for timestamp t at location l,
- α is the smoothing factor, a value between 0 and 1,
- x_{t-1} represents the actual state of environment at the previous timestamp,
- $\hat{Y_{t-1}^l}$ is the model's output for the previous timestamp and location.

The value of alpha determines the weight given to the most recent observation in the forecast. As alpha approaches 1, more weight is given to the most recent observations. As alpha approaches 0, more weight is given to the historical average of the series.

The forecasted value at time t+1 (next period) is a weighted average between the actual value at time t and the forecasted value at time t from the previous period. This makes the forecast "smoothed" towards recent observations but also accounts for the overall historical trend. We used *SimpleExpSmoothing* from *statsmodels* library with default parameters as a baseline.

4.2.2 Autoregressive Integrated Moving Average

ARIMA, or AutoRegressive Integrated Moving Average, is a forecasting method for time series data which accounts for three aspects: Autoregression (AR), Integration (I), and Moving Average (MA). The AR aspect represents the dependency of an observation on a number of lagged observations, as described by the formula:

$$\hat{Y}_t^l = C + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t \tag{4.3}$$

where:

- $\hat{Y}_t^l$ is the value we want to model for timestamp t at location l,
- C is a constant,
- ϕ_i are the parameters of the model,
- x_{t-i} are the lagged states of the environment,
- ε_t is the error term at timestamp t.

The Integration aspect indicates the differencing of observations, such as $\hat{Y}_t^l - \hat{Y}_{t-1}^l$, which is employed to make the time series stationary. Lastly, the MA aspect denotes the dependency between an observation and a residual error from a moving average model applied to lagged observations, expressed by the formula:

$$\hat{Y}_t^l = C + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \tag{4.4}$$

where ε_{t-i} are the error terms and θ_i are the parameters of the model. The combined ARIMA model, typically denoted as ARIMA(p, d, q), helps identify and predict the underlying pattern in a time series to forecast future values. In this work we used *ARIMA* from *statsmodels* library with default parameters as a baseline.

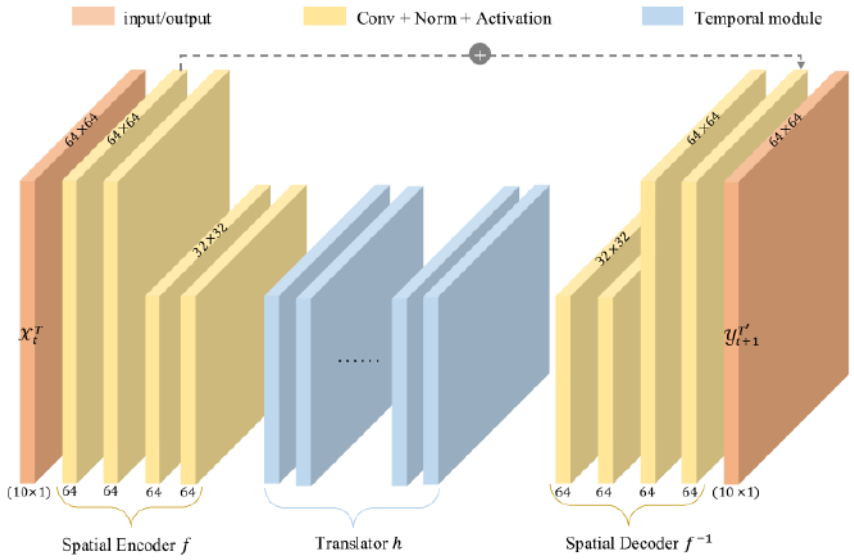


FIGURE 4.2: Simvp overall framework, source: Cheng Tan, 2022

4.3 Proposed method

Our forecasting model is strongly based on SimVP work. (Cheng Tan, 2022) In this work, the authors explore the potential of purely CNN-based models, which have not been as favored as the above RNN-based approaches. Existing CNN-based methods usually require advanced techniques like adversarial training, teacher-student distilling, and optical flow. In contrast, the authors propose a simple yet effective model called SimVP, which is based on convolutional networks and shortcut connections, and is trained with the mean square error (MSE) loss in an end-to-end manner.

The SimVP model consists of a spatial encoder, a spatiotemporal translator, and a spatial decoder. The spatial encoder encodes high-dimensional past frames into a low-dimensional latent space, and the spatial decoder decodes the latent space into the predicted future frames. The spatiotemporal translator learns both spatial dependencies and temporal variations from the latent space.

The spatiotemporal translator in SimVP is built with blocks of an Inception-like temporal module. Variants of the translator include the Inception-Unet Translator, Stacked-Unet Translator, Cross-Unet Translator, and Multi-scale Translator, each offering a different approach to capturing temporal dependencies and variations.

What interesting in this architecture is that spatiotemporal translator can be basically replaced by any U-Net (Olaf Ronneberger, 2015) like architecture and gives a space for further experiments to improve its architecture.

In comparison to baseline models that you need to fit for every given coordinate, we suggest fitting a single model, that could generalize through its spatial characteristics. Such a shift in methodology allows us to capture spatially distributed patterns, enabling superior generalization across various coordinates without requiring individual fitting. The proposed model, therefore, offers potential improvements in both efficiency and prediction accuracy.

Chapter 5

Experiments

The experiments were conducted using the SimVP model as delineated in Section 4.3, and the performance was benchmarked against several baseline models. The comparisons were carried out under a variety of conditions to ensure a comprehensive evaluation. Specific parameters of comparison encompassed the speed and accuracy of our model relative to the baseline counterparts. Subsequent to the quantitative evaluation, efforts were dedicated towards interpreting the output of the model, facilitating a deeper understanding of the results. This endeavor not only enabled validation of the model's effectiveness but also provided valuable insights into the dynamics of its capabilities.

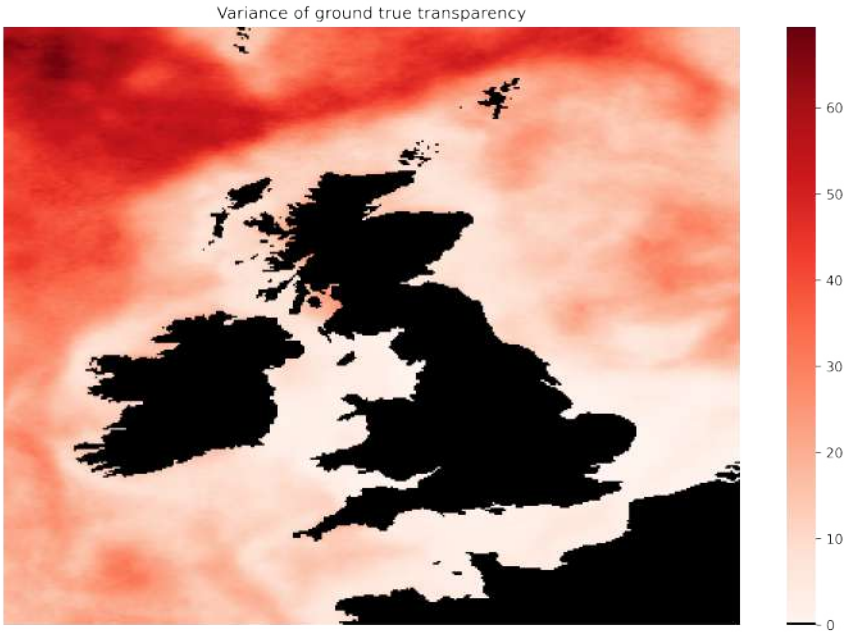


FIGURE 5.1: Variance of validation data. We calculate variance of every points within 2022-2023 year. The darker the color the more variable water transparency is in that location. Lighter colors suggest that variation in water transparency is relatively small

5.1 Metrics

To assess the predictive performance of our models, we primarily utilize two statistical measures: the Root Mean Square Error (RMSE) and the Relative Root Mean Square Error (RRMSE).

The RMSE measures the average magnitude of the error in our predictions. In essence, it is the standard deviation of the residuals or prediction errors. This measure provides us with an estimate of how far off our predictions are, on average. It is defined mathematically as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5.1}$$

Where:

- y_i represents the observed values,
- $\hat{y}_i$ signifies the predicted values,
- N is the number of observations.

The RRMSE, on the other hand, is a normalized version of the RMSE, expressing it relative to the range or variance of the observed data. This normalization allows for a more meaningful and fair comparison between datasets with distinct scales or variances. It can be computed as:

$$RRMSE = \frac{RMSE}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2}} \tag{5.2}$$

Our selection of metrics is primarily driven by their ease of interpretation, as is crucial in our study. For instance, we employ RMSE serves as an estimator of the standard forecast error in our case, which is expressed in meters. Therefore, an RMSE value of 2.1 indicates that our forecasted values diverge from the true values by an average of 2.1 meters.

In addition, RRMSE was chosen by affording the advantage of enabling comparisons of errors across different scales or units.

For instance, if we report an RRMSE value of 0.07, it means that our forecasted values deviate from the true values by an average of 7% relative to the size of the true values. This can be seen as our predictions being typically within 7% of the actual measurements, emphasizing the relative nature of this metric.

The RRMSE offers insight into the extent of the error in relation to the size of the values we are predicting, thereby providing a relative measurement of the forecast error. This allows us not only to understand the magnitude of the prediction errors but also their significance in relation to the overall scale of the values being predicted. A smaller RRMSE suggests a model that makes relatively smaller errors considering the values it is predicting, which enhances our understanding of the predictive model's performance.

In addition to these prediction quality metrics, we also monitor the computational efficiency of our models. Specifically, we document the time it takes to perform the train and inferencing on a Nvidia T4 Graphics Processing Unit (GPU). This is essential as it gives us insights into the practical viability and scalability of our models, especially when handling large datasets or real-world applications.

5.2 Models

Throughout this study, we employed an assortment of models constructed on the foundation of the SimVP architecture, as detailed in Section 4.3. We incorporate

different feature sets and also to adapt its size as per the model requirements. In our study, we trained and experimented with four distinct models, the specifics of which are presented in Table 5.1. Where in_shape refers to input shape, *S_hid* to size of hidden layer of Spatial encoder/decoder, *T_hid* size of hidden layers in Translator module, *T_N* - number of hidden layers in Translator module, *S_N* - number of hidden layers in Spatial encoder/decoder

Model name	*input features*	*in_shape*	*S_hid*	*T_hid*	*T_N*	*S_N*
Zsd_Only	Zsd	(10, 1, 32, 64)	32	256	8	2
Zsd_Only-L	Zsd	(10, 1, 32, 64)	128	512	18	4
ZsdChl	Zsd+Chl	(10, 2, 32, 64)	32	256	8	2
ZsdKD	Zsd+Kd	(10, 2, 32, 64)	32	256	8	2

TABLE 5.1: Different models that were trained during experimenting.

We train all models on same patch resolution that we described in section 3.2.4. Following Cheng Tan, 2022, we train models using the Adam optimizer (Kingma and Ba, 2017) with cosine learning rate sceduler (Loshchilov and Hutter, 2017). For every train we set batch size equals to 32, except of *ZSD-L* due to its larger memory usage batch size was set to 16. As a loss function we used simple mean square error (MSE), which authors of Cheng Tan, 2022 showed to be sufficient enough.

Our evaluation metrics of the various models is encapsulated in Table 5.2, where we incorporated parameters such as training time, number of epochs, and performance metrics. Given the constraints in computational capacity, adhering to the recommendation for 200 epochs for each model as suggested in Cheng Tan, 2022 was not feasible. Consequently, we opted for a reduced epoch count while training our models.

Num of Epochs	Model	RMSE	RRMSE	Train Time
29	Zsd_Only	2.23	0.163	64h
13	Zsd_Only-L	2.26	0.166	48h
55	ZsdChl	2.16	0.158	123h
54	ZsdKD	2.17	0.159	123h

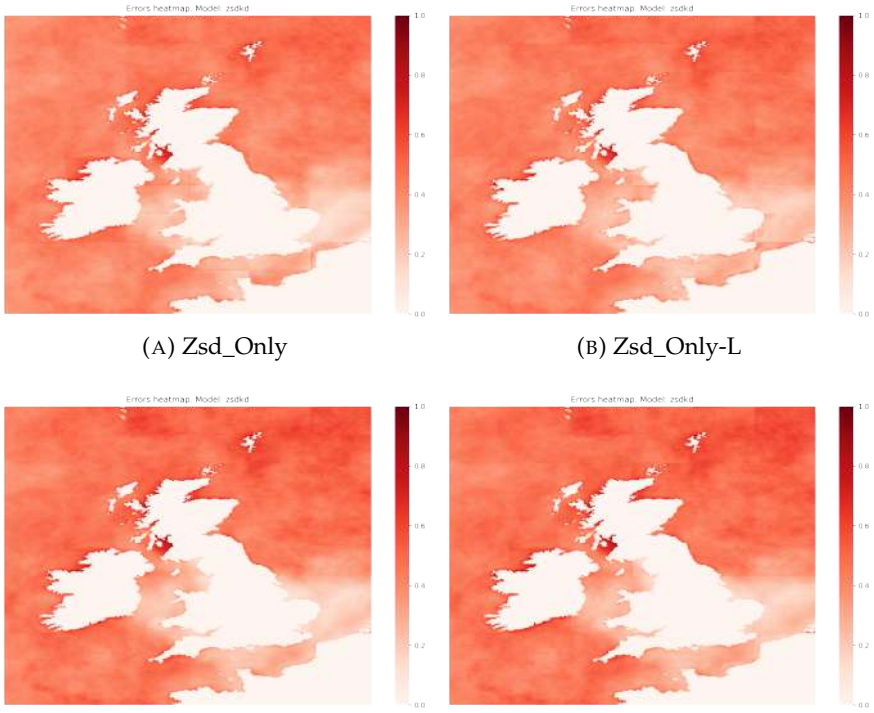
TABLE 5.2: Models Evaluation on validation dataset

From Figure 5.2, which presents a heatmap of the global average error, we can discern distinct patterns in the RMSE (Root Mean Square Error) of forecasted water transparency across our specified region of interest.

Most notably, there is a pronounced increase in RMSE observable in proximity to the coastlines. This suggests that the forecast model exhibits a higher degree of error when predicting water transparency near coastal regions. As we shift our focus towards the open ocean, the heatmap displays a more stable color gradient indicative of a relatively constant RMSE.

However, it should be noted that there are sporadic exceptions scattered throughout these open ocean areas. These anomalies could potentially be attributable to a multitude of factors that might be worth investigating further to optimize our model's predictive accuracy.

This analysis provides an initial understanding of the performance of our model, highlighting areas of strength and potential improvement. Further investigation into the reasons behind the increased RMSE near coastlines and sporadic errors in open



(A) Zsd_Only (B) Zsd_Only-L

(C) ZsdChl (D) ZsdKD

FIGURE 5.2: Models error on validation dataset

ocean regions is warranted to refine the predictive capabilities of our model. But now we can suggest that there could be several reasons to that, it could be caused by natural factors like tides or sediments from rivers as they are much more prominent closer to the coastline. Other reason could be human factors such as pollution from industrial waste. There also could be potential reason for this hidden in accuracy of optical measurements, it could have lower accuracy due to the reflection from land.

5.2.1 Train analysis

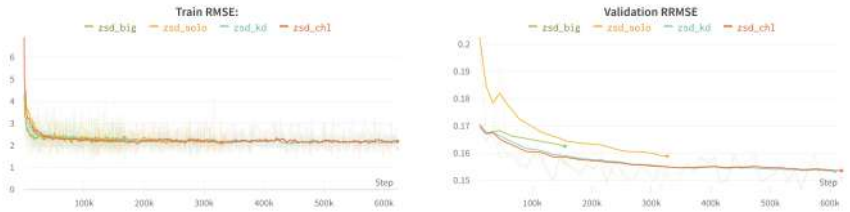
The process of training neural networks necessitates the careful monitoring of the progression of the training and the corresponding loss curves. This oversight serves as an indicator of the efficacy of the training, with any abrupt fluctuations potentially signaling underlying issues within the dataset or the architecture of the model itself.

Throughout the course of our models training, we logged the losses observed during both training and validation phases, as illustrated in Figure 5.3. From these illustrations, we can discern a noteworthy pattern. The training loss curve, as shown in Figure 5.3a, reached a point of convergence swiftly after an initial period. This suggests that our model was successfully able to minimize the discrepancy between its predictions and the actual target values during the training phase, thereby exemplifying its learning capacity.

Contrastingly, the validation loss curve, as depicted in Figure 5.3b, did not mirror this rapid convergence. Instead, the validation losses continued to decrement at a slow yet consistent pace. This sustained decline is indicative of a continuous, albeit slow, improvement in the model's ability to generalize its learning to unseen data.

The continuous decline in validation loss could potentially signify that the model is still improving its performance on the validation set, which, in turn, suggests that

the model may not have yet reached its optimal complexity. This suggestion can be confirmed by Cheng Tan, 2022, when they train their model for at least 200 epochs. Additionally they showed that train models for x3(600 epochs), x5(1000 epochs) and even x10(2000 epochs) more epochs still decrease validation error, which could be considered as a future improvement of our work.



(A) Train loss

(B) Validation loss

FIGURE 5.3: Models losses during training. Train loss is logged every iteration. Validation loss is logged after every epoch.

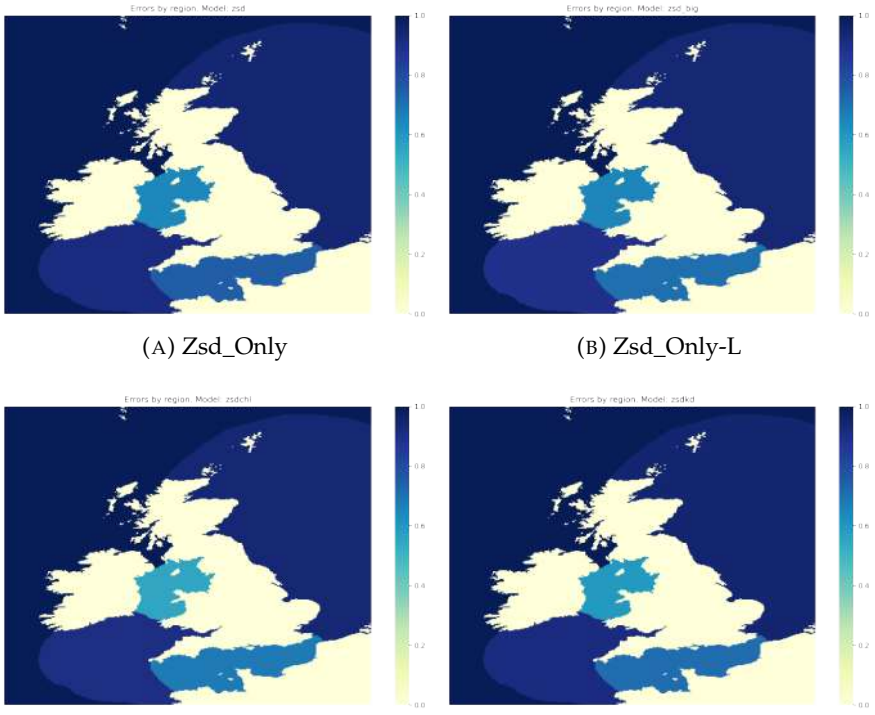
5.2.2 Evaluation on different marine environments

In an effort to gain a more nuanced understanding of our model's performance, we conducted an in-depth analysis that extended beyond the use of standard formal metrics. The first phase of this assessment involved manual labelling of our study region, delineating it into distinctive marine environments: the Atlantic Ocean, North Sea, Celtic Sea, Irish Sea, and English Channel. This stratification was instrumental in our subsequent analysis.

We proceeded to calculate the RMSE for each sub-region, a step that facilitated the examination of our model's performance in specific geographical contexts. This partitioned evaluation approach is advantageous as it provides insights into the model's predictive accuracy across diverse marine environments, thereby offering a more comprehensive understanding of the model's overall performance and robustness.

On Figure 5.4 we can see performance of every our model on different marine environments. The lighter a blue color is the less RMSE, for visualisation purpose we normalize RMSE values. From this we can clearly see that every model has the same per-region error distribution. Atlantic Ocean is the harder one for model to forecast, than we can see North and Celtic seas are almost same color, which means they are very similar to forecast, and than we have English Channel and Irish Sea in more brighter colors, when Irish Sea forecast has the lowest error. More detailed metrics can be found in table 5.3. We can state that some regions are more harder for our model to learn for example in Atlantic Ocean region on average model has 40% greater error than same model but in Irish Sea environment.

The observed results can largely be attributed to the inherent variability in water transparency across different regions, as depicted in Figure 5.1 which illustrates the variance of ground truth validation data. Upon closer inspection of this figure, it becomes evident that regions such as the Irish Sea and the English Channel, where our models demonstrate better performance in terms of RMSE, exhibit lower temporal variance. The smaller level of fluctuations in water transparency in these regions over time potentially simplifies the modeling task, thus leading to improved



(A) Zsd_Only (B) Zsd_Only-L

(C) ZsdChl (D) ZsdKD

FIGURE 5.4: Models error on validation dataset within different marine environments

Model	Atlanctic Ocean	North Sea	Celtic Sea	English Channel	Irish Sea
Zsd_Only	2.38	2.24	2.16	1.75	1.52
Zsd_Only-L	2.43	2.30	2.14	1.65	1.40
ZsdChl	**2.32**	**2.18**	**2.09**	**1.59**	**1.34**
ZsdKD	**2.32**	2.20	2.10	1.63	1.39

TABLE 5.3: Models Evaluation on validation dataset. Each value represents RMSE metric in corresponding marine environment

results. Therefore, the inter-regional disparity in RMSE performance can be reasonably explained by the natural variability in water transparency prevalent in these geographies.

5.2.3 Comparison to baselines

In Section 4.2, we elaborate on the construction of two baseline models, namely the Simple Exponential Smoothing (SES) and AutoRegressive Integrated Moving Average (ARIMA). These models serve as comparators, allowing us to benchmark the predictive capabilities of our specifically trained models. It is important to note that these models are exclusively temporal in nature.

Due to this temporal characteristic, evaluating these models requires an exhaustive fitting process at every point within our defined area. We trained our baselines model with a history window of 10, this is the same history window that we use to forecast with our trained models.

This procedure is computationally demanding, rendering it less efficient for larger datasets. Preliminary estimates indicate that a comprehensive evaluation across the full validation dataset would necessitate approximately a week of computational time.

In order to mitigate the computational demands of evaluating these models, we have employed a sampling strategy. We selected a subset of 100,000 data points from our validation dataset, utilizing a uniform random sampling method. This approach encompasses both the temporal aspect as well as spatial coordinates (longitude and latitude) to ensure a representative and unbiased sample of our overall data.

Subsequently, we leveraged this subset to evaluate the predictive performance of our models. The forecasts generated by our models were compared directly with the actual observations at these specific points. This methodology affords us an efficient yet comprehensive understanding of our model's capabilities when juxtaposed with the baseline SES and ARIMA models, without sacrificing the scientific rigour of our evaluation process.

Model	rmse	rrmse	GPU inference time (sec)	CPU inference time (sec)
ARIMA	3.080	0.230	-	1243.9
SES	2.626	0.197	-	334.6
Zsd_Only	2.3	1.172	2.6	5.03
Zsd_Only-L	2.333	0.175	3.4	14.0
ZsdChl	**2.236**	0.167	5.1	7.9
ZsdKD	2.240	0.168	5.2	7.9

TABLE 5.4: Comparison with baseline, on 100000 sampled points (baseline models that we used could only run on CPU).

As demonstrated in Table 5.4, the models developed in our research exhibit significantly enhanced performance relative to the baseline counterparts, particularly in terms of both RMSE and inference time. The RMSE of our models is considerably lower, implying better predictive accuracy and robustness. Furthermore, the markedly reduced inference time enhances the operational feasibility of these models in real-world scenarios, providing an efficient tool for rapid decision-making and analysis. Therefore, the models we have developed not only elevate the speed and performance of inference but also amplify their forecasting capabilities, thereby offering an efficacious solution for robust and efficient predictive modeling, which is expected to lead to more robust water management practices, cost savings, improved risk mitigation, and overall better understanding of our water bodies.

Chapter 6

Conclusions

6.1 Summary

In the context of this study, we have developed and trained multiple predictive models for seawater transparency, conducting extensive evaluation of their performance. Our work has led to several noteworthy contributions:

- We created a dataset for training and validating water transparency forecasting models in marine environments.
- To the best of our knowledge, we proposed the first application of modern spatio-temporal neural network architectures for ocean water transparency forecasting.
- We meticulously applied and examined our methodological approach, juxtaposing it with the baseline strategy. This involved comprehensive evaluations conducted across a diverse marine environments surrounding the Great Britain. This rigorous testing approach ensured a holistic comparison, taking into account the environmental conditions unique to different marine locales.
- Our approach demonstrated superior results against the established baseline approaches in both the speed of inference and overall prediction accuracy: Our best model *ZsdChl* showed a decrease of RMSE by 17.4% (2.63 -> 2.24), and increase in inference time efficiency in 66.3 times (334.6 -> 5.04 seconds) in comparison to best baseline *SES*.
- We validated our models on different regions and showed that our model perform worse in regions with high variance such as close to coastline and best on regions with less water transparency variance like in Irish Sea or English Channel. This points out that our model is vulnerable to regional variability, which gives us a window for further improvements with encountering regional characteristics. All validations were made on unseen data from future(in comparison to train data) which examines that our model is not overfitted and flexible enough to encompass real world future data.

The efficiency of these models indicates their potential as a valuable tool in advancing research in various domains. Moreover, they could be successfully leveraged for practical implementations in diverse sectors such as industrial diving, aquaculture, and the tourism industry, thereby enhancing safety standards and operational efficiency. We foresee that these models can contribute to the development of innovative solutions and procedures within these industries, thus paving the way for greater progress and safety measures.

6.2 Directions for future research

There are several potential areas of further improvement. One of them lies in enriching the dataset with additional features. Our research demonstrated that water transparency varied significantly across different geographical regions, indicating the impact of location-specific factors. Therefore, it may be beneficial to incorporate location features into our dataset, thereby integrating the influence of geographical variables into our analysis.

Another potential area for the improvement lies in extending the training period over a larger number of epochs. Our preliminary analysis indicates that the performance of the model could be significantly improved by allowing it to learn over an extended period, potentially leading to more refined and accurate predictions.

Furthermore, it's worth considering the expansion of the model's complexity as a feasible strategy for increasing its performance. This approach, characterized by increasing the model's complexity or size, has been substantiated by numerous studies, including our own research. This approach is predicated on the understanding that larger models, with their expanded capacity to capture and process information, can potentially yield more accurate results, thereby improving performance.

Our work has also shed the light on the conditions where the model underperforms. This include forecasting close to the coastline, which can be improved by expanding our dataset to include the information about the artifacts of existng coastal infrastructure, land use patterns, population density, etc. Additionally, the inclusion of real-time data about tidal cycles, wave activity, and local weather patterns can potentially enhance the model's predictive capability by accounting for the constant changes in coastal water environments.

Finally our current predictive model forecasts water quality for the following day. However, to provide more practical and actionable insights for policy makers, environmental managers, and the public, there's a need to expand this forecasting window beyond a single day. Predictions with a longer time horizon, for instance, a week or even a month ahead, could allow for better planning and more effective intervention strategies.

Bibliography

Aji Prasetya Wibawa Agung Bella Putra Utama, Hakkun Elmunsyah et al. (2022). *Time-series analysis with smoothed Convolutional Neural Network*.

Angela Lausch Stefan Erasmi, et al. (2016). *Understanding Forest Health with Remote Sensing -Part I—A Review of Spectral Traits, Processes and Remote-Sensing Characteristics*.

Bing Yu Haoteng Yin, Zhanxing Zhu (2017). *Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting*.

Cheng Tan Zhangyang Gao, Siyuan Li Stan Z. Li (2022). *SimVP: Towards Simple yet Powerful Spatiotemporal Predictive Learning*.

Copernicus (2023). *Discover Our Satellites*. URL: https://www.copernicus.eu/en/about-copernicus/infrastructure-overview/discover-our-satellites (visited on 06/10/2023).

Espeholt L., Agrawal S. Sønderby C. et al. (2022). "Deep learning for twelve hour precipitation forecasts". In.

Godson Ebenezer Adjovu Haroon Stephen, David James Sajjad Ahmad (2023). *Overview of the Application of Remote Sensing in Effective Monitoring of Water Quality Parameters*.

Heddam, Salim (2016). *Secchi Disk Depth Estimation from Water Quality Parameters: Artificial Neural Network versus Multiple Linear Regression Models?*

Joaquim Alves Gaspar HSH Prince Albert II of Monaco, Joseph Aschbacher et al. (2019). *The Blue Book*.

Jun Song Kim Il Won Seo, Donghae Baek (2019). "Seasonally varying effects of environmental factors on phytoplankton abundance in the regulated rivers". In.

Jun Yang Peng Gong, Rong Fu et al. (2013). *The role of satellite remote sensing in climate change studies*.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Krajick, Kevin (2022). "How Is Climate Change Affecting Ocean Waters and Ecosystems?" In.

Li G., Cheng L. Zhu J. et al. (2020). "Increasing ocean stratification over the past half-century." In.

Loredana Tecar, Bianca Balulescu Marica Buresin (2014). *Development analysis of the Timisoara based on satellite image*.

Loshchilov, Ilya and Frank Hutter (2017). *SGDR: Stochastic Gradient Descent with Warm Restarts*. arXiv: 1608.03983 [cs.LG].

NOAA (2023). "Why should we care about the ocean?" In.

NOAA, US National Marine Ecosystem (2020). *Chlorophyll-A*.

Olaf Ronneberger Philipp Fischer, Thomas Brox (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

Parra, Lorena (2019). *Remote Sensing and GIS in Environmental Monitoring*.

Pelton, Joseph (Nov. 2017). *The Expanding Use of Space in Communications, Navigation, Remote Sensing and Weather Satellites*. DOI: 10.1007/978-3-319-39273-8_3.

Platt, Trevor et al. (2017). "Primary Production: Sensitivity to Surface Irradiance and Implications for Archiving Data". In: *Frontiers in Marine Science*.

Qicheng Tang Mengning Yang, Ying Yang (2019). *ST-LSTM: A Deep Learning Approach Combined Spatio-Temporal Features for Short-Term Forecast in Rail Transit*.

Ricky J. Lee, Sarah L. Steele (2014). *Military Use of Satellite Communications, Remote Sensing, and Global Positioning Systems in the War on Terror*.

Robert J. W. Brewin Jaime Pitarch, Giorgio Dall'Olmo et al. (2023). *Evaluating historic and modern optical techniques for monitoring phytoplankton biomass in the Atlantic Ocean*.

Sabins, Floyd F (1999). *Remote sensing for mineral exploration*.

Sepp Hochreiter, Jürgen Schmidhuber (1997). *LONG SHORT-TERM MEMORY*.

Serhii Shevchuk Viktor Vyshnevskyi, Olena Bilous (2022). *The Use of Remote Sensing Data for Investigation of Environmental Consequences of Russia-Ukraine War*.

Thilo Wellmann Angela Lausch, Erik Andersson et al. (2020). *Remote sensing in urban planning: Contributions towards ecologically sound policies?*

Timo Toivanen Sampsa Koponen, Ville Kotovirta Matthieu Molinier Peng Chengyuan (2013). *Water quality analysis using an inexpensive device and a mobile phone*.

Vincent Le Guen, Nicolas Thome (2020). *Disentangling Physical Dynamics from Unknown Factors for Unsupervised Video Prediction*.

Wen Liu Fumio Yamazaki, Yoshihisa Maruyama (2019). *Detection of Earthquake-Induced Landslides during the 2018 Kumamoto Earthquake Using Multitemporal Airborne Lidar Data*.

Wernand, M. R. (2010). "On the history of the Secchi disc". In: *Journal of the European Optical Society - Rapid publications*.

Xingjian Shi Zhourong Chen, Hao Wang Dit-Yan Yeung (2015). *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*.

Yann LeCun Yoshua Bengio, Geoffrey Hinton (2015). *Deep Learning*.

Yipeng Liao Yun Li, et al. (2021). *Water Transparency Prediction of Plain Urban River Network: A Case Study of Yangtze River Delta in China*.

ZhongPing Lee Shaoling Shang, Chuanmin Hu et al. (2015). *Secchi disk depth: A new theory and mechanistic model for underwater visibility*.