Statistical Analysis for Decomposed Multivariate Time Series Data with an application to Water Discharge Forecasting

KHAWLA K. MAHMOOD

A thesis submitted in partial fulfilment of the
requirements of the University of Brighton
for the degree of Doctor of Philosophy

April 30, 2019

Dedicated To

My father's soul.

All Iraqi martyrs.

My mother for her endless love, support, and encouragement.

Acknowledgement
I greatly appreciate the financial support received through my PhD study for more than four years from Iraq/ Ministry of Higher Education and Scientific Research/ Middle Technical University/ Institute of Management Rusafa. This work would not have been possible without the academic support from my committee members. I am grateful to Andrew Fish, Katerina Tsakiri, Diwei Zhou, and Antonios Marsellos for their guidance and advice. Andrew's thoughts and comments were extremely helpful in producing this work.

I would like to thank my mother, my husband, my children, my brothers, my sisters, my nieces, and my nephews for their unlimited love, support, and encourage.
I would like also to thank my colleagues in the Institute of Management Rusafa, and also my colleagues in the University of Brighton, Ahmed, Jenny, and Safa.

Declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the original work of the author. The thesis has not been previously submitted to this or any other university for a degree, and does not incorporate any material already submitted for a degree.

Abstract

Constructing statistical forecasting models is essential for a wide range of purposes, such as, water resources planning and management, financial planning, and managing inventory and production systems. In order to improve the accuracy of flood forecasting, this thesis provides three strategies, which are based on combining models.

The three strategies are applied to enhance the prediction accuracy for three models for time series data. The common factor between these strategies is the decomposed data. The Kolmogorov-Zurbenko filter, which is a developed version of the Moving Average filter, is applied to extract the decomposed data which are the three components: the long-term (trend), the seasonal fluctuations, and the short-term (noise) component. These components are separated based on the assumption that there is a gap (difference) in their spectra. The events that last for a short time scale, which is ranged between 2 days to 3 weeks, represent the short-term series. The next scale is the seasonal variations with a period of 1 year. Finally, any scales with a period more than one year are related to the long-term component.

The first strategy which improves the prediction accuracy of the Combined Multiple Linear Regression (CMLR) is carried out via modelling the error (residual) terms of this model by using an Autoregressive Moving Average (ARMA) model. The inclusion of an ARMA model will handle the problem of Autocorrelation between the residual terms of a CMLR. A CMLR is a Multiple Linear Regression model constructed by combining the data of the three components that are embedded in time series data which can be extracted using a filtering technique.

The second strategy which improves the prediction accuracy of the Transfer Function-Noise model (TF-Noise) is carried out via using the data of three components rather than raw data. Unlike the MLR, the structure of this model is inherently designed to take into consideration the special nature of time series data. These two models, MLR and TF-Noise, are Frequentist statistics-based models.

The third strategy, however, is related to a Bayesian statistics-based model, which is the Bayesian Multiple Linear Regression, BMLR model. The enhancement of this model is carried out via specifying three likelihood functions and three prior distributions for the three components rather than specifying one likelihood and one prior distribution for the raw data. Besides, a Bayesian Vector Autoregressive (BVAR) model is used to fit the data of the short-term component. Bayesian analysis enables us to incorporate prior knowledge or evidence from previous experiences or studies via prior distributions.

One application for the aforementioned developed models is to predict the water discharge for a river. The water discharge series for three stations that are located in cities Cohoes, Utica, and Poughkeepsie, and related to the Mohawk and Hudson

rivers in New York State, US, are used. The independent variables are temperature, precipitation, wind speed, tide, and groundwater level. Using the decomposed data, the R Squared value, which is a measure of how well the model fits the data, for the MLR has been increased to become 0.67 compared to the R Squared for the raw data which is 0.48 for Cohoes city data. In the Bayesian analysis, the Metropolis-Hastings Markov Chain Monte Carlo (MH-MCMC) algorithm is used to estimate the parameters and then the mean value of daily water discharge (flow). The parameter's estimates and uncertainties computed using this algorithm are compared to those computed using maximum likelihood method which assumes that the model's parameters and residual terms are normally distributed. Similar results are obtained using these two methods.

The Bayesian models, which are constructed using the raw and the decomposed data, are compared based on the Deviance Information Criterion (DIC) while the Frequentist-based models are compared using AIC and other model selection methods. All the model selection methods follow the rule that states the smaller, the better. The results show that the forecasting models constructed using the three components outperform models constructed using the raw data. Also, the results for Bayesian models are in favour of the combined BMLR-BVAR rather than the combined BMLR where the DIC value declines from 6593.116 to 521.385.

Multivariate time series, MTS, datasets are very common in different financial and business, economic, and hydrological fields. In many cases, it is desirable to compare the similarity or dissimilarity of a group of MTS datasets. A considerable amount of literature has been published on the subject of similarity and dissimilarity for this type of datasets using the raw data. However, no study has been conducted to examine the dissimilarity between MTS using the decomposed data. In this study, a methodology that is based on a component-based distance measure is adapted to separately capture the dissimilarities between the components of a group of MTS datasets. This approach will help us to determine the factors that affect each component in the system of interest. One advantage of applying this method is its ability to provide the required guarantees that enable us to use a forecasting model for one object, for example, a city, to forecast future values for another city. Moreover, to support the decision of similarity, the greater the proximity value for one or more of parameter-based statistics, such as MSE values for MLR models, the more similar are the objects.

As long as we have MTS datasets, a covariance matrix-based Euclidean and Non-Euclidean distance measures is applied. The statistical approach used to provide a decision about the dissimilarity is the Hypothesis Testing. According to the results of the hypothesis testing for mean distance/dissimilarity and the MSE values for the MLR models, the similarities between the data of the cities of interest are detected. Eight Euclidean and Non-Euclidean distance measures which are Euclidean, Procrustes, Riemannian, Procrustes Shape (Full-Procrustes), Cholesky, Power, Log

Euclidean, and RiemannianLe are used. The performance of these distance measures is examined based on the order of distance from smallest to largest with respect to the pairs considered, which are Cohoes and Utica (cu) Cohoes and Poughkeepsie (cp) and Utica and Poughkeepsie (up) where the distance measures ProcrustesShape, Procrustes, and Cholesky provided the smallest distance values. Based on the clustering analysis results for the distance measures, the distance measures ProcrustesShape, Procrustes, and Cholesky, for example, have been clustered in one group. This would mention that there is no difference in their performance with respect to the pairs of cities, therefore, we are able to use any one of them to compute the distance between the datasets of cities.

To provide a full picture of the dissimilarity analysis for time series data, we present three new distance measures based on three features from the frequency domain. These features are the periodogram, power spectral density and the cross spectral density functions. Using these distance measures, a comparison between MTS datasets in terms of frequency content can be conducted. The distance is computed based on the $X^T X$ matrix which is a positive definite matrix built by using the periodograms for the variables that have been calculated using the Fast Fourier Transform (FFT). The other matrix used is the Power Spectral Density (PSDE) matrix that is constructed using the power spectral and the cross-spectral density functions for the variables. For these two matrices, the results have matched for the distance measures Procrustes, Riemannian, Power Euclidean, and Log-Euclidean.

Abbreviations

- AR Autoregressive Models.
- MA Moving Average Models.
- ARMA Autoregressive Moving Average Models.
- MLR Multiple Linear Regression Model.
- BMLR Bayesian Multiple Linear Regression Model.
- VAR Vector Autoregressive Model.
- BVAR Bayesian Vector Autoregressive Model.
- TF-Noise Transfer Function- Noise Model.
- SACF Sample Autocorrelation Function.
- SPACF Sample Partial Autocorrelation Function.
- SCCF Sample Cross Correlation Function.
- PSD Positive Semi-Definite.
- PSDE Power Spectral Density Function.
- TE Temperature.
- WS Wind Speed.
- PR Precipitation.
- TD Tide.
- GW Groundwater.
- IG Inverse Gamma Distribution.
- NIG Normal Inverse Gamma Distribution.
- NW Normal Wishart Distribution.
- NIW Normal Inverse Wishart Distribution.

- FT Fourier Transformation.
- DFT Discrete Fourier Transformation.
- FIR Finite Impulse Response.
- IIR Infinite Impulse Response

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In time series analysis, one of the categorizations of describing and forecasting future values is based on the principle of "how far into the future the event is to be forecast". This categorization implicitly refers to the importance of using the decomposition of time series that yields components with different time scales. The decomposition of time series means separating the short-term and seasonal signals from the trend (long-term component). This separation process provides "clean" data that can be used for examining and forecasting trends and the causes that lead to different kinds of trends. Also, this type of clean data can be used to study the climate change and the reasons behind the changes.

One of the most important and common solutions to carry out the signals' separation process is to use a filtering technique. Different filtering techniques can be used to separate the scales of motion in a time series, for example, the Moving Average (MA) and the Kolmogorov-Zurbenko (KZ) filters. The result of the filtering process is three components the long, seasonal, and the short-term component. These components for a number of variables (independent variables) are incorporated using a structure to build a combined model to describe the dependent variables.

One of the new contributions of this study is to improve the predictive accuracy of the Combined Multiple Linear Regression (CMLR) model via overcoming the issue of Autocorrelation between the residual terms of this model. The new model can be called CMLR-Noise model. The existence of these autocorrelations violates one of the most important assumptions, which is the residual terms have to be uncorrelated. This assumption has to be considered in a regression model to be validated.

The second new contribution is to enhance the predictive accuracy of the TF-Noise model which can be conducted using the decomposed data rather than the raw data. This model is chosen as it is inherently structured using the lagged variables for the input and the output variables, where often the lagged variables have an important impact on the forecasting process. The enhanced model, which can be called the

Combined TF-Noise model (CTF-Noise), is constructed using lagged variables for the three components.

In optimization context, and based on the most common strategies' types that are used to estimate the coefficients, the existing statistical methodologies are limited as the predictive framework is often based on the mean response predictions. In this case, the uncertainties of data, model's structure, and parameters, are ignored. This, in turn, leads to increase the risks of adopting incorrect decisions by the researchers. Based on this, the provision of statistical frameworks that explicitly handle this issue has to be considered. Statistically, one of the analyses that take the uncertainties of data, parameters, and model's mechanics into consideration is Bayesian analysis. Working within Bayesian statistical framework enables us to consider the prior knowledge, which is mathematically incorporated into the constructing model via a prior distribution. Bayesian paradigm creates posterior and posterior predictive distributions that will be used to provide some summary measures (inferences) for the parameters of interest, future values, and also the credible intervals. The third new contribution is to use the decomposed data to construct a Combined Bayesian Multiple Linear Regression (CBMLR). Using the decomposed data rather than raw data has increased the value of R Squared for Utica city from 0.43 to 0.56.

Examining the feature of similarity or dissimilarity between MTS datasets is an important topic in many fields. Different methods have been suggested to measure the similarity using the raw data. However, examining the similarity or dissimilarity between MTS datasets using the decomposed data has not been considered. This examination will provide a detailed picture of how much the datasets for each component are similar to each other. Besides, in case that the datasets are not similar, we are able to determine which component is responsible for obtaining this result. Knowing the responsible component will enable us to determine the factors and events that lead to such a decision of dissimilarity. Relying on the results, different tasks can be accomplished. For example, if we find that the MTS datasets for two cities are similar in terms of data of the three components and the raw data as well, we will be able to use a forecasting model for one of these cities to forecast future values for the other city.

In this study, we show the feasibility of using hypothesis testing for mean distances to make a decision on whether the behaviour of the raw and the three components data for a group of MTS datasets (objects) is similar. The distance is computed between the covariance matrices of the variables for the considered objects. The distance between covariance matrices is already used in different fields, such as diffusion tensor imaging. However, being used for the purpose of examining the dissimilarity between MTS datasets through hypothesis testing is a new contribution in this thesis. This method can be applied for any MTS datasets for a group of objects.

Furthermore, a parametric-based value, for example, Mean Square Error (MSE)

can also be used to confirm the decision of similarity of data.

Additionally, in this study, we present three new distance measures based on three features from the frequency domain. These features are the periodogram, power spectral density and the cross spectral density functions. These distance measures can be beneficial to compare the frequency content for MTS datasets.

One application for the new contributions, which are the developed models CMLR-Noise, CTF-Noise, and CBMLR, is to provide a hydrological forecast. A hydrological forecast provides the estimation process for future states (values) for hydrological phenomena. These forecasts are fundamental for performing different operations that are related to water's infrastructure and resources management. Additionally, they can be efficiently utilised to mitigate the risk of natural disasters such as droughts and floods. In the hydrological field, the amount of water that exceeds a measuring point or a gauging station in a river at a specific time is called water discharge. The aim is to provide more accurate forecasts for the amount of water discharge from a river for three stations.

The structure of this thesis is organised as follows. The remainder of this chapter, Chapter 1, Section 1.1 presents the definition and types of time series. Section 1.2 introduces a brief overview of the most common methods used to analyse hydrological data. Literature review is presented in Section 1.3. The kinds of floods are highlighted in Section 1.4. Section 1.5 provides a brief description of linear systems. General description of filters is displayed in Section 1.6. Section 1.7 provides a brief explanation of the Moving Average (MA) filter. Section 1.8 explains the KZ filter. The residual analysis is introduced in Section 1.9. The most common types of residual's plots are presented in Section 1.10. Moreover, in Section 1.11, the process of fitting residuals's data using an ARMA model, which is also known as a Box-Jenkins model, is highlighted. Information about the Sample Autocorrelation Function (SACF) and Sample Partial Autocorrelation Function (SPACF), is displayed in Section 1.12. Section 1.13 describes one of the most specified models, which is the Autoregressive model of order one, to the residuals in a regression model. Section 1.14 introduces the forecasting kinds. Section 1.15 presents the definition and tools for performing frequency analysis. Section 1.16 gives information about the data used in our study.

Chapter 2 displays two frequentist statistics-based methodologies, which are Multiple Linear Regression (MLR) and Vector Autoregressive (VAR) Models. The structures of these two models are used to build combined models. Moreover, our new model, which is a Combined Multiple Linear Regression-Noise (CMLR-Noise) model built using an ARMA process for the residuals, is provided in this chapter. The spectral content, which is represented by the periods for the studied series, is given in this chapter. Finally, the contribution for each component is also computed in this chapter.

Chapter 3 presents the new Combined Transfer-Function-Noise (CTF-Noise) model,

which is built using the three components the long-term component, the seasonal variations, and the short-term component. The tools that are used to determine the numerator and denominator for each term, as well as the Box-Jenkins model for the Noise series, are also presented in this chapter. This model will enable the use of a number of lagged variables. The prewhittened data are used in this chapter.

Chapter 4 provides the base of how to use Bayesian analysis to estimate the parameters of a combined model. The types of prior distributions and Bayesian Multiple Linear Regression (BMLR) are considered in this chapter. The Posterior and Posterior Predictive Distributions, Credible Intervals, the Highest Probability Density (HPD), and the ways used to Check and Compare Bayesian Models are also among the subjects considered in this chapter. The second part of this chapter is concerned with how to apply Bayesian Vector Autoregressive Model, BVAR, for the short-term component data. The Prior distributions, including Minnesota prior, are highlighted in this chapter. The contribution of each component is also produced in Chapter 4.

The essence of Chapter 5 is the application of Hypothesis Testing using the raw and the decomposed data. The work in this chapter is divided into three parts. The first part introduces a method to compare mean vectors of variables for two cities using Hoteling T-Squared test. In this part two cases are considered. The first case includes all variables, which are temperature, precipitation, wind speed, water discharge, tide, and groundwater level. In the second case, we ignore the hydrological impact by eliminating the variables of water discharge and groundwater level. Different results are obtained for the two cases.

The second part of this chapter presents a method to compare mean distance/dissimilarity for a group of objects (cities). The distances computed are between the covariance matrices of the variables, including the covariance matrices of the independent variables of the MLR for the raw data. In the third part of this chapter, which is carried out to gain insights into whether there is a difference between the results of dissimilarity using the time and frequency domains, three new dissimilarity measures are introduced. These measures are specifically designed using Euclidean and Non-Euclidean metrics for a number of frequency domain-based features. Euclidean and Non-Euclidean metrics here are applied for covariance matrices data in the time domain and Power Spectral Density (PSDE) and X^TX matrices of the periodograms for the variables in the frequency domain. The final conclusion and the future work are introduced in Chapter 6.

1.1 Time Series

A time series is a set of observations which are recorded or observed regularly or irregularly to measure the variation of a specific phenomenon through time. Generally, a time series can be classified into a number of kinds based on some features which are:

- Univariate and Multivariate time series, this type is determined by the number of the studied series.
- Discretely and continuously recorded observations, this kind is related to the time that is used to record the studied data. If a series of observations is defined and represented at any instant of time, this series will be continuous. The best examples for this type of signals are the sine and cosine functions. On the other hand, if a series of observations is defined and represented at certain times, the series will be discrete. This type of signals is also called digitalized signals.
- Stationary and Non-Stationary time series, where stationarity is a crucial feature that should be considered in a series.
- Deterministic and Stochastic time series: Some time series describes a regular and deterministic process, which can be modelled using a deterministic modelling technique. But when the behaviour of the series has an indeterministic pattern, this series is a stochastic process and should be treated and modelled using one of the models that take the uncertainty into account.

Typically, time series analysis can be utilised effectively to achieve many tasks. For example, explaining and describing the data, forecasting future values for the univariate and multivariate time series, clustering and classifying a group of series, and many other functions. Essentially, there are three distinguishable components, scales, which are Trend (long), Seasonality, and Short-term component (noise). The trend usually represents the long-term component behaviour of a time series as it is often identified with regular variations, which are slowly developing. This component can be efficiently modelled using a linear equation. With regard to the seasonality, this type of variations in the studied time series is attributed to the change in the time that is used to record the considered series, for instance, 4 seasons and 12 months. Also, some specific intervals in a year, for example, Christmas time, Weekends, Easter Holiday, and many other events can be regarded as seasonal factors. Finally, the short-term component can be considered as the most important part compared to the other components as it contains the “ unknown and indeterministic ” events that last for a short period of time; this component has attracted lots of researchers to investigate its nature and behaviour.

In fact, these three components are embedded in a time series data and in order to gain a clear insight on the trend in the underlying time series, it is essential to decompose this series [36, 83]. The isolation of the seasonal effects and the irregular signals can be enabled by applying a filtration technique. In most cases, to precisely determine the trend for a desired data, seasonal variations have to be removed a priori. For example, in some regions in the USA, the unemployment rate in June, due to the agricultural season, always decreases compared with May, so this temporary variation does not necessarily reflect a real decline in this rate. In such a case, the seasonal fluctuations have to be eliminated. This will definitely provide a clear insight about the real trend for the unemployment rate. There are different techniques to perform the process of decomposing, such as the filtering technique. Using a filtering mechanism will often produce one series, for example the long-term component, which will be used later to create the other required components. This strategy of decomposing and modelling has been applied widely in different fields such as economics, engineering, and meteorological studies [101, 100, 36, 122, 37].

Furthermore, for the purpose of modelling these various components (patterns), a variety of methods have been used, such as the regression and autoregressive of different orders models [15]. In general, there are two types of constructions that can be used to fit a model for the aforementioned components, which are the static and dynamic models. While the static models take into account only the current data, the past and current data are included in the dynamic models. For example, the multiple regression model is regarded as one of the most important and common types in the static modelling area. This model will enable the researchers to examine the relationship between a dependent and a number of independents variables. The error terms are also considered in this model.

On the other hand, an autoregressive, AR model, can be considered as a good example for the dynamic system. This model will consider the use of lagged versions of the variables in its structure.The dynamic models are essentially related to the lagged variables. In the time series analysis, Lagged variables are very important elements in the constructing models where often a value of the series of interest in a certain instant of time will have impact on its future values. The lagged variables can be of any order, and this order is selected based on measures such as the autocorrelation and partial autocorrelation functions. Although static system has been widely used for modelling the components long, seasonal, and short-term component, the dynamic structure can provide more accurate results for forecasting process.

Recently, many countries in the world have witnessed a flooding phenomenon. This process is often unexpected and most often attributed to climate change. The prediction of this phenomenon is a difficult task. Many researchers have dealt with this subject using different procedures [84, 112, 21]. The statistical models and flood histories have been used extensively for performing the forecasting process. Using

the statistical analysis, specifically time series models, for forecasting the amount of water discharge, for example from a river, has increasingly attracted the hydrologists. In particular, as long as we have data with time series form, we will need to exploit it in the process of building forecasting models [23, 69, 88, 29, 48, 28, 96, 100, 6].

In this thesis, a filter technique has been implemented to separate the scales embedded in the studied time series data. Also, for performing the task of constructing a model, different structures have been applied. There is no doubt that utilising the lags of the time series can significantly affect the efficiency and accuracy of these models [98].

1.2 Hydrological Forecast

In the hydrological field, researchers generally prefer working with the deterministic techniques rather than stochastic manners to compute the needed quantities [23]. This also can be observed with probabilistic methods. However, although they are more expensive in terms of time and the computational cost, the stochastic and probabilistic methods are required to forecast future values for a hydrological process [98]. These approaches take into account the non-linearity and the uncertainty of data for some hydrological systems such as water discharge from a river. A variety of stochastic methods have been used for processing water discharge data from a river [6, 96]. Most of them are accomplished using simulation process such as open channel and rainfall-runoff simulations. The two terms of stochastic and probability represent the randomness in a specific system. However, the majority of stochastic models are time-dependent models while probabilistic methods are independent of time [23].

Different factors affect the amount of water discharge from a river. Some of these factors are climatic conditions such as precipitation and temperature. Precipitation is regarded as one of the most important variables that affect the water discharge amount [21, 48, 28, 96]. Moreover, the water discharge system is a dynamic process and its data are often regularly recorded. Because of the nonlinearity and the uncertainty of hydrological systems data, the process of building a forecasting model may be a challenge in the operational hydrology area in spite of all the advances in the weather forecasting in the recent decades.

Often, one of the time series techniques is used to analyse the hydrological system. Several models have been built utilizing different kinds of techniques. Most of the constructed models are built without removing the short-term variations and the meteorological effects from the long-term trends. In time series analysis, due to its role to enhance the prediction accuracy, it is important to separate the undesirable data, such as short-term variations, by using one of the filtering techniques [100, 36, 118]. In this thesis, as the data for each variable have different time scales, the KZ filter is

used to separate the time scales for each variable into long, seasonal, and short-term component. Based on these components, a number of models have been proposed in an attempt to predict the water discharge accurately.

1.3 Literature Review

The organization of the literature review has been divided based on the methods considered:

- The Kolmogorov-Zurbenko Filter (KZ).
- Building and estimating the parameters of forecasting models using classical and Bayesian statistics.
- Forecasting the Amount of Water Discharge
- Similarity and Dissimilarity measures for time series data.

1.3.1 The Kolmogorov-Zurbenko (KZ) Filter

In meteorological and hydrological studies, to accurately analyse any time series data, the synoptic and seasonal fluctuations have to be removed. The resultant dataset is clean from any undesirable signals and can be used to precisely determine the trends, the climatic changes and the reasons behind these trends and changes [36]. Based on this perspective, Tsakiri et al. (2014) have applied the KZ filter in the analysis of the water discharge time series of the Schoharie Creek river, New York. This river is regarded as one of the two most essential tributaries for the Mohawk River [100]. In addition, taking into account the special nature for the region related to this river, two predictive models have been constructed to describe the potential amount of water discharge for the short-term component. These two models are one for the summer as flooding is often caused by extensive precipitation and the other for the winter where flooding is caused by rapid snowmelt. The parameters 29 days and 3 repetitions were used for the KZ filter to produce the global-term component. As a consequence, and compared to the pre-decomposition R Squared which was 59%, the percentage of the explanation for the regression model increased to become approximately 81%.

To investigate and forecast the global climate changes by using images of the spatial and temporal fluctuations of temperature, which were taken all over the world, Zurbenko and Lua (2012) smoothed out and interpolated gridded temperature data using the KZ filter [121]. Then, this smoothed series is used to form a global map for the long-term trend which represents 6 years, and EL Nino-like movements which represent 2-5 years. The data for this study was taken for the period 1893 to 2008.

Latitude and altitude variations were collected to enable the application of the KZ filter. The monthly observations are decomposed into different time scales depending on the spectral analysis. The results mentioned that the desired signals are constructed from data with high noise (high frequency signals) by using the KZ filter.

Wise and Comrie (2005) exploited the ability of the KZ filter to analyse the time series of the Particulate Matter (PM), Tropospheric Ozone (O3), and the Meteorological Conditions [115]. In order to enable the planners and managers of air quality to make the correct decisions and map for the future for the management of emissions and determine the policy for air quality, the meteorological signals must be separated from the two series O3 and PM. This separation will help to accurately investigate their trends, which are devoid from the climatic variables. The separating process was performed using the KZ filter with the parameters 15 days and 5 iterations as the window width and the number of iterations, respectively. That means, depending on the width and the repetition times for the filter, any cycle with a period of less than 33 days was removed. The details of determining the number of days for the cycles to be removed are presented in Section 1.8.

Milanchus et al. (1998) tried to evaluate the effectiveness of the ozone management efforts. However, because the meteorological fluctuations are embedded in the ozone data, it was necessary to filter out the meteorological effect to obtain " clean" data. So, to provide a precise evaluation, the KZ filter was applied to clean the ozone time series from the climatological conditions by isolating the ozone precursor emissions data [73]. As a consequence, the assessment of the effectiveness of the regular programs for the ozone air quality can be accomplished. They took the series of the average ozone concentrations for five states in the USA for the period 1984-1995. Also, hourly time series values for multiple meteorological variables, which were temperature, humidity, cloud cover, and wind speed were collected for the same period. To evaluate the influence of the control programs on the ozone air quality effectively, the different scales embedded in the time series of ozone and meteorological time series were separated.

By using the KZ filter, the authors were able to explain approximately 70% of the variations in the ozone data that are attributable to the meteorological variations. Depending on this result, the possibility of detecting and tracking variations for a meteorological adjusted ozone time series increases.

Moreover, Rao et al. (1997) described the space and time features for the ambient ozone series by using filtering technique [83]. To obtain effective results for the ozone problem, the seasonal and synoptic components have to be removed from the ozone time series. Two different cases were implemented which are $KZ_{15,5}$ and $KZ_{365,3}$ to derive the global component, which involves the long and seasonal components together, and the long-term component, respectively. The authors verified that using filtration mechanism solved the problem of identifying the percentages of variations

due to the meteorological fluctuations.

Furthermore, to separate the synoptic and seasonal variations, Eskridge et al. (1997) used four techniques, which are Anomalies, PEST, the KZ filter, and the Wavelet Transform [36]. In their research, the data of the temperature time series for Hong Kong was utilised. The results could be summarized as follows: By smoothing the data with the $KZ_{15,5}$ filter, all cycles with a period of less than 33 days were removed. On the other hand, to remove the yearly cycles, another two new parameters were selected, which were 365 days and 3 iterations. This filter was sufficient to isolate the yearly seasonal signals and all cycles with small time-scales, signals with less than 1.7 years, leaving behind only the long-term trend. Moreover, the variance percentage of the short-term component to the total variance was approximately 20%. The total variance was calculated for the temperature series before the filtering process. The variance's percentage for the long-term component to the total variance was almost 3%. On the other hand, the variance of the seasonal component again to the total variance was about 77% for the temperature time series for Hong Kong. Cleaning the raw data from the signals of noise and seasonal fluctuations is am essential step to calculate the long-term trend. With regard to the Anomalies and PEST methods, the two types of signals, which were the seasonal and synoptic, were not adequately removed. Finally, the Wavelet Transform technique produced the same results as the KZ filter. However, the KZ filter can be applied even though a number of values are missing, which would be impossible with all of the other three methods. In addition to this useful property, the computation process of the KZ filter is relatively easier compared to the same used techniques.

Moreover, employing the KZ filter to assess the temporal and spatial fluctuations in the ozone air quality by using data of ozone concentration from multiple monitoring sites in the United States, Rao et al. (1995) were able to verify that there was an improvement in the ozone air quality in most of the studied locations [81]. This was performed by considering the temperature-adjusted ozone time series data. That would suggest that studying the trends using an adjusted time series is much easier and provides more precise results than using raw data.

To moderate or filter out the effects of the meteorological variables from the ozone concentrations, Trivikrama and Zurbenko (1994) succeeded to detect the changes in the ozone air quality by separating the meteorological variations [82]. This was conducted by applying the $KZ_{29,3}$, which attenuated any cycle with less than 50 days for the ozone concentration and temperature time series. Simple linear regression was used to study the relationship between the filtered series, where the ozone series was the response variable and temperature represented the predictor, R Squared for this model was 0.83%. Since the temperature variable was taken into account when the process of constructing the model was performed, the residuals of this model represented the variations that can be attributed to the emissions.

1.3.2 Modelling Methods: Regression, Vector Autoregressive, and Transfer Function Models

Regression and Vector Autoregressive Models

Tsakiri et al. (2014) have used regression analysis to construct forecasting models for the long, seasonal, and short-term component for water discharge and also the final combined water discharge model [100]. The analysis by using this method has enhanced the R Squared value to become 0.81. Also, Tsakiri (2010) used the regression analysis and vector autoregressive model of order one to study the relationship between the ambient ozone, temperature, wind speed, and precipitation [101]. The results revealed that the explanation and the prediction for the response variable ambient ozone, has been improved, where the R Squared value became 0.88. Using the regression model, Trivikrama et al. (1995) were able to detect the downward pattern in the trend of the ozone concentration.

Transfer Function-Noise Model (TF-Noise) model

Bierkens et al. (1999) separated the groundwater time series data into two parts. One part to construct the transfer function-noise model, and the other to validate the resultant forecast model. The precipitation surplus series was used as the input variable. Accurate predictions and representative stochastic simulations of daily groundwater data were produced [13].

Furthermore, A statkie and Watt (1998) built univariate and multivariate models to predict daily streamflow data [6]. Non linear Nested Threshold AutoRegressive model (NETAR) was applied as a univariate model for the streamflow. For the multivariate method which used the non linear snow melt and the effective rain data as the input variables, the transfer function-noise model (TF-Noise) model for the streamflow was constructed. It was found that TF-Noise model would be better to describe the forecasting model than the others as the value of the MSE for this model was less than the MSE value for the NETAR model.

Harris and Liu (1993) applied multiple TFM to forecast the consumption of the electricity in the south east of the United States. Using the electricity consumption as the output variable and each of the electricity prices, heating degree days, cooling degree days, and percapita disposable income as the input variables, TF-Noise model was constructed [47].

Moreover, Khan (1990) used a multiple input transfer function model to predict the percentage of the gloss value for the coated aluminium that was exposed to an open environment [63]. Monthly temperature and relative humidity were selected to be the input variables. Two univariate time series models were built for the temperature and relative humidity variables to obtain the prewhitening values with the gloss's

data. The univariate and multivariate models were used. For the univariate method, autoregressive model of order one, AR(1), was suggested depending on the behaviour of the partial autocorrelation function. On the other hand, for the multivariate TF-Noise model, temperature and relative humidity were used to construct the forecasting model. The adjusted Root Mean Squared Error (RMSE) was used to diagnose which model would adequately accomplish this forecasting task. Based on the RMSE's values, the multivariate model was chosen to perform the forecasting process for the future values for the gloss data.

1.3.3 Forecasting the Amount of Water Discharge

Albostan and Onoz (2015) applied three different methods that have an ability to identify the chaotic behaviour for the daily water discharge series [3]. These methods are phase space reconstruction, local approximation, and correlation dimensions. Their study has been performed using the daily water discharge data for the Yesil Irmak River Basin for the period (1977-2002). The researchers divided this dataset into two groups. One was selected to perform the analysis and the second dataset was utilised to test the validity of the model. Taking into consideration that global warming may be regarded as a serious threat to nature, an accurate prediction can effectively provide a tool for facing the hazards of this important phenomenon [3]. Their study has successfully identified the nonlinearity for the data regarding the river flow. The predicted values have a relatively high correlation coefficient with the raw data which may confirm the validity for the suggested prediction model. Shijin et al. (2012) have built a hybrid forecasting model using Support Vector Machine to forecast the amount of water discharge from the Huaihe River [88]. This research has shown the possibility of improving the prediction process by partitioning of the water discharge series.

Moreover, Svetlíková et al. (2007) analysed the monthly water discharge, precipitation, and rainfall time series in the region of the Klastorske Luky wetland in Slovakia for the period between 1901 and 2004 [96]. At first, the trend, seasonal, and short-term component were investigated. The Autoregressive Moving Average Models (ARMA), Threshold AutoRegressive (TAR) were utilised to fit models for the components. The TAR with exogenous component and TAR that was associated with the Long Memory models were used to analyse the time series data for water discharge and rainfall. The best prediction models for the water discharge were the nonlinear TAR models and a group of the TAR and Long Memory equations. On the other hand, for the precipitation data, the ARMA models were the only suitable methods.

Damle and Ali (2007) investigated the possibility of forecasting the flood using Time Series Data Mining (TSDM) [28]. By exploiting the ability of chaos theory to

provide a structured explanation for the irregular nature for some of the Geophysical phenomena, such as floods and earthquakes, a novel approach to forecast flooding was presented. The results revealed that a successful prediction model for the flood was produced using the selected event characterization function. This function relies on a step-ahead function and the objective function.

1.3.4 Covariance Matrix and the Euclidean and Non-Euclidean Metrics

Recently, interest in the statistical analysis for the data that are located in a Non-Euclidean space (e.g. Riemannian Manifold) has been growing increasingly in a wide range of fields. For example, in Diffusion Tensor Imaging (DTI), the data is represented by a covariance matrix of dimension 3×3. Therefore, this data belongs to a Riemannian Manifold [34]. Also, Shape Analysis and the estimation of the covariance structures are also considered as examples for the data that belong to a Non-Euclidean space. The classical statistical methods that are often used to analyse data in an Euclidean space have been also applied for analysing data in the Riemannian Manifold space but with taking into consideration the specific nature of them. Estimation of even simple statistical tools such as mean and variance can be a difficult task for this type of data.

Zhou et al. (2016) have presented a number of non-Euclidean statistical methods that deal with Regularisation, interpolation, and visualisation for diffusion tensor [120]. Using the scale invariant power-Euclidean metric, a group of anisotropy measures that are specifically beneficial for visualisation has been introduced. A discussion of using weighted Procrustes methods for interpolation and smoothing for diffusion tensor has been also highlighted. A key relationship between the principal square root Euclidean metric and the size-and-shape Procrustes metric has been also established in the space of symmetric positive semi-definite tensors. The performance of a number of non-Euclidean metrics has been compared using a dataset of human brain diffusion-weighted magnetic resonance imaging. The results of the metrics Log-Euclidean, Euclidean Root, and Procrustes, which are non-Euclidean metrics, are much better than the results of the Euclidean metric.

With regard to covariance matrices, Li et al. (2015) have developed visual tracking dependent upon the log Euclidean Riemannian metric for statistics on the covariance matrices for image features. Also, to reflect the appearance changes for an object, a novel online log Euclidean subspace learning algorithm IRSL has been suggested [68].

Pigoli and Secchi (2012) applied the spatial statistical methods to investigate data, which has been collected from a non-Euclidean space [80]. Taking into account the case of the spatial dependence, they have proposed a new estimator to obtain

an estimation for the mean of the covariance matrix for two variables which are temperature and precipitation. They regarded the proposed mean covariance as a significant estimator when they compared it to one that did not consider the spatial dependence. Tsakiri (2010) investigated the effect of the noise for a covariance matrix, which was defined by using the spectral decomposition [101]. Also, she extended her research to cover the noise problem in the principle component analysis.

In addition, taking into consideration the Non-Euclidean nature for the space of the positive semi-definite symmetric matrices, Dryden et al. (2009) estimated the mean of a number of covariance matrices in DTI, where sample covariance matrices were used as the data for the analysis [34]. Moreover, this study concentrated on the Procrustes size and shape approach. Jian-Hong and Song-Gui (2006) have proposed a simple approach to provide the number of the distinct eigenvalues and the spectral decomposition for a covariance matrix by using the variance component model [60]. This approach relied on the partial ordering of a symmetric matrix and a relation matrix. This method was used to check the linear dependency between these distinct eigenvalues.

1.3.5 Similarity and Dissimilarity Measures

A considerable amount of literature has been published on the subject of investigating the similarity and dissimilarity for time series data. These publications can be divided into non-parametric and parametric-based distance measures.

Using Non Parametric Methods: Raw Data, Time and Frequency Domains Features

Several methods have been proposed to perform a comparison (similarity or dissimilarity) between time series data. Some of these approaches are only simple descriptive measurements applied by using a number of distance measures using the actual data. Methods that utilise some of time series functions have been also used to examine the property of similarity between the data of interest. Huang et al. (2016) have proposed a new similarity measure that relies on the distribution of the eigenvalues of the Hankel matrix [54]. The researchers have proved that this approach can satisfactorily work even with time series of different lengths, nonlinear features, and complex fluctuations.

With reference to use of time domain features, Kovaci (1996) (as cited in [26]) presented a distance metric between two time series using the autocorrelation and cross correlation functions. Moreover, using the estimated residuals for some linear and non-linear models, Tong and Dabas (1990) (as cited in [17]) examined the affinity between these fitted models. Bohte et al. (1980) ((as cited in [26]) used the

autocorrelation function to compute the distance between two time series.

On the other hand, using the frequency domain features, Vemulapalli and Jacobs (2015) have presented the learning Riemannian metric distance for the positive definite matrices[107]. They have evaluated this measure on face matching and clustering processes. The face matching and semi-supervised object classification results showed that the learned log-Euclidean geodesic distance outperforms classification results of other types of distances.

In addition, based on the dissimilarity measure that is defined using the Riemannian distance measure, Li et al. (2009) have classified EEG signals using a k nearest neighbour algorithm [68]. For each frequency, the power spectral density (power spectrum) can be considered as a point on a Riemannian manifold, in this case, the Euclidean distance will no longer be suitable for use. Instead, a geodesic distance function has to be established to measure the Riemannian distance between two points. Relying on the normalized periodogram, Caiado et al. (2006) have suggested a new distance measure to cluster the stationary and non stationary time series data [17]. They compared the new measure to different classical types of time series measures, where this proposed distance measure outperforms the others.

Investigating the Dissimilarity Using Parametric Methods

This type of analysis can be classified as an inferential analysis [26]. Here, rather than depending on the data of the time series itself and its features, the comparison process can be performed using the parameters of the fitted models, such as the parameters of the Autoregressive Moving Average Model (ARMA) models. In this respect, Kalpakis et al. (2001) considered the problem of clustering ARMA time series, which are fitted using ARMA models. They proposed the use of the Euclidean distance between the Linear Predictive Coding (LPC) cepstrum of two time-series as their dissimilarity measure [62].

Piccolo (1990) (as cited in [17]) proposed a metric that uses the coefficients of the autoregressive moving average (ARMA) model to measure the degree of dissimilarity between time series data. Piccolo utilised the Euclidean distance to calculate the dissimilarity between two vectors of the ARMA model's coefficients for the industrial production series. He also constructed a test statistic by using some of time series functions to examine whether two sets of data have been sampled from a common distribution.

Additionally, Thomson and De Souza (1985) provided a dissimilarity measure using Mahalanobis distance for the AR models and studied the properties of the distribution of this unit. This criterion was widely used in speech recognition field (as cited in [26]). Furthermore, Anderson (1993) used the spectral distributions as an example for frequency domain [72], and Melard et al. (1991) utilised the autocorrelation

function to perform the hypothesis testing as an example for time domain.

1.3.6 Bayesian Methods

Wang et al. (2017) have presented a Bayesian method using Metropolis-Hastings Markov Chain Monte Carlo, MH-MCMC, algorithm for forecasting daily river flow rate for Zhujiachuan River in China [112]. The results obtained have shown that the proposed Bayesian method is able to produce adequate credible intervals for flood quantiles that are in accordance with empirical estimates. Tutberidz and Japaridze (2017) have exploited the structure of VAR modelling method to construct a model for forecasting activities of business [104]. A macroeconomic model for Georgian economy was built based on few variables. Bayesian analysis has been used to estimate the parameters of this model where sensible forecasts for the variables of interest have been obtained.

Using Multi-Modal Neuroimaging Data, a BVAR Model for Multi-Subject Effective Connectivity Inference, has been estimated (Chiang et al. 2017) [24]. Simultaneous inferences on effective connectivity at both the subject-and group-level have been allowed by using a Bayesian variable selection method. In Tourism field, Roma et al. (2016) have proposed two BVAR models of order one, BVAR(1), for the sectorial regional employment and the sectorial regional Gross Value Added (GVA) in Algarve [85]. In their research a Minnesota prior has been applied. The results of their study revealed a number of important facts about the influence of the international financial crisis that happened in 2007 on the considered factors, which are the employment average and the GVA.

Lima et al. (2016) have produced an estimated local and regional Generalized Extreme Value distribution parameters for flood frequency analysis using a hierarchical Bayesian framework [69]. The findings of their research have shown that adequate credible intervals for flood quantiles have been obtained. To explicitly model and reduce uncertainties for local and regional Generalized Extreme Value (GEV) distribution parameters for flood frequency analysis, Carlos et al. (2016) have developed a hierarchical Bayesian GEV model in a multilevel, hierarchical Bayesian framework [69]. They handled the problem of using data from multiple stations with missing values and different periods of records. Moreover, Cuaresma et al. (2016) have developed a BVAR model to forecast a set of macroeconomic and financial variables [27]. A set of hierarchical priors have been used and they compared the accuracy of forecasting of this suggested model to a naive univariate model. The forecasts that are based on the new developed model tend to outperform forecasts from the univariate model according to the Root Mean Squared Error (RMSE) values. Herr and Krzysztofowicz (2015) have presented the theory and algorithms for Ensemble Bayesian Forecasting System (EBFS) to provide a general Bayesian technique for ensemble forecasting that

can be easily used for large basins [49].

Chan (2015) introduced a class of large BVAR models that are flexible to include non-Gaussian, heteroscedastic and serially dependent residuals [19]. Using MCMC method, a unified approach has been presented to estimate the parameters. Based on in-sample and out-of-sample data, the forecast performance of this approach has outperformed the performance of common standard method that assumes independent, homoscedastic Gaussian innovations. Their approach is more flexible in terms of satisfying the assumptions of covariance structures for the innovations.

Additionally, a large BVAR model that describes the complex relationship between the main components of the Harmonized Index of Consumer Prices and their determinants for the Euro area has been constructed by Giannone et al. (2014) [87]. Their model provided precise forecasts in real-time. Gupta et al. (2012) have taken into consideration the impact of monetary policy on the dynamics of the USA housing sector [44].

Based on the impulse-response functions obtained from a large-scale Bayesian Vector Autoregressive (LBVAR) model that used 143 monthly macroeconomic variables, a negative effect on the housing sector at the national level by contractionary monetary policy has been detected. In fact, most of the macroeconomic studies work with multivariate time series models such as VAR. This type of models is often accompanied by a large number of parameters, and as a consequence, over-parameterization problems occur [64].

Villani (2005) developed a numerical simulation algorithm for processing VAR for stationary time series that enables incorporating these prior beliefs adequately [108].

Reis and Stedinger (2005) were able to develop a Bayesian MCMC method that takes the uncertainty in parameters into consideration for constructing a posterior distribution for flood risk [84]. LeSage and Krivelyova (1999) developed a Bayesian prior distribution based on cross-sectional autoregressive models to be used for forecasting studies that involve spatial variables [66]. They created a spatial weighting matrix that imposes specific factors on the parameters of the VAR model. To evaluate the performance of this new prior, the results of the forecasting accuracy were compared to the Minnesota prior results where this new proposed prior confirmed its ability to enhance the accuracy of forecasting for macroeconomic variables.

The estimation is normally accomplished by using classical statistical approaches. taking into account the statistical uncertainties and involving some belief priors has not been exploited in the hydrological field [21]. To incorporate them, Chbab and Duits (1999) built a number of probability distributions using Bayes' theorem for estimating quantiles of discharges [21]. The results showed that applying Bayesian analysis has successfully provided estimation results that were better than those gained by the classical maximum-likelihood estimates.

1.4 Floods Types

There are five different types of floods Flash Floods, Coastal Floods, Urban Floods, River (Fluvial) Floods, and Ponding (Pluvial) Floods. In this thesis we concentrate on the Fluvial Flood. Naturally, Fluvial floods occur when the rainfall continues for a long period of time where the amount of water will exceed the capacity of the river's bank. This amount of water accumulates from different sources where sometimes the kind of soil is saturated or hard which makes the process of absorbing this amount of rainfall much harder. As a consequence, most of the rainfall will flow to the river as well as the rain that directly falls into the river. Sometimes, ice-jam and extensive snow melt can cause this kind of flooding.

Hydrology is the science of studying movement, distribution, and the quality of water on the earth. In hydrology, water discharge is defined as the amount of water that flows outside the river's bank, and this amount is often measured by m^3/s (cubic meters per second). During a year and in the short-term component, the volume rate for the water discharge from a river oscillates with different scales with respect to the rainfall periods. Moreover, in the flood management and surface water planning, obtaining a reliable estimation of the water discharge from a river can be a crucial task [48]. With an efficient estimation, which in turn leads to obtain accurate predictions for future values, the possibility of decreasing the hazardous of floods, by warning people and make some preparations, will be achieved. For the purpose of analysing the water discharge, the Mohawk and Hudson Rivers in the New York State have been selected to perform this study as they have long flood records and many studies have been conducted for investigating the reasons behind this critical event [86].

Since its inception before 10,000 years ago, the Mohawk River has witnessed unusual flood. There are two main types of events "Free water" and "break-up" which cause flood in this river [41, 42]. The former event appears in late summer and early autumn, where large amount of precipitation falls. The reason of the later event is related to the break-up of river ice, as a consequence to the high temperature; also melting snow and heavy rains commonly occur during winter and early spring.

1.5 Linear Systems

In general, studying linear systems can be useful not only for its significant role in determining the nature of the relationships between variables, but also for its ability to formalise many of the procedures for filtering the data of time series by removing some undesired components, such as seasonality and noise. Most literature about the linear systems is written from an engineering perspective, for example Bendat and Piersol, (2000) [10]. In addition to the identification process for the input and output

variables for the system, these studies specifically concentrate on some aspects, such as the digital communication and the control theory.

Typically, there are two types of linear systems time-invariant and time-variant systems. The term time-invariant means that whenever the time of the input variable changes, the time for the output will also directly change by the same amount. In other words, if the coefficients of a linear equation are constants, the defined system is a time-invariant one.

1.5.1 Linear Systems in the Time Domain

In the discrete time, where the observations are recorded for a sequential time, a time-invariant linear equation can be written as:

$$y_t = \sum_{k=-\infty}^{\infty} h_k x_{t-k} \tag{1.1}$$

where h_k is the weight function, which is also known as the impulse response function of the system. The weight function shows how the input and output are related to each other. This term (impulse response) stems from the idea that this function reflects the response of the studied system to an impulse input of a unit size. For example, assuming that the input x_t has a zero value for all t with the exception that x_t at time zero has the value one, can be written as follows:

$$x_t = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases}.$$

Then the value of the output at time t is obtained by:

$$y_t = \sum h_k x_{t-k} = h_t.$$

For example, the following special moving average filter

$$y_t = \frac{x_{t-1} + x_t + x_{t+1}}{3}$$

represents a linear system with an impulse response equals to:

$$h_k = \begin{cases} 1/3 & k = -1, 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

The Step Response function is an another way to describe linear system in the time domain, which is defined in the discrete time as follows:

$$S_t = \sum_{k \leq t} h_k.$$

The term step response stems from the idea that this function represents the response of the system (process) to a unit step change in the input at time zero.

1.6 Digital Filters

A digital filter is a type of filtration that can be applied to discrete-time signals. There are two kinds of digital filters that depend on the impulse response of the system and they are known as a Finite Impulse Response (FIR) and an Infinite Impulse Response (IIR). The Impulse Response term can be attributed to the mechanism of obtaining a new value (Filtered Value), which is computed by using past, current, and sometimes even future values may also be involved in this process. Mathematically, this can be written as follows:

$$y_t = \sum_{k=-\infty}^{\infty} h_k x_{(t-k)} \tag{1.2}$$

where x and y represent the input and output series.

The Moving Average (MA) filter is regarded as one of the simplest FIR digital filters, for example,

$$y_t = \frac{1}{3}(x_{t+1} + x_t + x_{t-1}) \tag{1.3}$$

is a 3 terms moving average, which depends on a Feed-Forward difference equation. Feed-Forward here refers to the range of the input signals, which do not contain any values from the output series. An equation that contains input and output values, where the values of this output series are extracted by using values of the input series, and sometimes past values of the output series, is called a difference equation. Using this type of filtering, FIR filter, for example, moving average method, will produce a new series with a pattern that is smoother than the input time series. Smoothing in time series analysis refers to the process of removing the irregular signals to clearly investigate the desired components such as a trend.

1.7 Moving Average Filter (MA)

Most time series contain different patterns and it can be more useful to categorize these patterns. Linear Filtering is regarded as one of the most important manners for extracting the components embedded in time series [37]. In fact, although there are a number of relatively effective digital filters, the MAF is regarded as one of the most important and widely used filtration mechanisms [37, 90]. This is attributed to its efficiency and easy calculation procedures. The definition of this filter implies that it can be applied by using the average of a number of points from the original

signal to generate each new point in the output signal. By using the equation form this can be written as:

$$y_i = \frac{1}{m}\sum_{j=0}^{m-1} x[i+j] \tag{1.4}$$

where x and y are the input and output signals, respectively, and m represents the number of considered points. There are two ways to apply the moving average filter, depending on the points from the input signals. The first way is when the number of points for averaging are taken for one side. For example, in a 7 point moving average filter, if we want to obtain the moving average for the point 70 in the output signal, the next equation can be used:

$$y[70] = \frac{x[70] + x[71] + x[72] + x[73] + x[74] + x[75] + x[76]}{7} \tag{1.5}$$

The second way is carried out by using a symmetrical group of points that are located around the output point. For the same example above, this can be written as:

$$y[70] = \frac{x[67] + x[68] + x[69] + x[70] + x[71] + x[72] + x[73]}{7} \tag{1.6}$$

In this case, the limits of the summation for the equation 1.4 will change from $j = 0$ to $m - 1$ into $j = -(m-1)/2$ to $(m-1)/2$. The symmetrical averaging requires that the value of m must be an odd number. Furthermore, based on the MA method, Kolmogorov and Zurbenko presented a developed version of this filter which is known as the Kolmogorov and Zurbenko filter (KZ) filter [121, 36]. The KZ filtering mechanism belongs to the class of the low pass filters which enable all signals with low frequencies to pass through it and at the same time attenuate all the signals with high frequencies. What is computed in the first iteration of the KZ filter will be the input series for the second iteration and so on.

1.8 The Kolmogorov-Zurbenko (KZ) Filter

In signal processing system, there are many kinds of filters such as the MA, the Centered Moving Average (CMA), and the KZ filter. The $KZ_{m,p}$ filter, where m and p are the window width and the number of iterations, respectively, is regarded as one of the most robust techniques for the linear filtration [100, 122, 83, 118]. This filter is the MA filter but with repetitive times. The two parameters of this filter are decided by the number of days that are needed to be filtered out, m, and the number of iterations, p, which is most often chosen to be in the range between 3 and 5 times

[115], [36]. To filter out periods of length less than N days, we will use the following criterion , [36]:

$$m \times \sqrt{p} \leq N.$$

For example, if the $KZ_{15,5}$ filter is used, then these two parameters will remove cycles of 33 days and less. That means, this criterion is carried out to determine the point of the desired cut off frequency for the number of periods, for example, the number of days, months, etc. Lumley and Panofsky 1964 (as cited in [36]) showed that the Transfer Function is used to examine the behaviour of the linear time invariant discrete time filters. Using the convolution theorem [36], this function is defined as the following:

$$\psi(\omega) = H(\omega)\phi(\omega) \tag{1.7}$$

where ϕ and ψ are the spectral densities of the original and filtered data, H is the transform function of the filter, and ω denotes the chosen frequency. Wei Yang and Igor Zurbenko (2010), [118], have stated that the energy transfer function of the KZ filter can be measured using the following criterion :

$$|\psi_{m,k}(\omega)|^2 = \left(\frac{1}{m}\frac{sin(m\omega/2)}{sin(\omega/2)}\right)^2$$

where ω is the chosen frequency. Flaum et al. (2012) have mentioned that the parameters of the KZ filter are researcher-specified [38]. Since the KZ filter belongs to the class of the low pass filtering techniques, its mechanism has been applied extensively to filter out the undesirable signals, as examples of its use, specifically in the environmental studies, [82], [81], [122], [73], [115], [100], [36], and [50].

As an example of how this filter separates the components, Figure 6.3 shows the data of the raw and the three components for the log water discharge for Utica city in New York state in the USA. When we examine the raw data in this figure (top left), we cannot extract a precise opinion about the long-term trend because of the noise (high frequency signals). But when we decompose the data using the KZ filter, a more clear insight for the trend can be obtained (top right). Let $X(t)$ be a vector of a real-valued time series, the KZ filter for the data of this series is the moving average filter with window size m, but this window size is repeated p times for each new resulting output. In other words, the output of each iteration will be dealt as an input series in the next iteration. The KZ filtering technique for the first iteration can be written as follows:

$$KZ_{m,k=1}[X(t)] = \frac{1}{m}\sum_{s=\frac{-(m-1)}{2}}^{\frac{(m-1)}{2}} X(t+s) \tag{1.8}$$

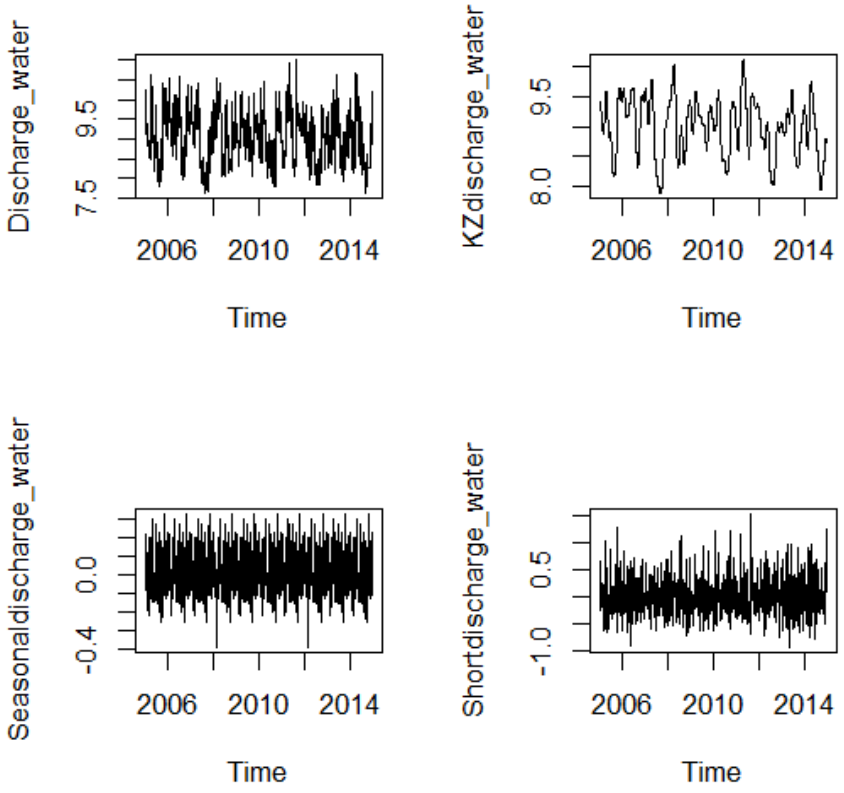

Figure 1.1: The Raw and the Three Components for the Water Discharge for Utica city Using the KZ Filter

where this is the first iteration with m window width. After that, the series obtained from the previous equation, Equation 1.8, will be used as an input for the second iteration as the following:

$$KZ_{m,k=2}[X(t)] = \sum_{s=\frac{-(m-1)}{2}}^{\frac{m-1}{2}} \frac{1}{m} KZ_{m,k=1}[X(t+s)] \tag{1.9}$$

and so on for the other iterations. As aforementioned the criterion $m\sqrt{p}$ can be used to determine the point of the desired cut off frequency for the number of periods, for example, the number of days, months, etc. The KZ filter is classified as one of the

low pass filters and the cut off frequency for it can be computed by using

$$2\sqrt{6} \times \sqrt{\frac{1-\alpha^{\frac{1}{2p}}}{m^2-\alpha^{\frac{1}{2p}}}}$$

where $\alpha \in (0,1)$ is a pre-specified value [118]. Since the KZ filtering technique is originally just a repetitive moving average method, it can deal with the problem of missing data. For this reason and also for the relatively easy calculation process, compared with the other types of the filtering techniques, the KZ filtration mechanism has been extensively used in different fields, especially in the environmental studies.

1.9 Residuals Analysis

To check the validity of the formalised regression model, we need to test whether the assumptions of the regression model are satisfied or not. The residual terms of the regression model can be exploited to accomplish this task. The validity of the constructed model means that the model has satisfactorily fitted the data. The residual is the difference between the observed and predicted values for the response variable. This can be written as:

$$e_t = y_t - \hat{y}_t$$

where the estimated value for the response variable y, which is $\hat{y}$, is computed by using different methods, such as the ordinary least square and maximum likelihood methods. The simple linear regression model, which is defined as the following:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.10}$$

can be easily extended to multiple linear regression analysis by adding some other predictors to this equation. From the linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, we extract that ϵ can be computed from $y - (\beta_0 + \beta_1 x)$. Using a sample of observations, the regression model can be written as $\hat{y} = b_0 + b_1 x + e_t$. Solving this equation provides $\hat{y}_t = b_0 + b_1 x_t$; this is the point estimate for $\beta_0 + \beta_1 x_t$. Consequently, the residual e_t represents the point estimate for the error ϵ_t.

Furthermore, to be validated, some assumptions have to be satisfied in the residuals of the regression model:

1. The error terms have to be normally distributed with mean zero and variance σ^2.

2. The error terms have to be uncorrelated (independent) to avoid the autocorrelation problem between them.

1.10 Plots of the Residuals

Depicting the residual series versus different variables can be considered as one of the most important ways to do checking process [15]. Almost every statistical package provides the possibility of plotting the residuals. Therefore, once the residuals are calculated, plotting them will be possible against:

1. Data of Xs.
2. Data of $\hat{y}$ (Fitted Values).
3. Time (if the data is time series data).

There are a number of assumptions that have to be achieved for the residuals:

- Assumption of the Constant Variance: One of the most significant advantages of the residuals plot is to examine the variance of these residuals. The values of x axis are the fitted values, $\hat{y}$ and the values of y axis are the residuals. There are three different scenarios for the variance. Firstly, the assumption of the constant variance for the residuals is violated if the residuals fluctuate around zero with a pattern that can imply that these residuals are increasingly spreading out as the x axis increases. Secondly, this assumption is also violated when the fluctuations of the residuals become narrower as the x axis increases. Finally, a constant variance appears when the residuals construct an approximately horizontal band, where the variance will remain more stable even when x axis increases.

- Assumption of the Appropriate Function Form: Typically, when the existing model does not correctly fit the data, the residuals plot can be considered as a helpful tool to detect this inappropriate model. Besides, the correct model may be determined by an alternative pattern that can be seen in the same plot. For instance, if we use a linear regression model to fit the data, and the residual plot displays a curve, this may be an indicator to the necessity of rebuilding the data by using the true relationship which is a non linear relation.

- Assumption of Normality: If the histogram and stem and leaf plots for the residuals display a plausible bell shape and symmetry about zero, that means the normal assumption is achieved for the residuals.

- Assumption of the Independency: Violation of the assumption of the independency of the residual terms is more likely to happen when the regression analysis is performed for time series data than cross sectional data. In this case, the residuals of this analysis can be serially correlated. These correlated values

lead to the problem of the autocorrelation. The patterns of the autocorrelated residuals construct two different types of autocorrelations.

1. Firstly, the time-ordered residual terms may have a positive autocorrelation. This positive autocorrelation can occur when a positive error terms in the time period t produces another positive error terms in the later time period $(t + k)$. Or, a negative error terms in the time period t is followed by another negative error terms in the later time period $(t+k)$. A positive autocorrelation generates a cyclical pattern over time.
2. Secondly, when a positive error terms in period t is followed by a negative error terms in time period $(t + k)$ and vice versa, that means a negative correlation exists, which generates an alternating pattern over time.

If regression analysis is performed for time series data and the residuals, which are depicted versus time, display a cyclical behaviour, this could imply existing of a positive autocorrelation between the residuals. The independence pattern, which provides no tendency to either a positive or a negative shape, is required to hold in the residuals of a regression model. In other words, the pattern of these residuals over time has to be randomly distributed. If this is the observed case, that means the regression analysis has successfully taken into account all the available information by using the fitted model and no other information can enrich the constructed model.

Moreover, the First-Order Autocorrelation is a type of either positive or negative pattern for the correlated residuals. This kind of autocorrelation exists when the error terms, ϵ_t, in the time period t is serially related to the error terms in the time period $(t - 1)$, ϵ_{t-1}, by the formula:

$$\epsilon_t = \phi_1 \epsilon_{t-1} + a_t$$

where ϕ_1 is the parameter of the correlation between the error terms that are separated by one time period and $a_1, a_2, \ldots$, are the values of the randomly and independently distributed error terms for this model with mean zero, $(\mu = 0)$, and a constant variance (σ^2).

1.11 Box-Jenkins Models for the Error Terms of the Regression Analysis in the Time Series

A time series regression model is often performed to predict values for a serially recorded data which has a deterministic nature. These models are suitable provided

that the parameters of the model do not change over time. In addition, the residual terms for these regression models are required to be randomly distributed with mean zero and constant variance for each time series period. Besides, for the residuals, the assumption of being statistically independent has to be available.

However, when a linear regression is employed to analyse the data, the situation that is most likely to happen is that this assumption is not achieved. Therefore, when we encounter such a situation, one of the Box-Jenkins models, which are also known as ARMA models, will be used to model the error terms. Once we select a model, we combine it with our regression model that we have already constructed, to formalise the final expression which will be used to forecast the future values. The mechanism for determining an appropriate model for the residuals series depends on the functions in Section 1.12.

1.12 Autocorrelation Functions

Examining the pattern of the Sample Autocorrelation Function (SACF) or Sample Partial Autocorrelation Function (SPACF) is a helpful procedure to decide which model of the Box-Jenkins will fit the residuals of the time series regression model adequately. The sample autocorrelation function at the lag k, r_k, for the values of a time series can be defined as follows:

$$r_k = \frac{\sum_{t=b}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=b}^{n}(z_t - \bar{z})^2} \tag{1.11}$$

where $\bar{z} = \frac{\sum_{t=b}^{n} z_t}{(n-b+1)}$, n is the number of the observed values for the studied time series, $z_b, z_{b+1}, \ldots, z_n$ are the values of the working time series. This statistic measures the linear relationship between the observations of a time series separated by a lag k (time units). The quantity r_k is always between -1 and 1, and whenever this quantity is near to 1, a strong linear relationship with a positive slope exists. Also, when the value is near to -1, this implies that the relationship between the values of the series is also strong but with a negative slope.

For a nonseasonal time series, the SACF can show a number of different patterns. The first type is the cutting-off pattern. This cutting off appears when r_k has a large value which appears as a spike when we plot the SACF. Theoretically, this leads to reject the null hypothesis which claims that the theoretical autocorrelation at lag k equals to zero. The symbol ρ_k is the theoretical autocorrelation at lag k. Therefore, if there is no spike after the lag k, we conclude that the SACF cuts off at lag k. The mathematical equation that can be used to detect the spikes in the SACF, is written as follows:

$$t_{r_k} = \frac{r_k}{S_{r_k}} \tag{1.12}$$

where S_{r_k} denotes the standard error for r_k. Whenever the absolute value for this quantity for the nonseasonal data is greater than 2, a spike exists in the SACF. On the other hand, the contrary situation for the cutting-off pattern is the dying down, which means that the pattern decreases in a fairly stable fashion. Generally, SACF has three principle ways to die-down:

- Exponentially damped pattern (with or without oscillation).
- A damped sine signal pattern.
- A combination of the two above mentioned patterns.

In addition, these dying-down patterns for the SACF tend to display two fashions, where the SACF either dies-down fairly quickly or extremely slowly.

1.13 The First-Order Autocorrelation

As long as the error terms for a time series regression model possess an autocorrelation pattern, the constructed model is inappropriate to fit the data. In such a case, fitting a model for these autocorrelated errors would be the best solution as the inclusion of this model will improve the forecasting performance of the constructed model and eliminate the autocorrelations between the error (residual) terms. Otherwise, wider prediction intervals will be obtained, and this is not desirable in the prediction process. However, when these autocorrelated values are handled, the prediction intervals will be narrower [15].

One of the most frequently encountered structures is the first-order autoregressive structure. This name is attributable to the nature of the relationship that relates the error terms in the period t, ϵ_t, with the other error terms in the previous period $t-1$, ϵ_{t-1}, as we have already mentioned to this relation in the previous expression, which is $\epsilon_t = \phi_1 \epsilon_{t-1} + a_t$.

1.14 Forecasting

Although there is a wide range of different forecasting techniques for predicting future time series data, no single mechanism is universally applicable [20]. The selection of any method depends on conditions or assumptions related to the given series. That means, for forecasting some values for the desired period, satisfying these assumptions

will be required to obtain an accurate forecasts. Furthermore, bearing in mind that any system could change over time, these assumptions should be modified so that any variations can be incorporated in the forecasting model, especially for the long-term component forecasting case. It is important to consider that the forecasting procedure is an extrapolation process, with all risks that will be accompanied [20].

In general, the forecasting methods are classified into two types Qualitative and Quantitative methods. We summarize them

- Qualitative Methods: Also known as Subjective Methods, forecasting by using one of these methods is largely dependent on an expert's opinion. We often need to use these techniques when there is no historical data, or it is scarce and not sufficient to give a precise forecast about the studied situation. For example, when the situation is related to a new product, it will be necessary to use this type of forecasting to enable the analyst, expert, to give his opinion. This expert could be one of the members from a sales force, or from a market research committee.

  Utilizing this kind of forecasting technique appears when we need to predict future values for a historical data when its pattern varies through time. The process of detecting this variation in the fashion of the desired data will be implemented by using one of the subjective mechanisms. These qualitative methods involve subjective curve fitting, Delphi Method, and Technological Comparisons [15].

- Quantitative Methods: These techniques are generally classified into: Univariate and Multivariate methods. The first group is specified when the studied system consists of a single time series without consideration for any other variables that could affect it. This univariate forecasting model requires historical data, past values for a specific period, to predict future events for this series. This should be implemented by analysing this historical data with one of the associated models, for example, autoregressive model of order one. Typically, this type is used if all the assumptions of the studied system are assumed to remain the same. But if these assumptions are vulnerable to change in the next periods, it will not be worth using this type of model to forecast future values.

  For the second group, which includes Multivariate Models (Causal Models), the idea is to incorporate any effect from other time series in the constructed model. After we determine those influential (independent) variables, a model that combines them with the dependent series would be built. Then, this constructed model will be exploited to predict future data depending on these related predictors. Using this kind of forecasting method is common, especially

in the business studies and the environmental area, as this type can reveal any relationships between the input and output series.

Moreover, in the forecasting process, we are able to produce either a point forecast, which gives a specific number, or a range of numbers with a confidence interval, which gives a possibility of including an error terms in the process of calculating the forecasting values. Extra attention is required to be paid to the necessity that the forecasting technique should be consistent with the pattern of data for each studied time series. In general, investigating these errors can help us to decide whether the forecasting technique is suitable to describe the pattern of the data or not.

When the forecasting method has suitably matched the pattern of the given data and produced a precise prediction for trend, seasonal, or cyclical component, these error terms should represent only the irregular (random) component. If this is not the situation, that means the forecasting technique has not taken into account all the existing information in this specific time series. For example, if these error terms construct an upward trend, this may suggest that there is an upward trend which has not been exploited in the present forecasting model. However, when the prediction expression has sufficiently accounted for all the embedded information, we need to measure the magnitude of these errors to check the accuracy of the forecasting model. This can be implemented by computing the Mean Squared Error (MSE), which can be calculated as:

$$\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}$$

where y_t and $\hat{y}_t$ are the actual and predicted values, respectively, and n is the sample size.

In the hydrological field, and as a consequence of the sophisticated nature of the data, forecasting remains a difficult and a critical task [29]. The data of this field is often non-linear, non-stationary, and also contains different scale characteristics. In fact, in order to increase and improve the accuracy of the prediction for the hydrological data, several novel models have been proposed to tackle and proceed with all the mentioned difficulties.

1.15 Frequency Domain

The time and frequency domains are used to analyse time series data The methods of the Time Domain are based on the principle of how the values of the series change over time. On the other hand, the methods of the Frequency Domain are based on the principle of how much of this series is located within each band of a previously determined bandwidth of frequencies. Furthermore, whenever we have one of these

domains, time or frequency, the process of converting one of them to be represented by the other can be carried out by using one of the transformation functions.

The Fourier series can be used to transfer a periodic function represented in time domain to another series represented using a number of frequencies. However, if the function of interest is aperiodic, Fourier Transformation (FT) can be utilised to transfer this non-periodic function to be represented using frequency rather than time. The FT is regarded as one of the most prominent functions that can be used to transfer a time signal to another signal calculated by using the frequency.

Also, to recover the time signal from the frequency domain, the Inverse Fourier Transformation (IFT) is typically applied. The Fourier Transformation describes the time signal as a sum of sinusoids functions, sine and cosine terms, using different frequencies [51, 59, 114]. So, if we have a signal that is recorded based on a range of discrete time, the Discrete Fourier Transformation (DFT) converts this signal into a new one, but with a range of discrete frequencies. The Fourier Transformation for discrete time functions would be also discrete function, but it is measured using the frequency. The DFT expresses each value in a time series by using complex sinusoids. A sequence of N numbers $x_0, x_1, \ldots, x_{N-1}$ is transformed into a new sequence includes also N numbers.

$$X_f = \sum_{t=0}^{N-1} X_t \cdot e^{-2\pi i f t/N}, \quad f \quad \text{is an integer} \tag{1.13}$$

where f is the frequency and $i = \sqrt{-1}$.

Typically, the domain for f lies within the range $[0, N-1]$, and this can be clearly seen when the DFT is computed by the Fast Fourier Transform (FFT). Furthermore, there are other domains for f such as $[\frac{-N}{2}, \frac{N}{2} - 1]$ if N is even; and if N is odd, the domain is $[\frac{-(N-1)}{2}, \frac{(N-1)}{2}]$ [93]. Moreover, to recover the time domain signal, the following equation can be used:

$$X_t = \frac{1}{N} \sum_{f=0}^{N-1} X_f \cdot e^{2\pi i f t/N}. \tag{1.14}$$

Let X_t represent a time series that contains a periodic sinusoidal component with a known period (wavelength). The natural model for this series can be written as follows:

$$X_t = A cos(\omega t + \phi) + Z_t \tag{1.15}$$

where ω is the Frequency for the variation for this sinusoidal function, A is the Amplitude for the variation, ϕ represents the Phase and Z_t refers to a Stationary Random Series. The angle $(\omega t + \phi)$ is typically measured by using the unit Radians, where 2π radians equals to 360 degree. Also, because ω is the number of radians per

unit time, it is called the Angular Frequency. The frequency f, which represents the number of cycles per unit time, equals to $\frac{\omega}{2\pi}$. The Period, also known as a wavelength, is the reciprocal of the frequency, it equal to $\frac{1}{f}$ or $\frac{2\pi}{\omega}$. Equation 1.15 is the simple form to describe the situation when the time series of interest is represented by one frequency [58, 114, 93]. In practice, however, most variations in the time series are associated with more than one frequency, so, in this case Equation 1.15 can be rewritten in the following form:

$$X_t = \sum_{j=1}^{k} A_j \cos(\omega_j t + \phi_j) + Z_t \tag{1.16}$$

where A_j, ω_j, and ϕ_j denote the amplitude, angular frequency, and phase, respectively, for the frequency j. Since the $cos(\omega t + \phi)$ equals to $\cos \omega t \cos \phi - \sin \omega t \sin \phi$, Equation 1.16, can alternatively be written by using the sine and cosine terms:

$$X_t = \sum_{j=1}^{k} (a_j \cos \omega_j t + b_j \sin \omega_j t) + Z_t \tag{1.17}$$

where

$$a_j = A_j \cos \phi_j \quad and \quad b_j = -A_j sin\phi_j.$$

So, by using either parametric or non-parametric methods for a time series with a finite set of measurements, the spectral content of this series can be determined.

The spectral analysis is one of the most widely used methods for analysing time series data in a wide range of fields, such as Oceanography, Geophysics, Astronomy, Economic, Marine Science, and Meteorology. This analysis characterizes the frequency content of a signal. The spectral content describes the distribution of the power of a signal through frequency. Also, the spectral analysis has an ability to discover any hidden periodicities (cyclical behaviour) in the data. Essentially, the non-parametric methods rely on the idea of partitioning the available data to be limited to a band of frequencies. On the other hand, the parametric analysis methods attempt to construct models that include some parameters which need to be estimated by one of the estimation methods.

1.15.1 Non-Parametric Methods

In time series analysis, there are many non-parametric spectral estimators that can be used to find the spectral content for the desired series, but the most common techniques are the Periodogram and Correlogram. Although these two estimators provide sensible high resolution for a signal, the results obtained by applying them are somewhat poor because of the high variance. This variance does not decrease

even when the size of data increases. The existence of high variance in the results of these two non-parametric methods encourages researchers to suggest some techniques to avoid this problem. However, this can lead to a reduced degree of resolution [93].

Periodogram and Correlogram Techniques

Periodogram Periodogram is a tool that is applied to describe and identify the dominant cycles in a time series. This tool is used to detect the periodicity or seasonality in a time series data [93]. Fourier analysis is used to rebuild the deterministic function by using a combination of sinusoid, sine and cosine (Trigonometric) waves. So, to examine whether or not a time series exhibits periodicity, plotting the periodogram or spectral density function against the period or frequency can be carried out. The following function can be used to compute the periodogram:

$$\phi_p(w) = \frac{1}{N}\sum_{t=1}^{N}(y_t e^{-i\omega t})^2 \tag{1.18}$$

where y_t, $t = 0, 1, 2, \ldots, N$, is a discrete time series whose values are a sequence of random variables that have mean equals to zero

$$Ey_t = 0. \tag{1.19}$$

Therefore, one of the major advantages of the periodogram is its ability of revealing any hidden periodicities that may be contained in the studied time series.

Correlogram The correlogram spectral estimators depend on the correlation coefficient [93] and can be written as follows:

$$\phi_c(w) = \sum_{k=-(N-1)}^{N-1} r_k^{\omega} e^{-iwk} \tag{1.20}$$

where r_k^{ω} is the estimated correlation at lag k. There are two ways to calculate the r_k^{ω}, which are:

$$r_k^{\omega} = \frac{1}{N-k}\sum_{t=k+1}^{N} y_t y_{t-k}^* \quad 0 \le k \le N-1 \tag{1.21}$$

and

$$r_k^{\omega} = \frac{1}{N}\sum_{t=k+1}^{N} y_t y_{t-k}^* \quad 0 \le k \le N-1. \tag{1.22}$$

1.16 Data

In this study, daily data is collected for Water Discharge (m^3/s), Temperature (F), Wind Speed (m/sec), Precipitation (mm/hr), Absolute Humidity (gram per cubic metre), Dew Point (F), Sea Level Pressure (millibar mb), Visibility Miles (m), and Cloud Cover ($okta$). Daily data for Groundwater Level and Tide are also chosen to be involved as predictors in this study. All these time series are collected for three cities Cohoes, Utica, and Poughkeepsie in New York State, US. The source for this data is the New York Department for Environmental Conservation for the period 2005 to 2014. This dataset is separated into two parts; the data for the period $2005 - 2013$ is used to construct the models, and the data for the year 2014 is utilised to validate the constructed models.

The Logarithm transformation is applied to the water discharge series to stabilize the variance of the values. Furthermore, because we have variables with different unit scales, we use the standardised data in the analysis instead of the original time series datasets. The standardisation procedure is implemented by subtracting each value for each variable from its mean, and then dividing by the standard deviation of the associated series.

Chapter 2

Regression and Vector Autoregressive Models For Forecasting Water Discharge

Following the Rao and Zurbenko (1994) method [36, 121, 100, 118], it is assumed that a time series for a variable can be separated as:

$$Y_t = LT_t + SE_t + SH_t$$

where Y is the original time series, LT is the long-term signal (wave), SE is the seasonal cycle, SH is the short-term (synoptic) component, and t is the time. This model is based on the assumption that there is a gap (difference) in the spectra of each component. The events that last less than 3 weeks represent the short-term component. The next scale is the seasonal variations which include any season-based event that repeats itself in a period of one year or less. Finally, any scales of periods of more than one year are related to the long-term component. This representation for a time series has been used extensively in the meteorological and environmental studies [100, 101].

A model is fit to the data of each component and the regression model is the most commonly used model to accomplish this task. The final combined model is constructed by combining the extracted components. However, the residual terms for the regression models for the three components and the final combined model often suffer from the autocorrelation problem. The R Squared value, the confidence interval, and the accuracy of prediction are affected by this issue. In this chapter we solve this problem by modelling the residual terms for the regression models of the components and the final combined regression model. Specifying one of the Autoregressive Moving Average (ARMA) models is the most common choice to fit a model for the residual terms. Therefore, the new contribution in this chapter is to fit an ARMA model to the

residual terms of the constructed models. Adopting this methodology has improved the prediction process according to the results of the model selection methods the Akaike Information Criterion, AIC, and the Schwarz Bayesian Criterion, SBC.

In this chapter, we have analysed the data of three cities which are Cohoes, Utica, and Poughkeepsie, following the same methodology. However, the new contribution in this chapter has been considered for the data of Utica city. The methodology can be summarised as follows:

- Building a regression model for the raw and the three components data, the long, seasonal, and the short-term component, for each city.
- Because the structure of a Vector Autoregressive (VAR) model is based on lag variables, this model will be a good choice to describe the data of the short-term component as it essentially represents the high frequency signals in a series. Based on this, the short-term component data has been analysed using two approaches, the regression and the VAR models.
- The residual terms for the regression models for Utica's city data have been expressed using an ARMA model.

In addition to the purpose of fitting a forecasting model for each city, some results from these analyses will be used later in the next chapters. For example, the results of the regression models for the three cities have been used in chapter 5 and also the results of the regression model for Cohoes city have been used in chapter 4.

The remainder of this chapter has been organised as follows. Section 2.1 provides information about the simple and multiple linear regression analysis. Sections 2.2 and 2.3 present the most common methods to analyse time series data, which are the Autoregressive (AR) and Moving Average (MA) processes. Besides, Section 2.4 introduces the Autoregressive Moving Average (ARMA) models. Section 2.5 gives a brief description of how to specify an ARMA model to the errors of a regression model. Sections 2.6, 2.7, and 2.8 provide applications for the data of Cohoes, Utica, and Poughkeepsie cities, respectively. Section 2.9 present a discussion for the results obtained. Section 2.10 presents the conclusion of this chapter.

Table 2.1 shows the methods applied for the three cities.

2.1 Regression Analysis

Regression analysis is one of the most common methodologies in statistics to predict values for a response (dependent) variable based on some predictors (independent) variables [61]. In addition to the prediction task, the regression analysis has been widely used to assess the influences of predictors on the response variable. Relying on

Table 2.1: Summary of the Applied Methods for the Studied Cities.

City	MLR	VAR	MLR with AR(1)
Cohoes	Raw and all Components	Short-Term Component	
Utica	Raw and all Components	Short-Term Component	MLR for the three components and AR(1) for the Residual terms
Poughkeepsie	Raw and all Components	Short-Term Component	

the number of independent variables, the regression analysis can be divided into two types, simple and multiple linear regression models (MLR). The former model requires two parameters to be employed, which are the intercept and the slope parameters. These parameters refer to the mean value of Y when X equals to zero and the amount of the change in the mean value of Y when the predictor X increases or decreases by one unit, respectively. In addition to the intercept term, the latter model needs more than one slope which are related to the independent variables. Let $X_1, X_2, \ldots, X_k$ be k independent variables which are associated with the dependent variable Y by the following regression model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} + \epsilon_t \tag{2.1}$$

where β_0, β_1, ..., β_k are the parameters of the model, and ϵ is the error term, which is assumed to possess the following properties:

- $E(\epsilon_t) = 0$.
- $Var(\epsilon_t) = \sigma^2$.
- $Cov(\epsilon_{t1}, \epsilon_{t2}) = 0$ for $t_1 \neq t_2$.

Generally, Y consists of a mean which relies on the X_i's, and a random error, ϵ, which represents the other possible variables which have not been explicitly accounted for in the constructed model. The recorded values for the independent variables are often regarded as fixed values. Furthermore, since the error term (and as a consequence the response variable) is considered as a random variable, so its behaviour

will be controlled by a collection of assumptions. When there are n independent observations of Y and the associated values of x_i, the full model will be as follows:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_k X_{1k} + \epsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_k X_{2k} + \epsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_k X_{nk} + \epsilon_n.$$

Sometimes the order of these independent variables can be changed to be first, second, and so on [61]. In matrix notation, the multivariate regression model can be written as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$Y = X\beta + \epsilon.$$

And the properties of the error term become

- $E(\epsilon) = 0$.
- $Cov(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I$.

2.2 Autoregressive Model

An Autoregressive (AR) model is a model that predicts the changes in the variable of interest based on two factors: the average c and the preceding (past) values y_{t-1}, y_{t-2}, …, y_{t-p}. This can be written mathematically as

$$\widehat{y_t} = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model and p is the order of the AR model. This model is commonly written as AR(p). The pattern of the sample partial autocorrelation function (SPACF) is typically used to decide which order has to be chosen for the AR model [15].

2.3 Moving Average Model

The Moving Average Model (MA) is constructed using the forecast errors (random shocks). Mathematically this can be written as follows:

$$y_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$

where c is the average of the series, $a_t, a_{t-1}, \ldots, a_{t-q}$ refer to the current and past random shocks and $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model, q is the order of the MA process. This model is commonly written as MA(q). The sample autocorrelation function (SACF) is responsible for determining the order that should be used for a MA model.

2.4 Autoregressive Moving Average Model

The result of combining the two models above is an Autoregressive Moving Average Model (ARMA) of order (p, q). The strategy of constructing this model is composed of three different stages. These stages are:

- Identification of a proper model for the considered data based on the SACF and SPACF.
- Estimation of the parameters of the constructed model using one of the estimation methods and diagnostic the adequacy of this model using the residual analysis.
- Forecasting future values for the response series.

Examining the patterns of the SACF and SPACF is the first step in the process of constructing an ARMA model. To make sure that the proposed model fits the data, testing the autocorrelation for the residual terms has to be performed. Therefore, if the autocorrelations for some lags are statistically significant, this would suggest that the fitted model is not adequate to capture all the information embedded in the studied series. As a consequence, it will be required to model the time series by using a more complex model [15, 105].

2.5 Linear Regression Model with ARMA Errors

Performing a regression analysis using time series variables often produces errors (residual terms) that have a time series structure. This, in turn, violates the required assumption of independent errors, which is shown in Section 2.1. Consequently, if the

autocorrelation problem has not been considered, this will influence the accuracy of the estimates of the regression coefficients and their standard errors. This problem can be avoided by providing a formula to express these errors. The formulation is built by specifying an ARMA model for the residuals. Mathematically, this can be constructed by adding an ARMA model (which is chosen based on the behaviour of the SACF and SPACF) to Equation 2.1 [116]. The structure for the residual terms can be expressed using one of the AR models. Therefore, the new model will be:

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + e_t \tag{2.2}$$

with

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + a_t$$

where $a_t \sim N(0, \sigma^2)$.
If we use the backshift operator $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, the model above can be rewritten as

$$\Phi(B) e_t = a_t.$$

Then, taking the inverse operator, $\Phi^{-1}(B)$, the model can be written as follows:

$$e_t = \Phi^{-1}(B) a_t.$$

Therefore, the final model can be written:

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \Phi^{-1}(B) a_t \tag{2.3}$$

where a_t is a white noise series.

This method of specifying an ARMA model for the errors of a MLR model has been not been applied for a combined MLR model. In this case, the new combined regression model with an ARMA errors model, which can be called CMLR+Noise, can be written as

$$y_t = C + Lt_{x1t} + SE_{x1t} + SH_{x1t} + Lt_{x2t} + SE_{x2t} + SH_{x2t} + \ldots + Lt_{xkt} + SE_{xkt} + SH_{xkt} + \Phi^{-1}(B) a_t \tag{2.4}$$

where Lt_t, SE_t, and SH_t are the variables of long, seasonal, and short-term components.

2.6 The Application for Cohoes' City Data

Figure A.1 in the Appendix illustrates the steps taken to build the developed models for Cohoes city.

2.6.1 The Analysis for Cohoes' City Raw Data

At first, the correlation matrix is calculated as shown in Table 2.2 to check the relationships between the water discharge (WD) and the meteorological variables Temperature (TE), Wind Speed (WS), Precipitation (PR), Groundwater level (GW), and tide (TD) series. The null hypothesis H0: $r = 0$ is applied against the alternative hypothesis H1: $r \neq 0$, where r is the correlation coefficient. Since the correlation coefficients with the water discharge series have P-values 0.84 and 0.61, which are greater than the significance level 0.05, the variables Sea Level Pressure and Visibility Miles are ignored. Based on the output of the correlation matrix, the variable Dew Point has been eliminated because of the high correlation coefficient with the Temperature variable. Also, depending on the results of the multiple linear regression analysis, the variables Absolute Humidity and Cloud Cover are also removed from the regression equation as their coefficients have P-values greater than the significance level. Therefore, the chosen variables that will be used to complete our analysis are temperature, wind speed, precipitation, tide, and groundwater level.

Table 2.2: The Correlation Matrix for the Raw Time Series of Cohoes City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.327	0.133	0.247	0.384	-0.567
TE	-0.327	1	-0.161	0.054	-0.056	0.079
WS	0.133	-0.161	1	0.064	0.155	-0.079
PR	0.247	0.054	0.064	1	-0.026	-0.011
TD	0.384	-0.056	0.155	-0.026	1	-0.495
GW	-0.567	0.079	-0.079	-0.011	-0.495	1

Additionally, as we have a number of predictors, it will be reasonable to build a number of nested regression models and choose the best fitting model based on the R Squared value. The nested regression models are shown in Table 2.3. Based on its diagnostic statistics, which are shown in the same table, we chose the first model which is constructed by 5 variables as the best regression model for the raw data. This model is shown in Equation 2.5.

$$\widehat{WD}_t = -0.27TE_t + 0.01WS + 0.25PR_t + 0.15TD_t - 0.48GW_t \tag{2.5}$$

where WD, TE, WS, PR, TD, and GW denote the Water Discharge, Temperature, Wind Speed, Precipitation, Tide, and Groundwater Level, respectively. The order of the absolute values of the coefficients from largest to smallest is GW, TE, PR, TD, and WS. The R Squared value for this model is 0.48.

Table 2.3: Applying Different Diagnostic Statistics to select the best MLR model.

Number	Adjusted R	R-Square	AIC	BIC	MSE	Variables
5	0.483	0.484	-2164.578	-2162.556	0.516	TE WS PR TD GW
4	0.469	0.469	-2075.705	-2073.961	0.530	TE WS PR GW
3	0.468	0.469	-2073.396	-2071.611	0.531	TE PR GW
4	0.415	0.416	-1762.559	-1761.726	0.584	TE WS TD GW
3	0.415	0.415	-1759.355	-1758.305	0.584	TE TD GW
3	0.404	0.404	-1699.077	-1698.164	0.595	TE WS GW
2	0.402	0.402	-1689.853	-1688.696	0.597	TE GW
4	0.39	0.400	-1671.038	-1670.46	0.600	WS PR TD GW
3	0.396	0.397	-1655.518	-1654.702	0.603	PR TD GW
3	0.385	0.386	-1596.08	-1595.396	0.614	WS PR GW
2	0.380	0.380	-1569.237	-1568.28	0.619	PR GW
3	0.341	0.342	-1369.152	-1368.96	0.658	WS TD GW
2	0.336	0.336	-1344.298	-1343.720	0.663	TD GW
2	0.33	0.330	-1313.189	-1312.662	0.670	WS GW
1	0.322	0.322	-1276.774	-1275.836	0.677	GW
3	0.316	0.317	-1246.359	-1246.423	0.683	TE PR TD
4	0.316	0.317	-1244.558	-1245.108	0.6837	TE WS PR TD
3	0.241	0.242	-906.057	-906.800	0.7581	TE WS TD
2	0.241	0.241	-904.671	-904.784	0.758	TE TD
3	0.216	0.217	-799.787	-800.733	0.783	WS PR TD
2	0.213	0.214	-788.077	-788.364	0.786	PR TD
3	0.181	0.181	-652.591	-653.80	0.818	TE WS PR
2	0.177	0.177	-638.951	-639.454	0.822	TE PR
2	0.153	0.153	-542.962	-543.601	0.8469	WS TD
1	0.147	0.147	-523.254	-523.137	0.852	TE
2	0.113	0.114	-393.126	-393.973	0.886	TE WS
1	0.107	0.107	-370.289	-370.328	0.892	TE
2	0.074	0.075	-251.759	-252.795	0.925	WS PR
1	0.060	0.061	-204.613	-204.815	0.939	PR
1	0.017	0.017	-56.271	-56.61	0.982	WS

2.6.2 The Periods for the Studied Variables for Cohoes' City

Before using any decomposition technique we need to examine the spectral aspects for each series to know how these variables behave (change) along the studied period based on the frequency domain. The periodic behaviour with a regular nature for any time series can be expressed as a combination of sine and/or cosine signals. Typically, the main goal of using a combination of signals of cosine, sine, or both of them is to identify the dominant frequencies (or periods) for the time series of interest. In spectral analysis, which represents the part of time series analysis that uses the frequency to examine the data of the series, the periodogram is considered as one of the most important statistical tools to implement the task of detecting the periodicity in a time series.

There are several methods to calculate the periodogram, which are often based on the Fourier Analysis. One of these methods is the Kolmogorov-Zurbenko Periodogram (KZP). Essentially, the KZP depends on the Fourier Transformation (FT). Working with a smoothed periodogram series is often preferable as it decreases the degree of variance. To obtain this smoothed series, different methods have been proposed to smooth the resultant periodogram. The DiRienzo and Zurbenko smoothing algorithm (DZ) is applied to smooth the periodogram for all the considered series [117]. Table 2.4 shows the periods for all the variables in our study. The spectral analysis for all

Table 2.4: Periods (days) for the Studied Variables by Using the DZ method for Cohoes City.

Variable	First Peak	Second Peak	Third Peak
Temperature	365	331	405
Precipitation	13	165	173
Groundwater	365	912	1825
Water Discharge	365	182	912
Tide	365	182	

the variables, except the precipitation, reveals that the dominant period is the period that consists of 365 days. In contrast, the main peak for the precipitation is at the period of 13 days.

2.6.3 The Decomposition of Cohoes' City Time Series

In order to precisely determine the trend for the desired series data, the seasonal variations and the random fluctuations have to be removed from the data. This act of removing can be accomplished using one of the decomposition techniques. Fundamentally, there are two common models to describe a decomposed time series,

which are the additive and the multiplicative models. The operation of separating any time series into different components that represent the various scales embedded in the time series values, has no theoretical basis [15]. That means, these two decomposition models are intuitively built. On the other hand, the criteria to select which model can be used to construct the time series depends on the behaviour of the seasonal variation for the series.

If the seasonality increasingly or decreasingly changes, the Multiplicative Model can be chosen to formulate the components of time series. Alternatively, if the changes of the seasonal component are characterized as a constant variation, this will lead to employ the Additive Model for the selected data. The decomposition of time series into the components, long, seasonal and short-term component, can be regarded as one of the most effective mechanisms to model time series data [121, 36].

In our previous regression model for the raw data, Equation 2.5, the R squared value, which is 0.48, is not high enough to select this model to perform the forecasting process. In order to examine the possibility of enhancing the preceding regression model, one of the decomposition filters is used to manipulate the process of isolating the various scales in the time series of interest. The KZ filter is used, where this filter has an efficient ability to cleanly separate the components of the studied time series [100, 101, 36, 118]. The decomposition expression for any time series by using the KZ filter can be written as follows:

$$Y_t = LT_t + SE_t + SH_t \tag{2.6}$$

where Y_t is the original series, LT_t, SE_t, and SH_t denote the Long, Seasonal, and the Short-term component, respectively. After the time series are partitioned with respect to the embedded scales, a multivariate model is required to express each component. In the next sections, we will model each extracted component using a MLR model in addition to the VAR model that is used to the data of the short-term component. Figure 2.1 shows the raw and the three extracted components.

2.6.4 The Prediction Modelling for Cohoes' City Long-Term Component

Fundamentally, the analysis of the long-term trend and the interpretation of the results may be a problematic and sophisticated task with the presence of different scales of motions in the time series. To avoid all the problems and accurately calculate the trend for WD for Cohoes city, we follow the Rao and Zurbenko's method that partitions the time series into three distinguishable components the Long, Seasonal, and the Short-Term [82, 36]. Depending on this filtration base, the filter $KZ_{15,5}$, which removes any cycle of a period of less than 33 days based on the criteria of

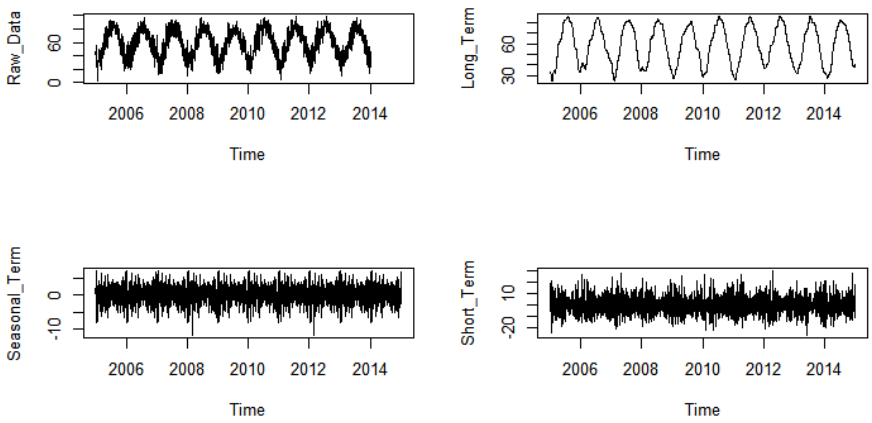

Figure 2.1: The Raw Data and the Three Components for the Temperature for Cohoes City Using the KZ Filter.

$15 \times 5^{1/2} \leq 33$, is applied for Cohoes city data. As a consequence, the result is the long-term component for each series [99, 100].

In fact, these two parameters, 15 days and 5 iterations, are chosen because they capture the required signals for the trend by achieving the highest total explanation value, R Squared. That means, we preserve the long-term component's data, and at the same time all the undesired waves (high frequency signals) are attenuated. On the other hand, to examine how the response variable and each of the meteorological conditions, tide, and groundwater level for the long-term component are correlated, the correlation matrix, which is shown in Table 2.5, is computed. The order of the absolute values of the correlation coefficients from the highest to the lowest with the WD variable are noticed for the variables GW, TD, TE, WS, and PR, respectively. For analysing the filtered data, a MLR model is formed as shown in the following

Table 2.5: The Correlation Matrix for the Long-Term Component for Cohoes City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.407	0.341	0.163	0.453	-0.623
TE	-0.407	1	-0.544	0.344	-0.065	0.087
WS	0.341	-0.544	1	-0.198	0.448	-0.235
PR	0.163	0.344	-0.198	1	-0.081	-0.040
TD	0.453	-0.065	0.448	-0.081	1	-0.512
GW	-0.623	0.087	-0.235	-0.040	-0.512	1

equation:

$$\widehat{WD}_{15,5}(t) = -0.44TE_{15,5}(t) - 0.07WS_{15,5}(t) + 0.29PR_{15,5}(t) + 0.25TD_{15,5}(t) - 0.47GW_{15,5}(t) \quad (2.7)$$

where WD, TE, WS, PR, TD, and GW denote the long-term components of water discharge, temperature, wind speed, precipitation, tide, and groundwater level, respectively. The total explanation for this model, which is measured using the R Squared value for Equation 2.7, is 0.62. This determination coefficient value means that 0.62 of the variability in the water discharge is attributed to the long-term component of the climatic variables, tide, and groundwater level.

2.6.5 The Prediction Modelling for Cohoes' City Seasonal-Term Component

The seasonal pattern, as the name may imply, is naturally related to the fluctuations that might happen by seasonal factors, such as quarter intervals, monthly periods, a day of a week, etc. The seasonality often occurs in fixed and known periods which can discriminate it from the cyclical behaviour which appears in non fixed periods. After we compute the long-term component, $KZ_{15,5}$, de-trending is the next step. This step can be conducted by subtracting the long-term series from the raw time series. Then, by using this de-trended series, the seasonal factors are computed relying on the period that is used to record these values.

Because our data is daily recorded, we need to filter this data depending on the day; after that the average for each day is computed. That means, the seasonal component is created. For instance, to calculate the seasonal effects for the first of January (01/01), we need to add all the available data for this day in the de-trended series and divide by the number of the (01/01) days in the specified period, which is 9 years $(2005 - 2013)$. By using equations, we can write this as:

$$Seasonal(t) = \frac{1}{9}\sum_{n=1}^{9}(Raw(t) - KZ(t))$$

where n is the number of times that each day appears through the range of data, for example 01/01 in the series occurred nine times. Furthermore, to build a regression model for these seasonal series, the correlation matrix is computed to select the significantly correlated variables. The correlation matrix is shown in Table 2.6. The highest two correlation coefficients with the WD are the coefficients of PR and GW level variables, then TD, WS, and TE, respectively. The regression analysis is carried out to construct a model to predict the seasonal values by using the significantly

Table 2.6: Correlation Matrix of the Seasonal-Term Component for Cohoes City.

	WD	TE	WS	PR	TD	GW
WD	1	0.039	0.061	0.454	-0.072	-0.44
TE	0.039	1	0.134	-0.104	-0.071	-0.007
WS	0.061	0.134	1	0.048	0.007	0.027
PR	0.454	-0.104	0.048	1	-0.017	0.016
TD	-0.072	-0.071	0.007	-0.017	1	-0.014
GW	-0.44	-0.007	0.027	0.016	-0.014	1

correlated variables with the water discharge as predictors. This model is shown in Equation 2.8. The determination coefficient for this model is approximately 0.42.

$$\widehat{WD}_{SE}(t) = 0.05TE_{SE}(t) + 0.47PR_{SE}(t) - 0.07TD_{SE}(t) - 0.41GW_{SE}(t) \quad (2.8)$$

where WD_{SE}, TE_{SE}, PR_{SE}, TD_{SE}, and GW_{SE} denote the seasonal components for the water discharge, temperature, precipitation, tide, and groundwater level, respectively. Also, as the parameter estimate for the variable WS has a P-value greater than the significance level, $\alpha = 0.05$, this predictor is eliminated from the regression model for the seasonal component.

2.6.6 The Prediction Modelling for Cohoes' City Short-Term Component

With regard to the components, the last step in the calculation process is related to obtain the short-term component series, which is sometimes called the synaptic, random, or irregular series. This component can be computed by subtracting the long-term trend and the seasonal factor from the raw series data. Mathematically speaking,

$$Short = Raw - Long - Seasonal.$$

The correlation matrix for the short-term component is computed and shown in Table 2.7. Examining this matrix reveals the order of correlation with the WD from highest to lowest, where based on this indicator of correlation the variables can be ordered as the following, PR, GW, TE, WS, and TD. Moreover, to examine the relationship between the response variable and the covariates of the short-term component, two distinctive modelling techniques have been used, which are:

1. Multiple Linear Regression analysis (MLR).

Table 2.7: The Correlation Matrix of the Short-Term Component for Cohoes City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.117	0.078	0.438	0.009	-0.377
TE	-0.117	1	0.049	-0.178	0.005	0.060
WS	0.078	0.049	1	0.109	-0.003	0.000
PR	0.438	-0.178	0.109	1	0.006	-0.013
TD	0.009	0.005	-0.003	0.006	1	-0.025
GW	-0.377	0.060	0.000	-0.013	-0.025	1

2. Vector Autoregressive Model (VAR).

Using regression analysis enables us to study the relationships between the current variables. On the other hand, because of the specific nature of the short-term component's data, as it includes the high frequency data, the use of the Vector Autoregressive Model (VAR) can often provide better results than regression model. The specificity of this component can be attributed to the high frequency data. So, we perform the two techniques and select the one that produces higher explanation value.

The Regression Analysis for Cohoes' City Short-Term Component

Applying MLR analysis for the data of the short-term component provides the following model:

$$\widehat{WD}_{SH}(t) = 0.421 PR_{SH}(t) - 0.358 GW_{SH}(t) \tag{2.9}$$

where WD_{SH} denotes the water discharge, PR_{SH} and GW_{SH} denote the predictors the precipitation and the groundwater level for the short-term component. Moreover, the remaining predictors, which are temperature, tide and wind speed are eliminated from the regression model as their coefficients have P-values that are greater than the significance level. This model explains about 0.32 of the variations in the response variable by using the data of the variables of precipitation and groundwater level.

Vector Autoregressive Model for Cohoes' City Short-Term Component

Since the short-term component is a random process as it is related to the events that quickly happen and finish like the fluctuations of the weather, one of the ARMA models is required to express the synoptic pattern. The structure of this methodology can be summarised by expressing each variable as a linear combination of past values (lags) of itself and the other predictors. The variables precipitation and temperature are the most important variables that affect the water discharge for the Mohawk

River [41]. The importance of these two variables can be supported by their significant correlation coefficients with WD. Based on this, the variables water discharge, temperature, and precipitation are used to construct the VAR model. The VAR(1) for the short-term component data model provided the smallest AIC and SBC values compared to the other constructing models, and is written as follows:

$$Y_t = \Phi\, Y_{t-1} + \epsilon_t \tag{2.10}$$

where

$$Y_t = \begin{bmatrix} WD_{SH_t} \\ TE_{SH_t} \\ PR_{SH_t} \end{bmatrix}.$$

The matrix Φ is a 3×3 matrix of the parameters of the first order autoregressive model, these parameters are the coefficients of the influence of the independent variables, which are the past values of the variables. The white noise series, ϵ_t, is a 3×1 vector, with a mean zero and a constant variance σ^2 and independent from the variables in the past period, mathematically speaking,

$$\mathbf{E}(\epsilon_t) = 0$$

$$\mathbf{E}(\epsilon_t \epsilon_s) = \begin{cases} \sigma^2 & \text{if } t = s \\ 0 & \text{if } t \neq s. \end{cases}$$

The autoregressive model of order one for each variable in the short-term component can be expressed as follows:

$$WD_{SH_t} = \phi_{11} WD_{SH_{t-1}} + \phi_{12} TE_{SH_{t-1}} + \phi_{13} PR_{SH_{t-1}} + \epsilon_{1t}$$

$$TE_{SH_t} = \phi_{21} WD_{SH_{t-1}} + \phi_{22} TE_{SH_{t-1}} + \phi_{23} PR_{SH_{t-1}} + \epsilon_{2t}$$

$$PR_{SH_t} = \phi_{31} WD_{SH_{t-1}} + \phi_{32} TE_{SH_{t-1}} + \phi_{33} PR_{SH_{t-1}} + \epsilon_{3t}$$

where $WD_{SH_{t-1}}, TE_{SH_{t-1}}$, and $PR_{SH_{t-1}}$ denote the values of the previous day for the variables in short-term components. By using the SAS program, the results of the VAR(1) model are listed as follows:

$$\begin{bmatrix} WD_{SH_t} \\ TE_{SH_t} \\ PR_{SH_t} \end{bmatrix} = \begin{bmatrix} 0.724 & 0.114 & 0.212 \\ -0.057 & 0.500 & -0.068 \\ -0.092 & 0.031 & 0.777 \end{bmatrix} \times \begin{bmatrix} WD_{SH_{t-1}} \\ TE_{SH_{t-1}} \\ PR_{SH_{t-1}} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{bmatrix}.$$

The model above enables us to predict the value of the current day for the variables water discharge, temperature, and precipitation, respectively, by using their values for the previous day. That is, the current value of the water discharge relies on

its previous day value as well as the values of the temperature and precipitation for the day before the targeted day. Also, this model reveals that the value of the coefficient of $WD_{SH_{t-1}}$ in the first equation, which is 0.724, is the most significant factor that influences computing the amount of WD_{SH_t}; after that the precipitation and temperature variables affect with factors 0.212 and 0.114, respectively.

Following the same method, the process of calculating the value of the current day for the other variables can be carried out. Therefore, the coefficients of the temperature in the matrix of the Model Parameter Estimates are -0.057, 0.500, and -0.068. The most important coefficient among these estimates is the coefficient of $TE_{SH_{t-1}}$ which is 0.500. Furthermore, the current value for the precipitation is related to each of the studied time series, which are water discharge, temperature, and precipitation for the previous day with about -0.092, 0.031, and 0.777, respectively. The main coefficient for forecasting the short-term component's value for the precipitation is the value of $PR_{SH_{t-1}}$, which is 0.777.

Additionally, in order to determine the percentage of the unexplained part by using this model, the residual data, which is originally related to the variables that have not been taken into account when we build this model, can be used [101]. As a consequence, the covariance matrix for the residuals (innovations) is estimated, this matrix is shown as follows :

$$\begin{bmatrix} 0.311 & -0.022 & 0.110 \\ -0.022 & 0.720 & -0.061 \\ 0.110 & -0.061 & 0.455 \end{bmatrix}.$$

Also, the process of calculating the unexplained percentage requires that the covariance matrix for the short-term component data for the variables WD, TE, and PR has to be estimated, this covariance matrix is shown as follows:

$$\begin{bmatrix} 1 & -0.101 & 0.426 \\ -0.101 & 1 & -0.172 \\ 0.426 & -0.172 & 0.99 \end{bmatrix}.$$

Then, the determinant for each covariance matrix should be calculated and after that we divide the determinant value of the innovation matrix of the VAR(1) by the value of the determinant of the covariance matrix of the short-term component, the result is the value 0.116. This value, 0.116, represents the part that has not been explained by using this model, as a consequence, the value 0.884 represents the variations in the water discharge that can be interpreted by using the values of the previous day of the water discharge itself, temperature, and precipitation variables.

2.6.7 The Contribution Percentages of the Components for Cohoes' City Data

The contribution percentages of the different scales of motions, which are embedded in a time series, can be computed by utilising the results of the KZ filtering mechanism. The results are shown in Table 2.8. Firstly, for the long-term component, which is expressed by Equation 2.7, the contribution percentage of the independent variables which are the meteorological conditions, tide, and groundwater level to the water discharge series data is about 0.44, (0.698×0.627). Furthermore, the seasonality effects, which are formed in Expression 2.8, have contributed with approximately 0.008, (0.021×0.421), to the response variable. Finally, since we have studied two modelling techniques for analysing the short-term component data, two percentages are gained:

- Firstly, the contribution percentage of the variables in the MLR model for the short-term component for the water discharge, which is expressed in Equation 2.9, is 0.05, (0.160×0.320), of the variations in water discharge data.
- Secondly, for the VAR(1) model, the R Squared value is 0.688, and the percentage of the explanation is 0.11, (0.160×0.688). This value of explanation is higher than the previous one for the MLR. That means, to predict the value of water discharge for the short-term component using the VAR(1) model provides more accurate results.

Table 2.8: The Results of the Variance and the Coefficient of Determination for all the components of the $KZ_{15,5}$ for Cohoes City.

	Variance	R Squared
Long-Term Component	0.698	0.627
Seasonal-Term Component	0.021	0.421
Short-Term Component	0.160	0.320

2.6.8 The Combining Process for Cohoes' City Components

Once we complete calculating all the required components, we can use them to fit the final combined forecasting model. We combine together the three separated components, which are shown in Equations 2.7, 2.8, and 2.9, in one model. The R Squared value for this new combined model is 0.67. When we combine the VAR(1) model for the short-term component data with the other two MLR models for the components

long and seasonal, the R Squared value for this combined model is 0.68. The difference between these R Squared values and the R Squared value before applying the KZ filter, which is 0.484, shows the ability of the KZ filter to improve the predictive performance for the MLR analysis.

2.7 The Analysis for Utica's City Data

Figure A.2 in the Appendix illustrates the steps taken to construct the developed models for Utica city. The novelty in this chapter is shown in this section using the data of Utica city.

2.7.1 The Analysis of MLR Model without an ARMA Process for the Errors for the Utica's City Raw Data

The correlation matrix has been computed for the daily raw data and the results are shown in Table 2.9. Some of the calculated correlation coefficients have significant relationship with the water discharge based on the P-values where these coefficients are related to the variables Temperature, Wind Speed, Precipitation, Tide, Cloud Cover and Groundwater Level. On the other hand, the remainder of the variables, which are Absolute Humidity, Sea Level Pressure, and Visibility Miles, would not be included in the process of building a linear model as their P-values are greater than 0.05. The highest correlation coefficient with WD is the correlation coefficients of the variable GW.

Table 2.9: The Correlation Matrix for the Raw Data for Utica City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.372	0.193	0.283	0.372	-0.451
TE	-0.372	1	-0.130	0.006	-0.071	0.075
WS	0.193	-0.130	1	0.149	0.080	-0.045
PR	0.283	0.006	0.149	1	-0.043	0.012
TD	0.372	-0.071	0.080	-0.043	1	-0.490
GW	-0.451	0.075	-0.045	0.012	-0.490	1

Figure 2.2 shows the daily raw data for the natural logarithm for the water discharge for Utica city for the considered period, 2005-2014. Furthermore, based on the available variables, a regression model was built for the raw data as shown in Equation 2.11. However, because of its non significant relation, based on the P-value of its coefficient, the cloud cover variable is eliminated from the formed model.

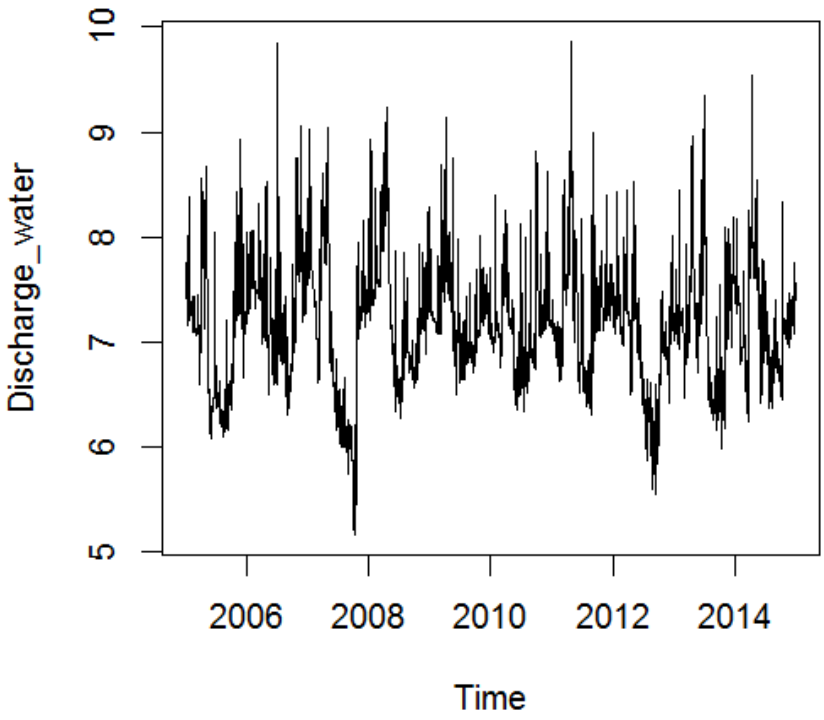

Figure 2.2: Time Series for Water Discharge for the Period 2005-2014 for Utica City.

$$\widehat{WD}_t = -0.29TE_t + 0.07WS_t + 0.28PR_t + 0.21TD_t - 0.36GW_t \tag{2.11}$$

where WD_t, TE_t, WS_t, PR_t, TD_t, and GW_t denote the raw data for the water discharge, temperature, wind speed, precipitation, tide, groundwater level, respectively. The R Squared value for this model is 0.45. This weak relationship between water discharge and the climatic variables, tide, and groundwater level is strengthened by using a decomposition technique for isolating the seasonal and short-term components from the studied series.

2.7.2 The Analysis of MLR with an ARMA Model for the Errors for the Utica's City Raw Data

The existence of some spikes in the ACF and PACF plots for the residual terms for the MLR model shown in Equation 2.11 in Figure 2.3 reveals that the constructed MLR is not adequate to fit the raw data for Utica city. These spikes indicate that there is some other information that has not been taken into consideration by this model. One solution to handle this problem is to specify an ARMA model for these residuals. The mechanism for this specification depends on examining the values of the ACF and PACF. In the PACF plot, as we have two spikes (one of them is large and the other is rather small), an AR model of order one sufficiently fits the data of the residual terms. This adequacy has been confirmed by plotting the residual terms for the final new model (MLR+AR(1)) as shown in Figure 2.4, where, apart from the first spike in the ACF plot, there are no obvious spikes appear in the plots of the ACF and the PACF.

The findings of the four model selection methods in Table 2.10 show that adding an AR(1) for the residuals of the regression model has extremely changed the results of the accuracy for the forecasting model based on the four tools used.

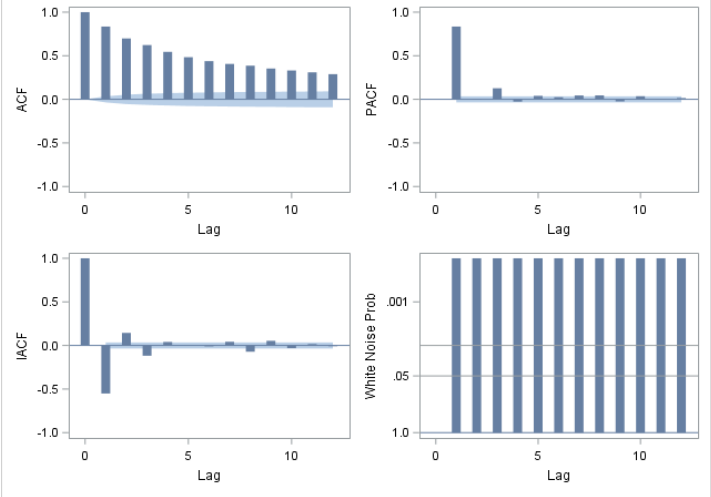

Figure 2.3: Residual Correlation Diagnostics for Water Discharge.

2.7.3 The Periods for the Studied Variables for Utica City

To spectrally examine the nature of the relationships between the studied variables, the DiRienzo and Zurbenko (DZ) smoothed function is applied to the pe-

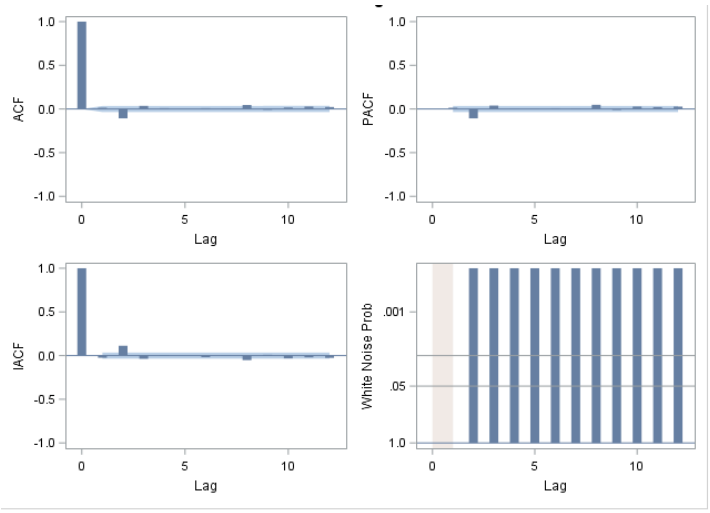

Figure 2.4: Residual Correlation Diagnostics for Water Discharge After Adding AR(1) Model.

Table 2.10: Model Selection Methods for the Raw Data for Utica City.

Tools	Regression Without Errors Model	Regression With Errors Model
Variance Estimate	0.562	0.126
Std Error Estimate	0.750	0.355
AIC	7443.578	2528.663
SBC	7480.165	2571.348

riodogram, which is calculated using the Kolmogorov-Zurbenko Fourier Transform (KZFT) method for all the variables. If we investigate the periods in Table 2.11, we can see that all the variables have relatively long periods that consist of at least 165 days, except the precipitation variable which possesses short-time periods, which are 35 and 12 days. These variations in the periods, which also refer to the changes in the frequencies of these variables, can indicate the necessity to decompose the studied time series. The isolation step enables us to individually analyse each component in our data. Figure 2.5 displays the spectrum of the smoothed series for water discharge using the DZ method with the parameter 0.0005 where the highest peak indicates that the dominant period for this variable is 365 days.

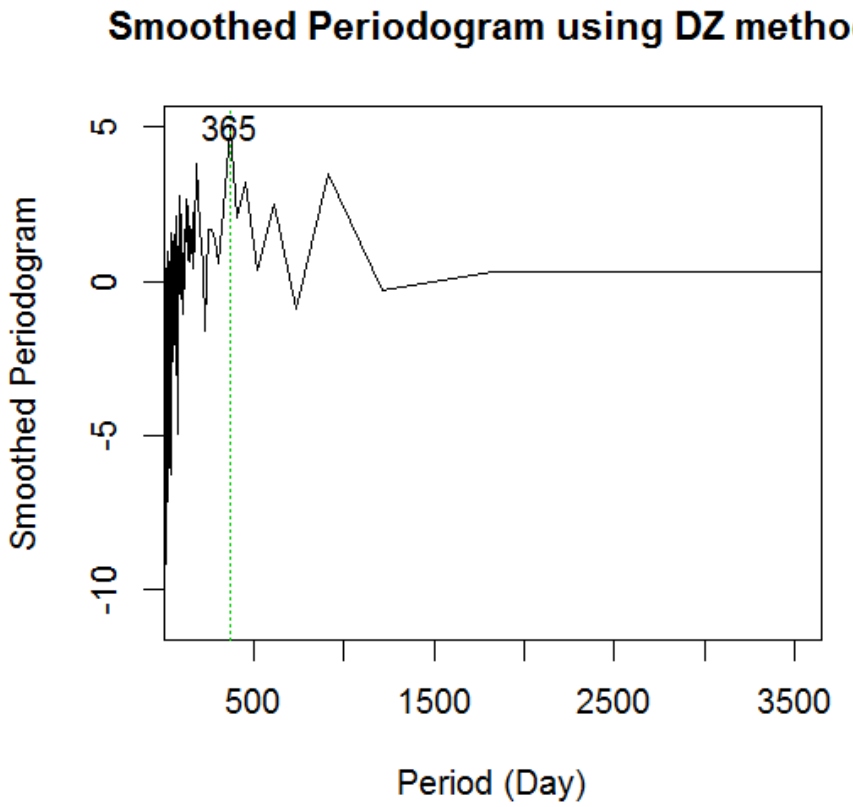

Figure 2.5: Power Spectrum of the Water Discharge Series for Utica City by using the DZ algorithm.

2.7.4 Decomposition of Utica's City Time Series

The available data for the variables of interest will be divided into three different parts, which can be defined mathematically as:

$$Y_t = LT_t + SE_t + SH_t$$

where Y_t, LT_t, SE_t, and SH_t denote the Raw, Long, Seasonal, and the Short-Term components, respectively.

Table 2.11: Periods for all the Studied Variables for Utica City by using the DZ method.

Variable	First Peak	Second Peak	Third Peak
Temperature	365	182	
Precipitation	35	12	
Groundwater	365	1217	
Water Discharge	365	182	912
Tide	365	182	
Wind Speed	365	165	

2.7.5 Prediction Modelling for Utica's City Long-Term Component without an Errors Model

For Utica's data, we obtained the long-term component for all the variables by removing all the high frequency signals variations. The high frequency fluctuations are separated by using the KZ filter of 29 days and 3 iterations as the window size and the number of iterations, respectively. These parameters 29 and 3 can allow for any cycle with more than 50 days to pass through it. The correlation coefficients for the long-term component for all variables are listed in Table 2.12. The correlation coefficients of the long-term component of water discharge series are relatively high with all the variables, except the correlation coefficient of the precipitation. This weak relationship is confirmed by the periods of these two variables, the water discharge and precipitation, as shown in Table 2.11. While the dominant periods for the water discharge are associated with cycles of at least 182 days, which is long period, the dominant period for the precipitation variable is only 35 days, which is clearly short-time period compared with the period of 182 days. Based on these results, it is expected that the precipitation variable can be more correlated with the short-term component of water discharge series rather than the long-term component of

Table 2.12: The Correlation Matrix of the Long-Term Component for the Variables of Utica City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.546	0.485	0.163	0.472	-0.527
TE	-0.546	1	-0.625	0.232	-0.090	0.085
WS	0.485	-0.625	1	-0.126	0.284	-0.150
PR	0.163	0.232	-0.126	1	-0.161	0.046
TD	0.472	-0.090	0.284	-0.161	1	-0.514
GW	-0.527	0.085	-0.150	0.046	-0.514	1

this variable, this will be shown later on in the correlation matrix for the short-term component.

For the purpose of analysing the long-term trend of the water discharge using the long-term components of the predictors, a MLR model is constructed. The expression of the model can be written as follows:

$$\widehat{WD}_{29,3}(t) = -0.492TE_{29,3}(t) + 0.077WS_{29,3}(t) + 0.415PR_{29,3}(t) + 0.377TD_{29,3}(t) - 0.263GW_{29,3}(t) \quad (2.12)$$

where $WD_{29,3}$, $TE_{29,3}$, $WS_{29,3}$, $PR_{29,3}$, $TD_{29,3}$, and $GW_{29,3}$ denote the long-term components for the water discharge, temperature, wind speed, precipitation, tide, and groundwater level, respectively. The term ϵ mentions to all other variables that are not taken into account when we build this regression model for the response variable. The coefficient of determination, R Squared, for this constructed model is 0.70 (the value of the R Squared is always between 0 and 100 %). Furthermore, all the parameter estimates for this model are significant, this is revealed by examining the P-values where all these values are 0.001.

2.7.6 Prediction Modelling for Utica's City Long-Term Component with an Errors Model

Having investigated the correlation analysis for the residual terms for the MLR model for the long-term component for the water discharge, constructing an ARMA model for these terms is necessary. We see from the values in Table 2.13 that better results have been obtained by augmenting the regression model with an AR(1) model for the residual terms, thereby accounting for the autocorrelation of the residual terms.

Table 2.13: Model Selection Method for the Long-Term Component.

Tools	Regression Without Errors Model	Regression With Errors Model
Variance Estimate	0.296	0.000
Std Error Estimate	0.544	0.026
AIC	5340.251	-14505.5
SBC	5376.838	-14462.9

2.7.7 Prediction Modelling for Utica's City Seasonal-Term Component without an Errors Model

The seasonal component series for each studied variable is found by performing the following two steps:

- Subtracting the long-term series, the $KZ_{29,3}$, from the raw data, and the result is a de-trended series.
- Filtering the de-trended series for each variable depending on the day, as our data is daily, and taking the average for each day in the studied data. Mathematically, this can be written as follows:

$$SE(t) = \frac{1}{9}\sum_{n=1}^{9}\left(Raw(t) - KZ_{29,3}(t)\right)$$

where 9 here is the number of times that each day appears through the range of data, which is 2005-2013. For example, if we want to calculate the average of the day $02-01$, the second of January, in all years chosen, we shall add all the values from the de-trended series which are associated with the day $02-01$ and after that divide it by 9.

We compute the correlation matrix for the seasonal variations to select the explanatory variables that will be used in the process of building a model. If we exam-

Table 2.14: The Correlation Matrix of the Seasonal-Term Component for Utica City.

	WD	TE	WS	PR	TD	GW
WD	1	0.057	0.206	0.377	0.074	-0.397
TE	0.057	1	0.206	0.008	-0.005	-0.048
WS	0.206	0.206	1	0.146	-0.029	0.000
PR	0.377	0.008	0.146	1	-0.001	0.088
TD	0.074	-0.005	-0.029	-0.001	1	-0.170
GW	-0.397	-0.048	0.000	0.088	-0.170	1

ine the correlation coefficients for the water discharge with all the predictors, we can highlight those with a high correlation coefficient with the water discharge, which are groundwater level, precipitation, and wind speed as shown in Table 2.14. For the purpose of explaining and predicting the seasonal component of water discharge, MLR analysis is used. The MLR model can be written as follows:

$$\widehat{WD}_{SE}(t) = -0.309TE_{SE}(t) + 0.151WS_{SE}(t) + 0.299PR_{SE}(t) + 0.223TD_{SE}(t) - 0.301GW_{SE}(t) \quad (2.13)$$

where $WD_{SE}, TE_{SE}, WS_{SE}, PR_{SE}, TD_{SE}, andGW_{SE}$ denote the seasonal components for the variables, water discharge, temperature, wind speed, precipitation, tide, and groundwater level, respectively. This model explains about 0.40 of the variations in the WD using the preceding variables. All the parameter estimates for this model are statistically significant relying on the P-values.

2.7.8 Prediction Modelling for Utica's City Seasonal-Term Component with an Errors Model

An AR model of order one has been assigned to the data of the residual terms for the MLR model. By examining the results in Table 2.15, the decision of including an AR model has enhanced the results of forecasting based on the model selection methods used.

Table 2.15: Model Selection Method for the Seasonal-Term Component.

Tools	Regression Without Errors Model	Regression With Errors Model
Variance Estimate	0.649	0.246
Std Error Estimate	0.805	0.496
AIC	7915.419	4730.703
SBC	7952.005	4773.387

2.7.9 Prediction Modelling for Utica's City Short-Term Component

To compute the short-term component series, we subtract the previous two components, which are the long and seasonal, from the raw data. Mathematically, this series can be written as follows:

$$SH_{29,3} = Raw - Long_{29,3} - Seasonal_{29,3}.$$

The correlation matrix is calculated for the short-time series for all the variables. The highest correlation coefficient with the WD in this matrix is the one that is associated with the precipitation series, 0.434, as shown in Table 2.16. The groundwater level and wind speed also have significant relationships with the short-term component of water discharge. In contrast to this situation, temperature and tide have a weak relation with the water discharge in the short-term component. To predict the short-term component of the water discharge, two methodologies are implemented, the Regression Analysis and the Vector Autoregressive Model, VAR.

Table 2.16: The Correlation Matrix of the Short-Term Component for Utica City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.021	0.150	0.434	0.016	-0.208
TE	-0.021	1	0.084	-0.145	0.021	0.059
WS	0.150	0.084	1	0.179	-0.000	-0.003
PR	0.434	-0.145	0.179	1	-0.017	-0.014
TD	0.016	0.021	-0.000	-0.017	1	-0.012
GW	-0.208	0.059	-0.003	-0.014	-0.012	1

2.7.10 The Regression model for Utica's City Short-Term Component Without an Errors Model

The MLR model is implemented where the water discharge series represents the response variable, and precipitation, groundwater level, and wind speed are the independent variables. This linear model is shown in Expression 2.14.

$$\widehat{WD}_{29,3}(t) = 0.059WS_{29,3}(t) + 0.426PR_{29,3}(t) - 0.209GW_{29,3}(t) \tag{2.14}$$

where $WD_{29,3}, WS_{29,3}, PR_{29,3}$, and $GW_{29,3}$ denote the water discharge, wind speed, precipitation, and groundwater level, respectively. In the MLR model above, the parameter estimate for the precipitation variable has a remarkable influence on the predicted amount of the water discharge in the short-term component. This result is identical with the spectrum analysis result for these variables, as shown in Table 2.11 where the dominant periods for the precipitation variable have short-times which are 35 and 12 days, respectively. The MLR model interprets about 0.23 of the variations in the short-term component of the water discharge.

2.7.11 The Regression model for Utica's City Short-Term Component with an Errors Model

As shown in Table 2.17, a considerable difference is observed after the inclusion of an AR(1) model for the residual terms of the MLR model for the short-term component.

The Vector Autoregressive Model for Utica's City Short-Term Component

Based on the tests of AIC and SBC for model selection, the Vector Autoregressive model of order 1, VAR(1), which is one of the VARMA models, has been chosen. In our study, the water discharge, temperature, and precipitation are the variables that

Table 2.17: Model Selection Method for the Short-Term Component.

Tools	Regression Without Errors Model	Regression With Errors Model
Variance Estimate	0.764	0.389
Std Error Estimate	0.874	0.624
AIC	8451.216	6238.872
SBC	8487.803	6281.556

construct the structure of the VAR(1). The results can be written as follows:

$$\begin{bmatrix} WD_{SH_t} \\ TE_{SH_t} \\ PR_{SH_t} \end{bmatrix} = \begin{bmatrix} 0.709 & 0.125 & 0.099 \\ -0.050 & 0.536 & -0.082 \\ 0.003 & 0.044 & 0.744 \end{bmatrix} \times \begin{bmatrix} WD_{SH_{t-1}} \\ TE_{SH_{t-1}} \\ PR_{SH_{t-1}} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{bmatrix}.$$

That means, to predict the value of the current day for the water discharge, we need to use the values of the previous day (lag1) for the water discharge, temperature, and precipitation. The highest parameter estimate in the water discharge equation is associated with the variable of lag one of the water discharge, 0.709, while the temperature and precipitation affect with approximately 0.125 and 0.099, respectively. Similarly, for predicting the value of the temperature for the present day, the coefficient estimate for the variable of the previous day, lag 1, is the highest parameter estimate, 0.536. Then each of the precipitation and the water discharge affect with -0.082, -0.050, respectively. Finally, in order to calculate the precipitation value for the current day, the variable of the previous value for this variable, lag 1, has a factor of 0.744, then 0.044 for the lag one of the temperature, and the parameter estimate 0.003 refers to the value of the lag 1 for the water discharge.

To determine how much this VAR(1) model explains from the variations in our data, we need to compute the covariance matrix for the innovations for this model as well as the covariance matrix for the short-term component for the studied series. The following matrix represents the covariance matrix for the residual terms (innovations) of the VAR(1).

$$\begin{bmatrix} 0.415 & 0.001 & 0.135 \\ 0.001 & 0.685 & -0.034 \\ 0.135 & -0.034 & 0.450 \end{bmatrix}.$$

In addition, the covariance matrix of the variables water discharge, temperature, and precipitation has also been computed and is shown as follows:

$$\begin{bmatrix} 0,99 & -0.014 & 0.431 \\ -0.014 & 0.99 & -0.145 \\ 0.431 & -0.145 & 0.98 \end{bmatrix}.$$

After these covariance matrices are calculated, the determinant for each of them is computed. Therefore, by dividing the determinant of the covariance matrix of the residual terms of the VAR(1) model by the determinant of the covariance matrix of the short-term component, the result is 0.14 where this refers to the unexplained part. Consequently, this means that the explained variance by using the vector autoregressive model of order one is about 0.86 of the total variations in our studied data.

2.7.12 The Contribution Percentages for the Components for Utica City

For testing the effectiveness of the decomposition process, a calculation procedure should be conducted to achieve this purpose. The process requires to multiply the value of the coefficient of determination, R Squared, for each component by the proportion of the variance of the water discharge series also for each component. The proportion of the variance is computed by dividing the variance of the water discharge for each component over the variance of the raw data of the water discharge. The R Squared values and the proportions of the variances are shown in Table 2.18.

Firstly, the contribution of the long-term component of the meteorological variables, which are temperature, precipitation, and wind speed, in addition to the other variables, which are tide and groundwater level, to the water discharge series is about 0.38, (0.54×0.70), from Equation 2.12. Moreover, 0.01, (0.045×0.4), is the percentage of the explanation that is attributed to the seasonal variations from Equation 2.13. With reference to the short-term component, since two distinguishable techniques are applied, we have two different figures for the contribution percentage for this component. Firstly, the 0.06, (0.27×0.23) is the total explanation of the short-term component using the regression analysis from Equation 2.14. Secondly, 0.15, (0.58×0.27) is the amount of the explanation in the data of the water discharge by modelling the short-term components for the studied variables using an VAR(1) model.

Table 2.18: The Results of the Variance and the Coefficient of Determination for all the Components of Utica City.

	Variance	R Squared
Long-Term Component	0.54	0.70
Seasonal-Term Component	0.045	0.4
Short-Term Component	0.27	0.23

2.7.13 The Combining Process for Utica's City Components

If we combine all the components, which are shown in Equations 2.12, 2.13, and 2.14, the result of the R Squared value is 0.56. This value is better than the R Squared value for the raw data, 0.45. However, the error terms for this combined MLR model are autocorrelated as shown in Figure 2.6. To remedy this problem of autocorrelated

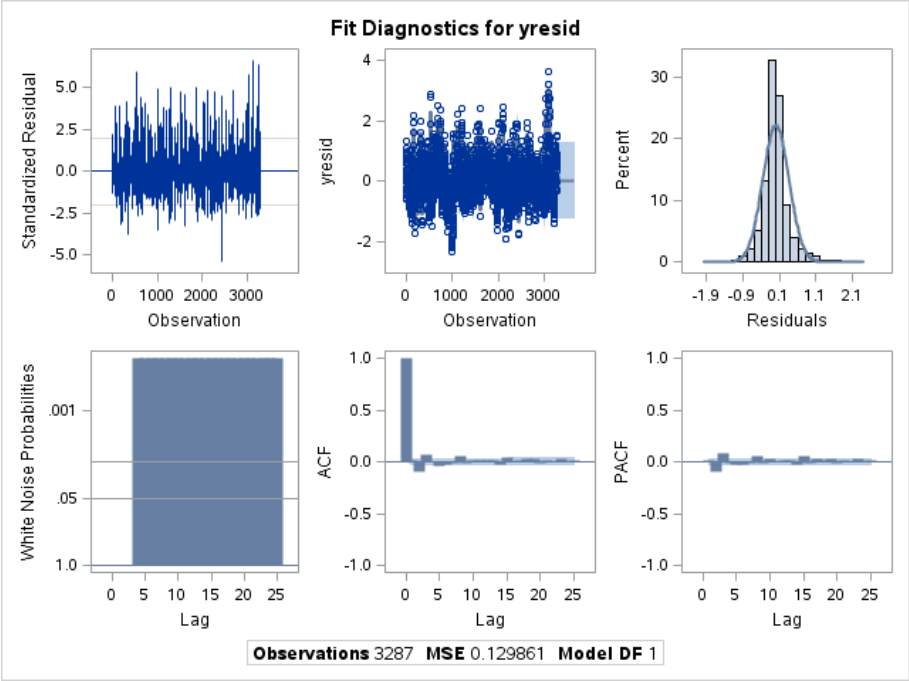

Figure 2.6: Fit Diagnostics for the Residuals of Combined Regression Model for the Water Discharge.

residual terms, an AR(1) model has been used to fit the residuals terms. The results of new combined MLR model with an AR(1) model for the residuals are shown in Table 2.19. Moreover, when we combine the regression models for the long and seasonal

Table 2.19: Model Selection Method for the Final Model.

Tools	Regression Without Errors Model	Regression With Errors Model
Variance Estimate	0.432	0.222
Std Error Estimate	0.657	0.471
AIC	6590.497	2453.619
SBC	6688.061	2557.28

components with the VAR(1) model for the short-term component, the result of the R Squared value becomes 0.66. Again, this value is also better than the R squared values 0.45 and 0.56 for the MLR for the raw data and the Combined MLR.

2.8 The Analysis for Poughkeepsie's City Data

The same variables have been chosen to construct MLR and VAR models for the data of Poughkeepsie city.

2.8.1 The Analysis for Poughkeepsie's City Raw Data

At the beginning, the correlation matrix is computed to choose the statistically significant variables that can affect the water discharge series. The variables Absolute Humidity, Sea Level Pressure, Visibility Miles, and Cloud Cover are removed from the analysis as they have non significant relations with the WD. Furthermore, because of its high correlation coefficient with the temperature variable, the variable Dew point is also eliminated from the analysis. The correlation matrix, which is shown in Table 2.20, reveals that almost all the remaining variables have high correlation coefficients with the dependent variable.

Table 2.20: The Correlation Matrix of the Raw Data for Poughkeepsie City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.422	0.144	0.395	0.290	-0.436
TE	-0.422	1	-0.126	0.019	-0.067	0.223
WS	0.144	-0.126	1	0.080	0.125	-0.090
PR	0.395	0.019	0.080	1	-0.048	0.058
TD	0.290	-0.067	0.125	-0.048	1	-0.393
GW	-0.436	0.223	-0.090	0.058	-0.393	1

For the purpose of predicting the long-term component for WD, the regression analysis is performed based on the explanatory variables. Because of its non significant P-value, wind speed variable has been removed from the model. The constructed model explains about 0.50 of the variations in the WD data, and this model is shown in Equation 2.15

$$\widehat{WD}_t = -0.347TE_t + 0.414PR_t + 0.159TD_t - 0.321GW_t. \tag{2.15}$$

2.8.2 The Periods for the Studied Variables for Poughkeepsie City

To investigate the spectral content of the studied data, the DiRienzo and Zurbenko (DZ) smoothed function is performed to smooth the periodogram, which is previously calculated by using the Kolmogorov-Zurbenko Fourier Transform (KZFT) method, of all the variables. If we examine the periods in Table 2.21, we can detect that most of the studied variables possess long periods except the precipitation variable. This can verify the necessity of applying a decomposition technique to separate each pattern in the studied data.

Table 2.21: The Periods of all the Studied Variables for Poughkeepsie City by using the DZ method.

Variable	First Peak	Second Peak	Third Peak
Temperature	365	182	
Precipitation	19	12	
Groundwater	365	912	260
Discharge Water	365	608	280
Tide	365	182	
Wind	365	3651	912

2.8.3 Decomposition of Time Series for Poughkeepsie City

Often, when a regression analysis, which is performed for a time series data, has a relatively low R Squared value, there is a possibility to enhance this model. This enhancing process will be conducted by re-analysing the same data but after applying one of the decomposition techniques for the original time series. Applying the $KZ_{29,3}$ filter for all the variables for Poughkeepsie city provides three patterns for each variable. In fact, this KZ filter removes all the high frequency signals (short-term component), which are signals with a period of less than 50 days. The parameters 29 and 3 have been chosen as they provide a regression model with the highest R Squared value.

2.8.4 Prediction Modelling for Poughkeepsie's City Long-Term Component

Separating the long-term signals provides a good opportunity to analyse it by using the regression analysis. This analysis is accomplished after computing the correlation matrix for all the long-term component variables, this matrix is shown in Table

2.22. Having investigated this correlation matrix and determined the variables that

Table 2.22: The Correlation Matrix for the Long-Term Component Data for Poughkeepsie City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.566	0.370	0.267	0.364	-0.620
TE	-0.566	1	-0.448	0.248	-0.080	0.269
WS	0.370	-0.448	1	-0.221	0.444	-0.344
PR	0.267	0.248	-0.221	1	-0.146	0.066
TD	0.364	-0.080	0.444	-0.146	1	-0.458
GW	-0.620	0.269	-0.344	0.066	-0.458	1

construct the prediction model, the regression analysis is performed and the resultant model is shown in Expression 2.16.

$$\widehat{WD}_t = -0.561TE_t + 0.444PR_t + 0.204TD_t - 0.414GW_t. \tag{2.16}$$

This model explains approximately 0.80 of the variations in the water discharge series. Furthermore, all the parameter estimates are statistically significant only the parameter estimate for the variable wind speed has a high P-value, so, for this reason this variable is not included in the regression model.

2.8.5 Prediction Modelling for Poughkeepsie's City Seasonal-Term Component

Once we obtain the long-term component, the process of calculating the seasonal factor can be enabled. Examining the relationships between the studied variables by using the correlation matrix, which is shown in Table 2.23, can help us to select the predictors that will be used to build the regression model for forecasting the seasonal component of the water discharge. The highest correlation coefficient has been observed for the data of precipitation. Equation 2.17 represents the prediction expression for the seasonality of water discharge. This forecasting model explains about 0.31 of the variations in water discharge series. The variable wind speed again has a non significant relationship based on the P-value.

$$\widehat{WD}_t = -0.082GW_t + 0.537PR_t + 0.062WS. \tag{2.17}$$

Table 2.23: The Correlation Matrix of the Seasonal Fluctuations Data for Poughkeepsie City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.088	0.090	0.547	0.009	-0.123
TE	-0.088	1	0.013	-0.125	-0.019	0.094
WS	0.090	0.013	1	0.045	0.003	-0.044
PR	0.547	-0.125	0.045	1	0.027	-0.068
TD	0.009	-0.019	0.003	0.027	1	0.040
GW	-0.123	0.094	-0.044	-0.068	0.040	1

2.8.6 Prediction Modelling for Poughkeepsie's City Short-Term Component

Two methods have been used to analyse the data of this component, which are the Regression Analysis and the Vector Autoregressive model of order one VAR(1). Before using these two methods, the correlation matrix was computed to investigate the relationships between the variables, Table 2.24 shows that the most related predictor to the water discharge is the precipitation variable.

Table 2.24: The Correlation Matrix of the Short-Term Component Data for Poughkeepsie City.

	WD	TE	WS	PR	TD	GW
WD	1	-0.053	0.116	0.577	0.020	0.057
TE	-0.053	1	0.004	-0.135	0.001	-0.008
WS	0.116	0.004	1	0.130	-0.007	0.035
PR	0.577	-0.135	0.130	1	-0.022	0.100
TD	0.020	0.001	-0.007	-0.022	1	0.012
GW	0.057	-0.008	0.035	0.100	0.012	1

The Regression Model for Poughkeepsie's City Short-Term Component

The multiple linear regression model is built by using the data of the short-term component. All the P-values for the coefficients of the independent variables are greater than the significance level, 0.05, except the P-value for the precipitation coefficient, so, the regression model has one predictor as shown in Equation 2.18. The explained variance using this model is about 0.33.

$$\widehat{WD}_t = 0.560 PR_t. \tag{2.18}$$

The Vector Autoregressive Model for Poughkeepsie City's Short-Term Component

The VAR(1) model is chosen, based on the values of the information criteria AIC and SBC, to construct the prediction model for the short-term component for Poughkeepsie city. The following expression shows the VAR(1):

$$\begin{bmatrix} WD_{SH_t} \\ TE_{SH_t} \\ PR_{SH_t} \end{bmatrix} = \begin{bmatrix} 0.735 & 0.028 & 0.187 \\ -0.030 & 0.502 & -0.043 \\ -0.066 & 0.017 & 0.774 \end{bmatrix} \times \begin{bmatrix} WD_{SH_{t-1}} \\ TE_{SH_{t-1}} \\ PR_{SH_{t-1}} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{bmatrix}$$

where WD, TE, and PR denote the water discharge, temperature, and precipitation, respectively. Basically, to predict the current value for the water discharge, the lag 1 for each variable, which means the value of the previous day, should be used to build the model. Having estimated the coefficients of the VAR(1), we can notice that each of which affects with about 0.735, 0.028, 0.187, respectively. Also, it is clear that the dominant parameter estimate has been observed for the lag1 for WD.

Similarly, it can be noticed that the most important coefficient that is related to the equation of the temperature is the coefficient of the variable of temperature itself but for the previous day value (lag 1). Finally, for the precipitation, again the highest parameter estimate is related to the lag 1 for the precipitation. The VAR(1) model explains about 0.90 of the variations in the water discharge series for the short-term component. This explained variance is computed by calculating the determinants for the covariance matrix of the residuals of the VAR(1) and the covariance matrix of the variables of the short-term component. Then, by dividing the determinant of the innovations (residuals) of the VAR(1) model over the determinant of the variables for the short-term component, the result is approximately 0.10. This value represents the unexplained amount of the variance when we use this model to predict the value of the water discharge. These two matrices are shown below where the first matrix displays the variance of the residuals of the VAR(1) model, and the second matrix is the covariance matrix of the variables of the short-term component.

$$\begin{bmatrix} 0.268 & 0.021 & 0.161 \\ 0.021 & 0.734 & -0.049 \\ 0.161 & -0.049 & 0.457 \end{bmatrix}.$$

$$\begin{bmatrix} 1 & -0.047 & 0.574 \\ -0.047 & 1 & -0.136 \\ 0.574 & -0.136 & 0.99 \end{bmatrix}.$$

2.8.7 The Contribution Percentages for the Components in Poughkeepsie City

In order to examine to what extent the decomposition process is effective to perform the separation of the embedded components, the contribution of each component should be calculated. The variances and the R Squared values are listed in Table 2.25. Firstly, for the long-term pattern, we multiply the proportion of the variance of the long-term component series by the value of the R Squared of this component, (0.66×0.80), and the result is 0.53.

The proportion of the variance of the water discharge for each component can be computed by dividing the variance of each component over the variance of the original water discharge series (water discharge series before the decomposition process). In a similar way, the contribution of the other components can be calculated. While the seasonal component contributes with about 0.005, the contribution of the short-term component by using the regression analysis approach is 0.06. However, the short-term component in the VAR(1) contributes with approximately 0.13. This value is obtained by multiplying the R Squared value of the VAR(1), which is 0.73, by the proportion of the variance of the short-term, which is 0.18.

Table 2.25: The Results of the Variance and the Coefficient of Determination for all the Components of the Variables for Poughkeepsie City.

	Variance	R Squared
Long-Term Component	0.66	0.80
Seasonal-Term Component	0.01	0.50
Short-Term Component	0.18	0.33

2.8.8 Combining Process for Poughkeepsie City's Components

Finally, to construct the final model, which combines all the components together, we firstly combined the variables of Equations 2.16, 2.17, and 2.18, the R Squared value for this model is 0.62. But when we combine the two regression models for the long and seasonal components with the AR(1) model for the short-term component, the R Squared value became 0.74. Both of these models are better than the regression model for the raw data based on the R Squared value to forecast the future values for the short-term component.

2.9 Discussion

The frequency content for the variables of the three cities is similar for all the variables, except the frequency content for the precipitation variable. While a cycle of 365 days is observed for all the studied variables, cycles of 13, 35, and 19 days are detected for the precipitation variable for the three cities, Cohoes, Utica, and Poughkeepsie, respectively. For the three cities, a small difference exists for the behaviours of the considered variables. This result is derived from the MLR model's coefficients and the correlation coefficients of the water discharge with the other predictors, which are temperature, wind speed, precipitation, tide, and groundwater level. Specifically, Cohoes and Utica's results are the closest to each other. For example, the effect's order of the independent variables on the response variable, WD, is similar. The effect of the variables of the raw data can be ordered from the highest regression coefficient to the lowest regression coefficient as follows: GW, TD, TE, PR, and WS. The same pattern can be noticed for the correlation coefficients of WD with the independent variables.

For the three cities and three components, while there is no specific pattern observed according to the regression models coefficients and the correlation coefficients, the difference between their regression and correlation coefficients with WD is small. However, for the short-term component, the precipitation variable is the most important one based on its regression coefficient and its correlation coefficient with WD.

With regard to the VAR(1) models for the short-term component for the three cities, these models are built using the variables of lag one for the three studied variables WD, TE, and PR. For each city, the model VAR(1) constructs three equations, each equation represents an AR(1) model for a variable. For water's discharge variable equations, WD, the highest coefficient is the coefficient of the WD_{t-1} variable, where for Cohoes, Utica and Poughkeepsie, we have 0.72, 0.70, and 0.735. Similarly, the highest coefficient in the temperature's variable equations, TE, is the coefficient of TE_{t-1} with 0.50, 0.53, 0.50. Finally, for the precipitation's variable equations, PR, the coefficient of PR_{t-1} is the highest coefficient with 0.77, 0.74, 0.74 for Cohoes, Utica, and Poughkeepsie.

For the three cities, the explained variances for the VAR(1) models are 0.88, 0.86, and 0.90. These percentages indicate that the constructing models adequately fit the data of these variables.

The contribution percentages for the three cities, Cohoes, Utica, and Poughkeepsie, for the long-term component are 0.44, 0.38, and 0.53; for the seasonal-term component are 0.008, 0.01, and 0.005; for the short-term component for regression model are 0.05, 0.06, and 0.06; and for regression model for the VAR(1) are 0.11, 0.15, and 0.13.

The results of R Squared values show that the performance of the combined MLR

models for the three cities outperform the performance of the MLR models constructed using the raw data. For the three cities, the R Squared values have been changed from 0.48, 0.45, 0.50 using the raw data to 0.67, 0.56, 0.62 using the decomposed data the long, seasonal, and short-term components.

For Utica city, based on the model selection methods used, the combined MLR model with an AR(1) model for the residual terms of the combined MLR is better than the combined MLR without an ARMA model. For example, the AIC value has been reduced from 7443.578 to 2528.663.

2.10 Conclusion

The Decomposition process that provides three components with different time scales has improved the prediction accuracy of MLR model. The three components are the long, seasonal, and the short-term component. To obtain these components, the Kolmogorov Zurbenko filter has been used. The principle of this filter is derived from the Moving Average filter. The spectral contents of the three components based on this filter can be distributed as follows. Any event lasts for a short time, most often ranging between 2 days to 3 weeks, will be included in the short-term component. Also, any event maximally lasts one year will be included in the seasonal component. Finally, any event needs more than one year to finish will be included in the long-term component. Therefore, based on the results of this study, the variations of the frequency content of the studied variable are one of the reasons that lead to "poor" R Squared values for the MLR model for the three cities. The evidence is that when a MLR model is built based on the decomposed data, the R Squared values have been increased with relatively high percentages for the three studied cities.

Furthermore, constructing an ARMA model using the residual terms has enhanced the accuracy of the combined MLR. This result has been extracted based on the results of the model selection methods of AIC and SBC.

For the three cities, based on the correlation matrices for the raw data, there is no high correlation coefficient for the precipitation variable with the WD. However, the precipitation's short-term component is highly related to the short-term component of the WD for the three cities. It seems that having different periods causes this result. The periods are computed using the DiRienzo and Zurbenko smoothing algorithm (DZ).

The long-term component has contributed with the highest percentage in the final combined MLR then the short-term component, and finally the seasonal fluctuations.

Chapter 3

Combined Transfer Function-Noise Model for Forecasting Water Discharge

In the previous chapter we used the regression model. The structure of this model is not specifically designed to deal with time series data. The special nature of time series data requires methods that specifically deal with them to avoid some problems, such as autocorrelation between residuals, and to exploit some features, such as lagged variables. In time series analysis, the inclusion of a number of lagged variables leads to improve the constructed models [15]. The Transfer Function-Noise (TF-Noise) model is one of the most important models that their structures inherently include lagged variables [110, 15]. A considerable amount of literature has been published on this model [13, 48, 63]. However, as far as we now, no study has considered the case when decomposed data are used to construct this model rather than raw data. In this chapter, a new developed model, which can be called Combined TF-Noise model (CTF-Noise), is constructed using the decomposed data. Based on the model selection methods used, this model yields better results than the TF-Noise model constructed using the raw data.

This chapter has been organised as follows. Section 3.1 presents a brief description of types of time series models. Section 3.2 gives an explanation about how to build a TF-Noise model. Section 3.3 provides a brief overview of the backshift operator in time series modelling. Section 3.4 presents the analysis of the Poughkeepsie's raw data using the TF-Noise model. Section 3.5 is concerned with the methodology employed to the decomposed data. Section 3.6 deals with the process of combining the components in one final model. Section 3.7 provides methods to evaluate the estimated Models for Poughkeepsie city. Section 3.8 provide an analysis for the data of Cohoes city. Section 3.9 analyses the data gathered for Utica city using the TF-Noise model. Section 3.10

presents the methods that are used to evaluate the constructed models for Utica city. Section 3.11 presents a brief discussion of the most important points obtained in this chapter. The conclusion of this chapter is presented in Section 3.12.

3.1 Transfer Function-Noise Model

Time series analysis is frequently used to analyse any data that has been regularly recorded in any dynamic system [37, 59]. This extensive use of the methods of time series analysis can be attributed to that these methods consider the stochastic nature for time series data. At the beginning, to determine which method among several time series techniques can be used, we firstly need to define the model for the data. In general, there are two types of models that are based on the number of the studied series, which are the Univariate and Multivariate Models. A univariate model is constructed using the current and past values for a single time series and often one of the Autoregressive Moving Average (ARMA) models is used to build the structure of this model. In this case, any other variables that can affect this series will not be involved in the analysis. For example, a model for a temperature series can be constructed using only the current and past values for this series via an AR(1) model.

However, in some fields such as hydrology and economics, a number of related variables that influence the time series of interest have to be included in the model's structure. For example, in an economics system, sales, which is the response variable, is often associated with the past and current values of the advertising variable, which is the independent variable. The ARMA models that contain one or more of the input variables with present and past values are also called the Transfer Function Models (TF) [98]. The TF models are also known as the Transfer Function-Noise (TF-Noise) Models, as a model for the residual terms has to be added to the structure of the function. The TF-Noise models are regarded as an extended case of the linear regression analysis. Therefore, sometimes they are also referred to as dynamic regression models [92].

Moreover, as there is a possibility to include a lagged variable for the inputs, this model is rather similar to the distributed lag model in the economics field. In fact, for these two kinds, univariate and multivariate, after determining the number of observations which are involved in the analysis, the resultant time series model can be exploited to:

- Perform the process of forecasting for the future values of the desired series.
- Perform Stochastic Simulation.

Multivariate time series analysis constructed using a TF-Noise model is one of the most important and common methods that can be used when the variable of interest

belongs to the hydrological field [13]. In this method, the response variable, which is often a hydrological one, for example a water discharge series for a catchment or a river, will be related to a number of input variables which will represent the independent variables (predictors) such as the precipitation variable.

The analysis by using the TF-Noise is preferred rather than using the univariate time series analysis for some significant aspects. Firstly, as the univariate time series model takes into account just the series of interest, using the input variables to describe the response variable can provide a physical interpretation for the relationship between the dependent and the independent variables. Secondly, applying the univariate time analysis requires that the studied series has to be stationary and does not contain any periodicity to obtain an acceptable results.

In contrast to this situation, the analysis by using the TF-Noise model does not actually need these requirements (stationarity and non periodicity) to be carried out. Also, when we use this function to implement the modelling task, the studied time series (response variable) will already be decomposed into different parts, where each input series (independent variables) will be linked directly to one of these parts. So, depending on these parts, we may be able to determine which input variable can be responsible for most of the variations in the considered time series.

In particular, the TF model has been utilised to construct a model for most of the hydrological variables such as groundwater level (Tankersle et al., 1993; Gehrels et al., 1994), stream flows (Hipel et al., 1975; Chow et al., 1983), and suspended sediment concentration (Gurnell and Fenn, 1984; Lemke, 1991) (as cited in [13]). Moreover, most applications of this function in the hydrological area have been done by using precipitation as an input series.

In general, if we have an output series at time t, Y_t, and an input series at time t, X_t, the Autoregressive Moving Average model (ARMA) of the order (p, q) for each series can be written as follows:

$$\phi(B)Y_t = \theta(B)\epsilon_t \quad \text{for output series,}$$

and

$$\phi(B)X_t = \theta(B)\epsilon_t \quad \text{for input series}$$

where B represents the backward shift operator, which is defined for the autoregressive and moving average models as follows:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \quad \text{for the autoregressive model,}$$

and

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_p B^p \quad \text{for the moving average model, respectively.}$$

The values θ_j and ϕ_j are the coefficients of the models, which are real constants. $\epsilon's$ denote the noise term, white noise, which has a normal distribution with zero mean and constant variance, σ^2. It is possible to regard the Box-Jenkins model, ARMA (p,q) model, as a linear dynamic process such that the white noise term, ϵ_t, is the input variable and Y_t is the output variable. This idea can be extended significantly to include some input series that are related to the output series, Y_t [110].

3.2 How to Build a Transfer Function-Noise and a Combined Transfer Function-Noise Models

Let X_t and Y_t denote the input and output series at time t. In general, TF model can be written as follows:

$$Y_t = \mu + \frac{C\omega(B)}{\delta(B)} B^b X_t + \epsilon_t \tag{3.1}$$

where

- μ is the mean term, which needs to be included in the model when the point estimate of it has an absolute t value that is greater than 2.
- Y_t is the output (response) series.
- X_t is the input series.
- t is the time.
- C represents the scale parameter.
- B is the backward shift operator, which is $BX_t = X_{t-1}$.
- $\omega(B)$ is the backshift operator for the weights (parameters) of the numerator of the transfer function model for the input variable, which is $\omega(B) = 1 - \omega_1 B - \ldots - \omega_s B^s$.
- $\delta(B)$ is also the backshift operator for the parameters but for the denominator of the TF model, which is $\delta(B) = 1 - \delta_1 B - \ldots - \delta_r B^r$.
- b is the delay time, which is tentatively identified by using the Sample Cross-Correlation Function (SCCF).
- ϵ_t is the residual series.

This model is illustrated in Appendix in Figure A.5.

This model with one input variable can be extended easily to include other independent variables, then it can be rewritten as the following:

$$Y_t = \mu + \sum_j \frac{C_j \omega_j(B)}{\delta_j(B)} B^{b_j} X_{j,t} + \epsilon_t. \tag{3.2}$$

where C_j represents the scale parameter, $X_{j,t}$ denotes the jth input time series. b_j denotes the delay time for the influence of the jth input series on the output series. $\omega_j(B)$ denotes the polynomial factors for the numerator of the TF model for the jth input series. $\delta_j(B)$ denotes the polynomial factors for the denominator of the TF model for the jth input series. Then, after the procedure of selecting an appropriate ARMA model for the residuals, ϵ_t, in Equation 3.2, is accomplished, the final model will be written as follows:

$$Y_t = \mu + \sum_j \frac{C_j \omega_j(B)}{\delta_j(B)} B^{b_j} X_{j,t} + \frac{\theta(B)}{\phi(B)} a_t \tag{3.3}$$

where a_t denotes the disturbance series that has zero mean and constant variance, σ^2.

The structure of our new model, a combined TF-Noise model, is similar to the structure of TF-Noise model except that each term for each input variable in the TF-Noise model needs to be decomposed into long, seasonal, and short-term component. For one variable, this can mathematically be written as follows:

$$Y_t = \mu + \frac{C_{LT} \omega_{LT}(B)}{\delta_{LT}(B)} B^{b_{LT}} X_{LTt} + \frac{C_{SE} \omega_{SE}(B)}{\delta_{SE}(B)} B^{b_{SE}} X_{SEt} + \frac{C_{SH} \omega_{SH}(B)}{\delta_{SH}(B)} B^{b_{SH}} X_{SHt} + \epsilon_t \tag{3.4}$$

where LT is the Long-Term component, SE is the Seasonal-Term component, and SH is the Short-Term component and the other notations are as defined above. ϵ_t needs to be formulated using an ARMA model as mentioned above in the TF-Noise model. This model of one input can be extended to involve other decomposed inputs.

3.3 The Backshift Operator

In time series analysis, Backshift operator is one of the most important operators that are used extensively in the constructed models. This term, as the name implies,

shifts the time of the observation backward by, for example, one period. This can be represented as follows:

$$BY_t = Y_{t-1}, \text{ for example } \quad BY_{70} = Y_{69}.$$

Similarly, B^2 can be written as $B^2Y_t = Y_{t-2}$ and so on. Generally,

$$B^kY_t = Y_{t-k}, \text{for example}, B^{12}Y_{60} = Y_{48}.$$

There are different uses for the Backshift operator, for example, to represent the general stationarity transformation for the seasonal and nonseasonal series. While the nonseasonal operator is $\bigtriangledown = 1 - B$, the seasonal operator will be $\bigtriangledown_L = 1 - B^L$, L refers to the number of the seasons in a year (L =4 for the quarterly data and $L = 12$ for the monthly data). The general representation of stationarity transformation is

$$z_t = \bigtriangledown_L^D \bigtriangledown^d Y_t^*$$

$$= (1 - B^L)^D (1 - B)^d Y_t^*$$

where D refers to the degree of the differencing process for the seasonal series, d denotes the differencing of the nonseasonal series, and Y^* is the transformed version of the original series. This transformation can be, for example, taking the natural logarithm for the raw data.

3.4 TF-Noise Modelling for Poughkeepsie's City Raw Data

Figure A.3 in Appendix illustrates the steps taken to construct the developed models. The process of constructing a TF-Noise model requires applying three sequential steps. These steps are listed as the following:

1. An input series is often autocorrelated and using the direct sample cross correlation function (SCCF) between the input and the output series provides a misleading indication of the relation between them [92]. Some solutions have been suggested to handle this problem. One solution is called prewhitening where the prewhitened values will be used rather than the raw values. The prewhitened values can be obtained by identifying a tentative model for the inputs, which are temperature, precipitation, wind speed, tide, and groundwater level. The identification's mechanism depends on the behaviour of the Sample Autocorrelation Function (SACF) and Sample Partial Autocorrelation Function (SPACF) for the input series. Whenever the SPACF for any input

variable tends to cut-off after few lags and at the same time the SACF exhibits a dying down pattern, the adequate model for this input variable should be one of the Autoregressive models (AR). The fashion of the dying down would exhibit either an exponential or a sinusoidal distribution.

Moreover, the number of the parameters for the AR model should be associated with the number of the significant correlation coefficients along the lag axis. The number of the significant correlation coefficients is determined based on the spikes in the SPACF plot. In contrast, when the SPACF decreases with extremely slow exponential or sinusoidal signals and the SACF cuts-off fairly quickly with a few lags, one of the Moving Average Models (MA) should be specified to the considered time series. If the behaviour of these two correlation functions, which are SACF and SPACF, tends to be similar to a dying down pattern, the chosen model should be one of the Autoregressive Moving Average Models (ARMA).

The specification of an ARMA model for each input variable can be explained as follows:

- For the temperature series, which has been transformed to a stationary series by applying the first differences, and based on the patterns of the SACF and SPACF, a MA(3) model has been specified. This model can be defined as the following:

$$TE_t = (1 - 0.285B - 0.293B^2 - 0.144B^3)a_t. \tag{3.5}$$

- For the Wind Speed series, which has been also transformed to a stationary series by taking the first differences, the MA(3) model has fitted the data adequately. This model can be written as the following:

$$WS_t = (1 - 0.798B - 0.194B^2 + 0.023B^3)a_t. \tag{3.6}$$

- For the precipitation series, the MA(4) model has been suggested based on the behaviours of the SACF and SPACF, and this model is written as:

$$PR_t = (1 + 1.080B + 0.75B^2 + 1.070B^3 + 0.083B^4)a_t. \tag{3.7}$$

- For the tide series, which has been stationarised by taking the first differences, a AR(3) model has been applied and is written as follows:

$$TD_t = \frac{1}{(1 - 0.902B - 0.055B^2 + 0.353B^3)}a_t. \tag{3.8}$$

- For the groundwater level series, the first differences has been applied to stationarise the series. Relying on the values of the SACF and SPACF, an AR(3) model has adequately fit the data where this model can be defined as follows:

$$GW_t = \frac{1}{(1 + 0.337B + 0.039B^2 - 0.039B^3)} a_t. \tag{3.9}$$

2. All the previous models have been used to obtain the prewhitened values for the input and output series. Having created them, the SCCF values between the water discharge and each of the inputs will be computed. The SCCF values will be utilised to determine the terms for each input variable in the TF-Noise model. Based on this, the process of identifying an introductory structure for the TF model and estimating the parameters of this model is the next step. By using the conditional least squares method to obtain the parameters values, the preliminary model can be written as follows:

$$\begin{aligned} WD_t = \frac{(0.061 + 0.061B)}{(1 + 1B)} TE_t + \frac{(0.202 - 0.186B)}{(1 + 0.192B)} PR_t \\ + \frac{(0.028 + 0.033B)}{(1 + 0.083B)} WS_t + \frac{(0.100 - 0.010B)}{(1 + 1B)} TD_t + \\ \frac{(-0.019 - 0.016B)}{(1 + 0.158B)} GW_t + \epsilon_t \end{aligned} \tag{3.10}$$

where WD, TE, PR, WS, TD, GW, and ϵ_t denote the water discharge, temperature, precipitation, wind speed, tide, groundwater level and the error term, respectively.

3. Up to this point of estimating the parameters of the constructed model, we have built the preliminary TF model by using the inputs series. By examining the SACF and SPACF of the residual terms series of the preliminary model, an ARMA model would be suggested for this data. Depending on the patterns of the SACF, which is extremely slowly dying down with an exponential signal, and SPACF, which exhibits 1 spike at lag 1 and cuts-off after this spike, an ARMA (1,0) model has been selected to describe the behaviour of the noise data. Then the final TF-Noise model can be written as the following:

$$\begin{aligned} WD_t = \frac{(0.069 - 0.026B)}{(1 - 0.699B)} TE_t + \frac{(0.196 - 0.182B)}{(1 - 0.475B)} PR_t + \frac{(0.027 + 0.031B)}{(1 + 0.069B)} \\ WS_t + \frac{(0.091 + 0.001B)}{(1 + 0.209B)} TD_t + \frac{(-0.018 - 0.015B)}{(1 + 0.148B)} GW_t + \frac{1}{(1 - 0.066B)} a_t \end{aligned} \tag{3.11}$$

where all the previous notations have been defined in the model above, and a_t is a white noise with mean zero and constant variance, $sigma^2$ respectively.

To investigate whether applying the TF-Noise model to the decomposed data can yield better results than these obtained using the raw data, the same steps above have been followed for each component. Then the three components have been combined to build the CTF-Noise model structure. The next section shows the analysis of the decomposed data.

3.5 Combined TF-Noise Modelling for Poughkeepsie's City Decomposed Data

After we decomposed the data of all the studied series by using the KZ filtering mechanism with the parameters 29 days and 3 iterations, the result is three components with different scales, which are the long, seasonal, and short-term, for each variable. The following subsections present the analysis for these components.

3.5.1 TF-Noise Modelling for Poughkeepsie's City Long-Term Component

- Firstly, in order to identify an appropriate model for each variable of the input variables, the SACF and SPACF were calculated and plotted. The behaviour of the temperature series, which has been transformed to a stationary series by taking the second differences, would be described by using an autoregressive of order five model, which is written as AR(5) or ARMA(5,0). This model has been chosen because the pattern of the SACF dies down extremely slowly and the fashion of the SPACF cuts-off fairly quickly after lag 5. The AR(5) model for the temperature is shown in the following equation:

$$TE_{LT}(t) = \frac{1}{(1 - 0.953B - 0.147B^2 + 0.033B^3 + 0.044B^4 + 0.056B^5)} a_{LTTE}(t) \quad (3.12)$$

  where $TE_{LT}(t)$ denotes the Long Term component of the temperature, B is the Backshift operator, and $a_{LTTE}(t)$ is the random shock (white noise) of this model. By investigating the P-values for the parameters of this model, we have removed the last three parameters as they have non significant P-values. So, the model can be written as follows:

$$TE_{LT}(t) = \frac{1}{(1 - 0.953B - 0.147B^2)} a_{LTTE}(t). \quad (3.13)$$

The AR model has completely fitted the temperature data as shown in Figure 3.1, where the correlation analysis for the residuals of this model shows that there is no spike at any lag. Since we need to examine the relationship between

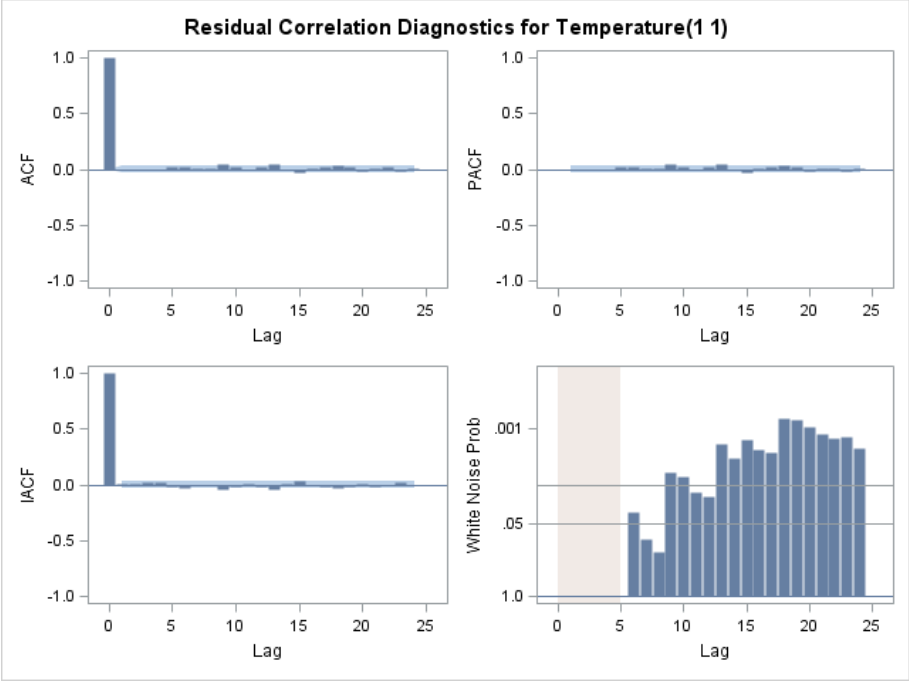

Figure 3.1: Residuals Correlation Diagnostics for the Temperature's Long-Term Component for Poughkeepsie's City.

the temperature and the water discharge series, the prewhitened values should be used instead of the raw data for both series. For this purpose, Equation 3.13 would be utilised to prewhiten the values of the temperature and the water discharge. Calculating the prewhitened values is necessary to enable computing the Sample Cross Correlation Function (SCCF) between the temperature and water discharge series. As a result, the decision of identifying an appropriate form for the TF model can then be taken more effectively and precisely.

Before we begin interpreting the SCCF, we need to check that there are no spikes (high SCCF values) at any negative lag. In case that there are such spikes, this would lead to an undesirable fact. The fact is that the past values of the output series, which is the water discharge, influence the future values of the input series, which is the temperature. No spikes appear in the negative side of the SCCF between the temperature and water discharge. By investi-

gating the pattern of the SCCF between the temperature and water discharge to determine the first spike, which in turn determines the value of b, it has appeared that this spike exists at lag 0, which in turns means that the effect of the temperature on the water discharge occurs on the same day. The value of b is the number of periods before the input series starts to influence the output series.

To identify the value of s, we have to examine the SCCF for any spikes between the first spike, which is the b value, and the starting of a clear dying down pattern. The value s is required to find the number of past values for the input variable that affect the output series. The clear pattern can have an exponential or a sinusoidal wave. For the temperature, this value will be set to 1, i.e $s = 1$, as we have one spike between the lag 0 and the starting point of the exponential pattern in this SCCF [15]. This implies that we need to use the operator $= (1 - \omega_1 B)$ for the numerator of the temperature in the TF model.

With regard to the denominator of the TF model, we need to determine the value r. This value represents the number of past values of the output series that affect on itself. Often, two choices are available. Either $r = 1$ which can be taken when an exponential pattern exists, or $r = 2$ when a sinusoidal fashion exists. For the temperature, it is ideally to select $r = 1$ as the clear dying down pattern in the SCCF has an exponential wave. So, the used operator will be $\delta B = (1 - \delta_1 B)$. To conclude, the temperature's term in the TF model for the long-term can be written as:

$$WD_{LT}(t) = \mu + \frac{C(1 - \omega_1 B)}{(1 - \delta_1 B)} B^0 TE_{LT}(t) + \epsilon_t \tag{3.14}$$

where WD_{LT} denotes the Long-Term of the water discharge, TE_{LT} is the Long Term of the temperature, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- Moreover, using the SACF and SPACF to identify a model for the wind speed, which has been transformed to a stationary series by using the first differences, we would describe the behaviour of this series by an AR(3) model. The estimation of the model's parameters has been implemented by the conditional least squares method as shown in Equation 3.15, and relying on the P-values the third term has been removed as this value was not significant.

$$WS_{LT}(t) = \frac{1}{(1 - 1.879B - 0.826B^2)} a_{LTWS}(t) \tag{3.15}$$

where WS_{LT} denotes the wind speed's long-term, B is the Backshift operator, and a_{LTWS} is the random shock (white noise) of this model. Investigating the SCCF values of this variable with the water discharge revealed that $b = 0$, which indicates that the effects of this variable on the WD appears at the same day (no lag). Additionally, the value of s is equal to 1, which means that we are led to tentatively choose the following model:

$$WD_{LT}(t) = \mu + C(1 - \omega_1 B)B^0 WS_{LT}(t) + \epsilon_t \tag{3.16}$$

where $WD_{LT}(t)$ denotes the water's discharge Long Term , WS_{LT} is the wind's speed Long Term, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- For the precipitation series, which has been transformed to a stationary series by applying the first differences, it is reasonable to consider the model AR(1,2,3,12) as there are spikes at lags 1,2,3, and 12 in the SPACF plot, where this model will be then used to prewhiten the values of the precipitation and the water discharge. The conditional least squares method has been utilised to estimate the parameters of the Autoregressive model, which can be written as follows:

$$PR_{LT}(t) = \frac{1}{(1 - 1.837B + 0.664B^2 + 0.183B^3 - 0.007B^{12})} a_{LTPR}(t) \tag{3.17}$$

where PR_{LT} denotes the precipitation's Long Term, B is the Backshift operator, and a_{LTPR} is the random shock (white noise) of this model. The SCCF values can provide the required information, which are the values of b, s, and r, to build the precipitation's term in the TF model of the long-term, where:

  - b=0 is the lag value for where the first spike has been seen. This would mean that the effect of the precipitation on the water discharge happens at the same time, as our data is daily so the effect occurs at the same day.
  - As long as we have encountered one spike (significant cross correlation coefficient) before the damped dying down pattern this would mean that s is equal to 1.
  - As a result for the damped exponential fashion, the value of r would be equal to 1.

  The precipitation's part in the TF model can be expressed as follows:

$$WD_{LT}(t) = \mu + \frac{C(1 - \omega_1 B)}{(1 - \delta_1 B)} B^0 PR_{LT}(t) + \epsilon_t \tag{3.18}$$

where WD_{LT} denotes the long-term of the water discharge, PR_{LT} is the precipitation's Long Term component, and ϵ_t is the error term.

- The tide series is also examined as an input variable, which is transformed to a stationary series by taking the second differences. The AR(3) model was the chosen model to represent this series, and this model can be written as follows:

$$TD_{LT}(t) = \frac{1}{(1 - 0.540B - 0.329B^2 - 0.125B^3)} a_{LTTD}(t) \tag{3.19}$$

where TD_{LT} denotes the tide's long-term, B is the Backshift operator, and a_{LTTD} is the random shock of this model. With regard to the form of the tide variable in the TF model, and because of $b = 0$ and $s = 1$, we will consider the following model:

$$WD_{LT}(t) = \mu + C(1 - \omega_1 B)B^0 TD_{LT}(t) + \epsilon_t \tag{3.20}$$

where $WD_{LT}(t)$ denotes the long-term of the water discharge, TD_{LT} is the long-term of the tide, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- For the groundwater level series, which has been stationarized by taking the second differences, the model that would adequately describe the data is the Autoregressive of order six, AR(6). The reason for selecting this model is attributed to the pattern of the SPACF, which cuts off after lag 6. The model is written as follows:

$$GW_{LT}(t) = \frac{1}{(1 - 0.988B - 0.291B^2 + 0.102B^5 + 0.105B^6)} a_{LTGW}(t) \tag{3.21}$$

where GW_{LT} denotes the groundwater's level long term, B is the Backshift operator, and a_{LTGW} is the random shock of this model. As they have non significant P-values, the third and fourth terms have been eliminated from the model. Additionally, in order to determine the term of the groundwater level in the TFM, we need to investigate the SCCF for the prewhitened values of this variable with the water discharge. According to the SCCF values, the first spike has also appeared at lag 0, which implies that the influence of the groundwater level on the water discharge would appear in the same day (no lag). The SCCF has a damped exponential pattern and the starting of this fashion begins after lag 1, which leads to select $s = 1$. That means we should use the operator $\omega(B) = (1 - \omega_1 B)$. Furthermore, as the pattern after lag 1 has been identified

as a damped exponential wave, it will be convenient to choose $r = 1$. These conclusions lead to tentatively consider the following TF model term for the groundwater level variable:

$$WD_{LT}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 GW_{LT}(t) + \epsilon_t \tag{3.22}$$

where WD_{LT} denotes the water discharge long-term, GW_{LT} is the long-term groundwater level, and ϵ_t is the error term.

- Secondly, according to the relations that have been captured by examining the SCCF patterns, the preliminary transfer function model can be defined as follows:

$$\begin{aligned} WD_{LT}(t) = \frac{(-13.83 - 18.244B)}{(1+0.181B)} TE_{LT}(t) + \frac{(0.058 - 0.019B)}{(1-0.917B)} PR_{LT}(t) \\ (0.311 + 0.456B) WS_{LT}(t) + (-41.498 - 43.984B) TD_{LT}(t) + \\ \frac{(-8.497 + 8.009B)}{(1+0.812B)} GW_{LT}(t) + \epsilon_t \end{aligned} \tag{3.23}$$

where $WD_{LT}, TE_{LT}, PR_{LT}, WS_{LT}, TD_{LT}, GW_{LT}$, and ϵ_t denote the long-term component for the studied variables.

- To complete our analysis using the TF-Noise model, the behaviour of the residual terms has to be examined and formalised using one of the ARMA models. The model AR(1) has adequately fit the data and the final structure of the TF-Noise model is written as follows:

$$\begin{aligned} WD_t = \frac{(5.572 - 0.917B)}{(1-0.256B)} TE_{LT}(t) + \frac{(0.204 - 0.148B)}{(1-0.774B)} PR_{LT}(t) \\ + (0.560 - 0.064B) WS_{LT}(t) + (-0.840 - 0.202B) TD_{LT}(t) + \\ \frac{(-0.255 + 0.244B)}{(1-0.946B)} GW_{LT}(t) + \\ \frac{1}{(0.560 - 0.064B)} a_{LT}(t). \end{aligned} \tag{3.24}$$

All the previous notations have been defined in the model above, and a_{LT} is a white noise series that has zero mean and constant variance, σ^2.

3.5.2 TF-Noise Modelling for Poughkeepsie's City Seasonal-Term Component

The TF-Noise model for the seasonal component can be constructed be applying the following steps.

- The First step is to obtain the prewhitened values for each variable. For temperature series, an ARMA(1,0) model is suitable to fit it based on the SACF and SPACF, and this model can be written as follows:

$$TE_{ST}(t) = \frac{1}{(1 - 0.562B)} a_{STTE}(t) \tag{3.25}$$

where TE_{ST} denotes the temperature's Seasonal-Term and a_{STTE} denotes the white noise series for this model. Equation 3.25 can be utilised to calculate the prewitened values for the temperature and the response variable. The SCCF between the temperature and water discharge has been calculated and used to determine the temperature's part in the TF model, which can be written as the following:

$$WD_{ST}(t) = \mu + \frac{C(1 - \omega_1 B)}{(1 - \delta_1 B)} B^0 TE_{ST}(t) + \epsilon_t \tag{3.26}$$

where WD_{ST} denotes the water's discharge Seasonal-Term, TE_{ST} is the temperature's Seasonal-Term component, and ϵ_t is the error term.

- The second examined predictor is the wind speed, therefore, having estimated the SACF and SPACF for the wind speed data, it would be more reasonable to build this data by using a MA(1) model, as shown in Equation 3.27

$$WS_{ST}(t) = (1 + 0.087B) a_{STWD}(t) \tag{3.27}$$

where WS_{ST} and a_{STWS} denote the Seasonal-Term of the wind speed and the white noise term, respectively. To determine the numerator and denominator factors for the wind speed's part in the TF model, the SCCF values between the wind speed and the water discharge will be examined to decide the appropriate operators. As a result, the term of the wind speed in the TF model can be written as follows:

$$WD_{ST}(t) = \mu + \frac{C(1 - \omega_1 B)}{(1 - \delta_1 B)} B^0 WS_{ST}(t) + \epsilon_t \tag{3.28}$$

where WD_{ST} denotes the water discharge Seasonal-Term, WS_{ST} is the wind speed Seasonal-Term, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- The precipitation series is the third analysed variable and the patterns of the SACF and SPACF suggested that the model MA(4) would adequately describe the precipitation data. Equation 3.29 shows the model for this predictor:

$$PR_{ST}(t) = (1 + 1.070B + 1.038B^2 + 1.025B^3 + 0.149B^4) a_{STPR}(t) \tag{3.29}$$

where $Precipitation_{ST}$ and a_{STPR} denote the precipitation's Seasonal-Term and the white noise series, respectively. Moreover, Equation 3.29 has also been used to obtain the prewitened values for the precipitation and the water discharge to calculate the SCCF. To determine the precipitation's portion in the TF model, we need to investigate the SCCF between the precipitation and the water discharge. Having examined the results of the SCCF, the next model can be chosen:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 PR_{ST}(t) + \epsilon_t \tag{3.30}$$

where WD_{ST} denotes the water discharge seasonal-term, PR_{ST} is the precipitation's seasonal Term, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- Moreover, according to the patterns of the SACF, which dies down extremely slowly, and the SPACF, which has five obvious spikes, the tide series can be represented by an AR(5) model, as shown in Equation 3.31

$$TD_{ST}(t) = \frac{1}{(1 - 1.427B + 0.275B^2 + 0.383B^3 + 0.145B^4 - 0.337B^5)} a_{STTD}(t) \tag{3.31}$$

where TD_{ST} and a_{STTD} denote the tide's seasonal-term and the white noise term, respectively. Equation 3.31 has been also used to compute the prewhitend values for the water discharge and the tide series. Inserting the suitable operators that have been determined by examining the SCCF values for the tide and the water discharge into the TF provides the next model:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 TD_{ST}(t) + \epsilon_t \tag{3.32}$$

where WD_{ST} denotes the Seasonal-Term of the water discharge, TD_{ST} is the Seasonal-Term component of the tide series, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- Finally, by investigating the behaviours of the SACF and SPACF for the groundwater, the model ARMA(2,0) is suggested to fit this time series data, as shown in Equation 3.33

$$GW_{ST}(t) = \frac{1}{(1 - 1.159B + 0.337B^2)} a_{STGW}(t) \tag{3.33}$$

where GW_{ST} and a_{STPR} denote the groundwater's level Seasonal-Term and the white noise term, respectively. Examining the behaviour of the SCCF between

the groundwater level and the water discharge leads to insert the following operators into the general equation:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 GW_{ST}(t) + \epsilon_t \tag{3.34}$$

where WD_{ST} denotes the water discharge seasonal-term, GW_{ST} is the groundwater level seasonal-term component, and ϵ_t is the error term, which has to be later substituted by one of the ARMA models.

- After we calculate the SCCF and identify the numerator and denominator polynomial factors for all the input series, the preliminary model for the seasonal component can be written as follows:

$$\begin{aligned} WD_{ST}(t) = {} & \frac{(0.005+0.004B)}{(1+0.991B)} TE_{ST}(t) + \frac{(0.529-0.382B)}{(1-0.954B)} PR_{ST}(t) \\ & + \frac{(0.058+0.031B)}{(1-0.753B)} WS_{ST}(t) + \frac{(0.104-0.052B)}{(1+0.686B)} TD_{ST}(t) + \\ & \frac{(-0.025-0.026B)}{(1-0.011B)} GW_{ST}(t) + \epsilon_t \end{aligned} \tag{3.35}$$

The notation ϵ_t is the error term for the model. Examining the residuals of this preliminary model reveals that an AR(1) model would adequately describe this series of the residuals, so, the final model, which combines the preliminary TF model and the AR(1) model, can be written as follows:

$$\begin{aligned} WD_{ST}(t) = {} & \frac{(-0.016-0.024B)}{(1+0.241B)} TE_{ST}(t) + \frac{(0.249+0.277B)}{(1+0.108B)} PR_{ST}(t) \\ & + \frac{(0.048-0.001B)}{(1-0.857B)} WS_{ST}(t) + \frac{(0.080-0.193B)}{(1+0.049B)} TD_{ST}(t) + \\ & \frac{(-0.094-0.027B)}{(1-0.350B)} GW_{ST}(t) + \frac{1}{(1-0.859B)} a_{ST}(t) \end{aligned} \tag{3.36}$$

where $a_{ST}(t)$ is a white noise series, which has zero mean and a constant variance. The final model, Equation 3.36 is adequate to represent the seasonal component water discharge for Poughkeepsie city, where this adequacy has been attributed to the pattern of the SACF and SPACF for the residuals for the final combined model. The model will also be used later to build the final forecasting expression by adding it to the long and short components.

3.5.3 TF-Noise Modelling for Poughkeepsie's City Short-Term Component

For the short-term component, the analysis can be summarised as follows. The values of the SACF and SPACF show that the ARMA model that would be specified to the temperature series is the ARMA(1,3). Having estimated the parameters of this model, it can be written as follows:

$$TE_{SH}(t) = \frac{(1 + 1.458B + 0.681B^2 + 0.163B^3)}{(1 + 0.903B)} a_{SHTE}(t) \tag{3.37}$$

where $Temperature_{SH}$ denotes the temperature's short-term component, and a_{STTE} denotes the error term of this model. The next step of studying the effect of this input series on the response variable is to calculate the SCCF between these two variables. Typically, to obtain significant information from the SCCF, prewhitened values should be used to compute this function. The prewhitened values for the temperature and the water discharge should be calculated by using the previous identified model, which is ARMA(1,3). Then, by examining the values of the SCCF, the numerator and denominator polynomial factors for the TF model of this input series can be identified. In this case, the temperature variable can be written with the factors(1,1), i.e s=1 and r=1.

Additionally, the other input variables should be treated by using the same manner. That means, we compute the SACF and SPACF to determine the model that can be utilised to prewitten the values of the input and output series, then we calculate the SCCF between each input variable and the output series. Thus, the preliminary model for this short-term component can be written as follows:

$$\begin{aligned} WD_{SH}(t) = {} & \frac{(0.037 - 0.020B)}{(1 - 0.935B)} TE_{SH}(t) + \frac{(0.479 - 0.287B)}{(1 - 0.902B)} PR_{SH}(t) \\ & + \frac{(0.037 - 0.061B)}{(1 - 0.922B)} WS_{SH}(t) + \frac{(0.007 - 0.009B)}{(1 - 1.963B)} GW_{SH}(t) + \epsilon_t \end{aligned} \tag{3.38}$$

To formalise our final model for the short-term component for the water discharge, the residuals of the preliminary model have been examined. The AR(4) model has been suggested as an appropriate model to fit the data. Therefore, the final model, which combines the preliminary and the residuals expression, can be written as fol-

lows:

$$WD_{SH}(t) = \frac{(0.049 + 0.043B)}{(1 + 0.908B)} TE_{SH}(t) + \frac{(0.292 + 0.217B)}{(1 - 0.118B)} PR_{SH}(t) + \frac{(0.047 - 0.071B)}{(1 - 0.822B)} WS_{SH}(t) + \frac{(0.011 + 0.003B)}{(1 - 0.951B)} GW_{SH}(t) + \frac{1}{(1 - 1.012B + 0.364B^2 - 0.193B^3 + 0.047B^4)} a_{SH}(t) \quad (3.39)$$

This model has been adequately fit this component as revealed by Figure 3.2.

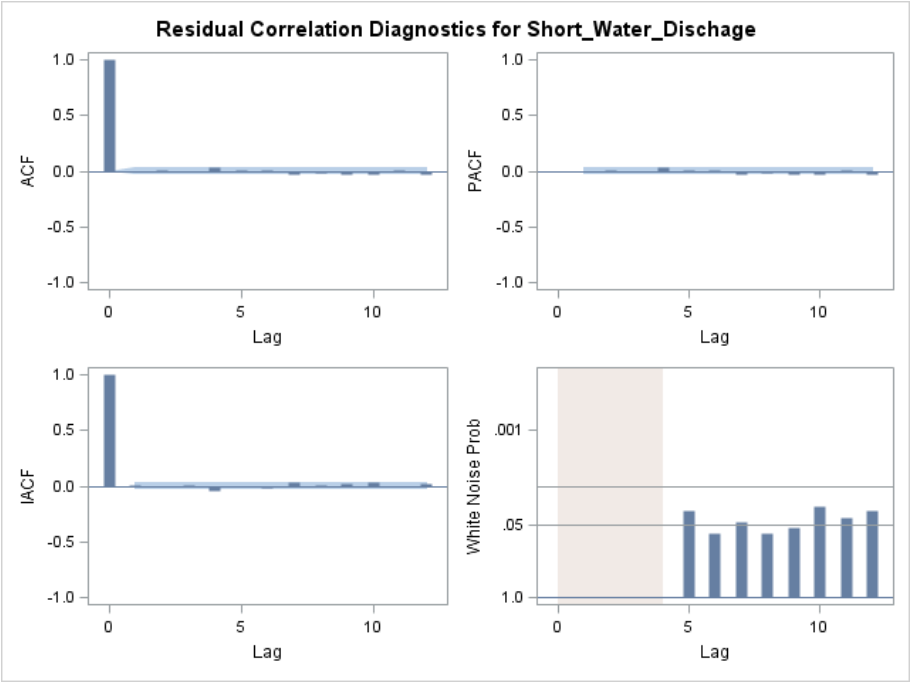

Figure 3.2: Residuals Correlation Diagnostics for the Water's Discharge Short-Term Component for Poughkeepsie's City.

3.6 The Final Combined TF-Noise Model for Poughkeepsie's City

Having constructed the three models of the three components, which are the long, seasonal, and short, the next step is to build the final TF-Noise model, (WD_{FI}), which combines these three patterns together. We have built the preliminary TF

model and then we examined the residuals correlation analysis for this model which reveals that specifying an AR model of order one would fit the noise's part of the model adequately. According to this specification, the final TF-Noise can be written as the following:

$$WD_{FI}(t) = \frac{(-6.705 + 7.125B)}{(1 + 0.624B)} PR_{FILT}(t) + \frac{(-15.922 + 15.190B)}{(1 + 0.610B)} GW_{FILT}(t) + \frac{(0.045 + 0.047B)}{(1 + 0.161B)} PR_{FISE}(t) + \frac{(-0.019 - 0.002B)}{(1 - 0.556B)} GW_{FISE}(t) + \frac{(0.146 + 0.133B)}{(1 + 0.048B)} PR_{FISH}(t) + \frac{(0.009 + 0.005B)}{(1 - 0.829B)} GW_{FISH}(t) + \frac{1}{(1 - 0.923B)} a_t \quad (3.40)$$

where $WD_{FI}, PR_{FILT}, GW_{FILT}, PR_{FISE}, GW_{FISE}, PR_{FISH}, GW_{FISH}$, and a_t denote the final water discharge, the long- term of the precipitation, groundwater level, and the seasonal component of the precipitation, groundwater level, and the short-term of precipitation, groundwater level, and a white noise series, respectively. It is obvious that the final structure of our model depends on the variables precipitation and groundwater level for the three components, long, seasonal, and short, and the others have no significant relation with the water discharge.

3.7 Evaluation of the Estimated Models for Poughkeepsie's City

To evaluate the obtained models, two different Goodness-of-fit statistics have been used. These statistics are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) or Schwarz Criterion (also known as SBC, SBIC). These statistics help with comparing the constructed models. For a specific dataset, the value of the Akaike's Information Criterion (AIC) has no meaning. However, this value becomes more interesting when we use it to compare the AIC of several competing models specified a priori. The output is shown in Table 3.1.

Typically, for any model obtained, Information Criteria (IC) includes the covariance matrix and the number of parameters, which will be used to compute the statistics that represent the information conveyed by the model by trying to balance the trade-off between the terms of a lack of fit and a penalty [8].
The AIC method provides an estimator for a measure of difference between the considered model and the 'true' model. The model that has the smallest AIC among several candidate models is considered as the best linear model. Akaike (1973) presented the

concept of IC as a statistical tool for conducting the task of model selection. This criteria is a function that is based on the sum of squared errors (SSE), the number of the parameters, k, and the number of the observations, n, where this measure of goodness of fit is defined as the following:

$$AIC = n \times ln\big[\frac{SSE}{n}\big] + 2k.$$

The following criteria is specifically designed for comparing regression models.

$$AIC = -2.lnL + 2 \times k,$$

where L is the likelihood function of the model. The criteria below is for comparing any models.

$$BIC = -2 \times lnL + 2 \times ln(nk)$$

By investigating the values of these statistics, which are shown in Table 3.1, it is obvious that the TF-Noise models built using the decomposed data are more accurate and as a result these models will be chosen to perform the forecasting process rather than TF-Noise models built using the raw data.

	AIC	SBC
TF-Noise for the Raw Data	254.878	358.529
TF-Noise for the Decomposed Data	138.373	260.322

Table 3.1: The Statistical Tests for the Model Selection for Poughkeepsie City.

3.8 Combined TF-Noise Modelling for Cohoes' City Decomposed Data

3.8.1 TF-Noise Modelling for Cohoes' City Long-Term Component

By examining the SACF and SPACF for the long-term component of the temperature variable for Cohoes city, it can be verified that an ARMA(4,0) is the suitable model to describe this data. The fashion of the former function dies down extremely slowly and the pattern of the latter function cuts off fairly quickly after the lag 4. Moreover, an estimated least square points have been obtained for the parameters of the AR(4) model, which are ϕ_1, ϕ_2, ϕ_3, and ϕ_4. Hence, the model can be written as follows:

$$TE_{LT}(t) = \frac{1}{(1 - 3.003B + 3.053B^2 - 1.092B^3 + 0.042B^4)} a_{LT}(t). \tag{3.41}$$

Because we need to study the relationship between the independent variable, temperature, and the dependent variable, water discharge, the Sample Cross Correlation Function (SCCF) should be computed. Typically, to find this type of correlation, prewhitening the values of the variables that will be used is an important step. Therefore, we will use the previous constructed model, AR(4), for the temperature variable to prewhiten its original values and the water discharge values. Because of the spikes that appear in the negative side, it seems that it is not permitted to include any lag for the temperature variable in our model.

Similarly, the previous steps can be applied to the other predictors, which are precipitation, groundwater level, wind speed, and tide, the results were as follows: An AR(3,0) is an adequate model to describe the behaviour of the precipitation series, as the SPACF has three obvious spikes and the SACF has a pattern that is exponentially dying down with a very slow movement. The estimated parameters can be listed in an AR(3) model as follows:

$$PR_{LT}(t) = \frac{1}{(1 - 2.979B + 2.973B^2 - 0.993B^3)} a_{LT}(t). \tag{3.42}$$

This model will be utilized to compute the prewhitened values that will enable us to calculate the SCCF values between the precipitation and water discharge variables. It is not possible to use any lag for the precipitation variable because there are two spikes in the negative side, which means that the past values for the water discharge (dependent variable) will affect the future values for the precipitation (independent variable) values. Furthermore, the groundwater level should be expressed by the model AR(1). So, the model will be written as follows:

$$GW_{LT}(t) = \frac{1}{(1 - 1B)} a_{LT}(t). \tag{3.43}$$

After we have investigated the SCCF between the groundwater's level and water discharge and because it has spikes in the negative side, it is not possible to use any lag for the groundwater as predictors in our model. The investigation of the SACF and SPACF of the tide variable has shown that this predictor can be represented by using an AR(2) model. Since it is not acceptable that we include the lags of the variables that have spikes in the negative side in SCCF, lags for the tide will not contribute in the constructed model.

Using the same steps, the SCCF for the wind speed variable can be analysed. Thus, the preliminary model for this component will be as follows:

$$\begin{aligned} WD_t = -0.475TE_{LT}(t) + 0.317PR_{LT}(t) - 0.479GW_{lt}(t) + 0.265TD_{LT}(t) \\ - 0.080WS_{Lt}(t) + \epsilon_t. \end{aligned} \tag{3.44}$$

Examining the SACF and SPACF of the residuals for the preliminary model shows the possibility of defining this series with AR(4,0) model. Hence, the final model for this long-term component, which combines the preliminary model and the residuals model, will be written as follows:

$$WD_{LT}(t) = 0.125TE_{LT}(t) + 0.137PR_{LT}(t) - 0.984GW_{Lt}(t) + 0.104TD_{LT}(t) + \frac{1}{1 - 2.429B + 1.311B^2 + 0.678B^3 - 0.560B^4} a_{LT}(t) \quad (3.45)$$

where $WD_t, TE_{LT}, PR_{LT}, GW_{Lt}, TD_{LT}, a_{LT}$ denote the water discharge, temperature, precipitation, groundwater, tide, and the error term, respectively. Error term has a normal distribution with zero mean and constant variance, σ^2.

3.8.2 TF-Noise Modelling for Cohoes' City Seasonal-Term Component

By investigating the SACF and SPACF for the temperature variable, it can be verified that, as the pattern of the SACF series dies down with a sinusoidal wave and the pattern of the SPACF cuts off fairly quickly after the lag 1, an ARMA(1,0) is an adequate model to describe this data. Moreover, an estimated least square point has been obtained for the parameter of the AR(1) model, which is ϕ_1, so, we can write the model as follows:

$$TE_{ST}(t) = \frac{1}{(1 - 0.533B)} a_{ST}(t). \quad (3.46)$$

Because we need to study the relationship between the temperature variable (independent), and the water discharge (dependent) variable, the SCCF, should be computed. We will use the previous constructed AR(1) model for temperature to prewhiten the temperature and the water discharge values. By investigating the SCCF, it seems that the first spike appears at lag 1, which means that the temperature will enter the forecasting model with the variable of lag 1. For the nominator of the temperature's term, value 2 will be used as there are two spikes between the first lag and the general pattern that follows the first lag. On the other hand, as long as the pattern for the series follows a sinusoidal fashion, the value 2 will be specified for the denominator. Similarly, the previous steps can be applied to the other predictors, which are precipitation, groundwater, wind speed, and tide, the results were as follows:

An autoregressive model with two parameters for the lags 1 and 5 is an adequate model to describe the behaviour of the precipitation series, as the SPACF has two obvious spikes and the SACF has a pattern which is dying down. The estimated parameters can be listed in the AR model as follows:

$$PR_{ST}(t) = \frac{1}{(1 - 0.718B + 0.053B^5)} a_{ST}(t). \tag{3.47}$$

This model will be utilized to compute the prewhitened values that will enable us to calculate the SCCF between the precipitation and water discharge. It is not possible to use any lag for the precipitation variable because there are three obvious spikes in the negative side, which means that past values for the water discharge (dependent) will affect the future values of the precipitation (independent) values. Furthermore, the groundwater level variable should be expressed by an AR(2) model as the SACF and SPACF revealed the possibility of using this autoregressive model of order 2. So, the model will be written as follows:

$$GW_{ST}(t) = \frac{1}{(1 - 1.433B + 0.509B^2)} a_{St}(t). \tag{3.48}$$

After we have investigated the SCCF for the groundwater's level and because it has spikes in the negative side, so, it is not possible to use any lag for the groundwater level as predictors in our model. Furthermore, investigating of the SACF and SPACF for the tide variable has shown that this predictor can be represented by using the model of AR(2). By estimating the parameters of the AR model of order two, the resultant equation will be written as follows:

$$TD_{ST}(t) = \frac{1}{(1 - 1.501B + 0.726B^2)} a_{ST}(t). \tag{3.49}$$

Furthermore, since it is not reasonable that we include lags for the variables that have spikes in the negative side in SCCF, lags for the tide variable will not enter in the forecasting model.

The last predictor is the wind speed. The SACF and SPACF for the wind speed refer to the ability of expressing this variable by using an Autoregressive Moving Average model, ARMA(2,2). And the estimated values for the parameters of the ARMA(2,2) model will be listed in the following equation:

$$WS_{St}(t) = \frac{(1 - 0.653B - 0.334B^2)}{(1 - 0.776B - 0.062B^2)} a_{ST}(t). \tag{3.50}$$

Again, to decide which lag will be used in the prediction model, the SCCF will be computed after obtaining the prewhitened values. The SCCF shows that lag 1, where the first spike has occurred, will be used in the forecasting model. Hence, the final preliminary model for the seasonal component can be written as follows:

$$WD_{ST}(t) = \frac{(0.083 + 0.120B + 0.055B^2)}{(1 + 0.192B - 0.349B^2)} TE_{t-1} + \frac{(0.073 + 0.069B)}{(1 + 0.590B - 0.408B^2)} WS_{t-1} + 0.461PR_t + \epsilon_t \quad (3.51)$$

where WD, TE, WS, and PR denote the water discharge, temperature, wind speed, and precipitation, respectively.

After we have constructed the preliminary model by using the input variables, which are temperature, precipitation, and wind speed, we need to determine an adequate ARMA model for the residuals that have been calculated using the preliminary model. The suggested model can be characterised by using the SACF and SPACF of the residuals. Since the SACF of the residuals values is dying down with a sinusoidal wave and the SPACF has two spikes at the lags 1 and 4 and cuts off fairly quickly after lag 4, AR with order 1 and 4 could be selected to describe this residuals series. Therefore, the final constructed model will contain the preliminary model as well as the AR model for the residuals. Mathematically, we can write the constructed model as follows:

$$WD_{ST}(t) = \frac{(0.083 + 0.120B + 0.055B^2)}{(1 + 0.192B - 0.349B^2)} TE_{t-1} + \frac{(0.073 + 0.069B)}{(1 + 0.590B - 0.408B^2)} WS_{t-1} + 0.461PR_t + \frac{1}{(1 - 0.837B + 0.071B^2)} a_{ST}(t) \quad (3.52)$$

where $a_{ST}(t)$ is a white noise series with zero mean and constant variance, σ^2.

3.8.3 TF-Noise Modelling for Cohoes' City Short-Term Component

Depending on the results of the SACF and SPACF for the studied variables, we have the following results:

- For the temperature, an ARMA(2,3) model can represent the data of this variable and the model is written as follows:

$$TE_{SH}(t) = \frac{(1 - 1.091B - 0.038B^2 + 0.138B^3)}{(1 - 1.583B + 0.667B^2)} a_{SH}(t). \quad (3.53)$$

- For the precipitation, ARMA(1,3) model is selected to describe the data of this variable, and the model can be written as follows:

$$PR_{SH}(t) = \frac{(1 + 0.88259B + 0.830B^2 + 0.764B^3)}{(1 - 0.050B)} a_{SH}(t). \tag{3.54}$$

- For the groundwater level, the model ARMA(1,1) is adequate to fit the data of this variable, and this model can be written as follows:

$$GW_{SH}(t) = \frac{(1 - 0.570B)}{(1 - 0.788B)} a_{SH}(t). \tag{3.55}$$

- For the tide variable, the chosen model, which is ARMA(3,0), can be written as follows:

$$TD_{SH}(t) = \frac{1}{(1 - 1.378B + 0.343B^2 + 0.299B^3)} a_{SH}(t). \tag{3.56}$$

- Finally, for the wind speed variable, ARMA(2,0) is suitable to describe the behaviour of this series, this model can be written as follows:

$$WS_{SH}(t) = \frac{1}{(1 - 0.180B + 0.107B^2)} a_{SH}(t). \tag{3.57}$$

- So, depending on the previous constructed models, prewhitening process for all the variables with the water discharge series has been performed, and the preliminary transfer function model is written as follows:

$$WD_{SH}(t) = \frac{-0.016 + 0.094B}{(1 - 0.567B)} TE_t + 0.408PR_t - 0.696GW_t + \frac{0.042 + 0.054B}{1 - 0.910B} WS_t + \epsilon_(t) \tag{3.58}$$

  where WD, TE, PR, GW, WS, and $\epsilon_(t)$ denote the water discharge, temperature, groundwater level, wind speed, and error term, respectively.

- To model the final prediction equation, the residuals from the preliminary model should be fitted to one of the ARMA models. Therefore, it is possible to fit them with an ARMA(1,0) model. The final model can be written as follows:

$$WD_{SH}t = \frac{-0.012 + 0.078B}{(1 - 0.719B)} TE_t + 0.195PR_t - 0.829GW_t + \frac{1}{1 - 0.772B} a_{SH}(t). \tag{3.59}$$

3.9 TF-Noise Modelling for Utica's City Raw Data

To investigate the effect of the TF-Noise model on Utica's city data, we follow the same steps that are applied to analyse Poughkeepsie's city data. In this case the amount of water discharge will be estimated for a station related to Mohawk River. Firstly, we will perform the analysis using the raw data. The water discharge series is not stationary as it is clear from Figure 3.3. This series has been stationarised

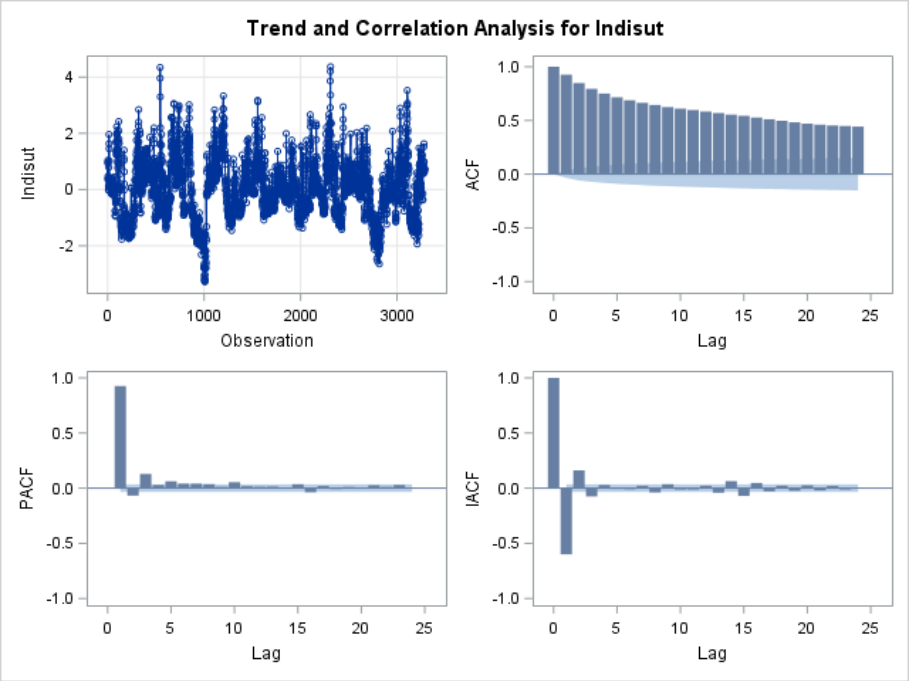

Figure 3.3: Diagnostic Plots for Water Discharge.

using the first differences. To build a TF model, we have begun our analysis by examining the input variables, which are temperature, wind speed, precipitation, tide, and groundwater level. For temperature, the first differences have been applied to obtain a stationary series. Based on the SACF and SPACF, a MA(4) model has been specified as we have four obvious spikes in the SACF and the pattern of the SPACF decreases fairly slowly. Then, the wind speed variable has been also transformed to a stationary series by applying the first differences. Relying on the behaviour of the SACF and SPACF, the MA(3) model has been assigned to the wind speed. The series of precipitation is stationary, so, we do not need to apply any differencing process. Based on the SACF and SPACF patterns, a MA(4) model has been suggested to fit the data of this variable.

With regard to the tide variable, the first differences have been taken to stationarize the tide series. An AR(3) model has been chosen to fit the data of this series. The final input variable that has to be examined is the groundwater variable, which has been stationarised by using the first differences. An AR(3) model has been applied to model the data of the groundwater level variable. Spikes have been noted in the negative side for the last two variables, tide and groundwater level, this has led to not incorporate them in the TF structure.

The following two models are the introductory and the final models with an AR(1) model for the residuals for the raw data.

$$WD_t = \frac{(0.022 + 0.101B)}{(1 - 0.633B)} TE_t + \frac{(0.166 - 0.155B)}{(1 - 0.295B)} PR_t + \frac{(0.041 + 0.042B)}{(1 + 0.080B)} WS_t + \epsilon_t \quad (3.60)$$

$$WD_t = \frac{(0.019 + 0.103B)}{(1 - 0.627B)} TE_t + \frac{(0.169 - 0.158B)}{(1 - 0.279B)} PR_t + \frac{(0.041 + 0.042B)}{(1 + 0.092B)} WS_t + \frac{1}{(1 + 0.050B)} a_t. \quad (3.61)$$

To examine the performance of the TF-Noise on the decomposed data, we have applied the same steps above to the three components as shown in the following subsections.

3.9.1 TF-Noise Modelling for Utica's City Long-Term Component

As it has been previously mentioned, calculating the functions of the SACF and SPACF can provide us with the required information for building the models for the studied variables. Relying on these two correlation functions, the results can be listed as the following:

- For temperature series, which has been stationarised by taking the second differences, the ARMA(2,0) model would reasonably represent the data of this variable, hence, this model can be written as follows:

$$TE_{LT}(t) = \frac{1}{(1 - 0.230B - 0.512B^2)} a_{LTTE}(t) \quad (3.62)$$

  where TE_{LT} is the temperature's long-term, and a_{LTTE} is a white noise series with mean of 0 and a constant variance, σ^2. With regard to the SCCF, the first

spike has occurred at lag 0, which means that $b = 0$. Also, further examination for the SCCF values has revealed that the numerator's operator of temperature in the TF should be equal to 1, i.e $s = 1$. Since the SCCF dies down in a damped exponential wave pattern, $r = 1$ should be assigned to the operator of water discharge variable.

$$WD_{LT}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 TE_{LT}(t) + \epsilon_t \tag{3.63}$$

where WD_{LT} and $TE_{LT}(t)$ are the water discharge and temperature Long Term series, respectively. The ϵ_t is the error term of this model that should be substituted later by one of the ARMA models.

- For the wind speed, the ARMA(2,0) model is selected to describe the data of this variable, and the model can be written as follows:

$$WS_{LT}(t) = \frac{1}{(1-0.482B-0.412B^2)} a_{LTWS}(t) \tag{3.64}$$

where WS_{LT} and a_{LTWS} denote the long-term of wind speed and a white noise series, respectively. The ability of this autoregressive model to describe the data of wind speed can be statistically checked by using the correlation analysis for the residuals. As shown in Figure 3.4, this model has successfully represented this data as long as there are no spikes in the SACF and SPACF. This would mean that there is no further information can be extracted from this data. With regard to the operators of the these two variables in the TF model, the next equation, Equation 3.65, can identify them:

$$WD_{LT}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 WS_{LT}(t) + \epsilon_t \tag{3.65}$$

where WD_{LT} and WS_{LT} denote the Long Terms for water discharge and wind speed series, respectively. The ϵ_t is the error term of this TF, which needs to be modelled by one of the Box-Jenkins Models, (ARMA) models.

- For precipitation, the model ARMA(5,0) is adequate to fit the data of this variable, and this model can be written as follows:

$$PR(t) = \frac{1}{(1-1.131B-0.141B^2+0.230B^3+0.176B^4-0.091B^5)} a_{LTPR}(t) \tag{3.66}$$

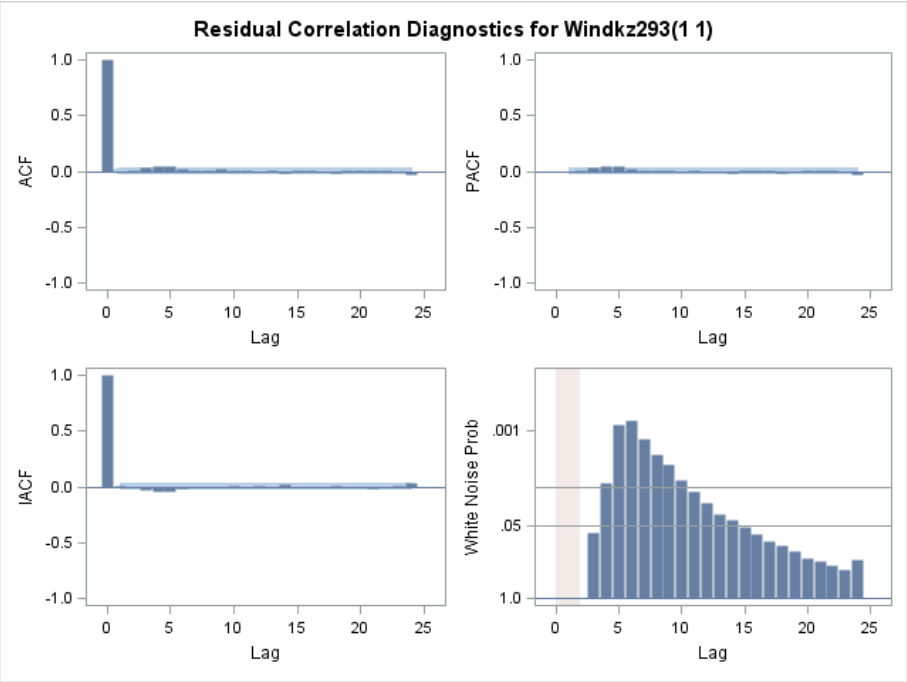

Figure 3.4: Cross Correlation of Water Discharge and Wind Speed of the Long-Term Component for Utica City.

where $PR(t)$ denote the long-term of the precipitation series, and a_{LTPR} is a white noise series. Examining the pattern of the SCCF between precipitation and water discharge reveals that the next form can be used for the part of precipitation in the TF:

$$WD_{LT}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 PR_{LT}(t) + \epsilon_t \tag{3.67}$$

where WD_{LT} is the long-term series of water discharge. The PR_{LT} is the long-term of precipitation series and ϵ_t is the error term of this equation.

- For tide series, which has been also transformed to a stationary series by applying the third differences, the model ARMA(2,2) was suggested to fit the data of this variable, and this model can be written as follows:

$$Tide_{LT}(t) = \frac{(1-1.692B+0.739B^2)}{(1-1.070B+0.084B^2)} a_{LTTD}(t) \tag{3.68}$$

where $Tide_{LT}$ and a_{LTTD} are the long-term of the tide and a white noise series, respectively. Using the SCCF values of water discharge and tide, the following

TF, Equation 3.69, can be applied. As shown in Figure 6.3, this equation can adequately identify the part of tide in the TF model.

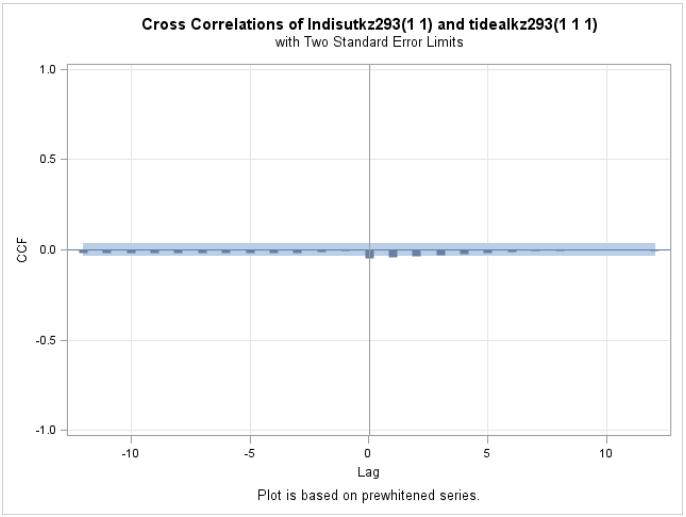

Figure 3.5: Cross Correlation of Water Discharge and Tide of the Long-Term Component for Utica City.

$$WD_{LT}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 TD_{LT}(t) + \epsilon_t \tag{3.69}$$

where $= WD_{LT}$ denotes the long-term of water discharge, TD_{LT} is the long-term of tide, and ϵ_t is the error term.

- For groundwater level, which has been converted to a stationary series by applying second differences, the model of ARMA(2,2) would be suitable to fit the data of this variable, and this model can be written as follows:

$$GW_{LT}(t) = \frac{(1 - 1.673B + 0.753B^2)}{(1 - 1.036B + 0.100B^2)} a_{LTGW}(t) \tag{3.70}$$

where GW_{LT} is the series of the Long Term of groundwater level. Also, a_{LTGW} is a white noise series. The values s=1 and r=1 would be assigned to the numerator and the denominator for the groundwater level part in the TF model.

- Therefore, by using the conditional least squares method to estimate the parameters of the input variables, the preliminary TF model can be written as

follows:

$$WD_{LT} = \frac{(0.859 - 0.112B)}{(1 + 0.551B)} TE_{LT}(t) + \frac{(-0.062 + 0.297B)}{(1 + 0.625B)} PR_{LT}(t) + \frac{(0.05624 + 0.056B)}{(1 - 1B)} WS_{LT}(t) + \frac{(0.961 - 1.465B)}{(1 - 0.819B)} TD_{LT}(t) + \frac{(0.819 - 2.066B)}{(1 - 0.954B)} GW_{LT}(t) + \epsilon_{LT}(t) \quad (3.71)$$

where WD_{LT}, TE_{LT}, PR_{LT}, WS_{LT}, TD_{LT}, GW_{LT}, and the ϵ_t denote the Long-Term components of water discharge, temperature, precipitation, groundwater level, tide, wind speed, and the error term, respectively.

- Then, to build the final structure for the TF model, the residuals of the preliminary model should be investigated and fitted to one of the Box-Jenkins models, ARMA models. The model ARMA(1,0) has been suggested to describe the data of these residuals. So, combining this model with the preliminary structure produces the full TF-Noise model, which is shown in the following expression:

$$WD_{LT} = \frac{(0.103 + 0.051B)}{(1 - 0.819B)} TE_{LT}(t) + \frac{(0.084 - 0.067B)}{(1 - 0.934B)} PR_{LT}(t) + \frac{(0.014 - 0.003B)}{(1 - 0.973B)} WS_{LT}(t) + \frac{(-1.385 - 0.848B)}{(1 - 0.691B)} TD_{LT}(t) + \frac{(-0.231 - 0.669B)}{(1 - 0.971B)} GW_{LT}(t) + \frac{1}{(1 - 0.999B)} a_t \quad (3.72)$$

where $a_{LT}(t)$ is a white noise series that has zero mean and a constant variance, σ^2.

3.9.2 TF-Noise Modelling for Utica's City Seasonal-Term Component

Depending on the results of the SACF and SPACF of the seasonal components for the considered variables, we extracted the following results:

- For the temperature, the AR(1,2,3,6,7,8) model would suitably fit the data of this variable and the model can be written as follows:

$$TE_{ST}(t) = \frac{1}{(1 - 0.678B + 0.226B^2 - 0.141B^3 - 0.043B^6 - 0.008B^7 + 0.153B^8)} a_{STTE}(t) \quad (3.73)$$

where TE_{ST} is the seasonal series of temperature, and a_{STTE} is a white noise series. Computing the SCCF between the two variables of temperature and water discharge reveals that there is a delay time, b=1, as the first spike has emerged at lag 1, and because of the presence of one spike at lag 2, the value of s equals to 1, and the r value equals to 1. Consequently, the appropriate part for the temperature variable in the TF can be written as follows:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^1 TE_{SE}(t) + \epsilon_{LT}(t) \tag{3.74}$$

where WD_{ST} denotes the seasonal-term of water discharge series and ϵ_{LT} is the error series.

- For the wind speed, the ARMA(2,2) model has been selected to describe the data of this variable, and the model can be written as follows:

$$WS_{ST}(t) = \frac{(1-0.662B-0.335B^2)}{(1-0.966B+0.045B^2)} a_{STWS}(t) \tag{3.75}$$

where WS_{ST}, denotes the seasonal-term of wind speed, and a_{STWS} is a white noise series. Based on the SCCF values, no delay time exists, i.e. $b = 0$, and the number of spikes between the first spike and the beginning of the dying down fashion is 1, this suggested that $s = 1$. Also, the dying down exponential pattern implies that the value of r for the wind speed variable in the TF model should be 1. Mathematically, this can be expressed as shown in Equation 3.76.

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 WS_{SE}(t) + \epsilon_t \tag{3.76}$$

where WD_{SE} is the Seasonal Term series of water discharge, and ϵ_t is the error term.

- For the precipitation, the adequate model is a moving average model with four consecutive lags, which are 1,2,3, and 4, i.e MA(4). This model can be written as follows:

$$PR_{ST}(t) = (1 + 1.135B + 1.1B^2 + 1.088B^3 + 0.143B^4) a_{STPR}(t) \tag{3.77}$$

where $Precipitation_{ST}$ is the seasonal-term series of precipitation, and a_{STPR} is a white noise series. Also, depending on the values of the SCCF, the value of b, which is the delay time, equals to 1, and the number of spikes that appear after this delay time is 1, i.e $s = 1$, and r is also equal to 1. Therefore, the precipitation portion in the TF model can be written as follows:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^1 PR_{SE}(t) + \epsilon_t \tag{3.78}$$

where WD_{ST} is the seasonal series of water discharge, and ϵ_t is the error term.

- For tide series, an autoregressive model with five lags has been determined for this variable, and this model can be written as follows:

$$TD_{ST}(t) = \frac{1}{(1 - 1.459B + 0.321B^2 + 0.371B^3 + 0.132B^4 - 0.329B^5)} a_{STTD}(t) \tag{3.79}$$

where TD_{ST} denotes the seasonal-term series for tide, and a_{STTD} is a white noise series. No delay time would be included and the number of spikes between the first spike and the dying down pattern is 1, so $s = 1$ and r is also 1, where the pattern has an exponential wave. This information can be illustrated in the following Equation:

$$WD_{ST}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 TD_{SE}(t) + \epsilon_t \tag{3.80}$$

where WD_{ST} is the seasonal-term series of water discharge, and ϵ_t is the error term.

- For the groundwater level, an autoregressive model with two lags, 1 and 2, has been chosen for this variable, and this model can be written as follows:

$$GW_{ST}(t) = \frac{1}{(1 - 1.493B + 0.529B^2)} a_{STGW}(t) \tag{3.81}$$

where GW_{ST} denotes the seasonal series of groundwater's level. Also, a_{STGW} is a white noise series. If we investigate the SCCF values of water discharge and groundwater level we can use $b = 0$, and the operators $= (1 - \omega_1 B)$ and $\delta B = (1 - \delta_1 B)$, for the numerator and denominator, respectively. This is shown in the next figure:

- By substituting all the aforementioned variables in the structure of the TF model, the preliminary model is built as it is shown in the following equation:

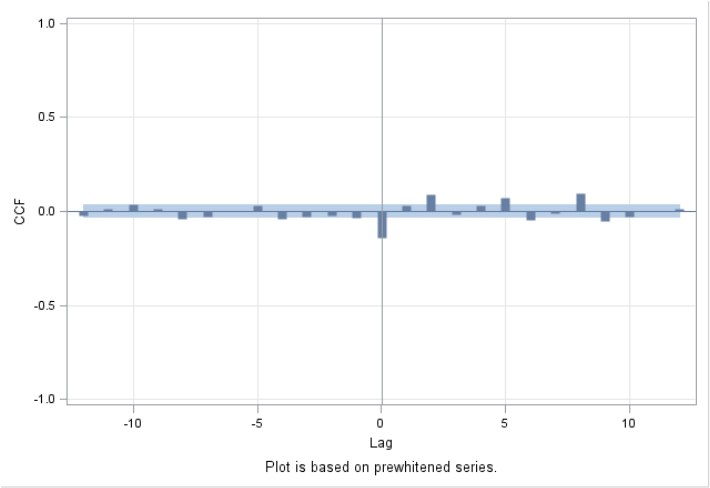

Figure 3.6: The Cross Correlation for the Water Discharge and Groundwater Level of the Seasonal-Term Component for Utica City.

$$WD_{ST}(t) = \frac{(0.111 + 0.022B)}{(1 - 0.495B)} B^1 TE_{ST}(t) + \frac{(0.381 + 0.263B)}{(1 + 0.991B)} B^1 PR_{ST}(t) + \frac{(0.136 - 0.027B)}{(1 - 0.791B)} WS_{ST} + \frac{(0.182 - 0.175B)}{(1 - 0.415B)} TD_{ST}(t) + \frac{(-1.044 + 0.738B)}{(1 - 0.332B)} GW_{ST}(t) + \epsilon_{ST} \quad (3.82)$$

where WD_{ST}, TE_{ST}, PR_{ST}, WS_{ST}, TD_{ST}, GW_{ST}, and ϵ_t denote the water discharge, temperature, precipitation, wind speed, tide, groundwater level, and the error term, respectively.

- Then, to build the final structure for the TF model of this seasonal component, the residuals of the preliminary model should be investigated and fitted to one of the Box-Jenkins models. The model ARMA(0,1) has been suggested to describe the data of these residuals. So, combining the ARMA(0,1) model with the preliminary structure produces the final TF-Noise model, which is shown in the following expression:

$$WD_{ST}(t) = \frac{(0.113 + 0.025B)}{(1 - 0.664B)} B^1 TE_{ST}(t) + \frac{(0.143 - 0.085B)}{(1 - 0.405B)} B^1 PR_{ST}(t) + \frac{(0.112 + 0.052B)}{(1 - 0.209B)} WS_{ST} + \frac{(0.065 - 0.087B)}{(1 - 0.415B)} TD_{ST}(t) + \frac{(-0.651 + 0.548B)}{(1 - 0.696B + 0.162B^2)} GW_{ST}(t) + \frac{1}{(1 - 0.796B)} a_{ST}(t). \quad (3.83)$$

This model has confirmed its validity where the correlation analysis that is based on the SACF and SPACF, for the residuals reveals that no spikes appear.

3.9.3 TF-Noise Modelling for Utica's City Short-Term Component

For the last component, which is the short-term, and relying on the functions of the SACF and SPACF, the results can be summarised as the following:

- For temperature series, the ARMA(2,3) model would sensibly fit the data of this variable and the model is written as follows:

$$TE_{SH}(t) = \frac{(1 - 0.992B - 0.166B^2 + 0.162B^3)}{(1 - 1.581B + 0.633B^2)} a_{SHTE}(t) \quad (3.84)$$

  where TE_{SH} denotes the short-term of the temperature series. Also, a_{SHTE} is a white noise series. This model has verified its ability to describe the data as shown in Figure 3.7, where there are no spikes at any lag. Examining the SCCF between the water discharge and temperature series reveals that the suitable form for this variable in the TF model can be written as the following:

$$WD_{SH}(t) = \mu + \frac{C(1 - \omega_1 B)}{(1 - \delta_1 B)} B^0 TE_{SH}(t) + \epsilon_t \quad (3.85)$$

  where WD_{SH} is the short-term component of water discharge series, and ϵ_t is the error term.

- For the wind speed, ARMA(1,2) model has been selected to describe the data of this variable and can be written as follows:

$$WS_{SH}(t) = \frac{(1 - 0.747B - 0.252B^2)}{(1 - 0.987B)} a_{SHWS}(t) \quad (3.86)$$

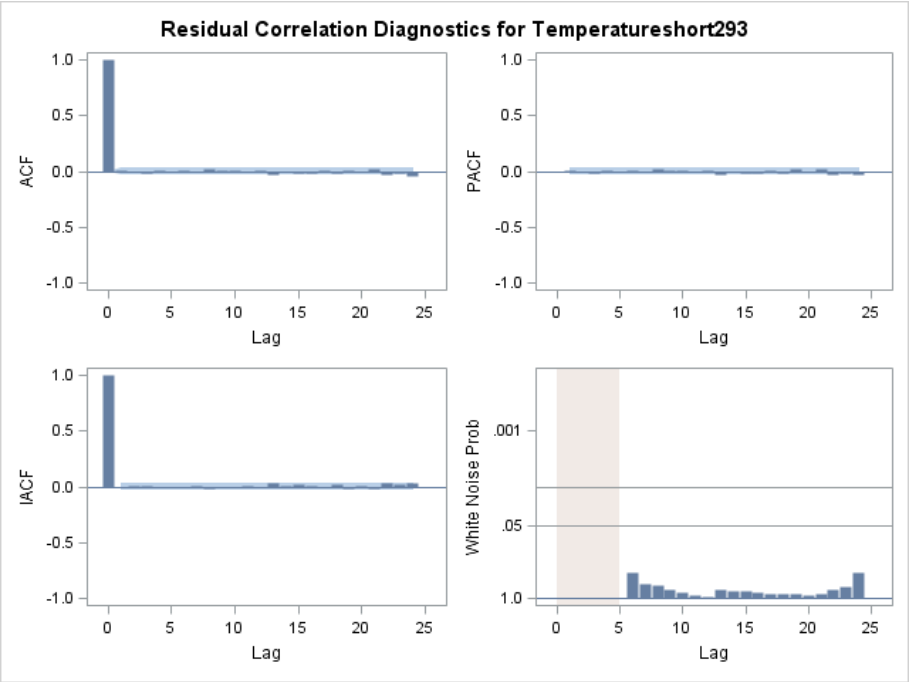

Figure 3.7: Residuals Correlation Diagnostics for the Temperature's Short-Term Component for Utica City.

where WS_{SH} denotes the short-term component of wind speed, and a_{SHWS} is a white noise series. The SCCF can identify the part of wind speed in the TF model, which can be described as follows:

$$WD_{SH}(t) = \mu + \frac{C(1-\omega_1 B)}{(1-\delta_1 B)} B^0 WS_{SH}(t) + \epsilon_t. \tag{3.87}$$

- For precipitation, investigating the SACF and SPACF for this variable provided us with the result of possibility of specifying an ARMA(4,0) model for the input variable as shown in Equation 3.88. The validity of this model has been confirmed using the correlation analysis for the residuals.

$$PR_{SH}(t) = \frac{1}{(1 + 1.041B + 1.032B^2 + 1.030B^3 + 0.051B^4)} a_{SHPR}(t) \tag{3.88}$$

where PR_{SH} denotes the Short-term component of precipitation, and a_{SHPR} is a white noise series. After examining the SCCF between precipitation and water discharge, we noticed that the first spike occurs at lag 1, i.e. $b = 1$. This means that it takes one day for the water discharge to be affected by

the precipitation variable. With regard to the factors of the numerator and denominator of this variable, and since there are no spikes between the b value, where $b = 1$, and the beginning of the dying down pattern for the SCCF, this implies that there is no operator for the numerator of precipitation. On the other hand, for the denominator, it should be reasonable to set up the operator of the water discharge series to be equal to 1. That means we need to use the next model:

$$WD_{SH}(t) = \mu + \frac{1}{(1 - \delta_1 B)} BPR_{SH}(t) + \epsilon_t \tag{3.89}$$

where WD_{SH} is the short-term of the water discharge and ϵ_t is the error term of this model.

- For the tide time series, having investigated the values of the SACF and SPACF for the tide, the suggested model can be an AR(3) model. This model has been chosen as we have three obvious spikes in the SPACF and at the same time the pattern of the SACF dies down with a damped sine wave fashion. The autoregressive model of order three is shown in Equation 3.90.

$$TD_{SH}(t) = \frac{1}{(1 - 1.347B + 0.440B^2 + 0.241B^3)} a_{SHTD}(t) \tag{3.90}$$

where $Tide_{SH}$ denotes the Short-term of tide, and a_{SHTD} is a white noise series. To study the relationship between tide and water discharge, we computed the SCCF using the prewhitened values. The result is no spikes at the negative lags, and a spike has been noticed at lag 3, which means that the influence of tide on water discharge appears after 3 days. To complete the analysis of this variable, factors for tide in the TF need to be determined, investigation of the SCCF provides us with $r = 0$ and $s = 1$ for the factors.

- The last considered variable in our studied time series is the groundwater level. The model AR(2) can be suggested to fit the data of this series; this chosen model has confirmed its validity by examining the correlation analysis for the residuals.

$$GW_{SH}(t) = \frac{1}{(1 - 1.518B + 0.568B^2)} a_{SHGW}(t) \tag{3.91}$$

where GW_{SH} denotes the short-term for the groundwater level, and a_{SHGW} is a white noise series. The groundwater's level portion in the TF structure can be defined as follows:

$$WD_{SH}(t) = \mu + \frac{1}{(1 - \delta_1 B)} B^0 GW_{SH}(t) + \epsilon_t \tag{3.92}$$

where WD_{SH} is the short-term of water discharge, and ϵ_t is the error term of this model.

- Therefore, the preliminary model can be written as follows:

$$WD_{SH}(t) = \frac{(-0.002 - 0.002B)}{(1 + 1B)} BTE_{SH}(t) + \frac{1}{(1 + 0.066B)} BPR_{SH}(t) + \frac{(0.100 - 0.044B)}{(1 - 0.303B)} WS_{SH}(t) + \frac{1}{(1 + 0.183B)} B^3 TD_{SH}(t) + \frac{1}{(1 + 0.812B)} GW_{SH}(t) + \epsilon_{SH} \quad (3.93)$$

where WD_{SH}, TE_{SH}, PR_{SH}, WS_{SH}, TD_{SH}, GW_{SH}, and ϵ_{SH} denote the short-term components of water discharge, temperature, precipitation, wind speed, tide, groundwater's level, and the error term, respectively. Inspecting the residuals for the preliminary model by using the SACF and SPACF leads to select the type of the ARMA model that can be used to model the behaviour of these (noise) data. The model AR(1) has been chosen to fit the noise data, so the full model can be written as follows:

$$WD_{SH}(t) = \frac{(-0.001 + 0.096B)}{(1 - 0.615B)} BTE_{SH}(t) + \frac{1}{(1 + 0.399B)} BPR_{SH}(t) + \frac{(0.091 + 0.052B)}{(1 - 0.17B)} WS_{SH} + \frac{1}{(1 + 0.991B)} B^3 TD_{SH}(t) + \frac{1}{(1 + 0.671B)} GW_{SH}(t) + \frac{1}{(1 - 0.721B)} a_{SH}(t) \quad (3.94)$$

All the previous notations have been already defined, and $a_{SH}(t)$ is a white noise series with zero mean and a constant variance.

3.9.4 The Final Combined TF-Noise Model for Utica City

In order to build the final combined model, the data of the three components, which are long, seasonal, and short-term, have been used to implement the TF model. The results of the TF model with no structure on the noise series are shown in Equation

3.95.

$$WD_{FI}(t) = \frac{(-0.101 - 0.234B)}{(1 - 0.409B)} TE_{FILT}(t) + \frac{(0.092 + 0.081B)}{(1 + 0.999B)} PR_{FILT}(t) + \frac{(-0.019 - 0.095B)}{(1 - 0.087B)} BTE_{FISE}(t) + \frac{(0.088 - 0.096B)}{(1 - 0.092B)} BPR_{FISE}(t) + \frac{(-0.307 - 0.064B)}{(1 - 0.150B)} GW_{FISE}(t) + \frac{(0.052 - 0.093B)}{(1 - 0.087B)} BTE_{FISH}(t) + \frac{(1)}{(1 - 0.081B)} BPR_{FISH}(t) + \frac{(1)}{(1 - 0.161B)} GW_{FISH}(t) + \epsilon_{FI}(t). \quad (3.95)$$

Examining the residual correlation analysis of this model reveals that specifying an AR model of order one would fit the noise part of the model adequately. According to this specification, the final TF-Noise model has been built. This final model can be written as the following:

$$WD_{FI}(t) = \frac{(0.365 - 0.393B)}{(1 - 0.957B)} TE_{FILT}(t) + \frac{(0.031 - 0.002B)}{(1 - 0.937B)} PR_{FILT}(t) + \frac{(0.034 - 0.003B)}{(1 - 0.787B)} BTE_{FISE}(t) + \frac{(0.032 - 0.018B)}{(1 - 0.293B)} BPR_{FISE}(t) + \frac{(-0.128 - 0.130B)}{(1 + 0.999B)} GW_{FISE}(t) + \frac{(0.064 + 0.002B)}{(1 - 0.536B)} TE_{FISH}(t) + \frac{(1)}{(1 - 0.179B)} PR_{FISH}(t) + \frac{(1)}{(1 - 0.016B)} GW_{FISH}(t) + \frac{1}{(1 - 0.870B)} a_t. \quad (3.96)$$

3.10 Evaluation of the Estimated Models for Utica City

The same previous statistics, which are used to evaluate the constructed models for Poughkeepsie city, have been also utilised to evaluate the constructed models for Utica city. The results are shown in Table 3.2. Based on these statistics, it is recommended to use the combined models, which are built using the decomposed data, as they have the lowest AIC and SBC.

3.11 Discussion

Using the TF-Noise model enables us to incorporate lagged variables for the inputs based on the SCCF between the output and the input variables. The second feature

	AIC	SBC
TF-Noise for the Raw Data	2516.409	2607.028
TF-Noise for the Decomposed Data	2448.503	2583.478

Table 3.2: The Statistical Tests for the Model Selection.

that distinguishes this model and makes it more robust and reliable is the capability of the inclusion of an ARMA model for the residuals. This will assure that there is no more information can be added to the model. To avoid the problem of autocorrelation between the studied variables, the prewhitened values have been used.

With regard to the stationarity of the variables for the raw data, same differences, which are first differences, have been applied to the variables for the two cities Poughkeepsie and Utica. Based on the patterns of the SACF and SPACF, nearly similar tentative models have been used to prewhiten the variables of the raw data. These models are: (1) Moving Average models of orders 3 and 4 have been chosen for the variables temperature, precipitation, and wind speed (2) Autoregressive models of order 3 have been determined for the variables tide and groundwater level. Also, the residuals' correlation analysis suggests an AR(1) for the noise part in the preliminary models for the two cities.

For the three components, the differences used to achieve stationarity for the variables for the two cities vary, but in most cases either first differences or second differences have been employed. With reference to the tentative models, the results oscillate between similarity and dissimilarity for the studied variables for the two cities. However, the residuals for all the preliminary models for all component have been described using AR(1) models. In addition, the separating process enables us to construct three TF-Noise models that can be used to separately forecast the long, seasonal, and short-term components.

3.12 Conclusion

For the two cities Poughkeepsie and Utica for the three components, the effect of the input variables, which are temperature, wind speed, precipitation, tide, and groundwater level, on the output series, which is the water discharge, appears on the same day. The only exception is that for the Utica's city seasonal component, one day is the pure time delay for the effect of the temperature and the precipitation on the water discharge. These results have been obtained relying on the SCCF values for these variables.

Different results have been noticed for data for Cohoes city, where the first lag for temperature and wind speed has been chosen to construct the effect of these two

variables on the water discharge for the seasonal variations. For all components, all the other variables have contributed without any lagged values.

Using the decomposed data has improved the accuracy of the TF-Noise forecasting model, which inherently includes lagged variables and an ARMA model for the residual terms in its structure. The TF-Noise models for each component are better than the MLR models. This result has been extracted based on the AIC and SBC values, for example, the AIC value for Utica city data has reduced from 5340.251 to -14605.7 for the long term component.

Chapter 4

Bayesian Inference for Water Discharge Modelling and Uncertainty Analysis

In many fields, such as water resources planning and management, decisions often need to be taken based on several uncertain factors. The studies of the impact of climate change and the calculations of water balance in an un-gauged basin are examples of these factors. The sources of these uncertainties are different, but the uncertainties of model's parameters and model's structure are the most common types [35]. Essentially, different variables affect the amount of water discharge from a river but the two most important variables are the precipitation and groundwater level. Based on this, including these variables in a probability model for the amount of water discharge is necessary.

Also, most processes in the hydrological system, for example, water discharge from a river, contain different embedded components in their data which are the long, seasonal, and the short-term components. To gain a better insight into a trend of a hydrological process, separating these components is an essential step [36, 83]. The forecasting models constructed using the three components outperform models constructed using raw data [100]. However, the model's parameters and structure uncertainties have not been taken into account as the estimation of the parameters of these models is often performed based on Frequentist statistics.

The essential feature for Bayesian analysis is the explicit use of probability distributions for quantifying the model's parameters and structure uncertainties [14]. Bayesian analysis enables us to incorporate evidence from previous experiences or studies via prior distribution. Based on a set of data, Bayesian analysis can be defined as the process of constructing a probability model and provide a number of summary measures. The constructed probability distribution, which is known as the

posterior probability distribution, will be then used to compute the posterior predictive distribution to calculate the required quantities such as predictions for future observations [14].

The innovation in this chapter is the combined Bayesian MLR and combined Bayesian MLR-VAR models, where as far as we know that Bayesian analysis has not been used to estimate the parameters of a combined model constructed using the three components of the long, seasonal, and the short-term components. To show the difference between classical Bayesian MLR model and the new combined Bayesian models, a comparison has been carried out. Based on this, the current chapter considers two cases. The first case is the analysis of the raw data. A number of Bayesian Multiple Linear Regression (BMLR) model are constructed using different hyperparameters for the prior distributions for the model parameters. The second case is the analysis of the decomposed data where two main Bayesian models are constructed. One of these models is exclusively built using current values, whereas the second model is built using current values for the long and seasonal components and a number of lagged variables for the short-term component. For these two types different hyperparameters are specified for the prior distributions of the parameters.

This chapter is organised as follows. Section 4.1 presents a brief overview of Bayesian analysis. Section 4.2 displays a number of known types of prior distributions for Multiple Linear Regression, MLR, model. Section 4.3 discusses posterior distribution. Section 4.4 presents a brief description of BMLR analysis. Section 4.5 focuses on the posterior predictive distribution. Section 4.6 presents credible intervals and Highest Probability Density (HPD) function. Section 4.7 highlights some common ways for checking and evaluating a Bayesian model.

Section 4.8 presents the key theoretical concepts of the Bayesian Vector Autoregressive, VAR, models. Section 4.9 shows the priors that are common in the VAR process. Section 4.10 lists the major steps in our methodology. Section 4.11 presents our methodology for the first case where no lagged variables are included, on a real dataset collected for Utica city. Section 4.12 provides the results of combining the three components.

Section 4.13 presents the analysis for the second case where a number of lagged variables are included for the short-term component. Section 4.14 displays the findings of the final combined Bayesian model with VAR for the short-term component. Section 4.15 presents the results. A summary of the use of Bayesian analysis to estimate a probability forecasting model structured as a combined model is presented in Section 4.16. The data for Utica and Cohoes cities have been used to construct Bayesian models as it has the highest similarity measure based on the results in Table 5.13 where we have exploited these results to specify prior distributions to the two cities.

4.1 Bayesian Analysis

Forecasting process in frequentist statistics is accompanied with many issues as highlighted by Aitchison (1964) "In the theory of statistical tolerance regions, as usually presented in frequentist terms, there are inherent difficulties of formulation, development and interpretation" [2]. The essence of the problem is attributed to the definition of the probability in frequentist statistics. Defining the probability in frequentist statistics as the long-run frequency of event does not fit in some cases. For example, when there is a new product, data will not be available to construct a forecasting model.

Applying Bayesian approaches in the modern statistical analysis is becoming increasingly popular with applications in different areas [94]. Bayesian analysis is used in many fields, for instance, Physical Sciences, Biological Studies, Medicine, and Image processing. Bayesian statistics express beliefs about unknown quantities by using probabilities that are assumed to be conditional on data. The unknown quantities here are the parameters of a model. Essentially, Bayesian methods depend on the Bayes theorem, which was developed in the $1700s$ by Thomas Bayes. If there are two events A and B, the Bayes theorem can be written as the following:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.1}$$

If the event A is replaced with a parameter θ, and the event B is also replaced with a sample data, y, then the Bayes theorem can be rewritten as follows:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}. \tag{4.2}$$

The denominator, $P(y)$, which is also known as the normalizing constant in Bayesian analysis, represents the marginal probability for y. This constant is used to ensure that the integration of the posterior probability, $P(\theta|y)$, equals to one. Because $P(y)$ is a constant and it is often ignored in the written formula, the Bayes theorem, which is known as the posterior probability, can be written as follows:

$$P(\theta|y) \propto P(y|\theta)P(\theta). \tag{4.3}$$

Here the likelihood function, $P(y|\theta)$, is regularly updated with the prior probability, $P(\theta)$, to construct a posterior distribution, $P(\theta|y)$ [71]. In other words, the likelihood function of the data is weighted by the prior to provide the posterior distribution.

From a Frequentist perspective, the parameter θ can be estimated from data, $\mathbf{y}= \{y_1, \ldots, y_n\}$. This can be carried out through utilizing a statistical model that is expressed by a density function $p(y|\theta)$ [14]. In contrast to the situation in the

classical statistics, Bayesian philosophy assumes that the parameter θ can not be exactly determined but instead the probability statements and distributions can be used to describe the uncertainty about the parameter.

The prior distributions can strongly or weakly affect the posterior distribution depending on whether the prior distributions are subjective or objective. The objective prior distributions are also known as non-informative, diffuse, and flat priors. Most often, the chosen flat prior is a uniform distribution, where $P(\theta) = 1$. This flat prior has no influence on the posterior; that means the distribution of the posterior is similar to the likelihood function. The only difference between the likelihood and the posterior is that in the former the random variable is y while in the latter the random variable is the parameter(s) of the model.

Based on the posterior distribution, some summary measures, such as, mean and variance for the parameter of interest, are computed. There are also some measures that are sometimes not straightforward such as calculating the probability that a parameter is greater than a specific value. Typically, this difficulty can be handled by using a simulation method where the simulated data can provide a solution for such questions. Moreover, some posterior distribution forms are not standard or unknown and this imposes using the simulation-based methods to estimate the posterior.

The prominence of Bayesian methods has been attributed to the computing advances in the late 20th century [94, 71]. Often these methods are practically performed by repeatedly drawing a number of samples from a target distribution to estimate the posterior. The Markov Chain Monte Carlo (MCMC) methods are considerably used to generate samples [94, 18, 64]. Generating conditional independent samples using the target distribution is the Markov Chain responsibility, MC, where MC is a stochastic process. And, Monte Carlo, MC, is a numerical integration technique, which is utilised to compute the integration. By gathering these two common techniques, a series of samples are generated from the posterior to compute the required quantities using Monte Carlo.

There are several algorithms that depend on the MCMC simulation, but the most popular algorithms are Gibbs, Metropolis-Hasting, and Metropolis algorithms. As aforementioned, the Bayesian analysis regards parameters as random variables that have distributions [94]. The posterior distribution is the key element in the Bayesian analysis and all of the statistical Bayesian inferences are derived from summary measures from this distribution. For example, point and interval estimates can be obtained from the mean and quantiles of the posterior distribution, respectively.

Normally, in Bayesian paradigm, the most difficult part in the analysis is the estimation of a Bayesian model. In classical analysis, the mechanism of Maximum Likelihood estimation (MLE) depends on computing the parameter values that maximize the likelihood function. Then, based on results of the MLE, point estimates for the standard errors of these estimates are computed. Having estimated these points,

a classical statistical test is then performed. This test can be conducted by determining a hypothesized value for the parameter that is yielded from the ML, and subtract this value from the estimated parameter and then divide by the estimated standard error.

In contrast to the classical analysis, which finds a point estimate and its standard error for the parameters of interest, Bayesian analysis provides a posterior distribution for the parameter. The major steps in the modern Bayesian inference can be summarised as the following:

1. Multiplying the likelihood by the prior density to compute the posterior distribution for the parameter.

2. Statistics such as mean, median, and variance can be calculated to summarise the knowledge about the parameter. Each statistic can be found by computing the required integration using the posterior distribution.

3. Often, the summary statistics cannot be computed analytically when we have unknown posterior distributions. To overcome this problem, generating a sample of data for the parameter of interest from the posterior distribution will help the researcher to compute some statistics.

4.2 Prior Distributions

Despite all the controversy surrounding its use, the prior distribution remains the most important part of Bayesian analysis; as it is the element that transforms the analysis from Frequentist to Bayesian. Prior information (knowledge) can be incorporated into the analysis with the information that is provided by the observed data to elicit the posterior distribution [97], [14], and [71]. Priori information is represented by the prior distribution and the information from the observed data is represented by the likelihood function.

The priori knowledge could be an opinion from an expert or some historical data or results from a previous research or experiment. Obviously, the chosen prior distribution can have a huge impact on the outputs, and it must be selected carefully. In case that the posterior distribution has a density function which comes from the same family of distributions for the prior, the prior distribution is then called a conjugated prior.

The important term "conjugacy" will be between the likelihood and the prior function. This term has a substantial effect in Bayesian analysis, where it often provides known posterior distributions. For example, if the prior distribution is a beta density function and the likelihood function can be described by a binomial distribution, the

resulting posterior is also a beta density with different parameters. In this case the likelihood and prior are called conjugate distributions. Once the posterior distribution is determined, some summary measures can be computed using the posterior's properties. These measures are often mean, median, variance, and confidence interval (credible interval). Using the credible interval, simple interpretation can be obtained where it can be easily said that the parameter, θ, falls between the lower and upper intervals.

The specification of an appropriate prior distribution can be divided into two types. The first type is to specify a point estimate, and because of that this point estimate is often not known with a complete certainty, the posterior distribution can be computed using different values for the prior. Afterwards, we can compare and select which estimated posterior probability is more reasonable. The second type is a determined distribution to define the prior. Fundamentally, there are two types of prior distribution which are always specified to the parameters of interest. These two types can be summarised as follows:

- Non-informative prior distribution: As the name may imply, this distribution has no or minimum impact on the resulting posterior. There are some common distributions that are classified into this type. Often, although the prior has an improper density, the posterior of $\boldsymbol{\theta}$ has a proper distribution. The terms proper and improper refer to the integrable and non-integrable distributions (the integration equals to 1). Typically, it is impossible to make inferences with improper posteriors. This type of non-informative (flat) distributions can be obtained when an equal likelihood on all the possible values of the parameter is specified. If there is no or weak knowledge about the parameters of interest, the reasonable choice will be a flat distribution. The decision of whether the posterior is proper or improper can be made by verifying whether the normalizing constant is finite or not for all y. The uniform distribution is the most popular flat distribution.

- Informative prior: This term can be assigned to a prior when the chosen distribution has a considerable influence on the posterior; that means, the posterior distribution will not be controlled by the likelihood function. The researcher needs to be very careful when this type of distributions is selected. For more information about the priors see [70].

4.3 Posterior Distributions for the Parameters

In Bayesian analysis, Equation 4.1 has to be solved. Based on the complexity of the resulting kernel for the posterior function, which is completely related to the

specified prior distribution, either analytically or using numerical methods, solutions can be derived. In the case of existence of simple forms (small number of parameters, linearity, and simple prior distributions), the function $P(\theta|y)$ can be analytically computed. However, in most cases, the joint posterior distribution is unidentified [65]. In such a case, applying sampling methodologies provide numerical solutions. Different reasons can lead to unidentified posterior distribution, where

- Different distributions are required if we have different parameters. In this case it will be easier to identify a conditional and/or marginal distribution for each type of parameters. For example, if we have a joint distribution density with two parameters, mean and variance, which is $P(\boldsymbol{\theta}|y) = P(\theta_1, \theta_2|y)$ and this function has unidentifiable form (not standard distribution), it is simpler to identify the conditional distribution $P(\theta_1|\theta_2, y)$ and the marginal function $P(\theta_2|y)$. Based on the product's rule, the posterior distribution can be computed as follows:

$$P(\boldsymbol{\theta}|y) = P(\theta_1, \theta_2|y) = P(\theta_1|\theta_2, y)P(\theta_2|y).$$

  In the first term of this function, which is $P(\theta_1|\theta_2, y)$, θ_2 is considered as a constant, and it can be ignored in the computations. For the second term, which is $P(\theta_2|y)$, the parameter(s) θ_1 will be integrated out of the joint posterior distribution as follows:

$$P(\theta_2|y) = \int_{\theta_1} P(\theta_1, \theta_2|y) d\theta_1.$$

- Some types of joint posterior densities are not integrable with respect to one or more of parameters of interest. In this case, we are not able to apply the procedure presented above. However, Gibbs sampling can be used to simulate samples from the joint posterior. The Gibbs sampling is presented in the Appendix.

- The magic algorithm that can work with any type of joint posterior distribution, whether it is simple or complicated, is the MCMc. This method and some other common methods that are used in Bayesian analysis are introduced in Appendix.

4.4 Bayesian Multiple Linear Regression

This section introduces Bayesian analysis for the multivariate normal distribution and the related parameters in the MLR. Suppose that $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is a realizations vector of the response variable, which can be represented using an $(n \times 1)$ vector.

And $\mathbf{X} = (x_1, x_2, \ldots, x_p)$ is a matrix that contains the explanatory variables, where the dimension of this matrix is $n \times p$, n is the number of the observations, and p is the number of the covariates. If the constant term is added to the model, the dimension of the matrix $\mathbf{X}$ will be $n \times k$, where $k = p + 1$. The $\mathbf{X}$ matrix is commonly known as the design matrix and the values for the constant term will be 1 in the first column of the design matrix.

The Bayesian approach for the MLR can be summarised as follows:

$$y_i|(\mu_i, \sigma^2, X) \sim N(\mu_i, \sigma^2 I_n)$$

$$i = 1, 2, \ldots, n$$

and

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mathbf{x_i}'\boldsymbol{\beta}$$

where

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p).$$

The unknown parameters for the MLR model include the coefficients and the variance of the model, where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma^2]$. So, $p(\boldsymbol{\theta}|X)$ represents the joint prior of the parameters of interest, which are $\boldsymbol{\beta}$ and σ^2. The joint posterior distribution for $\boldsymbol{\beta}$ and σ^2 is

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) \times p(\sigma^2|\mathbf{y}).$$

Given σ^2, the conditional distribution of $\boldsymbol{\beta}$ can be defined as follows:

$$p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) \sim N(\hat{\boldsymbol{\beta}}, \sigma^2 V)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$.

Additionally, the marginal distribution for σ^2 has an Inverse Gamma (IG) function, which can be written as follows:

$$\sigma^2|\mathbf{y} \sim IG\left(\frac{n-k}{2}, \frac{(n-k)s^2}{2}\right)$$

where

$$s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Therefore, the posterior distribution is a Normal Inverse Gamma function (NIG).

In Bayesian Multiple Linear Regression, BMLR, it is supposed that data are drawn from a normal (Gaussian) distribution, so, the likelihood function will be:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma^2}|y, x) = \prod_{i=1}^{N} P(y_i|x_i, \boldsymbol{\beta}, \boldsymbol{\sigma^2}). \tag{4.4}$$

Given the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, $P(y_i|x_i, \boldsymbol{\beta}, \sigma^2)$ is the conditional probability density function of y_i that is induced by the conditional distribution of ϵ_i. Here, x_i is refereed to as a fixed quantity. If the errors (disturbances) are independent, Gaussian, and homoscedastic, then Equation 4.4 can be rewritten as the following:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2|y, x) = \prod_{i=1}^{N} \boldsymbol{\phi}(y_i; x_i\boldsymbol{\beta}, \boldsymbol{\sigma}^2) \tag{4.5}$$

where $\boldsymbol{\phi}(y_i; x_i\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ is the Gaussian probability density function that is evaluated at y_i with mean $x_i\boldsymbol{\beta}$ and variance $\boldsymbol{\sigma}^2$. Using Bayes' theorem, the joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ can be computed as follows:

$$p(\boldsymbol{\beta}, \boldsymbol{\sigma}^2|y, x) = \frac{p(\boldsymbol{\beta})p(\boldsymbol{\sigma}^2)\ell\ (\boldsymbol{\beta}, \boldsymbol{\sigma}^2|y, x)}{\int_{\boldsymbol{\beta},\boldsymbol{\sigma}^2} p(\boldsymbol{\beta})p(\boldsymbol{\sigma}^2)d\boldsymbol{\beta}d\boldsymbol{\sigma}^2} \propto p(\boldsymbol{\beta})p(\boldsymbol{\sigma}^2)\ell\ (\boldsymbol{\beta}, \boldsymbol{\sigma}^2|y, x). \tag{4.6}$$

The prior distribution of $\boldsymbol{\beta}$ should be replaced with $\boldsymbol{\beta}|\boldsymbol{\sigma}^2$ when the parameter(s) $\boldsymbol{\beta}$ relies on $\boldsymbol{\sigma}^2$. The joint posterior distribution is analog to any other joint probability distribution for a random variable. With reference to the posterior distribution, the integrals of functions of parameters are used to compute the estimates and inferences of parameters. If the posterior is the kernel of a known probability distribution, then integrals of the parameters can be analytically tractable. Most often, known kernels exist when a conjugated prior is utilised where this conjugated prior leads to a known posterior function. In this case, a number of moments of the distribution of interest are known and can be later used to estimate the parameters. On the other hand, if unknown kernels arise, then analytically intractable posterior is obtained where integrals of the parameters can not be analytically computed. Consequently, numerical integration mechanisms are required to compute the integrals. Most of numerical integration can be implemented, under some conditions, by using MCMC sampling. Typically, to carry out Monte Carlo estimation, many samples have to be drawn from a probability distribution, and for each draw, a suitable function has to be applied, then the obtaining draws have to be averaged to approximate the integral.

There are some cases where the mean is known and the variance is unknown in normal distribution. This is an important case as an example of the estimation of a scale parameter in Bayesian analysis. If the random variable y is distributed normally with known mean and unknown $\boldsymbol{\sigma}^2$, then the likelihood of this random variable, which includes n of independent and identical observations, is:

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n} \exp\big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\big) \\ &= (\sigma^2)^{-n/2}\exp(-\frac{n}{2\sigma^2}v) \end{aligned} \tag{4.7}$$

where $v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)$ is a sufficient statistics. It is clear that the corresponding conjugate prior distribution is the Inverse-Gamma (IG), so:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp -\beta|\sigma^2$$

where α ad β are the hyperparamters of this prior. In Bayesian analysis, the main object for forecasting is the posterior predictive distribution, where the distribution of the future values $y_{T+1}, \ldots, y_{T+H}$ is conditional on the observed data Y_T, $t = 1, \ldots, T$, i.e. $p(y_{T+1}, \ldots, y_{T+H}|y_T)$. To predict the unknown future values for the desired event, all the relevant information can be calculated using the posterior predictive distribution. These future values can be captured by choosing the related feature from the posterior predictive distribution where this feature can be mean, mode, or median.

4.5 Posterior Predictive Distribution

In Bayesian analysis, the distribution of unobserved data (prediction) conditional on observed data is called the posterior predictive distribution. Given the observed data vector $\mathbf{y}$, the parameters vector $\boldsymbol{\beta}$, and the unobserved data $\mathbf{y}_{pred}$, the posterior predictive distribution can be defined to be the following:

$$\begin{aligned} p(\mathbf{y}_{pred}|\mathbf{y}) &= \int p(\mathbf{y}_{pred}, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{y}_{pred}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \end{aligned} \tag{4.8}$$

In addition, given $\boldsymbol{\theta}$, the assumption that the observed and unobserved observations are conditionally independent has to be achieved. According to this assumption, the PPD (Equation 4.8) can be rewritten as the following:

$$p(\mathbf{y}_{pred}|\mathbf{y}) = \int p(\mathbf{y}_{pred}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

So, the PPD is an integral of the product of the likelihood function $p(\mathbf{y}_{pred}|\boldsymbol{\theta})$ and the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ [14, 65, 21]. Since we have a new distribution, which is the posterior predictive, a determined number of samples will be generated. The prior predictive distribution, which is also known as the marginal function of the observations, is not similar to the PPD. The difference is that the prior predictive distribution is an integral of the product of the likelihood function and the prior distribution of the parameter, which can be defined as the following:

$$p(\mathbf{y}_{pred}) = \int p(\mathbf{y}_{pred}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where the prior predictive distribution is not conditional on the observed data. In the Bayesian paradigm, the PPD can be effectively exploited to check whether the constructed model is consistent with data as discussed by Gelman et al. (2013), [14].

4.6 Credible Intervals and Highest Probability Density

The main motivation for Bayesian thinking is that it provides a more common-sense interpretation for a number of statistical conclusions. For example, a Bayesian (probability) interval, which is commonly known as Credible Interval (CI), for an unknown parameter, can be directly considered as the interval that contains the unknown parameter. This interpretation is completely different from the one considered in the frequentist analysis, where the interval is interpreted based on a sequence of similar inferences that are made in repeated practice.

Different products can be yielded from Bayesian posterior inference, one of them is summarizing the posterior marginal densities for the parameters of interest [22]. It is known that by tabulating $100(1-\alpha)\%$ posterior credible intervals, the marginal posterior distribution for the parameter considered can be summarised. This quantity can be obtained analytically or by using MCMC method. However, $100(1-\alpha)\%$ highest probability density (HPD) interval is a highly recommended criterion, specifically, when the marginal posterior distribution is not symmetric [22, 14]. The HPD has two major properties which they are:

- Inside the interval, the density for every point is higher than that for every point outside the considered interval.
- The interval has the shortest length for a given probability.

Therefore, after computing the posterior marginal distribution and a number of samples have been drawn, credible or HPD intervals can be derived. The interval for a parameter θ can be written as the following:

$$\int_{\delta 1}^{\delta 2} p(\theta|\mathbf{y})d\theta.$$

4.7 Checking and Comparing Bayesian Models

Having constructed and estimated a posterior distribution for all estimands for a Bayesian model, it is fundamental to accomplish the task of assessing the fitted model

to the data. Most of the methods used for this purpose are based on using the posterior predictive distribution. To test the fitting of the resultant model to the considered data, some quantities from the posterior predictive distribution, such as mean, standard deviation, and maximum and minimum values, have to be tested. Then, based on a chosen test, for example a t-test, a hypothesis testing for the difference between the real and 'simulated' values of one or more of these test quantities, can be applied. This method is almost the most common way to check the adequacy of a Bayesian model to the data of interest [14]. Indeed, the inclusion of all knowledge about a specific problem in a probability distribution is rather difficult, therefore, it is required to examine what features or aspects have not been captured by the model [14].

For a scientific problem, often more than one model can fit the data adequately. Based on this fact, there is a question arises which is: to what extent do posterior inferences vary when the current model is replaced by another new probability model?. The new model can be different in many aspects, for example, the specified prior, the hyperparameters, the sampling distribution, or the number of the considered variables, for example, the covariates in regression analysis. Applying the posterior predictive distribution to an external data to make predictions for new data is also preferable as a procedure of checking and is commonly known as an External Validation.

4.7.1 Deviance Information Criterion (DIC)

The natural way to compare different constructed models is to apply a criterion that relies on the principle of trade-off between the fitting of the model to the studied data and the corresponding complexity of the model. Based on this, Spiegelhalter et al. (2002) proposed a criterion, which is known as Deviance Information Criterion, DIC, to compare between Bayesian models [91]. This criterion can be summarised as:
DIC= Goodness of Fit+ Model Complexity.
The first term, goodness of fit, can be calculated as the following:

$$D(\theta) = -2LogL(data|\theta).$$

The second term, model complexity, can be computed by estimating the effective number of parameters in the model. Mathematically, this can be written as

$$\begin{aligned} P_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \bar{D} - D(\bar{\theta}) \end{aligned}$$

where $\bar{D}$ is the posterior mean deviance and $D(\bar{\theta})$ is the deviance that is evaluated at the posterior mean of the parameters. This criterion can also be rewritten in a way

that is similar to AIC criterion as the following:

$$DIC = D(\bar{\theta}) + 2P_D$$
$$= \bar{D} + P_D.$$

These two quantities can be calculated using MCMC.

4.8 Bayesian Vector Autoregressive

In case that there is a vector of random variables of k-dimensional time series of interest and to practically understand the dynamic relationships over time among them, analysing and modelling them together is necessary [64, 18]. Also, by using additional information that is available from the associated series, the forecasts process accuracy for individual series will be enhanced.

Let $\mathbf{y_t} = (y_{1t}, y_{2t}, \ldots, y_{kt})'$, $t = 1, 2, \ldots, T$ denote a vector of time series of k-dimension. A p^{th}-order Vector Autoregressive Process, commonly denoted as VAR(p), can be written as follows:

$$\mathbf{y_t} = \boldsymbol{\delta} + \boldsymbol{\Phi_1} \mathbf{y}_{t-1} + \ldots + \boldsymbol{\Phi_p} \mathbf{y_{t-p}} + u_t$$

where $(\delta_1, \ldots, \delta_k)'$ is a vector of constants and $\boldsymbol{\Phi_j} = \boldsymbol{\Phi_1} \ldots \boldsymbol{\Phi_p}$ is a matrix of dimension $k \times k$. Also, u_t is the error term, where $u_t \sim N(0, \Sigma_u)$. By changing the vectors and matrices with scalars, we will simply provide an AR model of order p.

The simplest form of this modelling technique is the VAR(1) with two variables (time series A and B), which can be written as the following:

$$\begin{bmatrix} y_{A,t} \\ y_{B,t} \end{bmatrix} = \begin{bmatrix} \delta_{A,0} \\ \delta_{B,0} \end{bmatrix} + \begin{bmatrix} \phi_{A,11} & \phi_{B,12} \\ \phi_{A,21} & \phi_{B,22} \end{bmatrix} \times \begin{bmatrix} y_{A,t-1} \\ y_{B,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{A,t} \\ \epsilon_{B,t} \end{bmatrix}.$$

That means,

$$y_{A,t} = \delta_{A,0} + \phi_{A,11} y_{A,t-1} + \phi_{B,12} y_{B,t-1} + \epsilon_{A,t}$$
$$y_{B,t} = \delta_{B,0} + \phi_{A,21} y_{A,t-1} + \phi_{B,22} y_{B,t-1} + \epsilon_{B,t}.$$

By using matrix notations, a VAR model can be written as the following:

$$Y = \boldsymbol{\Phi} \mathbf{Z} + \mathbf{U} \tag{4.9}$$

where $Y = (y_1, y_2, \ldots, y_T)$, $\boldsymbol{\Phi} = (\boldsymbol{\delta}, \boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}, \ldots, \boldsymbol{\Phi_p})$, and $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_{T-1})$, with $Z_{t-1} = (1, y'_{t-1}, y'_{t-2}, \ldots, y'_{t-p})'$. By vectorising (applying the "vec" operator), the previous equation can be rewritten as follows:

$$\mathbf{y} = (\mathbf{Z}' \otimes I_k) \boldsymbol{\varphi} + \mathbf{u} \tag{4.10}$$

where $\boldsymbol{\varphi} = vec(\boldsymbol{\Phi})$, $\mathbf{y} = vec(Y)$, and $\mathbf{u} = vec(\mathbf{U})$.

4.9 Prior Distributions for Bayesian Vector Autoregressive

The specification of a diffuse (non-informative) prior distribution is the simplest version that can produce Bayesian inferences that are almost similar to the MLE method. A uniform distribution for the $\boldsymbol{\varphi}$ and a Jeffreys' prior for the Σ_u distribution can be specified. That means, $p(\boldsymbol{\varphi}, \Sigma_u) \sim |\Sigma_u^{-(k+1)}|$.

In a way that is similar to the specification of a multivariate normal distribution to the parameters of a MLR model, a multivariate normal distribution with known mean vector $\boldsymbol{\varphi_0}$ and covariance matrix $V_{\boldsymbol{\varphi}}$, can be specified to the parameters vector $\boldsymbol{\varphi}$, where it can be written as the following:

$$\boldsymbol{\varphi} \sim N(\boldsymbol{\varphi_0}, V_{\boldsymbol{\varphi}}).$$

Therefore, given Σ_u, which is the covariance matrix of the model, the prior distribution can be written as the following:

$$p(\boldsymbol{\varphi}) = \left(\frac{1}{2\pi}\right)^{k(kp+1)/2} |V_{\boldsymbol{\varphi}}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\varphi} - \boldsymbol{\varphi_0})' V_{\boldsymbol{\varphi}}^{-1} (\boldsymbol{\varphi} - \boldsymbol{\varphi_0})\right].$$

In addition, from Equation 4.9 and based on the assumptions of the error term, each of the T observed response vector y of size k, $t = 1, 2, \ldots, T$ is also independent and identically distributed (i.i.d). It is also supposed that these vectors follow a multivariate normal distribution (MN) given the vector of the coefficients of the VAR model, which are $\boldsymbol{\varphi}$ and $\boldsymbol{\Sigma}_u$ [64]. Then the joint density of the T vectors of error defines the Gaussian likelihood function, which is:

$$\ell(\mathbf{y}|\boldsymbol{\varphi}) = \left(\frac{1}{2\pi}\right)^{kT/2} |I_T \otimes \Sigma_u|^{-1/2} \times \exp\left[-\frac{1}{2}[\mathbf{y} - (\mathbf{Z}' \otimes I_k)\boldsymbol{\varphi}]'(I_T \otimes \Sigma_u^{-1})[\mathbf{y} - (\mathbf{Z}' \otimes I_k)\boldsymbol{\varphi}]\right].$$

Multiplying these two functions, the prior and the likelihood, the posterior density will be yielded:

$$\begin{aligned} p(\boldsymbol{\varphi}|\mathbf{y}) &\propto p(\boldsymbol{\varphi})\ell(\boldsymbol{\varphi}|\mathbf{y}) \\ &\propto \exp\left\{-\frac{1}{2}\Big[\big[V_{\boldsymbol{\varphi}}{}^{-1/2}(\boldsymbol{\varphi} - \boldsymbol{\varphi_0})\big]'\big[V_{\boldsymbol{\varphi}}{}^{-1/2}(\boldsymbol{\varphi} - \boldsymbol{\varphi_0})\big] \right. \\ &\quad + \{(I_T \otimes \Sigma_u^{-1/2})\mathbf{y} - (Z' \otimes \Sigma_u^{-1/2})\boldsymbol{\varphi}\}' \\ &\quad \left.\times \{(I_T \otimes \Sigma_u^{-1/2})\mathbf{y} - (Z' \otimes \Sigma_u^{-1/2})\boldsymbol{\varphi}\}\Big]\right\}. \end{aligned} \tag{4.11}$$

By defining the following:

$$\omega = \begin{bmatrix} V_{\boldsymbol{\varphi}}^{-1/2}\boldsymbol{\varphi_0} \\ (I_T \otimes \Sigma_u^{-1/2})\mathbf{y} \end{bmatrix}$$

and

$$\Omega = \begin{bmatrix} V_{\boldsymbol{\varphi}}^{-1/2} \\ (Z' \otimes \Sigma_u^{-1/2}) \end{bmatrix},$$

Equation 4.11 can be rewritten as the following:

$$= -\frac{1}{2}(\omega - \Omega\boldsymbol{\varphi})'(\omega - \Omega\boldsymbol{\varphi})$$

$$= -\frac{1}{2}\big[(\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}})'\Omega'\Omega(\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}}) + (\omega - \Omega\bar{\boldsymbol{\varphi}})'(\omega - \Omega\bar{\boldsymbol{\varphi}})\big].$$

Therefore, after simplifying, the posterior distribution is:

$$p(\boldsymbol{\varphi}|\mathbf{y}) \propto \exp\bigg[-\frac{1}{2}(\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}})'V_{\boldsymbol{\varphi}}^{-1}(\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}})\bigg] \tag{4.12}$$

where the mean of the posterior density is:

$$\bar{\varphi} = \big[V_{\boldsymbol{\varphi}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})\big]^{-1}\big[V_{\boldsymbol{\varphi}}^{-1}\boldsymbol{\varphi_0} + (Z \otimes \Sigma_u^{-1})\mathbf{y}\big]$$

and the covariance matrix of the posterior is:

$$\bar{V_{\boldsymbol{\varphi}}} = \big[V_{\boldsymbol{\varphi}}^{-1} + (Z'Z \otimes \Sigma_u^{-1})\big]^{-1}.$$

It is clear that the density above, Equation 4.12, can be recognised as a multivariate normal distribution with mean $\bar{\varphi}$ and covariance matrix $\bar{V_{\boldsymbol{\varphi}}}$. In other words, the posterior density of $\boldsymbol{\varphi}$ is $N(\bar{\varphi}, \bar{V_{\boldsymbol{\varphi}}})$. Because the distribution of this density is of a known form, it will be easy to use it to draw samples. On the other hand, if the covariance matrix of the model is not known, a distribution has to be specified to describe the behaviour of this matrix. The Inverse Wishart distribution (IW) is the common choice [64]. Practically it is easier to work with the precision, which is simply the inverse of the variance, rather than the variance itself. In this case, the Wishart distribution will be used. For the VAR model, which is expressed in Equation 4.9, if the $u_t \sim N(0, \Sigma_u)$, the following multivariate normal distribution can be specified

$$\boldsymbol{\varphi}|\Sigma_u \sim MN(\boldsymbol{\varphi_0}, V_{\boldsymbol{\varphi}} = V \otimes \Sigma_u)$$

and

$$\Sigma_u \sim IW(S, n).$$

In case that we consider the precision matrix instead of the covariance matrix, we will have

$$\Sigma_u^{-1} \sim W_k(S^{-1}, n).$$

Based on the priors that have been assigned to the coefficients and the variance of the model, the resulting posterior distribution is a normal-inverse Wishart distribution, sometimes also called a Gaussian-inverse Wishart distribution, which can be written as the following: $\boldsymbol{\varphi}|\Sigma_u, y \sim MN(\boldsymbol{\varphi I}, V_{\boldsymbol{\varphi}} = V \otimes \Sigma_{\boldsymbol{\varphi}})$, $\Sigma_u|y \sim IW_k(S, \tau)$.

4.9.1 The Minnesota Prior Distribution

The Litterman prior, which is also known as Minnesota Prior, is a Gaussian prior that can be specified for the parameters of a VAR model. This prior density was proposed by Litterman (1986) and Doan, Litterman, and Sims (1984) [64]. By utilising this distribution, the VAR estimates will be shrunk towards a multivariate random walk distribution. This prior is extensively used in the economics time series analysis [66]. This prior can be defined by specifying the following mean and covariance matrix for the parameters of interest:

- For each equation, the prior mean for the first lag for the endogenous variable of interest will be set to 1, or any other number, and all other prior means for the other variables will be zero.
- The prior variance of the intercept terms will be set to infinity and the prior variance for the other terms in the coefficients matrix, $\boldsymbol{\Phi_i}$, will be set to the following two values:

$$vij, l = \begin{cases} (\frac{\alpha}{l})^2, & \text{if i=j,} \\ (\frac{\alpha\theta\sigma_i}{l\sigma_j})^2, & \text{if } i \neq j. \end{cases} \tag{4.13}$$

  where α is the prior standard deviation for the parameter, σ_i^2 is the i^{th} diagonal entry of Σ_u, and θ is the term that controls the relative tightness of the prior variance for the other lagged variables in the equation of interest.

For example, a VAR model of order 2 that all its slope parameters are set up to their prior mean would be as follows:

$$y_{A,t} = \underbrace{0}_{\infty} + \underbrace{1 \times y_{A,t-1}}_{(\alpha)} + \underbrace{0 \times y_{B,t-1}}_{(\alpha\theta\sigma_1/\sigma_2)} + \underbrace{0 \times y_{A,t-1}}_{(\alpha/2)} + \underbrace{0 \times y_{B,t-1}}_{(\alpha\theta\sigma_1/2\sigma_2)} + u_{1t}. \tag{4.14}$$

$$y_{B,t} = \underbrace{0}_{\infty} + \underbrace{0 \times y_{A,t-1}}_{(\alpha\theta\sigma_2/\sigma_1)} + \underbrace{1 \times y_{B,t-1}}_{(\alpha)} + \underbrace{0 \times y_{A,t-1}}_{(\alpha\theta\sigma_2/2\sigma_1)} + \underbrace{0 \times y_{B,t-1}}_{(\alpha/2)} + u_{1t}. \tag{4.15}$$

The numbers in the parenthesis are the prior standard deviations for the coefficients.

4.10 Bayesian Multiple Linear Regression Methodology (BMLR)

The major steps of our methodology can be summarised as the following:

1. To apply BMLR, we need to specify prior distributions for the parameters of the MLR model which they are the coefficients (β) and the variance (σ^2).

2. Specifying a Multivariate Normal distribution (MN) for the coefficients and an Inverse Gamma for the variance has been conducted.

3. Non-informative and informative prior distributions for β have been assigned.

4. Different hyper-parameters have been specified for the prior density to construct various BMLR and combined BMLR models.

5. Two non-informative priors are applied for the coefficients of the model. These two priors are (1) normal distribution with zero mean vector and high variance for all the considered variables (2) uniform distribution with the maximum likelihood estimates (MLE) of the regression parameters for Utica city as the starting values for the simulation process. Generally, a non-informative distribution is recommended in case that there are many observations and only a few parameters where it can provide acceptable results [14]. On the other hand, and as an attempt to see the impact of the similarity analysis for two cities on the BMLR, an informative prior distribution is specified. This prior is a MN distribution for the coefficients (β) with hyper-parameters taken from the MLE of the regression parameters and their variance-covariance matrix for another city, here is Cohoes city, to be the hyper-parameters for the city of interest, which is here Utica city.

4.11 The Application

4.11.1 The Study Region Data and Bayesian Analysis

Utica is a city in New York State, USA, located on the Mohawk River. The length of this River is 149-mile-long (240 km), where it is the largest tributary for the Hudson river, and drains out at about 3.412 square miles (8.837 square km). A few miles north of the Albany city, exactly in the Capital District, the Mohawk River flows into the Hudson river. The Schoharie and West Canada Creeks are the main tributaries of the Mohawk River, and this river has a long record of flooding [86].

In order to select the significant variables, the correlation matrix needs to be computed. According to their low correlation coefficient, the variables of Absolute Humidity, Dew Point, Sea Level Pressure, Visibility Miles, and Cloud Cover, are ignored. The dataset is separated into two parts; the data for the period $2005 - 2013$ is used to construct the models and the data for year 2014 is utilised to validate and assess the constructed models. The former is statistically known as the training data, and the latter is known as the testing data. Figure A.4 in Appendix clarifies the steps taken to construct the developed models.

4.11.2 BMLR Analysis for the Raw Data

To fit a BMLR model for the water discharge, WD, we need to determine the following:

1. The likelihood function for the water discharge conditional on the considered covariates; this function can be written as the following:

$$L(WD|\boldsymbol{X}, \boldsymbol{\beta}) = Normal(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I_n}).$$

   Here, the six regression parameters in the likelihood function are $\beta_0, \ldots, \beta_5$.

2. The priors for the parameters, which are the regression coefficients and the variance of the model.

The WD has the following density function:

$$WD \sim Normal(\boldsymbol{\mu}, \sigma^2)$$

where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, X is the design matrix for the covariates.

Three analyses have been performed to construct a Bayesian model for the WD:

- Firstly, a flat distribution for the prior of the parameters needs to be chosen. A non-informative independent normal prior distribution with mean zero and variance 1000000 has been specified for each parameter, *i.e.* $\boldsymbol{\beta} \sim N(0, 1000000 I_6)$. Different values have been tried to establish the covariance matrix for the coefficients, for example, $1, 10, 100, 1000, 10000, 100000$, and 1000000, the same MSE for all cases is obtained. A special case appeared when we considered the parameters 3000 and $10 I_n$ for the mean vector and the covariance matrix, respectively, where the MSE reduces with a remarkable percentage.

- Secondly, in a Bayesian analysis, an expert opinion or information from relative studies, such as previous experiments or researches, need to be used to determine a prior distribution. After defining the priors, their hyper-parameters have to be set up to some magnitudes. Based on this, the findings of the similarity analysis that are shown in Chapter 5 in Section 5.5.1 can be exploited effectively to specify hyper-parameters for the priors to apply an informative Bayesian analysis for Utica city. The results of similarity show that the data for the two cities of Utica and Cohoes have the highest similarity measure compared to the other considered cities.

  The specification can be carried out by selecting the coefficients of the MLR for Cohoes city and their variances as hyper-parameters for the prior of the parameters. The parameters are the coefficients of BMLR and their prior distribution is the multivariate normal distribution. Hence, based on Chapter 2 Section 2.6, the mean vector for the coefficients of BMLR is taken from Equation 2.5 in subsection 2.6.1. Therefore, the vector of the coefficients is: $\mu_0 = (0, -0.29, 0.01, 0.26, 0.13, -0.47)$ for the intercept and the explanatory variables, respectively. With regard to the covariance matrix (Σ_0) of the coefficients, the covariance matrix of the coefficients of the MLR model for the Cohoes city model, has been used. The prior distribution of the parameters (coefficients), therefore, is $\boldsymbol{\beta} \sim N(\mu_0, \Sigma_0)$.

- Finally, a non-informative uniform distribution with constants for all the regression coefficients and their variances and a non-informative gamma prior distribution for the normal scale (variance) parameter of the model have been used; with these prior distributions, the maximum likelihood estimates (MLE) for the regression parameters for Utica city, which occurs in Chapter 2, Section 2.7, Equation 2.11, and their covariance matrix will be used as the starting values for the simulation process [56, 57].
  Having decided the values for both the non-informative and informative prior distributions, Bayesian models (posterior distributions) for Utica city have been computed as follows.

  - The first BMLR model is constructed by multiplying the likelihood function of the response variable by the non-informative prior and the obtained model can be called the Non-Informative BMLR (NI-BMLR). The parameters of this model are almost similar to the parameters of the classical regression analysis.
  - By multiplying the likelihood function of the WD by the informative prior that is based on the MLR model for Cohoes city, the second BMLR model is built and can be called the Informative BMLR (IN-BMLRCohoes). The

results of this model are rather different from the findings of the classical regression.

– The structure of the third BMLR model depends on multiplying the likelihood function of the WD by the third non-informative prior, which is the uniform distribution. This model can be called the Non-Informative BMLR (NI-BMLRUtica). The results that have been computed by the Random Walk Metropolis algorithm and Gibbs sampling are shown in Table 4.1 [94].

Parameter	MLR	NI-BMLR	IN-BMLRCohoes	NI-BMLRUtica
IN	0	0	0	0
TE	-0.29	-0.29	-0.30	-0.29
WS	0.07	0.07	0.04	0.07
PR	0.28	0.29	0.27	0.28
TD	0.21	0.21	0.15	0.21
GW	-0.36	-0.36	-0.40	-0.36

Table 4.1: The Coefficients of MLR, NI-BMLR, IN-BMLRCohoes and NI-BMLRUtica for the Raw Data.

When a flat (non-informative) prior distribution for the parameters is utilised, the means of the posterior distribution for the coefficients of the model are almost identical to the Maximum Likelihood Estimates (MLE) of MLR as shown in Table 4.1. It is obvious that the parameters of MLR, NI-BMLR, NI-BMLRUtica models are similar.

There are three plots that are commonly used to assess the convergence of the posterior in Bayesian analysis. Figure 5.2 shows these three diagnostic plots for assessing the convergence of the generated samples for the parameter of Temperature. The first plot shown above is called the trace, this plot visualises the behaviour of the Markov chain for the sampled values of the parameter of interest. It seems that this series is almost stabilized and constant over the graph. Furthermore, apart from the large spike at lag 0, the autocorrelation plot indicates no degree of autocorrelation for the posterior samples, which refers to a good mixing. Finally, the kernel density plot, which is also known as the posterior density, estimates the posterior marginal distribution for the temperature's parameter. In conclusion, these plots imply that the Markov Chain has successfully converged to the desired posterior.

In Table 4.2, although the results of the posterior distribution for the parameters are almost similar to the results of the Frequentist sampling distribution for these parameters, two completely different interpretations are produced for the confidence

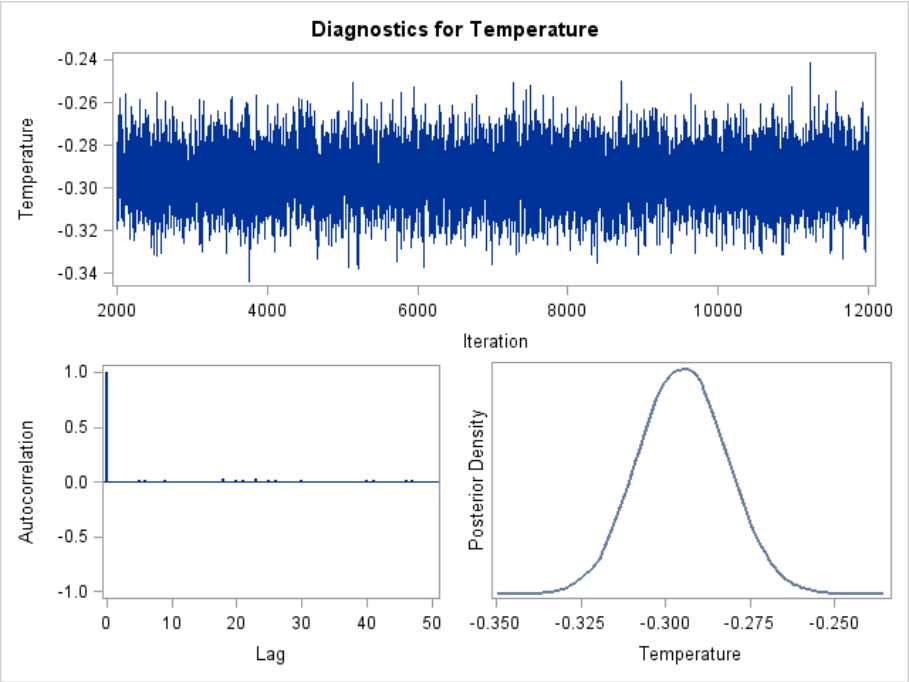

Figure 4.1: Diagnostic Plots for Temperature.

and credible intervals in Frequentism and Bayesianism statistics. In Bayesian statistics the interpretation would be, given the observed data, there is a 90% confidence probability that the true values for these parameters fall within the range of the computed credible regions. On the other hand, in frequentist statistics, the interpretation would be, there is a 90% confidence probability that when we compute the confidence interval from data of this sort, the true values for these parameters will fall within the calculated ranges. We have applied two confidence probabilities, which are 90% and 95%. Moreover, for the purpose of model selection, in Table 4.3 and compared to the other models, the MSE for the IN-BMLRCohoes has the highest value. This result might be due to the small variances for the coefficients in the MLR for Cohoes city. This has been noticed when we tried different values for the covariance matrix of the coefficients. The model with the smallest DIC value shows the best fit to the data compared to other models. Hence, we can conclude that the IN-BMLRUtica model can be chosen as it has the smallest DIC value. Figure 4.2 shows the credible intervals for the parameters of the BMLR model for the raw data.

Table 4.2: Confidence Interval Comparison Between MLE and Bayesian Methods for the Estimation of the Parameter of interest.

			90%			95%		
	Parameter	Estimate	Upper	Lower	Dif	Upper	Lower	Dif
Bayesian	Intercept	0.00	-0.02	0.02	-0.04	-0.01	0.03	-0.05
	TE	-0.29	-0.34	-0.30	-0.04	-0.31	-0.27	-0.04
	WS	0.07	0.05	0.09	-0.04	0.05	0.10	-0.04
	PR	0.28	0.26	0.31	-0.04	0.26	0.31	-0.04
	TD	0.21	0.16	0.22	-0.05	0.18	0.24	-0.05
	GR	-0.36	-0.35	-0.30	-0.04	-0.39	-0.3	-0.05
MLE	Intercept	0.00	-0.02	0.02	-0.04	-0.02	0.02	-0.04
	TE	-0.29	-0.34	-0.29	-0.04	-0.31	-0.26	-0.05
	WS	0.07	0.05	0.09	-0.04	-0.01	0.03	-0.05
	PR	0.28	0.26	0.31	-0.04	0.23	0.28	-0.04
	TD	0.21	0.17	0.22	-0.04	0.11	0.16	-0.05
	GW	-0.36	-0.35	-0.30	-0.04	-0.49	-0.44	-0.05

Table 4.3: DIC Values for the Raw Data with Different Types of Priors.

Model	NI-BMLR	IN-BMLRCohoes	NI-BMLRUtica
DIC	7445.291	7476.317	7439.595
MSE	0.59	0.63	0.59

4.11.3 BMLR Model for the Decomposed Data

For constructing a BMLR model for the decomposed data, which are the components the long, seasonal, and short-term component, a MN distribution can be chosen as a prior density for the parameters $\boldsymbol{\beta}$ for each component. The hyper-parameters are based on the coefficients of the MLR models and their covariance matrices for the components of the long, seasonal, and short-term component from subsections 2.6.1, 2.6.4, 2.6.5, and 2.6.6 for Utica city; 2.7.1, 2.7.5, 2.7.7, and 2.7.9 for Cohoes city. Also, an IG distribution has been specified for the variance of the three models as shown below.

Prediction Modelling for the Long-Term Component

Follow the same steps listed in the above, we chose the non-informative and informative prior distributions for the parameters of the BMLR model. The hyper-parameters are either based on the coefficients and their variances from the MLR model, 2.7, or constant values. The results for the long-term trend are shown in Table 4.4.

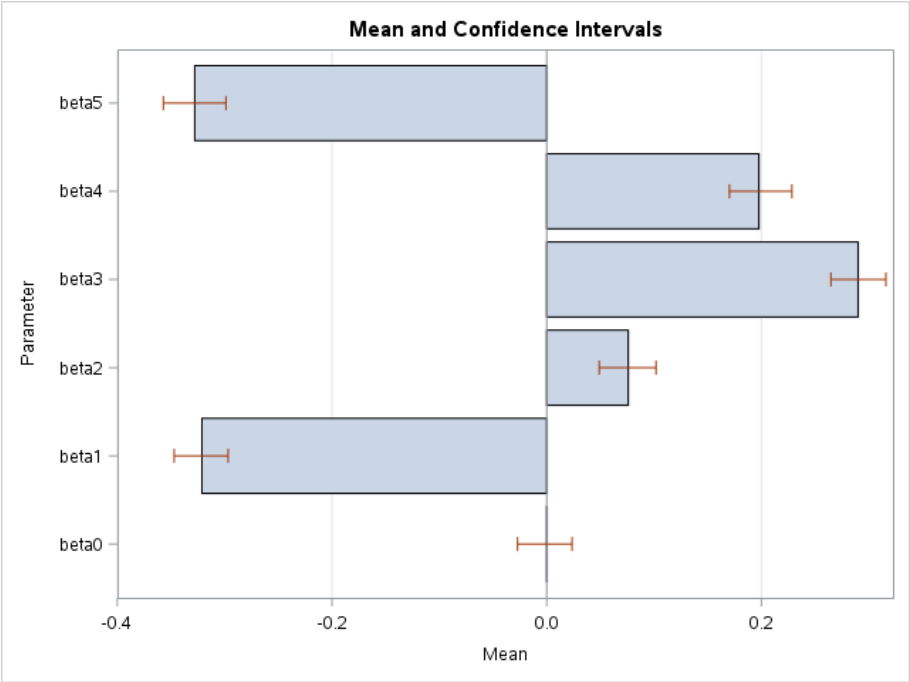

Figure 4.2: The Confidence Intervals for the Parameters of Raw Data for Utica City.

The MSE values for the testing data, which is the daily data for the year 2014, for the long-term component, are shown in Table 4.5. Examining these findings leads us to conclude that using the hyper-parameters that depend on the values of MLR model for Cohoes city has the highest MSE value.

Prediction Modelling for the Seasonal Variations

A MN and a uniform, and an IG distributions have been specified for the coefficients and the variance for the seasonal model, respectively. The MN distributions are constructed using different hyper-parameters for the mean vector and covariance matrix for the coefficients.

- The first type of hyper-parameters is based on zero mean vector and $1000000I_6$ covariance matrix for the coefficients; this is a non-informative prior distribution that constructs NI-BMLR.

- The second type of hyper-parameters is based on the coefficients of the seasonal MLR model for Cohoes city and their covariance matrix from subsection 2.6.5 Equation 2.8; this is an informative prior density that constructs the IN-BMLRCohoes.

Table 4.4: The Coefficients of MLR, NI-BMLR, IN-BMLRCohoes and NI-BMLRUtica, for the Long-Term Component.

Parameter	MLR	NI-BMLR	IN-BMLRCohoes	IN-BMLRUtica
IN	0	0	0	0
TE	-0.52	-0.52	-0.54	-0.52
WS	0.07	0.07	0.008	0.06
PR	0.37	0.37	0.35	0.37
TD	0.30	0.30	0.27	0.30
GW	-0.33	-0.33	-0.39	-0.33

Table 4.5: MSE values for the Long-Term Component.

Model	NI-BMLR	IN-BMLRUticaCohoes	NI-BMLRUtica
MSE	0.47	0.49	0.47

- The final prior is a non-informative prior which is an uniform distribution with the maximum likelihood estimates (MLE) of the regression parameters for the seasonal component for Utica city from subsection 2.7.7 Equation 2.13 as the starting values for the simulation process using Gibbs sampling. This will produce the NI-BMLRUtica model. The results are shown in Table 4.6. Again, because the means of the coefficients are almost identical for the MLR, NI-BMLR, and NI-BMLRUtica, the MSE value is 0.64 for these models. However, the result for the IN-BMLRCohoes model is different.

Table 4.6: The Coefficients of MLR, NI-BMLR, IN-BMLRCohoes, and NI-BMLRUtica, for the Seasonal Variations.

Parameter	MLR	NI-BMLR	IN-BMLRCohoes	IN-BMLRUtica
IN	0	0	0	0
TE	0	0	0.04	0
WS	0.14	0.14	0.08	0.14
PR	0.39	0.39	0.43	0.39
TD	0	0	-0.03	0
GW	-0.43	-0.43	-0.44	-0.43

Prediction Modelling for the Short-Term Component

The short-term component has been analysed using a manner that is similar to the method above where the results of MLR from subsections 2.6.6 and 2.7.9 for Equations

2.9 and 2.14 have been used to accomplish this analysis. The coefficients results are shown in Table 4.7 and the results for the MSE are shown in Table 4.8. In Table 4.7, except the third column, there is no difference between the results of the coefficients. In Table 4.8, identical values for the MSE have been obtained for the NI-BMLR and NI-BMLRUtica and different MSE value has been noticed for IN-BMLRCohoes where this value is the smallest one, and this is the first time in our analysis we obtain better results using Cohoes city data rather than Utica city data.

Table 4.7: The Coefficients of MLR, NI-BMLR, IN-BMLRCohoes, and NI-BMLRUtica for the Short-Term Component.

Paramter	MLR	NI-BMLR	IN-BMLRCohoes	NI-BMLRUtica
IN	0	0	0	0
TE	0.05	0.05	0.01	0.05
WS	0.06	0.06	0.04	0.06
PR	0.42	0.42	0.42	0.42
TD	0.01	0.01	0.00	0.01
GW	-0.20	-0.20	-0.29	-0.20

Table 4.8: MSE values for the Short-Term Component.

Model	NI-BMLR	IN-BMLRCohoes	NI8-BMLRUtica
MSE	0.74	0.73	0.74

4.11.4 Contribution Percentages for the Decomposed Data

The contribution of the different scales of motions, which are embedded in a time series, can be computed by utilising the results of the KZ filtering mechanism. The output for this analysis are shown in Table 4.9. Firstly, for Bayesian analysis for the

Table 4.9: Results of the Variance and the Coefficient of Determination.

	Variance	R Squared
Long-Term Component	55.86	0.70
Seasonal-Term Component	4.36	0.35
Short-Term Component	26.29	0.23

long-term pattern, the proportion of the variance for the long-term component series is multiplied by the value of the R-squared for this component, (55.86×0.70), and the result is 39.57. Mathematically, the proportion of the water discharge variance

for each component can be computed by dividing the variance of each component by the variance of the original water discharge series (the water discharge series before the decomposition process). Using the same procedure, the contributions of the other components can be computed. While the seasonal component contributes with about 1.53 (4.36×0.35), the contribution of the short-term component using the regression analysis approach is 6.22 (26.29×0.23).

4.12 Combined Bayesian Multiple Linear Regression (CBMLR) Model

Table 4.10 shows the results of the three posterior distributions that have been constructed by combining the three components, long, seasonal, and short, together. It is clear that almost same results have been obtained by specifying the non-informative and the informative prior distributions that are based on the MLR results for Cohoes and Utica cities. For the Non-Informative Combined BMLR (NI-CBMLR) model, which has been constructed by specifying a non-informative prior distribution, the MSE has reduced to 0.43 compared to the MSE of the BMLR model that has been built by specifying the non-informative prior distribution for the raw data, which was 0.56, as shown in Table 4.11. This result clearly indicates that the decomposition technique has successfully improved the forecasting process of the water discharge for Utica city.

Component	NI-CBMLR	IN-CBMLRCohoes	NI-CBMLRUtica
Intercept	0.000859	0.000476	-0.00035
LT-Temperature	-0.415	-0.413	-0.413
LT-Wind speed	0.044	0.047	0.047
LT-Precipitation	0.310	0.309	0.308
LT-Tide	0.213	0.213	0.213
LT-Ground	-0.218	-0.219	-0.218
SE-Precipitation	0.092	0.097	0.097
SE-Groundwater	-0.185	-0.184	-0.185
SH-Precipitation	0.198	0.204	0.204
SH-Ground-Water	-0.168	-0.169	-0.167

Table 4.10: The Coefficients of NI-CBMLR, IN-CBMLRCohoes, and NI-CBMLRUtica Models for Long-Term (LT), Seasonal-Term (SE), and Short-Term (SH).

Table 4.11: DIC Values for the Combined Models with Different Types of Priors.

Model	NI-CBMLR	IN-CBMLRCohoes	NI-CBMLRUtica
MSE	0.43	0.45	0.43
DIC	6616.294	6619.684	6616.405

4.13 Bayesian Vector Auto Regressive (BVAR) Model for Short-Term Component

Bayesian VAR model has been used to analyse the data of the short-term component. This model can adequately represent data that has short-time periods. It has been previously shown that the two variables of the precipitation and groundwater level have significant relationships with the water discharge for the short-term component using the correlation matrix. Therefore, the BVAR model has been built using these three variables. Essentially, this model is just a multivariate model as three variables are considered as dependent (response) variables.

The parameters of this model are the AR coefficients and the covariance matrix of the model. For the priors of these two parameters, the multivariate normal distribution (MN) and Inverse-Wishart distribution (IW) have been specified. Practically, it is easier to apply the precision matrix instead of the covariance matrix, in this case, the Wishart distribution, W, will be used instead of the Inverse-Wishart. Different hyper-parameters have been assigned to the coefficients of the BVAR model.

For the VAR model, which is expressed in Equation 4.9, if $u_t \sim N(0, \Sigma_u)$, the following multivariate normal distribution can be specified

$$\boldsymbol{\varphi}|\Sigma_u \sim MN(\boldsymbol{\varphi_0}, V_{\boldsymbol{\varphi}} = V \otimes \Sigma_u)$$

and

$$\Sigma_u \sim IW(S, n).$$

In case that we consider the precision matrix rather than the variance matrix, we will have

$$\Sigma_u^{-1} \sim W_k(S^{-1}, n).$$

Based on the priors that have been assigned to the coefficients and the variance of the model, the resulting posterior distribution is Normal-Inverse-Wishart distribution.

By imposing prior distributions on the VAR parameters, the BVAR can be used successfully to overcome problems of over-fitting (over-parametrization) and collinearity that appear when VAR models are applied [16]. For the purpose of fitting a BVAR model, the Minnesota prior, which is a special case of the conditional Normal-Inverse-Wishart distribution, has been used. Based on the Minnesota prior, we are able to

Table 4.12: The Parameter Estimates for the Short-Term of the Precipitation (PR) Using BVAR model.

PR			
Parameter	Estimate	P	Variable
AR11,1	0.84926	**0.0001**	PR(t-1)
AR11,2	0.01189	0.4791	WD(t-1)
AR11,3	-0.01098	0.8054	GW(t-1)
AR21,1	-0.10647	**0.0001**	PR(t-2)
AR21,2	0.16495	**0.0001**	WD(t-2)
AR21,3	-0.01172	0.8622	GW(t-2)
AR31,1	-0.13533	**0.0001**	PR(t-3)
AR31,2	-0.20655	**0.0001**	WD(t-3)
AR31,3	0.06508	0.1051	GW(t-3)

assign any value for the mean vector while two parameters control the covariance matrix of the parameters of the VAR model. These two parameters are α and θ. The first parameter, α, is the standard deviation of the parameter of interest. As the value of this parameter increases, the BVAR of order p model becomes similar to a VAR of order p model. The other parameter, which is θ, controls the relative tightness of the prior variance for the lags in the equation of the required variable. The value of this parameter is in the interval (0,1) and whenever this value approaches 1, the chosen BVAR of order p model approaches a VAR of order p model [16, 64].

The mean vector is set up based on the coefficient estimates from the classical VAR for the lagged variables for the targeted variable in its equation and the value zero has been specified for all the other coefficients. With regard to the hyper-parameters of the covariance matrix of the parameters, α and θ, the values 0.9 and 0.1 have been selected. In addition to choose the values for these parameters, the covariance matrix of the model (disturbance) terms needs to be estimated to compute the elements of the covariance matrix of the parameters. The covariance matrix of the disturbance terms is a diagonal matrix that is estimated using equation-by-equation AR models. The results of this analysis for the three response variables, which are precipitation, water discharge, and groundwater level, for the short-term component using BVAR model are shown in Tables 4.12, 4.13, and 4.14, respectively. The P-values for the significant coefficients have been highlighted in the these Tables.

Using the Random Walk Metropolis MCMC algorithm, the results of the posterior distributions for the parameters of the BVAR are shown in the plots 4.3, 4.4, and 4.5.

Table 4.13: The Parameter Estimates for the Short-Term Component of Water Discharge (WD) Using BVAR model.

WD			
Parameter	Estimate	P	Variable
AR12,1	0.22612	**0.0001**	PR(t-1)
AR12,2	0.82271	**0.0001**	WD(t-1)
AR12,3	-0.0367	0.4185	GW(t-1)
AR22,1	-0.26464	**0.0001**	PR(t-2)
AR22,2	-0.14498	0.0001	WD(t-2)
AR22,3	0.06492	0.3446	GW(t-2)
AR32,1	0.10595	**0.0001**	PR(t-3)
AR32,2	0.0278	0.1102	WD(t-3)
AR32,3	-0.01976	0.6286	GW(t-3)

Table 4.14: The Parameter Estimates for the Short-Term Component of Groundwater Using BVAR model.

GW			
Parameter	Estimate	P	Variable
AR13,1	-0.01272	**0.0221**	PR(t-1)
AR13,2	-0.05377	**0.0001**	WD(t-1)
AR13,3	1.49904	**0.0001**	GW(t-1)
AR23,1	0.0175	**0.007**	PR(t-2)
AR23,2	0.02406	**0.0002**	WD(t-2)
AR23,3	-0.59698	0.0001	GW(t-2)
AR33,1	-0.01101	**0.025**	PR(t-3)
AR33,2	0.00305	0.5351	WD(t-3)
AR33,3	0.04237	**0.0147**	GW(t-3)

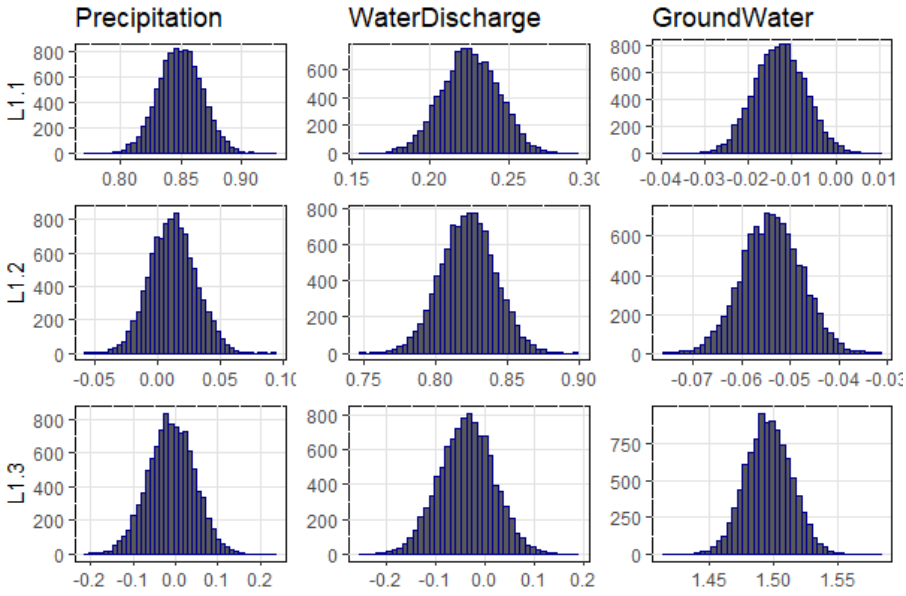

Figure 4.3: Posterior Distributions for the Parameters for the First Lag.

4.14 The Final Combined Bayesian model with BVAR for the Short-Term Component

For the purpose of constructing the final model, the Bayesian analysis has been used for analysing the combined model, and the results are shown in Table 4.15. Table 4.16 shows the results of the diagnostic statistics, which are DIC and MSE, and based on the fact that the smallest DIC and MSE test statistics indicate the best fitting model, the second model, is the best model. This model has been built using the current variables for the long and seasonal component as well as the first three lags for the variables of water discharge, precipitation, and groundwater level for the short-term component. The DIC value for this model is 521.385. The second best model, which has the DIC value 684.782, has been constructed using the current values of the long and seasonal components and the first three lags of the variables of precipitation and groundwater level. This would mean that the inclusion of a number of lagged variables has improved the forecasting model based on the given diagnostic statistics. Similar results can be noticed for the MSE values.

In a similar way, Bayesian analysis has been applied to the raw and the decomposed data for Cohoes and Poughkeepsie cities. For the raw data, the same three kinds of priors that have been used to analyse the data of Utica city have been also used. These prior distributions are: (1) a non-informative independent nor-

Table 4.15: Results for Long-Term (LT), Seasonal-Term (SE), and Short-Term (SH) Components for Utica City.

Variables	LT Co.	SE Co.	Variables	SH Co.
intercept	0.17	0	Lag1PR	0.114
TE	0.027	-0.0007	Lag1WD	0.483
WS	0.242	0.049	Lag1GW	-0.188
PR	0.203	0.066	Lag2PR	-0.212
TD	0.011	0.12	Lag2WD	-0.11
GW	-0.217	-0.176	Lag2GW	0.114
			Lag3PR	0.065
			Lag3WD	0.061
			Lag3GW	-0.045

Table 4.16: Model Diagnostic Checks For the Final Model with BVAR for the Short-Term Component.

	DIC	MSE
Full model without Lags for the WD in the short-term Co.	684.782	0.244
Full model with Lags for the variables of the short-term Co.	521.385	0.102
Full model without lags for the short-term Co.	6593.116	0.33

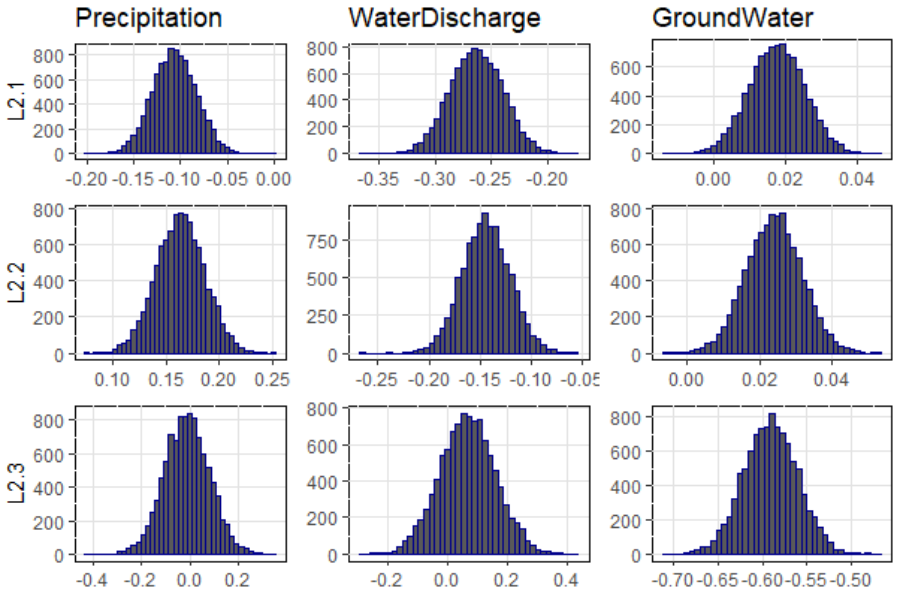

Figure 4.4: Posterior Distributions for the Parameters for the Second Lag.

mal prior distribution with mean zero and variance 1000000 for each parameter, i.e. $\boldsymbol{\beta} \sim N(0, 1000000 I_6)$, (2) an informative independent normal prior distribution with mean and variance that are based on the results of the regression analysis for Utica city, (3) a non-informative uniform distribution with constants for all the regression coefficients and their variances and a non-informative gamma prior distribution for the normal scale (variance) parameter of the model. The nature of the results for Bayesian analysis for these two cities is relatively similar to the results for Utica city in terms of: (1) the parameters estimates, where approximately the same regression coefficients obtained using the first and third prior distributions and these coefficients are, in turn, approximately similar to the classical MLR model's coefficients. (2) The DIC values for Cohoes city data for the two BMLR models constructed using the first and third priors are 7130.62 and 7117, respectively. (3) Compared to the coefficients resultant from using the first and third prior distributions, however, the parameters estimates using the second prior distribution are slightly different. These parameters estimates are -0.2946, 0.01, 0.2605, 0.1401, -0.4717 for temperature, wind speed, precipitation, tide, and groundwater level, respectively, also the DIC value for this model is 7165.738. For Poughkeepsies city, the coefficients are also approximately similar to the coefficients of the classical MLR for the first and third priors. However, the results of using the third prior are different, where the values are -0.3463, 0.03, 0.4274, 0.1596, -0.3204 for temperature, wind speed, precipitation, tide, and groundwater level, respectively. The DIC values are 7661.340, 7690.768, and 7650.323 for

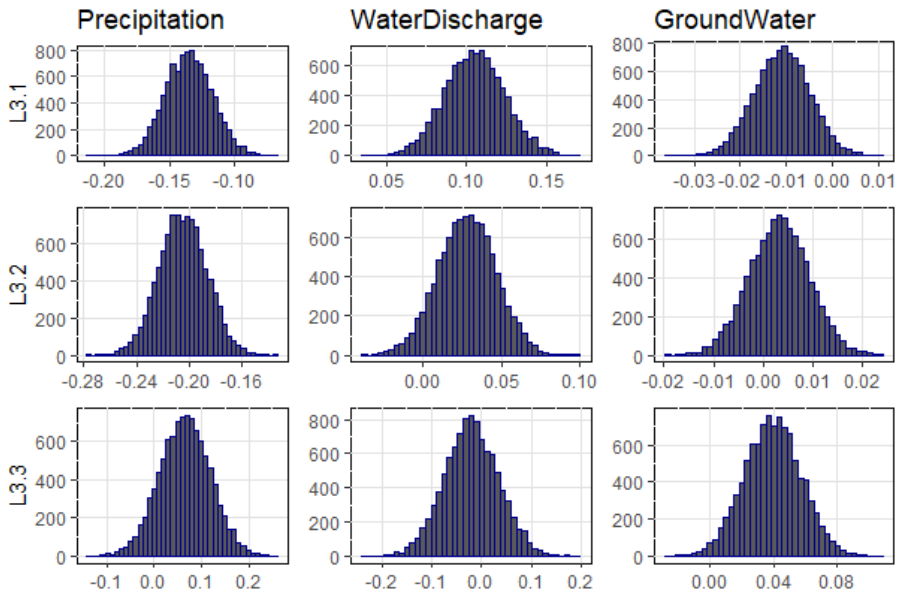

Figure 4.5: Posterior Distributions for the Parameters for the Third Lag.

the three models.

For the decomposed data for Cohoes and Poughkeepsie cities, the above-mentioned three prior distributions have been applied to each component, which are the long, seasonal, and the short-term components. With regard to the coefficients, using the first and third prior distributions, approximately the same parameters estimates have been obtained, which are similar to the classical MLR model's coefficients for the three components. For Cohoes city for the long-term component using the second prior, the coefficients are -0.51, -0.09, 0.32, 0.25, -0.45 for temperature, wind speed, precipitation, tide, and groundwater level, respectively. For the short-term component, the BVAR model has been also applied. For Cohoes and Poughkeepsie cities, the final combined models are constructed using the BMLR model for the long and seasonal components and BVAR model for the short-term component. The DIC values are 6376.183 and 5255.887 for Cohoes and Poughkeepsie cities, respectively. The results are shown in Tables 4.17 and 4.18.

4.15 Results

To forecast a daily future value for water discharge, we have built different models. These models can be summarised as follows:

1. A BMLR model has been constructed using the raw data where all the coeffi-

Table 4.17: Results for Long-Term (LT), Seasonal-Term (SE), and Short-Term (SH) Components for Poughkeepsie City.

Variables	LT Co.	SE Co.	Variables	SH Co.
intercept	0	0	Lag1PR	0.108
TE	-0.442	0.06	Lag1WD	0.547
WS	-0.007	0.015	Lag1GW	0.015
PR	0.353	0.089	Lag2PR	-0.114
TD	0.143	0.030	Lag2WD	-0.165
GW	-0.334	0.011	Lag2GW	0.015
			Lag3PR	-0.03
			Lag3WD	0.091
			Lag3GW	0.013

Table 4.18: Results for Long-Term (LT), Seasonal-Term (SE), and Short-Term (SH) Components for Cohoes City.

Variables	LT Co.	SE Co.	Variables	SH Co.
intercept	0	0	Lag1PR	0.153
TE	-0.441	0.024	Lag1WD	0.457
WS	-0.069	0.011	Lag1GW	-0.043
PR	0.295	0.066	Lag2PR	-0.154
TD	0.174	0.044	Lag2WD	0.109
GW	-0.357	-0.165	Lag2GW	0.015
			Lag3PR	0.01047
			Lag3WD	0.06092
			Lag3GW	-0.13725

cients are significant and the R-squared value is 0.43. That means 0.43 of the variations are due to the considered covariates, which are temperature, wind speed, precipitation, tide, and groundwater level.

2. In an attempt to improve this relatively weak relationship, the KZ filter is used to decompose the data into long, seasonal, and short-term components. Based on these components, a number of BMLR models have been fitted using different prior distributions.

3. For the decomposed data, the BMLR models constructed by using informative and non-informative prior distributions provide the R-squared values that are shown in Table 4.19.

Table 4.19: R Squared Values for the Constructed Models.

Model	IN-Combined-BMLRCohoes	NI-Combined-BMLRUtica
R-Squared-Long	0.69	0.70
R-Squared-Seasonal	0.35	0.34
R-Squared-Short	0.24	0.23

4. Calculating the contribution percentages of the Long, Seasonal, and Short-Term components for the BMLR models has revealed that the order of the contribution from the highest to the lowest in the data is assigned to the long, short, and seasonal component, respectively.

5. By combining the three components to elicit the CBMLR model, the R squared value becomes 0.56. Also, the significant coefficients are temperature, precipitation, tide, and groundwater level for the long-term component; precipitation, tide, and groundwater level for the seasonal data; precipitation and groundwater level for the short-term component's data.

6. With reference to the results of similarity, we have noticed that, although the statistical distance (SD) between Utica and Cohoes is the lowest, the value of the MSE was not the smallest for the raw data. Practically, we have discovered that the reason for the high MSE for the raw data using the Cohoes parameters may be attributed to the small values for the variances of the coefficients. When we replaced them with relatively large values, for instance, 10 and 100, we obtained results that are similar to the results of the MLE method. We also discovered that the covariances between the coefficients have no influence on the results as much as the variances, for example, we tried different values, small and large, but the results did not change.

However, for the decomposed data, the results were better. This would mean that we are able to use the parameters of the regression models for Cohoes city as hyper-parameters for the priors for Utica city in case that data are not available for Utica city. These findings have been also noticed on the data of Cohoes city when we used the parameters of the regression model for Utica city as hyper-parameters for Cohoes city and also a simulated data. Therefore, to use the results of the similarity analysis, we need to check the variances between the coefficients and it will be suitable to use the results of the decomposed data to specify hyper-parameters and also the raw data after checking the variances for the parameters.

7. When the BVAR model is applied to the short-term component, better result has been obtained for the final combined Bayesian model compared to the results obtained using the combined BMLR. This is revealed by using the DIC and MSE procedures.

8. Based on our analysis, it is not recommended to assign small values (near to zero) as variances for the coefficients.

4.16 Conclusion

The provision of an accurate flood forecasting model is a vital task. All forecasts include uncertainty and one of the most successful methods of dealing with this uncertainty is the use of a structure that inherently considers this uncertainty. The uncertainty in a hydrological forecast is principally associated with the meteorology, specifically, what is related to the precipitation variable, model parameters, and model's structure.

Based on this, the structure of Bayesian methods is the optimal structure to be used as the uncertainties are explicitly considered. The main elements in the Bayesian analysis are the posterior and posterior predictive densities for a set of variables of interest. The variables are the parameters of the model and the future values. Applying the Bayesian analysis can lead to distributions either with known or unknown kernels. Both types can be solved using a specific sampling technique such as Gibbs sampling by generating a number of sampled data.

It is important to ensure that the results of the simulation process using MCMC is converged. The most important feature in the Bayesian analysis is the ability of obtaining credible intervals or HPD intervals, which can be computed directly (by using the probability density function) or indirectly (by using generated data from MCMC method). The predictive distribution can also be exploited to supply a risk-based approach, i.e. to produce the guarantee that the variable of interest will be

within a determined range of data in the future. The predictive distribution is of particular interest as it is considered as the most useful tool to assess the constructed model.

In this chapter we develop a methodology that takes into account the uncertainties concerning the model's parameters and structure and also the effect of the embedded components in the time series considered. Similar results are obtained using the MH-MCMC and MLE methods for the parameters of interest and the confidence intervals for the raw data. By applying the Combined BMLR and BVAR models, the DIC and MSE values have declined with a remarkable discrepancy compared to these values of the raw data. For instance, the DIC values decline from 7445.291 for MLR constructed using NI-BMLR for the raw data to 521.385 using BMLR for the long and seasonal components and BVAR(3) model for the short-term component. Also, for each predicted data, credible and HPD intervals are obtained. In addition, for each parameter, a HPD interval is computed. It is also important to mention that the similarity between the confidence and credible intervals results can be attributed to the type of the prior distribution specified to the parameters. Based on the results of the three cities, there is no significant impact of the similarity results on the forecasting accuracy of water discharge values based on the DIC values. The result is derived when an informative prior distribution with hyper-parameters related to the MLR results for the cities that have the highest similarity measure is used.

Chapter 5

Hypothesis Testing For Dissimilarity Analysis

Often, in time series analysis, it is required to examine whether the population means for $p \times 1$ random vectors of variables for two subjects (multivariate time series datasets) are equal. In this case, the null hypothesis is $\mu_1 = \mu_2$, where μ_1 and μ_2 are the mean vectors for the variables for the two examined subjects. Also, examining whether the population means of dissimilarity/distance for a number of subjects are equal is an important topic [40, 62]. The null hypothesis for dissimilarity analysis is $\mu_1 = \mu_2$, where μ_1 and μ_2 are the mean distances for the two subjects of interest. The dissimilarity analysis is the corner stone for most of the data mining and pattern recognition methods.

If the null hypothesis using, for example, mean or median of raw data is rejected, it is often not possible to know which component is responsible for this rejection. The determination of this component will, in turn, lead to know the variables and the events that are responsible for this rejection. By filtering out the impact of undesirable signals using a filtering technique, a number of issues in different analysis, such as dissimilarity analysis and data mining techniques (clustering, classification, etc), can be avoided. For example, to cluster a number of countries based on the effectiveness of regulatory programs and initiatives in improving ozone air quality, it is crucial to use an ozone air series that is devoid of any meteorological fluctuations. Filtering out the influence of meteorological fluctuations on ozone concentrations using a filtering technique will help to detect changes in ozone air quality due to changes in emissions.

In this chapter, we present a new method to detect the component(s) that influences the results of hypothesis testing that are performed using the raw data for two cases. Firstly, we compare two mean vectors for variables for two multivariate time series datasets using Hotelling T-Squared test. Secondly, we compare two mean distances (dissimilarities) using the one way Analysis of Variance (ANOVA)

and Kruskal Wallis tests for two multivariate time series datasets. In the context of multivariate time series data mining, the essential problem is how to represent the data of a time series [40]. One of the most common methods is converting the time series to another domain for dimensionality reduction. Covariance matrix is an example for this converting process [34]. Based on this, covariance matrices are used to compute the distance between two multivariate time series. For the raw and the decomposed data, the distance between two covariance matrices are computed using a number of Euclidean and Non-Euclidean metrics.

Sometimes, for reasons such as wars or natural disasters, it is not feasible to provide a forecasting model for a specific city. So, we examine the feasibility of using a forecasting model for one city to forecast future values for another city. However, to be able to use this option, we need to check whether the data of cities of interest are similar. Based on this, some of the hypothesis testing findings can be used to make the decision of similarity. In particular, the hypothesis testing results that are related to the covariance matrices structured by using the independent variables for a regression model are used. To support the claim of proximity between the data of the cities, some parameter-based statistics can be used. For example, the Mean Square Error, MSE, which is computed for a regression model constructed using the same independent variables that are used to build the covariance matrices, can be utilised. Also, to ensure that the data for the two cities are similar, two samples T-Test has been used to compare means for each variable for the two cities of interest. The questions that are required to be answered are the following:

- Does hypothesis testing for comparing mean vectors for variables using the raw data convey the full picture about the differences between the studied groups? Equivalently, it can be rewritten as the following: Is there any hidden information about two multivariate time series datasets and has not been captured by comparing the mean vectors for the variables or mean distances using the raw data?

- Could applying hypothesis testing for the data for the three components show different results to what has been obtained using the raw data?

- Are we able to determine the component(s) and then the variables that are responsible for the differences (if they exist) if testing the raw data reveals some differences between the considered groups?

- If the effect of a special group of variables, such as hydrological variables, is ignored, could the results of hypothesis testing change?

This chapter is organised as follows. Section 5.1 provides a brief description of the decomposed data, comparing mean vectors for variables, and dissimilarity analysis.

Section 5.2 shows how to examine whether two mean vectors for variables for two subjects are equal using two samples Hotelling T-Squared test. Section 5.3 introduces a brief overview of a number of common types of distance measures for Univariate and Multivariate time series. Section 5.4 presents a brief overview of hypotheses testing. Section 5.5 displays the methodology for comparing two mean vectors for the variables for two subjects. This section is divided into two subsections. Subsection 5.5.1 explains the two samples Hotelling T-Squared for the raw and the three components' data. Subsection 5.5.2 considers the case when the impact of the responsible data, which is the hydrological data, is removed. The hydrological data is represented by the variables of the water discharge and groundwater level.
Section 5.6 is also subdivided into two subsections. Subsection 5.6.1 applies a number of Euclidean and Non-Euclidean metrics to the covariance matrices for the raw and the decomposed data to examine the similarity between two subjects. The second subsection, 5.6.2, seeks to assess the impact of removing the hydrological data on the dissimilarity results. In Section 5.7 a comparison between the distance measures is performed.

So far we have dealt with the dissimilarity analysis using time domain's data and methods. As long as our data is time series data, it is necessary to examine the dissimilarity between the subjects using frequency domain data and methods. To examine the dissimilarity based on the frequency domain's data, we present three new developed dissimilarity measures and their applications to a group of multivariate time series datasets. These three dissimilarity measures are covered in Section 5.8. In the first subsection, 5.8.1, we show how to construct a Power Spectral Density (PSDE) matrix and also how to apply the existing Euclidean and Non-Euclidean metrics to pairs of these matrices, and using this as a measure of similarity of the matrices in the frequency domain. In the second subsection, 5.8.2, we use the Eigenvalues for this matrix, PSDE matrix, for the raw data as a dissimilarity measure using the Euclidean distance. In the third subsection, 5.8.3, we discuss how to use the X^TX matrix for the Periodograms of the variables for the raw data as a dissimilarity measure also in the frequency domain. Only the raw data has been used to examine the performance of these new developed dissimilarity measures. A discussion of the results obtained is presented in Section 5.9. Finally, Section 5.10 presents the conclusion of this chapter.

5.1 The Decomposed Data, Comparing Mean Variables, and Dissimilarity Analysis

In univariate and multivariate time series analysis, separating different scales of motions, which is known as the decomposition or filtering process, is often a pivotal technique that is required to be applied in the process of building a time series model

[121, 36]. Although the decomposed data has confirmed its ability to provide better results in the modelling process than the raw data in terms of forecasting accuracy [101, 100], no study, as far as we know, has attempted to shed a light on its use in another area, such as hypothesis testing for comparing means, similarity analysis, data mining, and machine learning algorithms. In this chapter we extend the use of the decomposed data to be included in two of the aforementioned applications, which are the hypothesis testing and the dissimilarity analysis.

Typically, the first step in the analysis of multivariate datasets is computing the mean vector and the covariance matrix. Both of these statistics, mean vector and covariance matrix, can be exploited to study the relationships between two subjects (multivariate time series datasets) [61]. The most common test to perform this comparison is the two samples Hotelling T-Squared test. This test is typically applied based on the mean vectors for the variables for two subjects [46].

On the other hand, the covariance matrix, which contains the variances of the considered variables along the main diagonal and the covariances between each pair of variables in the off-diagonal entries, can also be used to do a comparison to investigate the dissimilarity of datasets between two or more subjects. There is a growing body of literature that recognises the importance of the statistical covariance analysis, [5], in similarity-based techniques in different areas such as diffusion tensor imaging [120, 34] and longitudinal (panel) data [34]. The covariance matrix for a multivariate probability distribution is always a Positive Semi-Definite (PSD) matrix [34, 5]. This kind of matrices can also be seen in different areas, for example, many applications in the computer vision include features that can be represented using this matrix.

According to Dryden et al. (2009), the natural space for the covariance matrix is a Riemannian space [34]. Based on this, it is more suitable to use a Non-Euclidean metric rather than a Euclidean metric to calculate the distance. Additionally, since time series data is a special case of longitudinal data [53], using a Riemannian metric is a more suitable choice to compute the dissimilarity magnitude between two multivariate time series datasets than the Euclidean metrics. By applying hypothesis testing, we can use the distance between two covariance matrices to deduce whether the datasets for two subjects are similar. Different Euclidean and Non-Euclidean metrics have been used to assess whether they provide the same decision for the null hypothesis. In case that the main analysis is clustering or classifying a number of multivariate time series datasets, the performance of each Euclidean and Non-Euclidean metric can be assessed.

The distance between time series datasets is a fundamental subject in several fields, such as meteorology, business, economics, finance, medicine, hydrology, astronomy, seismicity, and many others [113, 17]. In fact, the results of similarity or dissimilarity are also necessary to apply most of the data mining techniques, which is an increasingly important area in applied statistics [26, 95]. As a result, a variety

of methods have been proposed to handle the vast amount of information that is available on the web. For example, in meteorology we may be interested in clustering or classifying a number of countries based on temperature time series data. In economics, it is required to classify or discriminate some countries depending upon some time series indicators, such as unemployment or inflation rates. Therefore, an increasing interest towards similarity and dissimilarity measures for comparing time series data has been noticed [40]. In addition, this increase is also attributed to the advances in the computers and numerical algorithms [113].

The topic of identifying the similarities and dissimilarities between time series data has been specifically studied in the clustering, classification, and discriminant analyses literature [68, 95, 17]. Moreover, in multivariate time series analysis, specifically in economics and environmental studies, often a large number of time series datasets needs to be analysed. This, in turn, means large memory space and long time are required to process them. In some data mining and machine learning techniques, such as similarity analysis, classification, and clustering, the working data is not necessarily the raw data. In many cases, it can be replaced with another data form, for example, covariance and correlation matrices. This process of replacement has been already extensively applied in many applications [40].

5.2 Two Samples Hotelling T-Squared Test

Let R be a set of N subjects. For each subject, p outcome variables $X_{..1}, \ldots, X_{..p}$ at different times are measured. The sequence for the subject i at time j can be written as the following:

$$x_{ij.} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{pmatrix}$$

where $i = 1, 2, \ldots N$, $j = 1, 2, \ldots, t_i$, and $k = 1, 2, \ldots, p$.

The null and alternative hypothesis for the Hotelling T-Squared test are:

$$H0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$
$$H1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

where $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1p})$ is a p-dimensional vector, where $\mu_{11}, \mu_{12}, \ldots, \mu_{1p}$ are the means for the variables for the first subject; $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \ldots, \mu_{2p})$ is a p-dimensional vector, where $\mu_{21}, \mu_{22}, \ldots, \mu_{2p}$ are the means for the variables for the second subject. The null and alternative hypotheses can also be written in a vector form, for example,

the null hypothesis can be replaced with:

$$H0 : \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix}.$$

The formula for the two samples Hotelling T-Squared test is defined as the following:

$$T^2 = (\bar{\mathbf{x}_1} - \bar{\mathbf{x}_2})' \left\{ S_P (\frac{1}{t_1} + \frac{1}{t_2}) \right\}^{-1} (\bar{\mathbf{x}_1} - \bar{\mathbf{x}_2}) \tag{5.1}$$

where $\bar{\mathbf{x}_1}$ is the sample mean vector for the first subject and $\bar{\mathbf{x}_2}$ is the sample mean vector for the second subject, the dimension for these vectors is $p \times 1$. Also, S_P is a pooled covariance matrix, which can be defined as follows:

$$S_p = \frac{(t_1 - 1)S_{\mathbf{x}1} + (t_2 - 1)S_{\mathbf{x}2}}{(t_1 - 1) + (t_2 - 1)}$$

where S_{x1} and S_{x2} are the covariance matrices for the two subjects of interest, t_1 and t_2 are the samples sizes for the two subjects.
However, to account for the effect of the serial dependency (autocorrelation) as the data here is time series data, two solutions are common. The first solution is carried out by calculating the Hotelling T squared value using the residuals rather than the raw data. These residuals are obtained using one of the time series models, such as ARMA models. The second solution is by reconstructing the covariance matrix using Standard Deviation and Correlation matrices, which has been used here, for more information see [106, 76, 32]. In this case, the T-squared value can be written as follows:

$$T^2 = (\bar{\mathbf{x}_1} - \bar{\mathbf{x}_2})' \S^{-1} (\bar{\mathbf{x}_1} - \bar{\mathbf{x}_2}) \tag{5.2}$$

where $\S^{-1}$ is the sample covariance matrix that is expressed using the standard deviation and first-Order Autocorrelation matrix. Moreover, if t_1 and t_2 are large enough, the value of T^2 will have a chi-squared distribution with p degrees of freedom.

5.3 Types of Distance Measures based on the series number

Distance measures are split into two kinds based on the number of the considered series. These types can be summarised as follows.

5.3.1 Distance Measures for Univariate Time Series

There are many similarity, dissimilarity, and distance measures specifically proposed to deal with univariate time series datasets. For example:

- Piccolo (1990) introduced a metric for the ARIMA models as the Euclidean distance between the parameters of the ARIMA models [17, 26]. Let TS_1 and TS_2 be two time series; $\pi_{j,x}$ and $\pi_{j,y}$ be two parameter vectors for the two time series. This metric can be written as:

$$d_{PIC}(TS_1, TS_2) = \sqrt{\sum_{j=1}^{n}(\pi_{j,x} - \pi_{j,y})^2}.$$

- Galeano and Peña (2000) [17] suggested the Sample Autocorrelation function (SACF)-based dissimilarity metric, which can be defined as:

$$d_{SACF}(TS_1, TS_2) = \sqrt{(\rho_{\hat{TS_1}} - \rho_{\hat{TS_2}})'\Omega(\rho_{\hat{TS_1}} - \rho_{\hat{TS_2}})}$$

  where $\rho_{\hat{TS_1}}$ and $\rho_{\hat{TS_2}}$ are two SACF vectors for a number of lags, and Ω is the weight matrix. If Ω equals to identity matrix, then the resulting formula is simply the Euclidean distance. Also, when Ω is the inverse of the variance-covariance matrix between the two SACF vectors, the resulting formula is the Mahalanobis distance, which is

$$d_{MAH}(TS_1, TS_2) = \sqrt{(\rho_{\hat{TS_1}} - \rho_{\hat{TS_2}})'S^{-1}(\rho_{\hat{TS_1}} - \rho_{\hat{TS_2}})}.$$

- The Sample Partial Autocorrelation, SPACF, and Inverse Sample Autocorrelation Function, ISACF, can also be used as a dissimilarity measure between two time series.

- Caiado et al. (2006) proposed the periodogram-based dissimilarity metric. This metric uses the Euclidean distance for the periodograms for the two time series of interest as a dissimilarity measure [17]. This metric can be defined as:

$$d_P(TS_1, TS_2) = \sqrt{\sum_{j=1}^{n/2}\left[P_{TS1}(\omega_j) - P_{TS2}(\omega_j)\right]}$$

  where $P_{TS1}(\omega_j)$ and $P_{TS2}(\omega_j)$ are the periodograms that have been calculated using the Fourier Transform for $n/2$ frequencies, where ω_j denotes the frequency.

5.3.2 Distance Measures for Multivariate Time Series

Multivariate time series (MTS) datasets can be found in different fields, such as finance, meteorology and hydrology. Examining the similarity between Multivariate time series datasets has been widely considered [113, 40]. Each series possess a number of characteristics in time and frequency domains and they are often used to build distance measures. Different output can be obtained to represent the distance (dissimilarities) between two matrices. A single value, a vector, and a matrix can be obtained based on the used data and the distance measure. Calculating pairwise distances between the rows of two design matrices returns a distance matrix. The design (data) matrix is a matrix in which each row represents an individual object and each column represents the value for this object with the corresponding variable. Fundamentally, the distance matrix is used in data minings or pattern recognition algorithms, such as clustering and classification. To measure the distance between two Positive Semi-Definite matrices (PSD), by using, for example, the Euclidean or Mahalanobis measures, a single value is returned. However, the Euclidean distance suffers from many defects on PSD matrices [5].

Recently, researchers have shown an increased interest in the use of Non-Euclidean metrics to study some data minings techniques such as clustering and classification [68]. This kind of metrics takes into consideration the Non-Euclidean nature for the space for positive semi-definite symmetric matrices [107]. Covariance matrix is one of the most important and used structures in the applications of Non-Euclidean metrics[34, 5]. Using Non-Euclidean metrics has confirmed its ability to provide better results in a number of analyses, such as cluster and classification, specifically in diffusion tensor imaging [34] than Euclidean-based metrics [68].
In multivariate time series analysis where often large datasets need to be analysed, it will be practically more suitable to use any shortened form of data instead of the raw data [40]. In light of this, using a sample of covariance matrices as the data to be analysed would be a reasonable choice. Below there is a number of distance measures particularly used with the PSD matrices [107, 34], for example, the covariance matrix in the time domain and the power spectral density matrix in the frequency domain.

- The Euclidean Distance between two covariance matrices, S_1 and S_2, can be written as:

$$D_{EU}(S_1, S_2) = \| S_1 - S_2 \| = \sqrt{trace(S_1 - S_2)^T (S_1 - S_2)} \tag{5.3}$$

  where $\| Y \| = \sqrt{trace(Y^T Y)}$ is the Euclidean distance, which is also known as the Frobenius norm.

- The logarithm Euclidean distance, which is proposed by Arsigny et al. (2007),

is defined as follows:

$$D_{LE}(S_1, S_2) = \| \ Log(S_1) - Log(S_2) \ \| \tag{5.4}$$

where the Logarithm of a covariance matrix S can be obtained using the spectral decomposition form for a matrix. The spectral decomposition form for any PSD matrix is defined as $U \wedge U^T$, where U is an orthogonal matrix composed of the eigenvectors for the matrix S, $\wedge$ is a diagonal matrix composed of the eigenvalues for the matrix S [5]. In this case the log for the matrix S, log S, is $U(log\wedge)U^T$. Using the spectral decomposition form is also common for computing the exponential for a PSD matrix.

- The Riemannian metric, which is another logarithm-based distance for any PSD matrix, can be written as follows:

$$D_{RE}(S_1, S_2) = \| \ log(S_1^{-1/2} S_2 S_1^{-1/2}) \ \| \ . \tag{5.5}$$

- The RiemannianLe metric, [33] which is Riemannian metric for the multiplication of each covariance matrix by its transpose, can be written as follows:

$$D_{RELE}(S_1, S_2) = \frac{1}{2} \ \| \ log(M^{-1/2} K M^{-1/2}) \ \| \tag{5.6}$$

where $M = S_1 S_1^T$, $K = S_2 S_2^T$.

- The Cholesky distance, which can be computed using the factorization for the covariance matrix using the Cholesky decomposition, where $S = LL^T$, $L = chol(S)$ is a lower triangular matrix with positive diagonal entries, can be written as:

$$D_{CH}(S_1, S_2) = ||chol(S_1) - chol(S_2)||. \tag{5.7}$$

- Another type of decomposition that can be used when the considered matrix is PSD matrix is the square root where $S^{1/2} = U \wedge^{1/2} U^T$, so, the Root Euclidean distance is defined as follows:

$$D_{RE}(S_1, S_2) = \| \ S_1^{1/2} - S_2^{1/2} \ \| \ . \tag{5.8}$$

- For two covariance matrices S_1 and S_2 with dimension $k \times k$, the Non-Euclidean size and shape metric, which is known as Procrustes size-and-shape, between them can be defined as:

$$D_{PR}(S_1, S_2) = \inf_{R \in O(k)} \| \ L_1 - L_2 R \ \| \tag{5.9}$$

where L_i is a decomposition form for a covariance matrix S_i such that $S_i = L_i L_i^T$, $i = 1, 2$. The decomposition form can be either Cholesky where L_i is a lower triangular matrix with positive diagonal entries, that means $L_i = chol(S_i)$ or the matrix square root where $L = S^{1/2} = U \wedge^{1/2} U^T$, S here is represented by the spectral decomposition form where $S = U \wedge U^T$.

- Full Procrustes Shape metric can be written as follows:

$$D_{PRSH}(S_1, S_2) = \inf_{R \in O(k), V \geq 0} \parallel \frac{L_1}{||L_1||} - V L_2 R \parallel \tag{5.10}$$

where $S1 = (L_1 R)(L_1 R)^T$ for any $R \in O(k)$, R an orthogonal matrix [34].

- Power Euclidean metric can be written as follows:

$$D_{PO}(S_1, S_2) = \frac{1}{\alpha} \parallel S_1^{\alpha} - S_2^{\alpha} \parallel . \tag{5.11}$$

$$D_{POS}(S_1, S_2) = \parallel S_1^{1/2} - S_2^{1/2} \parallel . \tag{5.12}$$

5.4 Hypothesis Testing

In our research, hypothesis testing is the required analysis to obtain a decision about(1) comparing mean vectors for variables for the two subjects of interest and (2) comparing mean distances (dissimilarities) for a number of subjects using covariance matrices. For the first case, the null hypothesis claims that two mean vectors for variables for two subjects are equal as mentioned in section 5.2. This hypothesis is tested using the Hotelling T-Squared test. Similarly, for the second case, the null hypothesis claims that the mean distances for a number of subjects (multivariate datasets) are equal. This hypothesis is tested using either one way ANOVA method, which is used when the data are normally distributed, or the Kruskal Wallis test, which is used when the data are not normally distributed. Therefore, the first step needs to be accomplished is to check whether the data are normally distributed. The null and the alternative hypotheses will be:
H0: Distance data are normally distributed.
H1: Distance data are not normally distributed, respectively.
Moreover, the null hypothesis for the one way ANOVA test is:
H0: Mean distances for all groups are equal, this can be written as
$\mu_1 = \mu_2 = \mu_3$,

and the alternative hypothesis will be
H1: At least one group has a different mean.
However, if the studied data are not normally distributed, the common alternative test is the Kruskal Wallis.
The null hypothesis for this test is:
H0: All the median distances are equal, and the alternative hypothesis is:
H1: At least one median distance is different.

5.5 Methodology for Comparing Two Mean Vectors for Two Subjects

5.5.1 Comparing Two Mean Vectors for Two Cities

In this study, daily recorded measurements for the period between 2005 and 2013 for the variables of temperature, precipitation, wind speed, tide, groundwater level, and water discharge are collected. The considered cities here are Cohoes, Utica, and Poughkeepsie, which are located in New York state, US. The rivers are the Mohawk and Hudson, while the former passes through Cohoes and Utica cities, the latter passes through Poughkeepsie city. The raw data for all the cities and all variables have been decomposed to obtain long, seasonal, and short-term components as shown in chapter 2 Sections 2.6, 2.7, and 2.8. As the Hotelling T-Squared test deals with two subjects (two cities) each time, the analysed pairs of cities are: Cohoes and Utica (cu), Cohoes and Poughkeepsie (cp), and Utica and Poughkeepsie (up). The software used in this chapter are SAS (9.4), Minitab(18), Matlab (7.8), R program (3.4.0), and Excel 2013. The findings can be summarised as follows:

1. For the raw data, the analysed cities pairs are shown in Table 5.1. According to the P-values of Hotelling T-Squared test, three decisions, which are Not Reject, Not Reject, and Reject, have been taken for the pairs cu, cp, and up, respectively. Based on these results, the datasets for the two pairs cu and cp have been generated from similar populations (distributions). As we have different cities, we are interested to know if the decisions taken for the null hypothesis are consistent with the geographical distances, in terms of the more closer, the more similar, for each pair of cities. The geographical distance between the two cities for the pairs cu, cp, is short and perhaps this is one of the reasons that lead to the result of Not Rejecting the null hypotheses. Similarly, the decision of Rejecting the null hypothesis for the pair (up) may be attributed to the geographical distance, as this pair has the longest distance compared to the other pairs, and also different rivers passes through these two

cities, Utica and Poughkeepsie.

The same test has been carried out to the datasets for the long, seasonal, and short-term components to gain more insight about what will be resulted if we apply the Hotelling T-Squared test for the data for the three pairs given. This step, in particular, will help us to know if the datasets for the three components for the two pairs cu and cp are originality not dissimilar, which led to not reject the null hypothesis, or perhaps there is a difference for at least one component and has not been captured by testing the raw data. Also, for the pair up, we will be able to know which component(s) is responsible for the differences, which led to the decision of not rejecting the null hypothesis.

The Pairs of Cities	T-Squared	P-Values	GD	Decision
cu	4.253	0.642	88 mile	Not Reject
cp	7.199	0.302	99 mile	Not Reject
up	14.659	0.023	170 mile	Reject

Table 5.1: The T-Squared and P-Values for the Raw Data.

2. For the Long-Term Component, the results are shown in Table 5.2. Three "Not Reject" decisions have been made based on the P-values for the pairs cu, cp, and up, respectively. This would suggest that the long-term component for the three cities has the same pattern for the data for the period 2005 to 2013. This result is consistent with the results of the raw data for the first two pairs, but for the pair up, the result is contradictory.

The Pairs of Cities	T-Squared	P-Values	Geo	Decision
cu	3.788	0.705	88 mile	Not Reject
cp	1.936	0.925	99 mile	Not Reject
up	6.946	0.325	170 mile	Not Reject

Table 5.2: The T-Squared and P-Values for the Long-Term Component.

3. For the Seasonal data, Table 5.3 shows the results for the Hoteling T-Squared test and the P-values. Based on the P-values, the decisions made for the seasonal fluctuations are opposite to findings produced using the raw data for the pair cp. That means there is a significant difference between the seasonal data, but this difference has not been captured when the raw data is investigated. The low contribution percentages for the seasonal components for the considered cities, Cohoes and Poughkeepsie, which are shown in Sections 2.6.7 and 2.8.7,

might be the reason that this difference has not been captured by analysing the raw data. However, for the pairs cu and up, the analysis for the seasonal components has provided similar findings to that produced using the raw data.

The Pairs of Cities	T-Squared	P-Values	Geo	Decision
cu	3.253	0.51	88 mile	Not Reject
cp	31.154	0.01	99 mile	Reject
up	16.592	0.03	170 mile	Reject

Table 5.3: The T-Squared and P-Values for the Seasonal Variations.

4. For the Short-Term Component, one decision, which is the null hypothesis can not be rejected, has been made for the data for the short-term components for the two pairs cu and cp. With regard to the results of this component for the pair up, the null hypothesis has been rejected. Table 5.4 displays the results for this component.

The Pairs of Cities	T Squared	P-Values	Geo	Decision
cu	7.810	0.252	88 mile	Not Reject
cp	8.562	0.199	99 mile	Not Reject
up	18.582	0.03	170 mile	Reject

Table 5.4: The T-Squared and P-Values for the Short-Term Component.

The hypothesis testing results for the data of the raw and the three components are shown in Table 5.5. Examining the results in Table 5.5 reveals that the components

Type of Data	cu	cp	up
Raw	Not Reject	Not Reject	Reject
Long	Not Reject	Not Reject	Not Reject
Seasonal	Not Reject	Reject	Reject
Short	Not Reject	Not Reject	Reject

Table 5.5: Results for the Null Hypotheses.

that lead to not reject the null hypothesis for the pair cu are the three components and for the pair cp are the long and short-term components. Even though the raw data for the pair Cohoes and Poughkeepsie (cp) has similar pattern based on the Hoteling T-Squared test for the raw data, the seasonal variations for this pair has a different behaviour. For the pair (up), we can say that the impact of the seasonal and the short-term components is more than the impact of the long-term component.

Having known that the seasonal and the short-term components are responsible for the rejection of the null hypothesis for the pair up, we examined the variables for the seasonal and short-term components for these two cities. Two statistics, which are the mean and the variance for each variable, which are temperature, wind speed, precipitation, water discharge, tide, and groundwater level, have been computed and examined. The means and variances for the water discharge and groundwater level, which represent the hydrological effect, for these two cities, Utica and Poughkeepsie, are relatively different. This result has been supported by applying a two sample T-test for these two variables. A significant difference has been detected based on the P-values for the two variables WD and GW.

5.5.2 Comparing Two Mean Vectors Without the Hydrological Effect

In this subsection we examine the situation when the effect of the hydrological data is removed. The previous approach will be carried out, the difference is the mean vectors tested will include four variables, temperature, precipitation, wind speed, and tide. That means the data of water discharge and groundwater level is no longer included in the analysis. The results are shown as follows.

1. For the Raw Data, three not rejecting decisions for the null hypothesis have been made based on the P-values. Table 5.6 shows these decisions.

The Pairs of Cities	T-Squared	P-Values	Geo	Decision
cu	5.887	0.207	88 mile	Not Reject
cp	2.266	0.686	99 mile	Not Reject
up	9.232	0.18	170 mile	Not Reject

Table 5.6: The T-Squared and P-Values for the Raw Data Without the Hydrological Effect.

To check the impact of the decomposition of time series, we apply the same test to the data of the long, seasonal, and short-term component.

2. For the Long-Term Data, the values of the Hotelling T-Squared test and P-values in Table 5.7 lead to not reject the null hypotheses. This result is similar to the preceding finding when the influence of the hydrological data has been considered. This would mean that the means vectors for the data of the long-term component for each pair of cities are not affected by the data of water discharge and groundwater level.

The Pairs of Cities	T-Squared	P-Values	Geo	Decision
cu	3.136	0.535	88 mile	Not Reject
cp	3.514	0.475	99 mile	Not Reject
up	3.243	0.518	170 mile	Not Reject

Table 5.7: The T-Squared and P-Values for the Long Term Component Without the Hydrological Effect.

3. For the Seasonal Variations, there is an obvious response to the step of removing the hydrological data. While we have rejected the null hypotheses in the case of the inclusion of the hydrological data, we are not able to reject it when we remove this influence. This could be an evident reason to say that the differences between the mean vectors for the two pairs cp and up can be attributed to the data of the water discharge and groundwater level.

The Pairs of Cities	T-Squared	P-Values	Geo	Decision
cu	2.6601	0.6162	88 mile	Not Reject
cp	2.878	0.5784	99 mile	Not Reject
up	1.9642	0.7424	170 mile	Not Reject

Table 5.8: The T-Squared and P-Values for the Seasonal Variations Without the Hydrological Effect.

4. For the Short-Term component, one decision has been taken, which is the null hypotheses have not been rejected, for the data of this component as shown in Table 5.9.

The Pairs of Cities	T Squared	P-value	Geo	Decision
cu	3.116	0.5386	88 mile	Not Reject
cp	6.773	0.1484	99 mile	Not Reject
up	3.642	0.4566	170 mile	Not Reject

Table 5.9: The T-Squared and P-Values for the Short-Term Component Without the Hydrological Effect.

If we examine the results for this test in Table 5.10, we can see that the findings for the raw and the three components data, for the pairs cu, cp, and up have one decision, which is the null hypothesis is not rejected. Based on this, the reason for the variations between the data of the three cities was because of the hydrological data. When this data has been removed, the null hypotheses for the data for the raw and the three components have not been rejected.

Data	cu	cp	up
Raw	Not Reject	Not Reject	Not Reject
Long	Not Reject	Not Reject	Not Reject
Sesaonal	Not Reject	Not Reject	Not Reject
Short	Not Reject	Not Reject	Not Reject

Table 5.10: Results for the Null Hypotheses Without the Hydrological Effect.

5.6 Methodology for Dissimilarity Analysis

5.6.1 Dissimilarity Analysis for the Raw and the Decomposed Data

For each city, nine covariance matrices have been computed where each matrix has been calculated using the daily data for each year for the period 2005-2013. In our research, we have chosen eight distance measures to compute the statistical distance (SD). These measures are: Euclidean, Procrustes, Riemannian, Procrustes Shape, Cholesky, Power, Log Euclidean, and RiemannianLe. Except the Euclidean distance, all the other distance measures are Non-Euclidean metrics. The SD here is computed between two covariance matrices associated with the considered pair of cities. That means, for each pair and for each distance measure, 9 distance values have been computed as we have daily data for nine years.

The first step in the analysis is to check the normality for the distance values for all the distance measures used. Figure 5.1 shows the plot of normality for the Riemannian distance values. Based on the significance level, which is 0.05, the null hypothesis is not rejected and the data are normally distributed, where the P-value is 0.319. Having checked that the distribution for these distances is the normal distribution, the one way ANOVA test has been applied for all distance measures. Also, for each distance measure for the raw data, as each pair contains 9 distance values, the total distance values to be tested using the one way ANOVA test are 27 values.

According to the P-values, the decision of rejecting the null hypothesis, which is $\mu_1 = \mu_2 = \mu_3$, has been taken, where all the P-values are less than 0.05 as shown in Table 5.11. This means that there is at least one pair has a mean distance that is different from the other. In other words, there are significant differences between the considered pairs of cities based on the available time series datasets. The reason for this result is perhaps attributable to the data of the pair up, where as it has been previously mentioned in section 5.5.1 that the data for the pair up has not been generated from similar populations. We have applied Tukey test, which depends on the differences of means to determine whether the mean difference between each two

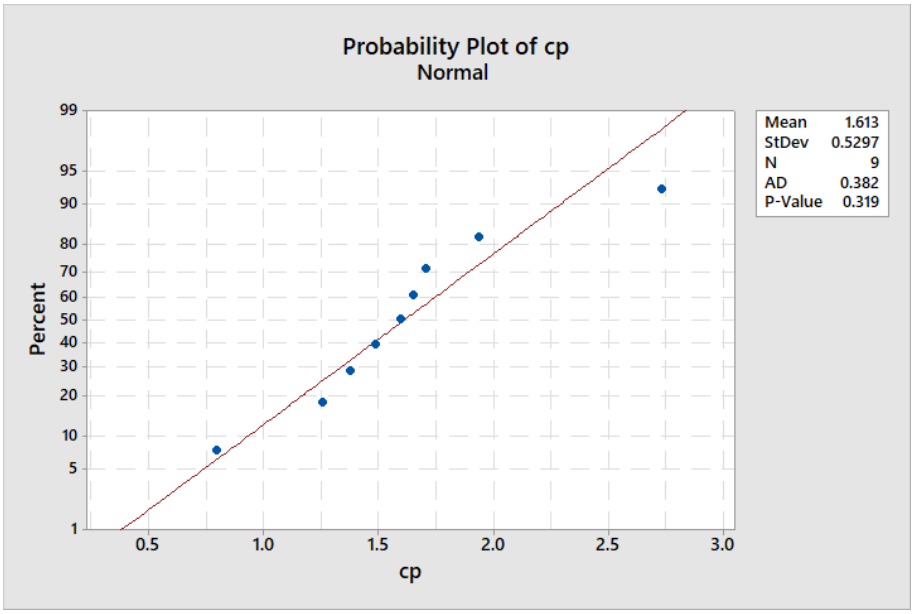

Figure 5.1: Probability Plot for the Riemannian Distance Measure for the Raw Data for the Pair cp.

pairs is significant, the two pairs cu and cp have been classified in one group, and the pair up in a different group.
We have separately checked the statistics of mean and variance for each variable for the raw data for the pairs cu and cp. There are no obvious differences between these statistics for all the variables for Cohoes and Utica cities. This result has been supported using the two samples T-test for all the variables for the same two cities. There are no significant differences between the data of these two cities.

In Figure 5.2, we can visualize these differences by examining the Box plots for the three pairs of cities for the Log Euclidean distance. The highest median distance in this figure is only 1.55 for the pair cp. Also, for this pair of cities, the interquartile range, which is the width of the box plot for this group, is 0.54. The pair cu has the lowest median, which is approximately 0.82, and also the lowest interquartile range, which is about 0.26. The median for the pair up is 1.48, and the highest interquartile range in this plot has been noticed for this pair, up.

To find the component(s) that has led to this result, the same steps above for dissimilarity analysis have been repeated but by using the decomposed data.

Distance Measure	P-Values
Procrustes	0.001
Riemannian	0.001
ProcrustesShape	0.005
Cholesky	0.001
Power	0.001
Euclidean	0.009
LogEuclidean	0.004
RiemannianLe	0.001

Table 5.11: The P-Values for the One Way ANOVA test for the Distance Measures for the Raw Data.

Dissimilarity Analysis for the Long-Term Component

- For nine years for the period between 2005 and 2013 and for each distance measure, 27 covariance matrices have been computed for the long-term component series for the three pairs of cities. For example, the covariance matrix for the long-term component data for Utica city for year 2005 can be written as follows:

$$\begin{bmatrix} 0.99 & 0.07 & 0.32 & -0.65 & -0.11 & 0.59 \\ 0.07 & 1.00 & 0.30 & 0.33 & 0.23 & -0.07 \\ 0.32 & 0.30 & 1.00 & 0.21 & -0.36 & 0.41 \\ -0.65 & 0.33 & 0.21 & 0.99 & 0.44 & -0.64 \\ -0.11 & 0.23 & -0.36 & 0.44 & 1.00 & -0.67 \\ 0.59 & -0.07 & 0.41 & -0.64 & -0.67 & 0.99 \end{bmatrix}$$

  The dimension of this array is 6×6 for the variables temperature, wind speed, precipitation, water discharge, tide, and groundwater level, respectively.

- In order to examine the amount of dissimilarity between each pair of cities, the eight distance measures have been applied. Calculating the SD provides a set of distance values for the considered nine years for each pair. For example, the Log-Euclidean distance values for the long-term component for each pair of cities are shown in Table 5.12.

- In the context of hypothesis testing, since there is one categorical variable, which is the pairs of cities, and the data are normally distributed, the one way ANOVA test has been used.

Thus, based on the P-values, which are less than 0.05 for all the distance measures used, the results for the long-term component are statistically significant by testing

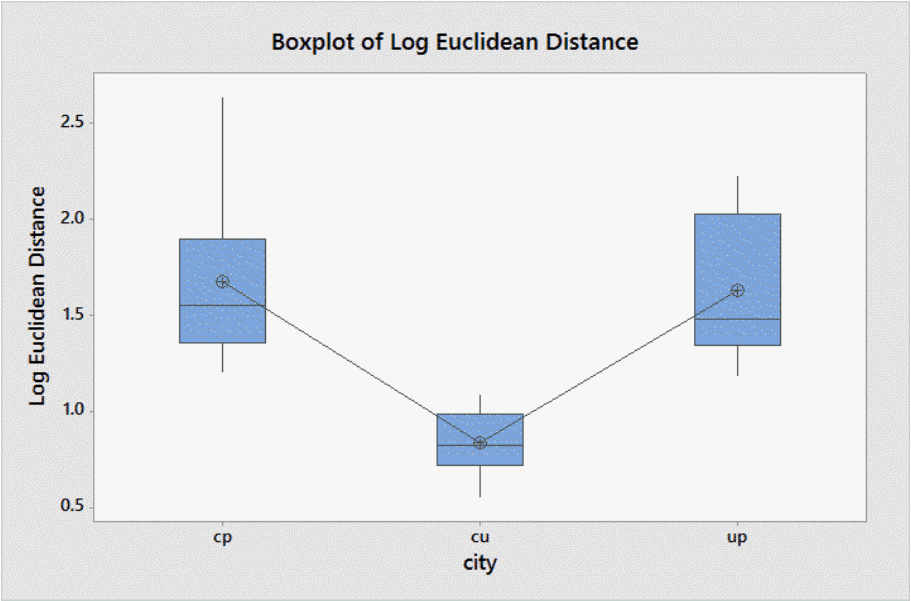

Figure 5.2: The Box Plots for the Log-Euclidean Distance Measure for the Raw Data.

the mean distances for all pairs given. For the long-term component for Cohoes and Utica cities, we have examined the mean and variance for each variable and applied two samples T-test for each variable. No statistically significant difference has been noticed between the data of Cohoes and Utica.

Dissimilarity Analysis for the Seasonal Fluctuations

Testing the normality assumption for the distance measures values for the seasonal fluctuations has led to reject the null hypothesis that claims the distance measure values for the seasonal data have a normal distribution for the three pairs. For example, Figure 5.3 shows the probability plot for the values of the power Euclidean distance measure for the seasonal variations for the pair cu. Based on the P-values, the decision of not rejecting the null hypothesis has been taken. In this case, we are not able to use the one way ANOVA test. As an alternative, the Kruskal Wallis test has been applied. Based on the P-values for this test the decision of not rejecting the null hypothesis has been made. For the seasonal component for Cohoes and Utica cities, we have applied two samples T-test. Again, no statistically significant difference has been noticed between the the data of these two cities.

	cu	cp	up
2005	0.81	2.37	2.65
2006	2.35	1.15	2.61
2007	1.84	2.09	2.43
2008	0.59	1.51	1.45
2009	1.12	1.12	1.65
2010	1.99	1.37	2.74
2011	0.99	2.11	2.19
2012	1.51	1.02	2.02
2013	1.33	1.79	2.01

Table 5.12: The Log-Euclidean Distance Measure for the Long-Term Component Data.

Dissimilarity Analysis for the Short-Term Component

Since the values for all the distance measures kinds for the short-term component data are normally distributed, we use the one way ANOVA test. The null hypothesis has been also rejected for all the distance measure kinds. No statistically significant difference has been obtained using the two samples T-test for the short-term component data for Cohoes and Utica cities.

Comparison Between the Geographical and Statistical Distances

Often it is expected that whenever two cities are close to each other, the behaviour for some variables such as climatic and hydrological data, has a similar pattern. To investigate if this expectation is applicable using our data, we have carried out a comparison. Briefly this comparison is performed to check if the order (low to high) of the mean SD is identical to the order (low to high) for the Geographical Distance (GD) for each pair of cities. This examination has been carried out for the raw, long, seasonal, and short-term component data and the results are shown in Table 5.13.

Type	Geo	Raw	long	seasonal	short
cp	88.5 mile	1.19	1.167	0.76	1.04
cu	99.5 mile	0.68	0.953	0.46	0.68
up	170 mile	1.28	1.487	0.86	1.17

Table 5.13: The Geographical and Statistical Distances (Euclidean).

In Dissimilarity analysis, whenever we have a small distance value obtained by one of distance measures, the degree of similarity will be high. The results can be

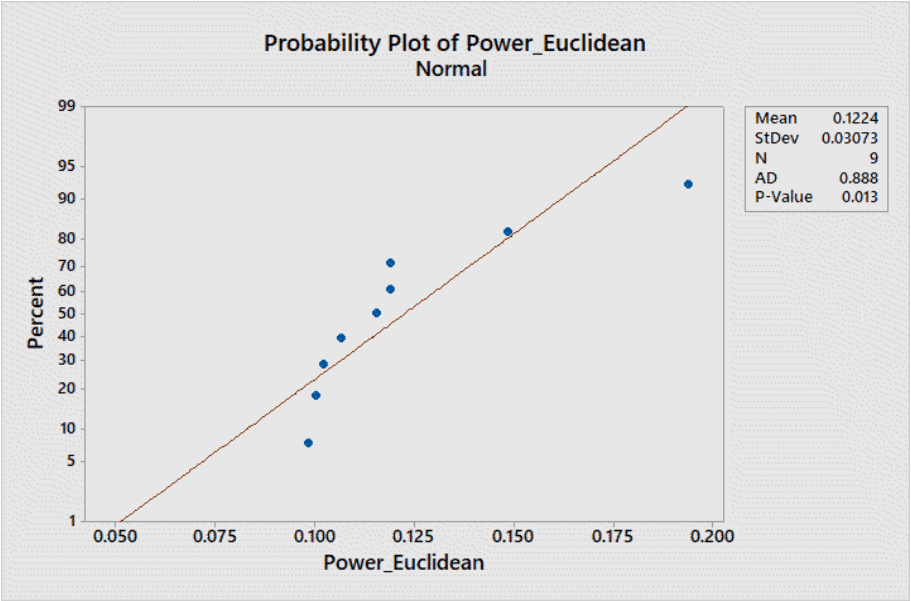

Figure 5.3: Probability Plot for the Power Euclidean Distance Measure for the Seasonal Fluctuations for cu.

summarised as follows:

- In Table 5.13, same order has been noticed for the geographical and statistical distances for the third pair up. The geographical distance between these two cities is the highest, and the mean SD are also the highest mean among all the other pairs. For the first two pairs, a reverse situation has been detected. Although the GD between Cohoes and Utica cu is higher than the distance between Cohoes and Poughkeepsie cp, the patterns of their climatic and hydrological variables are more similar.

- In Table 5.13, the same results for the raw data have been noticed for the long and seasonal components.

- For the short-term component, as shown in Table 5.13, the pair cu also has displayed the minimum mean SD, then, the pairs cp and up.

To sum up, for the raw and the three components, the pair cu has the highest degree of similarity even though the distance between them is not the lowest. This, in turn, indicates that the data for these two cities are more similar than the other two pairs. The same river, which is the Mohawk River, passes through the two cities for this pair and this perhaps leads to this result.

5.6.2 Dissimilarity Analysis Without the Hydrological Effect

Having rejected the null hypothesis when all the studied variables are included and tested using the one way ANOVA method, the mean and variances for the variables for the three components have been examined. The mean and variance for the two variables of water discharge and groundwater level for the two components for the seasonal and the short-term component for the two cities, Utica and Poughkeepsie, are relatively different. Based on this, we removed these two variables from the datasets, then we followed the same steps above. Hence, the new datasets are climatic datasets contain the data of the series temperature, precipitation, wind speed. Additionally, tide data has also been included.

Consequently, the covariance matrix for this group of variables will be a matrix of dimension 4×4. The covariance matrices have been computed for the raw and the three components. In order to obtain a decision about the dissimilarity between all the pairs after eliminating the effect of water discharge and groundwater level, the hypothesis testing should be performed. The normality assumption for all the distance measures for the raw data is no longer achieved here. In this case the Kruskal Wallis test has been chosen to test the null hypothesis instead of one way ANOVA test. The Kruskal Wallis test depends upon the median of the data. The calculated P-values for all the distance measures are shown in Table 5.14. For the raw data, as all the P-probabilities for all the considered distance measures have values that are greater than 0.05, the null hypothesis has not been rejected. This result is different compared to the case of including all the variables. For the long-term component, which its data are normally distributed, when we applied the one way ANOVA test, we obtained the P-values shown in Table 5.14, where these values are greater than 0.05, except ProcrustesShape measure. This would suggest that there are no significant differences over the studied pairs for the long-term component.

For the seasonal variations, the Kruskal Wallis has been used and the results for the P-values are shown in Table 5.14. The P-values for all the distance measures are greater than 0.05; this has led to not reject the null hypothesis. This, in turn, means that there is no significant difference between the pairs for the seasonal fluctuations.

For the short-term component, the values of all the distance measures are not normally distributed. This has led to use the Kruskal Wallis test. The results obtained using this criterion are shown in Table 5.14. It is obvious that these results are not similar to the case when the hydrological variables have been involved where here the null hypotheses have not been rejected. This means, apart from the hydrological data, that the short-term component for the cities has a similar pattern.

Type	Raw	Long	Seasonal	Short
Procrustes	0.84	0.06	0.40	0.55
Riemannian	0.46	0.06	0.64	0.49
ProcrustesShape	0.61	0.03	0.20	0.68
Cholesky	0.79	0.12	0.53	0.64
Power	0.82	0.07	0.32	0.61
Euclidean	0.83	0.07	0.11	0.62
LogEuclidean	0.52	0.06	0.44	0.49
RiemannianLe	0.46	0.06	0.23	0.51

Table 5.14: The P-Values for all the Distance Measures for the Raw and Three Components Data Without the Hydrological Effect.

Comparison Between the Geographical and Statistical Distances

- For the raw data, Table 5.15 shows that the order of the GD, (low to high), has not completely matched the order of the SD. The SD for the raw and the three components has been calculated by taking the mean for the distances that have been computed by using the Power Euclidean distance measure for the nine years considered. Even though the GD for the pair cp is smaller than the GD for the pair cu, the SD for the cp is higher, 0.57. This indicates that the magnitude of similarity between the data of Cohoes and Utica is higher than the similarity between the data of the two cities of Cohoes and Poughkeepsie. Additionally, the orders of the GD and SD for the third pair up are similar where both of them are in the third order except the distance for the short-term.

- For the long-term component, both GD and SD show the same order. This means that the closer the two cities are to each other, the higher the amount of similarity between their data is.

- For the seasonal variations, the order is similar to the order of the long-term component as shown in Table 5.15.

- For the short-term component, the orders (low to high) of the GD and SD distances for the pairs have not matched. The highest SD has appeared for the pair cp, which has the smallest GD. The pair up has the second highest SD while the pair of cu has the smallest value, where this is shown in Table 5.15.

With regard to the dissimilarity analysis results based on the covariance matrices for the independent variables of the MLR models for the raw and the three components data for the three cities, the P-values for the one way ANOVA test are greater

Type	Geo	Raw	Long	Seasonal	Short
cp	88.5 mile	0.57	0.99	0.1	0.49
cu	99.5 mile	0.46	1.15	0.12	0.42
up	170 mile	1.55	1.487	0.18	0.46

Table 5.15: The Geographical and Statistical Distances (Power Euclidean) for the Raw and Components Data Without the Hydrological Effect.

than the significance level, 0.05. These results lead to not reject the null hypothesis that claims there is no difference between the mean distances for the three pairs. The mean distance is calculated using the same Euclidean and Non-Euclidean distance measures.

Relying on the results of different hypothesis testing for the pairs, specifically, cu, it would be possible to use the MLR models for the raw and the three components interchangeably for forecasting future values for the Cohoes and Utica cities. That means, if there is no possibility to obtain a forecasting model for one of these cities, using the forecasting model for the other will provide acceptable results. To support this result, the greater the proximity value for one or more of parameters-based statistics, such as MSE for MLR models, the more similar the objects are.

5.7 Comparison Between the Distance Measures

In order to compare the behaviour of the eight distance measures, the values of these measures for the year 2005 for the raw and the three components long, seasonal, and short-term component for all pairs cu, cp, and up have been transformed (mapped) into the range [0,1] using the Max-Min normalization technique. A number of slightly distinct patterns have been observed when we sorted the mapped data from the smallest to the largest which could be summarised as follows:

- For the raw data, a consistent pattern can be observed for the three pairs cu, cp, and up where the distance measures can be ordered from smallest to largest as follows: ProcrustesShape, Procrustes, Cholesky, Euclidean, Power, LogEuclidean, Riemannian, RiemannianLe. Figure 5.4 shows this consistency over the three pairs where the mapped values for the eight distance measures for year 2005 have been plotted.

- For the long-term component data, no common behaviour can be detected for the three pairs. However, based on the sorted data, the first three measures ProcrustesShape, Procrustes, and Cholesky and also the last two measures Rie-

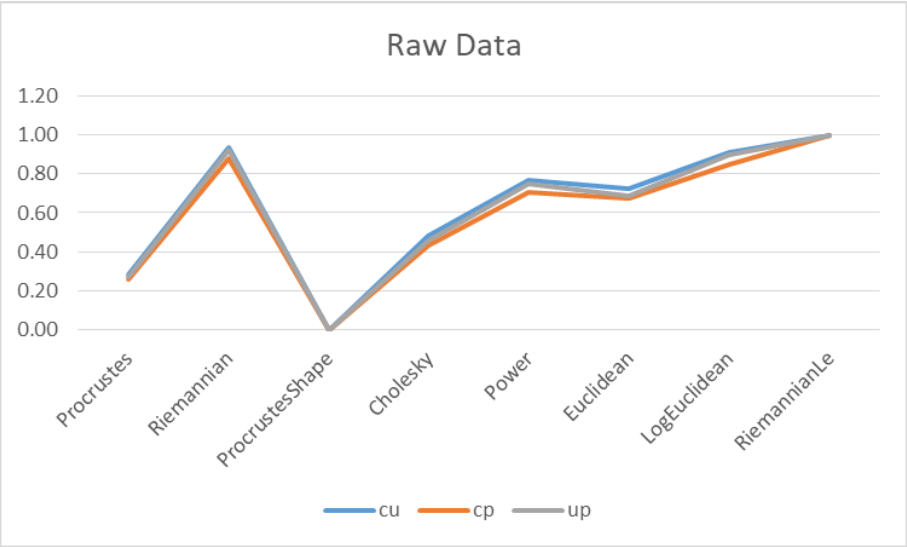

Figure 5.4: The Mapped Raw Values for the Eight Distance Measures for Year 2005.

mannian and RiemannianLe have similar pattern of variation over the three pairs as shown in Figure 5.5.

- For the Seasonal Variations, the same results for the raw data have been obtained for the seasonal fluctuations where the measures have been sorted from the smallest to the largest as follows: ProcrustesShape, Procrustes, Cholesky, Euclidean, Power, LogEuclidean, Riemannian, and RiemannianLe.

- For the short-term component, compared to the raw and seasonal results, slightly different findings have occurred for this component, where the only difference is the order of the two distance measures Power and Euclidean.

Furthermore, for the data of the raw and each component, all the eight distance measures are clustered according to their performance with respect to the three pairs cu, cp, and up. The results of the clustering process are summarised using a Dendrogram which is a plot that arranges the clusters produced by a hierarchical clustering as a tree consists of many u-shaped lines that connect data points (distance measures) to each other.

The dendrograms are created using two clusters. Figures 5.6 and 5.7 show the dendrograms for the raw and short-term component data. Also, dendrograms for the long and seasonal components are shown in Figures A.6 and A.7 in Appendix. A small difference can be noticed for the dendrograms plotted. Figure 5.6 shows the Dendrogram for the raw data where the measures Procrustes, Cholesky, and ProcrustesShape construct one cluster while the remaining measures construct the second

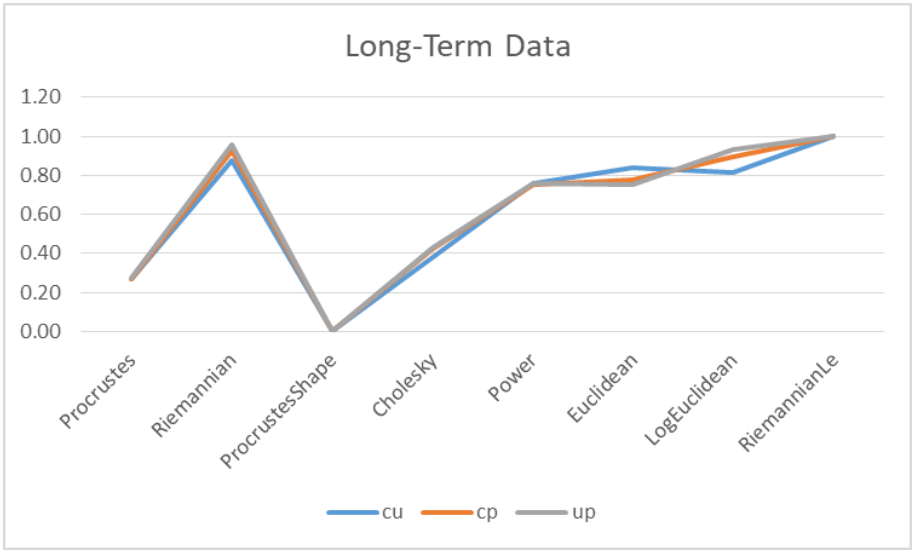

Figure 5.5: The Mapped Long-Term Values for the Eight Distance Measures for Year 2005.

cluster. This graph reveals that the similarity level between the measures included in each cluster separately (within groups) is greater than the similarity level between the two clusters (between groups). Moreover, in the short-term component's dendrogram, the first cluster (left) is composed of three distance measures Procrustes, Cholesky, and ProcrustesShape which means that these measures are close to each other. Based on the connection points, however, the measures Procrustes and Cholesky are more similar to each other than they are to the ProcrustesShape measure. The second cluster includes Riemannian, LogEuclidean, RiemannianLe, Power, and Euclidean distance measures. The measures Riemannian and LogEuclidean are more similar to each other than they are to the RiemannianLe measure. Also, Power and Euclidean are more similar to each other than they are to the other measures.

5.8 Frequency Domain Positive Semi-Definite Matrices-Based Metrics

In this part of this chapter we present three new developed dissimilarity metrics all of them constructed using features from the frequency domain. These metrics have been applied to the raw data in an attempt to examine their behaviours using multivariate time series datasets. Hypothesis testing has been also used to obtain a decision about whether mean distances for the frequency content for the groups (pairs) are equal.

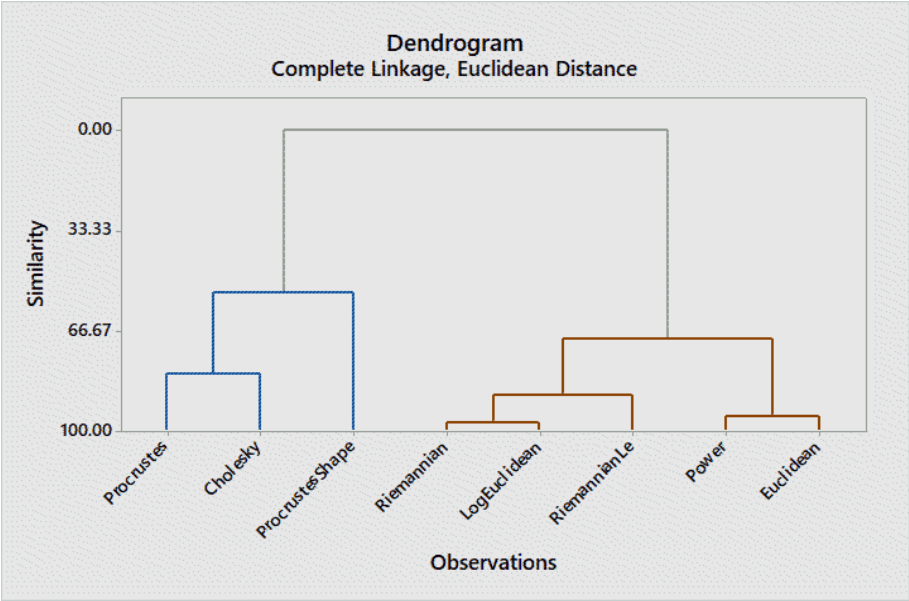

Figure 5.6: The Dendrogram for the Raw Data for the Eight Distance Measures for Year 2005.

The null hypothesis here will be the mean distance for the three pairs cu, cp, and up, are equal. The mean distance here is computed based on the frequency as we use the Power Spectral Density (PSDE) Matrix and the Periodogram for the variables. The PSDE and periodograms for the variables are calculated using the frequency. The new developed metrics and their applications are summarised in the following subsections.

5.8.1 Dissimilarity Analysis Using the Power Spectral Density Matrix (PSDE)

Each signal possesses some features in the time and frequency domains. The most common time domain signal's attributes are the autocorrelation, partial autocorrelation, and the cross-correlation. On the other hand, the functions that are often used to describe the signals in the frequency domain are the power spectral density (auto spectrum) and the cross power spectral density (cross spectrum) functions. Using these two functions, the Power Spectral Density Matrix (PSDE) matrix will be created. The structure of the PSDE matrix is similar to the covariance matrix structure, where the diagonal elements are the auto spectrum (spectrum of the signal with itself) and the off-diagonal entries are the cross-spectrum between two signals. This matrix is built using one frequency ω, where $\omega \in [\omega_{min}, \omega_{max}]$.

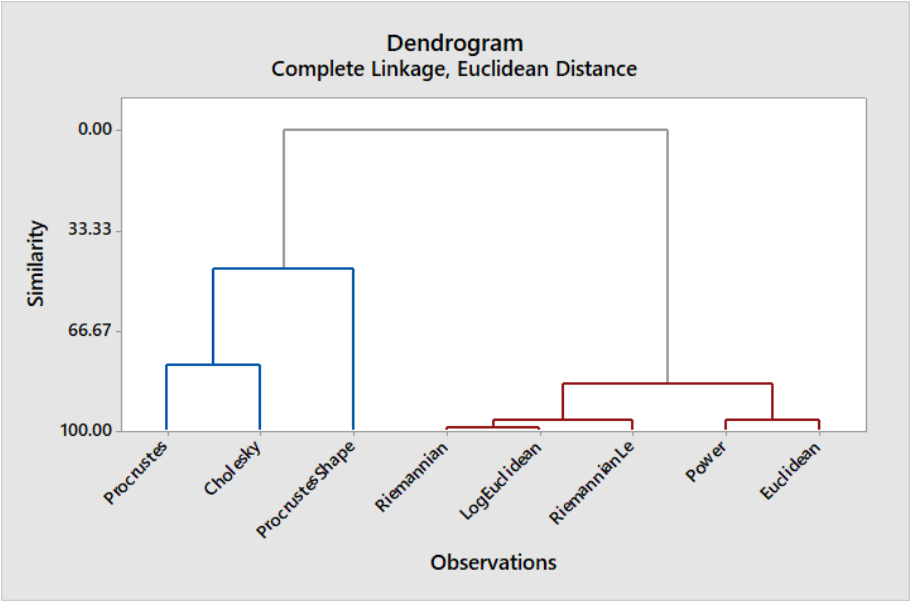

Figure 5.7: The Dendrogram for the Short-Term Data for the Eight Distance Measures for Year 2005.

Most of time series (signals) can be expressed using the sine and cosine functions for the frequencies that construct it [58] as shown in Equation 5.13.

$$y_t = \frac{a_0}{2} + \sum_{k=1}^{p-1} f_k(a_k cos\omega_k t + b_k sin\omega_k t) \tag{5.13}$$

and

$$f_k = \begin{cases} 1/2 & \text{if} N \text{is even and} \quad k = p-1 \\ 1 & \text{if } N \text{ otherwise} \end{cases}$$

where

- t is the time subscript, where $t = 1, 2, \ldots, N$
- y_t is the time series.
- N is the number of observations.
- p is the number of frequencies for Fourier decomposition, where $p = \frac{N+2}{2}$ if N is odd and $p = \frac{N+1}{2}$ if N is even.
- k is the frequency index, $k = 1, 2, \ldots, p-1$.

- a_0 is the mean where $a_0 = 2\bar{x}$.
- a_k are the cosine coefficients.
- b_k are the sine coefficients.
- ω_k are the Fourier frequencies, where $\omega_k = \frac{2\pi k}{N}$.

The periodogram is defined as:

$$P_k = N/2\bigg((a_k^2 + b_k^2)\bigg).$$

In addition, the cross periodogram between two time series x and y can be written as follows:

$$P_k^x y = \frac{N}{2}\bigg(a_k^x a_k^y + b_k^x b_k^y\bigg) + i\frac{N}{2}\bigg(a_k^x b_k^y - b_k^x a_k^y\bigg)$$

As a function for frequency, not time, the Power Spectral Density function (PSDE), also known as the Auto power spectral density, presents the frequency content of a signal. In other words, this function determines how the variations (energy) of the signal oscillates based on the frequency. The PSDE for x is defined as follows:

$$J_{11}(\omega) = \sum_{j=-p}^{p} W_j P_{k+j}^x \tag{5.14}$$

where W is a vector of $(2p+1)$ smoothing weights, and P is the periodogram. The Cross Spectral Density (CSD) is the Fourier Transform, FT, for the cross correlation function between two series, r_{xy}. The CSD can be written as follows:

$$P_{12}(\omega) = \sum_{n=-\infty}^{\infty} r_{xy}(n)\exp(-i\omega n) \tag{5.15}$$

where i represents the imaginary unit. If $x = y$, the CSD transforms to the PSDE. The matrix below is the PSDE matrix for Poughkeepsie city for the data of the variables Temperature, Wind Speed, and Precipitation, for year 2013. In our analysis, we have included all the considered variables, which are, in addition to the variables above, Water Discharge, Tide, and Groundwater level, respectively. The diagonal elements are real values and represent the auto spectral density of the variables, and the off-diagonal entries are complex numbers and represent the cross spectral for the variables

$$\begin{bmatrix} 11.88 + 0.00i & -0.66 - 1.00i & 3.37 - 0.19i \\ -0.66 + 1.00i & 1.64 + 0.00i & 0.00 + 0.21i \\ 3.37 + 0.19i & 0.00 - 0.21i & 1.47 + 0.00i \end{bmatrix}.$$

Seven Euclidean and Non-Euclidean distance measures have been used to calculate the distance for each considered pair using the PSDE. These measurements are: Euclidean, Procrustes, Riemannian, Procrustes Shape (Full-Procrustes), Cholesky, Power Euclidean, Log Euclidean. Table 5.16 displays the Riemannian Distance (RD) for the PSDE matrices for the raw data for the three pairs.

Year	RD	Pair	RD	Pair	RD	Pair
2005	3.51	cu	6.32	cp	7.73	up
2006	5.04	cu	3.73	cp	6.26	up
2007	6.24	cu	8.49	cp	6.84	up
2008	2.26	cu	5.75	cp	4.97	up
2009	4.41	cu	4.14	cp	4.95	up
2010	4.54	cu	4.27	cp	5.81	up
2011	4.28	cu	5.90	cp	6.27	up
2012	4.89	cu	3.26	cp	5.87	up
2013	3.95	cu	5.34	cp	5.85	up

Table 5.16: The Riemannian Distance (RD) Measure for the PSDE Matrices.

Additionally, as shown in Table 5.17, applying the one way ANOVA test produces two different groups of decisions. In the first group, which includes the measures of Procrustes, fullprocrustes, and Cholesky, the null hypothesis has not been rejected based on the P-values. On the other hand, in the second group, which consists of the measures of Riemannian, Euclidean, LogEuclidean, and Root Euclidean (power), the null hypothesis has been rejected. Just to mention, the distance measures in the first group are mathematically calculated using the Cholesky decomposition. Hence, according to these results, some of the distance measures, first group, show that there is significant difference between the frequency-based mean distance for the three pairs. However, the results for the second group indicate that there is no significant difference between the frequency content for the three pairs.

5.8.2 Using the Eigenvalues for the PSDE matrix as a Dissimilarity Measure

Once we construct a PSDE matrix, the eigenvalues for this matrix can be computed. The eigenvalues can be used to construct a dissimilarity measure. Because the results are vectors of eigenvalues which belong to a Euclidean space, we have used the Euclidean distance to compute the distances between the vectors. Mathematically, this

Distance-Type	P-Value
Procrustes	0.718
Riemannian	0.02
Full-procrustes	0.453
Cholesky	0.17
PowerEuclidean	0.003
Euclidean	0.003
Log-Euclidean	0.018

Table 5.17: The P-Values for the Distance Measures for the PSDE Matrices.

can be written as follows:

$$d_{EI}(TS_1, TS_2) = \sqrt{\sum_{j=1}^{k}(EIj_{TS1} - EIj_{TS2})^2}$$

where TS_1 and TS_2 are the two considered subjects (multivariate time series datasets). Also, EI_{TS1} and EI_{TS2} are the eigenvalue vectors for the PSDE matrices for these two multivariate time series datasets, and j is the index for the eigenvalues. We have checked that this Eigenvalue-based distance satisfies the properties of being a metric. These properties are the symmetry, where $d_{EI}(TS_1, TS_2) = d_{EI}(TS_2, TS_1)$, non-negativity, where $d_{EI}(TS_1, TS_2) > 0$, and the triangle inequality property, where $d_{EI}(TS_1 + TS_2) \leq d_{EI}TS1 + d_{EI}TS_2$. Table 5.18 shows the eigenvalues for the PSDE matrices for Cohoes City for the studied years. Table 5.19 displays the dissimilarity

Year	e1	e2	e3	e4	e5	e6
2005	29.92	11.34	2.53	0.42	0.13	0.08
2006	16.23	8.33	2.84	0.42	0.18	0.08
2007	37.80	17.76	1.24	0.32	0.11	0.00
2008	27.03	11.88	1.57	0.48	0.15	0.05
2009	18.14	10.00	1.46	0.49	0.17	0.03
2010	29.09	13.17	1.76	0.35	0.09	0.03
2011	18.33	13.11	2.90	0.61	0.07	0.02
2012	28.22	10.79	2.54	0.49	0.15	0.02
2013	20.17	12.17	1.94	0.31	0.09	0.02

Table 5.18: The Eigenvalues for the PSDE matrices for Cohoes City.

analysis using the Euclidean distance for the raw data for the three pairs.

Because there are three pairs and the data are not normally distributed, the Kruskal-Wallis test has been used. Based on the P-value, 0.01, the null hypothesis

Year/Pair	cu	cp	up
2005	1.08	19.00	19.17
2006	2.89	3.05	1.23
2007	1.71	9.95	11.38
2008	1.17	9.34	10.34
2009	0.95	2.71	3.47
2010	3.53	8.43	11.77
2011	1.34	3.97	2.88
2012	3.81	5.76	9.48
2013	1.56	1.85	2.23

Table 5.19: The Euclidean Distance for the Eigenvalues Vectors for the PSDE Matrices.

has been rejected which, in turn, means that there is at least one median is different from the medians of the other pairs.

5.8.3 X^TX Matrix for the Periodograms as a Dissimilarity Measure

Typically, in statistics, specifically in regression analysis, a matrix X that its columns represent the variables and its rows represent the studied observations is called a design matrix. When we multiply the transpose of this matrix, which is (X^T) by the matrix itself, the result is a new symmetric square matrix with dimensions determined by the number of variables. Additionally, this new constructed matrix is a positive definite matrix and it is essentially used to estimate the parameters of the regression model using the least squares method. Here we build a similar matrix but for the periodograms of the variables rather than the raw data to be used in the dissimilarity analysis by applying Euclidean and Non-Euclidean metrics.

The periodograms for the variables have been calculated using the Fast Fourier Transform (FFT). The matrix below is the X^TX matrix for the periodograms for the Cohoes city variables for the year 2006.

$$\begin{bmatrix} 5813.04 & 597.73 & 617.19 & 1333.44 & 4429.10 & 369.62 \\ 597.73 & 378.35 & 282.79 & 475.01 & 873.53 & 115.23 \\ 617.19 & 282.79 & 1190.23 & 894.55 & 891.94 & 166.18 \\ 1333.44 & 475.01 & 894.55 & 4077.67 & 6133.98 & 909.32 \\ 4429.10 & 873.53 & 891.94 & 6133.98 & 12228.12 & 1346.42 \\ 369.62 & 115.23 & 166.18 & 909.32 & 1346.42 & 214.08 \end{bmatrix}.$$

Table 5.20 shows the values for the Riemannian distance measure for the X^T

matrices for the three pairs of cities. As long as the data are normally distributed,

Year	RE	Pair	RE	Pair	RE	Pair
2005	2.08	cu	5.22	cp	4.96	up
2006	5.23	cu	5.45	cp	4.14	up
2007	3.03	cu	5.08	cp	4.93	up
2008	4.20	cu	5.18	cp	5.67	up
2009	3.19	cu	4.64	cp	5.57	up
2010	4.95	cu	3.02	cp	5.13	up
2011	1.53	cu	3.10	cp	3.37	up
2012	2.73	cu	6.86	cp	6.71	up
2013	6.35	cu	5.95	cp	5.18	up

Table 5.20: The Riemannian Distance for the X^TX Matrices for the Periodograms for Each Pair of Cities.

the one way ANOVA test is carried out and the results for P-value are shown in Table 5.21. Again, the results extracted using the X^TX matrices provide two groups.

Distance-Measure	P-Values
Procrustes	0.06
Riemannian	0.04
ProcrustesShape (Full Procrustes)	0.013
Cholesky	0.029
PowerEuclidean	0.039
Euclidean	0.04
LogEuclidean	0.021
RiemannianLe	0.103

Table 5.21: The P-values for the ANOVA Test.

Based on the P-values for the first group, which includes Procrustes and RiemannianLe measures, the null hypothesis has not been rejected. That means, there is no significant difference between the frequency content for three pairs. However, the results for the second group, which includes Riemannian, ProcrustesShape (Full Procrustes), Cholesky, Power Euclidean, Euclidean, and LogEuclidean, lead to reject the null hypothesis that claims the periodogram-based mean distances are equal for the three pairs.

5.9 Discussion

In the first part of this chapter we dealt with the two sample Hotelling T-Squared test to obtain a decision about whether two multivariate time series datasets have been generated from similar populations. For the raw, long, seasonal, and short-term component data, two cases related to whether the hydrological effect, which is represented by the data of the two series, water discharge and groundwater level, have been considered. For the raw data, the null hypothesis, which claims that the mean vectors for all the pairs are equal, has been rejected for one pair, up, and has not been rejected for the other two pairs, cu and cp.

When we extended the analysis to include the data of the three components as an attempt to detect which component has led to this result of not rejecting the null hypothesis for the pair up, we observed that the null hypotheses for the data of the seasonal and short-term components have not been rejected. Based on this result, these two components, seasonal and short, have led to this rejection. For the pair cu, the null hypotheses for the raw and the three components have not been rejected. For the pair cp, by examining the three components, the mean vectors for the seasonal fluctuations were not equal. But these results have not been detected when we analysed the raw data, perhaps because of its low contribution percentage compared to the long and short-term components as discussed in the contribution percentages in Sections 2.6.7 and 2.8.7.

The differences between the results may be attributed to the GD or having same river passing through each pair; where we have noticed that the GD for the pairs cu and cp is shorter than GD for the pair up. However, when we have eliminated the impact of the hydrological data, the null hypothesis for the raw data for the three pairs has not been rejected. This would suggest that the raw data for the three cities has same patterns based on the climatic variables and the tide variable. That means there are no significant differences between the multivariate time series datasets for the period given for the three pairs of cities if we only consider the climatic and tide variables. This result has been extracted by applying two samples Hotelling T-Squared test for comparing the means of temperature, precipitation, wind speed, and tide.

On the other hand, the results for the second part of this chapter, which is the hypothesis's testing for the dissimilarity analysis part, provide one decision, which is the null hypothesis, has been rejected based on the P-values as shown in Table 5.11. This analysis has been conducted using one way ANOVA test for the means for eight distance measures for the three pairs. The ANOVA one way method tests the three pairs together. The null hypothesis for the raw data has been rejected. This result could be attributed to the inclusion of the pair up within the range of the tested pairs where it has been verified previously by the analysis of comparing means, Hotelling

T-Squared test, that the pair up has significant differences. Similar results have been noticed for the null hypotheses for the three components.

The results of the seasonal component could be supported by the findings of the previous analysis of comparing two mean vectors for this component as the null hypotheses have been rejected for two pairs. Similarly, the rejection of the null hypothesis for the short-term component using the one way ANOVA test was perhaps expected as the pair up is one of the examined pairs, where the null hypothesis for this component has been rejected in the Hotelling T-Squared test. However, the results for the long-term component for the two analyses, Hotelling T-Squared and ANOVA, were different.

The last part in this discussion for the dissimilarity analysis is related to the case when the impact of the hydrological data is removed. As has been previously mentioned, applying the Hotelling T-Squared test, has led to not reject the null hypotheses for the raw and the three components data. Provided that the results of similarity between the data of the two cities of interest using the hypothesis testing are in favour of no significant difference between them, the possibility of using a forecasting model for one city to forecast future values for another city is achievable. Besides, existence of one or more of parameter-based statistics, such as MSE values, that are approximately similar, will also support the similarity decision.

The last part in this chapter is devoted to the new three developed dissimilarity measures. For the outputs of hypothesis testing using the Euclidean distance for the three dissimilarity measures, which are power spectral density matrix, Eigenvalues of the PSDE, and X^TX matrix of the periodograms, the decision of rejecting the null hypothesis has been taken. However, if we compare the results of hypothesis testing using the PSDE and X^TX, we will see that in addition to the discrepancy within each case, where the null hypothesis has been rejected for some distance measures and has not been rejected for the others, there are some differences between these two dissimilarity measures. The results for the distance measures Procrustes, Riemannian, Power Euclidean, and Log-Euclidean are almost identical. For the other distance measures, different dissimilarity decisions have been made. Within hypothesis testing context, there is no possibility to assess the performance of these dissimilarity measures. However, the accomplishment of assessing task will be possible if we apply these distance measures to one of data mining techniques, such as clustering and classification.

With regard to the comparison between the geographical and statistical distances, firstly, for the case of involving the whole studied variables, the results for the third pair, which is up, show that there is no difference between the geographical and statistical distances for the raw and decomposed data. The first two pairs have shown reverse results in terms of geographical and statistical distances. Even though the GD is smaller for the pair cp, the SD between the two cities of this pair is higher than the SD for the pair cu. When the hydrological impact has been ignored, the third pair

has still displayed similar results to what has been obtained when the effect of the hydrological data has been involved for the raw, long, and seasonal data. However, the highest SD for the short-term component has been observed for the pair cp. For the raw and short-term component data, the SD values for the pair cp are higher than the SD for the pair cu.

5.10 Conclusion

In this chapter, the main tool is the hypothesis testing which has been used to examine whether two multivariate time series datasets (1) have been generated from similar distributions, (2) are similar. Based on the results for this study, if the null hypothesis for the two cases above is rejected, it is recommended to use the decomposed data to provide more accurate and detailed decisions which clearly determine the component(s) that are responsible for rejecting the null hypothesis.

Additionally, the geographical distance and having the same river passes through two cities are factors that can significantly influence the results of a comparison process for the data of interest. Sometimes, there are variables that have an obvious impact on the results of the comparison and when we eliminate them, different outcomes can be obtained. These findings are produced by using the two samples Hotelling T-Squared, one way ANOVA, and Kruskall Wallis tests.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this study, decomposition of time series improves the forecasting accuracy for a number of Bayesian and Frequentist-based methods. These methods are Multiple Linear Regression (MLR), Transfer-Function-Noise (TF-Noise), and Bayesian Multiple Linear Regression models. It also has been noticed that the difference between the models MLR, TF-Noise, and BMLR is not big in terms of the accuracy of water discharge predictions. However, the models BMLR-BVAR have better results compared to them. These results have been obtained based on daily data collected for three cities located in New York State, which are Cohoes, Utica, and Poughkeepsie, and two rivers, which are Mohawk and Hudson Rivers. These data are collected for the variables temperature, precipitation, wind speed, tide, water discharge, and groundwater level for the period 2005-2014, where the data of year 2014 has been used to evaluate the constructed models. The main conclusions from the study are: On the basis of MSE values, better results for the three models MLR, TF-Noise, and BMLR models have been obtained by using the decomposed data rather than the raw data. For example, for Cohoes city, the MSE value has reduced from 0.51 for the MLR constructed using the raw data to 0.40 for the CMLR constructed using the decomposed data. However, even better results have been obtained for the combined models, combined MLR and combined BMLR, when we modelled the short-term component using a VAR(1) model. For example for Utica city, the value of MSE has declined from 0.56 for BMLR using the raw data to 0.10 for BMLR-BVAR model using the decomposed data.

Also, based on the results of this research, the performance of the combined MLR that is augmented with an autoregressive model for the random error outperforms the combined MLR without an autoregressive model for the errors, thereby accounting

for the autocorrelation of the errors. Based on the results of the three cities, there is no significant impact of the similarity results on the forecasting accuracy of water discharge values based on the DIC values. This result is derived when an informative prior distribution with hyper-parameters that are related to the results of MLR model for the city that has the highest similarity measure to the city of interest is used.

The decomposed data, which are known as the components, are three the long, seasonal, and the short-term component, extracted using the KZ filter. There are two parameters that control the behaviour of the KZ filter, which are the window width and the number of iterations. Different parameters can be chosen based on the number of days that need to be filtered out and the resulting R Squared value, where the parameters that lead to produce the highest R Squared value can be selected. For example, in our analysis and based on the resultant R Squared values, we have specified the parameters 29 days and 3 iterations, and 15 days and 5 iterations for the three cities. Based on the KZ filter, the contribution of the three scales of motions to the total variance for the response variable can be computed. The KZ filter is one of the best techniques that can be used to detect and track changes in a time series due to its simplicity, accuracy, and its ability to deal with the missing values.

In each chapter a comparison process has been conducted between the models constructed using the raw data and the models constructed using the decomposed data. Chapters 2 and 3 present the analyses that are carried out using the frequentist-based statistical methods and Chapter 4 deals with the Bayesian-based statistical method. Typically, the assumption that the residual terms have to be uncorrelated is required to be satisfied in the constructed model to obtain accurate forecasting results. However, this assumption is often not achieved when the analysed data are time series data. This statistical problem has been extensively considered in MLR models built using the raw data. But, in CMLR model, which is constructed using decomposed time series data, the problem of autocorrelation between the residual terms of this model has not previously been considered. In Chapter 2 we have developed a CMLR-Noise model constructed using a MLR model for the data of the three components the long, seasonal, and the short-term component, with an AR(1) model for the residual terms that are serially correlated. This model has substantially improved the prediction accuracy by removing the impact of the autocorrelation between the residual terms. Based on the model selection methods used, the forecasting models constructed using the decomposed data and an AR(1) model for the disturbances outperform the forecasting models constructed using the raw data. For example, for Utica's city data, the AIC value reduced from 6590.497 to 2453.619 when we added an AR(1) model to the CMLR model.

In Chapter 3 we have shown that the CTF-Noise model constructed using the decomposed data provides better results than the TF-Noise model constructed using the raw data. For instance, for Utica city the AIC value reduced from 2516.409

for the TF-Noise model using the raw data to 2448.503 for the CTF-Noise model using the decomposed data. The prewhitened values have been used rather than the original values for the raw and the decomposed data. The prewhitening process is applied to provide a filter that can be used to transfer the input and output series into a white noise series which is devoid of autocorrelations. The process of obtaining the prewhitened values has been carried out by identifying a tentative model for the inputs where the mechanism of the identification depends on the behaviour of the SACF and SPACF for the input series. Also, to examine the relationship between the input and the output variables, the SCCF function has been used. The CTF-Noise model involves a number of lagged variables which can be determined based on the SCCF. The structure of this model is built using a difference equation. An AR(1) model is specified to fit the data of the residual terms for the final model.

The third procedure considered to enhance the prediction accuracy, which is presented in Chapter 4, is carried out by using Bayesian analysis to estimate the parameters of the models constructed using the decomposed data and also the final combined model. Bayesian analysis enables us to overcome an important issue in the forecasting process, which is the uncertainties in data, parameters, and the structure of the model. We have used the BMLR model and the BVAR of order 3 structure to fit models for the decomposed data. Firstly, the BMLR has been utilised to model the data of the three components, the long, seasonal, and the short-term component and also to fit the final combined BMLR model. Secondly, the BMLR has been used to model the data of the long and seasonal components, and the model BVAR of order 3 has been used to model the data of the short-term component and the result is the final combined BMLR-BVAR(3) model. Both models produced better results than the raw data-based BMLR model according to the MSE and DIC values. For Utica's city data, the DIC values reduced from 6593.116 for the combined BMLR to 521.385 for the BMLR-BVAR(3). The outperformance of the proposed model can be attributed to the inclusion of the three lagged variables, which are lags 1, 2, and 3, for the variables water discharge, precipitation, and groundwater level for the short-term component.

The proposed combined Bayesian models have a number of desirable features. Firstly, the variance of the models, σ^2, is treated as a random and an unknown value to incorporate the uncertainty of the parameters. For BMLR, we used the Inverse Gamma (IG) distribution to describe the behaviour of the variance (or we can use the Gamma distribution for the precision (the inverse of variance)). Secondly, the multivariate normal (MN) distribution has been used to describe the behaviour of the model's coefficients with two parameters, which are the mean vector and the variance-covariance matrix for the coefficients. Additionally, the likelihood is originally distributed with a MN distribution. So, using these two distributions, the multivariate normal for the mean vector (coefficients) and IG for the variance of the

model, together constructs a conjugate prior which is the Normal-Inverse-Gamma distribution (NIG). The product of the likelihood by this conjugate prior will yield a posterior function that has a NIG distribution.

Similarly, for the BVAR of order 3, we have used the Inverse Wishart, IW, distribution to generate data for the unknown variance-covariance matrix for the model, and the Minnesota prior to generate data for the coefficients. Using of the Minnesota prior will reduce the number of the parameters in the BVAR model. In other words, this prior will handle the problem of the overfitting which is one of the most important issues in the VAR models. The Minnesota prior is a special case of the conditional Normal-Inverse-Wishart, NIW prior. The covariance matrix of the coefficients of the VAR model that is used in the Minnesota prior is a diagonal matrix estimated using "equation-by-equation" AR models. Also, the likelihood function has a MN distribution. The resulting posterior distribution, therefore, is a Normal-Inverse-Wishart distribution. Even though the results of estimation using Bayesian Analysis either with non-informative or informative priors are nearly similar to the estimates that are produced using the Maximum Likelihood (MLE) method, the accuracy of Bayesian estimates is much better based on the credible intervals.

Also, working within Bayesian framework enables us to change the classical method of obtaining predictions for new outputs. In Bayesian analysis, this is implemented by using the resultant predictive distribution by using either numerical simulations or mathematical derivations. The predictive distribution has been computed using the Random Walk MCMC algorithm. Besides, the predictive distribution is of particular interest where it provides better insight about the model and the quality of the method used. The credible intervals are another reason to prefer using Bayesian analysis instead of Frequentist analysis. The credible intervals produce a range of values such that the value of the parameter of interest falls within this range with a pre-specified probability.

Frequentist-based models, which are regression and vector autoregressive models, have been applied for all the three cities. In addition to the models regression and vector autoregressive, the transfer function-noise model has been also applied for Poughkeepsie, Cohoes, and Utica cities. Bayesian-based models, which are Bayesian regression and Bayesian vector autoregressive models, have been also applied to forecast the water discharge for the three cities. The methodologies used are applicable for any time series data that are composed of embedded components, in particular, in the economics field, where different variables need to be analysed at the same time.

To forecast daily future values for the water discharge for Poughkeepsie city, three different models have been constructed. Figure 6.1 shows the three developed models that have been graphed with the original values (blue line) of the water discharge. The first model (MLR, red line, MSE=0.49) is fitted using the MLR model for the

raw data. This model is written as follows:

$$\widehat{WD}_t = -0.347TE_t + 0.414PR_t + 0.159TD_t - 0.321GW_t. \quad (6.1)$$

where WD, TE, PR, TD, and GW, denote the water discharge, temperature, precipitation, tide, and groundwater level, respectively. The second model (Combined MLR (CMLR), green line, MSE=0.37) has been built by combining the variables of Equations 2.16, 2.17, and 2.18.

The third model (CTF-Noise, darkred line, MSE=0.34) has been constructed using the CTF-Noise for the decomposed series for the long, seasonal, and short-term components, which is shown in Equation 3.40, which has been built using the variables of Equations 3.24, 3.36, and 3.38. Based on the MSE values, the CTF-Noise model is the best model as it provides the lowest MSE value. Overall, apart from the data of March and October, our forecasts agree with the original values well, showing an upward, downward, and then upward trends starts from the beginning of the year and captured the seasonality toward the end of the year. It appears that the performance of CMLR is more close to CTF-Noise models than to MLR model, which might be attributed to the use of the decomposed data in these two models, CMLR and CTF-Noise.

Figure 6.2 shows the original values (blue line, with a straight-line to describe the missing values that have been estimated from mid-October until mid-December) for the water discharge for Cohoes city for year 2014 along with the (1) Multiple Linear Regression model (MLR, red line, MSE=0.51) (2) Combined Multiple Linear Regression model (CMLR, green line, MSE=0.40), and (3) Combined Transfer Function-Noise model (CTF-Noise, darkred line, MSE=0.38). The CTF-Noise model is the best constructed model depending on the MSE values.

For Utica city, Figure 6.3 shows the raw data of the water discharge (Original values, blue line) along with (1) CMLR model (red line) constructed using Equations 2.12, 2.13, 2.14, and AR(1) for the residual terms (2) CTF-Noise model (green line) constructed using Equations 3.71, 3.82, 3.93, and AR(1) model for the residual terms (3) combined BMLR-BVAR model (darkred line) constructed by applying BMLR model for the data of the long and seasonal components with a non-informative independent normal prior distribution, and BVAR model for the data of the short-term component with Minnesota Prior distribution. Based on the MSE values, the performance of BMLR-BVAR model is the best compared to the other models, where the MSE values for these models are 0.222, 0.116, and 0.102, respectively.

With regard to the results of Chapter 5, it has been shown that a decision about the cities' data similarity can be obtained using the hypothesis testing. To support the decision of data similarity, the MSE values of the MLR can also be considered. Knowing that the data for two or more objects are similar can help us to use an object's forecasting model to forecast the future values for another object.

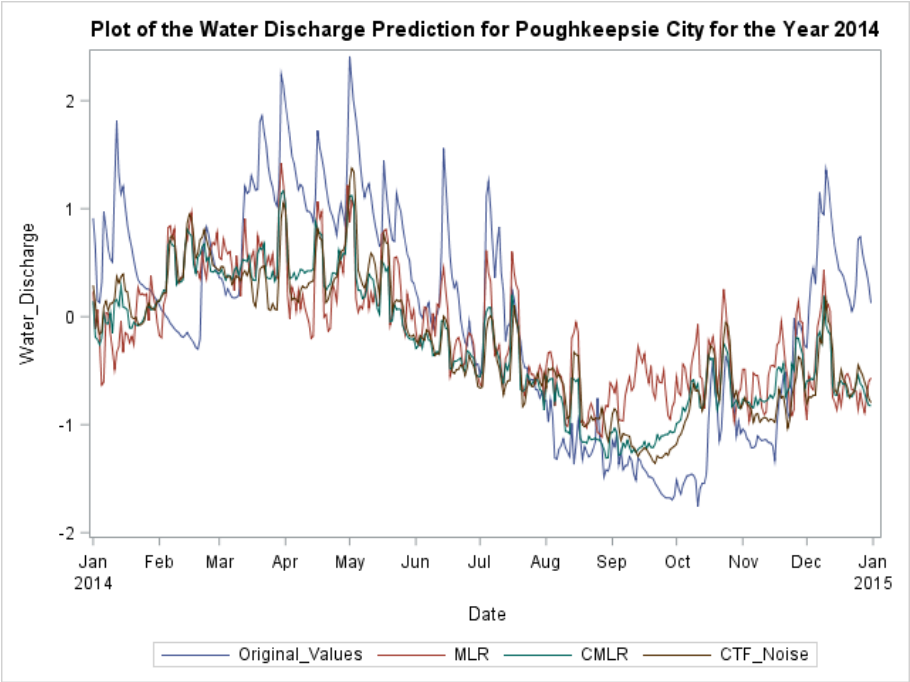

Figure 6.1: The Original Data and the Three Developed Models, Multiple Linear Regression (MLR), Combined Multiple Linear Regression (CMLR), and Combined Transfer Function-Noise (CTF-Noise), for the Water Discharge for Poughkeepsie city.

In this study we show that there is no difference between the mean distance of the raw and the decomposed data of the three cities based on the one way ANOVA test for the mean distance/dissimilar. Using a number of Euclidean and non-Euclidean distance measures, the distance between the covariance matrices of the variables of the studied cities is computed. The variables considered are the independent variables that have been used to construct MLR models for the cities of interest using the raw data. It has been also noted that the MSE values of the MLR models of the raw data are close to each other. So, considering these results, we are able to use one of the MLR models for one city to forecast future values for another city.

Also, we show the possibility of determining the component(s) and then the variable(s) that are responsible for rejecting the null hypothesis. The null hypothesis here claims that the mean distances are equal. This determination is not possible if we apply hypothesis testing using the raw data. Two mean types have been considered; the first one is the mean vector of variables of a city where the Hotelling T-Squared test has been used while the second mean, which is a scalar, has been computed using the mean of distances for a number of years. The distance has been computed using a number of Euclidean and non-Euclidean distance measures for covariance matrices

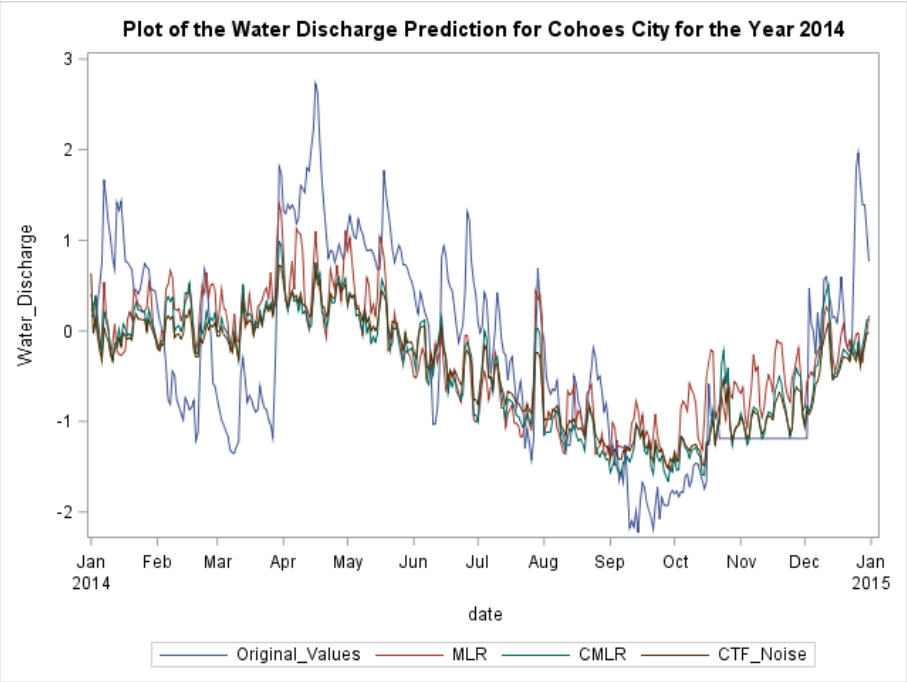

Figure 6.2: The Raw Data and the Three Developed Models for the Water Discharge for Cohoes city.

for cities' variables. The one way ANOVA and Kruskal Wallis tests have been applied in case that the data are normally or not normally distributed, respectively. Using this methodology will essentially enable us to know the events that are behind each component and lead to the result that there is a significant difference between the data of the studied cities. Based on their performance with respect to the three pairs, the same order has been observed from smallest to largest for the eight distance measures used. ProcrustesShape, Procrustes, and Cholesky have provided the smallest distances for the raw and the decomposed data.

Finally, to get the full picture of dissimilarity for multivariate time series datasets, part of this thesis has been devoted to examine the dissimilarity between multivariate time series datasets based on the frequency domain. In Chapter 5 we present three new developed distance measures. Three features from Frequency domain are used. These features are the periodogram, power spectral density and the cross spectral density functions. The data of these features are described by two matrices and a vector. The first matrix is the $X^T X$ matrix of the periodograms of the variables of temperature, precipitation, tide, wind speed, water discharge, and groundwater level, this matrix is a Positive Definite matrix. The second matrix, which is the Power Spectral Density Matrix (PSDE), has been built using power spectral density and

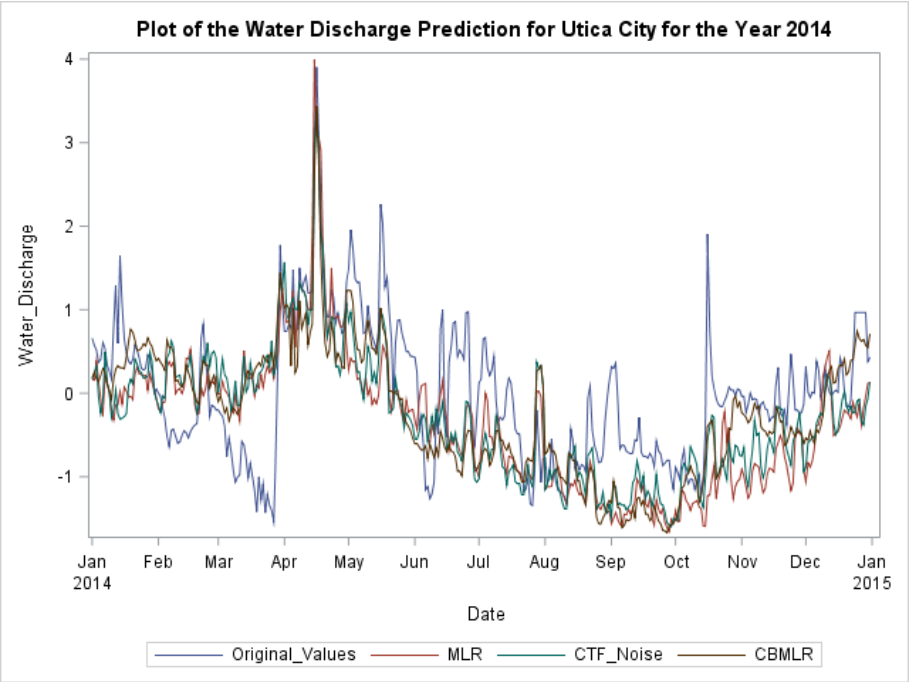

Figure 6.3: The Raw Data and the Three Developed Models for the Water Discharge for Utica city.

cross spectral density functions for the studied variables.

Using these matrices, we have applied the same distance measures that have been previously used with the covariance matrices. The distance measures are Euclidean, Procrustes, Riemannian, Procrustes Shape, Cholesky, Power, Log Euclidean, and RiemannianLe. We also use the Eigenvalues vectors of the PSDE matrix to derive a distance measure using the Euclidean distance. The results obtained using these new developed distance measures are not completely matched the dissimilarity results that are obtained using the covariance matrices, which are time domain-based data.

In case that we have a reasonable number of objects, for example, cities, the behaviour of the previous and new distance measures can be assessed using one of data mining techniques, such as clustering and classification analyses. Overall, the major contributions of the thesis are the development of new statistical models for accurate forecast for water discharge and new dissimilarity measures for time series data based on a number of frequency domain features.

6.2 Future Work

- Constructing a combined model by applying VAR process for the three components and comparing the results with the VAR for raw data.
- Constructing a combined Frequency-Domain Linear Regression for the three components and comparing it with a Frequency-Domain Linear Regression for raw data.
- Constructing a combined model by applying neural network model for the three components and comparing the results with a neural network model for the raw data.
- Applying Bayesian analysis for estimating a combined model constructed using TF-Noise model and comparing it with a TF-Noise model built by using raw data and estimated by using Bayesian analysis. The application of Bayesian analysis will significantly reduce the number of resultant parameters, which is known as the over fitting problem.
- Applying Bayesian Hierarchical analysis for a combined model constructed using TF-Noise model and comparing it with a TF-Noise model constructed using raw data and estimated by using Bayesian analysis.
- Applying BVAR model for the three components, long, seasonal, and short, and construct the final combined Bayesian VAR model and then compare the results with the results of normal Bayesian.
- Applying Bayesian hierarchical modelling, which is also known as multilevel or random-effects models, for a combined model could lead to provide results that are more significant compared to normal Bayesian analysis. Hierarchical Bayesian means that the hyperparameters of the model themselves will have distributions.
- Investigate the possibility of constructing normal (using raw data) and a combined Frequency-Domain Linear Regression using Bayesian analysis.
- In multivariate time series analysis and in case that there is a number of cities that enables us to perform one of data mining techniques, such as clustering and classification, it will be possible to assess the performance of these different distance measures.
- Also, for any data mining analysis, there is a possibility of performing a comparison between the results of a distance measure that is based on the covariance

matrix for the raw data and the distance measure that is based on a combined covariance matrix (similar to the idea of a pooled covariance matrix but by using the covariance matrices of long, seasonal, and short instead of the covariance matrices for the two tested groups).

The process of applying some analysis such as clustering and classification has many advantages. For example, it is well known that with times series data there is a need to perform some specific transformations, for instance, the seasonal adjustments. Therefore, it will be easier to implement these transformations for a group of series without examining each series separately.

- It is worth to apply the methodologies proposed in this chapter in some fields, specifically in the marketing, economic, environment.
- Applying the KZ filter to the covariance matrices and then apply one of the pattern recognition analyses, cluster or classification, and then make a comparison between the results of these analyses before and after applying KZ to the covariance matrix.

Bibliography

[1] Ahelegbey, D., Billio, M., and Casarin, R. Bayesian Graphical Models for Structural Vector Autoregressive Processes. *Journal Of Applied Econometrics 31* (2016), 357–386.

[2] Aitchison, J. Two Papers on the Comparison of Bayesian and Frequentist Approaches to Statistical Problems of Prediction : Bayesian Tolerance Regions. *Journal of the Royal Statistical Society.Series B (Methodological) 26*, 2 (2016), 161–175.

[3] Albostan, A., and Önöz, B. Implementation of Chaotic Analysis on River Discharge Time Series. *Energy and Power Engineering*, March (2015), 1225–1231.

[4] Alvarez, I., Niemi, J., and Simpson, M. Bayesian Inference for a Covariance Matrix. *Ann. Statist. 20* (2014), 1669–1696.

[5] Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Log-Euclidean Metrics for Fast and Simple Calculus on Diffusion Tensors. *Magnetic Resonance in Medicine 56*, 2 (2006), 411–421.

[6] Astatkie, T., and Watt, W. Multiple-Input Transfer Function Modeling of Daily Streamflow Series Using Nonlinear Inputs. *Water Resources Research 34*, 10 (1998), 2717–2725.

[7] Barnard, J., McCulloch, R., and Meng, X. Modelling Covariance Matrices in Terms of Standard Deviations and Correlations With Applications to Shrinkage. *Statistica Sinica 10*, 4 (2000), 1281–1311.

[8] Beal, D. SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria. *SAS Global Forum*, 1 (2005).

[9] Bendat, J. A General Theory of Linear Prediction and Filtering. *Journal of the Society for Industrial and Applied Mathematics 4*, 3 (1956), 131–151.

[10] Bendat, J., and Piersol, A. *Random Data Analysis and Measurement Procedures*. John Wiley & Sons, Inc., 2000.

[11] Bernardo, J., and Smith, A. *Bayesian Theory*. John Wiley & Sons, Inc., 2009.

[12] Berry, D. *Statistics: A Bayesian Perspective*. Duxbury, 1996.

[13] Bierkens, M., Knotters, M., and Geer, F. Calibration of Transfer Function-Noise Models to Sparsely or Irregularly Observed Time Series. *Water Resources Research 35*, 6 (1999), 1741–1750.

[14] Bishop, C. *Pattern Recognition and Machine Learning*. Springer-Verlag Berlin, Heidelberg, 2006.

[15] Bowerman, B., O'Connell, R., and Koehler, A. *Forecasting, Time Series, and Regression*, fourth ed. Belmont, CA: Thomson Brooks/Cole., USA, 2005.

[16] Brocklebank, J., Dickey, D., and Choi, B. *SAS for Forecasting Time Series*. SAS institute, 2018.

[17] Caiado, J., Crato, N., and Peña, D. A Periodogram-Based Metric for Time Series Classification. *Computational Statistics and Data Analysis 50*, 10 (2006), 2668–2684.

[18] Canova, F. Bayesian VARs. In *Methods for Applied Macroeconomic Research*. Princeton University Press, 2011, ch. 10.

[19] Chan, J. Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure. *Journal of Business & Economic Statistics* (2018), 1–12.

[20] Chatfield, C. *The Analysis of Time Series- An Introduction*, 5th editio ed. Chapman and Hall CRC, London., 1996.

[21] Chbab, E., Noortwijk, J., and Duits, M. Bayesian Frequency Analysis of Extreme River Discharge. *Proceedings of International Symposium on Flood Defence* (2000).

[22] Chen, M., and Shao, Q. Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics 8*, 1 (1999), 69–92.

[23] Cheng, D., and Tadashi, Y. A Study on Uncertainty of Discharge in River Channel Using Stochastic Partial Differential Equation. *Procedia Engineering 154* (2016), 595–600.

[24] Chiang, S., Guindani, M., Yeh, H., Haneef, Z., Stern, J., and Vannucci, M. Bayesian Vector Autoregressive Model for Multi-Subject Effective Connectivity Inference Using Multi-Modal Neuroimaging Data. *Human Brain Mapping 38*, 3 (2017), 1311–1332.

[25] Christiano, L. Christopher A . Sims and Vector. *Scand. J. of Economics 114*, 4 (2012), 1082–1104.

[26] Corduas, M., and Piccolo, D. Time Series Clustering and Classification by the Autoregressive Metric. *Computational Statistics and Data Analysis 52*, 4 (2008), 1860–1872.

[27] Cuaresma, J., Feldkircher, M., and Huber, F. Autoregressive Models: A Bayesian Approach. *Journal Of Applied Econometrics 31*, 4 (2016), 1371–1391.

[28] Damle, C., and Yalcin, A. Flood Prediction using Time Series Data Mining. *Journal of Hydrology 333*, 2-4 (2007), 305–316.

[29] Di, C., Yang, X., and Wang, X. A Four-Stage Hybrid Model for Hydrological Time Series Forecasting. *PLoS ONE 9*, 8 (2014).

[30] Dickey, D. Case Studies in Time Series. *Sugi 28 Paper 252-*, 1–9.

[31] Dickey, D. Regression with Time Series Errors. *Journal of the American Statistical Association 79* (1984), 118–124.

[32] Dong, K. *High-dimensional covariance matrix estimation with application to Hotelling ' s tests*. PhD thesis, Hong Kong Baptist University, 2015.

[33] Dryden, I., Hill, B., Wang, H., and Laughton, C. Covariance Analysis for Temporal Data, with Applications to DNA Modelling. *Stat* (2017), 218–230.

[34] Dryden, I., Koloydenko, A., and Zhou, D. Non-Euclidean Statistics for Covariance Matrices With Applications to Diffusion Tensor Imaging. *The Annals of Applied Statistics 3*, 3 (2009), 1102–1123.

[35] Engeland, K., Xu, C., and Gottschalk, L. Assessing Uncertainties in a Conceptual Water Balance Model using Bayesian Methodology. *Hydrological Sciences Journal 50*, 1 (2005), 45–63.

[36] Eskridge, R., Ku, J., Rao, S., Porter, P., and Zurbenko, I. Separating Different Scales of Motion in Time Series of Meteorological Variables. *Bulletin of the American Meteorological Society 78*, 7 (1997), 1473–1483.

[37] Falk, M., Marohn, F., Michel, R., Hofmann, D., Macke, M., Tewes, B., and Dinges, P. *A First Course on Time Series Analysis*. University of Wurzburg, 2012.

[38] Flaum, J., Rao, S., and Zurbenko, I. Moderating the Influence of Meteorological Conditions on Ambient Ozone Concentrations. *Journal of the Air & Waste Management Association 46* (2012), 35–46.

[39] Flury, B., and Riedwyi, H. *Multivariate Statistics: a Practical Approach*. London, Chapman and Hall, 1998.

[40] Fu, T. A Review on Time Series Data Mining. *Engineering Applications of Artificial Intelligence 24*, 1 (2011), 164–181.

[41] Garver, J. Ice Jam Flooding on the Lower Mohawk River and the 2018 Mid-Winter Ice Jam Event. In *The 2018 Mohawk Watershed Symposium* (2018), no. March, Union College, Schenectady NY, pp. 12–18.

[42] Garver, J., and Cockburn, J. A Historical Perspective of Ice Jams on the Lower Mohawk River. In *The 2009 Mohawk Watershed Symposium* (2009), Union College, Schenectady NY.

[43] Geman, S., and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on pattern analysis and machine intelligence 6*, 6 (1984), 721–741.

[44] Gupta, R., Jurgilas, M., Kabundi, A., and Miller, S. Monetary Policy and Housing Sector Dynamics in a Large-Scale Bayesian Vector Autoregressive Model. *International Journal of Strategic Property Management 16*, 1 (2012), 1–20.

[45] Hamilton, J. *Time Series Analysis*. Princeton University Press, USA, 1995.

[46] Hammer, Ø., Harper, D., and Ryan, P. PAST-PAlaeontological STatistics, ver. 1.89, 2009.

[47] Harris, J., and Liu, L. Dynamic Structural Analysis and Forecasting of Residential Electricity Consumption. *International Journal of Forecasting 9*, 4 (1993), 437–455.

[48] Hasanah, Y., Herlina, M., and Zaikarina, H. Flood Prediction using Transfer Function Model of Rainfall and Water Discharge Approach in Katulampa Dam. *Procedia Environmental Sciences 17* (2013), 317–326.

[49] Herr, H., and Krzysztofowicz, R. Ensemble Bayesian Forecasting System Part I: Theory and Algorithms. *Journal of Hydrology 524* (2015), 789–802.

[50] Hogrefe, C., Rao, S., and Zurbenko, I. Detecting Trends and Biases in Time Series of Ozonesonde Data. *Atmospheric Environment 32*, 14-15 (1998), 2569–2586.

[51] Holmes, W. *Time Series: Sir Maurice Kendall and J. Keith Ord.* Edward Arnold, Great Britain, 1992.

[52] Holt, C., and Shore, R. *Bayesian Analysis in Economic Theory and Time Series Analysis.* PhD thesis, 1980.

[53] Hsiao, C. Panel Data Analysis - Advantages and Challenges. *Sociedad de Estad ıstica e Investigaci on Operativa 00*, 0 (2007), 1–63.

[54] Huang, X., Ghodsi, M., and Hassani, H. A Novel Similarity Measure Based on Eigenvalue Distribution. *Transactions of A. Razmadze Mathematical Institute 170*, 3 (2016), 352–362.

[55] Inc, S. I. Introduction to Bayesian Analysis Procedures. In *SAS/STAT® 14.1 User's Guide Introduction to Bayesian Analysis Procedures.* SAS Institute Inc., Cary, NC, USA, 2008, ch. 7, pp. 127–161.

[56] Inc, S. I. Bayesian Analysis of a Linear Regression Model. 2009.

[57] Inc, S. I. Getting Started: GENMOD Procedure: Bayesian Analysis of a Linear Regression Model, 2009.

[58] Inc, S. I. The Spectral Procedure. In *SAS/STAT®9.2 User's Guide.* , SAS Institute Inc., Cary, NC, USA, 2014, ch. 26, pp. 973–997.

[59] Janacek, J., and Swift, L. *Time series: Forecasting, Simulation, Applications.* Ellis Horwood Limited, 1993.

[60] Jian-Hong, S., and Song-Gui, W. The Spectral Decomposition of Covariance Matrices for the Variance Components Models. *Journal of Multivariate Analysis 97*, 10 (2006), 2190–2205.

[61] Johnson, R., and Wichern, D. *Applied Multivariate Statistical Analysis*. Pearson; 6 edition, USA, 2008.

[62] Kalpakis, K., Gada, D., and Puttagunta, V. Distance Measures for Effective Clustering of ARIMA Time-Series. In *Proceedings 2001 IEEE International Conference on Data Mining* (2001), pp. 273–280.

[63] Khan, M. Transfer Function Model for Gloss Prediction of Coated Aluminum Using the AIMA Procedure. In *In Proceedings of the Fifteenth Annual SAS Users Group International Conference* (1990), Kuwait Institute for Scientific Research Introduction, pp. 517–522.

[64] Kilian, L., and Helmut, L. Bayesian VAR Analysis. In *Structural Vector Autoregressive Analysis*. Cambridge University Press, 2017, ch. 5.

[65] Lebrun, P. *Bayesian Design Space Applied to Pharmaceutical Development*. PhD thesis, University DE LIÈGE., 2012.

[66] LeSage, J., and Krivelyova, A. A Spatial Prior for Bayesian Vector Autoregressive Models. *Journal of Regional Science 39*, 2 (1999), 297–317.

[67] Li, Y., and Wong, K. Signal Classification by Power Spectral Density: An Approach via Riemannian Geometry. 2012.

[68] Li, Y., Wong, K., and DeBruin, H. EEG Signal Classification Based on a Riemannian Distance Measure. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)* (Toronto, 2009), pp. 268–273.

[69] Lima, C., Lall, U., Troy, T., and Devineni, N. A Hierarchical Bayesian GEV Model for Improving Local and Regional Flood Quantile Estimates. *Journal of Hydrology 541* (2016), 816–823.

[70] Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. Prior Distributions. In *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman and Hall/CRC, 2012, ch. 5, pp. 81–102.

[71] Lynch, S. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer Science & Business Media., 2007.

[72] Melard, G., Paesmans, M., and Roy, R. Consistent Estimation of the Asymptotic Covariance Structure of Multivariate Serial Correlations. *Journal of Time Series Analysis 12*, 4 (1991), 351–361.

[73] Milanchus, M., Rao, S., and Zurbenko, I. Evaluating the Effectiveness of Ozone Management Efforts in the Presence of Meteorological Variability. *Journal of the Air and Waste Management Association 48*, 3 (1998), 201–215.

[74] Montgomery, D., Jennings, C., and Kulahci, M. *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons, 2015.

[75] Montgomery, D., Johnson, L., and Gardiner, J. *Forecasting and Time Series Analysis*. McGraw-Hill Companies, 1990.

[76] NCSS, L. Hotelling ' s Two-Sample T squared. In *PASS Sample Size Software*. 1998, ch. 600, pp. 1–11.

[77] Neagu, R., and Zurbenko, I. Algorithm for Adaptively Smoothing the Log-Periodogram. *Journal of the Franklin Institute 340*, 2 (2003), 103–123.

[78] Pennec, X., Fillard, P., and Ayache, N. A Riemannian Framework for Tensor Computing. *HAL RR-5255, INRIA* (2004), 34 pages.

[79] Perrott, M., and Cohen, R. An Efficient Approach to ARMA Modeling of Biological Systems with Multiple Inputs and Delays. *IEEE Transactions on Biomedical Engineering 43*, 1 (1996), 1–14.

[80] Pigoli, D., and Secchi, P. Estimation of the Mean for Spatially Dependent Data Belonging to a Riemannian Manifold. *Electronic Journal of Statistics 6* (2012), 1926–1942.

[81] Rao, S., Zalewsky, E., and Zurbenko, I. Determining Temporal and Spatial Variations in Ozone Air Quality. *Journal of the Air & Waste Management Association (1995) 45*, 1 (1995), 57–61.

[82] Rao, S., and Zurbenko, I. Detecting and Tracking Changes in Ozone Air Quality. *Air & Waste 44*, 9 (1994), 1089–1092.

[83] Rao, S., Zurbenko, I., Neagu, R., Porter, P., Ku, J., and Henry, R. Space and Time Scales in Ambient Ozone Data. *Bulletin of the American Meteorological Society 78*, 10 (1997), 2153–2166.

[84] Reis, D., and Stedinger, J. Bayesian MCMC Flood Frequency Analysis with Historical Information. *Journal of Hydrology 313*, 1-2 (2005), 97–116.

[85] Romão, J., Guerreiro, J., and Rodrigues, P. Tourism Growth and Regional Resilience : The ' Beach Disease ' and the Consequences of the Global Crisis of 2007. *Tourism Economics 22*, 4 (2016), 699–714.

[86] Scheller, M., Luey, K., and Garver, J. Major Floods on the Mohawk River (NY): 1832-2000. Tech. rep., Geology Department Union College, Schenectady NY (USA), 2008.

[87] Schumacher, C. Comments on "Short-Term Inflation Projections: A Bayesian Vector Autoregressive Approach". *International Journal of Forecasting 30*, 3 (2014), 645–647.

[88] Shijin, L., Lingling, J., Yuelong, Z., and Ping, B. A Hybrid Forecasting Model of Discharges Based on Support Vector Machine. *Procedia Engineering 28* (2012), 136–141.

[89] Smith, A., and Gelfand, A. Sampling Based Approaches to Calculating Marginal Densities. Tech. rep., the Office of Naval Research Herbert, Stanford, California, 1989.

[90] Smith, S. Moving Average Filters. In *The scientist and engineer's guide to digital signal processing*. 2003, ch. 15, pp. 277–284.

[91] Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. Bayesian Measures of Model Complexity anf Fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 64*, 4 (2002), 583–639.

[92] Stage, F., and Statements, U. *The ARIMA Procedure*. SAS Institute Inc., Cary, NC, USA, 2014.

[93] Stoica, P., and Moses, R. *Spectral Analysis of Signals*. Prentice Hall, Inc, Upper Saddle River, New Jersey 07458, 2004.

[94] Stokes, M., Chen, F., and Gunes, F. An Introduction to Bayesian Analysis with SAS/STAT® Software, 2014.

[95] Struzik, Z., and Siebes, A. The Haar Wavelet Transform in the Time Series Similarity Paradigm. In *European Conference on Principles of Data Mining and Knowledge Discovery* (Berlin, Heidelberg., 1999), Springer, pp. 12–22.

[96] Svetlíková, D., Komorníková, M., Kohnová, S., Szolgay, J., and Hlavčová, K. Analysis of Discharge and Rainfall Time Series in the Region of the Káštorské lúky Wetland in Slovakia. In *XXIVth conference of the Danubian countries on the hydrological forecasting. Conference E-papers. Bled* (2008), vol. 120, pp. 215–265.

[97] Theodoridis, S., and Koutroumba, K. *Pattern Recognition*. Academic Press, 2008.

[98] Tiao, G., and Box, G. Modeling Multiple Time Series with Applications. *Journal of the American Statistical Association 76*, 376 (1981), 802–816.

[99] Tsakiri, K., Marsellos, A., and Kapetanakis, S. Artificial Neural Network and Multiple Linear Regression for Flood Prediction in Mohawk River, New York. *Water (Switzerland) 10*, 9 (2018).

[100] Tsakiri, K., Marsellos, A., and Zurbenko, I. An Efficient Prediction Model for Water Discharge in Schoharie Creek, NY. *Journal of Climatology 2014* (2014), 1–10.

[101] Tsakiri, K., and Zurbenko, I. *Effect of Noise in Principal Component Analysis With An Application to Ozone Pollution*. PhD thesis, State University of New York, 2010.

[102] Tsay, R. *Analysis of Financial Time Series*. A Wiley-Interscience Publication JOHN WILEY & SONS, INC., 2006.

[103] Turner, R., and Sahani, M. Time-Frequency Analysis as Probabilistic Inference. *IEEE Transactions on Signal Processing 62*, 23 (2014), 6171–6183.

[104] Tutberidze, D., and Japaridze, D. Macroeconomic Forecasting Using Vector Autoregressive Approach. *Bulletin of Taras Shevchenko National University of Kyiv. Economics 2*, 191 (2017), 42–49.

[105] Vandaele, W. *Applied Time Series and Box-Jenkins Models*. Orlando ; London : Academic Press, c1983., 1983.

[106] Vanhatalo, E., and Kulahci, M. The Effect of Autocorrelation on the Hotelling. *Wiley Online Library*, September 2014 (2015).

[107] Vemulapalli, R., and Jacobs, D. Riemannian Metric Learning for Symmetric Positive Definite Matrices. *arXiv preprint arXiv:1501.02393.* (2015).

[108] Villani, M. Steady-State Priors for Vector Autoregressions. *Econometrics, Applied 24*, 4 (2019), 630–650.

[109] Vladimir, M., Auranen, K., and Halloran, M. Introduction to Bayesian Inference. Tech. rep., 8th Summer Institute in Statistics and Modeling in Infectious Diseases, 2016.

[110] Vol, E., and Nos, I. Transfer Function Models. In *Time Series Analysis, Fourth Edition.*, vol. 33. John Wiley & Sons, Inc., 1997, ch. 11, pp. 439–472.

[111] Volatility, S. *Methods for Empirical Macroeconomics By Gary Koop and Dimitris Korobilis*, vol. 3. 2010.

[112] Wang, H., Wang, C., Wang, Y., Gao, X., and Yu, C. Bayesian Forecasting and Uncertainty Quantifying of Stream Flows Using Metropolis–Hastings Markov Chain Monte Carlo Algorithm. *Journal of Hydrology 549* (2017), 476–483.

[113] Wang, J., Zhu, Y., Li, S., Wan, D., and Zhang, P. Multivariate Time Series Similarity Searching. *The Scientific World Journal 2014* (2014), 1–8.

[114] Wei, W. *Time Series Analysis: Univariate and Multivariate Methods*. Pearson Addison Wesley, 2006.

[115] Wise, E., and Comrie, A. Extending the Kolmogorov–Zurbenko Filter: Application to Ozone, Particulate Matter, and Meteorological Trends. *Journal of the Air and Waste Management Association 55*, 8 (2005), 1208–1216.

[116] Wooldridge, J. Basic Regression Analysis with Time Series Data. In *Econometric Analysis of Cross Section and Panel Data*. MIT press., 2012, ch. 10, pp. 1–17.

[117] Yang, W., and Zurbenko, I. Package ' kzft ' Kolmogorov-Zurbenko Fourier Transform and Applications, 2007.

[118] Yang, W., and Zurbenko, I. Kolmogorov-Zurbenko Filters. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 3 (2010), 340–351.

[119] Zhou, D. *Statistical Analysis of Diffusion Tensor Imaging*. PhD thesis, University of Nottingham., 2010.

[120] Zhou, D., Dryden, I., Koloydenko, A., Audenaert, K., and Bai, L. Regularisation, Interpolation and Visualisation of Diffusion Tensor Images using Non-Euclidean Statistics. *Journal of Applied Statistics 43*, 5 (2016), 943–978.

[121] Zurbenko, I., and Luo, M. Restoration of Time-Spatial Scales in Global Temperature Data. *American Journal of Climate Change 1* (2012), 154–163.

[122] Zurbenko, I., Porter, P., Rao, S., Ku, J., Gui, R., and Eskridge, R. Detecting Discontinuities in Time Series of Upper-Air Data: Development and Demonstration of an Adaptive Filter Technique. *Journal of Climate 9*, 12 II (1996), 3548–3560.

Appendix A

Appendix

For univariate linear regression where we have one predictor, the likelihood function for an observation y_i can be written as follows:

$$\ell\ (y_i|\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\big(-\frac{(y_i-\beta x_i)^2}{2\sigma^2}\big). \tag{A.1}$$

The full likelihood for all observations will be

$$\ell\ (\mathbf{y}|\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\big(-\sum_{i=1}^{n}\frac{(y_i-\beta x_i)^2}{2\sigma^2}\big). \tag{A.2}$$

A.1 Gibbs Sampling

The integrals of the posterior distribution produce the quantities of interest, such as mean, median, and variance. Hence, the main key in the Bayesian analysis is to calculate the integrals which in turn produce some inferences about the distribution. For simple posterior distributions, the integrals can be solved using some numerical integrations or even by using a paper and a pencil. But when the type of the posterior is relatively complex, such as a multivariate distribution (with many parameters), the best treatment will be the MCMC, which is the predominant method in Bayesian inference.

Gibbs sampler was proposed in the early 1990s by Geman and Geman, 1984, and Gelfand and Smith, 1990, [43], [89]. The major step in this sampling mechanism is to reduce the calculations of sampling from a high (multivariate) dimensional joint density into a number of samples of low (univariate) dimensional conditional distributions. In other words, rather than one joint sample with dimensional p, the number of the considered parameters p one dimensional samples will be generated. Then,

the posterior summaries of interest can be calculated by using the samples that have been drawn from the posterior distribution. Priori to use the samples to compute the required posterior summaries, it is fundamental to examine the convergence. This examination can be implemented by using either virtual or mathematical procedures. To illustrate Gibbs sampler, assume that our target distribution is $p(\theta|y)$ with $\theta = (\theta_1, \theta_2)$ and we are able to sample from $p(\theta_1|\theta_2, y)$ and $p(\theta_2|\theta_1, y)$. The process will begin by setting an initial values for the considered parameters, these initial values are θ_1^0 and θ_2^0, any iteration in Gibbs sampler involves the following two steps:

- Generate θ_1^n from $p(\theta_1|\theta_2^{n-1}, y)$.
- Generate θ_2^n from $p(\theta_2|\theta_1^n, y)$.

In Bayesian statistics, the Gibbs Sampling (GS) is regarded as one of the most extensively used methods to sample a sequence of data from the posterior distribution. This algorithm is originally a special case from the Metropolis-Hastings algorithm and can be used when:

- The posterior distribution under study is a multivariate distribution, and there is no possibility to sample using the two previous approaches, inversion and rejecting methods.
- There is feasibility to sample from the conditional distribution for each parameter.

A.2 The Inversion Method of Sampling

It is preferable and common to use an inversion method to draw a sample of data from a univariate distribution. Essentially, the inversion method follows two main steps:

- From the uniform distribution, a random number u between 0 and 1 is drawn.
- Computes $z = F^{-1}(u)$ from the $f(x)$.

To generate a number of observations from a uniform (0,1), we use the density function of the uniform and after that we will need to use the sampled data to calculate the integral from zero to z.

When there is a density that has no specific routines which allow to sample from it, drawing a sample of data from $u \sim U(0, 1)$ and computing z from $u = \int_L^z f(x)dx$ can be implemented. This inversion method is efficient, however, there are two major

limitations that affect using it as a general technique to draw samples from a posterior distribution. The first limitation is its failure to analytically derive the inverse function, then this method can not be used. The second limitation appears when the density under study is a multivariate density function. For example, if there are two variables, x and y, the reasonable solution to proceed the problem of having two unknown variables in one equation is to choose a value for one of these variables and then utilises the inversion method for drawing from the conditional distribution of the other variable. This procedure will significantly transfer the original two bivariate process into one of the sampling methods for a univariate conditional distribution. In fact, this idea is the basis for Gibbs sampling.

A.3 The Rejection Method of Sampling

There are different methods to be used when $F^{-1}(u)$ can not be computed. To some extent, the rejection method of sampling is regarded as one of the most important methods after the inversion approach. To sample from a specific distribution, $f(x)$, this technique can be defined by the following three steps:

- Suppose that a value z can be easily generated from a distribution $g(x)$ such that the values of $m * g(x)$ are relatively greater than $f(x)$ at approximately all points, where m is a constant.
- Calculate the ratio $R = \frac{f(z)}{m*g(z)}$.
- Sample $u \sim U(0, 1)$. When $R \geq u$, the value of z will be accepted as a draw from $f(x)$. Otherwise, repeat the process from step 1.

The term $m * g(x)$ is usually called an "envelop function". This method has two limitation points. First, determining an envelope function sometimes may not be easy. Second, the results may not be efficient.

A.4 Metropolis Algorithm

The Metropolis algorithm is one of the most important and common algorithms and this method is the foundation for the MCMC method. This technique is based on identifying a symmetric proposal distribution, which is known as a transition function which is:

$$p(\boldsymbol{\theta^r}|\boldsymbol{\theta^{(r-1)}}) = p(\boldsymbol{\theta^{(r-1)}}|\boldsymbol{\theta^r}).$$

Using this function a number of samples are drawn where these draws either rejected or not based on a pre-specified decision rule.

A.5 Metropolis-Hasting Algorithm (MH)

The MH algorithm is a method used to generate samples from a probability distribution, which is the full joint density function. The feature that discriminates this method is its ability to work with multivariate distributions and also with this method we do not need an envelope function. The major steps of this technique are:

- For the parameter θ, a starting value will be established: $\theta^{j=0} = L$. Then, set $j = 1$.
- Draw a candidate parameter, which is θ^c, from a proposal density, which is $p(.)$.
- Then, the ratio $G = \frac{f(\theta^c)p(\theta^{j-1}|\theta^c)}{f(\theta^{j-1})p(\theta^c|\theta^{j-1})}$.
- The computed ratio, G, will be compared with a Uniform(0,1) random draw u. If $G \geq u$, then we will set $\theta^j = \theta^c$. Otherwise, $\theta^j = \theta^{g-1}$.
- Set $j = j + 1$, and then to step 2 to draw the required number of samples.

The starting values in the first step could be derived from MLE or even an arbitrary numbers. The MCMC theory says that the algorithm's stationary distribution will be the required posterior distribution, regardless of the starting values selected. The stationary distribution is the distribution in which the Markov chain produced by the algorithm converges. For more information see [71].

A.6 MCMC

As a consequence to the limitations of the inversion and rejection methods, the need to apply an efficient algorithm that can proceed all the limitations in the two previous approaches appears. In the last few decades, Markov Chain Monte Carlo (MCMC) algorithms have been extensively applied to facilitate the sampling process from any complex density function. The first part of this mechanism, Markov Chain, is responsible about generating new value from the posterior distribution, given that the previous value is known. This will lead to simulate a sequence of data from the posterior distribution. The second part of the name is related to the process of calculating the integral for each simulated value.

Gibbs Sampling Using the Inversion Method

To use this method, the conditional distribution for each variable has to be computed.

A.7 Distributions

- Normal Distribution, where the Density function is:

$$f(\theta)\frac{1}{\sqrt{2\pi}\sigma}\exp\big(-\frac{(\theta-\mu)^2}{2\sigma^2}\big). \tag{A.3}$$

  In MCMC it is preferable to use the parameter τ instead of σ^2 to simplify the calculations. So, Equation A.3 can be rewritten as the following:

$$\frac{\tau}{\sqrt{2\pi}}\exp\left(-\frac{\tau(\theta-\mu)^2}{2}\right). \tag{A.4}$$

- Gamma Distribution, where the Density function of this distribution can be written as the following:

$$f(\theta)=\frac{\beta^\alpha}{\Gamma\alpha}\theta^{\alpha-1}\exp(-\beta\theta). \tag{A.5}$$

  The parameters of this distribution are α and β which they are the shape and scale parameters, respectively.

- Inverse Gamma Distribution: When we need to specify a distribution to the reciprocal of a random variable that is distributed according to the Gamma distribution, the Inverse Gamma is the required distribution. In Bayesian analysis, when a non-informative prior is used, this density function appears as the marginal posterior for the unknown variance of a normal model.

$$f(\theta)=\frac{1}{\beta^\alpha\Gamma\alpha}\theta^{-(\alpha+1)}\exp -1/\beta\theta. \tag{A.6}$$

- Chi-Squared distribution, where the density function can be written as:

$$f(\theta)=\frac{1}{\Gamma(v/2)2^{v/2}}\boldsymbol{\theta}^{(v/2)-1}\exp -\theta/2. \tag{A.7}$$

- The mean and the precision of the posterior are usually a combination of the mean and precision of the prior and likelihood functions. If $y|mu\sim N(\mu,\sigma^2)$ and $\mu\sim N(\mu_0,\sigma_o^2)$, then

$$\mu\sim N\left(\frac{\sigma_o^2}{(\sigma^2/n)+\sigma_o^2}y+\frac{\sigma^2}{(\sigma^2/n)+\sigma_o^2}\mu_0\right),\left(\frac{1}{\sigma_o^2}+\frac{n}{\sigma^2}^{-1}\right).$$

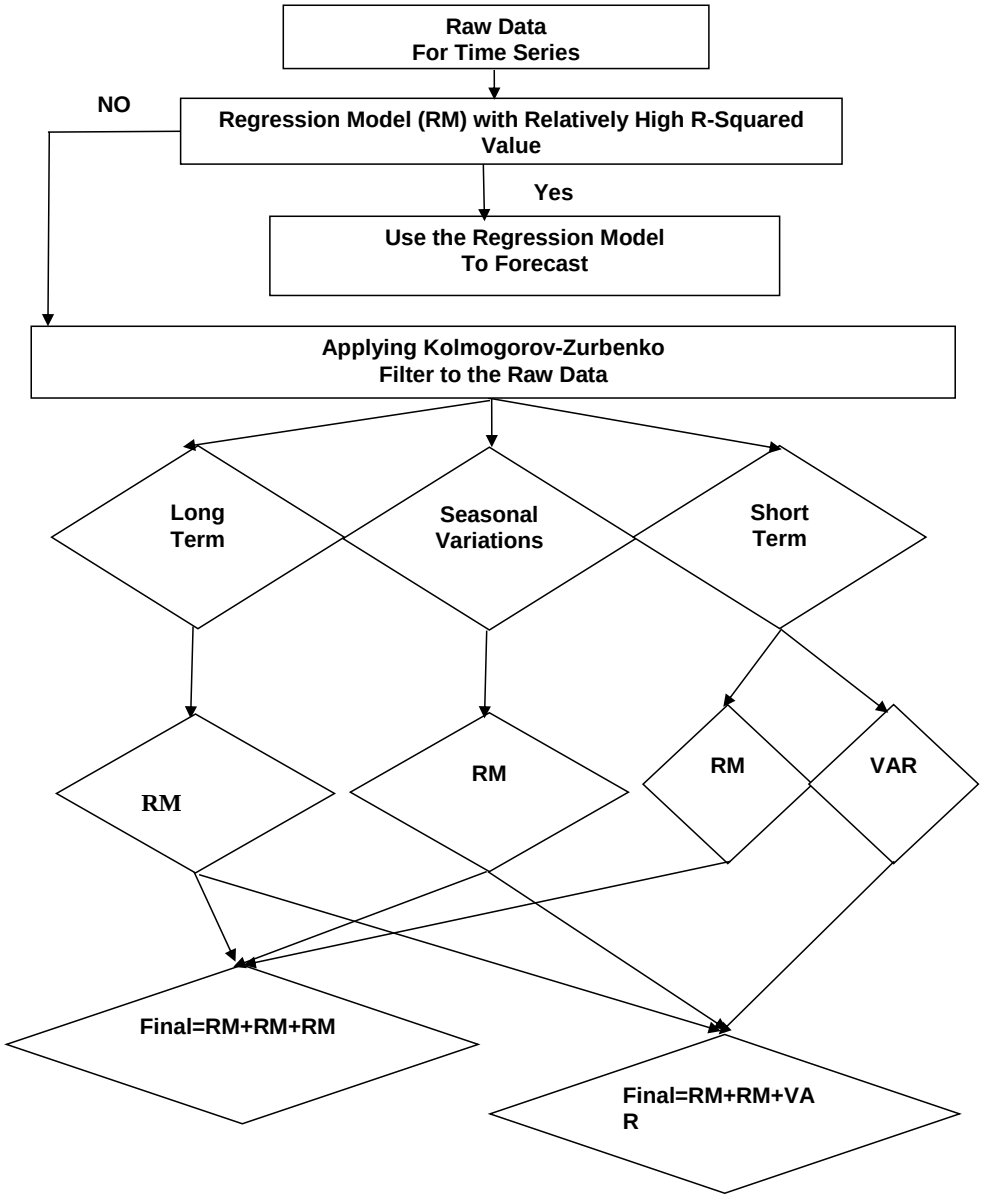

Figure A.1: Chart Illustrates the Steps Taken to Construct the Developed Models for Cohoes City.

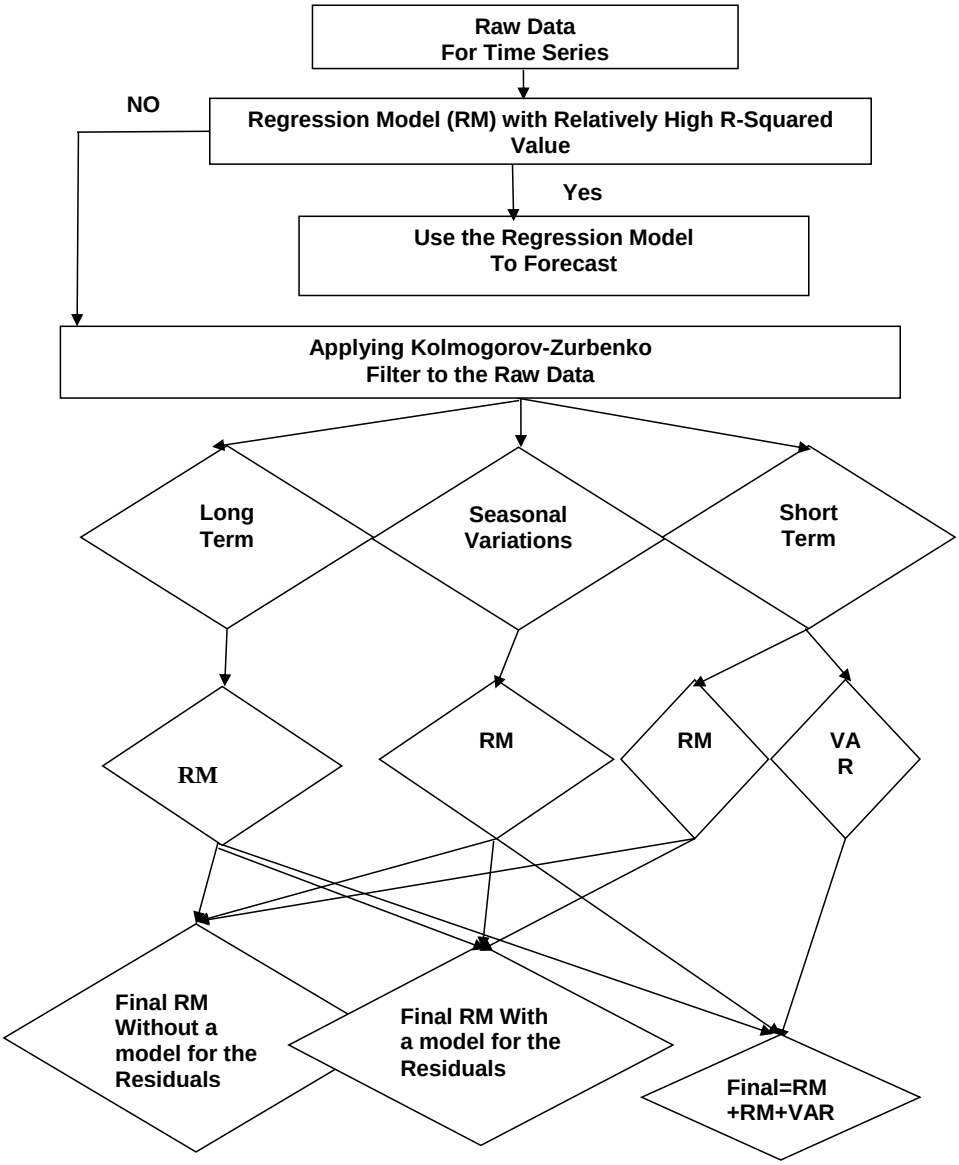

Figure A.2: Chart Illustrates the Steps Taken to Construct the Developed Models for Utica City.

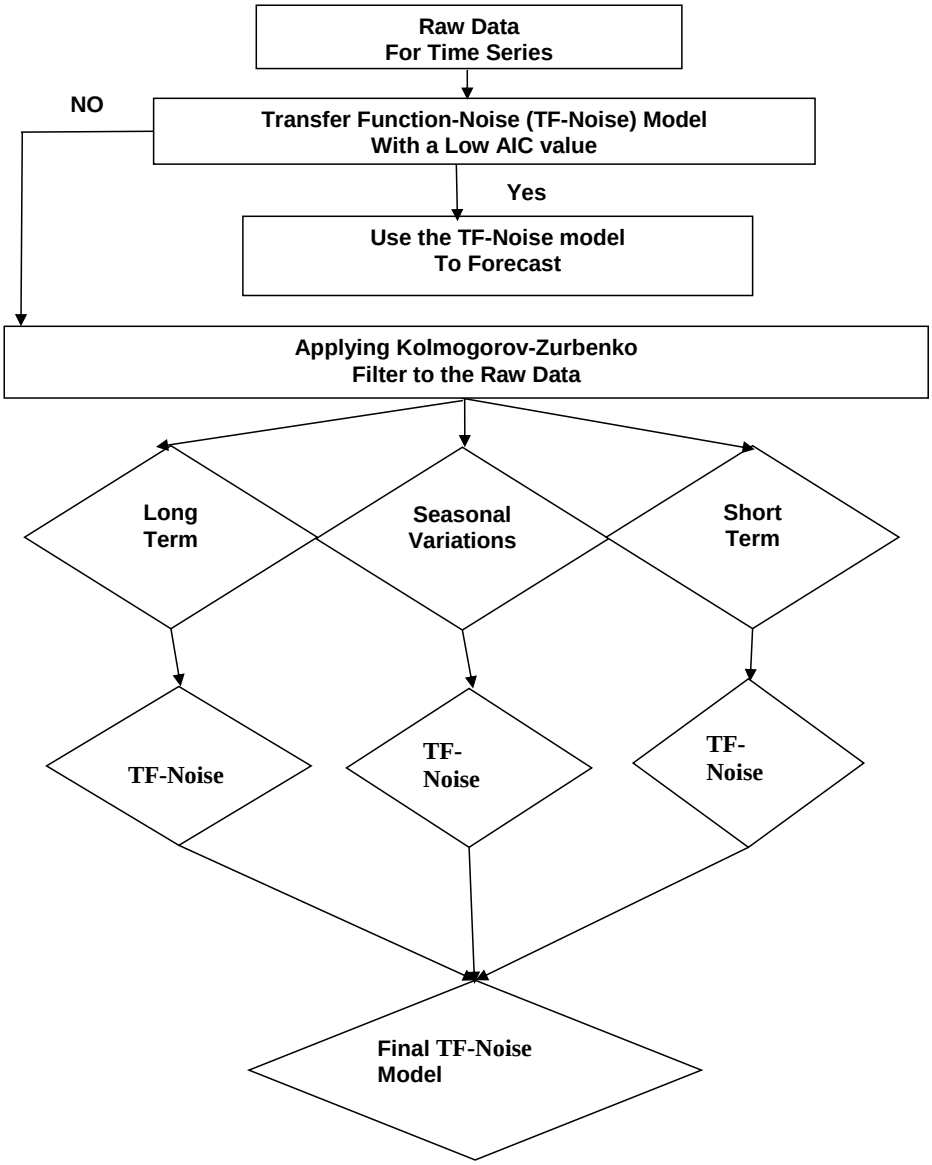

Figure A.3: Chart Illustrates the Steps Taken to Construct the Developed Models for Poughkeepsie City.

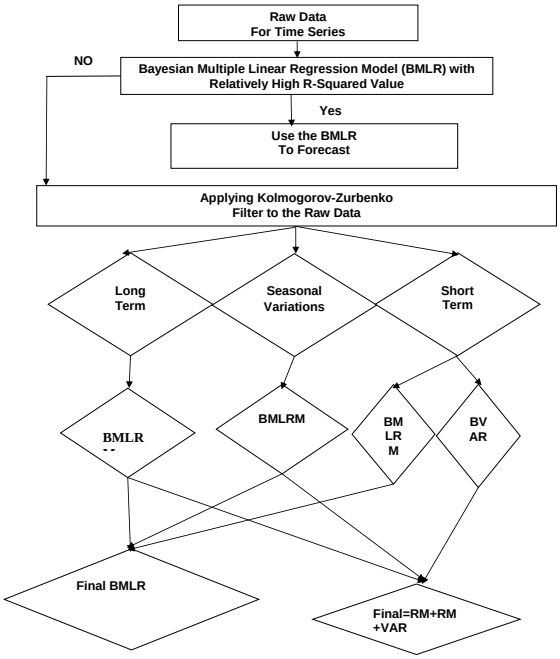

Figure A.4: Chart Illustrates the Steps Taken to Construct the Developed Models.

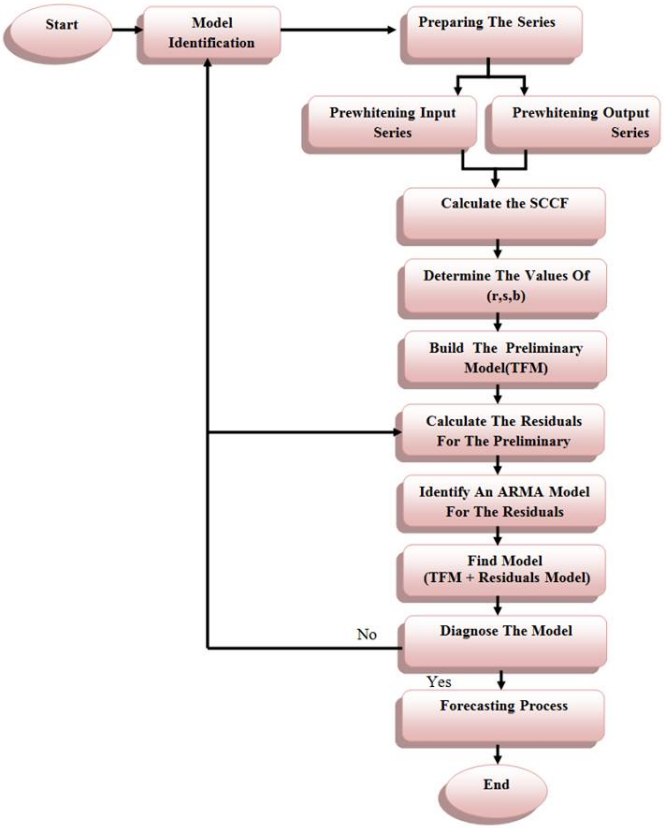

Figure A.5: Chart Illustrates the Steps Taken to Construct the Developed Models.

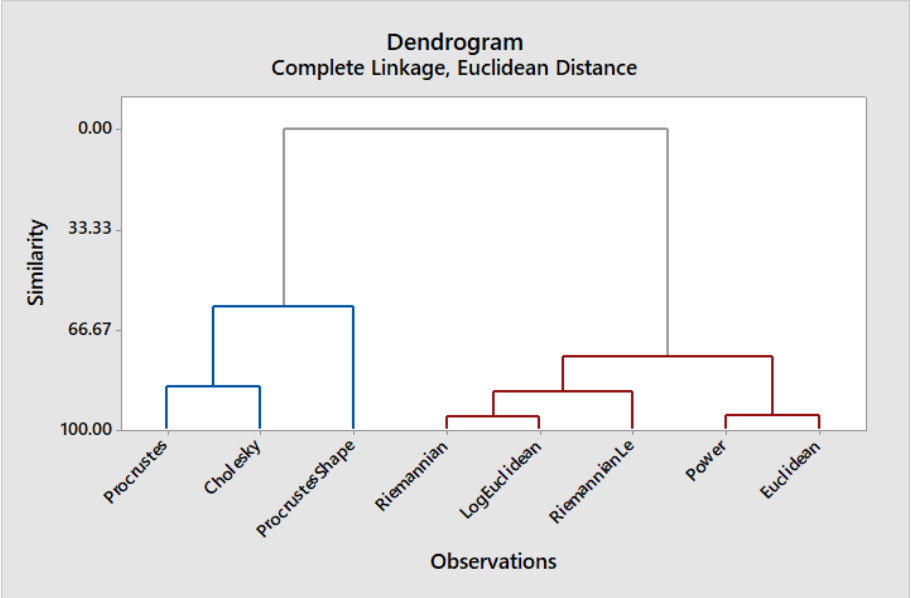

Figure A.6: The Dendrogram for the Long-Term Data for the Eight Distance Measures for Year 2005.

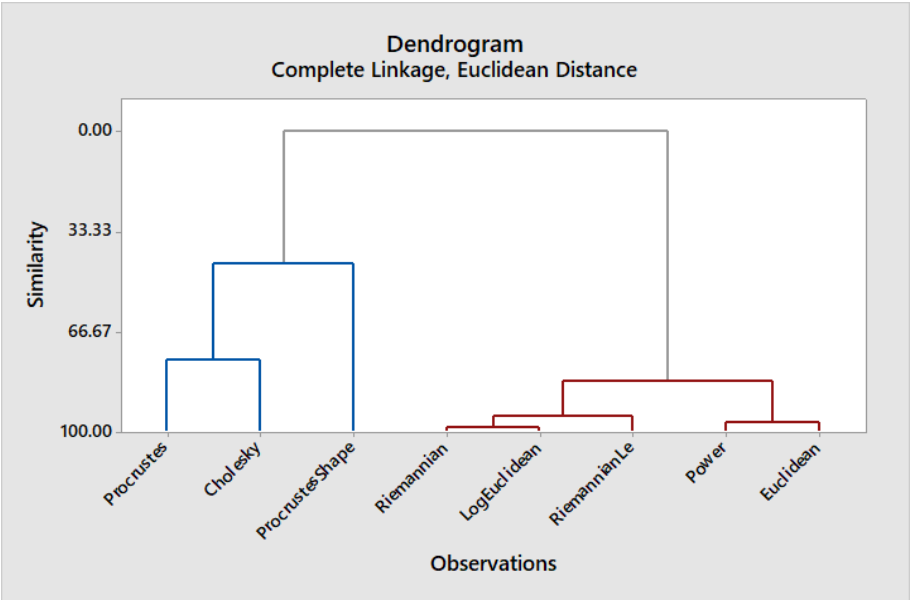

Figure A.7: The Dendrogram for the Seasonal Data for the Eight Distance Measures for Year 2005.