MSc Thesis Applied Mathematics

A Machine Learning Approach for Estimating Gross Primary Productivity Using Sentinel-2 Data

Karisma

Chair and Supervisor:
Prof. Dr. A. J. Schmidt-Hieber

Committee:
Dr. Annika Betken
Dr. Ing. A.B. Zander

External Supervisors:
Ir. Anna Spinosa
Ir. Mario Alberto Fuentes Monjaraz
Dr. Valeria Mobilia

July, 2024

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.　**Deltares**

Acknowledgement

I would like to extend my deepest gratitude to those who have supported me throughout my master's journey in Applied Mathematics at The University of Twente.

First and foremost, I would like to thank my supervisor, Prof. Dr. A. J. Schmidt-Hieber, who also served as the chair of my defense committee. His invaluable insights and understanding, especially during times when I struggled to meet deadlines, have been instrumental in the completion of this thesis. I am also grateful to the defense committee members, Dr. Annika Betken and Dr. Ing. A. B. Zander, for their constructive feedback and guidance.

I extend my sincere thanks to Deltares for providing me with the incredible opportunity to conduct my research. I also want to thank LPDP for funding my master's degree. This journey would not have been possible without their generous scholarship. My heartfelt appreciation goes to my Deltares supervisor, Ir. Anna Spinosa, whose patience and support were particularly crucial during the initial stages of my internship when I faced significant challenges. I am deeply grateful to Ir. Mario Alberto Fuentes Monjaraz for his unwavering assistance and for teaching me extensively about remote sensing, which was vital for data collection. I also thank Dr. Valeria Mobilia, who stepped in to help in Anna's absence, and Jana Lim, for her continuous support and thorough reviews of my thesis. My gratitude extends to all my colleagues at the Data Science and Water Management Department, especially Paul, the head of the department, and my fellow interns. Their camaraderie and support have been invaluable.

I am profoundly thankful to my family for their unwavering support from afar. To my mother, Ayik Suryandhari, my sister Karimah, my brother Muhammad Rowi, my aunties, my grandmother, my nephew Baim, and my niece Shanum, your encouragement has been my driving force. I would also like to acknowledge my Applied Mathematics fellows: Nina, Sophie, Fabio, Jose, Steven, and Renske. Your companionship throughout the entire master's program, especially during exam weeks and our memorable international lunches, has been a source of strength and joy for me.

To my Indonesian friends in the Netherlands, particularly those in the Mangan group and the Serba Serbi group, thank you for making me feel at home and for adding color to my life here. Special thanks to Helda, Nisa, and Luluk for their constant advice, willingness to listen to my problems, and for being my travel companions.

Finally, I would like to thank myself for not giving up, for striving to make my dreams come true, and for setting new goals despite the challenges. Studying in the Netherlands has been a transformative experience, and I am proud of my perseverance and achievements. Thank you all for being part of this incredible journey.

Above all, I am grateful to Allah SWT. I couldn't have done this without divine guidance. This is also my present to my late father, Abdul Karim. I miss him so much.

<div style="text-align: right">

Enschede, 24 July 2024
Karisma

</div>

Abstract

Predicting Gross Primary Productivity (GPP) across diverse ecosystems is essential for understanding the global carbon cycle and managing environmental resources effectively. This study evaluates the effectiveness of three different models, namely SARIMAX, XGBoost, and LSTM in estimating GPP using a combination of in-situ measurements and remote sensing data across various European ecosystems. The research consist of two main stages: the development of site-specific models to understand individual site characteristics and the creation of a unified model capable of generalizing predictions across different ecosystems without further site-specific adjustments. Our findings indicate that XGBoost consistently outperformed other models, showing superior prediction accuracy and robustness, particularly when generalized across multiple sites. SARIMAX and LSTM models also demonstrated useful capabilities, though with some limitations in specific contexts such as catastrophic forgetting in LSTM and poor performance in peak GPP predictions by SARIMAX. The inclusion of specific remote sensing indices, like the modified normalized difference vegetation index (MNDVI) and the enhanced vegetation index (EVI), significantly improved model performance across varied ecosystems. This study underscores the potential of integrating machine learning techniques with traditional ecological modeling approaches to enhance the prediction of GPP, which can significantly contribute to ecological management and climate change mitigation strategies. Future work should focus on refining these models' ability to handle diverse data sets and improve their predictive reliability across global ecosystems.

Keywords: Gross Primary Productivity, Sentinel-2, SARIMAX, XGBoost, LSTM

Contents

List of Abbreviations

ACF	Autocorrelation Function
AOT	Aerosol Optical Thickness
ARIMA	Autoregressive Integrated Moving Average
BOA	Bottom Of Atmosphere
BPPT	Back Propagation Through Time
CIr	Red-edge Chlorophyll Index
DBF	Deciduous Broadleaf Forest
DL	Deep Learning
EC	Eddy Covariance
ENF	Evergreen Needleleaf Forest
EVI	Enhanced Vegetation Index
GPP	Gross Primary Productivity
ICOS	The Integrated Carbon Observation System
INRA	National Institute for Agricultural Research
LSTM	Long Short Term Memory
LSWI	Land Surface Water Index
LUE	Light Use Efficiency
MAE	Mean Absolute Error
ML	Machine Learning
MNDVI	Modified Normalized Difference Vegetation Index
MNDWI	Modified Normalized Difference Water Index
MSE	Mean Squared Error
NEE	Net Ecosystem Exchange
NIR	Near Infrared
NDII	Normalized Difference Infrared Index
NDVI	Normalized Difference Vegetation Index
NPP	Net Primary Production
PACF	Partial Autocorrelation Function
PAR	Photosynthetically Active Radiation
R^2	Coefficient of determination
RECO	Ecosystem Respiration
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARIMAX	Autoregressive Integrated Moving Average with Exogenous variable
SCL	Scene Classification
SR	Surface Reflectance
TOA	Top Of Atmosphere
VI	Vegetation Index

VPD	Vapor Pressure Deficit
XGBoost	Extreme Gradient Boosting
WV	Water Vapour

Chapter 1

Introduction

Gross Primary Productivity (GPP) indicates the amount of carbon dioxide fixed in an ecosystem through vegetation photosynthesis [1]. Thus, it serves as a critical measure in understanding the global carbon cycle. Understanding and accurately predicting GPP is essential for numerous ecological and environmental management applications, including climate change mitigation, agriculture optimization, and biodiversity conservation. For instance, GPP is a vital indicator of forest productivity, and the prediction of GPP from climate variables can aid the management of forests and the prediction of future growth under climate change scenarios [2]. Despite its significance, the accurate prediction of GPP across diverse and geographically dispersed ecosystems is challenging due to the complexity of ecosystem responses to varying environmental conditions.

At present, there is no standard protocol for quantifying GPP. Existing methods are complex and are typically based on either in-situ or remotely-sensed measurements. The most widely used in-situ measurement methods are open chambers (manual measurements) and Eddy Covariance (EC) techniques [3]. The EC method is extensively used in different regional networks to measure Carbon Dioxide (CO_2), water, and energy exchanges between the biosphere and the atmosphere. Currently, over 250 EC sites worldwide are actively monitoring carbon exchange at high temporal resolution, typically every half hour, across various biomes and climatic conditions [4]. The other remotely-sensed measurement methods involve the indirect GPP calculations by deriving vegetation indices from Earth Observations. Satellites from missions such as Landsat, MODIS, and Sentinel have been used to derive Vegetation Indices (VIs). VIs are formulas that reduce multispectral scanner data to a single number for assessing vegetation characteristics like species, leaf area, stress, and biomass [5]. Over the past 30 years, vegetation indices have been refined and used extensively for monitoring vegetation health, growth levels, and stress, among other conditions [6].

Modeling plays a critical role in predicting GPP. It enables researchers to simulate and understand complex ecological processes with greater precision. Recent advancements in remote sensing technology have significantly enhanced the capability to estimate GPP using Sentinel-2 data. Pabón et al. explored the use of red-edge-based and near-infrared-based VIs alongside machine learning techniques to predict GPP at multiple sites in Europe and North America with EC measurements. They concluded that incorporating multiple spectral bands and their interactions significantly enhanced model performance, with random forests achieving a R^2 value of 0.71, despite some challenges in predicting GPP during extreme weather conditions [7].

The potential of Sentinel-2 data for estimating GPP was also assessed by Spinosa et al. [8] and compared with MODIS data. The high spatial resolution of Sentinel-2 data

provided a significant advantage in modeling GPP regarding precision. Similarly, Astola et al. [9] also conducted a comparative study of Sentinel-2 and Landsat 8 data. The result demonstrated that the models based on Sentinel-2 data showed better performance than those using Landsat 8 data for all forest variables, indicated by lower RMSE values and higher R^2 values. In contrast, a study in the European Nordic region comparing the performance of Sentinel-2 against MODIS data in estimating GPP indicated that the finer resolution of Sentinel-2 (10 m) did not significantly outperform MODIS (250 m and 500 m). This suggests that factors other than spatial resolution, such as model selection and environmental variables, are crucial in improving GPP estimation accuracy [10]. In addition, the availability of longer time series and increased data volume from Sentinel-2 can contribute to developing more robust models compared to MODIS.

There are also increasing number of research studies utilizing Machine Learning (ML) to predict GPP. For example, Chen et al. [11] developed a model based on artificial neural networks model that effectively predicts GPP in coniferous forests, with climatic factors being the dominant contributors including net radiation, vapor pressure deficit (VPD), air temperature, and soil water content or CO_2 concentration. Wang et al. [12] assesed six machine learning algorithms for predicting GPP in the Mongolian Plateau. Additionally, Chang et al. [13] evaluated GPP across nine ChinaFlux sites using a random forest model based on light use efficiency (LUE). The latter authors developed a different model for each of the sites. The authors included seasons as categorical data, as GPP follows clear seasonal patterns, thus eliminating the time dependence of the data. These studies showed that most models exhibited better performance in capturing the temporal variations and GPP magnitude in mixed forests and evergreen needleleaf forests compared to evergreen broadleaf forests and grasslands. Tree based methods were also employed to predict GPP in [14].

Despite the promising results from integrating machine learning with remote sensing data, significant challenges persist. Existing models for predicting GPP are mainly site specific, as they were developed and calibrated for individual locations with little adaptability to different ecosystems [15, 16, 17]. This approach, while effective on a local scale, fails to address the need for scalable and generalizable models that can operate across multiple sites with varying ecological characteristics. This research aims at addressing this gap by introducing a novel approach consisting in the development of a unified model that uses ML and Deep Learning (DL) techniques to predict GPP across different ecosystem types by combining in-situ data and remote sensing data. The model uses vegetation indices and other relevant variables and is designed to generate accurate predictions for various sites without needing site-specific recalibration. This approach not only aims at addressing the limitations of existing models, but also at enhancing scalability of the GPP prediction, making it applicable globally.

On the other hand, this advancement introduces new challenges, as the influence of GPP by the ecosystem type adds layers of complexity to the model's application. For example, forests with large, dense tree canopies are characterized by higher GPP compared to grasslands due to their greater leaf area and photosynthetic capacity. In turn, grasslands are highly responsive to climatic elements such as temperature and rainfall, making the spatial patterns and evolutionary trends of grassland GPP more intricate [12]. Moreover, in grasslands and evergreen needle-broad forests, moisture variables and photosynthetically active radiation (PAR) were found to have substantial impacts on GPP [13].

In order to address the aforementioned challenges would be beneficial to explore methods that have proven successful in other areas with time series data which is also used in this research. Time series analysis is a vital tool for analyzing sequential data, and it

has led to major progress in fields such as finance [18], public health [19] and other fields. The AutoRegressive Integrated Moving Average (ARIMA) model is one such well-known method that has been used to predict future trends based on historical patterns [20, 21, 22].

The ARIMA model was popularized by George Box in [23]. Extending this approach, the AutoRegressive Integrated Moving Average with Exogenous variable (SARIMAX) model incorporates exogenous variables, enhancing ARIMA's applicability. It has been successfully applied in various sectors including hydrology [24], ecology [25], and healthcare [26, 27]. This method can be applied to predict GPP by incorporating in-situ measurements and satellite-derived VIs as exogenous variables. GPP exhibits seasonal variations, particularly influenced by ecological and climatic conditions that change throughout the year. Integrating these indices as exogenous factors allows the SARIMAX model to leverage spatial data alongside temporal observations from in-situ measurements.

ML and DL approaches have also shown outstanding performance in time series analysis across various fields [28, 29, 30, 31]. Among the great diversity, Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) networks are particularly prominent. XGBoost, an advanced implementation of gradient boosting, has been validated for its superior performance and robustness in handling structured data [32], offering critical insights through feature importance scores. Additionally, XGBoost showed better model performance than the random forest for simulating Net Ecosystem Exchange (NEE) based on the FLUXNET dataset, especially in terms of computational efficiency [33]. On the other hand, LSTMs are well-suited for modeling temporal sequences due to their capability to maintain long-term dependencies, essential for accurate forecasting in time series data [34].

This study aims to compare statistical methods for time series analysis, ML, and DL methodologies. Specifically, the study will compare SARIMAX, XGBoost, and LSTM in terms of their effectiveness in generalizing across different ecosystems. The approach involves developing individual models for each site in the initial stage to understand the specific site characteristics. In the later stage, a unified model will be developed to predict GPP across different ecosystem types. Throughout both stages, the study will compare the models' ability to generalize across different ecosystems. In essence, this study aims to answer the following questions:

1. How well do the SARIMAX, XGBoost, and LSTM models predict the GPP at individual sites?

2. How well do the SARIMAX, XGBoost, and LSTM models perform when they are trained on combined data from multiple sites across Europe?

3. How do different indicators from remote sensing and in-situ data contribute to predicting GPP across diverse ecosystems in European sites?

4. How do the SARIMAX, XGBoost, and LSTM models perform in terms of generalization when applied to sites not included in the training data?

The thesis is structured into six chapters, starting with the introduction, followed by a theoretical background on GPP and the methodologies in Chapter 2. Chapter 3 and 4 focus on data collection, preprocessing, model design, and experimental setup. The results, model evaluation, and discussion are described and discussed in Chapter 5. Finally, the last chapter presents the conclusions, limitations, and recommendations for further research.

Chapter 2

Background

2.1 Gross Primary Productivity

Gross Primary Productivity (GPP) is a fundamental ecological metric that represents the total amount of CO_2 that plants and algae capture from the atmosphere through photosynthesis at the ecosystem scale [35]. The chemical reaction of photosynthesis is

$$6CO_2 + 6H_2O + \text{light} \rightarrow C_6H_{12}O_6 + 6O_2. \tag{2.1}$$
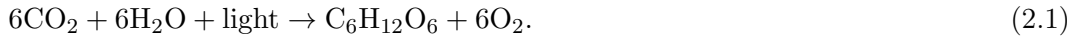
Carbon dioxide (CO_2) and water (H_2O), in the presence of light, are transformed into glucose ($C_6H_{12}O_6$) and oxygen (O_2). The glucose produced serves as the primary energy storage molecule, fueling various biological activities within the plant, including growth and reproduction, while the oxygen produced is released into the atmosphere [36].

The availability of water significantly impacts the photosynthesis rates. Light and solar radiation are also crucial indicators, especially Photosynthetically Active Radiation (PAR), which refers to the specific range of solar radiation between 400 and 700 nm that can be utilized by photosynthetic organisms for the process of photosynthesis [37]. The amount and quality of light influence the efficiency and rate of photosynthesis. There are also other indicators that influence this process, such as temperature, chlorophyll, and soil nutrients. Temperature influences enzymatic activities critical for photosynthesis. Each ecosystem has a temperature range that optimizes these processes.

GPP is measured over a specific area and time period, typically expressed in grams of carbon per square meter per day (gCm^2day^{-1}). From GPP, other important ecosystem metrics can be derived, such as Net Primary Production (NPP) and Net Ecosystem Exchange (NEE). NPP accounts for the CO_2 released by plants through respiration (e.g the process by which plants consume some of the captured carbon, in the form of glucose, to maintain their metabolic functions). Therefore, NPP is expressed as

$$NPP = GPP - R_{plant}, \tag{2.2}$$

where R_{plant} denotes the plant respiration. NEE, is a broader measure that includes all carbon exchanges between the ecosystem and the atmosphere. It accounts for not only GPP but also all respiratory processes, including plant respiration and the respiration of all other organisms within the ecosystem (R_{total}). Mathematically, NEE is expressed as

$$NEE = GPP - R_{total}. \tag{2.3}$$

As well as GPP, NPP, NEE and R_{total} are typically obtained using various methods, such as EC towers that measure CO_2 fluxes at high frequencies or through modeling approaches that integrate field observations. Understanding the relationship between GPP,

NPP, and NEE provides a strong foundation for assessing the impact of environmental changes, such as climate variations, land use changes, and management practices, on ecosystem carbon dynamics.

2.1.1   In-situ Measurement and Sentinel-2 for Predicting GPP

The quantification of GPP may be done via various means, including in-situ measurements and remote sensing techniques. In-situ measurements refer to the direct collection of data within the natural environment. This approach allows for accurate and immediate data acquisition about the atmospheric, terrestrial, or aquatic conditions at specific locations. Among the primary techniques for in-situ measurements of greenhouse gas exchanges, the EC method is particularly notable. The EC method and the use of closed chamber are the most widely used methods.

The EC method is particularly useful for analyzing turbulent greenhouse gas fluxes at a landscape scale [38]. Currently, measurements of CO_2 and water vapor fluxes via EC are conducted routinely on a global scale, as exemplified by networks like FLUXNET. FLUXNET is a global network of micrometeorological flux measurement sites that measure the exchanges of carbon dioxide, water vapor, and energy between the biosphere and atmosphere [39]. The network's extensive data collection, standardized processing, and integration with remote sensing and modeling efforts makes it a cornerstone for understanding and predicting ecosystem responses to environmental changes.

Complementing the earlier discussed ground-based measurements through remote sensing, the Copernicus Sentinel-2 mission, part of the European Space Agency's Earth observation program, further enhances the monitoring of Earth's ecosystems. This constellation of satellites is specifically designed to provide comprehensive and dynamic insights into Earth's vegetation and landscapes, thereby enhancing the spatial and temporal resolution of data available for ecosystem analysis. The mission features two identical satellites, Sentinel-2A and Sentinel-2B, with plans to launch Sentinel-2C in 2024, all orbiting on the same trajectory. These satellites are equipped with an innovative, wide swath, high-resolution multispectral sensor which captures data across 13 spectral bands. This advanced imaging capability offers unprecedented views of the land and detailed analyses of plant health and vitality. Sentinel-2A was launched on 23 June 2015, followed by Sentinel-2B on 7 March 2017. These satellites provide essential data for monitoring changes in land use, managing natural resources, and aiding in disaster management. With a current revisit time of every five days, the Sentinel-2 mission offers frequent monitoring capabilities [40].

Sentinel-2 data is available in two primary processing levels: Level-1C and Level-2A, each serving distinct purposes. Level-1C data offers ortho-rectified Top-Of-Atmosphere (TOA) reflectance values. This data is geometrically corrected for terrain and sensor distortions and is organized in 110x110km tiles using the UTM/WGS84 coordinate system. It includes per-pixel radiometric measurements in TOA reflectances, making this data suitable for applications that require geometrically corrected imagery but still need further atmospheric correction for most analytical purposes [41].

Level-2A data provides atmospherically corrected Bottom-Of-Atmosphere (BOA) reflectance values, now referred to as Surface Reflectance (SR). This data is corrected for atmospheric effects, ensuring more accurate surface reflectance information. The Level-2A data is resampled to consistent spatial resolutions of 10 m, 20 m, and 60 m across different spectral bands and includes additional outputs such as Aerosol Optical Thickness (AOT) maps, Water Vapor (WV) maps, and Scene Classification (SCL) maps. These classifications help categorizing pixels based on various parameters like cloud cover, vegetation, and

water. Level-2A data is ideal for direct analysis in environmental monitoring, agriculture, and land use applications. Users can also generate Level-2A data from Level-1C using the Sentinel-2 Toolbox or the Sen2Cor processor [42]. The descriptions of Sentinel-2 Bands are described in Table 2.1.

TABLE 2.1: Descriptions of Sentinel-2 Bands [43]

Name	Scale	Resolution (m)	Wavelength (nm)	Description
B1	0.0001	60	443.9	Aerosols
B2	0.0001	10	496.6	Blue
B3	0.0001	10	560.0	Green
B4	0.0001	10	664.5	Red
B5	0.0001	20	703.9	Red Edge 1
B6	0.0001	20	740.2	Red Edge 2
B7	0.0001	20	782.5	Red Edge 3
B8	0.0001	10	835.1	NIR
B8A	0.0001	20	864.8	Red Edge 4
B9	0.0001	60	945.0	Water vapor
B11	0.0001	20	1613.7	SWIR 1
B12	0.0001	20	2202.4	SWIR 2

2.2 Vegetation Indices

2.2.1 Vegetation Health Indices

Vegetation Health Indices derived from remote sensing data such as Sentinel-2 are crucial tools for estimating GPP as they reflect the photosynthetic activity and biomass of vegetation. Indices such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) are particularly useful for their ability to indicate vegetation greenness. These indices have been extensively utilized for monitoring changes in vegetation growth [44].

NDVI calculates the contrast between the red light absorbed by chlorophyll and the infrared light reflected by the leaf cellular structure [45]. NDVI tends to saturate in dense vegetation areas, therefore the Modified Normalized Difference Vegetation Index (MNDVI) was developed. MNDVI utilizes narrow bands in the red edge region, enhancing correlations with biomass as compared to standard NDVI [46]. It also includes mid-infrared that is sensitive to moisture content and to the structural integrity of plant cells, making MNDVI especially valuable for detecting plant stress that might affect internal water content and cell structure, such as frost damage [47].

The EVI is calculated using three spectral bands namely the red, near-infrared (NIR), and blue bands. EVI includes a blue band to correct for aerosol scattering in the atmosphere, making the index more reliable under various atmospheric conditions. EVI was developed to address some of the issues with NDVI, particularly its tendency to saturate in high biomass areas [48]. However, Biudes et al. [17] found that NDVI was able to capture the seasonal variation in leaf area development better than EVI. There are several types of EVI, namely EVI2. EVI2 simplifies the EVI formula by eliminating the need for a blue band, making it suitable for sensors that do not have a blue band. EVI2 has the best similarity to EVI when atmospheric effects are negligible, and atmospheric correction data (necessary for the blue band in EVI) is lacking [49]. The formula to compute NDVI,

MNDVI, EVI, and EVI2 derived from Sentinel-2 bands are as follows:

$$\text{NDVI} = \frac{\text{B8} - \text{B4}}{\text{B8} + \text{B4}}, \tag{2.4}$$

$$\text{MNDVI} = \frac{\text{B7} - \text{B5}}{\text{B7} + \text{B5}}, \tag{2.5}$$

$$\text{EVI} = \frac{\text{B8} - \text{B4}}{\text{B8} + 6\text{B4} - 7.5\text{B2} + 1}, \tag{2.6}$$

$$\text{EVI2} = \frac{\text{B8} - \text{B4}}{\text{B8} + 2.4\text{B4} + 1}, \tag{2.7}$$

where the values of the B's can be accessed from Table 2.1.

Another important index related to vegetation health is the Chlorophyll Index of Red Edge (CIr). It is widely acknowledged that chlorophyll is the universal basis for expressing photosynthetic rate in vegetation. Chlorophyll significantly influences the amount of light absorbed by vegetation [50]. CIr was proposed in [51] and retrieved from reflectance in the red edge spectral band between 700 and 720 nm and in the NIR between 760 and 800 nm using a nondestructive technique. A study in [8] shows that the combination of CIr and rainfall data has the strongest correlation with in-situ measurements of GPP. The CIr can be computed as

$$\text{CIr} = \frac{\text{B7}}{\text{B5}} - 1. \tag{2.8}$$

2.2.2 Water Content Indices

The measurement of water content in vegetation and soil is crucial for understanding ecosystem health and for managing water resources. While there is not direct measures of water mass, water content plays a vital role in assessing water availability and its impact on plant physiological processes and GPP. Remote sensing data can provide insight on water content.

A widely used index is the Land Surface Water Index (LSWI) which reflects land surface water conditions [52]. LSWI is derived from a combination of NIR and short infrared spectral band (SWIR) which is sensitive to vegetation water content and soil moisture [17]. Another significant water content index is the Normalized Difference Infrared Index (NDII) which make use of the 1610 nm band from Sentinel-2 and is a little more robust in dryland environments than the Normalized Difference Water Index (NDWI) [53]. NDWI itself, utilized in remote sensing to assess vegetation liquid water, employs the green (560nm) and NIR (865nm) bands, making it sensitive to changes in liquid water content of vegetation canopies while being less affected by atmospheric aerosol scattering than NDVI [54].

Another commonly used water content index is the Modification of Normalized Difference Water Index (MNDWI). It uses the green band and SWIR band proposed by [55] to effectively enhance and accurately extract water body information by suppressing built-up land noise and revealing subtle water features more efficiently than NDWI, especially in areas dominated by built-up land. Comparative analyses in [56] indicated that MNDWI outperforms NDWI in identifying water features intertwined with vegetation in satellite imagery. The formulas to compute MNDWI, LSWI, and NDII are as follows:

$$\text{LSWI} = \frac{\text{B8} - \text{B11}}{\text{B8} + \text{B11}}, \tag{2.9}$$

$$\text{NDII} = \frac{\text{B8} - \text{B12}}{\text{B8} + \text{B12}}, \tag{2.10}$$

$$\text{MNDWI} = \frac{\text{B3} - \text{B11}}{\text{B3} + \text{B11}}. \tag{2.11}$$

2.3 Time Series Prediction

A time series is a sequence of observations recorded at specific and equally spaced time intervals, as defined by Brockwell in [57]. It is mathematically represented as X_t, where X_t denotes the value of the series at time t. The objective of time series prediction is to estimate future values $\hat{X}_{t+h}$, where $h > 0$, using historical observations. The general mathematical formulation of this prediction is as follows:

$$\hat{X}_{t+h} = f(X_t, X_{t-1}, \ldots, X_{t-p}, \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}), \tag{2.12}$$

where $\hat{X}_{t+h}$ denotes the predicted value at time $t + h$, f is a function representing the model, X_{t-i} are past values of the series, and ϵ_{t-j} are past error terms.

For effective time series prediction, it is essential to understand the concept of stationarity. A time series is said to be stationary if its statistical properties, such as mean, variance, and autocorrelation, remain constant over time [58]. Stationarity is a crucial assumption in many time series models, including AutoRegressive (AR), Moving Average (MA) and AutoRegressive Moving Average (ARMA), because it simplifies the analysis and helps to ensure the reliability of the predictions.

There are several statistical tests available to determine the stationarity of a time series. One commonly employed test is the Augmented Dickey-Fuller (ADF) test [59]. The null hypothesis of the ADF test states that the time series possesses a unit root, meaning it is non-stationary. On the other hand, the alternative hypothesis suggests that the series is indeed stationary. The ADF test involves estimating the following regression:

$$\Delta X_t = \mu + \nu t + \xi X_{t-1} + \rho_1 \Delta X_{t-1} + \cdots + \rho_p \Delta X_{t-p} + \epsilon_t. \tag{2.13}$$

Here, $\Delta X_t = X_t - X_{t-1}$ is the first difference of the series, μ is a constant, νt is a time trend, and ξ, ρ_i are coefficients. The test focuses on the coefficient ξ. If ξ is significantly different from zero, the null hypothesis of a unit root is rejected, indicating that the series is stationary. If a time series is not stationary, differencing can be applied to make it stationary [60]. Differencing involves subtracting the current value of the series from its previous value. The first-order difference of a time series $\{X_t\}$ is given by

$$\Delta X_t = X_t - X_{t-1} = (1 - B)X_t, \tag{2.14}$$

where B is the backshift operator, $BX_t = X_{t-1}$. Higher-order differencing can be used if needed.

2.4   SARIMAX

The AR and MA components are fundamental to understanding time series analysis, leading to the development of the SARIMAX model, which integrates these components into a comprehensive modeling framework. The AR component models the current value of the series as a linear function of its previous values [60]. This can be expressed as

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t, \tag{2.15}$$

where X_t denotes the time series value at time t, ϕ_i are the coefficients representing the influence of past values, p is the order of the AR model, and ϵ_t is the white noise error term. In contrast, the MA component models the current value of the series as a function of the past error terms [61]. This relationship is expressed as

$$X_t = \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}. \tag{2.16}$$

Here, θ_j are the coefficients measuring the impact of past forecast errors ϵ_{t-j} on the current value, and q is the order of the MA model, representing the number of past errors included in the model. Combining these AR and MA components results in the ARMA model. The ARMA model captures both the influence of past values and the impact of past errors, making it particularly useful for time series that exhibit characteristics of both AR and MA processes. The mathematical representation of an ARMA model is given by

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t. \tag{2.17}$$

In this equation, c is the constant term, X_t is the time series value, ϕ_i and θ_j are the coefficients, and p and q are the orders of the AR and MA components, respectively. If the series exhibits non-stationarity, differencing is applied first to make the series stationary, leading to an ARIMA model. The ARIMA model incorporates an integration component represented using the backshift operator B. The ARIMA model is expressed as

$$(1 - B)^d X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t. \tag{2.18}$$

Here, $(1 - B)^d$ represents the differencing operator, where d is the order of differencing needed to make the series stationary. The terms c, ϕ_i, θ_j, p, q, and ϵ_t retain their previous definitions.

Many real-world time series exhibit seasonal patterns that repeat at regular intervals. To capture these seasonal effects, the ARIMA model is extended to the Seasonal ARIMA (SARIMA) model. Seasonal differencing is used to remove seasonal trends, defined as

$$(1 - B^s) X_t = X_t - X_{t-s}, \tag{2.19}$$

where s is the seasonal period. The SARIMA model incorporates both non-seasonal and seasonal components as in Equation (2.20). In this equation, $(1 - B)^d$ represents the non-seasonal differencing operator, $(1 - B^s)^D$ is the seasonal differencing operator with seasonal period s and seasonal differencing order D. The terms Φ_I and Θ_J are the seasonal AR and MA coefficients, respectively, with orders P and Q. The seasonal terms X_{t-Is} and ϵ_{t-Js} capture the seasonal lags, where X_{t-Is} represents the value of the time series from

I seasonal periods ago, and ϵ_{t-Js} represents the error term from J seasonal periods ago. The overall model is

$$(1-B)^d(1-B^s)^D X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{I=1}^{P} \Phi_I X_{t-Is} + \sum_{J=1}^{Q} \Theta_J \epsilon_{t-Js} + \epsilon_t. \quad (2.20)$$

While SARIMA effectively models time series with seasonal patterns, it does not account for external factors that might influence the series. To address this, SARIMA can be extended to the SARIMAX model, which incorporates exogenous variables representing external influences. Let Z_t as a vector consists of k exogenous variables at time t. Each element in Z_t, from Z_{1t} to Z_{kt}, corresponds to a different external factor that may influence the dependent variable of the time series. The column vector form of Z_t is represented mathematically as

$$Z_t = \begin{bmatrix} Z_{1t} \\ Z_{2t} \\ \vdots \\ Z_{kt} \end{bmatrix}. \quad (2.21)$$

Associated with these exogenous variables is a vector of coefficients β, which quantifies the influence of each variable in Z_t on the time series. This vector β aligns with the elements of Z_t, providing a weight or impact level for each exogenous variable. The coefficient vector β is shown as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}. \quad (2.22)$$

Together, $\beta^\top Z_t$ represents the linear combination of the exogenous variables and their coefficients:

$$\beta^\top Z_t = \beta_1 Z_{1t} + \beta_2 Z_{2t} + \cdots + \beta_k Z_{kt}, \quad (2.23)$$

incorporating the term $\beta^\top Z_t$, the SARIMAX model [62] is defined by

$$(1-B)^d(1-B^s)^D X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{I=1}^{P} \Phi_I X_{t-Is} + \sum_{J=1}^{Q} \Theta_J \epsilon_{t-Js} + \beta^\top Z_t + \epsilon_t. \quad (2.24)$$

The inclusion of exogenous variables, such as $\beta^\top Z_t$, allows the SARIMAX model to account for external influences on the time series. This capability makes the model particularly robust and adaptable to real-world scenarios where external factors significantly influence outcomes. To clarify the parameter substitution process in the SARIMAX model, a hypothetical example with defined parameters is provided in Appendix A1.

Selecting appropriate parameters for the SARIMAX model is crucial for accurate modeling and forecasting. This process involves determining the orders of the AR, MA, seasonal AR, and seasonal MA components, as well as the differencing orders. Tools like the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots can help in this selection process [63]. The PACF plot, in particular, is useful for identifying the

order of the AR component. The PACF measures the correlation between a time series and its lagged values, after removing the effects of intermediate lags. Significant spikes in the PACF plot indicate the appropriate lag order p for the AR component.

Similarly, the ACF plot helps to identify the order of the MA component. The ACF measures the correlation between a time series and its lagged values without removing intermediate effects. Significant spikes in the ACF plot suggest the appropriate lag order q for the MA component. For seasonal components, the ACF and PACF plots can be examined at lags that are multiples of the seasonal period s. This helps to determine the seasonal AR and MA orders P and Q.

2.5 XGBoost

A decision tree is a hierarchical structure used for decision making. It splits the data into subsets based on certain criteria at each node. The objective is to split the population into homogeneous sets based on the most significant input variables. The decision tree can be applied to both classification and regression tasks. For classification tasks, the target variable is typically discrete, while for regression tasks (also called regression trees), the target variable is continuous [64].

A tree structure consists of a root node, internal nodes, and terminal nodes (leaves) [65]. The root node constitutes the initial point from which the top-down population is recursively partitioned into two or more homogeneous subsets. Each internal node embodies a "test" or "decision" on an attribute, with each branch representing the outcome of the test. This process, known as splitting, divides a node into two or more sub-nodes. When a sub-node further splits into additional sub-nodes, it is referred to as a decision node. Terminal nodes, or leaves, are nodes that are not further split and represent the final outcome of the tree.

Mathematically, a decision tree can be represented as a function $f_t(x)$, where t denotes the iteration number and x represents the input features. The function $f_t(x)$ is constructed by recursively partitioning the input space and fitting a simple prediction model, such as a constant value, within each partition. The general form of $f_t(x)$ can be written as:

$$f_t(x) = \sum_{v=1}^{V} w_v I(x \in R_v), \tag{2.25}$$

where V is the number of nodes in the tree, w_v is the prediction value assigned to the v-th node, R_v represents the region (partition) corresponding to the v-th node, and $I(x \in R_v)$ is an indicator function that evaluates to 1 if x falls into the region R_v, defined as

$$I(x \in R_v) = \begin{cases} 1 & \text{if } x \text{ is in region } R_v, \\ 0 & \text{otherwise.} \end{cases} \tag{2.26}$$

This function essentially checks if the input feature vector x belongs to the region R_v defined by the node v. For regression tasks, this w_v is typically the mean of the target values y for the data points that fall into the region R_v, which can be computed as:

$$w_v = \frac{1}{|R_v|} \sum_{(x_i, y_i) \in R_v} y_i, \tag{2.27}$$

where $|R_v|$ is the number of data points in region R_v, and y_i are the target values of these data points. Each region R_v is defined by a series of splits based on the features x.

For example, a split might partition the data based on whether a particular feature x_j is less than or greater than a threshold θ. In the context of time series, features x_j could include past values of the target variable, moving averages, or other lagged predictors. This partitioning continues recursively until the stopping criteria, such as max_depth or min_child_weight, are met. These criteria are used to control the tree's growth and prevent overfitting. Overfitting occurs when a model learns patterns specific to the training data, resulting in high accuracy on the training set but poor performance on unseen test data [66, 67].

The max_depth controls the maximum number of levels in the tree, which limits its complexity. A larger max_depth allows the model to capture more intricate patterns in the data but increases the risk of overfitting. Conversely, a smaller max_depth reduces the risk of overfitting but may lead to underfitting if the model is too simple.

The min_child_weight specifies the minimum sum of sample weights needed in a child node. It helps prevent the model from learning overly specific patterns by requiring that any node that is split must have a substantial amount of data. This ensures that each split results in meaningful partitions and mitigates overfitting by avoiding the creation of nodes with too few data points.

A single tree is usually not sufficient for building a complex model, therefore models such as XGBoost, which uses decision trees as weak learners, were proposed for better performance. XGBoost was proposed to create a scalable, efficient, and high-performance tree boosting system that addresses limitations of existing methods, particularly in handling large datasets and improving model accuracy in diverse machine learning challenges [32].

According to [68], XGBoost is able to build a robust classifier using weak classifiers and has several advantages over other gradient boosting algorithms, including efficiently handling missing values, preventing overfitting, and decreasing running time through parallel and distributed calculations. XGBoost employs a gradient boosting framework, functioning as an ensemble of decision trees constructed sequentially [69]. Gradient boosting is a powerful ensemble technique that improves predictive accuracy by sequentially adding weak learners, typically decision trees, to an ensemble.

The construction of an XGBoost model begins with a simple initial model and iteratively adds decision trees to improve predictions by minimizing a specified loss function. The initial model $F_0(x)$ is typically a simple estimator of the output, often chosen as the mean of the target variable y. This initial prediction is set to minimize the initial loss across all training samples, and it is mathematically defined as

$$F_0(x) = \arg\min_\alpha \sum_{i=1}^{n} L(y_i, \alpha), \tag{2.28}$$

where α is a constant representing the initial prediction, $L(y_i, \alpha)$ is the loss function measuring the error between the actual target value of the i-th training sample y_i, and n is the number of training samples. Once the initial model is set, XGBoost improves the model iteratively by adding new decision trees. At each iteration t, the model is updated by adding a new tree $f_t(x)$, leading to the updated model

$$F_t(x) = F_{t-1}(x) + f_t(x). \tag{2.29}$$

To simplify the notation, let $\hat{y}_i^{(t)}$ denote the prediction for sample i at iteration t. Thus, $\hat{y}_i^{(t)} = F_t(x_i)$ and the model update becomes

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i). \tag{2.30}$$

Here, $\hat{y}_i^{(t-1)}$ is the model's prediction after $t - 1$ iterations, and $f_t(x_i)$ is the new tree added at iteration t to correct the errors from $\hat{y}_i^{(t-1)}$. To effectively determine the contribution of each new tree $f_t(x)$, XGBoost calculates the gradient and Hessian of the loss function with respect to the previous predictions. For each training sample i, the gradient g_{it} and the Hessian h_{it} at iteration t are defined as

$$g_{it} = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \tag{2.31}$$

and

$$h_{it} = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}, \tag{2.32}$$

respectively.

In these equations, g_{it} represents the gradient of the loss function with respect to the predictions at sample i during iteration t, indicating the direction and magnitude of the error for each prediction. The Hessian h_{it} represents the curvature of the loss function, indicating how sensitive the loss is to changes in the prediction. These derivatives provide detailed guidance on how to adjust the model to reduce the loss effectively.

The concept of pseudo-residuals is central to gradient boosting. Pseudo-residuals are the negative gradients of the loss function with respect to the predictions from the current model. They represent the necessary adjustments needed to correct the predictions. For sample i at iteration t, the pseudo-residual is given by

$$r_{it} = - \left[\frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \right]_{\hat{y}_i^{(t-1)}} = -g_{it}. \tag{2.33}$$

These pseudo-residuals r_{it} serve as the target values for training the new tree $f_t(x)$. To approximate the change in the loss function due to the addition of the new tree $f_t(x)$, XGBoost uses a second-order Taylor expansion around the previous predictions $\hat{y}_i^{(t-1)}$:

$$L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx L(y_i, \hat{y}_i^{(t-1)}) + \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} f_t(x_i)$$

$$+ \frac{1}{2} \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} f_t(x_i)^2 \tag{2.34}$$

$$= L(y_i, \hat{y}_i^{(t-1)}) + g_{it} f_t(x_i) + \frac{1}{2} h_{it} f_t(x_i)^2. \tag{2.35}$$

In this expansion, the term $L(y_i, \hat{y}_i^{(t-1)})$ is the current loss (from the previous iteration) and is constant with respect to $f_t(x_i)$. Therefore, it does not influence the optimization process of $f_t(x_i)$ and can be omitted from the final objective function.

XGBoost extends gradient boosting by introducing a regularization component to the optimization problem, thus enhancing model performance and control overfitting. The XGBoost objective function includes a regularization term that penalizes the complexity of the trees [70]. The regularization term is

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2, \tag{2.36}$$

where T is the number of leaves in the tree f_t, w_j are the weights associated with the leaves, γ is a parameter that penalizes the number of leaves, and λ is a parameter that penalizes the magnitude of the leaf weights. Combining the approximated loss with a regularization term, the objective function for the t-th iteration becomes

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} \left[g_{it} f_t(x_i) + \frac{1}{2} h_{it} f_t(x_i)^2 \right] + \Omega(f_t). \tag{2.37}$$

The summation aggregates the contributions of all training samples, ensuring that the tree $f_t(x)$ improves predictions across the entire dataset. Finally, the combined objective function that XGBoost minimizes at each iteration is

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} \left[g_{it} f_t(x_i) + \frac{1}{2} h_{it} f_t(x_i)^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2. \tag{2.38}$$

This detailed formulation and systematic approach show how XGBoost incrementally builds a robust model by adding trees that correct residuals from previous iterations, using gradients and Hessians to guide the optimization, and regularization to prevent overfitting. By focusing on the incremental improvements and penalizing complexity, XGBoost effectively balances model accuracy and generalizability.

XGBoost offers the advantage of interpretability by calculating feature importance. Feature importance helps in understanding how different variables impact the model's predictions. XGBoost determines feature importance through three key metrics: weight, gain, and cover [71].

The weight term indicates the number of times a feature is used to make splits in nodes across all the trees in a model. A higher weight suggests that the feature is crucial for decision-making or predictions due to its frequent use.

Cover refers to the average number of samples affected when a feature is used to make a split. It provides insight into the extent of data influenced when a particular feature is used in node splitting. The cover metric considers the second order gradient of the training data, which relates to the impact of changes in the model's predictions on the overall error or accuracy when this feature is utilized.

Gain measures the effectiveness of a feature in improving the model's accuracy. It quantifies the average enhancement in the loss function when a feature is used to split nodes. By analyzing these metrics, one can evaluate which features are most influential and how they contribute to the predictive accuracy of the model, offering a comprehensive understanding of the role each feature plays in the model's outcomes.

2.6 LSTM

Before delving into the specifics of Long Short-Term Memory (LSTM) networks, it is essential to understand the foundational concepts of neural networks and recurrent neural networks (RNNs). These concepts provide the necessary background to appreciate the advancements and functionalities of LSTM networks.

Neural networks are computational models inspired by the human brain's neural structure. Each neural network consists of neurons organized into layers: the input layer, hidden layers, and the output layer [72]. Neurons in a layer are connected to neurons in subsequent layers by weights w_{ij}, where i and j are indices representing the neurons in the previous and current layers, respectively. A neuron computes a weighted sum of its inputs, adds a bias term b_j, and applies an activation function ϕ. The output y_j of neuron j is given by

$$y_j = \phi \left(\sum_i w_{ij} x_i + b_j \right), \tag{2.39}$$

where x_i represents the input from neuron i in the previous layer. Common activation functions include the sigmoid function, the hyperbolic tangent (tanh) function, and the rectified linear unit (ReLU).

Recurrent Neural Networks (RNNs) extend traditional neural networks by incorporating connections that form cycles, enabling them to maintain hidden states that capture information from previous timesteps [73]. This recurrent connection allows RNNs to model sequential data. The hidden state h_t at time step t is computed from the current input x_t and the previous hidden state h_{t-1}:

$$h_t = \phi(W_{hh} h_{t-1} + W_{xh} x_t + b_h), \tag{2.40}$$

where W_{hh} and W_{xh} are weight matrices representing the connections from the hidden state and input at the previous timestep to the hidden state at the current timestep, respectively. b_h is a bias vector, and ϕ is the activation function.

RNNs employ Backpropagation Through Time (BPTT) for training. The performance of RNNs is optimized using a loss function, which measures the discrepancy between predicted outputs and actual targets. This loss function guides the backpropagation process. For regression tasks, the Mean Squared Error (MSE) is typically utilized as the loss function. During training, the network is unrolled through time, and the gradients of the loss function with respect to each parameter are computed. These gradients are subsequently used to update the weights and biases through an optimization algorithm such as gradient descent.

Despite their effectiveness, RNNs often encounter difficulties in learning long-term dependencies due to vanishing and exploding gradient problems. In standard RNNs, the hidden state at each time step is computed recursively by multiplying the hidden state from the previous time step by a weight matrix and adding the current input. This recursive process can result in vanishing or exploding gradients during BPTT, as the gradients are products of many small or large terms.

Consider the gradient of the loss function $\mathcal{L}$ with respect to the weight matrix W. For simplicity, assume the recurrent connection is unrolled over t' time steps:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{t=1}^{T} \left(\frac{\partial \mathcal{L}}{\partial h_t} \prod_{k=t}^{T} \frac{\partial h_k}{\partial h_{k-1}} \frac{\partial h_{k-1}}{\partial W} \right), \tag{2.41}$$

Here, $\frac{\partial \mathcal{L}}{\partial W}$ denotes the gradient of the loss function with respect to the weight matrix W, and h_t represents the hidden state at time step t. The term $\prod_{k=t}^{T} \frac{\partial h_k}{\partial h_{k-1}}$ represents the product of gradients of the hidden states from time step t to t'. If these gradients are less than one, their product can shrink exponentially, leading to vanishing gradients. Conversely, if they are greater than one, their product can grow exponentially, leading to exploding gradients.

LSTM networks were introduced to address the limitations of standard RNNs by Hochreiter and Schmidhuber (1997) [34]. Specifically, LSTMs were designed to mitigate the issues of vanishing and exploding gradients by incorporating a more complex architecture that allows better control over the flow of information. Error flow within the LSTM is managed using 'gates' which regulate the flow of information alongside the hidden state, used in standard RNNs [74]

An LSTM block commonly incorporates a memory cell, an input gate, an output gate, and a forget gate, alongside the hidden state utilized in standard RNN [34]. The overall structure is depicted in Figure 2.1.

The forget gate, decides which parts of the cell state are no longer needed [75]. The gate combines the current input $x_j^{(t)}$, the previous hidden state $h_j^{(t-1)}$, and a bias term b_i^f, applying a sigmoid function to produce values between 0 and 1 for each number in the cell state. The forget gate is
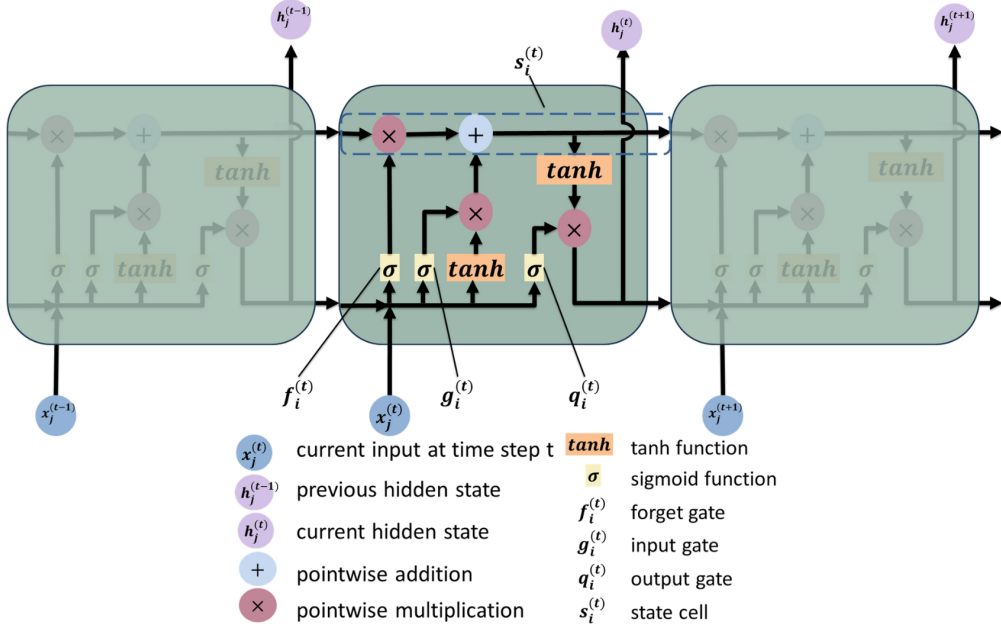

FIGURE 2.1: Structure of LSTM blocks. The figure illustrates the flow of information through the forget gate, input gate, cell state update, and output gate.

$$f_i^{(t)} = \sigma \left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)} \right). \tag{2.42}$$

These values effectively determine the proportion of each component of the cell state that should be carried forward. Concurrently, the input gate $g_i^{(t)}$ identifies new information to be added to the cell state [76]. It also uses a sigmoid function to filter incoming data, deciding which values are important to keep:

$$g_i^{(t)} = \sigma \left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)} \right). \tag{2.43}$$

The candidate cell state $\tilde{s}_i^{(t)}$ is created using the tanh function, which ensures the values are in the range of -1 to 1:

$$\tilde{s}_i^{(t)} = \tanh \left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)} \right). \tag{2.44}$$

The cell state $s_i^{(t)}$ is then updated by an operation that combines the output of the forget gate and the input gate. The forget gate output scales the previous cell state, while the input gate scales the new candidate values that are added to the state:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \tilde{s}_i^{(t)} \tag{2.45}$$

The output gate $q_i^{(t)}$ determines the final output of the LSTM unit at time step t. It filters the cell state through another sigmoid function to decide which parts of the cell state will be used to generate the output hidden state $h_i^{(t)}$:

$$q_i^{(t)} = \sigma \left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)} \right) \tag{2.46}$$

The output $h_i^{(t)}$ is then computed as the product of the output of the tanh function applied to the cell state and the output of the output gate:

$$h_i^{(t)} = \tanh(s_i^{(t)}) \cdot q_i^{(t)} \tag{2.47}$$

The parameters b, U, and W represent biases and weights that are learned during the training process. These are critical for adapting the LSTM's performance to specific sequences or data patterns, allowing it to effectively capture long-term dependencies in time-series data or sequential inputs [77].

Similar to RNNs, LSTM networks also use BPTT to train the network [78]. During training, the loss function is minimized over multiple epochs. An epoch refers to one complete pass through the entire training dataset. Monitoring the training loss across epochs helps ensure that the model is learning and not overfitting. Adjusting hyperparameters such as the learning rate and batch size can help improve training performance.

This mechanism of employing gates allows LSTMs to effectively mitigate the vanishing and exploding gradient problems, making them particularly well-suited for tasks involving long-term dependencies. This architecture and its parameterization allow LSTMs to perform exceptionally well in tasks involving sequence prediction and time-series forecasting, where understanding long-term patterns is essential.

Chapter 3

Methodology

This chapter describes the methodology used in this study to develop a machine learning model to predict GPP by combining in-situ and satellite data. An overview of the process is shown in Figure 3.1. The methodology is divided into four main parts. Data Collection, Data Preparation, Model Development, and Model Evaluation. The Model Development stage consists of two stages: site-specific or individual model, and unified or global model. The process is explained in more detail in the following sections.
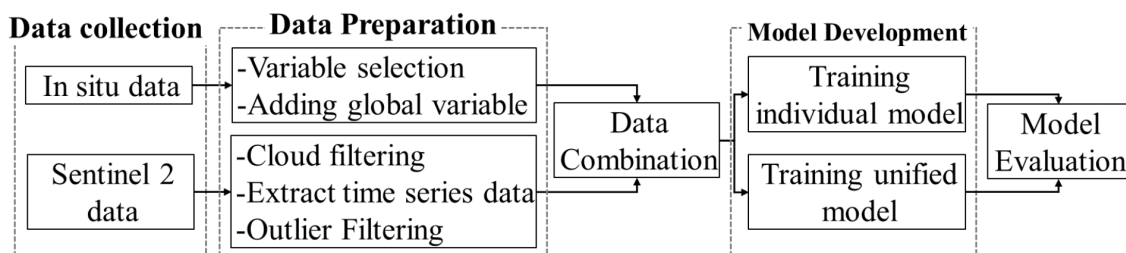
FIGURE 3.1: Diagram Overview

3.1 Area of Interest

The objective of this study is to develop a unified model capable of predicting GPP in various sites across Europe. As an initial step, a base model has been be developed and trained using different types of ecosystem data available in The Integrated Carbon Observation System (ICOS) [79]. ICOS is a European-wide greenhouse gas research infrastructure that provides standardized data on greenhouse gas concentrations in the atmosphere, as well as on carbon fluxes between the atmosphere, land, and oceans.

The chosen ecosystem sites include evergreen needleleaf forest (ENF), deciduous broadleaf forest (DBF), cropland, and grassland. These selections are based on several considerations. The most important requirement is the amount of available data. Only sites with a minimum of 2 years of data records were considered to ensure there was enough data for the model training process. Another consideration is the selection of different sites and locations to achieve sparsely distributed geographies both horizontally and vertically, especially for training sites. There are two sets of site combinations, with one set serving as training sites for building the model for individual and unified sites, while the other set serves as testing sites to evaluate the performance of the unified model in predicting GPP on unseen sites. However, both sets will have the same ecosystem types group as the model is only trained on those ecosystems. The selected sites for this study are shown
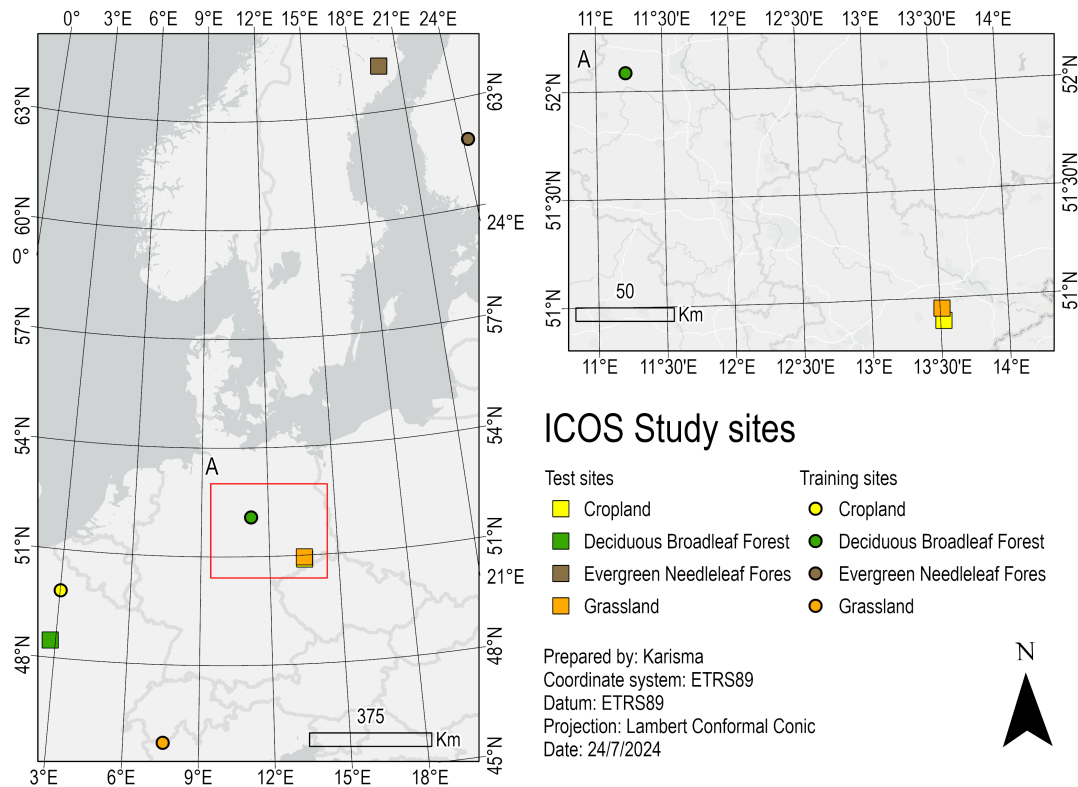
FIGURE 3.2: Map of ICOS stations used in this study, with training sites marked by squares and test sites by circles. Different ecosystems are indicated by color: cropland (yellow), DBF (green), ENF (brown), and grassland (orange).

in Figure 3.2. The training site consists of Estrees-Mons A28, Hohez Holz, Hyytiala, and Torgnon while the testing sites are Klingenberg, Fontainebleau-Barbeau, Svartberget, and Grillenburg. The site description is available in the subsequent subsections.

3.1.1 Training Sites

Estrees-Mons A28

The Estrees-Mons A28 station, bearing the ICOS code FR-EM2, is situated in the Picardie region of France, approximately 200 km northeast of Paris. This station is part of a National Institute of Agricultural Research (INRA) research station encompassing 150 ha dedicated to experimental arable farming. The experiments include both annual food crops and perennial bioenergy crops such as miscanthus and switchgrass. Positioned on deep loamy soil, the station operates within a marine west coast climate zone.

Geographically, the station is located at a latitude of 49.87211 and a longitude of 3.02065, at an elevation of 85 m. It experiences an average annual temperature of 10.8°C, receives an average annual precipitation of 680 mm, and observes an average annual incoming shortwave radiation of 125 Wm^{-2}. This ICOS Associated Ecosystem Station is instrumental in advancing research on cropland management and sustainability, particularly focusing on the potential of perennial bioenergy crops in agricultural practices [80]. The flux tower of Estrees-Mons A28 is shown in Figure 3.3.

FIGURE 3.3: The FR-EM2 tower taken from [80].

Hohes Holz

The Hohes Holz station, marked by the ICOS code DE-HoH, is located at the northern boundary of the Bode water catchment area in Germany. This ICOS station was labeled on 22 May 2019, and resides within a marine west coast climate zone. The site is an alluvial forest characterized by a diverse assembly of dominant tree species including European beech (*Fagus sylvatica L.*), sessile oak (*Quercus petraea*), silver birch (*Betula pendula*), European hornbeam (*Carpinus betulus*), with plantations of Norway spruce (*Picea abies*) and European larch (*Larix decidua*). The soil type here is classified as Luvisoles.
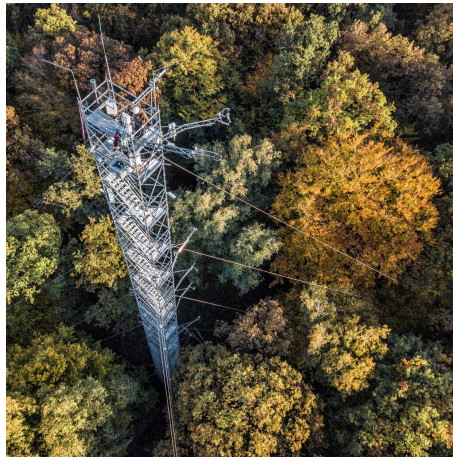

FIGURE 3.4: The DE-HoH tower taken from [81].

Located at a latitude of 52.08656 and a longitude of 11.22235, Hohes Holz stands at an elevation of 193 m. It experiences an average annual temperature of 9.1°C, and an average annual precipitation of 563 mm. This station is pivotal for research in deciduous broadleaf forests, focusing on ecosystem responses and dynamics within a temperate climate context [81]. The flux tower of Hohes Holz is shown in the Figure 3.4.

Hyytiala

The Hyytiala (FI-Hyy) Forestry Field Station is located in Finland, 220 km northwest of Helsinki. It operates under ICOS and is managed by the University of Helsinki. The

station is situated in a relatively uniform Scots pine (*Pinus sylvestris L.*) forest with slightly hilly terrain. Its exact coordinates are 61.8474 latitude and 24.2947 longitude, and it is positioned at an elevation of 181 m above sea level. The climate of the area is classified as subarctic, and the primary ecosystem is Evergreen Needleleaf Forests (ENF). The average annual temperature is 3.5°C, the average annual precipitation is 711 mm, and the average annual incoming shortwave radiation is 100 W/m^2 [82]. The flux tower of Hyytiala is shown in the Figure 3.5.

FIGURE 3.5: The FI-Hyy tower taken from [82].

Torgon

The Torgnon station, with the code IT-Tor, is situated in the small village of Torgnon in Valle d'Aosta, nestled in the western Alps of Italy. This station is part of ICOS. It lies within the subarctic climate zone, on abandoned grassland dominated by Nardus stricta in the European Alps. The site is positioned at an altitude of 2168 m above sea level, with geographical coordinates of 45.8444 latitude and 7.57805 longitude. The climate at this location has an average annual temperature of 3.27 °C, an average annual precipitation of 945 mm, and an average annual incoming shortwave radiation of 165 W/m^2 [83]. The flux tower of Torgnon is shown in the Figure 3.6.

FIGURE 3.6: The IT-Tor tower taken from [83].

3.1.2 Testing Sites

Klingenberg

The Klingenberg station, designated with the ICOS code DE-Kli, is located in the mountain foreland of the Erzgebirge in Germany. Klingenberg is set in a distinctively flat area, with the exception of a moderately sloped western direction, at an elevation of 478 m. The site, marked by its coordinates at 50.89306°N, 13.52238°E, has been conducting EC measurements since May 2004, with the reference setup established at a height of 3.5 m above ground. Klingenberg functions as an intensively managed agricultural site, operating a 5-year crop rotation system that includes rapeseed, winter wheat, forage maize, spring barley, winter barley, and occasionally, a catch crop [84]. The flux tower of Klingenberg is shown in the Figure 3.7

FIGURE 3.7: The DE-Kli tower taken from [84].

This region is characterized by a humid continental climate with warm summers , which supports the main ecosystem of croplands. The area experiences a mean annual temperature of 8.2°C, average annual precipitation of 756 mm, and receives an average annual incoming shortwave radiation of 131 Wm^{-2}. This comprehensive setup allows for detailed study and understanding of agricultural ecosystems in suboceanic or subcontinental climates.

Fontainebleau-Barbeau

The Fontainebleau-Barbeau station, designated by the ICOS code FR-Fon, is located approximately 50 km southeast of Paris, France. It is situated within the Atlantic Biogeographical Region and is characterized by a deciduous oak forest ecosystem, making it a prime example of temperate oceanic climate zones, representative of a large northwestern part of Europe. The forest at Fontainebleau-Barbeau is predominantly made up of sessile oak (*Quercus petraea*) and English oak (*Quercus robur*), with an understory of hornbeam coppice. The station, located at latitude 48.476357 and longitude 2.780096, at an elevation of 103 m, experiences a mean annual temperature of 11.44°C. It receives an average annual precipitation of 678.99 mm and an average annual incoming shortwave radiation of 134.97 Wm^{-2}. This setup facilitates comprehensive studies on deciduous broadleaf forest dynamics, ecology, and climate interactions in a marine west coast climate zone [85]. The Fontainebleau-Barbeau flux tower is shown in Figure 3.8.

FIGURE 3.8: The FR-Fon tower taken from [85].

Svartberget

The Svartberget station, identified by the ICOS code SE-Svb, is situated approximately 70 km west of the Gulf of Bothnia, northwest of Umeå in northern Sweden. This class 2 station in the ICOS network was officially labeled on May 22, 2019. The station is deeply embedded within the Svartberget Experimental Forest, which includes about 1520 ha of forest land dedicated to research since 1909. Located in a boreal forest, the area is characterized by a landscape of ridges, valleys, and lakes that extend from the northwest to the southeast. The predominant vegetation includes evergreen needleleaf forests, making it a vital site for studying subarcticecosystem dynamics.

FIGURE 3.9: The SE-Svb tower taken from [86].

Geographically, the station is located at latitude 64.25611 and longitude 19.7745, with an elevation of 267 m. The climate is subarctic, with a mean annual temperature of just 1.8°C, annual precipitation averaging 614 mm, and an average annual incoming shortwave radiation of 93.4 Wm^{-2}. These conditions make Svartberget an ideal location for observing and understanding the complex interactions within boreal forest ecosystems under cold climate pressures [86]. The Svartberget flux tower is shown in Figure 3.9.

Grillenburg

The Grillenburg station, identified by ICOS code DE-Gri, is situated in Germany. This 40 ha permanent grassland has been unfertilized since 1987 and is managed for fodder and hay production, with occasional grazing. The flora includes couch grass, meadow foxtail, yarrow, common sorrel, and white clover. The soil is a mix of silty and clayey loam. It is located in a suboceanic or subcontinental climate zone, the station's environment is marked by restricted fetch due to surrounding forests. Positioned at latitude 50.95004 and longitude 13.51259, with an elevation of 385 , it experiences a humid continental climate with a mean annual temperature of 8.6°C, precipitation of 872 mm, and incoming shortwave radiation of 126 Wm^{-2}. This provides a key site for studying grassland ecosystems under temperate climate conditions [87]. The Grillenburg flux tower is shown in Figure 3.10.
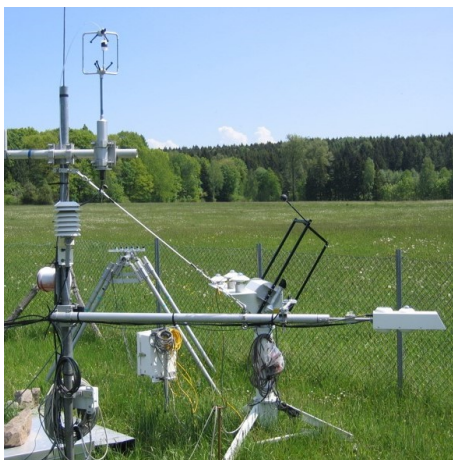

FIGURE 3.10: The DE-Gri tower taken from [87].

3.2 Data Collection

3.2.1 in-situ Measurements

As mentioned before, the in-situ data for this study was gathered from ICOS [79]. In particular, the ARCHIVE data product for ecosystem measurements was used in this study. The carbon fluxes and meteorological measurements provided are processed using the standard FLUXNET procedure of the FLUXNET2015 collection [88]. The archive consists of various in-situ measurements with different time intervals including half-hourly, daily, weekly, monthly, and yearly data, each complete with descriptions of the variables.

This study utilized the daily data provided from the ARCHIVE data product, which has been gap-filled using multiple methods. While the archive data product contains numerous variables, only variables that were available at all chosen study sites were used. These are shown in Table 3.1. Most of the variables used were gap-filled time series with good quality flags (as indicated by the '_F suffix')

The GPP_DT_ VUT_USTAR50 was chosen as the target variables for the prediction task. This value was obtained by applying a spike detection algorithm using multiple friction velocity (USTAR) and taking its median. The _VUT represent the USTAR threshold are found for each year and _DT means the data was partitioned using daytime method. More detailed information regarding the processing step of the data can be found in [88].

TABLE 3.1: in-situ variables that were used in this study.

Variables	Explanation	Unit
H_F_MDS	Sensible heat turbulent flux	Wm^{-2}
LE_F_MDS	Latent heat turbulent flux	Wm^{-2}
LW_IN_F	incoming (down-welling) longwave radiation	Wm^{-2}
PA_F	Atmospheric Pressure	kPa
P_F	Precipitation	mmd^{-1}
SW_IN_F	Shortwave radiation	Wm^{-2}
TA_F	Air Temperature	°C
VPD_F	Vapor Pressure saturation Deficit	hPa
WS_F	Wind speed	ms^{-1}
GPP_DT_VUT_USTAR50	Gross Primary Production	$gCm^{-2}d^{-1}$

3.2.2 Remote Sensing Data

Sentinel-2 were used in this study to derive the VIs, which are essential for estimating GPP. The study in [48] shows that VIs related to greenness were more effective during the wet phase of the growing season, whereas indices related to water content performed better in the dry phase as they are more sensitive to water availability than the other vegetation indices. Therefore, both vegetation related to greenness and water were used in this study. The VIs used in this study were NDVI, MNDVI, EVI, EVI2, CIr, LSWI, MNDWI, NDII. Additionally, the original spectral bands from the Sentinel-2 satellite were collected to facilitate the derivation of further indices if required.

To speed up the preprocessing process, Google Earth Engine was used to preprocess the data. The Level 2A data Sentinel-2 images were retrieved with defined locations and time periods. The time period used was based on the available data from the in-situ data since the ground truth data are from in-situ measurements, while the location corresponded to the coordinates of the flux tower with a certain buffer area. The buffer area was defined as a circle area with a certain radius with the coordinate of the flux tower as a center. The radius of the buffer area varied for each site to ensure that the retrieved data covered only the desired ecosystem. The radius per each site was defined manually by checking the satellite image for each area of interest.

Firstly, any available images in the defined coordinates within the defined time period were retrieved. Thereafter, these were filtered based on cloud coverage. Specifically, the 'Q60' quality band was used to mask out pixels affected by opaque and cirrus clouds. This was followed by calculating the cloud-covered area and the total area ratio, and performing cloud filtering of 30%, meaning that images with cloud coverage of more than 30% were excluded. Finally, non-vegetation pixels were removed from the images by masking pixels with 'SCL' band values different to 4 [89].

After applying the cloud filter, the pixel value within the buffer area of the selected images was averaged so that only one daily value, which had images, was obtained. The process was followed by interpolation to fill the gap dates for which there were no available images. This resulted in a daily time series data for each variable retrieved from the Sentinel-2 data for each site. To illustrate, for the Torgon site, the available images between January 1, 2016 until December 31, 2019 were 341. After cloud filtering by 30%, 115 images were collected, resulting in 115 data points, which were gap-filled to create full daily data.

In the data processing workflow, the presence of noise observed was first addressed. To mitigate the impact of outliers on subsequent analyses, a z-score technique was employed,

retaining only data within the range of $-2\sigma_{\text{data}}$ to $+2\sigma_{\text{data}}$, which effectively excluded extreme values from the dataset. Here, σ_{data} represent the standard deviation of the data. Linear interpolation was then performed to estimate missing or unrecorded data points. Following interpolation, a Savitzky-Golay filter [90] was applied to smooth the time series, enhancing the clarity and quality of the data for further analysis.

3.3 Data Preprocessing

This section outlines the additional data preprocessing that was performed after collecting both in-situ and Sentinel satellite data. The preprocessing involved integrating the in-situ measurements with the corresponding Sentinel observations for each site. Specifically, this merging process entailed aligning the datasets by matching the observation dates from both sources for each location. The datasets were merged to ensure that each in-situ data point corresponded directly to a satellite observation on the same date. The completeness and availability of the data for each site after this matching process are detailed in Table 3.2.

TABLE 3.2: Availability of Data after Matching Process

	Site	Start date	End date	Records
Training site	Estrees-Mons A28	26 May 2017	15 October 2023	2334
	Hohes Holz	1 January 2019	26 September 2023	1730
	Hyytiala	1 January 2018	22 September 2023	2091
	Torgon	6 July 2017	31 December 2023	2370
Testing site	Klingenberg	1 January 2018	25 September 2023	2093
	Font. Barbeau	1 January 2019	7 September 2023	1770
	Svartberget	1 Januari 2019	21 October 2023	1754
	Grillenburg	24 April 2017	28 September 2023	2348

The subsequent step in the data preprocessing involved the incorporation of global variables, which served as proxies for the distinct ecosystems of each site. These variables included elevation, latitude, and longitude. Ecosystem type and season, derived from the months of data collection, were also utilized as categorical explanatory variables. Seasons were categorized as follows [13]: Winter (December to February), Spring (March to May), Summer (June to August), and Autumn (September to November). Both the type of ecosystem and the season were treated as categorical data; these categories were subsequently encoded using one-hot encoding to facilitate their integration into the analysis.

3.3.1 Experimental Setup

Experiment 1: Individual Site

In the first stage of model development, individual sites were analyzed using three distinct methods: SARIMAX, XGBoost, and LSTM. The performance of these models was assessed and compared using a predefined evaluation matrix to determine the most effective approach for each site. Following this initial phase, the second stage involved the integration of these individual models using transfer learning techniques. This approach facilitated the creation of a comprehensive unified model capable of encompassing all sites.

Prior to training, the datasets from each site were partitioned into training and testing subsets with a split ratio of 80:20. This structured approach not only ensures consistency in

model training but also enhances the reliability of the model evaluation process across different ecological contexts. The SARIMAX model was implemented using the StatsModels library for Python [91, 92]. The XGBoost implementation utilized the XGBoost package for Python [93]. The LSTM algorithm was executed using the Keras library for Python [94]. This study employed a shallow network consisting of two LSTM layers, followed by a dropout layer and an output layer. Dropout is a common method to enhance generalization in neural networks by randomly deactivating neurons during training. This prevents dependency on any single neuron and promotes learning of diverse features [95]. The training was conducted using the Adam optimizer and mean squared error as the loss function. The batch size for the training process was set at 64.

TABLE 3.3: Tuning Parameter Values

Algorithm	Param	Details [93] [91]	Default	Param Val
SARIMAX	(p,d,q)	Order of the model for the number of AR parameters, differences, and MA parameters	(1, 0, 0)	(1,0,1), (1,1,1), (3,0,1), (3,0,3), (3,1,1), (3,1,3), (7,0,3), (7,0,7), (7,1,3), (7,1,7), (14,0,7), (14,1,7), (30,0,3), (30,1,3)
	(P,D,Q,s)	Order of the seasonal component of the model for the AR parameters, differences, MA parameters, and periodicity	(0, 0, 0, 0)	(3, 1, 3, 7), (3, 1, 3, 14), (3, 1, 3, 30)
XGBoost	η	Learning rate: step size shrinkage used in update to prevents overfitting	0.3	0.01, 0.1, 0.2, 0.3
	α	L1 regularization term on weights	0	0, 0.1, 1, 5
	γ	minimum split loss: minimum loss reduction required to make a further partition on a leaf node of the tree	0	0, 0.1, 0.2, 0.5, 1
	md	Maximum depth of a tree	6	3, 4, 6, 8, 10
	mcw	Minimum sum of instance weight (hessian) needed in a child	1	1, 2, 5, 10
	L	Time lags	-	3, 5, 7
LSTM	epoch	The number of times the entire training dataset is processed during training	-	50, 100, 150
	Unit1	LSTM unit of the first layer	-	32, 64, 128
	Unit2	LSTM unit of the second layer	-	32, 64, 128
	L	Time lags	-	3, 5, 7, 14
	dropout	dropout rate		0.1, 0.2, 0.3

To optimize the performance of the individual site models, hyperparameter tuning was conducted. For XGBoost, the hyperparameters tuned included min_child_weight (mcw),η, α, γ, max_depth (md), and L. For the LSTM, the hyperparameters adjusted

were the LSTM units of each layer, the number of epochs, and the dropout ratio. For SARIMAX, the hyperparameters were the order and seasonal order. A detailed explanation of each parameter and the values used in this study are presented in Table 3.3.

To determine the hyperparameters for the SARIMAX method, each time series was subjected to a stationarity check using the ADF Test. This analysis facilitated the selection of the parameter d, representing the integration order required to achieve stationarity. Additionally, the ACF and PACF graphs were quickly examined to identify suitable hyperparameters. The tuning of parameters was conducted manually through a grid search scheme. It is important to note that in the SARIMAX model, the seasonal and non-seasonal AR components must be distinct. Any combinations of orders that did not meet this criterion were excluded from consideration. For example, the configuration with an order of (7, 0, 3) and a seasonal order of (3, 1, 3, 7) was skipped.

For the training of site-specific models, all vegetation indices that were collected, specifically NDVI, MNDVI, EVI, EVI2, CIr, LSWI, MNDWI, NDII were used along with available in-situ measurement data detailed in Table 3.1. The target variable was GPP_DT_VUT_USTAR50. Global variables were excluded from the training dataset for each site-specific model. The rationale for this exclusion was that the values for global data remained constant within the same site, rendering them redundant for model training as they provided no additional insight to enhance model performance. Additionally, the study explored the feature importance of the XGBoost model to identify which indicators most significantly influenced the prediction of GPP.

Experiment 2: Unified Model

After constructing individual models and evaluating their performance based on the metrics specified in Section 3.3.2, the subsequent phase involved building a unified model. This model was developed using an incremental learning framework, which entailed continuously training the model with new data. The objective was to enhance the generalization capabilities of the unified model across different ecosystem types, thereby increasing its adaptability for further improvements. With this approach, adding new sites to the model did not require starting training from scratch; instead, the model utilized the knowledge previously acquired from earlier training sessions. This strategy also addressed the challenge posed by the varying time periods of data availability from different sites.

As with the individual models, data for each site were split into 80:20 ratios for training and testing, respectively. The model was initially trained using data from the first site. Subsequently, the pre-trained model underwent further training with data from the second site, and so on. The training process of the unified model is illustrated in Figure 3.11.

Initially, training data from the first site (Train Data 1) were used to produce the first model (Model 1). This model was then evaluated using test data from the same site (Test Data 1). Following the evaluation, Model 1 was retrained using training data from the second site (Train Data 2), resulting in Model 2. Model 2 was evaluated using both the test data from the second site (Test Data 2) and the test data from the first site (Test Data 1) to check for consistency and improvement. This process continued through the 4th training session, culminating in the development of Model 4. Model 4, serving as the unified model, was then assessed to determine the performance enhancements achieved through this retraining approach.

The hyperparameters for each algorithm were initialized during the first training, and these settings remained consistent throughout all training processes at each site. The selection of hyperparameters was based on the optimal parameters identified through the tuning process of the individual models. Selection was primarily guided by a majority
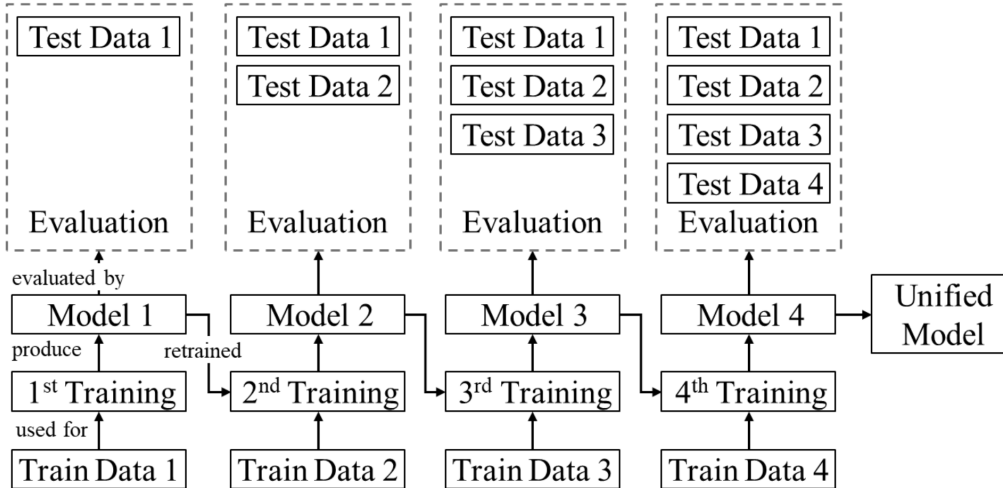
FIGURE 3.11: Training Process of Unified Model

vote from the best parameters across all sites, although this was not the exclusive method used; there was also exploration of parameters to enhance performance. However, systematic exploration, such as comprehensive parameter tuning, was not conducted for the unified model due to time constraints. During the exploratory phase, global variables were also utilized as features to determine whether their inclusion could enhance the model's performance.

The incremental learning was employed for constructing unified models using XGBoost and LSTM. However, this approach was not suitable for the SARIMAX model. Once a SARIMAX model is trained, updating it with new data typically required retraining from scratch due to the tight coupling of the model's state with the entire time series. The model's parameters, optimized based on the complete dataset, make it difficult to pretrain on a subset and achieve reliable generalization to future data. This inherent limitation of the SARIMAX model represents a significant constraint of the study. This challenge was addressed by combining all training datasets for the SARIMAX method and inserting zero-padding between datasets from different sites to maintain data separation. The training process was then executed once.

3.3.2 Evaluation Metrics

To evaluate the predictive model performance, four evaluation metrics were used, namely Mean Squared Error (MSE), RMSE, Mean Absolute Error (MAE), and coefficient of determination (R^2). The RMSE is more appropriate for representing model performance than the MAE when the error distribution is expected to be Gaussian, but a combination of metrics is often necessary to assess model performance [96]. The R^2 value is often used to show how well the regression model fits the data. A value near 1 indicates a good fit, and one near 0 indicates a poor fit. In other words, if the regression model can explain most of the variation in the response data, then it fits the data well [97].

The equations for MSE, RMSE, MAE, and R^2 are presented in Equations (3.1) through (3.4), respectively. In these formulas, m denotes the total number of data points, y_i represents the observed value of GPP for the i-th data point, $\hat{y}_i$ indicates the corresponding predicted value of GPP, and $\bar{y}$ is the mean of the observed GPP values across all data points. In this study, the objective is to minimize the values of MSE, RMSE, and MAE, indicating closer agreement between observed and predicted values, while a higher R^2 value signifies

a better fit of the model to the data.

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{3.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{3.2}$$

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{3.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2} \tag{3.4}$$

Chapter 4

Results and Discussion

4.1 Individual Site

4.1.1 SARIMAX

Table 4.1 presents the evaluation metrics for the tuned SARIMAX model using parameters from Table 3.3. It displays the three best-ranked metrics for each site based on combinations of parameters. SARIMAX perform best to predict GPP at Estrees Mons 28 using the parameter combination $(3, 0, 3)$ for order and $(3, 1, 3, 7)$ for seasonal adjustments, achieving the lowest MSE of 6.59 and the highest R^2 value of 0.65. This combination slightly outperformed the other configurations tested, indicating a strong seasonal pattern that fits a weekly cycle. The same parameter combinations also perform best in Hohes Holz where it reached an R^2 value of 0.88 and the lowest RMSE of 1.91 which is better than in Estrees Mons.

TABLE 4.1: Performance Metrics for SARIMAX Method Across Various Sites

Site	Order	Seasonal	MSE	RMSE	MAE	R^2
Estrees-Mons A28	(7, 0, 3)	(3, 1, 3, 14)	6.72	2.59	1.95	0.64
	(14, 0, 7)	(3, 1, 3, 30)	6.64	2.58	2.01	0.64
	(3, 0, 3)	(3, 1, 3, 7)	**6.59**	**2.57**	**1.91**	**0.65**
	(3, 1, 3)	(3, 1, 3, 30)	8.84	2.97	1.97	0.53
Hohes Holz	(3, 0, 1)	(3, 1, 3, 7)	3.70	1.92	**1.37**	0.88
	(3, 0, 3)	(3, 1, 3, 7)	**3.65**	**1.91**	1.40	**0.88**
	(1, 0, 1)	(3, 1, 3, 7)	3.70	1.92	1.37	0.88
Hyytiala	(3, 0, 1)	(3, 1, 3, 30)	1.30	1.14	0.88	0.88
	(1, 0, 1)	(3, 1, 3, 14)	1.64	1.28	1.00	0.85
	(1, 0, 1)	(3, 1, 3, 30)	**1.09**	**1.04**	**0.79**	**0.90**
Torgnon	(3, 0, 1)	(3, 1, 3, 30)	1.02	1.01	0.74	0.88
	(14, 0, 7)	(3, 1, 3, 30)	**0.95**	**0.98**	**0.71**	**0.89**
	(7, 0, 7)	(3, 1, 3, 30)	1.01	1.01	0.74	0.88

In Hyytiala, the model with configuration $(1, 0, 1)$ and $(3, 1, 3, 30)$ showed the best performance, with an R^2 value of 0.90, and the lowest RMSE and MAE values of 1.04 and 0.79, respectively. Similarly, for Torgnon, the optimal seasonal order parameters were $(3, 1, 3, 30)$, indicating a better fit for models capturing monthly seasonal effects. The $(14, 0, 7)$ parameters yielded the most effective outcomes for the Torgnon site, underscoring the effectiveness of using a more extensive lag.
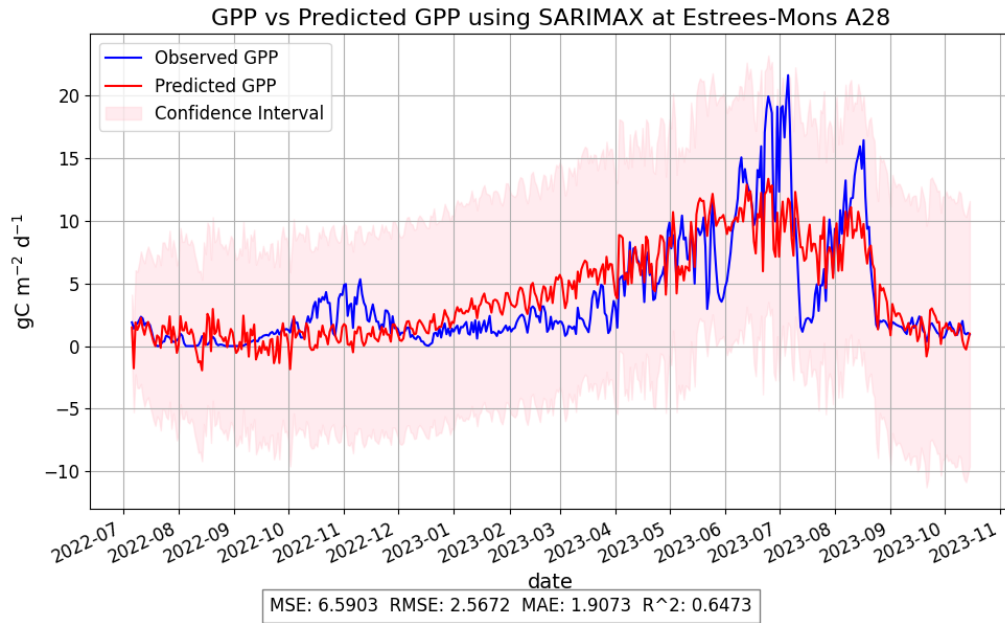
FIGURE 4.1: Predicted vs. Observed Value of GPP at Estrees Mons. The blue line represents observed GPP values, and the red line indicates predicted GPP values.
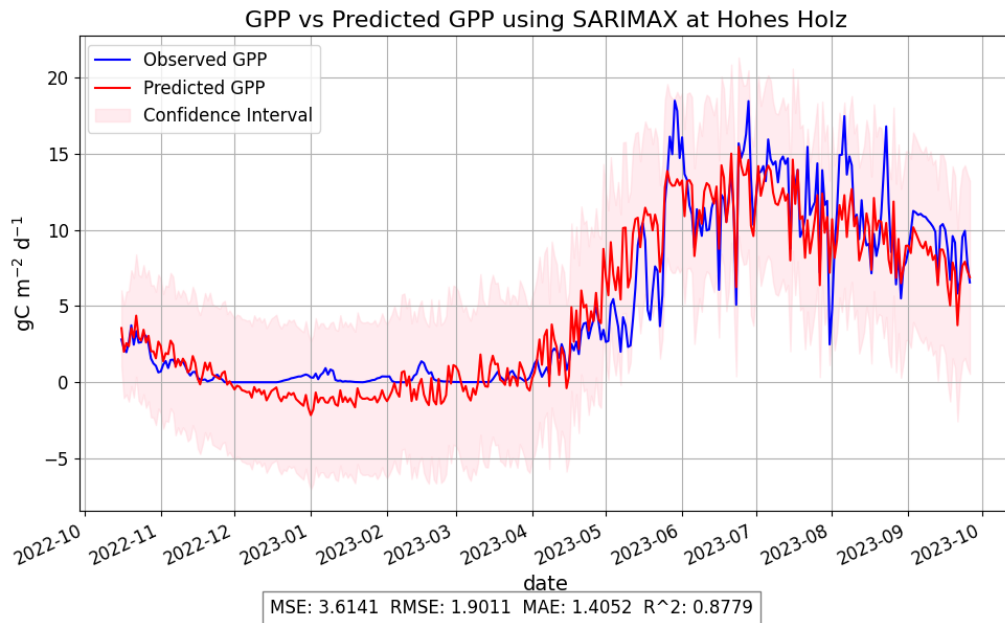

FIGURE 4.2: Predicted vs. Observed Value of GPP at Hohes Holz. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

The Torgnon site exhibits the lowest MSE, RMSE, and MAE values compared to other sites, as shown in Table 4.1. However, these lower error metrics may not fully reflect the relative effectiveness of the models across different sites due to variations in the range of GPP values at each site. Specifically, the GPP values at these sites vary widely, ranging from a minimum of zero to maximums of 28.71 $\text{gCm}^{-2}\text{d}^{-1}$ at Estrees Mons, 20.17 $\text{gCm}^{-2}\text{d}^{-1}$ at Hohes Holz, 12.85 $\text{gCm}^{-2}\text{d}^{-1}$ at Hyytiala, and 14.18 $\text{gCm}^{-2}\text{d}^{-1}$ at Torgnon. Such high values of GPP, especially in croplands, are expected and align with
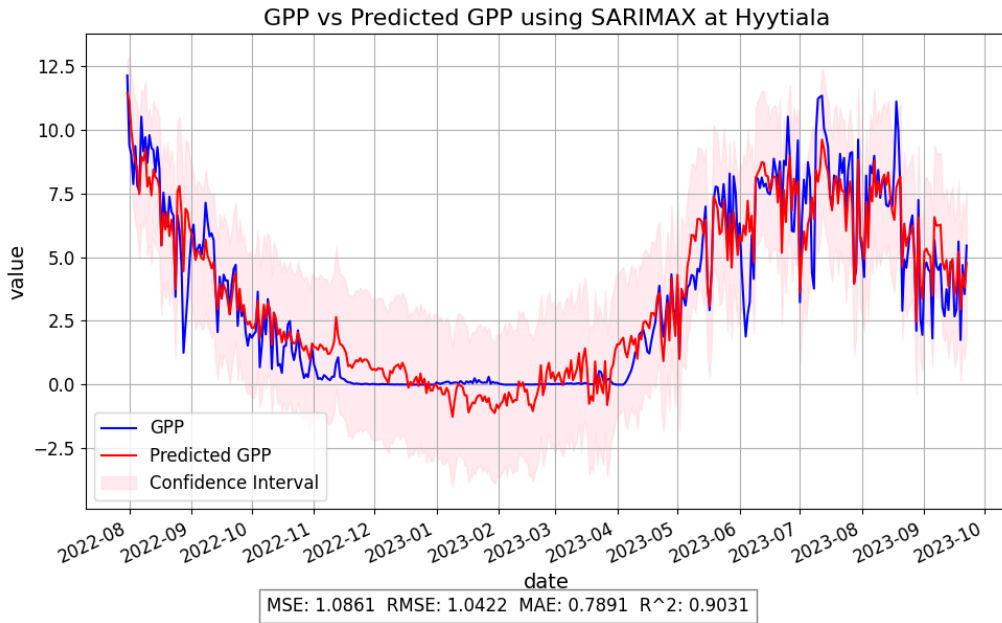
FIGURE 4.3: Predicted vs. Observed Value of GPP at Hyytiala. The blue line represents observed GPP values, and the red line indicates predicted GPP values.
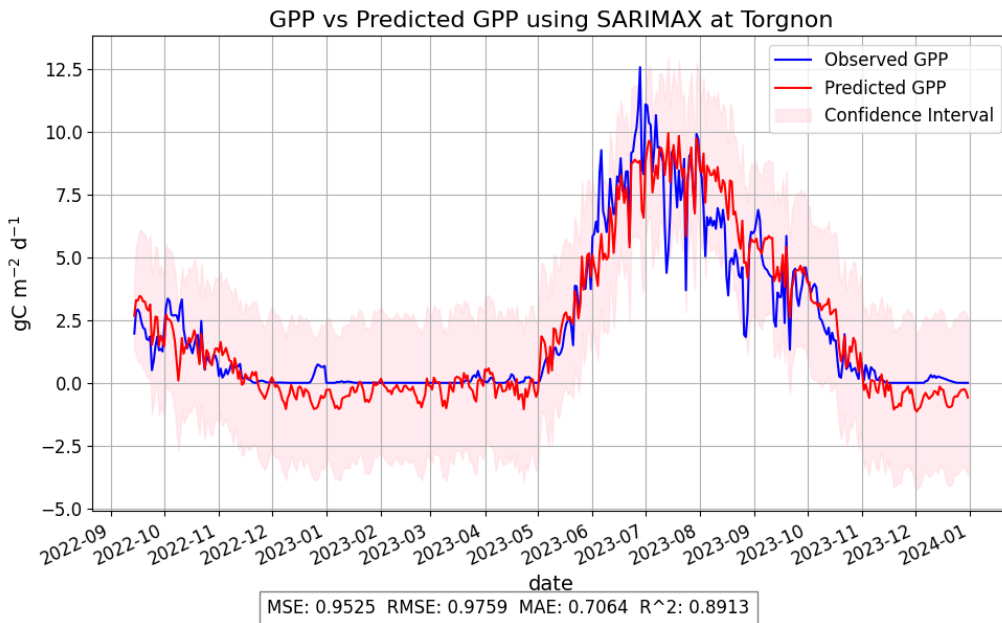

FIGURE 4.4: Predicted vs. Observed Value of GPP at Torgnon. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

observed carbon fixation rates in productive agricultural settings, as documented in [98]. Additionally, the magnitude of MSE can increase significantly with higher target values. The larger the value to predict, the greater the potential squared error, since no model predicts without error, as discussed in [99]. This phenomenon accounts for the higher MSE observed at Estrees Mons compared to other sites.

Across all sites, the SARIMAX model occasionally underestimates GPP (Figures 4.1 - 4.4), yielding predicted values that are less than zero. Such outcomes are not feasible

in the context of GPP, which represents the rate of carbon fixation in an ecosystem and inherently cannot be negative. These negative predictions suggest potential limitations in the model's parameterization or structural inadequacies in accurately capturing the lower bounds of GPP under certain conditions.

4.1.2 XGBoost

The data presented in Table 4.2 presents the three best-ranked performance including MSE, RMSE, MAE, and R^2 of the XGBoost algorithm based on the tuning parameter for each site. Estrees-Mons A28 shows the highest R^2 of 0.73 and the lowest RMSE and MSE values of 2.26 and 5.11, respectively, achieved with the configuration of md = 6, mcw = 10, $\eta = 0.01$, $\gamma = 1$, and $\alpha = 1$. This model configuration reflects a conservative yet effective model setup. The high gamma value significantly reduces the likelihood of overfitting by demanding considerable loss reductions for further tree splits, which is advantageous in complex regression tasks where maintaining generalizability is crucial.

TABLE 4.2: Performance Metrics for XGBoost Method Across Various Sites

Site	(md,mcw)	(L, η, γ, α)	MSE	RMSE	MAE	R^2
Estrees-Mons A28	(6, 5)	(5, 0.3, 0.1, 0.1)	5.93	2.43	1.57	0.68
	(6, 5)	(5, 0.3, 0.2, 0.1)	5.86	2.42	**1.58**	0.69
	(6, 2)	(5, 0.3, 0.5, 1)	5.87	2.43	1.59	0.69
	(6, 10)	(5, 0.01, 1 , 1)	**5.11**	**2.26**	1.84	**0.73**
	(6, 10)	(5, 0.01, 0.1, 1)	5.13	2.26	1.85	0.73
Hohes Holz	(6, 5)	(5, 0.1, 0, 1)	**2.73**	**1.65**	1.10	**0.91**
	(6, 10)	(7, 0.3, 0.1, 0)	2.75	1.66	1.09	0.91
	(6, 10)	(7, 0.3, 0.5, 0.1)	2.80	1.67	**1.07**	0.91
	(6, 10)	(7, 0.3, 1, 0)	2.83	1.68	1.08	0.90
Hyytiala	(4, 1)	(7, 0.1, 0.1, 0.1)	**1.00**	**1.00**	0.63	**0.91**
	(4, 2)	(7, 0.1, 0.1, 0.1)	1.02	1.01	**0.62**	0.91
	(4, 5)	(7, 0.1, 0.5, 0.1)	1.05	1.02	0.63	0.90
Torgnon	(6, 2)	(7, 0.2, 0, 5)	**0.79**	**0.89**	0.52	**0.91**
	(6, 5)	(7, 0.2, 1, 5)	0.79	0.89	0.53	0.91
	(6, 5)	(7, 0.2, 0.1, 5)	0.80	0.89	**0.52**	0.91
	(6, 2)	(7, 0.2, 0.1, 5)	0.81	0.90	0.52	0.91

Hohes Holz has excellent prediction accuracy, achieving the highest R^2 of 0.9079 and the lowest RMSE and MSE of 1.65 and 2.73, respectively, using parameters md = 6, mcw = 5, $\eta = 0.1$, and $\alpha = 1$. In Hyytiala, the model with md = 4, mcw = 1, $\eta = 0.1$, $\gamma = 0.1$, and $\alpha = 0.1$ achieves an R^2 of 0.91, with the lowest RMSE and MSE of 1.00 and 1.00, respectively.

Lastly, Torgnon demonstrates the highest R^2 of 0.91 and the lowest RMSE and MSE of 0.89 and 0.79, respectively, under the parameters md = 6, mcw = 2, $\eta = 0.2$, and $\alpha = 5$. This site's best parameter indicates a strong management of the bias-variance trade-off, facilitated by moderate depth and high regularization. High α values contribute to a substantial reduction in model complexity by penalizing non-zero coefficients, effectively mitigating potential overfitting while maintaining high predictive performance.

Similar to SARIMAX, XGBoost performs worse at the Estrees Mons site compared to other locations. As illustrated in Figure 4.5, it is evident that XGBoost tends to
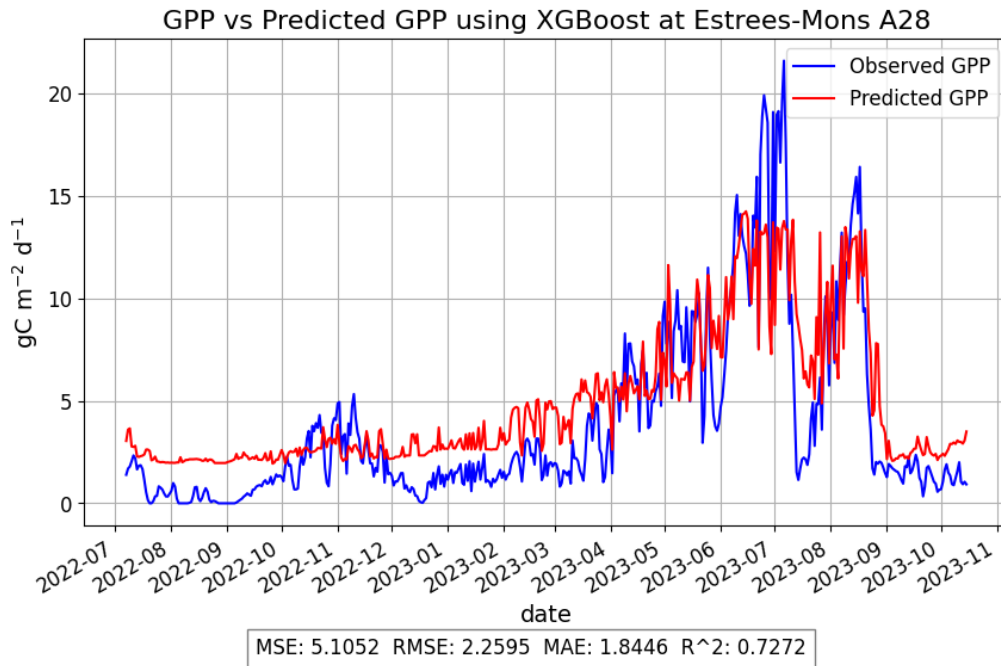
FIGURE 4.5: Predicted vs. Observed Value of GPP from Estrees-Mons A28. The blue line represents observed GPP values, and the red line indicates predicted GPP values.
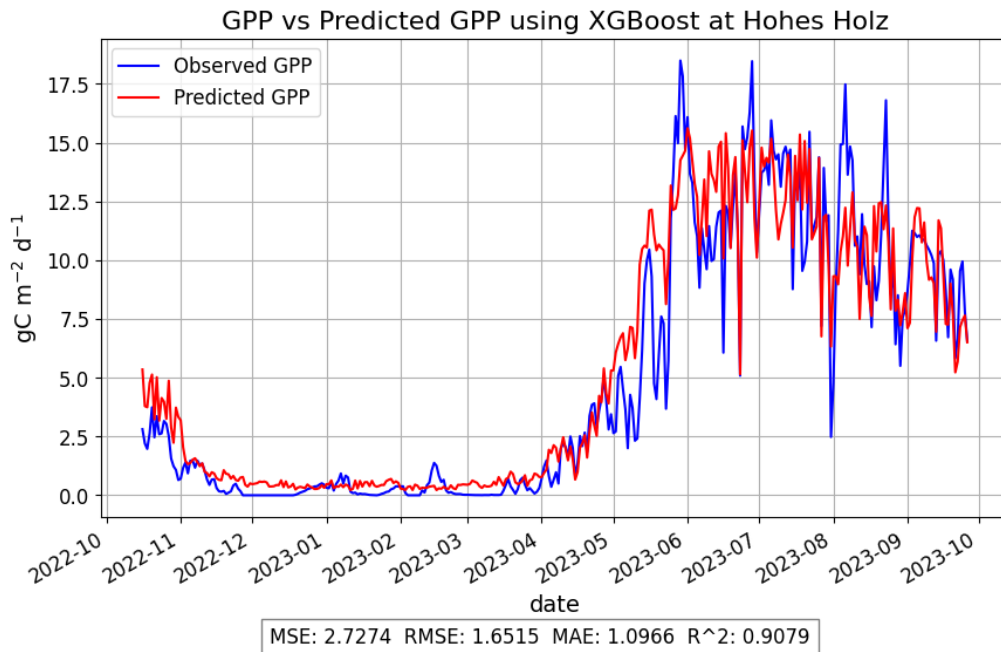

FIGURE 4.6: Predicted vs. Observed Value of GPP from Hohes Holz. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

overestimate the GPP values when they are low in Estrees Mons, while it generally predicts the GPP values accurately at other sites despite some discrepancies at extreme values.

Figure 4.9 displays the feature importance of each model from different sites, showing the ten most important features based on the relative value of the gain metrics. This feature
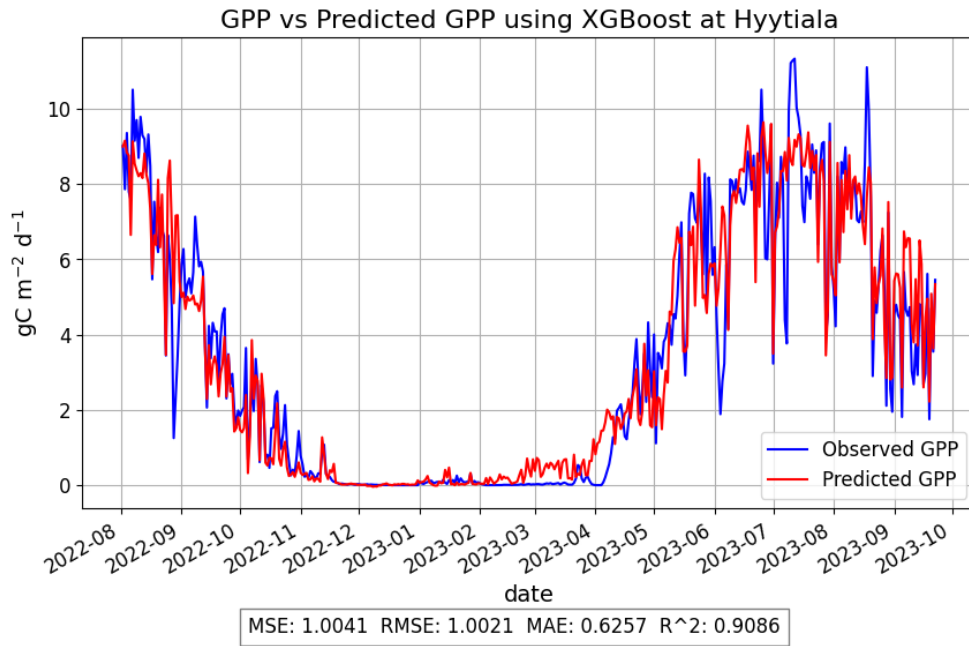
FIGURE 4.7: Predicted vs. Observed Value of GPP from Hyytiala. The blue line represents observed GPP values, and the red line indicates predicted GPP values.
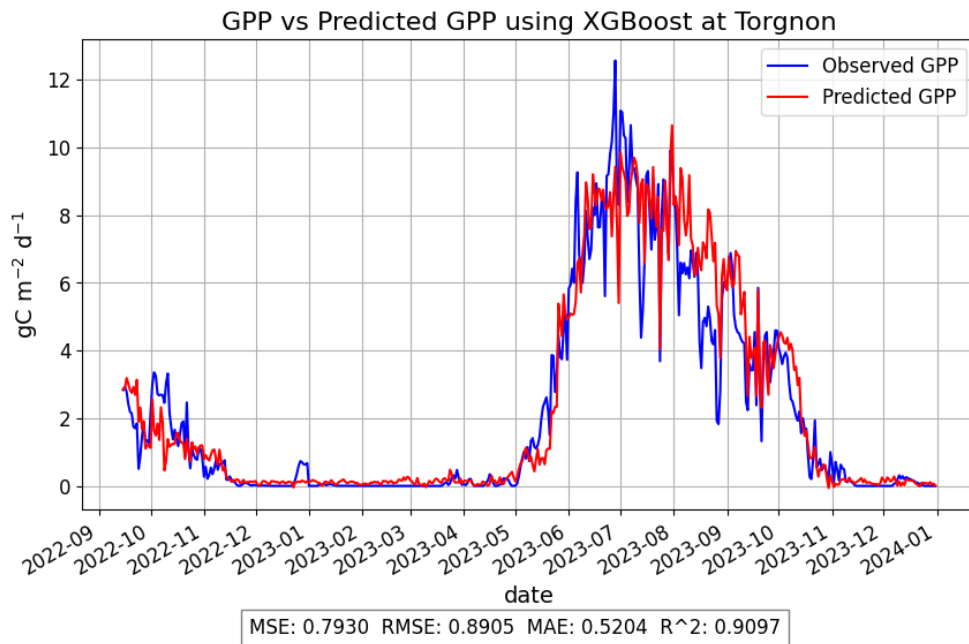

FIGURE 4.8: Predicted vs. Observed Value of GPP from Torgnon. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

importance can help provide a better understanding of which indicators are most important in each ecosystem. From the figure, it is evident that the latent heat turbulent flux is the most important feature in Estrees Mons A28. Similarly, this feature also contributes to the highest gain value in Hohes Holz.

In Hyytiala, air temperature (TA) emerges as the most significant feature, likely due to the site location in the Nordic region. Temperature plays a crucial role as there is an opti-
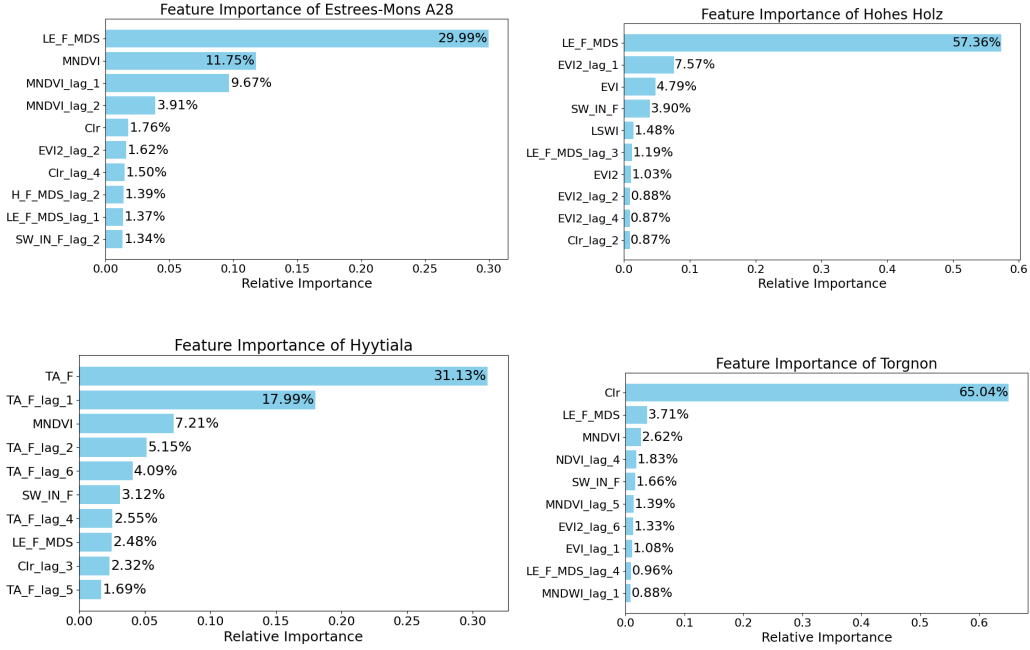
36

FIGURE 4.9: Feature Importance Scores for XGBoost Models Across Different Sites. For details on specific variables, refer to Table 3.1 and Section 2.2.

mal temperature range for vegetation to carry out photosynthesis. A study in [100] found that air temperature was the major limiting factor for photosynthesis in early spring, autumn, and winter. This indicates the significant impact of temperature on photosynthetic rates during seasons characterized by lower temperatures, which are common in the Nordic climate of Hyytiala. Additionally, this finding aligns with the results in [11], which found that net radiation (not used in this study) and TA are important features in predicting GPP in evergreen needleleaf forests in subarctic climate conditions.

In Torgnon, CIr demonstrates the highest gain value. This finding aligns with the results of the study in [101], which found that GPP based on CIr exhibited the highest correlation and low uncertainties with GPP from EC across grassland sites. The occurrence of EVI and NDVI as feature importance also aligns with the study in [12] that evaluated the key impact factors for predicting GPP in grassland.

4.1.3 LSTM

Table 4.3 presents the performance metrics of the LSTM model across different sites. This table specifically showcases the parameter settings that resulted in the three best-ranked performance metrics for predicting GPP, including the lowest MSE, RMSE, MAE, and the highest R^2. Each configuration displayed represents the three best outcomes for each site after parameter tuning, highlighting the effectiveness of different LSTM configurations in capturing the variability and trends in GPP data across different sites.

In Estrees-Mons, the parameter set with $L = 3$, Epochs $= 50$ and Unit1 $= 128$, Unit2 $= 32$, Dropout $= 0.2$ achieves the best performance in terms of MSE 6.07, RMSE 2.46, and R^2 0.68. The relatively low dropout rate seems to maintain enough complexity in the model to learn effectively without overfitting. The lowest MAE is observed with a larger lag of $L = 14$ and higher epochs, which implies better average performance but not necessarily
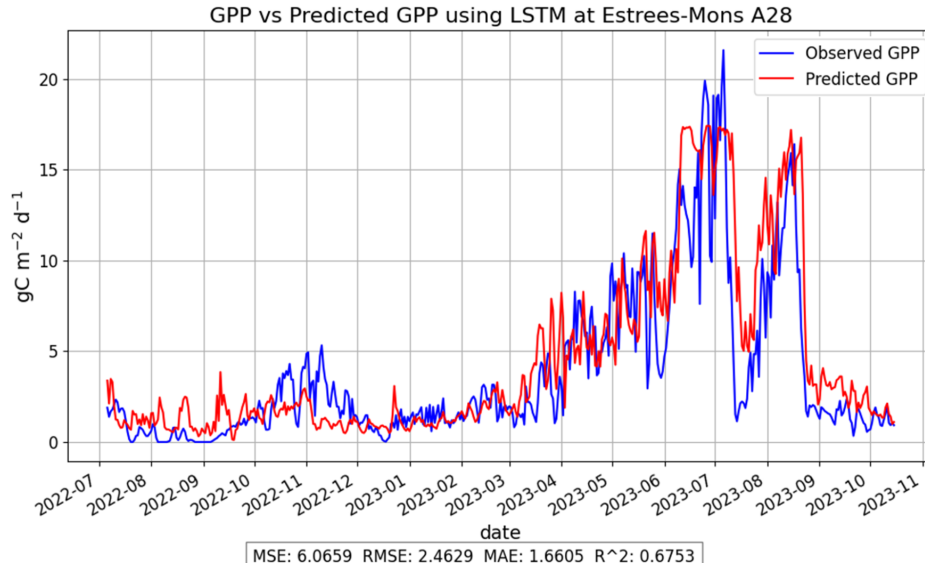
FIGURE 4.10: Predicted vs. Observed Value of GPP from Estrees-Mons A28. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

TABLE 4.3: Performance Metrics for LSTM Method Across Various Sites

Site	L, Epoch	(Unit1, Unit2, Dropout)	MSE	RMSE	MAE	R^2
Estrees-Mons A28	(3,50)	(128,32,0.2)	**6.07**	**2.46**	1.66	**0.68**
	(3,50)	(64,64,0.3)	6.68	2.58	1.78	0.64
	(3,50)	(32,64,0.1)	6.93	2.63	1.74	0.63
	(14,150)	(128,64,0.3)	7.20	2.68	**1.58**	0.62
	(14,50)	(128,128,0.3)	8.30	2.88	1.70	0.56
Hohes Holz	(14,50)	(32,128,0.1)	**4.03**	**2.01**	**1.30**	**0.86**
	(14,50)	(32,32,0.2)	4.80	2.19	1.46	0.84
	(3,100)	(32,32,0.3)	4.92	2.22	1.38	0.83
	(7,50)	(64,32,0.2)	5.00	2.24	1.37	0.83
	(7,50)	(64,32,0.3)	5.05	2.25	1.35	0.83
Hyytiala	(5,50)	(64,32,0.3)	**1.69**	**1.30**	0.89	**0.85**
	(5,50)	(64,32,0.1)	1.71	1.31	0.87	0.85
	(5,150)	(32,32,0.3)	1.77	1.33	0.89	0.84
	(7,50)	(32,128,0.2)	1.77	1.33	0.87	0.84
	(14,50)	(32,32,0.3)	1.79	1.34	**0.84**	0.84
Torgnon	(5,50)	(32,128,0.3)	**0.97**	**0.99**	0.59	**0.89**
	(5,50)	(64,128,0.3)	0.98	0.99	0.58	0.89
	(5,50)	(32,64,0.3)	1.00	1.00	0.58	0.89
	(3,100)	(32,32,0.1)	1.00	1.00	0.57	0.89
	(3,50)	(128,64,0.1)	1.02	1.01	0.57	0.88
	(3,50)	(32,32,0.3)	1.04	1.02	**0.57**	0.88

improved peak performance.

In Hohes Holz, the combination $L = 14$, Epochs $= 50$ and Unit1 $= 32$, Unit2 $=$
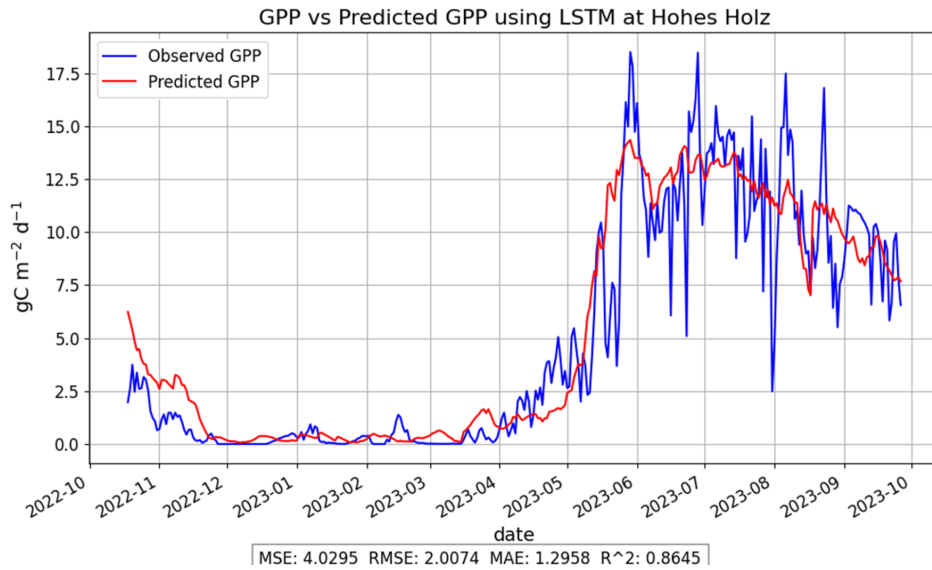
38

FIGURE 4.11: Predicted vs. Observed Value of GPP from Hohes Holz. The blue line represents observed GPP values, and the red line indicates predicted GPP values.
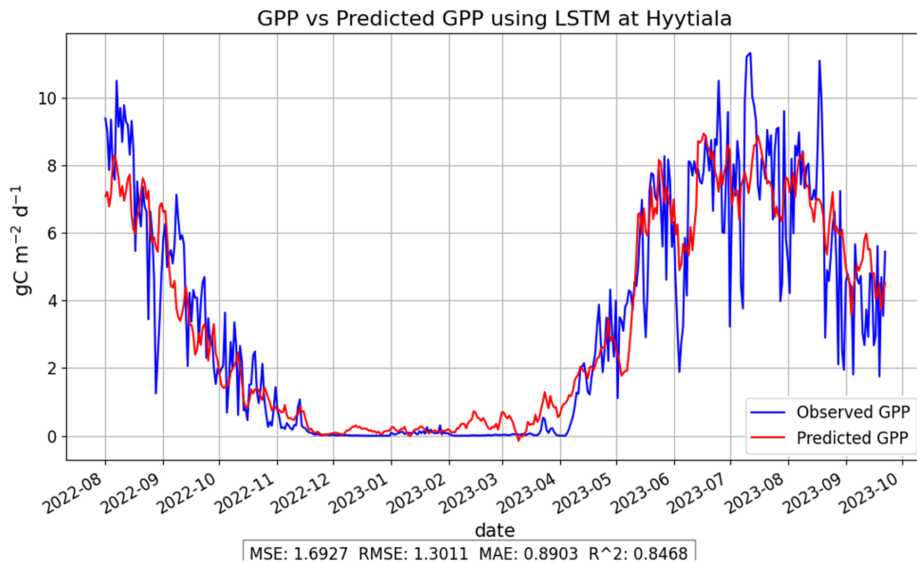

FIGURE 4.12: Predicted vs. Observed Value of GPP from Hyytiala. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

$128, \text{Dropout} = 0.1$ outperforms other configurations across all metrics with MSE of 4.03, RMSE of 2.01, MAE of 1.30, and an R^2 of 0.86. This configuration indicates that increasing the units in the second layer while reducing the dropout rate can enhance the model's predictive accuracy, possibly by allowing the network to refine its feature extraction capabilities more effectively.

For Hyytiala, the optimal parameters are $L = 5$, Epochs $= 50$ and Unit1 $= 64$, Unit2 $= 32, \text{Dropout} = 0.3$, with MSE, RMSE, and R^2 values of 1.69, 1.30, and 0.85 respectively. At Torgnon, the lowest MSE (0.97), RMSE (0.99), and highest R^2 (0.89) are achieved with $L = 5$, Epochs $= 50$ and Unit1 $= 32$, Unit2 $= 128, \text{Dropout} = 0.3$.
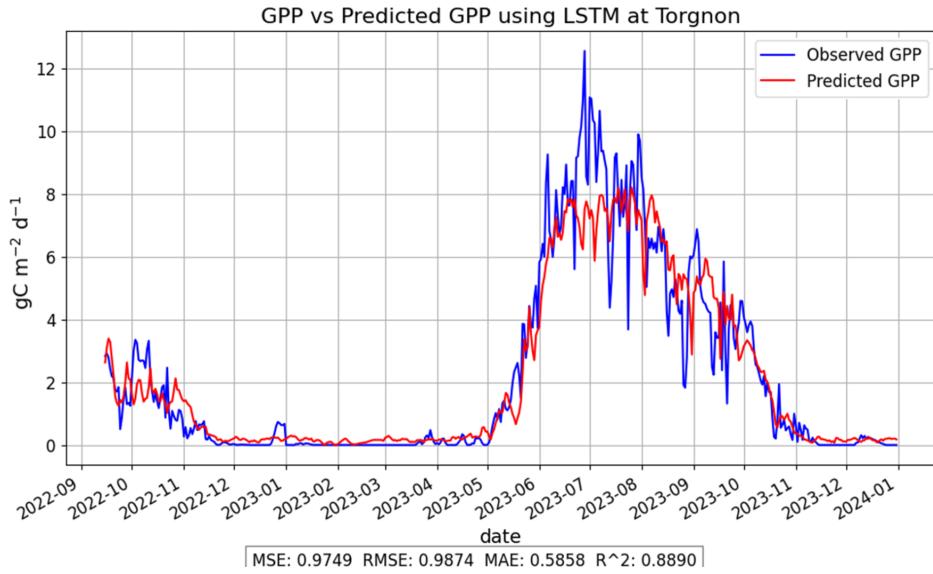
39

Figure 4.13: Predicted vs. Observed Value of GPP from Torgnon. The blue line represents observed GPP values, and the red line indicates predicted GPP values.

4.1.4 Comparison

In general, XGBoost produced the best evaluation metrics if compared to SARIMAX and LSTM in all sites, as it produced the highest R^2 and the lowest MSE, RMSE, and MAE(Table 4.4). For three out of four sites, XGBoost achieved an R^2 of 0.91, except for Estrees-Mons. This could be due to the type of ecosystem, as GPP in croplands depends not only on seasonal patterns, but also on the harvest time, which can vary based on the type of crops. Accurate prediction is challenging because the ecosystem type in question is not 'natural', introducing further variability that is not easily captured by standard models.

Table 4.4: Performance metrics by site and method

Site	MSE	RMSE	MAE	R^2	Methods
	6.59	2.58	1.91	0.65	SARIMAX
Estrees-Mons A28	**5.11**	**2.26**	1.84	**0.73**	**XGBoost**
	6.07	2.46	**1.66**	0.68	LSTM
	3.65	1.91	1.40	0.88	SARIMAX
Hohes Holz	**2.73**	**1.65**	**1.10**	**0.91**	**XGBoost**
	4.03	2.01	1.30	0.86	LSTM
	1.09	1.04	0.79	0.90	SARIMAX
Hyytiala	**1.00**	**1.00**	**0.63**	**0.91**	**XGBoost**
	1.69	1.30	0.89	0.85	LSTM
	0.95	0.98	0.71	0.89	SARIMAX
Torgnon	**0.79**	**0.89**	**0.52**	**0.91**	**XGBoost**
	0.97	0.99	0.59	0.89	LSTM

While XGBoost performs best in terms of the evaluation metrics, LSTM has better capability on predicting extreme values, particularly high values, in all sites (Figure 4.1, 4.5, 4.10). There is an unusual pattern at Estrees Mons in July 2023 where the GPP value

dropped drastically from around $21.5gCm^{-2}d-1$ on July 6th to around $1.1gCm^{-2}d-1$ on the July $14th$ and then increased again on 14th of August to $15.9gCm^{-2}d-1$.

The sharp drop in GPP could correspond to the harvesting of crops, which would significantly reduce the photosynthetic biomass available in the field. This would result in a temporary reduction in GPP, while not being explained by the in-situ explanatory variables. This decreasing pattern has been also observed in previous years. Following the harvest, the subsequent planting of fast-growing crops could have contributed to the quick recovery in GPP values observed in August. The SARIMAX model cannot properly capture this, while XGBoost has shown better performance in this regard and has the best evaluation metrics on this site. However, among the various algorithms tested, the LSTM model more effectively captures these patterns, particularly in predicting both the significant lows and the subsequent high peaks more accurately than the other models.

The predictive performance of the individual models at the Hyytiala evergreen needle-leaf forest site demonstrates improvement over the results reported by Wang et al. [102], which utilized a Light Use Efficiency (LUE) modeling approach. The LUE model achieved a minimum RMSE of 1.43 and an optimal R^2 of 0.76 at an evergreen needleleaf forest site in China. In comparison, the XGBoost model used in the current study outperforms the linear regression model applied to daily GPP prediction in Hyytiala by Cai et al. [10], which utilized Sentinel-2 data to achieve an R^2 of 0.89 and an MAE of 0.73. The results of this study show an R^2 of 0.91 and an MAE of 0.63, indicating slightly higher accuracy. However, the RMSE reported by Cai et al. at 0.98 is slightly better than the 1.00 obtained in this study.

While XGBoost consistently delivers the best performance metrics across all sites, SARIMAX also yields comparable results. According to Table 4.4, SARIMAX outperformed LSTM in predicting GPP at Hohes Holz, Hyytiala, and Torgnon. This demonstrates that deep learning methods are not always superior to classical statistical approaches. The performance of SARIMAX is closely matched with that of XGBoost at these sites, with only slight differences. However, a significant drawback of SARIMAX is its longer training time. For instance, training and testing SARIMAX for the Hyytiala site took about 31 minutes, whereas XGBoost completed the same task in approximately 4 seconds. This efficiency advantage of XGBoost is also noted by [32], highlighting its scalability and faster runtime.

Another distinction observed in the figures is that the predicted values of GPP from the LSTM model tend to be smoother compared to those from XGBoost and SARIMAX. This smoothness indicates that LSTM exhibits better generalization capabilities, particularly in capturing seasonality. While this characteristic might be advantageous for constructing a global model that generalizes well across different sites, it is crucial to balance this with the need for accuracy in the GPP predictions.

4.2 Unified Model

4.2.1 SARIMAX

Table 4.5 details the performance of the SARIMAX model in the training phase for a unified model across four different sites. The results were achieved using a parameter order of $(3, 0, 1)$ and a seasonal order of $(3, 1, 3, 7)$, incorporating global variables such as latitude, longitude, elevation, ecosystem type, and season as exogenous factors. Table 4.6 represents the model's performance during the testing phase at the test sites. The inclusion of these global variables improved MSE and RMSE scores. Although the average

R^2 for the training phase decreased from 0.68 to 0.65, there was an improvement in model performance on the testing sites, where the average R^2 increased from 0.54 to 0.60 for unseen data, maintaining the same parameter configuration. Additionally, the average MSE, RMSE, and MAE also showed improvements.

TABLE 4.5: Performance metrics of SARIMAX in training phase

Site	MSE	RMSE	MAE	R^2
Estrees-Mons A28	8.01	2.83	2.03	0.57
Hohes Holz	3.88	1.97	1.59	0.87
Hyytiala	7.06	2.66	2.29	0.37
Torgnon	1.87	1.37	1.13	0.79
Average	5.20	2.21	1.76	0.65

TABLE 4.6: Performance metrics of SARIMAX in testing phase

Site	MSE	RMSE	MAE	R^2
Klingenberg	6.13	2.48	1.67	0.66
Fontainebleau-Barbeau	4.69	2.17	1.69	0.78
Svartberget	7.40	2.72	2.34	0.38
Grillenburg	5.97	2.44	1.93	0.60
Average	6.05	2.45	1.91	0.60

Despite improvements in the unified model's performance metrics, the SARIMAX model showed significantly lower effectiveness in predicting GPP at certain sites. Particularly, in Hyytiala, the model reached the lowest R^2 value of 0.37 in relation to other methods, indicating poor predictive accuracy. This site also had the highest MAE of 2.29 and a high MSE of 7.06. A similar decline in performance was observed in Svartberget, which, like Hyytiala, is an evergreen needleleaf forest. This pattern suggests that the model struggles particularly with this ecosystem type.

The adjustment of the seasonal period (parameter s) from 30 to 7 could be contributing to these challenges. In the individual site model, a seasonal period of 30 yielded the best performance, suggesting that a longer seasonal period is necessary for this type of ecosystem, as also seen in Torgnon. However, even though Torgnon showed the best results among the training sites in terms of MSE, RMSE, and MAE in the unified model, this was still a drop from the performance observed with a seasonal period of 30 in the individual model settings.

When the model was applied to the test site Grillenburg, which has the same ecosystem type as Torgnon, the performance metrics degraded further. Meanwhile, in terms of R^2, Hohes Holz displayed the best parameters, achieving an R^2 of 0.87. This setting also led to the best results in MSE and RMSE during testing at Fontainebleau-Barbeau, another deciduous broadleaf forest site.

4.2.2 XGBoost

Table 4.7 shows the metrics performance of XGboost during the training phase of the unified model and Table 4.8 shows the performance of the unified model when applied to an unseen site. Pattern emerged in the evaluation of models across multiple training sessions (table 4.7). After each model was retrained with data from a subsequent site, its performance on test data from previously trained sites tended to decrease. For example,

at Estrees Mons A28, the MSE increased significantly from 5.7261 after the first training session to 9.8288 in the second training session, before settling down to 6.3473 in the final session. Similarly, the R^2 value decreased from 0.6941 to 0.4749 after the second training, indicating a loss of predictive accuracy, though it improved slightly in subsequent sessions.

TABLE 4.7: Performance metrics by site and method

Site	Metrics	1st Train	2nd Train	3rd Train	4th Train
Estrees Mons A28	MSE	5.73	9.83	7.28	6.35
	RMSE	2.39	3.14	2.70	2.52
	MAE	1.91	2.31	1.95	1.73
	R^2	0.69	0.47	0.61	0.66
Hohes Holz	MSE		4.29	3.77	4.07
	RMSE		2.07	1.94	2.02
	MAE		1.57	1.29	1.33
	R^2		0.86	0.87	0.86
Hyytiala	MSE			2.76	2.31
	RMSE			1.66	1.52
	MAE			1.20	1.03
	R^3			0.75	0.79
Torgnon	MSE				3.05
	RMSE				1.75
	MAE				1.06
	R^2				0.65
Average MSE					3.95
Average RMSE					1.95
Average MAE					1.29
Average R$\hat{2}$					0.74

TABLE 4.8: Performance metrics of XGBoost in testing phase

Site	MSE	RMSE	MAE	R^2
Klingenberg	5.6739	2.3820	1.4258	0.6840
Fontainebleau-Barbeau	4.6266	2.1510	1.5113	0.7811
Svartberget	2.7404	1.6554	1.1514	0.7690
Grillenburg	5.1500	2.2694	1.6033	0.6565
Average	4.5477	2.1144	1.4230	0.7227

This observation suggests that while the model adapted to new site data, it simultaneously lost some of its predictive capabilities for earlier sites. This phenomenon could be indicative of overfitting to the new site data or an inability of the model to generalize effectively across diverse environmental conditions. Further investigation into this trend revealed that each successive training session, though aimed at enhancing the model's robustness, might have inadvertently introduced complexities that diminished its performance on previously well-modeled sites.

By the final training session, despite some recovery in metric scores at specific sites like Estrees Mons A28 and Hyytiala, where the MSE improved to 6.35 and 2.31 respectively, the overall trend indicated that the model's performance level s dropped in relation to its performance at the earlier training phases. The average MSE across sites by the last

training was 3.9471, average RMSE was 1.9516, average MAE was 1.29, and average R^2 was 0.74, illustrating the challenge of balancing model fit across multiple datasets.

The evaluation of the unified model on new sites provides important insights into its ability to generalize and adapt across different ecological settings. For example, Klingenberg, which has a similar ecosystem to Estrees Mons, showed an R^2 value of 0.68, indicating moderate adaptability. Even though the model initially struggled during training at Estrees Mons, it later showed improvement. This pattern suggests that while the model can adapt to changes, it may still need refinements to handle specific environmental characteristics effectively.

In the case of Fontainebleau-Barbeau, the model achieved an R^2 value of 0.7811, indicating strong performance and better generalization compared to other sites like Grillenburg and Svartberget. This higher performance aligns with the earlier trend where the model showed improvement after adapting to similar ecosystems during training, achieving an R^2 value of 0.7690 at Svartberget.

Despite these successes, the model's performance at Grillenburg, with an R^2 value of 0.6565, highlights the challenges in fully generalizing the model across all similar ecosystems. The variability in performance metrics, with mean squared error (MSE) ranging from 2.7404 at Svartberget to 5.6739 at Klingenberg, emphasizes the model's nuanced response to different environmental conditions, even within the same type of ecosystem.

The average performance metrics across these unseen sites reinforce the model's moderate adaptability, with an MSE of 4.5477, root mean squared error (RMSE) of 2.1144, mean absolute error (MAE) of 1.4230, and an R^2 value of 0.7227. These figures demonstrate the model's ability to maintain reasonable accuracy, while also highlighting areas where further model tuning could enhance its robustness and consistency across different ecological settings.

4.2.3   LSTM

Table 4.9 sumarizes the LSTM's performance metrics during the training phase of the unified model. Additionally, Table 4.10 provides insights into the model's performance at an unseen site. The training order of sites differed from the sequence used with XGBoost, significantly affecting the LSTM's performance. Unlike the alphabetical arrangement in XGBoost, rearranging the training sequence had improved the LSTM's performance, suggesting that the initial training sites heavily influenced the model's parameters, potentially biasing learning towards these conditions.

This influence was further evidenced by the performance degradation observed when testing data from previously trained sites. Specifically, Hohes Holz and Estrees Mons A28 showed worsening metrics in subsequent training phases. For instance, Hohes Holz experienced an increase in MSE, RMSE, and MAE, with the third training yielding an MSE of 21.22 and a low R^2 of 0.28. Similarly, Hyytiala exhibited poor metrics in the third training phase. This decline could be attributed to the model excessively adapting to Torgnon, which had initially demonstrated superior performance with an R^2 of 0.87. Such patterns suggest a bias towards more recently trained sites, which overshadowed the learnings from earlier sites.

Despite recovery in performance for some sites in the fourth training phase, Torgnon's metrics contrasted sharply, with MSE escalating from 1.09 to 14.88 and the R^2 turning negative, indicating poor regression performance as discussed in [103]. This inconsistency across different sites pointed to the model's susceptibility to catastrophic forgetting, where it failed to retain previously learned information upon new data exposure [104].

Table 4.9: Performance metrics by site and method

Site	Metrics	1st Train	2nd Train	3rd Train	4th Train
Hohes Holz	MSE	7.57	9.18	21.22	6.90
	RMSE	2.75	3.03	4.61	2.63
	MAE	1.87	2.02	3.01	1.69
	R^2	0.74	0.69	0.28	0.77
Estrees Mons A28	MSE		9.77	15.86	7.41
	RMSE		3.13	3.98	2.72
	MAE		2.01	2.70	2.02
	R^2		0.48	0.15	0.60
Torgnon	MSE			1.10	14.89
	RMSE			1.05	3.86
	MAE			0.61	3.34
	R^2			0.87	-0.70
Hyytiala	MSE				2.35
	RMSE				1.53
	MAE				0.99
	R^2				0.79
Average MSE					7.89
Average RMSE					2.69
Average MAE					2.01
Average R^2					0.36

Table 4.10: Performance metrics of LSTM in testing phase

Site	MSE	RMSE	MAE	R^2
Klingenberg	7.34	2.71	1.80	0.59
Fontainebleau-Barbeau	4.99	2.23	1.60	0.76
Svartberget	3.16	1.78	1.10	0.73
Grillenburg	6.59	2.57	1.78	0.56
Average	5.52	2.32	1.57	0.66

According to [105], freezing certain units from a previously trained model during subsequent training sessions had been shown to preserve learned information and prevent overfitting to new data. Additionally, studies in [106] suggested that regularization might mitigate catastrophic forgetting in neural networks. Thus, to address this issue, mitigation actions were taken by modifying the network architecture during the training process. Instead of using the same architecture in all training phases, the LSTM layer from the previous training was preserved to retain information from the previous training sites, and a new LSTM layer and a dropout layer were added as forms of regularization.

However, in this thesis, this mitigation strategy failed to improve performance. The model continued to exhibit signs of forgetting previously learned information, particularly during the training phase. This phenomenon is, documented in the literature, where regularization alone proved insufficient to fully counteract catastrophic forgetting in neural networks [107]. Therefore, the author suggested further exploration to overcome catastrophic forgetting, potentially exploring more varied architecture either through deeper layering or additional forms of regularization besides dropout layers.

Despite the poor average evaluation metrics during the final training phase, with an

R^2 of 0.3649, the model showed improved performance when used on the testing sites. This suggests that, although the model has difficulty generalizing across training sites, it still has the ability to adapt and perform reasonably well in entirely new environments. The unified LSTM model outperformed the unified SARIMAX model at the unseen site, achieving an average R^2 of 0.66. Similar to XGBoost, the LSTM performed better for deciduous broadleaf forest and evergreen needleleaf forest ecosystem types at the testing sites. Unlike SARIMAX, adding global variables did not improve the performance of XGBoost and LSTM. This lack of improvement could be due to the model considering global variables as less important, especially in the initial training phases, where the values remained the same within the same site.

4.3 Comparison

Figures 4.14-4.16 depict the observed GPP and predicted GPP using the SARIMAX, XG-Boost, and LSTM unified models at the Klingenberg, Fontainebleau-Barbeau, Svarberget, and Grillenburg sites, respectively. In Klingenberg, XGBoost demonstrated the best performance with the highest R^2 value of 0.6840 and the lowest MSE, RMAE, and MAE of 5.67, 2.38, 1.43, respectively. This affirms its effectiveness in capturing trends. However, it faced challenges in accurately predicting the highest GPP peaks, a difficulty also encountered by SARIMAX and LSTM. SARIMAX notably struggled to accurately predict GPP values, occasionally resulting in negative values and challenges in managing amplitude fluctuations. Additionally, there were cases of overestimated patterns by LSTM in the month of September in most of the years (see Figure 4.14).
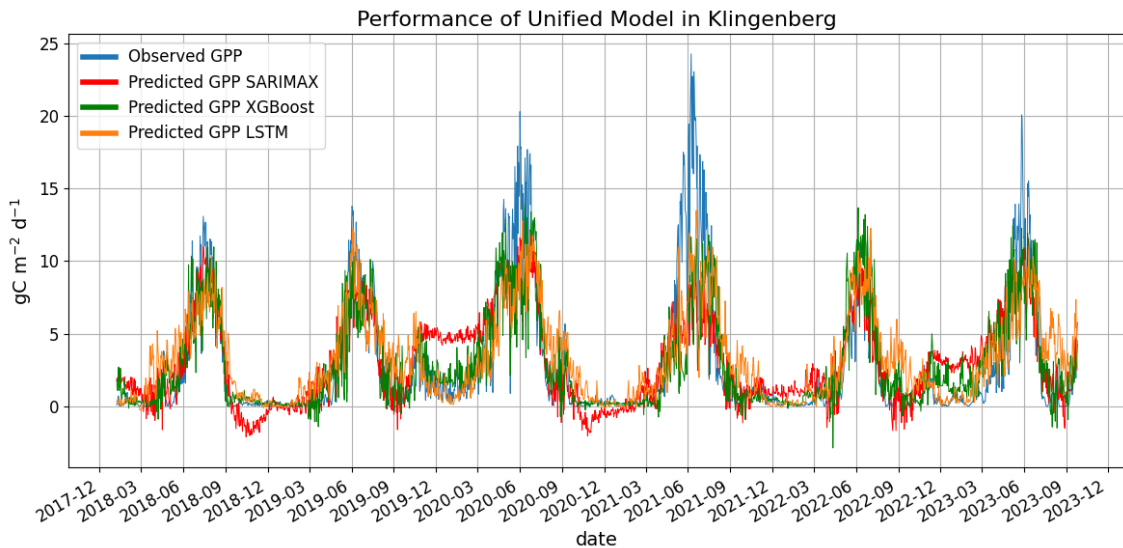

FIGURE 4.14: Predicted vs. Observed Value of GPP from Klingenberg. Observed GPP in blue, SARIMAX predictions in red, XGBoost predictions in green, and LSTM predictions in orange.

The unified model for Fontainebleau-Barbeau demonstrated good performance from all models, with R^2 values higher than 0.76, showing their adaptability to seasonal patterns and fluctuations in GPP data. However, all methods struggled in predicting GPP values at the beginning of 2019 and tended to underestimate. SARIMAX also showed inaccurate predictions for the decreasing values in 2019 and 2022.
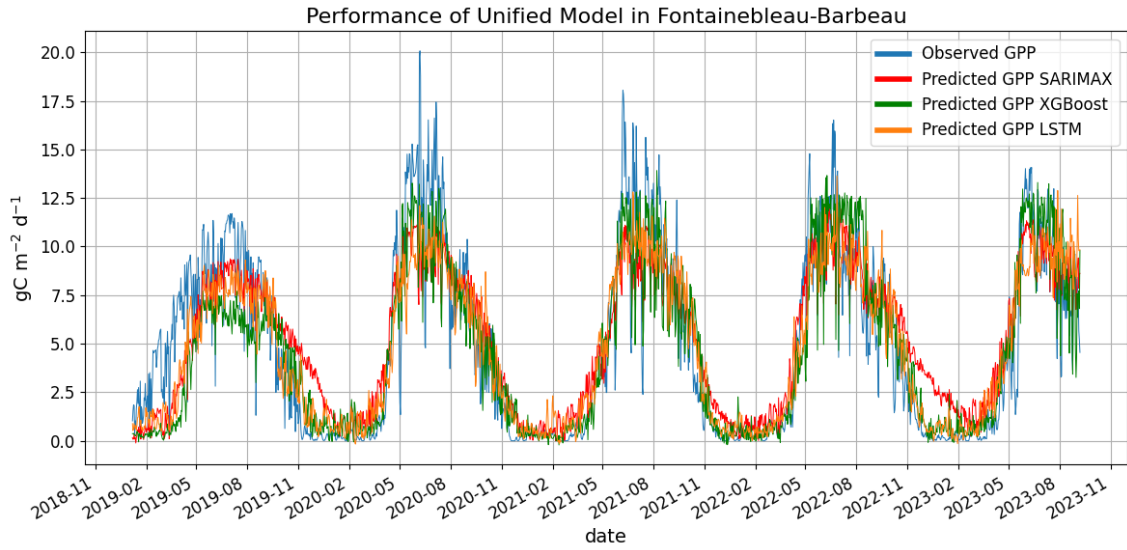
FIGURE 4.15: Predicted vs. Observed Value of GPP from Fontainebleau-Barbeau. Observed GPP in blue, SARIMAX predictions in red, XGBoost predictions in green, and LSTM predictions in orange.
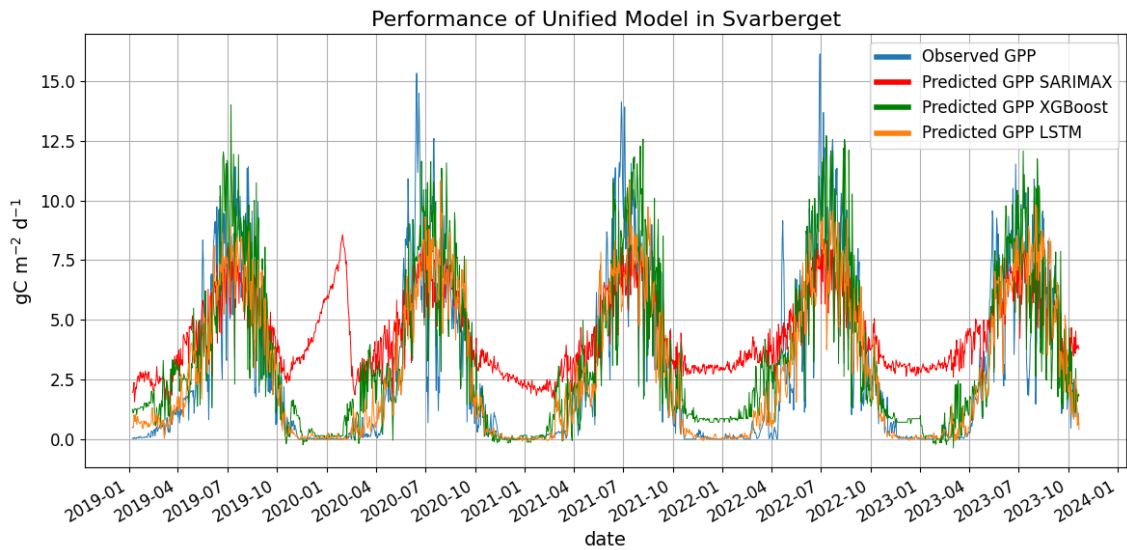

FIGURE 4.16: Predicted vs. Observed Value of GPP from Svartberget. Observed GPP in blue, SARIMAX predictions in red, XGBoost predictions in green, and LSTM predictions in orange.

At Svartberget, XGBoost and LSTM showed strong performances, with R^2 values of 0.77 and 0.73, respectively. XGBoost tended to overestimate low GPP values from October 2021 until April 2024. SARIMAX's performance was notably poor, overestimating low GPP values and incorrectly predicting peaks between October 2019 and April 2020.

In the case of Grillenburg, XGBoost achieved the best evaluation metrics among the other methods. Although it occasionally predicted higher values than the other methods, it struggled with predicting peak values. Both XGBoost and SARIMAX tended to overestimate low GPP values while LSTM more accurately predicted the low values, despite missing some peak values.
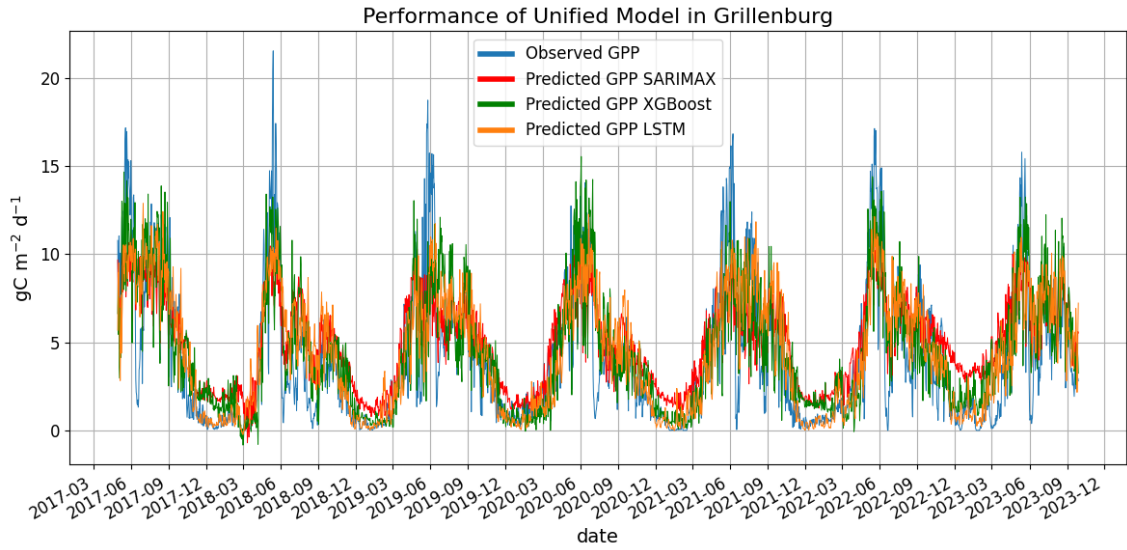
47

FIGURE 4.17: Predicted vs. Observed Value of GPP from Grillenburg. Observed GPP in blue, SARIMAX predictions in red, XGBoost predictions in green, and LSTM predictions in orange.

Overall, the models demonstrated variable success in predicting GPP values across different ecosystem types. XGBoost consistently outperformed others in terms of R^2, achieving the highest average R^2 value of 0.72 in unseen sites, indicating robust performance in various conditions despite its tendency to underestimate peak and overestimate low GPP values during off-peak seasons. LSTM excelled in predicting lower GPP values more accurately than the other models. SARIMAX, however, struggled with peak predictions and inappropriate negative value predictions, indicating limitations in handling non-linear, complex patterns typical of ecological GPP data. While XGBoost and LSTM offer potential tools for predicting GPP values, SARIMAX may require substantial adjustments or restructuring for such applications. All unified models generally performed well, especially in natural forest environments like deciduous broadleaf sites, with R^2 values above 0.76, showcasing adaptability to seasonal GPP variations. SARIMAX performed poorly in evergreen needleleaf forests, while both LSTM and XGBoost required improvements for better predictions in grassland sites. This highlights the need for further refinement to enhance the models' ability to handle variability within a unified framework, preserving the novelty of predicting GPP across diverse ecosystems.

Chapter 5

Conclusion

In this study, predictive models for quantifying GPP in various ecosystem types across Europe were developed, marking an initial attempt at modeling GPP values across diverse ecosystems. This represents a novel approach in the field of environmental modeling. These models were based on SARIMAX, XGBoost, and LSTM methods and their performances were evaluated and compared. This was achieved by combining in-situ measurements and remote sensing data. The study consisted of two main stages: building site-specific models and a unified model for all sites. The earlier stage aimed to gain a better understanding of each site and to explore hyperparameters in order to achieve the best performance. The second phase was designed to build a general model that could predict the GPP in various ecosystem sites without the need for further adjustment by specific sites. The findings are discussed in relation to the posed research questions as follows:

1. **Effectiveness at Individual Model Stage**: The results indicate that all three methods performed similarly in predicting GPP at the site-specific level. Specifically, the XGBoost method consistently outperformed the others in terms of prediction accuracy, with R^2 values more than 0.9 in three out of four sites and demonstrating the lowest error metrics. Although SARIMAX showed better performance than LSTM in some sites, the LSTM model exhibited potential in capturing extreme GPP values in cropland areas.

2. **Performance of Unified Model**: When models were trained using combined data from multiple sites, XGBoost again demonstrated the best performance with an average R^2 of 0.7413 and average $RMSE$ of 1.9516 during the training phase. In contrast, SARIMAX performed poorly at the ENF site, while LSTM continued to struggle with catastrophic forgetting.

3. **Contribution of Data Indicators**: It was found that different types of remote sensing and in-situ data indicators had varying impact on model performance across different ecosystems. The study revealed that XGBoost and LSTM did not show improvement with the inclusion of global variables such as the location of the site, season, and ecosystem type, indicating that these factors were either adequately captured by other variables or less critical than expected in influencing GPP prediction in different sites. The feature importance of XGBoost, based on the gain, highlighted Latent Heat flux as an important factor, especially in cropland and DBF. In ENF, Air temperature was identified as the most important feature, and CIr was important in Grassland. Additionally, MNDVI, EVI, and EVI2 also appeared in the feature importance in various sites.

4. **Model Adaptability in Test Sites**: All methods faced challenges in accurately predicting extreme values, particularly at peak levels. SARIMAX and XGBoost tended to overestimate low GPP values, with SARIMAX performing poorly in predicting peak patterns. Despite challenges in training, LSTM showed better adaptability in test sites, indicating potential for generalizing across untrained locations. However, XGBoost remained the most effective model in predicting GPP in unseen sites based on the evaluation metrics.

The findings from this study highlight the potential of employing machine learning to build a unified model for predicting GPP values across various sites by combining in-situ measurements and remote sensing data. This approach has significant implications for environmental management, particularly in optimizing resource allocation, improving biodiversity conservation strategies, and enhancing climate change mitigation efforts. Site-specific models showed state-of-the-art performance with R^2 values of 0.91, especially in non-cropland sites. Croplands were particularly challenging due to their high GPP range and less natural variability, making accurate predictions difficult. In the unified model, deciduous broadleaf forests exhibited good performance across all methods, while SARIMAX performed poorly in evergreen needleleaf forests. Both LSTM and XGBoost required improvements for better predictions in grassland sites.

While the models demonstrated promising results, their performance can still be improved. Developing strategies to mitigate catastrophic forgetting in LSTM models is crucial, potentially through more sophisticated regularization techniques or revisiting training sessions to reinforce previous data. Further exploration and possibly integrating additional ecosystem types are also needed to enhance generalization. Future research should explore the scalability of these models in global ecosystems beyond the European context. Extending the validation of these models across a broader range of geographical locations and ecosystem types would help refine their applicability and reliability.

Bibliography

[1] Q. Lu, H. Liu, L. Wei, *et al.*, "Global prediction of gross primary productivity under future climate change," *Science of The Total Environment*, vol. 912, p. 169239, 2024.

[2] A. R. Weiskittel, N. L. Crookston, and P. J. Radtke, "Linking Climate, Gross Primary Productivity, and Site Index across Forests of the Western United States," *Canadian Journal of Forest Research*, vol. 41, no. 8, pp. 1710–1721, 2011.

[3] L. Yu, H. Wang, G. Wang, *et al.*, "A Comprehensive Review on Detection of Cyber-Attacks: Data Sets, Methods, Challenges, and Future Research Directions," *Environmental pollution*, vol. 181, pp. 81–90, 2013.

[4] D. Papale, M. Reichstein, M. Aubinet, *et al.*, "Towards a Standardized Processing of Net Ecosystem Exchange Measured with Eddy Covariance Technique: Algorithms and Uncertainty Estimation," 2006.

[5] C. Perry and L. Lautenschlager, "Functional Equivalence of Spectral Vegetation Indices," *Remote Sensing of Environment*, vol. 14, pp. 169–182, 1984.

[6] N. Silleos, T. Alexandridis, I. Gitas, *et al.*, "Vegetation Indices: Advances Made in Biomass Estimation and Vegetation Monitoring in the Last 30 Years," *Geocarto International*, vol. 21, pp. 21 – 28, 2006.

[7] D. E. Pabon-Moreno, M. Migliavacca, M. Reichstein, *et al.*, "On the Potential of Sentinel-2 for Estimating Gross Primary Production," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[8] A. Spinosa, M. A. Fuentes-Monjaraz, and G. E. Serafy, "Assessing the Use of Sentinel-2 Data for Spatio-Temporal Upscaling of Flux Tower Gross Primary Productivity Measurements," *Remote Sensing*, vol. 15, 2 2023.

[9] H. Astola, T. Häme, L. Sirro, *et al.*, "Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region," *Remote Sensing of Environment*, vol. 223, pp. 257–273, 3 2019.

[10] Z. Cai, S. Junttila, J. Holst, *et al.*, "Modelling Daily Gross Primary Productivity with Sentinel-2 Data in the Nordic Region–Comparison with Data from MODIS," *Remote Sensing*, vol. 13, no. 3, 2021.

[11] Y. Chen, X. Xu, C. Huang, *et al.*, "Selection of Prediction Factors of Gross Primary Productivity Based on Artificial Neural Network," in *2022 International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIIPCC)*, pp. 426–429, 2022.

[12] H. Wang, W. Shao, Y. Hu, *et al.*, "Assessment of Six Machine Learning Methods for Predicting Gross Primary Productivity in Grassland," *Remote Sensing*, vol. 15, no. 14, 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[13] X. Chang, Y. Xing, W. Gong, *et al.*, "Evaluating Gross Primary Productivity over 9 ChinaFlux Sites Based on Random Forest Regression Models, Remote Sensing, and Eddy Covariance Data," *Science of The Total Environment*, vol. 875, p. 162601, 2023.

[14] D. P. Sarkar, B. Uma Shankar, and B. Ranjan Parida, "A Novel Approach for Retrieving GPP of Evergreen Forest Regions of India Using Random Forest Regression," *Remote Sensing Applications: Society and Environment*, vol. 33, p. 101116, 2024.

[15] Y. Zhou, T. Hilker, W. Ju, *et al.*, "Modeling Gross Primary Production for Sunlit and Shaded Canopies Across an Evergreen and a Deciduous Site in Canada," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 1859–1873, 4 2017.

[16] X. Yu, Z. Wu, and X. Guo, "Investigating the Potential of GIMMS and MODIS NDVI Data Sets for Estimating Gross Primary Productivity in Harvard Forest," in *MultiTemp 2013: 7th International Workshop on the Analysis of Multi-temporal Remote Sensing Images*, pp. 1–4, 2013.

[17] M. S. Biudes, G. L. Vourlitis, M. C. S. Velasque, *et al.*, "Gross Primary Productivity of Brazilian Savanna (Cerrado) Estimated by Different Remote Sensing-Based Models," *Agricultural and Forest Meteorology*, vol. 307, p. 108456, 2021.

[18] R. Tsay, "Analysis of Financial Time Series," *Technometrics*, vol. 48, pp. 316 – 316, 2005.

[19] S. Turner, A. Karahalios, A. Forbes, *et al.*, "Design Characteristics and Statistical Methods Used in Interrupted Time Series Studies Evaluating Public Health Interventions: A Review," *Journal of clinical epidemiology*, 2020.

[20] J. Fattah, L. Ezzine, Z. Aman, *et al.*, "Forecasting of Demand Using ARIMA Model," *International Journal of Engineering Business Management*, vol. 10, 2018.

[21] Y. Lu, "Crime Prediction Utilizing ARIMA Model," *BCP Business Management*, 2023.

[22] D. Lee, D. Lee, M. jae Choi, *et al.*, "Prediction of Network Throughput using ARIMA," *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 1–5, 2020.

[23] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day, 1970.

[24] M. M. Fathi, A. G. Awadallah, A. M. Abdelbaki, *et al.*, "A New Budyko Framework Extension Using Time Series SARIMAX Model, journal = Journal of Hydrology," vol. 570, pp. 827–838, 2019.

[25] Y. Guo, X. Lai, and M. Gan, "Cyanobacterial Biomass Prediction in a Shallow Lake Using the Time Series SARIMAX Models," *Ecological Informatics*, vol. 78, p. 102292, 2023.

[26] T. D. Tolcha, "The State of Africa's Air Transport Market Amid COVID-19, and Forecasts for Recovery," *Journal of Air Transport Management*, vol. 108, p. 102380, 2023.

[27] N. Kumar, V. Jain, K. Joshi, *et al.*, "Prediction of Epidemic Disease Cases Using ARIMA and SARIMAX Models, year=2023," in *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, pp. 201–205.

[28] G. J. Streefland, F. Herrema, and M. Martini, "A Gradient Boosting Model to Predict the Milk Production," *Smart Agricultural Technology*, vol. 6, 12 2023.

[29] M. Niazkar, A. Menapace, B. Brentan, *et al.*, "Applications of XGBoost in Water Resources Engineering: A Systematic Literature Review," *Environmental Modelling Software*, vol. 174, p. 105971, 2024.

[30] W. A. Gadzama, D. Gabi, M. S. Argungu, *et al.*, "The Use of Machine Learning and Deep Learning Models in Detecting Depression on Social Media: A Systematic Literature Review," *Personalized Medicine in Psychiatry*, vol. 45-46, p. 100125, 2024.

[31] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar, *et al.*, "State-of-the-Art Review on Energy and Load Forecasting in Microgrids Using Artificial Neural Networks, Machine Learning, and Deep Learning Techniques, journal = Electric Power Systems Research," vol. 225, p. 109792, 2023.

[32] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug 2016.

[33] J. Liu, Y. Zuo, N. Wang, *et al.*, "Comparative Analysis of Two Machine Learning Algorithms in Predicting Site-Level Net Ecosystem Exchange in Major Biomes," *Remote Sensing*, vol. 13, no. 12, 2021.

[34] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] M. Fennessy and J. Cronk, "Primary Production and Respiration: Ecological Processes in Wetlands," in *The Wetland Book* (C. Finlayson *et al.*, eds.), Dordrecht: Springer, 2018.

[36] B. Jagannathan and J. Golbeck, "Photosynthesis: Microbial," in *Encyclopedia of Microbiology (Third Edition)* (M. Schaechter, ed.), pp. 325–341, Oxford: Academic Press, third edition ed., 2009.

[37] N. Azbar and D. Levin, "6.48 - Biohydrogen Production from Agricultural Agrofood-Based Resources," in *Comprehensive Biotechnology (Second Edition)* (M. Moo-Young, ed.), pp. 629–641, Burlington: Academic Press, second edition ed., 2011.

[38] P. Deb Burman, N. Shurpali, S. Chowdhuri, *et al.*, "Eddy Covariance Measurements of CO2 Exchange from Agro-ecosystems Located in Subtropical (India) and Boreal (Finland) Climatic Conditions," *Journal of Earth System Science*, vol. 129, no. 43, p. 1, 2020.

[39] D. Baldocchi, E. Falge, L. Gu, *et al.*, "FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and

Energy Flux Densities," *Bulletin of the American Meteorological Society*, vol. 82, pp. 2415–2434, 2001.

[40] European Space Agency, "Sentinel-2: Observing the Earth." https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2, 2023. Accessed: 2023-05-14.

[41] European Space Agency, *Sentinel-2 MSI User Guide - Processing Levels*, 2024. Accessed: 2024-05-14.

[42] European Space Agency, *Sentinel-2 MSI Technical Guide - Level-2A Algorithms and Products*, 2024. Accessed: 2024-05-14.

[43] Google Developers, "Copernicus Sentinel-2 Surface Reflectance Harmonized Dataset." https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED, 2023. Accessed: 2023-05-14.

[44] T. Zhang, J. Zhou, P. Yu, *et al.*, "Response of Ecosystem Gross Primary Productivity to Drought in Northern China Based on Multi-Source Remote Sensing Data," *Journal of Hydrology*, vol. 616, p. 128808, 2023.

[45] C. Wu, Z. Niu, Q. Tang, *et al.*, "Remote Estimation of Gross Primary Production in Wheat Using Chlorophyll-Related Vegetation Indices," *Agricultural and Forest Meteorology*, vol. 149, no. 6-7, pp. 1015–1021, 2009.

[46] O. Mutanga and A. K. Skidmore, "Narrow Band Vegetation Indices Overcome the Saturation Problem in Biomass Estimation," *International Journal of Remote Sensing*, vol. 25, no. 19, pp. 3999–4014, 2004.

[47] C. Jurgens, "The Modified Normalized Difference Vegetation Index (mNDVI): A New Index to Determine Frost Damages in Agriculture Based on Landsat TM Data," *International Journal of Remote Sensing*, vol. 18, no. 17, pp. 3583–3594, 1997.

[48] K. Noumonvi, M. Ferlan, K. Eler, *et al.*, "Estimation of Carbon Fluxes from Eddy Covariance Data and Satellite-Derived Vegetation Indices in a Karst Grassland (Podgorski Kras, Slovenia)," *Remote Sensing*, vol. 11, no. 6, p. 649, 2019.

[49] Z. Jiang, A. R. Huete, K. Didan, *et al.*, "Development of a Two-Band Enhanced Vegetation Index Without a Blue Band," *Remote Sensing of Environment*, vol. 112, pp. 3833–3845, 2008.

[50] Y. Peng and A. A. Gitelson, "Remote Estimation of Gross Primary Productivity in Soybean and Maize Based on Total Crop Chlorophyll Content," *Remote Sensing of Environment*, vol. 117, pp. 440–448, 2012.

[51] A. A. Gitelson, A. Viña, S. B. Verma, *et al.*, "Relationship Between Gross Primary Production and Chlorophyll Content in Crops: Implications for the Synoptic Monitoring of Vegetation Productivity," *Journal of Geophysical Research*, vol. 111, no. D08S11, 2006.

[52] G. Fu, Z. Shen, X. Zhang, *et al.*, "Modeling Gross Primary Productivity of Alpine Meadow in the Northern Tibet Plateau by Using MODIS Images and Climate Data," *Acta Ecologica Sinica*, vol. 30, pp. 264–269, 2010.

[53] M. J. Hill, "Vegetation Index Suites as Indicators of Vegetation State in Grassland and Savanna: An Analysis with Simulated SENTINEL 2 Data for a North American Transect," *Remote Sensing of Environment*, vol. 137, pp. 94–111, 2013.

[54] B. cai Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sensing of Environment*, vol. 58, no. 3, pp. 257–266, 1996.

[55] X. Han-qiu, "A Study on Information Extraction of Water Body with the Modified Normalized Difference Water Index (MNDWI)," *Journal of Remote Sensing*, no. 5, pp. 589–595, 2005.

[56] K. V. Singh, R. Setia, S. Sahoo, *et al.*, "Evaluation of NDWI and MNDWI for Assessment of Waterlogging by Integrating Digital Elevation Model and Groundwater Level," *Geocarto International*, vol. 30, no. 6, pp. 650–661, 2014.

[57] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Switzerland: Springer, 3 ed., 2016.

[58] O. Ryan, J. M. B. Haslbeck, and L. Waldorp, "Non-Stationarity in Time-Series Analysis: Modeling Stochastic and Deterministic Trends," Jul 2023.

[59] R. Harris, "Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test," *Economics Letters*, vol. 38, no. 4, pp. 381–386, 1992.

[60] J. Korstanje, *The AR Model*, pp. 45–69. Berkeley, CA: Apress, 2021.

[61] S. A. hady Soliman and A. M. Al-Kandari, "9 - Electric Load Modeling for Long-Term Forecasting," in *Electrical Load Forecasting* (S. A. hady Soliman and A. M. Al-Kandari, eds.), pp. 353–406, Boston: Butterworth-Heinemann, 2010.

[62] J. Korstanje, *The SARIMAX Model*, pp. 125–131. Berkeley, CA: Apress, 2021.

[63] W. H. Tang and A. Röllin, "Model identification for ARMA time series through convolutional neural networks," *Decision Support Systems*, vol. 146, p. 113544, 2021.

[64] A. H. Shahid and M. Singh, "Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 3, pp. 638–672, 2019.

[65] T. Bellini, "Chapter 2 - One-Year PD," in *IFRS 9 and CECL Credit Risk Modelling and Validation* (T. Bellini, ed.), pp. 31–89, Academic Press, 2019.

[66] S. Mutasa, S. Sun, and R. Ha, "Understanding artificial intelligence based radiology studies: What is overfitting?," *Clinical Imaging*, vol. 65, pp. 96–99, 2020.

[67] J. Subramanian and R. Simon, "Overfitting in prediction models – Is it a problem only in high dimensions?," *Contemporary Clinical Trials*, vol. 36, no. 2, pp. 636–641, 2013.

[68] J. Luo, Z. Zhang, Y. Fu, *et al.*, "Time Series Prediction of COVID-19 Transmission in America Using LSTM and XGBoost Algorithms," *Results in Physics*, vol. 27, p. 104462, 2021.

[69] H. Ahmetoglu and R. Das, "A Comprehensive Review on Detection of Cyber-Attacks: Data Sets, Methods, Challenges, and Future Research Directions," *Internet of Things*, vol. 20, p. 100615, 2022.

[70] N. Zhai, P. Yao, and X. Zhou, "Multivariate Time Series Forecast in Industrial Process Based on XGBoost and GRU," pp. 978–979.

[71] C. Zhang, D. Wang, L. Wang, *et al.*, "Cause-Aware Failure Detection Using an Interpretable XGBoost for Optical Networks," *Opt. Express*, vol. 29, pp. 31974–31992, Sep 2021.

[72] S. Kalogirou, "Neural Network Modeling of Energy Systems," in *Reference Module in Earth Systems and Environmental Sciences*, Elsevier, 2013.

[73] N. Alsadi, S. A. Gadsden, and J. Yawney, "Intelligent estimation: A review of theory, applications, and recent advances," *Digital Signal Processing*, vol. 135, p. 103966, 2023.

[74] Z. Han, J. Zhao, H. Leung, *et al.*, "A Review of Deep Learning Models for Time Series Prediction," *IEEE Sensors Journal*, vol. 21, pp. 7833–7848, Mar 2021.

[75] A. P. M. Diniz, P. M. Ciarelli, E. O. T. Salles, *et al.*, "Use of Deep Neural Networks for Clogging Detection in the Submerged Entry Nozzle of the Continuous Casting," *Expert Systems with Applications*, vol. 238, p. 121963, 2024.

[76] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, *et al.*, "4 - Selected Approaches to Supervised Learning," in *Computational Learning Approaches to Data Analytics in Biomedical Applications* (K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, eds.), pp. 101–123, Academic Press, 2020.

[77] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. http://www.deeplearningbook.org.

[78] A. Ahmad, S. Ismail, and D. Samaon, "Recurrent Neural Network with Backpropagation Through Time for Speech Recognition," in *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004.*, vol. 1, pp. 98–102 vol.1, 2004.

[79] Integrated Carbon Observation System, "ICOS Community Portal." https://www.icos-cp.eu/. Accessed: 2023-05-14.

[80] Integrated Carbon Observation System, "ES_FR-EM2 Station - ICOS Carbon Portal." https://meta.icos-cp.eu/resources/stations/ES_FR-EM2, 2024. Accessed: 2024-06-10.

[81] Integrated Carbon Observation System, "ES_DE-HoH Station - ICOS Carbon Portal." https://meta.icos-cp.eu/resources/stations/ES_DE-HoH, 2024. Accessed: 2024-06-10.

[82] Integrated Carbon Observation System, "ICOS Station FI-Hyy." https://meta.icos-cp.eu/resources/stations/ES_FI-Hyy, 2024. Accessed: 7 May 2024.

[83] Integrated Carbon Observation System, "ICOS Station IT-Tor." https://meta.icos-cp.eu/resources/stations/ES_IT-Tor, 2024. Accessed: 7 May 2024.

[84] Integrated Carbon Observation System, "DE-Kli Station Information." `https://meta.icos-cp.eu/resources/stations/ES_DE-Kli`, 2024. Accessed: June 28, 2024.

[85] Integrated Carbon Observation System, "FR-Fon Station Information." `https://meta.icos-cp.eu/resources/stations/ES_FR-Fon`, 2024. Accessed: June 28, 2024.

[86] Integrated Carbon Observation System, "SE-Svb Station Information." `https://meta.icos-cp.eu/resources/stations/ES_SE-Svb`, 2024. Accessed: June 28, 2024.

[87] Integrated Carbon Observation System, "DE-Gri Station Information." `https://meta.icos-cp.eu/resources/stations/ES_DE-Gri`, 2024. Accessed: June 28, 2024.

[88] G. Pastorello, C. Trotta, E. Canfora, *et al.*, "The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data," *Scientific Data 2020 7:1*, vol. 7, pp. 1–27, jul 2020.

[89] Copernicus Sentinel Hub, "Sentinel-2 Level-2A Data Documentation." Copernicus Sentinel Hub Documentation, 2024. Accessed: June 28, 2024.

[90] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Analytical Chemistry*, vol. 36, pp. 1627–1639, July 1964. Publisher: American Chemical Society.

[91] "SARIMAX — statsmodels 0.13.2 documentation." `https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html`, 2024. Accessed: 2024-06-10.

[92] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python," in *9th Python in Science Conference*, 2010.

[93] "XGBoost Python Package." `https://xgboost.readthedocs.io/en/stable/python/index.html`, 2024. Accessed: 2024-06-10.

[94] "LSTM layer - Keras Documentation." `https://keras.io/api/layers/recurrent_layers/lstm/`, 2024. Accessed: 2024-06-10.

[95] L. Alzubaidi, J. Zhang, A. Humaidi, *et al.*, "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions," *Journal of Big Data*, vol. 8, p. 53, 2021.

[96] T. Chai and R. Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)," *Geoscientific Model Development Discussions*, vol. 7, pp. 1525–1534, 2014.

[97] S. M. Ross, "CHAPTER 12 - Linear Regression," in *Introductory Statistics (Third Edition)* (S. M. Ross, ed.), pp. 537–604, Boston: Academic Press, third edition ed., 2010.

[98] C. Hu, S. Hu, L. Zeng, *et al.*, "Estimation of Daily Maize Gross Primary Productivity by Considering Specific Leaf Nitrogen and Phenology via Machine Learning Methods," *Remote Sensing*, vol. 16, no. 2, 2024.

[99] A. Tiwari, "Chapter 2 - Supervised Learning: From Theory to Applications," in *Artificial Intelligence and Machine Learning for EDGE Computing* (R. Pandey, S. K. Khatri, N. kumar Singh, and P. Verma, eds.), pp. 23–32, Academic Press, 2022.

[100] S. H. Wu, P.-E. Jansson, and P. Kolari, "The role of air and soil temperature in the seasonality of photosynthesis and transpiration in a boreal Scots pine ecosystem," *Agricultural and Forest Meteorology*, vol. 156, pp. 85–103, 2012.

[101] S. Lin, J. Li, Q. Liu, *et al.*, "Evaluating the Effectiveness of Using Vegetation Indices Based on Red-Edge Reflectance from Sentinel-2 to Estimate Gross Primary Productivity, journal = Remote Sensing," vol. 11, no. 11, 2019.

[102] Y. Wang, R. Li, J. Hu, *et al.*, "Daily estimation of gross primary production under all sky using a light use efficiency model coupled with satellite passive microwave measurements," *Remote Sensing of Environment*, vol. 267, p. 112721, 2021.

[103] D. Chicco, M. Warrens, and G. Jurman, "The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE, and RMSE in Regression Analysis Evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021.

[104] H. Jung, J. Ju, M. Jung, *et al.*, "Less-forgetting Learning in Deep Neural Networks," *ArXiv*, vol. abs/1607.00122, 2016.

[105] S. Ede, S. Baghdadlian, L. Weber, *et al.*, "Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI," 2022.

[106] A. E. Khatib and F. Karray, "Preempting Catastrophic Forgetting in Continual Learning Models by Anticipatory Regularization," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2019.

[107] R. Kemker, M. McClure, A. Abitino, *et al.*, "Measuring catastrophic forgetting in neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018.

Appendix

A1 Example of SARIMAX

To illustrate the application of the SARIMAX model, consider a scenario where the objective is to model and forecast the daily Gross Primary Production (GPP) of an ecosystem. The GPP data exhibits seasonal patterns, with peaks during specific times of the year due to variations in sunlight and temperature. Additionally, external factors such as daily precipitation and chlorophyll content are believed to significantly influence GPP.

In this context, suppose the non-seasonal orders are specified as $(p, d, q) = (1, 1, 1)$. Here, $p = 1$ indicates that the model includes one autoregressive (AR) term, meaning the model uses the previous day's value (X_{t-1}) to predict the current day's value. The parameter $d = 1$ signifies that the time series needs to be differenced once to achieve stationarity, thereby removing any trends or non-stationary components. The parameter $q = 1$ denotes that the model incorporates one moving average (MA) term, meaning the model uses the previous day's forecast error (ϵ_{t-1}) to adjust the prediction for the current day's value.

For the seasonal component, the parameters are $(P, D, Q, s) = (1, 1, 1, 7)$, where $P = 1$ indicates that the model includes one seasonal autoregressive term, meaning the model uses the value from the same day of the previous week (X_{t-7}) to account for seasonal patterns. The parameter $D = 1$ means that seasonal differencing is applied once to remove seasonal patterns, while $Q = 1$ signifies that the model includes one seasonal moving average term, which utilizes the forecast error from the same day of the previous week (ϵ_{t-7}). The seasonal period $s = 7$ corresponds to weekly seasonality in daily data. The exogenous variables, which include daily precipitation (Z_{1t}) and chlorophyll content (Z_{2t}), are represented as:

$$Z_t = \begin{bmatrix} Z_{1t} & Z_{2t} \end{bmatrix}. \tag{1}$$

The vector of coefficients β, which quantifies the impact of each exogenous variable, is given by:

$$\beta = \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}. \tag{2}$$

The linear combination of the exogenous variables and their corresponding coefficients is expressed as:

$$\beta^\top Z_t = \beta_1 Z_{1t} + \beta_2 Z_{2t}. \tag{3}$$

To fit the SARIMAX model, we first apply the differencing operators to achieve stationarity in the time series X_t. The non-seasonal differencing operator is given by Equation (2.14). This operator transforms X_t into:

$$\Delta X_t = (1 - B)X_t = X_t - X_{t-1}. \tag{4}$$

The seasonal differencing operator as in Equation (2.18), where $s = 7$ is the seasonal period. This operator transforms X_t into:

$$\Delta_s X_t = (1 - B^7)X_t = X_t - X_{t-7}. \tag{5}$$

When both non-seasonal and seasonal differencing are applied, the combined differencing operator is $(1 - B)^d (1 - B^s)^D$. For $d = 1$ and $D = 1$, this becomes:

$$\Delta^d \Delta_s^D X_t = (1 - B)(1 - B^7)X_t. \tag{6}$$

Expanding this, we first apply the seasonal differencing:

$$(1 - B^7)X_t = X_t - X_{t-7}. \tag{7}$$

Then, applying the non-seasonal differencing operator to the result:

$$(1 - B) \left[(1 - B^7)X_t \right] = (1 - B) \left[X_t - X_{t-7} \right]. \tag{8}$$

Expanding this, we have:

$$(1 - B) \left[X_t - X_{t-7} \right] = (X_t - X_{t-1}) - (X_{t-7} - X_{t-8}). \tag{9}$$

Thus, the differenced series Y_t after applying both non-seasonal and seasonal differencing is:

$$Y_t = (X_t - X_{t-1}) - (X_{t-7} - X_{t-8}). \tag{10}$$

With the differenced series Y_t now stationary, the SARIMAX model can be expressed as:

$$Y_t = c + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1} + \Phi_1 Y_{t-7} + \Theta_1 \epsilon_{t-7} + \beta_1 Z_{1t} + \beta_2 Z_{2t} + \epsilon_t. \tag{11}$$

Here, c is a constant term, ϕ_1 is the coefficient for the non-seasonal AR term, θ_1 is the coefficient for the non-seasonal MA term, Φ_1 is the coefficient for the seasonal AR term, and Θ_1 is the coefficient for the seasonal MA term. The exogenous variables and their coefficients are included as $\beta^\top Z_t$, where Z_t represents daily precipitation and chlorophyll content, and ϵ_t denotes the error term of the model.