THESIS

TIME SERIES ANALYSIS OVER SPARSE, NON-STATIONARY DATASETS WITH VARIATIONAL MODE DECOMPOSITION AND TRANSFER LEARNING

Submitted by

Katherine Patterson

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2025

Master's Committee:

Advisor: Shrideep Pallickara
Co-Advisor: Sangmi Pallickara

Allan Andales

ABSTRACT

TIME SERIES ANALYSIS OVER SPARSE, NON-STATIONARY DATASETS WITH VARIATIONAL MODE DECOMPOSITION AND TRANSFER LEARNING

Data volumes have been growing exponentially across many domains. However, in fields such as ecology and environmental monitoring, data remains sparse, creating unique challenges. One such challenge is detecting extreme events (sudden spikes or anomalies in the data) and understanding their causes based on spatiotemporal patterns. The difficulty is exacerbated by time lags between an observed outlier and its underlying trigger, making causal attribution and forecasts difficult. These challenges have implications, particularly for environmental protection and regulatory compliance. This thesis explores the issue of time-series analysis over sparse, non-stationary datasets to support outlier detection and forecasts. We mitigate non-stationarity using variational mode decomposition (VMD) to break the signal into multiple seasonal components. To tackle the challenges of long-term seasonality, we leverage information obtained from the frequency domain regarding dominant lagged relationships within these signals. Finally, we leverage transfer learning to warm-start models at spatial extents where the data are sparse. We validate these ideas in the context of nutrient runoff into surface waters, where identifying and explaining anomalies is critical for the protection of ecosystems. Challenges arise due to three main factors: (1) nutrient time series are naturally non-stationary, which complicates the identification of underlying patterns; (2) temporal models often struggle over an entire season's span; and (3) water quality measurements are often sporadic and sparse. Results showed that the historical similarity mapping of these spatiotemporal profiles and their frequency-motivated seasonality characteristics improved prediction performance in each target series. Additionally, the final proposed model captured more series fluctuations than the base models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The continued growth of agricultural yields is essential for maintaining a productive and reliable community. However, achieving and sustaining this growth involves a complex discussion about the proper balance between food security and ecosystem protection. Efforts to enhance agricultural yields often rely on commercial fertilizers and manure that provide additional nutrients, primarily nitrogen and phosphorus, to promote crop growth. Inevitably, these commercial products run off into nearby lakes and freshwater bodies due to storms or excess application. While fertilizers have boosted productivity in many agricultural organizations, nutrient runoff contributes significantly to water pollution. This excess of nutrients is one of the most common causes of environmental degradation in freshwater ecosystems, such as the increase in hypoxic zones or areas with low oxygen levels. In recent years, this increase in hypoxic zones has motivated more research into the ecological impacts of water pollution and the underlying causes of these nutrient imbalances [1]. These hypoxic zones primarily result from a process known as eutrophication, which occurs when water bodies become overly enriched with nutrients. This nutrient overload leads to excessive growth of algae and aquatic plants, reducing oxygen levels and limiting light penetration in the water. Developing a stronger understanding of how agricultural practices contribute to overabundant nutrient levels is essential to protect our ecosystems' health and ensure reliable access to clean water.

Agricultural runoff is classified as a non-point source of pollution, meaning it is difficult to manage or monitor. This difficulty arises because various factors naturally influence nutrient levels in freshwater bodies, such as climate, flow rate, and human activities [2]. The complex relationships among nutrients and their ecosystems suggest that there remains significant potential for improvement despite considerable research in modeling and predicting these interactions. If we can develop a reliable method to predict fluctuations in nutrient levels, we can conclude that the impact of agricultural runoff is monitorable and, thus, manageable.

1.1 Challenges

Building a predictive model for nutrient time series forecasting presents several key challenges, each of which complicates the modeling process in distinct ways. These include:

1. Long-term nutrient data often exhibit nonstationary characteristics i.e. their mean and variance fluctuate over time. This instability makes it more difficult for models to detect consistent patterns or trends, reducing predictive reliability.

2. Temporal models often struggle with information loss as the look-back (or retrospective) window size increases. This limitation weakens the model's ability to capture seasonal trends, particularly those that unfold over extended timeframes.

3. Achieving long-term temporal consistency in water quality measurements at a given location is often challenging. This limited data availability complicates efforts to validate model performance in a consistently defined real-world setting. This places greater demands on both data collection and evaluation strategies.

1.2 Research Questions

The analysis in this paper investigates the following research questions:

RQ-1: *How can we identify spatial extents of interest that are impacted by external events – especially when those effects take time to reveal themselves?* In particular, how can we identify water sources affected by agricultural activities?

RQ-2: *How can we effectively model relationships between sensed variables to forecast outcomes for spatial extents of interest?* In particular, in agricultural settings, we are interested in capturing nonlinear relationships between meteorological conditions and nutrient concentrations at particular water bodies.

RQ-3: How can we leverage this model at scale in the face of data scarcity? Notably, not all spatial extents may have rich measurement data for nutrients of interest.

1.3 Approach Summary

Our methodology is built upon three interconnected elements with each addressing a distinct yet complementary aspect of the modeling process. The first element involves applying data augmentation methods in tandem with techniques designed to mitigate the non-stationarity inherent in the data. This enhances the stability and robustness of the dataset, ensuring that subsequent modeling efforts are less prone to drift and inconsistency. The second element identifies pattern profiles derived from historical data. These profiles capture recurring seasonal trends across spatial extents, providing a basis for informed predictions. With these preparatory steps in place, the third element tackles a critical challenge: overcoming data scarcity when training models over spatial extents. To address this, we employ transfer learning strategies that enable models to be warm-started using models trained over spatial extents that are relatively data rich.

We combine various data augmentation methods with machine learning techniques to overcome common challenges in water quality forecasting. The machine learning model leveraged for prediction is the Long Short-Term Memory (LSTM) model because of its temporal pattern recognition and internal memory structure. The basis of the model design implements a bidirectional LSTM coupled with a cross-attention mechanism between the target sequence and the available input features. We mitigate the challenge of non-stationarity in nutrient time series through variable decomposition, namely Variational Mode Decomposition (VMD), a standard data augmentation method to decompose a signal into its frequency components or trends and increase stationarity. Each signal decomposition trains on individual models, and their predictions' summation serves as the target measure's final prediction. The prediction model additionally leverages an engineered input feature that assists in capturing the seasonal trends of the series. We use the frequency domain and the signal decomposition's dominant frequencies to inform on seasonality. This engineered input feature is thus a moving average calculation of the nutrient series whose sliding window uses the identified dominant frequencies (or dominant lagged relationships) as its window size.

We define spatiotemporal pattern profiles based on historical data to accurately relate the current prediction sequence to its expected seasonality characteristics. These profiles comprise the

available input features and the frequency-motivated nutrient moving average. We train a K-Nearest Neighbors (KNN) model to define a comparison space. The spatiotemporal profiles for the current prediction dataset then leverage these KNN models to identify its nearest historical neighbor. For each point in the current predictive sequence, the KNN mapping results are an additional input feature set that includes the historical pattern's moving average (seasonality) and the signal's corresponding decomposition residual to aid in error correction.

Finally, a significant challenge is accurate prediction in the face of data scarcity, which this research addresses through transfer learning. The transfer learning approach allows for improved prediction in a sparse dataset because the model can leverage previous knowledge from training in a similar environment. Thus, each signal decomposition model is first trained in a data-rich environment unaffected by agricultural runoff, allowing for more consistency in pattern capture. We then transfer this previously learned knowledge to a data-scarce environment affected by agricultural runoff. The new data environment can also leverage the defined spatiotemporal pattern profiles, granting access to a more robust and accurate KNN mapping.

1.4 Thesis Contributions

Our methodology provides a framework for forecasting measurements across spatial extents, particularly in environments where data are scarce. We have validated these ideas in the context of predicting nutrient concentrations in water bodies where there are time-lagged relationships between activities (e.g., fertilization) and nutrient levels. Our specific contributions include the following:

- We have developed a set of quantitatively driven techniques designed to determine whether agricultural activities and runoff have a measurable impact on a given water body.
- We have introduced a mechanism for capturing seasonality while constructing spatiotemporal pattern profiles. This mechanism mitigates the well-known decline in accuracy when temporal models are applied with increasingly long look-back (or retrospective) windows.

- Finally, to enhance nutrient forecasting in data-scarce environments, we employ transfer learning strategies that enhance model training (via warm-starts) and improve the accuracy of spatiotemporal pattern mapping.

1.5 Organization of Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we discuss related works considering both model techniques and data augmentation methods. Chapter 3 describes the methodology behind the model design and implementation. Chapter 4 presents benchmarks for each research question and discusses the model's prediction results. Finally, Chapter 5 provides concluding remarks alongside key takeaways from the performance benchmarks, followed by limitations in the current work and future research opportunities in nutrient time series forecasting.

Chapter 2

Related Works

2.1 Deep Learning Implementations

Due to the complex interactions in water quality relationships, research often utilizes machine learning models for predictive tasks. The Random Forest algorithm has previously been implemented in nutrient forecasting because of its ability to highlight the importance of different features. However, the predictive capabilities of Random Forest and its variants, such as the Boosted Regression Tree, do not inherently capture temporal dependencies, which are essential for water quality prediction. Moreover, tree-based models struggle to account for the relationships between input features themselves in addition to their connections to the target variable [3]. However, the Random Forest algorithm has assisted in the feature selection task for time series, which is useful in cases where a plethora of environmental feature data is available [4].

Deep neural networks, on the other hand, are valuable for capturing temporal dependencies and non-linear relationships among environmental features [5]. Recurrent Neural Networks (RNNs) are desirable because they can retain memory through multiple time steps. One common enhancement of the RNN architecture is the Long Short-Term Memory (LSTM) model, which utilizes information gates to capture longer-term temporal dependencies. Previous studies have shown that LSTM models excel in spatiotemporal predictions [6, 7]. For instance, Hu et al. assessed LSTM predictions within the context of mariculture, the artificial cultivation of marine life. This setting allows us to observe LSTMs in a real-world application while benefiting from consistent measurements over a short period [8]. Additionally, implementing a bidirectional LSTM can improve prediction performance. Bi-directionality is advantageous because the current measurement relates to past and future environmental states. By enabling the model to understand these dual temporal relationships, prediction accuracy can be improved [9, 10]. LSTM-based models continue to show

promise in water quality predictions, leading to various hybrid models that aim to address LSTM shortcomings [11–13].

2.2 Data Augmentation Methods

A key challenge in nutrient time series forecasting is its non-stationarity, as these series often lack consistent patterns due to unpredictable or unmeasured events. Additionally, because LSTMs struggle to identify meaningful long-term patterns as the look-back window size increases, this poses challenges in nutrient forecasting that require capturing seasonal cycles. Data conversion methods, particularly variable decomposition, are commonly used to transform non-stationary series into stationary components to improve prediction accuracy. Research has demonstrated that techniques such as Empirical Mode Decomposition (EMD) can decompose time series into intrinsic mode functions (IMFs) that represent various frequencies and can be utilized for time series forecasting [14–16].

Liu et al. conducted an extensive review of variable decomposition research coupled with wind energy forecasting, highlighting the benefits of variable decomposition strategies in overall model accuracy [17]. Putra et al. demonstrated the strengths of implementing Variational Mode Decomposition (VMD) in time series forecasting, stating that EMD is susceptible to noise and mode mixing, i.e., mixing distinct patterns [18]. This VMD model notably incorporated a linear framework from the LTSF-Linear models proposed by Zeng et al. [19]. LTSF-Linear models are 'embarrassingly simple' one-layer linear models that leverage historical patterns in the data rather than short-term previous knowledge. Results of Putra et al.'s LTSF-Linear related research emphasize the DLinear sub-model, which extracts trends and seasonal components using historical moving averages. These experiments indicate that combining DLinear with VMD yields the highest accuracy, effectively capturing seasonal trends without depending solely on complex models. Thus, leveraging this concept of historical moving averages can compel us to capture and incorporate seasonal trends without expecting complex machine learning models to recognize these long-term patterns.

In a continued consideration of data augmentation, Fallah et al. provide a methodological overview of computational intelligence in forecasting, highlighting four short-term forecasting methods: similar patterns, variable selection, hierarchical forecasting, and weather station selection [20]. These approaches improve forecasting performance, suggesting that focusing on effective data augmentation methods is essential, especially when predicting non-stationary data.

2.3 Spatiotemporal and Scientific Data Management

Managing data and metadata [21] for scientific data collections [22, 23] has been tackled in diverse computational settings such as web services, large-scale peer-to-peer grids [24–26], and collaborative environments [27, 28]. Data storage frameworks specifically targeting spatiotemporal data management issues include systems such as Galileo [29], Trident [30], and Atlas [31, 32]. These frameworks often sit alongside support for queries that are ad hoc [33], geometry constrained [34], and expressive [35]. A key characteristic in such systems is that the data organization schemes (indexes, dispersion, and metadata management) and aligned closely with the data characteristics: geocoding and timestamps associated with each observation. This organization, in turn, is leveraged to support queries at low latencies.

Our methodology does not preclude the use of custom metadata management schemes nor does it place any restrictions on spatiotemporal data management schemes. Finally, though we consider data sparse environments, nothing in our methodology precludes applicability in voluminous data environments. This adaptability ensures that our methodology remains broadly applicable, accommodating a wide range of data landscapes without rigid assumptions.

2.4 Generalization

The question left unanswered thus far is how to enable these models to generalize in the face of data scarcity. A commonly proposed solution to this problem is transfer learning. When a model trains on a set of data, it, in theory, captures the underlying patterns of its testing environment. Suppose we have another dataset with a similar environment but too few data points to effectively

train its model. Can we leverage another model's existing knowledge to aid the smaller dataset's prediction performance? Ma et al. showed the applicability of transfer learning in the temporal domain, aiming to transfer knowledge from a smaller temporal resolution to a larger one. With a bidirectional LSTM-based model, transfer learning minimized error magnitude for larger temporal resolutions [36]. Sarmas et al. conducted a more extensive review of transfer learning in the context of solar power forecasting. This study implemented three strategies of transductive transfer learning where the source and target tasks are the same. Results showed improved precision for each transfer learning method compared with other standard models [37]. Chen et al. demonstrated transfer learning's ability to impute consecutive missing data in the context of water quality [38]. Thus, transfer learning is efficacious in improving a smaller dataset's performance. The capability of generalization continues to be imperative, as long-term or robust datasets will not always be available.

Chapter 3

Methodology

3.1 Optimal Dataset Selection

This research utilizes water quality data from the Environmental Protection Agency (EPA). The EPA provides water quality measurements collected by local governments, non-governmental organizations, and volunteers nationwide. While this broad availability allows for extensive water quality monitoring, it also increases the risk of inconsistencies in monitoring site accuracy, temporal availability, and spatial coverage. The challenge, therefore, is to identify a water body that exhibits the effects of agricultural runoff and has long-term, consistent data availability. To define our search criteria, we queried the EPA for water bodies located in Colorado from 2000 to 2020, providing the opportunity for decades of data. We specifically focused on water bodies with consistent measurements of nitrogen and phosphorus, the primary nutrients in aquatic ecosystems and commonly found in agricultural fertilizers. Additionally, we utilize gridMET data to incorporate weather features for prediction assistance. The input features included from EPA and gridMET are listed in Table 3.1.

Table 3.1: Model Input Features

Input Feature	Unit
Specific Conductance	us/cm
Water Temperature	deg C
Precipitation	mm daily total
Wind Speed	m/s
Minimum Air Temperature	deg C
Maximum Air Temperature	deg C
Specific Humidity	mass fraction

In addition to leveraging EPA and gridMET data sources, data from the National Land Coverage Database (NLCD) was also incorporated. Considering the surrounding land categorization

for each water body enables us to identify locations likely to be affected by agricultural runoff. We calculated the percentages of different land types in its vicinity for each water body with a substantial number of temporally consistent nutrient measurements. We defined a 'bounding box' around each water body based on its EPA-documented minimum and maximum latitude-longitude coordinates. This bounding box is then extended by 20 km in all four directions, and each present land type is represented as a percentage. We aim to prioritize water bodies whose surrounding land coverage is primarily cultivated crops and pasture/hay.

Table 3.2: Water Bodies of Interest and Data Availability

Water Body	No. Nitrogen Measurements	No. Phosphorus Measurements	Time Frame	Top Land Coverage
Lake Granby	2274	807	2001 – 2020	40.4% Evergreen Forest
Barr Lake	351	223	2002 – 2005	37.5% Cultivated Crops

For this study, two freshwater bodies were selected in Colorado to utilize in modeling, whose details are displayed in Table 3.2. Although we already favor water bodies with the highest available nutrient measurement consistency, linear interpolation is also performed on these datasets to provide daily data measurements for the model. The time series plots for each water body nutrient series are stored in the appendix, depicting where the original measurements reside and how we interpolated the series. After this interpolation, decades of measurements are available to leverage for analysis. The first water body of focus is Lake Granby, which has data availability from 2001 to 2020 and is considered unaffected by agricultural activities because evergreen forests mainly surround it. The second is Barr Lake, where nearly 40% of the surrounding land is cultivated crops, with data available from 2002 to 2005. These selections aim to build an appropriate testing environment for RQ3, which focuses on scalability amidst data scarcity. This approach allows us to train our model using abundant data from an environment with fewer expected 'unpredictable' events and apply this knowledge in a less stable environment. We can assess the expected differences in stability between these environments using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) statis-

tical metric. This metric provides insight into the level of stationarity (or lack thereof) within a time series. We can then use this information as a proxy metric to classify a non-stationary series as more impacted by agricultural activities, thereby validating our defined approach for RQ1.

3.2 Addressing Non-Stationarity

An expected characteristic of these nutrient time series is non-stationarity. Mitigating non-stationarity is often accomplished with variable decomposition. Because of its proven performance with LSTMs and time series forecasting, our methodology implements Variational Mode Decomposition (VMD) as the decomposition technique. VMD separates the complex time series into Intrinsic Mode Functions (IMFs) modulated in amplitude and frequency. With multiple signals now representing different frequency bands of the series, we can better understand differing trends present in the data. The original signal is obtained simply by summing the IMF series and the residuals. An example of this decomposition, shown with Lake Granby's nitrogen series, is depicted in Figure 3.1. Each of the resulting signal decompositions is stored in the appendix.

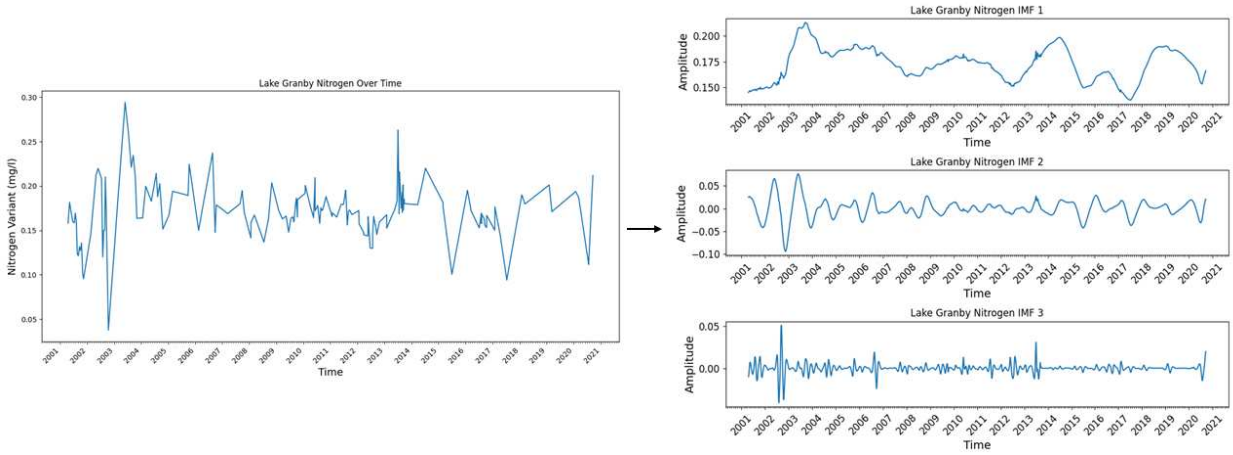

Figure 3.1: Lake Granby nitrogen original series to VMD transformation.

Hyper-parameters implemented for VMD are defined as follows: α = *2000;* τ = $1e-4$*; K = 3; DC = False; init = 1, tol* = $1e-7$. The α parameter is the bandwidth limit criterion, thus controlling the degree of smoothness of the decomposition, and we chose a higher value of

2000 to reduce noise and capture the signal's dominant low-frequency trends. The τ parameter is a noise tolerance parameter, where the smaller chosen value of $1e-4$ provides a more precise decomposition. Importantly, the K parameter defines the number of signal decomposition modes. A value of 3 was chosen based on visual inspection, as a larger mode value than this resulted in multiple similar mode series, signaling that the modes may be decomposing the same underlying relationship and providing unnecessary repetition. The *DC* component represents a signal's zero-frequency or average value. This component was set to false because nutrient time series are expected to exhibit non-stationarity and lack a consistent mean. In continuation, setting the *init* parameter to 1 initializes the frequencies uniformly, which is generally more stable compared to random initialization when *init* is set to 0. Lastly, the *tol* parameter influences the stopping threshold for the VMD optimization, where a lower tolerance, such as $1e-7$, ensures convergence at a stable solution.

The decomposition of the original signal into its frequency components can provide valuable insights into the underlying seasonal trends of the data. Musbah et al. demonstrated how leveraging the plot of a Fast Fourier Transform (FFT) can easily show where each IMF dominant frequency resides and thus its dominant lagged relationship [39]. Research soon to be coupled with this quantitative seasonal knowledge from decomposition lies in a continued discussion of the LTSF-Linear models proposed by Zeng et al. [19]. Its research was conducted to challenge the already-accepted concept that complex transformers are effective for forecasting time series. This paper sheds light on the shortfalls of self-attention, namely that its positional encoding and temporal embeddings cannot mitigate all temporal information loss and thus may not be ideal in forecasting a temporally dependent series. When the transformer's performance is compared to the LTSF-Linear models, the simpler linear models that leverage historical patterns often outperform the complex self-attention-based transformer. The results of this research against transformers in time series can continue to motivate the concept that it is not only a highly complex model that will enable the prediction capabilities required but also the chosen methods for data augmentation that truly help optimize performance, thus developing and improving strides to accomplish RQ2. A

depiction of the FFT results, after first differencing the series twice to isolate low-frequency (or long-term) trends, is shown in Figure 3.2. Each of the resulting FFT plots is also stored in the appendix.

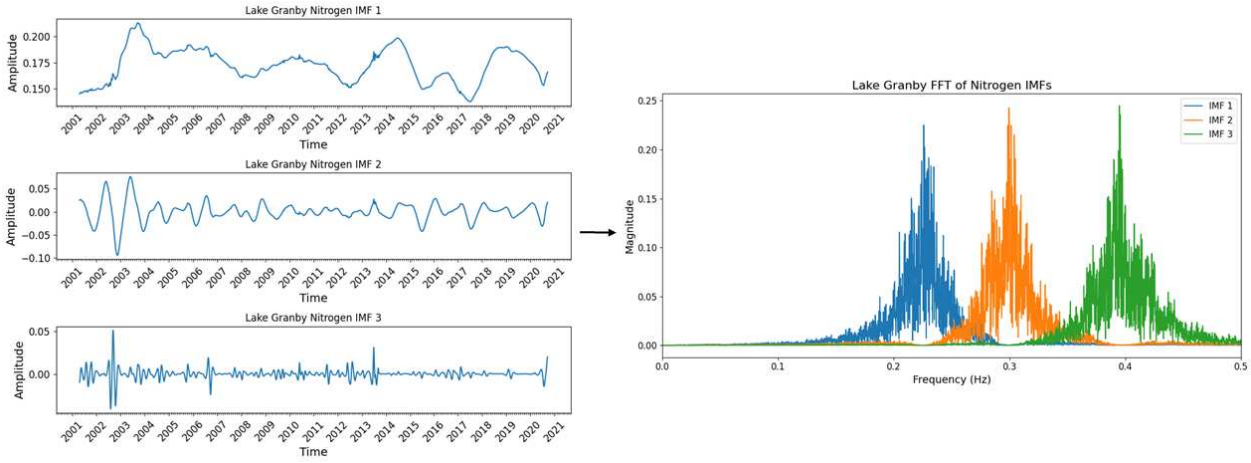

Figure 3.2: Lake Granby nitrogen IMFs to FFT transformation after differencing the series twice.

3.3 Capturing Seasonality

Capturing seasonality is equivalent to capturing patterns or fluctuations that occur regularly and cyclically. Seasonal trends often refer to patterns with longer periodicity. Because LSTMs struggle when the look-back window size increases and exhibit temporal loss, it can be challenging to identify these seasonal trends with the model technique alone. Thus, capturing seasonality is achieved by engineering an additional input feature based on historical knowledge of the environment. This concept is incorporated similarly to the DLinear model of the LTSF-Linear family, whose models leverage historical measurements rather than short-term preceding knowledge and have successfully captured seasonality well. Aswanuwath et al. previously implemented a VMD and FFT-motivated machine learning model with an additional LTSF-Linear layer. They successfully demonstrated the highest performance accuracy with the coupled VMD and DLinear efforts in the context of electricity peak load forecasting [40]. Thus, with the quantitative dominant frequency knowledge of our decomposed signal and the research supporting the effectiveness of

incorporating historical moving averages, we can capture the seasonal trends that the LSTM lacks to capture independently.

To create and leverage historical knowledge of seasonality, we must divide the available data into a 'historical' and 'current' subset. Deciding the best data split is open to interpretation, but this methodology implements a half-split-by-date due to improved model performance with increased historical subset size. Additionally, this split avoids data leakage and ensures the model does not train on the data characteristics it aims to predict. For each dataset, moving averages are calculated based on the frequency characteristics of the nutrient time series decomposition, where the identified dominant lagged relationship influences the window size. This series of nutrient moving averages is joined with its corresponding input features to represent a spatiotemporal profile of the environment, as depicted in Figure 3.3. These pattern profiles can conceptually represent various possible spatiotemporal patterns of the ecosystem. Each signal decomposition is represented by different spatiotemporal pattern sets, with the moving average being the only axis that differs between them. It is important to note the differing definitions in the window span for the current or historical subsets. In the historical dataset, the date is indexed as the middle of the window to capture the seasonal patterns around its environment. However, we must ensure that the model does not utilize leaking prediction data in the input features for the current portion of the dataset. Thus, the current dataset's moving average window is set purely with its preceding values.

With these spatiotemporal pattern profiles, we can use a similarity metric to map the current series' expected seasonality characteristics to our historical knowledge of trends. We implement this similarity mapping with a K-Nearest Neighbors (KNN) model that trains on the historical spatiotemporal profiles and is the domain space used to map the most similar profile to the 'current' environment. The result of this similarity mapping is the nearest historical neighbor's corresponding moving average (seasonal component). The residuals of the signal decomposition are also included as an input feature. This coupled representation of spatiotemporally mapped moving average and VMD residual is the final engineered input feature to capture seasonality. Incorporating the profile's corresponding residual can serve as an error-adjusting mechanism to capture the

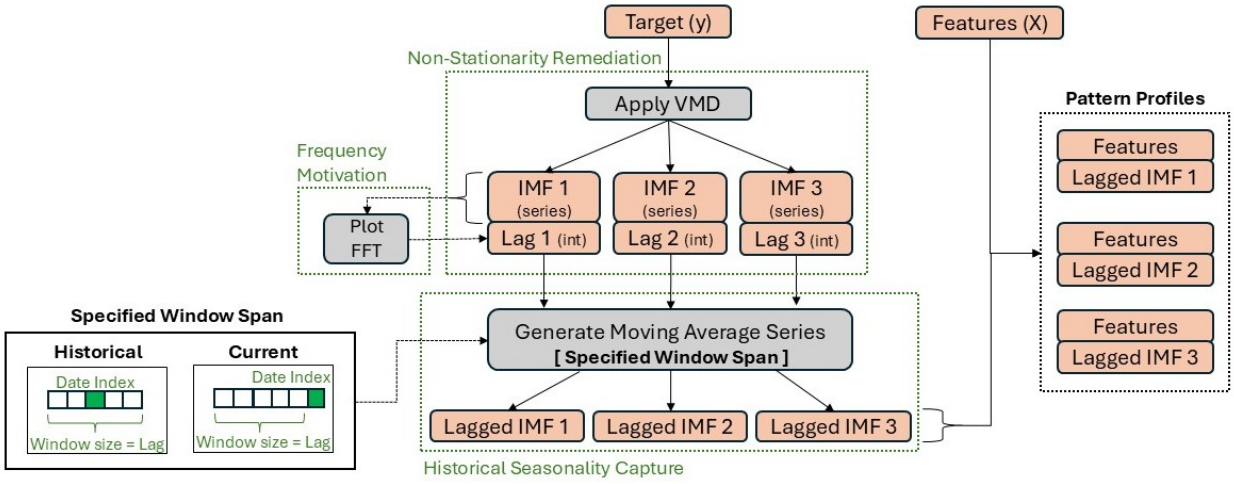

Figure 3.3: Developing Spatiotemporal Pattern Profiles for KNN

missing components of the environment indirectly. While the additional input features include the residuals of the VMD signals, they are not considered in the spatiotemporal profiles, as that would require knowledge of the current prediction's status and thus result in data leakage.

3.4 Model Design

The base of this model implements a bidirectional LSTM with the target series defined as nitrogen or phosphorus. The extended design of the model will leverage the defined tactic of VMD, where each IMF is predicted independently of the other, and the final prediction is the sum of these IMF model outputs. We previously incorporated seasonality by leveraging the frequency-motivated historical moving averages and a KNN-assisted similarity metric mapping of historical spatiotemporal patterns. The overall design of this model is depicted in Figure 3.4, which abstracts the historical and current pattern profiles as defined in Figure 3.3.

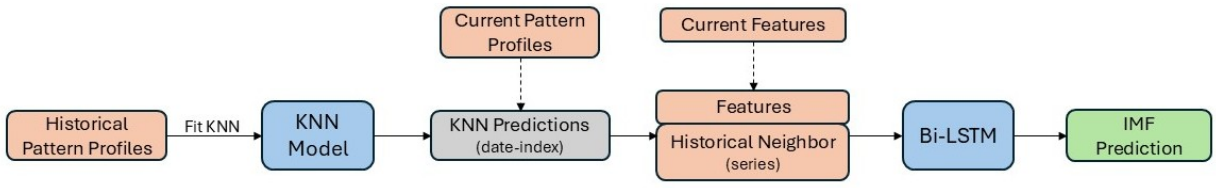

Figure 3.4: Diagram of Input Feature Development and Model Training

Within the Bi-LSTM models, we uniquely incorporate a cross-attention mechanism amidst previous research into its effectiveness. Shih et al. shed light on the shortfalls of self-attention in multivariate time series forecasting because it fails to detect functional temporal patterns due to the averaging of multivariate information across multiple time steps [41]. This previous research alternatively proposed an attention mechanism where the context vector is instead the weighted sum of the row vectors across time steps, thus better capturing temporal information. Therefore, the proposed model of this research implements the specialized incorporation of cross-attention.

Table 3.3: Model Hyper-Paramaters

Hyperparameter	Value
Sequence Length	14
Loss Function	Huber Loss
Optimizer	Adam
Learning Rate	0.001
Dropout	0.2
Batch Size	20
Hidden Dimension	64
Output Dimension	1
Epochs	100

Finally, Table 3.3 defines the base hyper-parameters used for model testing. These decisions are motivated partly by testing and comparing performance in previous work. When tested in tandem with other values, such as 7 and 30, the sequence length 14 performed the best and seems a capable middle ground for the LSTM to minimize temporal loss while capturing the relevant historical impacts. The Huber loss function is optimal for datasets with possible outliers present, as it is more robust regarding their effects. When predicting time series with expected fluctuations, robustness to outliers is favorable. The learning rate is initially set to 0.001 to allow a proper level of learning but also incorporates a dynamically increased learning rate that decreases by a factor of 0.5 for every 15 epochs in the absence of validation loss improvement. The primary purpose of the dynamically decreasing learning rate, along with a dropout value, is to avoid overfitting the current

dataset and support generalization. Previous works have additionally shown that a large batch size did not improve performance, and thus, it is set comparatively low at 20. There have also been consistent trends through previous work and model testing on our current series that too high or too low of a hidden dimension degrades performance. The best-performing hidden dimension was 64, thus used to define our baseline hyper-parameters. The model's output dimension is set to 1 and thus predicts the following time step. Lastly, 100 epochs were sufficient for training the base and the transferred models.

3.5 Transfer Learning

We use the same tactics of historical seasonality to assist in the transfer learning implementation and improve prediction in a sparse data environment. The spatiotemporal profile mapping also uses the pre-trained KNN models, incorporating another channel of knowledge transfer opportunity. Implementing a transfer learning concept alongside the KNN model not only provides us with extensive comparative capabilities to decades of spatiotemporal profiles but also enables us to utilize the smaller dataset in its entirety, as there is no longer a need to leverage its historical data for seasonality profile matching or be concerned about data leakage.

Transferring the learning of the LSTM-based model is implemented by first training the source model on the large and non-agriculturally affected dataset (Lake Granby). This trained model is then utilized for the smaller, agriculturally affected dataset, warm-starting the model's understanding of its prediction environment. The input features of the smaller dataset leverage Lake Granby's spatiotemporal profiles, influencing only the features that represent the historical moving average and its VMD residual. We initially freeze each model for 15 epochs at the start of this transfer learning. Thus, the model has not yet begun adjusting its understanding of the new dataset's characteristics. Freezing layers of the model at the beginning of training is a strategy to allow the data first to leverage the knowledge of the pre-trained model while also having the opportunity to gradually adapt to its specific domain.

Chapter 4

Performance Benchmarks

4.1 Stationarity Benchmarks

The benchmark used to test the stationarity of the two focused water bodies is the KPSS Statistic, which tests against the null hypothesis that a series is stationary. A high KPSS statistic with a low p-value indicates that we can reject the null hypothesis and classify the series as non-stationary. These results are shown in Table 4.2. The KPSS statistic can theoretically range from 0 to infinity, where 0 indicates perfect stationarity. The threshold KPSS values are influenced by the chosen significance level, whose threshold is typically set to 10%, 7.5%, 5%, and 2.5%. In this benchmark, we decided on a significance value of 5%, whose corresponding threshold values are displayed in the table.

Table 4.1: KPSS statistics to quantitatively represent the level of stationarity or non-stationarity of target series nitrogen (N) and phosphorus (P)

Water Body	Type	Threshold	KPSS stat		KPSS p-value	
	N	P	N	P	N	P
Lake Granby	Level	0.463	0.298	3.268	0.10	0.01
Barr Lake	Level	0.463	0.774	1.248	0.01	0.01
Lake Granby	Trend	0.146	0.129	0.505	0.10	0.01
Barr Lake	Trend	0.146	0.130	1.248	0.01	0.01

We can notice that the nitrogen series for Lake Granby has a low KPSS statistic and a high p-value, indicating that this series is stationary. However, the phosphorus series for Lake Granby exhibits non-stationarity, which is still possible in any nutrient time series. In continuation, we can see that both nitrogen and phosphorus series for Barr Lake have a high KPSS statistic and a low p-value. With these metrics, we can conclude that both nutrient series for Barr Lake are non-stationary. The results of the KPSS metric quantitatively define the expected lack of stationarity in

our agriculturally impacted water body (Barr Lake). In contrast, stationarity is not as strong in the non-agriculturally impacted water body (Lake Granby). Thus, this benchmark validates the goal of RQ1, effectively identifying water bodies affected by agricultural runoff.

4.2 Incremental Model Performance

This section provides performance benchmarks for incremental improvements incorporated into the model. Table 4.2 shows the metrics for several incremental model designs. While reviewing these metrics, it is important to note that the relative scales of nitrogen and phosphorus measurements are 0.1 - 3.0 mg/l and 0.005 - 0.05 mg/l, respectively. Because the nutrient scales are relatively small, we can classify any change in the final metrics as meaningful. The natural incremental additions of bi-directionality and attention to the LSTM do not see meaningful improvements until the feature-based attention is incorporated.

Furthermore, the implementation and separation of IMF training performed better overall when compared with the models predicting the original signal. The final proposed model that maps historical moving averages with KNN-motivated spatiotemporal profile similarity performed with the highest accuracy in almost all metrics. The proposed model performs with a more distinct accuracy improvement on the nitrogen series rather than the phosphorus series.

Table 4.2: Lake Granby: Incremental model performance metrics for nitrogen (N) and phosphorus (P)

Model	MAE (mg/L)		MAPE (%)		RMSE (mg/L)	
	N	P	N	P	N	P
LSTM	0.0167	0.0018	9.89	17.22	0.0201	0.0023
BiLSTM	0.0159	0.0023	9.52	20.42	0.0196	0.0026
Att-BiLSTM	0.0154	**0.0014**	9.25	**12.40**	0.0204	**0.0019**
IMF Att-BiLSTM	0.0146	**0.0014**	8.84	12.52	0.0195	**0.0019**
Historical IMF Att-BiLSTM	0.0169	0.0017	10.46	15.99	0.0198	0.0022
Historical KNN-IMF Att-BiLSTM	**0.0115**	**0.0014**	**7.41**	13.64	**0.0166**	**0.0019**

MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ MAPE $= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \times 100$ RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

4.3 Proposed Model Performance

Performance metrics, such as MAE, can help determine how well the model predicts the series. However, viewing the original time series against the predicted ones in a visual plot is much more valuable. Determining if the model captures the series' fluctuations shown in Figure 4.1 is much easier. For example, the predicted series plotted against the original series for the incremental Att-BiLSTM model can capture and display oscillatory patterns. Still, its predictions are too temporally symmetric around the mean and do not capture the full extent of fluctuations. We can notice that, with the proposed model, the nitrogen series performs exceptionally well, as this model can capture the plunge in the series that previous iterated models could not. Performance for phosphorus, while predicting around the correct mean of the series, noticeably struggles to recognize the plunge in the series. In continuation, Figure 4.2 displays the prediction performance of Barr Lake, a location deemed to be impacted by agricultural runoff with substantially less available data for training. Because of this lack of data, the model trained with Barr Lake data struggles to capture the present trends accurately.

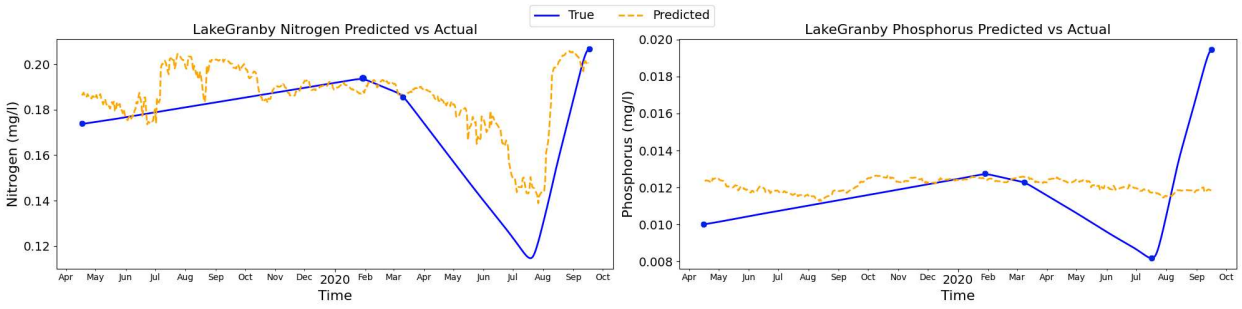

Figure 4.1: Lake Granby Predicted vs Actual Time Series (April 2019 - October 2020). The model used in this prediction is the Historical KNN-IMF Att-BiLSTM.

These results trend toward improved nitrogen performance compared to phosphorus. A possible reason for this is the fundamental difference in the transport mechanisms of these nutrients. Nitrogen is particularly mobile due to its solubility and low soil absorption and thus is more affected by weather characteristics like rainfall or wind. Phosphorus, conversely, is more commonly transported through events such as erosion or sediment displacement. The data sources utilized in

this study to develop input features better capture the influences of nitrogen transport rather than phosphorus transport.

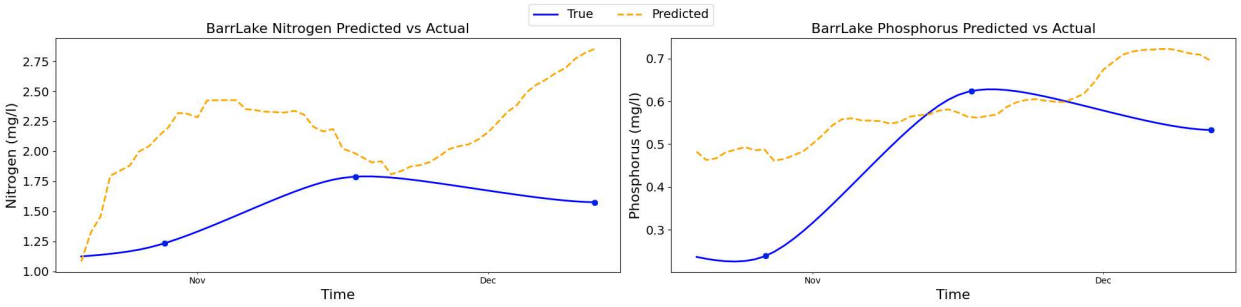

Figure 4.2: Barr Lake Predicted vs Actual Time Series(October 2005 - December 2005). The model used in this prediction is the Historical KNN-IMF Att-BiLSTM.

4.4 Transfer Learning Performance

Transfer learning was applied to the proposed model to improve prediction performance for Barr Lake, a location with limited data availability. The prediction performance results amidst this transfer learning implementation are depicted in Figure 4.3. When compared against Figure 4.2, where the model does not leverage transfer learning, we can see the improved trend capturing of the model. Leveraging transfer learning from the KNN mappings improved the model's prediction performance to a notable degree. This improvement can be attributed to the plethora of feature profiles generated from Lake Granby's extensive historical data, allowing more opportunities to map similar profiles accurately. Additionally, because we are not directly leveraging historical data from Barr Lake, we can include its full dataset in the model training, as we no longer have to worry about data leakage.

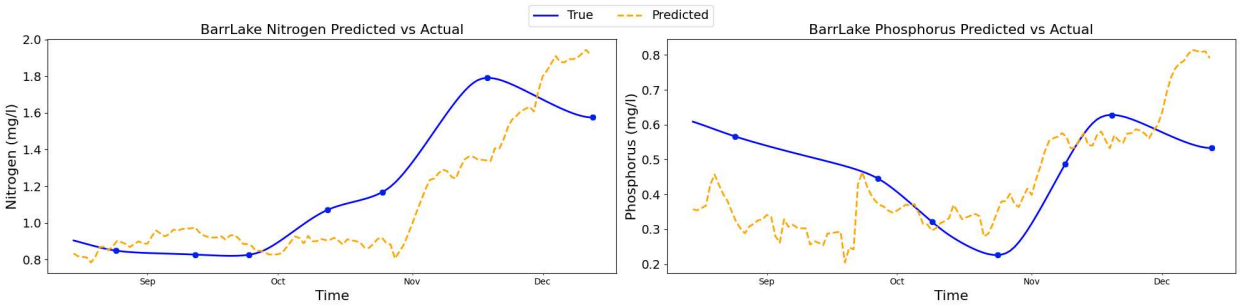

Figure 4.3: Transfer Learning: Predicted vs Actual Time Series (August 2005 - December 2005). The model used in this prediction is the Historical KNN-IMF Att-BiLSTM. The base model is trained with the Lake Granby nutrient series. This pre-learned knowledge is then leveraged in prediction for the Barr Lake nutrient series.

Chapter 5

Conclusion

In conclusion, this research proposes a hybrid approach of algorithmic and modeling techniques to mitigate common challenges in water quality data and improve prediction performance in sparse, non-stationary datasets. We mitigate the non-stationarity of nutrient time series through variable decomposition, namely Variational Mode Decomposition (VMD). The relationships between VMD and its dominant frequencies (i.e., dominant lags) incorporate an understanding of seasonality that machine learning models have struggled to provide due to their limited window size. We enhance the understanding of seasonality with similarity metrics by developing historical spatiotemporal profiles motivated by environmental features and mapping their similarity through a KNN model. Finally, we leverage transfer learning to warm-start a model's understanding of its environment and improve prediction performance in sparse datasets. Overall performance showed a higher prediction accuracy for nitrogen series over phosphorus series. We hypothesize that the elevated accuracy of the nitrogen series is due to the nutrients' differing transport mechanisms and that the chosen incorporated input features more closely represent nitrogen transport mechanisms, like daily weather. It is possible that incorporating more phosphorus-related measurements as input features, such as turbidity or soil erosion, will improve its prediction performance.

5.1 Limitations

One limitation of the current research is the lack of temporal consistency in the target series. While interpolation techniques can remedy the requirement of consistent measurements for LSTM training, there is still a loss of relational information when the recorded data is inconsistent. Additionally, because the datasets utilized are not temporally consistent and thus linearly interpolated, it is difficult to determine whether the predicted fluctuations not shown in the original series occur or result from a lack of accuracy. Another limitation of this research was the comparatively minimal number of input features incorporated. EPA measurement data seldom had consistent measure-

ments of other water quality parameters and the presence of features in both modeled locations, as transfer learning models must appropriately fit the initially trained input features. Thus, more weather measurements were incorporated from gridMET, while many other water quality measurements were not temporally attainable. Additionally, the gridMET weather parameters were compressed into a single daily value by averaging gridMET's 4 km tile resolution, which can introduce a level of error in its relationship to the target series.

5.2 Future Work

Future research on nutrient prediction should incorporate a wider variety of environmental measures alongside the proposed model architecture. Including additional environmental characteristics will also enable the application of feature selection techniques, enhancing the model's understanding of ecological relationships. Furthermore, using frequency-motivated historical moving averages could benefit other modeling techniques beyond LSTM variations. Since this methodology captures historical spatiotemporal profiles, it is plausible that the model reflects more true fluctuations than were measured in reality. This plausibility suggests an opportunity for successful implementation in the context of data imputation.

Bibliography

[1] Sravani Pericherla, Manoj Kumar Karnena, and Saritha Vara. A review on impacts of agricultural runoff on freshwater resources. *International Journal on Emerging Technologies*, 11:829–833, 04 2020.

[2] Keerthana Suresh, Ting Tang, Michelle T H van Vliet, Marc F P Bierkens, Maryna Strokal, Florian Sorger-Domenigg, and Yoshihide Wada. Recent advancement in water quality indicators for eutrophication in global freshwater lakes. *Environmental Research Letters*, 18(6):063004, 06 2023.

[3] Ali O. Alnahit, Ashok K. Mishra, and Abdul A. Khan. Stream water quality prediction using boosted regression tree and random forest models. *Stochastic Environmental Research and Risk Assessment*, 36(9):2661–2680, 2022.

[4] Hui Wang, Jingxuan Sun, Jianbo Sun, and Jilong Wang. Using random forests to select optimal input variables for short-term wind speed forecasting models. *Energies*, 10(10), 2017.

[5] Ali Najah Ahmed, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, and Ahmed Elshafie. Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578:124084, 2019.

[6] Steefan Contractor and Moninya Roughan. Efficacy of feedforward and lstm neural networks at predicting and gap filling coastal ocean timeseries: Oxygen, nutrients, and temperature. *Frontiers in Marine Science*, 8, 2021.

[7] Zhongyao Liang, Rui Zou, Xing Chen, Tingyu Ren, Han Su, and Yong Liu. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. *Journal of Hydrology*, 581:124432, 2020.

[8] Zhuhua Hu, Yiran Zhang, Yaochi Zhao, Mingshan Xie, Jiezhuo Zhong, Zhigang Tu, and Juntao Liu. A water quality prediction method based on the deep lstm network considering correlation in smart mariculture. *Sensors*, 19(6), 2019.

[9] Qiang Zhang, Ruiqi Wang, Ying Qi, and Fei Wen. A watershed water quality prediction model based on attention mechanism and bi-lstm. *Environmental Science and Pollution Research*, 29(50):75664–75680, 2022.

[10] Qinghong Zou, Qingyu Xiong, Qiude Li, Hualing Yi, Yang Yu, and Chao Wu. A water quality prediction method based on the multi-time scale bidirectional long short-term memory network. *Environmental Science and Pollution Research*, 27(14):16853–16864, 2020.

[11] Yurong Yang, Qingyu Xiong, Chao Wu, Qinghong Zou, Yang Yu, Hualing Yi, and Min Gao. A study on water quality prediction by a hybrid cnn-lstm model with attention mechanism. *Environmental Science and Pollution Research*, 28(39):55129–55139, 2021.

[12] Yumeng Wang, Ke Liu, Yuejun He, Pengfei Wang, Yuxin Chen, Hang Xue, Caiyi Huang, and Lin Li. Enhancing air quality forecasting: A novel spatio-temporal model integrating graph convolution and multi-head attention mechanism. *Atmosphere*, 15(4), 2024.

[13] Xinghan Xu and Minoru Yoneda. Multitask air-quality prediction based on lstm-autoencoder model. *IEEE Transactions on Cybernetics*, 51(5):2577–2586, 2021.

[14] Jing Bi, Zexian Chen, Haitao Yuan, and Jia Zhang. Accurate water quality prediction with attention-based bidirectional lstm and encoder–decoder. *Expert Systems with Applications*, 238:121807, 2024.

[15] Shengyue Chen, Jinliang Huang, Peng Wang, Xi Tang, and Zhenyu Zhang. A coupled model to improve river water quality prediction towards addressing non-stationarity and data limitation. *Water Research*, 248:120895, 2024.

[16] Yituo Zhang, Chaolin Li, Yiqi Jiang, Lu Sun, Ruobin Zhao, Kefen Yan, and Wenhui Wang. Accurate prediction of water quality in urban drainage network with integrated emd-lstm model. *Journal of Cleaner Production*, 354:131724, 2022.

[17] Hui Liu and Chao Chen. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy*, 249:392–408, 2019.

[18] Hafizh Raihan Kurnia Putra, Novanto Yudistira, and Tirana Noor Fatyanosa. Variational mode decomposition and linear embeddings are what you need for time-series forecasting, 2024.

[19] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, Jun. 2023.

[20] Seyedeh Narjes Fallah, Mehdi Ganjkhani, Shahaboddin Shamshirband, and Kwok-wing Chau. Computational intelligence on short-term load forecasting: A methodological overview. *Energies*, 12(3), 2019.

[21] Sangmi Lee Pallickara, Shrideep Pallickara, Milija Zupanski, and Stephen Sullivan. Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pages 573–580. IEEE, 2010.

[22] Sangmi Lee Pallickara, Marlon Pierce, Qunfeng Dong, and ChinHua Kong. Enabling large scale scientific computations for expressed sequence tag sequencing over grid and cloud computing clusters. In *PPAM 2009 Eight International Conference on Parallel Processing and Applied Mathematics Wroclaw, Poland*, 2009.

[23] Sangmi Lee Pallickara, Shrideep Pallickara, and Marlon Pierce. Scientific data management in the cloud: A survey of technologies, approaches and challenges. *Handbook of Cloud Computing*, pages 517–533, 2010.

[24] Geoffrey Fox, Shrideep Pallickara, Marlon Pierce, and Harshawardhan Gadgil. Building messaging substrates for web and grid applications. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1833):1757–1773, 2005.

[25] GC Fox, Hasan Bulut, Kangseok Kim, Sung-Hoon Ko, Sangmi Lee, Sangyoon Oh, Shrideep Pallickara, Xiaohong Qiu, Ahmet Uyar, Minjun Wang, et al. Collaborative web services and peer-to-peer grids. *SIMULATION SERIES*, 35(1):3–12, 2003.

[26] Yogesh L Simmhan, Sangmi Lee Pallickara, Nithya N Vijayakumar, and Beth Plale. Data management in dynamic environment-driven computational science. In *Grid-Based Problem Solving Environments: IFIP TC2/WG 2.5 Working Conference on Grid-Based Problem Solving Environments: Implications for Development and Deployment of Numerical Software July 17–21, 2006, Prescott, Arizona, USA*, pages 317–333. Springer, 2007.

[27] Sangmi Lee, Sung Hoon Ko, and Geoffrey C Fox. Adapting content for mobile devices in heterogeneous collaboration environments. In *International Conference on Wireless Networks*, pages 211–217, 2003.

[28] Geoffrey C Fox, Sung Hong Ko, Kang-Seok Kim, Sangyoon Oh, and Sangmi Lee. Integration of hand-held devices into collaborative environments. In *International Conference on Internet Computing*, pages 231–250, 2002.

[29] Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara. Galileo: A framework for distributed storage of high-throughput data streams. In *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, pages 17–24. IEEE, 2011.

[30] Matthew Malensek, Walid Budgaga, Ryan Stern, Shrideep Pallickara, and Sangmi Lee Pallickara. Trident: Distributed storage, analysis, and exploration of multidimensional phenomena. *IEEE Transactions on Big Data*, 5(2):252–265, 2018.

[31] Daniel Rammer, Sangmi Lee Pallickara, and Shrideep Pallickara. Atlas: A distributed file system for spatiotemporal data. In *Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing*, pages 11–20, 2019.

[32] Daniel Rammer, Thilina Buddhika, Matthew Malensek, Shrideep Pallickara, and Sangmi Lee Pallickara. Enabling fast exploratory analyses over voluminous spatiotemporal data using analytical engines. *IEEE Transactions on Big Data*, 8(1):213–228, 2019.

[33] Matthew Malensek, Sangmi Pallickara, and Shrideep Pallickara. Fast, ad hoc query evaluations over multidimensional geospatial datasets. *IEEE Transactions on Cloud Computing*, 5(1):28–42, 2015.

[34] Matthew Malensek, Sangmi Pallickara, and Shrideep Pallickara. Evaluating geospatial geometry and proximity queries using distributed hash tables. *Computing in Science & Engineering*, 16(4):53–61, 2014.

[35] Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara. Expressive query support for multidimensional data in distributed hash tables. In *2012 IEEE Fifth International Conference on Utility and Cloud Computing*, pages 31–38. IEEE, 2012.

[36] Jun Ma, Jack C.P. Cheng, Changqing Lin, Yi Tan, and Jingcheng Zhang. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment*, 214:116885, 2019.

[37] Elissaios Sarmas, Nikos Dimitropoulos, Vangelis Marinakis, Zoi Mylona, and Haris Doukas. Transfer learning strategies for solar power forecasting under data scarcity. *Scientific Reports*, 12(1):14643, 2022.

[38] Zeng Chen, Huan Xu, Peng Jiang, Shanen Yu, Guang Lin, Igor Bychkov, Alexey Hmelnov, Gennady Ruzhnikov, Ning Zhu, and Zhen Liu. A transfer learning-based lstm strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system. *Journal of Hydrology*, 602:126573, 2021.

[39] Hmeda Musbah and Mo El-Hawary. Sarima model forecasting of short-term electrical load data augmented by fast fourier transform seasonality detection. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4, 2019.

[40] Lalitpat Aswanuwath, Warut Pannakkong, Jirachai Buddhakulsomsiri, Jessada Karnjana, and Van-Nam Huynh. A hybrid model of vmd-emd-fft, similar days selection method, stepwise regression, and artificial neural network for daily electricity peak load forecasting. *Energies*, 16(4), 2023.

[41] Shun-Yao Shih, Fan-Keng Sun, and Hung-yi Lee. Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, 108(8):1421–1441, 2019.

Appendix A

Nutrient Time Series Plots

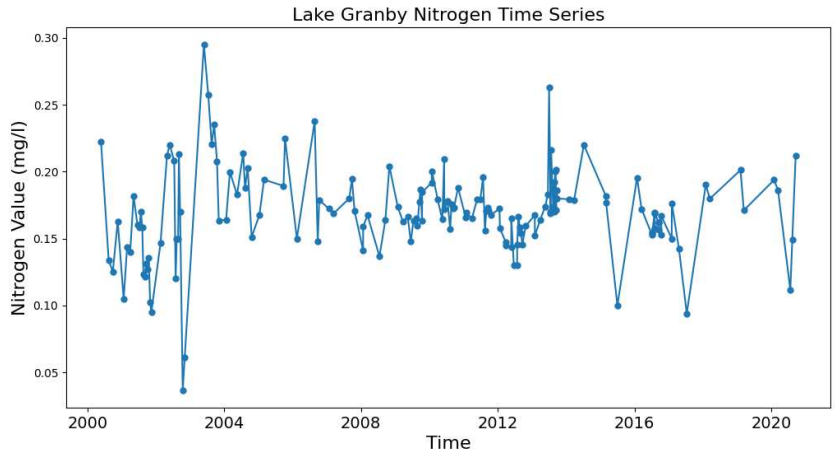

Figure A.1: Lake Granby Interpolated Nitrogen Time Series. The marked dots represent where measurements are available.

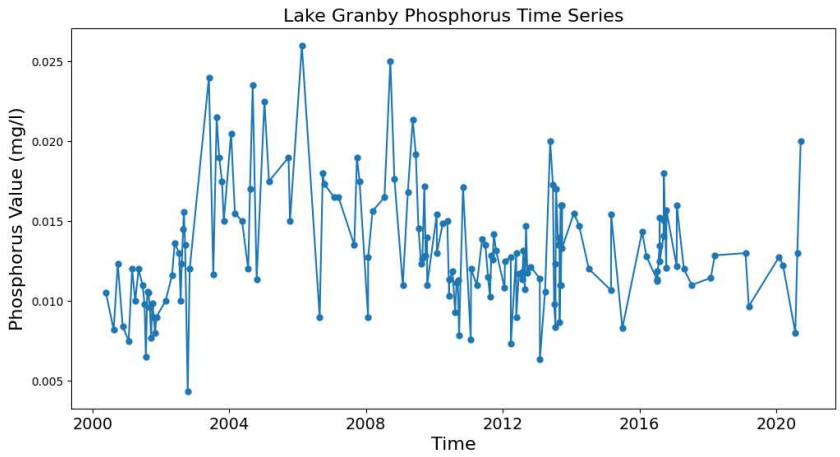

Figure A.2: Lake Granby Interpolated Phosphorus Time Series. The marked dots represent where measurements are available.

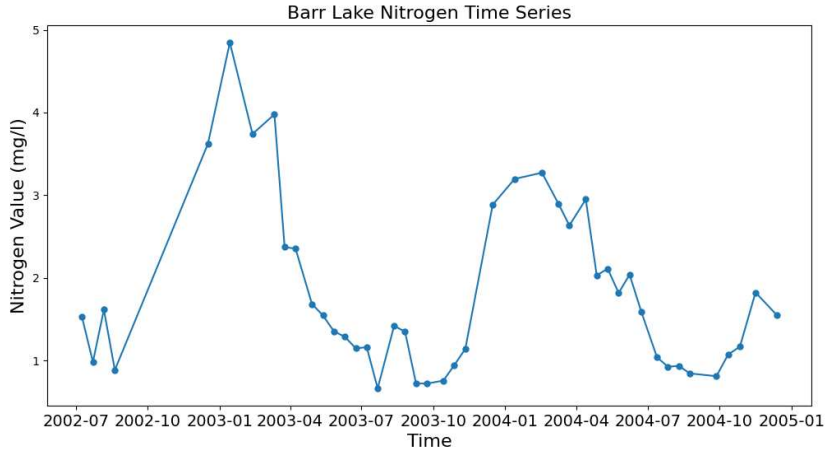

Figure A.3: Barr Lake Interpolated Nitrogen Time Series. The marked dots represent where measurements are available.

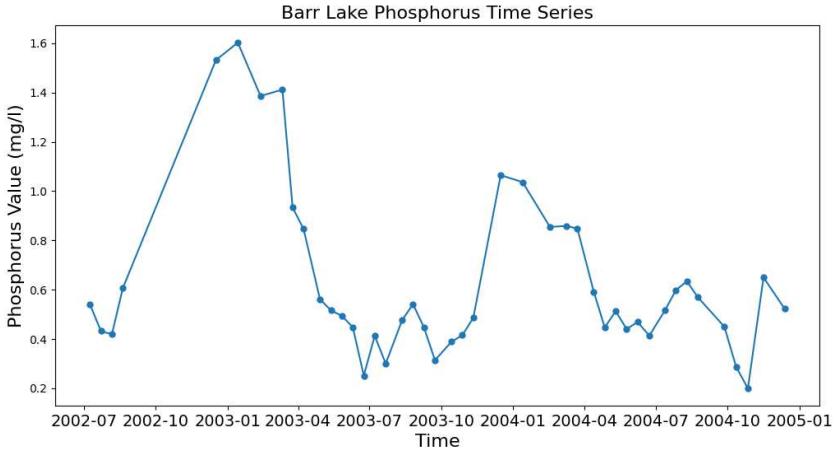

Figure A.4: Barr Lake Interpolated Phosphorus Time Series. The marked dots represent where measurements are available.

Appendix B

Nutrient Time Series Variational Mode Decomposition Plots

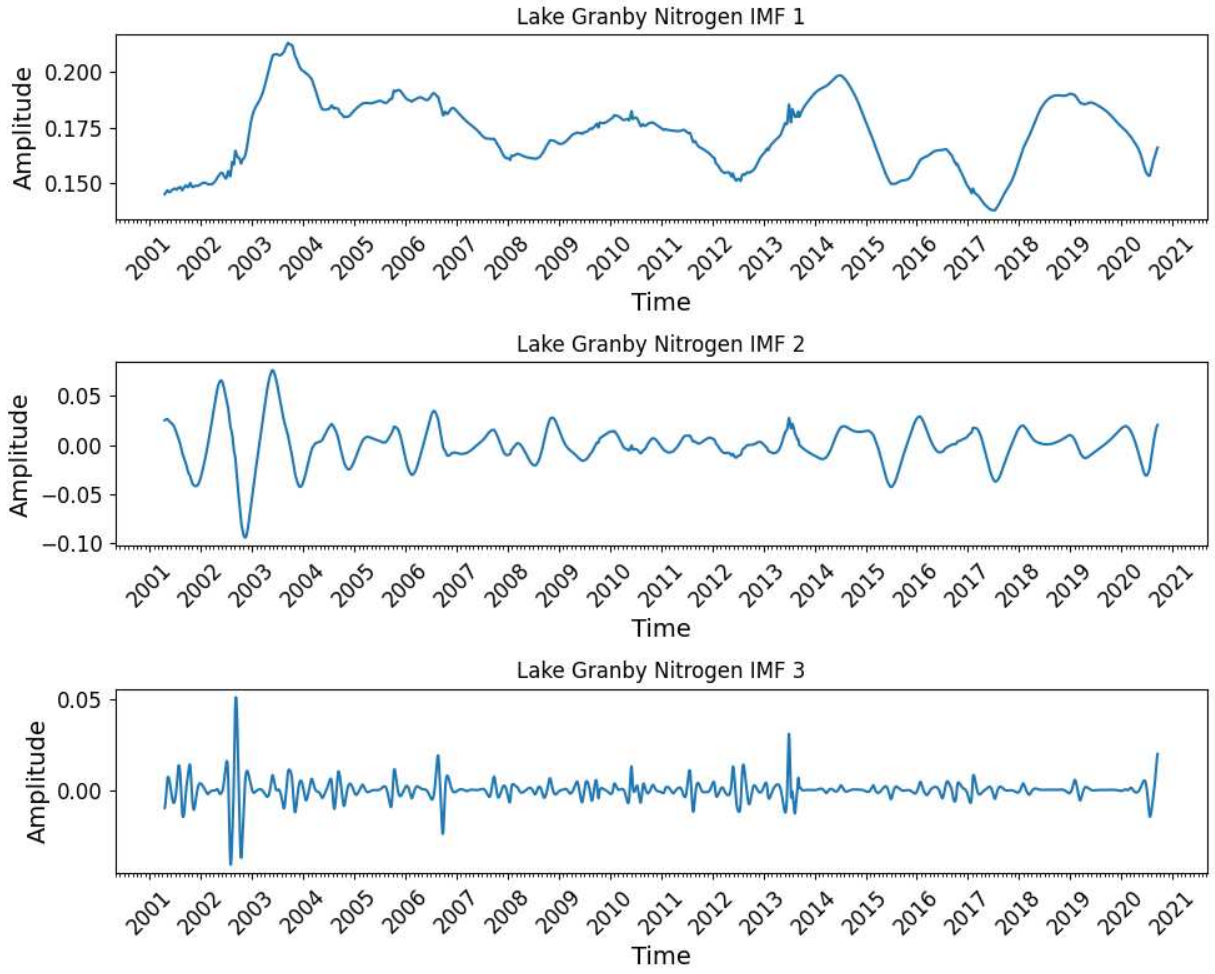

Figure B.1: Lake Granby Nitrogen Series Intrinsic Mode Functions

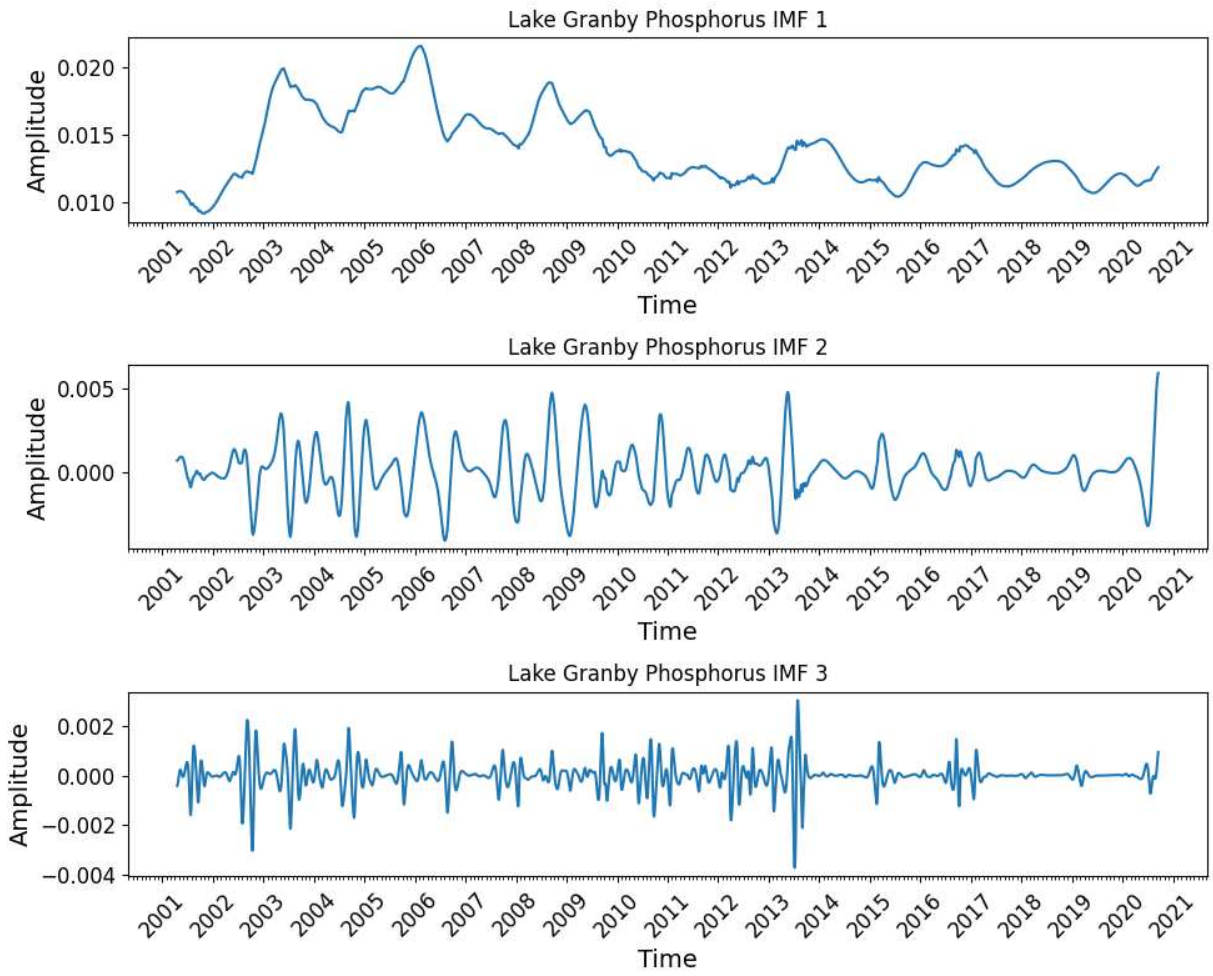

Figure B.2: Lake Granby Phosphorus Series Intrinsic Mode Functions

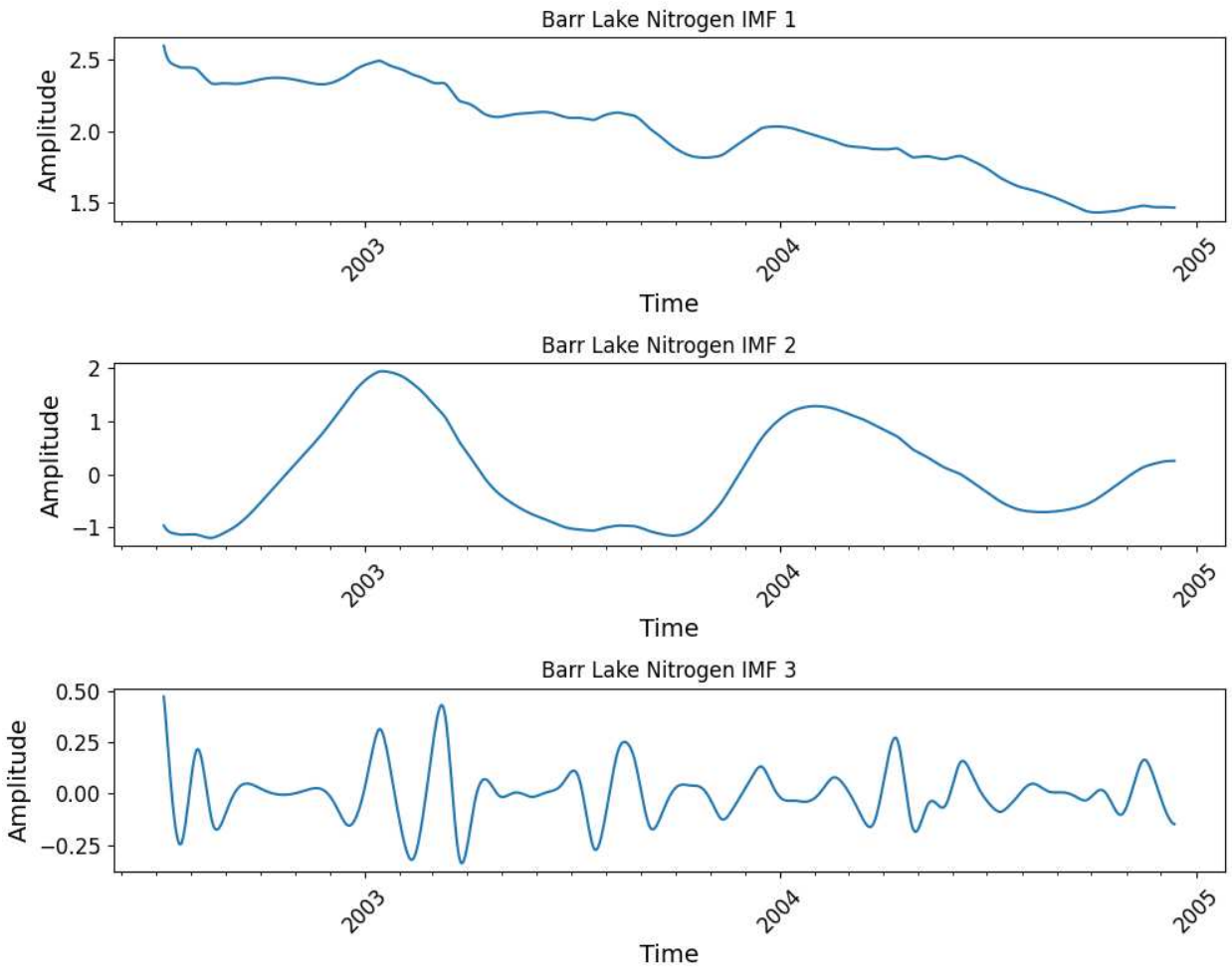

Figure B.3: Barr Lake Nitrogen Series Intrinsic Mode Functions

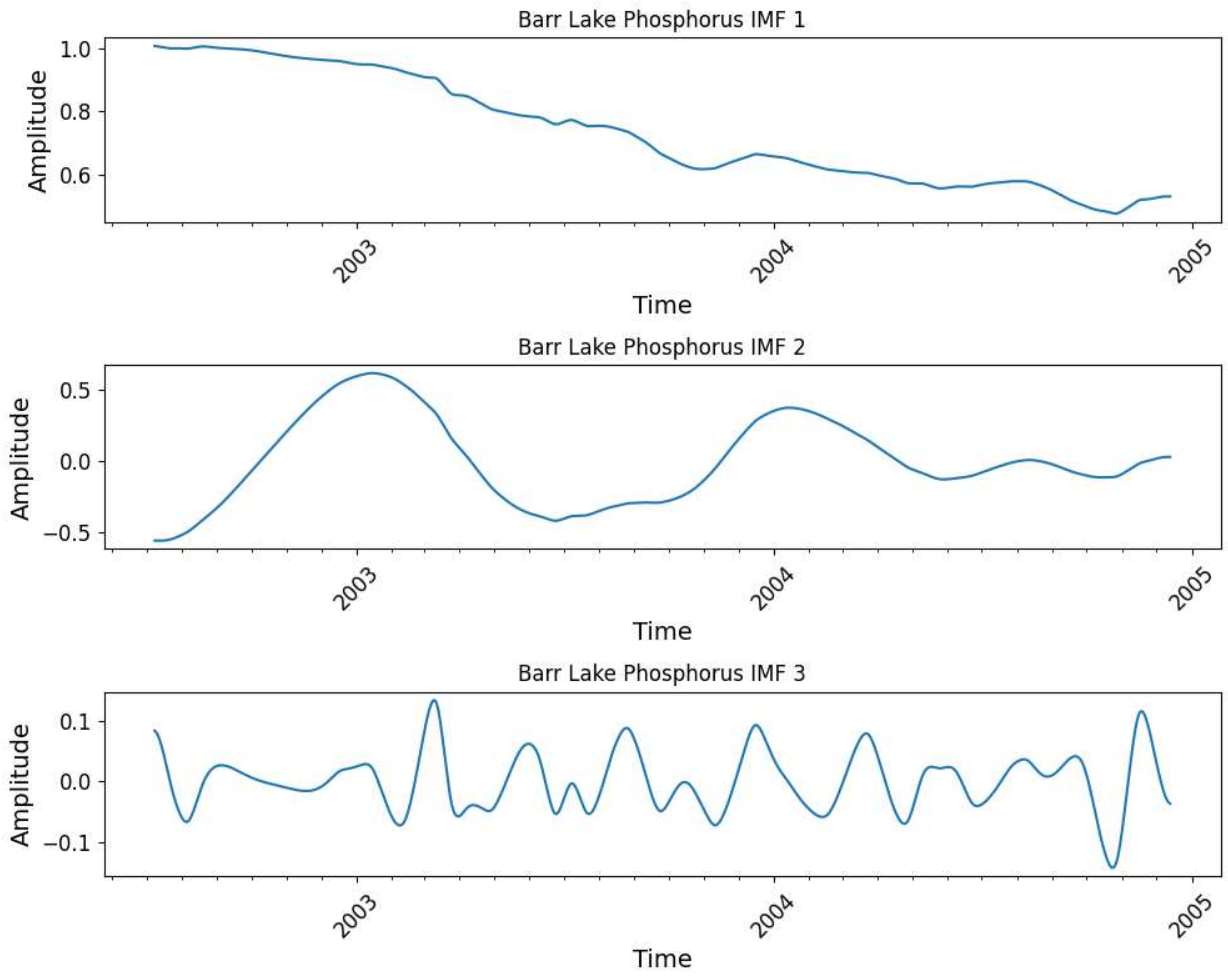

Figure B.4: Barr Lake Phosphorus Series Intrinsic Mode Functions

Appendix C

Nutrient Time Series Fast Fourier Transform Plots

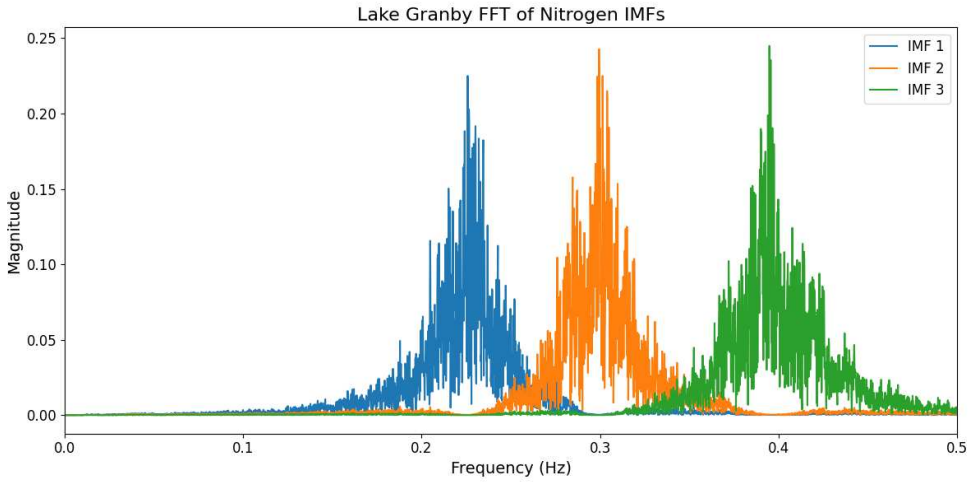

Figure C.1: Lake Granby Nitrogen Series IMF Fast Fourier Transform after differencing the original series twice.

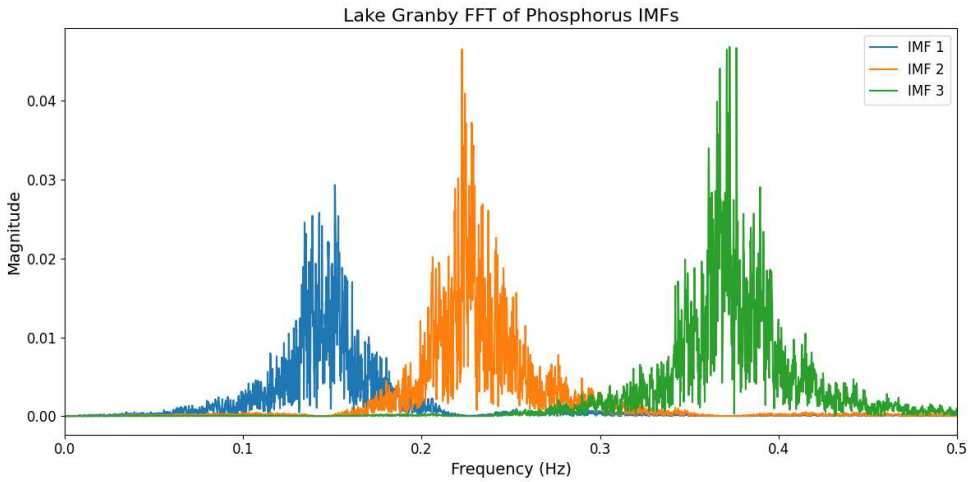

Figure C.2: Lake Granby Phosphorus Series IMF Fast Fourier Transform after differencing the original series twice.

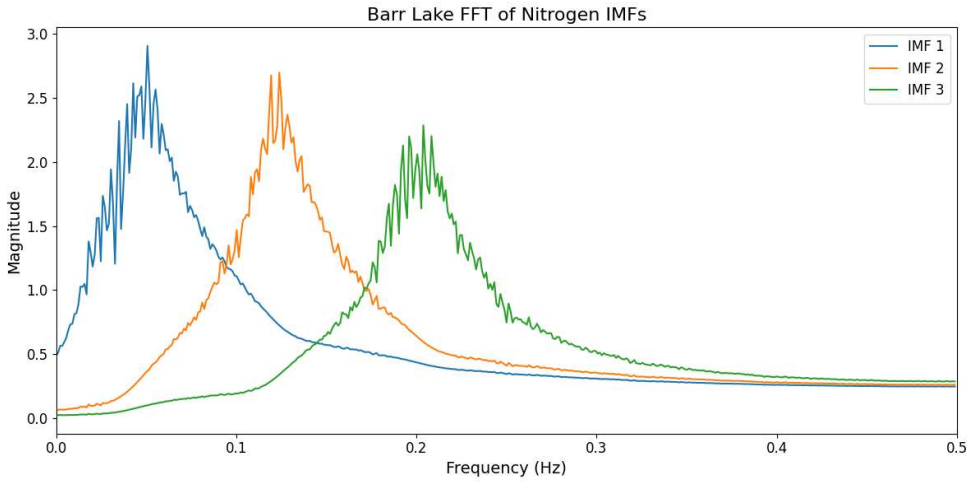

Figure C.3: Barr Lake Nitrogen Series IMF Fast Fourier Transform after differencing the original series twice.

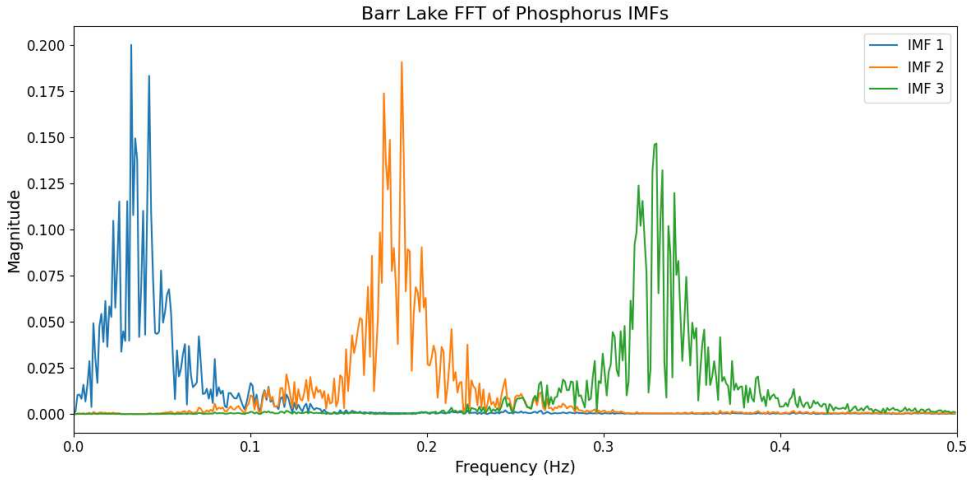

Figure C.4: Barr Lake Phosphorus Series IMF Fast Fourier Transform after differencing the original series twice.